Linear Algebra with Applications

W. Keith Nicholson

Version
2023-A-D

WITH OPEN TEXTS

Open Texts and Editorial

Digital access to our high-quality texts is entirely FREE! All content is reviewed and updated regularly by subject matter experts, and the texts are adaptable; custom editions can be produced by the Vretta-Lyryx editorial team for those adopting Vretta-Lyryx assessment. Access to the original source files is also open to anyone!

Support

Access to our in-house support team is available 7 days/week to provide prompt resolution to both student and instructor inquiries. In addition, we work one-on-one with instructors to provide a comprehensive system, customized for their course. This can include managing multiple sections, assistance with preparing online examinations, help with LMS integration, and much more!

Online Assessment

Vretta-Lyryx has been developing online formative and summative assessment (homework and exams) for more than 20 years combined, and as with the textbook content, questions and problems are regularly reviewed and updated. Students receive immediate personalized feedback on their mistakes to guide their learning. Student grade reports and performance statistics are also provided, and full LMS integration, including gradebook synchronization, is available.

Instructor Supplements

Additional instructor resources are also available to users. Product dependent, these include: full sets of adaptable slides, solutions manuals, and multiple choice question banks with an exam building tool.

Linear Algebra with Applications

Open Edition

Distribution Version 2023 — Revision A-D

AUTHOR

W. Keith Nicholson, University of Calgary

CONTRIBUTIONS

Tim Alderson, University of New Brunswick
Daniel Brinkman, San Jose State University
Karl-Dieer Crisman, Gordon College
Benjamin E. Dozier, Cornell University
Jérôme Fortier, McGill University
David S. Gerstl, Farmingdale State College
Robert Gross, Boston College
Mark Fels, Utah State University
Chris Sangwin, University of Edinburgh
Alistair Savage, University of Ottawa
Csilla Tamas, Langara College
Vretta-Lyryx

Be a champion of OER!

Contribute suggestions for improvements, new content, or errata:

- A new topic
- A new example
- An interesting new question
- Any other suggestions to improve the material

Contact Vretta-Lyryx at info@lyryx.com with your ideas.

Linear Algebra with Applications

Open Edition

by W. Keith Nicholson — Distribution Version 2023 — Revision A-D

Attribution

To redistribute all of this book in its original form, please follow the guide below:

The front matter of the text should include a "License" page that includes the following statement.

> This text is *Linear Algebra with Applications* by W. Keith Nicholson and Vretta-Lyryx. View the text for free at `https://lyryx.com/linear-algebra-applications/`

To redistribute part of this book in its original form, please follow the guide below. Clearly indicate which content has been redistributed

The front matter of the text should include a "License" page that includes the following statement.

> This text includes the following content from *Linear Algebra with Applications* by W. Keith Nicholson and Vretta-Lyryx Inc. View the entire text for free at `https://lyryx.com/linear-algebra-applications/`.
> <List of content from the original text>

The following must also be included at the beginning of the applicable content. Please clearly indicate which content has been redistributed from the Vretta-Lyryx text.

> This chapter is redistributed from the original *Linear Algebra with Applications* by W. Keith Nicholson and Vretta-Lyryx Inc. View the original text for free at `https://lyryx.com/linear-algebra-applications/`.

To adapt and redistribute all or part of this book in its original form, please follow the guide below. Clearly indicate which content has been adapted/redistributed and summarize the changes made.

The front matter of the text should include a "License" page that includes the following statement.

> This text contains content adapted from the original *Linear Algebra with Applications* by W. Keith Nicholson and Vretta-Lyryx Inc. View the original text for free at `https://lyryx.com/linear-algebra-applications/`.
> <List of content and summary of changes>

The following must also be included at the beginning of the applicable content. Please clearly indicate which content has been adapted from the Vretta-Lyryx text.

> This chapter was adapted from the original *Linear Algebra with Applications* by W. Keith Nicholson and Vretta-Lyryx Inc. View the original text for free at `https://lyryx.com/linear-algebra-applications/`.

Citation

Use the information below to create a citation:

Author: W. Keith Nicholson
Contributing Author: Claude Laflamme
Publisher: Vretta-Lyryx Inc.
Book title: Linear Algebra with Applications
Book version: 2023-A-D
Publication date: July 1, 2023
Location: Toronto, Ontario, Canada
Book URL: `https://lyryx.com/linear-algebra-applications`

For questions or comments please contact editorial@lyryx.com

Linear Algebra with Applications

Open Edition

Revision History: Distribution Version 2023 — Revision A-D

Revision	Contributor	Significant Changes
2023 A	T. Alderson:	• An example of standard basis vectors in $\mathbb{R}^3$ added to Section 2.6.
	D. Brinkman:	• Section 3.3 has been split into three sections 3.4, 3.5 and 3.6.
	K-D. Crisman:	• The various definitions of standard basis have been clarified and consolidated.
		• The notation for points and vectors has been clarified.
	D.E. Dozier:	• Theorem 6.3.2 has been clarified.
		• The proof of Lemma 6.6.1 has been expanded.
		• The statement of Theorem 10.5.2 has been corrected.
	J. Fortier:	• The statement of Theorem 5.2.67 has been corrected and clarified.
	D. Gerstl:	• The diagram for Exercise 1.5.2 has been corrected.
		• Various typos have been corrected.
	R. Gross:	• Various typos have been corrected.
	D. Morris:	• Various typos have been corrected.
	C. Sangwin:	• Theorem 5.4.2 on rank and nullity has been expanded.
		• Various typos have been corrected.
2021 A		• Front matter has been updated including cover, Vretta-Lyryx with Open Texts, copyright, and revision pages. • An attribution page has been added. • Typo and other minor fixes have been implemented throughout.
2019 A		• New Section on Singular Value Decomposition (8.6) is included. • New Example 2.3.2 and Theorem 2.2.4. Please note that this will impact the numbering of subsequent examples and theorems in the relevant sections. • Section 2.2 is renamed as *Matrix-Vector Multiplication*. • Minor revisions made throughout, including fixing typos, adding exercises, expanding explanations, and other small edits.
2018 B		• Images have been converted to LaTeX tikz throughout. • Text has been converted to LaTeX with minor fixes throughout. • Full index has been implemented.
2018 A		• Text has been released with a Creative Commons license.

Table of Contents

Foreward

Mathematics education at the beginning university level is closely tied to the traditional publishers. In my opinion, it gives them too much control of both cost and content. The main goal of most publishers is profit, and the result has been a sales-driven business model as opposed to a pedagogical one. This results in frequent new "editions" of textbooks motivated largely to reduce the sale of used books rather than to update content quality. It also introduces copyright restrictions which stifle the creation and use of new pedagogical methods and materials. The overall result is high cost textbooks which may not meet the evolving educational needs of instructors and students.

To be fair, publishers do try to produce material that reflects new trends. But their goal is to sell books and not necessarily to create tools for student success in mathematics education. Sadly, this has led to a model where the primary choice for adapting to (or initiating) curriculum change is to find a different commercial textbook. My editor once said that the text that is adopted is often everyone's third choice.

Of course instructors can produce and maintain their own lecture notes, and have done so for years, but this remains an onerous task and difficult for others to benefit. The publishing industry arose from the need to provide authors with copy-editing, editorial, and marketing services, as well as extensive reviews of prospective customers to ascertain market trends and content updates. These are necessary skills and services that the industry continues to offer.

Authors of open educational resources (OER) including (but not limited to) textbooks and lecture notes, cannot afford this on their own. But they do have two great advantages: The cost to students is significantly lower, and open licenses return content control to instructors. Through editable file formats and open licenses, OER can be developed, maintained, reviewed, edited, and improved by a variety of contributors. Instructors can now respond to curriculum change by revising and reordering material to create content that meets the needs of their students. While editorial and quality control remain daunting tasks, great strides have been made in addressing the issues of accessibility, affordability and adaptability of the material.

For all the above reasons I have decided to release my text under an open license, even though it was published for many years through a traditional publisher.

However supporting students and instructors in a typical first year College or University classroom requires much more than a textbook. Thus, while anyone is welcome to use the distributed text at no cost, I also decided to work closely with colleagues at the University of Calgary and help create Lyryx Learning almost 20 years ago. The original idea was to develop quality but affordable formative online assessment and other educational software to assist students. Revenues are then used to sustain the project including editorial for the open textbook. Lyryx is now part of Vretta-Lyryx, and they continue working with authors, contributors, and reviewers to ensure instructors need not sacrifice quality and rigour when adopting an open text.

I believe this is the right direction for mathematical publishing going forward, and I look forward to being a part of how this new approach develops.

W. Keith Nicholson, Author
University of Calgary

Preface

This textbook is an introduction to the ideas and techniques of linear algebra for first- or second-year students with a working knowledge of high school algebra. The contents have enough flexibility to present a traditional introduction to the subject, or to allow for a more applied course. Chapters 1–4 contain a one-semester course for beginners whereas Chapters 5–9 contain a second semester course (see the Suggested Course Outlines below). The text is primarily about real linear algebra with complex numbers being mentioned when appropriate (reviewed in Appendix A). Overall, the aim of the text is to achieve a balance among computational skills, theory, and applications of linear algebra. Calculus is not a prerequisite; places where it is mentioned may be omitted.

As a rule, students of linear algebra learn by studying examples and solving problems. Accordingly, the book contains a variety of exercises (over 1200, many with multiple parts), ordered as to their difficulty. In addition, more than 375 solved examples are included in the text, many of which are computational in nature. The examples are also used to motivate (and illustrate) concepts and theorems, carrying the student from concrete to abstract. While the treatment is rigorous, proofs are presented at a level appropriate to the student and may be omitted with no loss of continuity. As a result, the book can be used to give a course that emphasizes computation and examples, or to give a more theoretical treatment (some longer proofs are deferred to the end of the Section).

Linear Algebra has application to the natural sciences, engineering, management, and the social sciences as well as mathematics. Consequently, 18 optional "applications" sections are included in the text introducing topics as diverse as electrical networks, economic models, Markov chains, linear recurrences, systems of differential equations, and linear codes over finite fields. Additionally some applications (for example linear dynamical systems, and directed graphs) are introduced in context. The applications sections appear at the end of the relevant chapters to encourage students to browse.

SUGGESTED COURSE OUTLINES

This text includes the basis for a two-semester course in linear algebra.

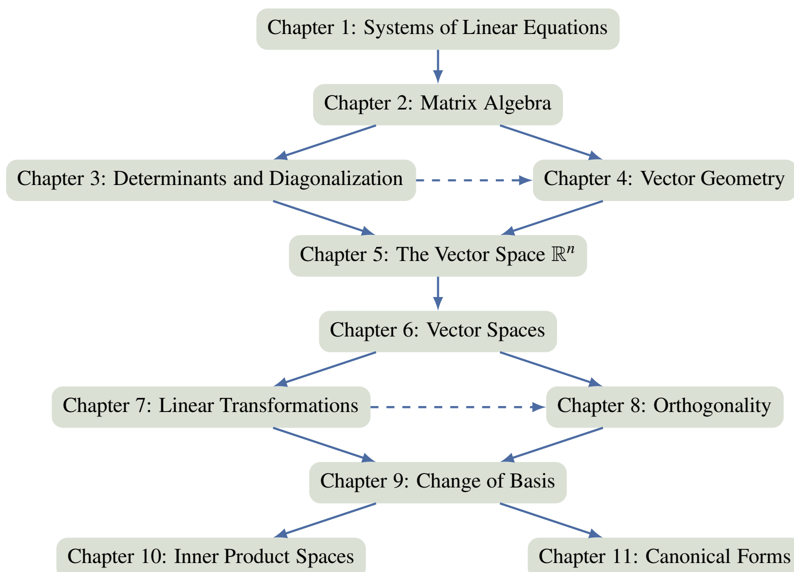
- Chapters 1–4 provide a standard one-semester course of 35 lectures, including linear equations, matrix algebra, determinants, diagonalization, and geometric vectors, with applications as time permits. At Calgary, we cover Sections 1.1–1.3, 2.1–2.6, 3.1–3.3, and 4.1–4.4 and the course is taken by all science and engineering students in their first semester. Prerequisites include a working knowledge of high school algebra (algebraic manipulations and some familiarity with polynomials); calculus is not required.

- Chapters 5–9 contain a second semester course including $\mathbb{R}^n$, abstract vector spaces, linear transformations (and their matrices), orthogonality, complex matrices (up to the spectral theorem) and applications. There is more material here than can be covered in one semester, and at Calgary we cover Sections 5.1–5.5, 6.1–6.4, 7.1–7.3, 8.1–8.7, and 9.1–9.3 with a couple of applications as time permits.

- Chapter 5 is a "bridging" chapter that introduces concepts like spanning, independence, and basis in the concrete setting of $\mathbb{R}^n$, before venturing into the abstract in Chapter 6. The duplication is balanced by the value of reviewing these notions, and it enables the student to focus in Chapter 6 on the new idea of an abstract system. Moreover, Chapter 5 completes the discussion of rank and diagonalization from earlier chapters, and includes a brief introduction to orthogonality in $\mathbb{R}^n$, which creates the possibility of a one-semester, matrix-oriented course covering Chapter 1–5 for students not wanting to study the abstract theory.

CHAPTER DEPENDENCIES

The following chart suggests how the material introduced in each chapter draws on concepts covered in certain earlier chapters. A solid arrow means that ready assimilation of ideas and techniques presented in the later chapter depends on familiarity with the earlier chapter. A broken arrow indicates that some reference to the earlier chapter is made but the chapter need not be covered.



HIGHLIGHTS OF THE TEXT

- **Two-stage definition of matrix multiplication.** First, in Section 2.2 matrix-vector products are introduced naturally by viewing the left side of a system of linear equations as a product. Second, matrix-matrix products are defined in Section 2.3 by taking the columns of a product AB to be A times the corresponding columns of B. This is motivated by viewing the matrix product as composition of maps (see next item). This works well pedagogically and the usual dot-product definition follows easily. As a bonus, the proof of associativity of matrix multiplication now takes four lines.

- **Matrices as transformations.** Matrix-column multiplications are viewed (in Section 2.2) as transformations $\mathbb{R}^n \to \mathbb{R}^m$. These maps are then used to describe simple geometric reflections and rotations in $\mathbb{R}^2$ as well as systems of linear equations.

- **Early linear transformations.** It has been said that vector spaces exist so that linear transformations can act on them—consequently these maps are a recurring theme in the text. Motivated by the matrix transformations introduced earlier, linear transformations $\mathbb{R}^n \to \mathbb{R}^m$ are defined in Section 2.6, their standard matrices are derived, and they are then used to describe rotations, reflections, projections, and other operators on $\mathbb{R}^2$.

- **Early diagonalization.** As requested by engineers and scientists, this important technique is presented in the first term using only determinants and matrix inverses (before defining independence and dimension). Applications to population growth and linear recurrences are given.

- **Early dynamical systems.** These are introduced in Chapter 3, and lead (via diagonalization) to applications like the possible extinction of species. Beginning students in science and engineering can relate to this because they can see (often for the first time) the relevance of the subject to the real world.

- **Bridging chapter.** Chapter 5 lets students deal with tough concepts (like independence, spanning, and basis) in the concrete setting of $\mathbb{R}^n$ before having to cope with abstract vector spaces in Chapter 6.

- **Examples.** The text contains over 375 worked examples, which present the main techniques of the subject, illustrate the central ideas, and are keyed to the exercises in each section.

- **Exercises.** The text contains a variety of exercises (nearly 1175, many with multiple parts), starting with computational problems and gradually progressing to more theoretical exercises. Select solutions are available at the end of the book or in the Student Solution Manual. There is a complete Solution Manual is available for instructors.

- **Applications**. There are optional applications at the end of most chapters (see the list below). While some are presented in the course of the text, most appear at the end of the relevant chapter to encourage students to browse.

- **Appendices.** Because complex numbers are needed in the text, they are described in Appendix A, which includes the polar form and roots of unity. Methods of proofs are discussed in Appendix B, followed by mathematical induction in Appendix C. A brief discussion of polynomials is included in Appendix D. All these topics are presented at the high-school level.

- **Self-Study.** This text is self-contained and therefore is suitable for self-study.

- **Rigour.** Proofs are presented as clearly as possible (some at the end of the section), but they are optional and the instructor can choose how much he or she wants to prove. However the proofs are there, so this text is more rigorous than most. Linear algebra provides one of the better venues where students begin to think logically and argue concisely. To this end, there are exercises that ask the student to "show" some simple implication, and others that ask her or him to either prove a given statement or give a counterexample. I personally present a few proofs in the first semester course and more in the second (see the Suggested Course Outlines).

- **Major Theorems.** Several major results are presented in the book. Examples: Uniqueness of the reduced row-echelon form; the cofactor expansion for determinants; the Cayley-Hamilton theorem; the Jordan canonical form; Schur's theorem on block triangular form; the principal axes and spectral theorems; and others. Proofs are included because the stronger students should at least be aware of what is involved.

CHAPTER SUMMARIES

Chapter 1: Systems of Linear Equations.

A standard treatment of gaussian elimination is given. The rank of a matrix is introduced via the row-echelon form, and solutions to a homogeneous system are presented as linear combinations of basic solutions. Applications to network flows, electrical networks, and chemical reactions are provided.

Chapter 2: Matrix Algebra.

After a traditional look at matrix addition, scalar multiplication, and transposition in Section 2.1, matrix-vector multiplication is introduced in Section 2.2 by viewing the left side of a system of linear equations as the product $A\mathbf{x}$ of the coefficient matrix A with the column $\mathbf{x}$ of variables. The usual dot-product definition of a matrix-vector multiplication follows. Section 2.2 ends by viewing an $m \times n$ matrix A as a transformation $\mathbb{R}^n \to \mathbb{R}^m$. This is illustrated for $\mathbb{R}^2 \to \mathbb{R}^2$ by describing reflection in the x axis, rotation of $\mathbb{R}^2$ through $\frac{\pi}{2}$, shears, and so on.

In Section 2.3, the product of matrices A and B is defined by $AB = \left[\begin{array}{cccc} A\mathbf{b}_1 & A\mathbf{b}_2 & \cdots & A\mathbf{b}_n \end{array} \right]$, where the $\mathbf{b}_i$ are the columns of B. A routine computation shows that this is the matrix of the transformation B followed by A. This observation is used frequently throughout the book, and leads to simple, conceptual proofs of the basic axioms of matrix algebra. Note that linearity is not required—all that is needed is some basic properties of matrix-vector multiplication developed in Section 2.2. Thus the usual arcane definition of matrix multiplication is split into two well motivated parts, each an important aspect of matrix algebra. Of course, this has the pedagogical advantage that the conceptual power of geometry can be invoked to illuminate and clarify algebraic techniques and definitions.

In Section 2.4 and 2.5 matrix inverses are characterized, their geometrical meaning is explored, and block multiplication is introduced, emphasizing those cases needed later in the book. Elementary matrices are discussed, and the Smith normal form is derived. Then in Section 2.6, linear transformations $\mathbb{R}^n \to \mathbb{R}^m$ are defined and shown to be matrix transformations. The matrices of reflections, rotations, and projections in the plane are determined. Finally, matrix multiplication is related to directed graphs, matrix LU-factorization is introduced, and applications to economic models and Markov chains are presented.

Chapter 3: Determinants and Diagonalization.

The cofactor expansion is stated (proved by induction later) and used to define determinants inductively and to deduce the basic rules. The product and adjugate theorems are proved. Then the diagonalization algorithm is presented (motivated by an example about the possible extinction of a species of birds). As requested by our Engineering Faculty, this is done earlier than in most texts because it requires only determinants and matrix inverses, avoiding any need for subspaces, independence and dimension. Eigenvectors of a 2×2 matrix A are described geometrically (using the A-invariance of lines through the origin). Diagonalization is then used to study discrete linear dynamical systems and to discuss applications to linear recurrences and systems of differential equations. A brief discussion of Google PageRank is included.

Chapter 4: Vector Geometry.

Vectors are presented intrinsically in terms of length and direction, and are related to matrices via coordinates. Then vector operations are defined using matrices and shown to be the same as the corresponding intrinsic definitions. Next, dot products and projections are introduced to solve problems about lines and planes. This leads to the cross product. Then matrix transformations are introduced in $\mathbb{R}^3$, matrices of projections and reflections are derived, and areas and volumes are computed using determinants. The chapter closes with an application to computer graphics.

Chapter 5: The Vector Space $\mathbb{R}^n$.

Subspaces, spanning, independence, and dimensions are introduced in the context of $\mathbb{R}^n$ in the first two sections. Orthogonal bases are introduced and used to derive the expansion theorem. The basic properties of rank are presented and used to justify the definition given in Section 1.2. Then, after a rigorous study of diagonalization, best approximation and least squares are discussed. The chapter closes with an application to correlation and variance.

This is a “bridging” chapter, easing the transition to abstract spaces. Concern about duplication with Chapter 6 is mitigated by the fact that this is the most difficult part of the course and many students welcome a repeat discussion of concepts like independence and spanning, albeit in the abstract setting. In a different direction, Chapter 1–5 could serve as a solid introduction to linear algebra for students not requiring abstract theory.

Chapter 6: Vector Spaces.

Building on the work on $\mathbb{R}^n$ in Chapter 5, the basic theory of abstract finite dimensional vector spaces is developed emphasizing new examples like matrices, polynomials and functions. This is the first acquaintance most students have had with an abstract system, so not having to deal with spanning, independence and dimension in the general context eases the transition to abstract thinking. Applications to polynomials and to differential equations are included.

Chapter 7: Linear Transformations.

General linear transformations are introduced, motivated by many examples from geometry, matrix theory, and calculus. Then kernels and images are defined, the dimension theorem is proved, and isomorphisms are discussed. The chapter ends with an application to linear recurrences. A proof is included that the order of a differential equation (with constant coefficients) equals the dimension of the space of solutions.

Chapter 8: Orthogonality.

The study of orthogonality in $\mathbb{R}^n$, begun in Chapter 5, is continued. Orthogonal complements and projections are defined and used to study orthogonal diagonalization. This leads to the principal axes theorem, the Cholesky factorization of a positive definite matrix, QR-factorization, and to a discussion of the singular value decomposition, the polar form, and the pseudoinverse. The theory is extended to $\mathbb{C}^n$ in Section 8.7 where hermitian and unitary matrices are discussed, culminating in Schur’s theorem and the spectral theorem. A short proof of the Cayley-Hamilton theorem is also presented. In Section 8.8 the field $\mathbb{Z}_p$ of integers modulo p is constructed informally for any prime p, and codes are discussed over any finite field. The chapter concludes with applications to quadratic forms, constrained optimization, and statistical principal component analysis.

Chapter 9: Change of Basis.

The matrix of general linear transformation is defined and studied. In the case of an operator, the relationship between basis changes and similarity is revealed. This is illustrated by computing the matrix of a rotation about a line through the origin in $\mathbb{R}^3$. Finally, invariant subspaces and direct sums are introduced, related to similarity, and (as an example) used to show that every involution is similar to a diagonal matrix with diagonal entries ± 1.

Chapter 10: Inner Product Spaces.

General inner products are introduced and distance, norms, and the Cauchy-Schwarz inequality are discussed. The Gram-Schmidt algorithm is presented, projections are defined and the approximation theorem is proved (with an application to Fourier approximation). Finally, isometries are characterized, and distance preserving operators are shown to be composites of a translations and isometries.

Chapter 11: Canonical Forms.

The work in Chapter 9 is continued. Invariant subspaces and direct sums are used to derive the block triangular form. That, in turn, is used to give a compact proof of the Jordan canonical form. Of course the level is higher.

Appendices

In Appendix A, complex arithmetic is developed far enough to find nth roots. In Appendix B, methods of proof are discussed, while Appendix C presents mathematical induction. Finally, Appendix D describes the properties of polynomials in elementary terms.

LIST OF APPLICATIONS

- Network Flow (Section 1.4)
- Electrical Networks (Section 1.5)
- Chemical Reactions (Section 1.6)
- Directed Graphs (in Section 2.3)
- Input-Output Economic Models (Section 2.8)
- Markov Chains (Section 2.9)
- Polynomial Interpolation (in Section 3.2)
- Population Growth (Examples 3.3.1 and 3.5.1, Section 3.3)
- Google PageRank (in Section 3.3)
- Linear Recurrences (Section 3.4; see also Section 7.5)
- Systems of Differential Equations (Section 3.5)
- Computer Graphics (Section 4.5)

- Least Squares Approximation (in Section 5.6)
- Correlation and Variance (Section 5.7)
- Polynomials (Section 6.5)
- Differential Equations (Section 6.6)
- Linear Recurrences (Section 7.5)
- Error Correcting Codes (Section 8.8)
- Quadratic Forms (Section 8.9)
- Constrained Optimization (Section 8.10)
- Statistical Principal Component Analysis (Section 8.11)
- Fourier Approximation (Section 10.5)

ACKNOWLEDGMENTS

It is also a pleasure to recognize the contributions of several people over the many years all the way since the early days of this text. Now that the text has an open license, we have a much more fluid and powerful mechanism to incorporate comments and suggestions.

The editorial group at Vretta-Lyryx invites instructors and students to contribute to the text, and we will post revisions and credits in a separate revision page.

W. Keith Nicholson
University of Calgary

Chapter 1

Systems of Linear Equations

1.1 Solutions and Elementary Operations

Practical problems in many fields of study—such as biology, business, chemistry, computer science, economics, electronics, engineering, physics and the social sciences—can often be reduced to solving a system of linear equations. Linear algebra arose from attempts to find systematic methods for solving these systems, so it is natural to begin this book by studying linear equations.

If a, b, and c are real numbers, the graph of an equation of the form

$$ax+by=c$$

is a straight line (if a and b are not both zero), so such an equation is called a *linear* equation in the variables x and y. However, it is often convenient to write the variables as $x_1, x_2, \dots, x_n$, particularly when more than two variables are involved. An equation of the form

$$a_1x_1+a_2x_2+\cdots+a_nx_n=b$$

is called a **linear equation** in the n variables $x_1, x_2, \dots, x_n$. Here $a_1, a_2, \dots, a_n$ denote real numbers (called the **coefficients** of $x_1, x_2, \dots, x_n$, respectively) and b is also a number (called the **constant term** of the equation). A finite collection of linear equations in the variables $x_1, x_2, \dots, x_n$ is called a **system of linear equations** in these variables. Hence,

$$2x_1-3x_2+5x_3=7$$

is a linear equation; the coefficients of x_1, x_2, and x_3 are 2, -3, and 5, and the constant term is 7. Note that each variable in a linear equation occurs to the first power only.

Given a linear equation $a_1x_1+a_2x_2+\cdots+a_nx_n=b$, a sequence $s_1, s_2, \dots, s_n$ of n numbers is called a **solution** to the equation if

$$a_1s_1+a_2s_2+\cdots+a_ns_n=b$$

that is, if the equation is satisfied when the substitutions $x_1=s_1, x_2=s_2, \dots, x_n=s_n$ are made. A sequence of numbers is called **a solution to a system** of equations if it is a solution to every equation in the system.

For example, $x=-2$, $y=5$, $z=0$ and $x=0$, $y=4$, $z=-1$ are both solutions to the system

$$\begin{aligned} x+y+\ \ z&=3\\ 2x+y+3z&=1 \end{aligned}$$

A system may have no solution at all, or it may have a unique solution, or it may have an infinite family of solutions. For instance, the system $x+y=2$, $x+y=3$ has no solution because the sum of two numbers cannot be 2 and 3 simultaneously. A system that has no solution is called **inconsistent**; a system with at least one solution is called **consistent**. The system in the following example has infinitely many solutions.

Example 1.1.1

Show that, for arbitrary values of s and t,

$$\begin{aligned} x_1 &= t - s + 1 \\ x_2 &= t + s + 2 \\ x_3 &= s \\ x_4 &= t \end{aligned}$$

is a solution to the system

$$\begin{array}{rcrcrcrcr} x_1 & - & 2x_2 & + & 3x_3 & + & x_4 & = & -3 \\ 2x_1 & - & x_2 & + & 3x_3 & - & x_4 & = & 0 \end{array}$$

Solution. Simply substitute these values of x_1, x_2, x_3, and x_4 in each equation.

$$\begin{aligned} x_1 - 2x_2 + 3x_3 + x_4 &= (t - s + 1) - 2(t + s + 2) + 3s + t = -3 \\ 2x_1 - x_2 + 3x_3 - x_4 &= 2(t - s + 1) - (t + s + 2) + 3s - t = 0 \end{aligned}$$

Because both equations are satisfied, it is a solution for all choices of s and t.

The quantities s and t in Example 1.1.1 are called **parameters**, and the set of solutions, described in this way, is said to be given in **parametric form** and is called the **general solution** to the system. It turns out that the solutions to *every* system of equations (if there *are* solutions) can be given in parametric form (that is, the variables x_1, x_2, $\dots$ are given in terms of new independent variables s, t, etc.). The following example shows how this happens in the simplest systems where only one equation is present.

Example 1.1.2

Describe all solutions to $3x - y + 2z = 6$ in parametric form.

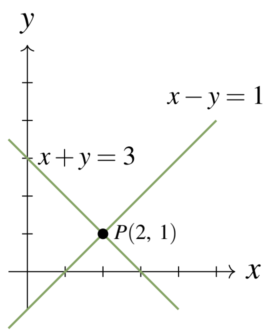
Solution. Solving the equation for y in terms of x and z, we get $y = 3x + 2z - 6$. If s and t are arbitrary then, setting $x = s$, $z = t$, we get solutions

$$\begin{aligned} x &= s \\ y &= 3s + 2t - 6 \qquad s \text{ and } t \text{ arbitrary} \\ z &= t \end{aligned}$$

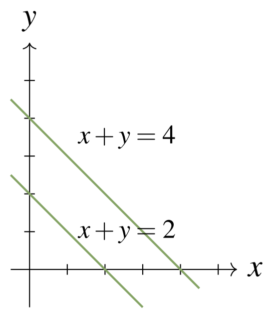
Of course we could have solved for x: $x = \frac{1}{3}(y - 2z + 6)$. Then, if we take $y = p$, $z = q$, the solutions are represented as follows:

$$\begin{array}{rcl} x & = & \frac{1}{3}(p - 2q + 6) \\ y & = & p \qquad\qquad\qquad p \text{ and } q \text{ arbitrary} \\ z & = & q \end{array}$$

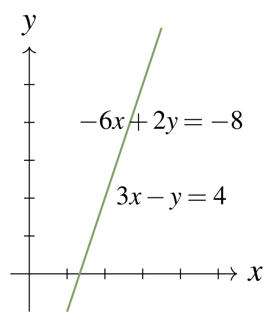
The same family of solutions can "look" quite different!



(a) Unique Solution
$(x=2,\, y=1)$



(b) No Solution



(c) Infinitely many solutions
$(x=t,\, y=3t-4)$

Figure 1.1.1

When only two variables are involved, the solutions to systems of linear equations can be described geometrically because the graph of a linear equation $ax+by=c$ is a straight line if a and b are not both zero. Moreover, a point $P(s,\ t)$ with coordinates s and t lies on the line if and only if $as+bt=c$—that is when $x=s$, $y=t$ is a solution to the equation. Hence the solutions to a *system* of linear equations correspond to the points $P(s,\ t)$ that lie on *all* the lines in question.

In particular, if the system consists of just one equation, there must be infinitely many solutions because there are infinitely many points on a line. If the system has two equations, there are three possibilities for the corresponding straight lines:

1. *The lines intersect at a single point. Then the system has a* unique solution *corresponding to that point.*
2. *The lines are parallel (and distinct) and so do not intersect. Then the system has* no solution.
3. *The lines are identical. Then the system has* infinitely many solutions*—one for each point on the (common) line.*

These three situations are illustrated in Figure 1.1.1. In each case the graphs of two specific lines are plotted and the corresponding equations are indicated. In the last case, the equations are $3x-y=4$ and $-6x+2y=-8$, which have identical graphs.

With three variables, the graph of an equation $ax+by+cz=d$ can be shown to be a plane (see Section 4.2) and so again provides a "picture" of the set of solutions. However, this graphical method has its limitations: When more than three variables are involved, no physical image of the graphs (called hyperplanes) is possible. It is necessary to turn to a more "algebraic" method of solution.

Before describing the method, we introduce a concept that simplifies the computations involved. Consider the following system

$$\begin{array}{rcrcrcrcr}
3x_1 & + & 2x_2 & - & x_3 & + & x_4 & = & -1 \\
2x_1 & & & - & x_3 & + & 2x_4 & = & 0 \\
3x_1 & + & x_2 & + & 2x_3 & + & 5x_4 & = & 2
\end{array}$$

of three equations in four variables. The array of numbers[1]

$$\left[\begin{array}{rrrr|r}
3 & 2 & -1 & 1 & -1 \\
2 & 0 & -1 & 2 & 0 \\
3 & 1 & 2 & 5 & 2
\end{array}\right]$$

occurring in the system is called the **augmented matrix** of the system. Each row of the matrix consists of the coefficients of the variables (in order) from the corresponding equation, together with the constant

[1] A rectangular array of numbers is called a **matrix**. Matrices will be discussed in more detail in Chapter 2.

term. For clarity, the constants are separated by a vertical line. The augmented matrix is just a different way of describing the system of equations. The array of coefficients of the variables

$$\left[\begin{array}{rrrr} 3 & 2 & -1 & 1 \\ 2 & 0 & -1 & 2 \\ 3 & 1 & 2 & 5 \end{array}\right]$$

is called the **coefficient matrix** of the system and $\left[\begin{array}{r} -1 \\ 0 \\ 2 \end{array}\right]$ is called the **constant matrix** of the system.

Elementary Operations

The algebraic method for solving systems of linear equations is described as follows. Two such systems are said to be **equivalent** if they have the same set of solutions. A system is solved by writing a series of systems, one after the other, each equivalent to the previous system. Each of these systems has the same set of solutions as the original one; the aim is to end up with a system that is easy to solve. Each system in the series is obtained from the preceding system by a simple manipulation chosen so that it does not change the set of solutions.

As an illustration, we solve the system $x+2y=-2$, $2x+y=7$ in this manner. At each stage, the corresponding augmented matrix is displayed. The original system is

$$\begin{array}{rcr} x+2y & = & -2 \\ 2x+y & = & 7 \end{array} \qquad \left[\begin{array}{rr|r} 1 & 2 & -2 \\ 2 & 1 & 7 \end{array}\right]$$

First, subtract twice the first equation from the second. The resulting system is

$$\begin{array}{rcr} x+2y & = & -2 \\ -3y & = & 11 \end{array} \qquad \left[\begin{array}{rr|r} 1 & 2 & -2 \\ 0 & -3 & 11 \end{array}\right]$$

which is equivalent to the original (see Theorem 1.1.1). At this stage we obtain $y=-\frac{11}{3}$ by multiplying the second equation by $-\frac{1}{3}$. The result is the equivalent system

$$\begin{array}{rcr} x+2y & = & -2 \\ y & = & -\frac{11}{3} \end{array} \qquad \left[\begin{array}{rr|r} 1 & 2 & -2 \\ 0 & 1 & -\frac{11}{3} \end{array}\right]$$

Finally, we subtract twice the second equation from the first to get another equivalent system.

$$\begin{array}{rcr} x & = & \frac{16}{3} \\ y & = & -\frac{11}{3} \end{array} \qquad \left[\begin{array}{rr|r} 1 & 0 & \frac{16}{3} \\ 0 & 1 & -\frac{11}{3} \end{array}\right]$$

Now *this* system is easy to solve! And because it is equivalent to the original system, it provides the solution to that system.

Observe that, at each stage, a certain operation is performed on the system (and thus on the augmented matrix) to produce an equivalent system.

Definition 1.1 Elementary Operations

The following operations, called ***elementary operations****, can routinely be performed on systems of linear equations to produce equivalent systems.*

I. *Interchange two equations.*

II. *Multiply one equation by a* nonzero *number.*

III. *Add a multiple of one equation to a* different *equation.*

Theorem 1.1.1

Suppose that a sequence of elementary operations is performed on a system of linear equations. Then the resulting system has the same set of solutions as the original, so the two systems are equivalent.

The proof is given at the end of this section.

Elementary operations performed on a system of equations produce corresponding manipulations of the *rows* of the augmented matrix. Thus, multiplying a row of a matrix by a number k means multiplying *every entry* of the row by k. Adding one row to another row means adding *each entry* of that row to the corresponding entry of the other row. Subtracting two rows is done similarly. Note that we regard two rows as equal when corresponding entries are the same.

In hand calculations (and in computer programs) we manipulate the rows of the augmented matrix rather than the equations. For this reason we restate these elementary operations for matrices.

Definition 1.2 Elementary Row Operations

The following are called ***elementary row operations*** *on a matrix.*

I. *Interchange two rows.*

II. *Multiply one row by a* nonzero *number.*

III. *Add a multiple of one row to a* different *row.*

In the illustration above, a series of such operations led to a matrix of the form

$$\left[\begin{array}{cc|c} 1 & 0 & * \\ 0 & 1 & * \end{array}\right]$$

where the asterisks represent arbitrary numbers. In the case of three equations in three variables, the goal is to produce a matrix of the form

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & * \\ 0 & 1 & 0 & * \\ 0 & 0 & 1 & * \end{array}\right]$$

This does not always happen, as we will see in the next section. Here is an example in which it does happen.

Example 1.1.3

Find all solutions to the following system of equations.

$$\begin{aligned} 3x + 4y + z &= 1 \\ 2x + 3y &= 0 \\ 4x + 3y - z &= -2 \end{aligned}$$

Solution. The augmented matrix of the original system is

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

To create a 1 in the upper left corner we could multiply row 1 through by $\frac{1}{3}$. However, the 1 can be obtained without introducing fractions by subtracting row 2 from row 1. The result is

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

The upper left 1 is now used to "clean up" the first column, that is create zeros in the other positions in that column. First subtract 2 times row 1 from row 2 to obtain

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

Next subtract 4 times row 1 from row 3. The result is

$$\left[\begin{array}{rrr|r} 1 & 1 & 1 & 1 \\ 0 & 1 & -2 & -2 \\ 0 & -1 & -5 & -6 \end{array}\right]$$

This completes the work on column 1. We now use the 1 in the second position of the second row to clean up the second column by subtracting row 2 from row 1 and then adding row 2 to row 3. For convenience, both row operations are done in one step. The result is

$$\left[\begin{array}{rrr|r} 1 & 0 & 3 & 3 \\ 0 & 1 & -2 & -2 \\ 0 & 0 & -7 & -8 \end{array}\right]$$

Note that the last two manipulations *did not affect* the first column (the second row has a zero there), so our previous effort there has not been undermined. Finally we clean up the third column. Begin by multiplying row 3 by $-\frac{1}{7}$ to obtain

$$\left[\begin{array}{rrr|r} 1 & 0 & 3 & 3 \\ 0 & 1 & -2 & -2 \\ 0 & 0 & 1 & \frac{8}{7} \end{array}\right]$$

Now subtract 3 times row 3 from row 1, and then add 2 times row 3 to row 2 to get

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & -\frac{3}{7} \\ 0 & 1 & 0 & \frac{2}{7} \\ 0 & 0 & 1 & \frac{8}{7} \end{array}\right]$$

The corresponding equations are $x = -\frac{3}{7}$, $y = \frac{2}{7}$, and $z = \frac{8}{7}$, which give the (unique) solution.

Every elementary row operation can be **reversed** by another elementary row operation of the same type (called its **inverse**). To see how, we look at types I, II, and III separately:

Type I *Interchanging two rows is reversed by interchanging them again.*

Type II *Multiplying a row by a nonzero number k is reversed by multiplying by $1/k$.*

Type III *Adding k times row p to a different row q is reversed by adding $-k$ times row p to row q (in the new matrix). Note that $p \neq q$ is essential here.*

To illustrate the Type III situation, suppose there are four rows in the original matrix, denoted R_1, R_2, R_3, and R_4, and that k times R_2 is added to R_3. Then the reverse operation adds $-k$ times R_2, to R_3. The following diagram illustrates the effect of doing the operation first and then the reverse:

$$\left[\begin{array}{c} R_1 \\ R_2 \\ R_3 \\ R_4 \end{array}\right] \rightarrow \left[\begin{array}{c} R_1 \\ R_2 \\ R_3 + kR_2 \\ R_4 \end{array}\right] \rightarrow \left[\begin{array}{c} R_1 \\ R_2 \\ (R_3 + kR_2) - kR_2 \\ R_4 \end{array}\right] = \left[\begin{array}{c} R_1 \\ R_2 \\ R_3 \\ R_4 \end{array}\right]$$

The existence of inverses for elementary row operations and hence for elementary operations on a system of equations, gives:

Proof of Theorem 1.1.1. Suppose that a system of linear equations is transformed into a new system by a sequence of elementary operations. Then every solution of the original system is automatically a solution of the new system because adding equations, or multiplying an equation by a nonzero number, always results in a valid equation. In the same way, each solution of the new system must be a solution to the original system because the original system can be obtained from the new one by another series of elementary operations (the inverses of the originals). It follows that the original and new systems have the same solutions. This proves Theorem 1.1.1. □

1.2 Gaussian Elimination

The algebraic method introduced in the preceding section can be summarized as follows: Given a system of linear equations, use a sequence of elementary row operations to carry the augmented matrix to a "nice" matrix (meaning that the corresponding equations are easy to solve). In Example 1.1.3, this nice matrix took the form

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & * \\ 0 & 1 & 0 & * \\ 0 & 0 & 1 & * \end{array}\right]$$

The following definitions identify the nice matrices that arise in this process.

Definition 1.3 Row-Echelon Form (Reduced)

A matrix is said to be in ***row-echelon form*** *(and will be called a* ***row-echelon matrix****) if it satisfies the following three conditions:*

1. *All* ***zero rows*** *(consisting entirely of zeros) are at the bottom.*
2. *The first nonzero entry from the left in each nonzero row is a* 1*, called the* ***leading 1*** *for that row.*
3. *Each leading* 1 *is to the right of all leading* 1*s in the rows above it.*

A row-echelon matrix is said to be in ***reduced row-echelon form*** *(and will be called a* ***reduced row-echelon matrix****) if, in addition, it satisfies the following condition:*

4. *Each leading* 1 *is the only nonzero entry in its column.*

The row-echelon matrices have a "staircase" form, as indicated by the following example (the asterisks indicate arbitrary numbers).

$$\left[\begin{array}{c|cccccc} 0 & 1 & * & * & * & * & * \\ 0 & 0 & 0 & 1 & * & * & * \\ 0 & 0 & 0 & 0 & 1 & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

The leading 1s proceed "down and to the right" through the matrix. Entries above and to the right of the leading 1s are arbitrary, but all entries below and to the left of them are zero. Hence, a matrix in row-echelon form is in reduced form if, in addition, the entries directly above each leading 1 are all zero. Note that a matrix in row-echelon form can, with a few more row operations, be carried to reduced form (use row operations to create zeros above each leading one in succession, beginning from the right).

Example 1.2.1

The following matrices are in row-echelon form (for any choice of numbers in $*$-positions).

$$\left[\begin{array}{ccc} 1 & * & * \\ 0 & 0 & 1 \end{array}\right] \left[\begin{array}{cccc} 0 & 1 & * & * \\ 0 & 0 & 1 & * \\ 0 & 0 & 0 & 0 \end{array}\right] \left[\begin{array}{cccc} 1 & * & * & * \\ 0 & 1 & * & * \\ 0 & 0 & 0 & 1 \end{array}\right] \left[\begin{array}{ccc} 1 & * & * \\ 0 & 1 & * \\ 0 & 0 & 1 \end{array}\right]$$

The following, on the other hand, are in reduced row-echelon form.

$$\left[\begin{array}{ccc} 1 & * & 0 \\ 0 & 0 & 1 \end{array}\right] \left[\begin{array}{cccc} 0 & 1 & 0 & * \\ 0 & 0 & 1 & * \\ 0 & 0 & 0 & 0 \end{array}\right] \left[\begin{array}{cccc} 1 & 0 & * & 0 \\ 0 & 1 & * & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right]$$

The choice of the positions for the leading 1s determines the (reduced) row-echelon form (apart from the numbers in $*$-positions).

The importance of row-echelon matrices comes from the following theorem.

Theorem 1.2.1

Every matrix can be brought to (reduced) row-echelon form by a sequence of elementary row operations.

In fact we can give a step-by-step procedure for actually finding a row-echelon matrix. Observe that while there are many sequences of row operations that will bring a matrix to row-echelon form, the one we use is systematic and is easy to program on a computer. Note that the algorithm deals with matrices in general, possibly with columns of zeros.

Theorem: Gaussian[2] Algorithm[3]

Step 1. If the matrix consists entirely of zeros, stop—it is already in row-echelon form.

Step 2. Otherwise, find the first column from the left containing a nonzero entry (call it a), and move the row containing that entry to the top position.

Step 3. Now multiply the new top row by $1/a$ to create a leading 1.

Step 4. By subtracting multiples of that row from rows below it, make each entry below the leading 1 zero.

This completes the first row, and all further row operations are carried out on the remaining rows.

Step 5. Repeat steps 1–4 on the matrix consisting of the remaining rows.

The process stops when either no rows remain at step 5 or the remaining rows consist entirely of zeros.

Observe that the gaussian algorithm is recursive: When the first leading 1 has been obtained, the procedure is repeated on the remaining rows of the matrix. This makes the algorithm easy to use on a computer. Note that the solution to Example 1.1.3 did not use the gaussian algorithm as written because the first leading 1 was not created by dividing row 1 by 3. The reason for this is that it avoids fractions. However, the general pattern is clear: Create the leading 1s from left to right, using each of them in turn to create zeros below it. Here are two more examples.

Example 1.2.2

Solve the following system of equations.

$$\begin{array}{rcrcrcr} 3x & + & y & - & 4z & = & -1 \\ x & & & + & 10z & = & 5 \\ 4x & + & y & + & 6z & = & 1 \end{array}$$

Solution. The corresponding augmented matrix is

$$\left[\begin{array}{rrr|r} 3 & 1 & -4 & -1 \\ 1 & 0 & 10 & 5 \\ 4 & 1 & 6 & 1 \end{array}\right]$$

Create the first leading one by interchanging rows 1 and 2

$$\left[\begin{array}{rrr|r} 1 & 0 & 10 & 5 \\ 3 & 1 & -4 & -1 \\ 4 & 1 & 6 & 1 \end{array}\right]$$

[2] Carl Friedrich Gauss (1777–1855) ranks with Archimedes and Newton as one of the three greatest mathematicians of all time. He was a child prodigy and, at the age of 21, he gave the first proof that every polynomial has a complex root. In 1801 he published a timeless masterpiece, *Disquisitiones Arithmeticae*, in which he founded modern number theory. He went on to make ground-breaking contributions to nearly every branch of mathematics, often well before others rediscovered and published the results.

[3] The algorithm was known to the ancient Chinese.

Now subtract 3 times row 1 from row 2, and subtract 4 times row 1 from row 3. The result is

$$\left[\begin{array}{rrr|r} 1 & 0 & 10 & 5 \\ 0 & 1 & -34 & -16 \\ 0 & 1 & -34 & -19 \end{array}\right]$$

Now subtract row 2 from row 3 to obtain

$$\left[\begin{array}{rrr|r} 1 & 0 & 10 & 5 \\ 0 & 1 & -34 & -16 \\ 0 & 0 & 0 & -3 \end{array}\right]$$

This means that the following reduced system of equations

$$\begin{aligned} x \quad + 10z &= 5 \\ y - 34z &= -16 \\ 0 &= {-3} \end{aligned}$$

is equivalent to the original system. In other words, the two have the same solutions. But this last system clearly has no solution (the last equation requires that x, y and z satisfy $0x + 0y + 0z = -3$, and no such numbers exist). Hence the original system has no solution.

Example 1.2.3

Solve the following system of equations.

$$\begin{aligned} x_1 - 2x_2 - x_3 + 3x_4 &= 1 \\ 2x_1 - 4x_2 + x_3 &= 5 \\ x_1 - 2x_2 + 2x_3 - 3x_4 &= 4 \end{aligned}$$

Solution. The augmented matrix is

$$\left[\begin{array}{rrrr|r} 1 & -2 & -1 & 3 & 1 \\ 2 & -4 & 1 & 0 & 5 \\ 1 & -2 & 2 & -3 & 4 \end{array}\right]$$

Subtracting twice row 1 from row 2 and subtracting row 1 from row 3 gives

$$\left[\begin{array}{rrrr|r} 1 & -2 & -1 & 3 & 1 \\ 0 & 0 & 3 & -6 & 3 \\ 0 & 0 & 3 & -6 & 3 \end{array}\right]$$

Now subtract row 2 from row 3 and multiply row 2 by $\frac{1}{3}$ to get

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

This is in row-echelon form, and we take it to reduced form by adding row 2 to row 1:

$$\left[\begin{array}{rrrr|r} 1 & -2 & 0 & 1 & 2 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

The corresponding reduced system of equations is

$$\begin{aligned} x_1 - 2x_2 \quad + \quad x_4 &= 2 \\ x_3 - 2x_4 &= 1 \\ 0 &= 0 \end{aligned}$$

The leading ones are in columns 1 and 3 here, so the corresponding variables x_1 and x_3 are called leading variables. Because the matrix is in reduced row-echelon form, these equations can be used to solve for the leading variables in terms of the nonleading variables x_2 and x_4. More precisely, in the present example we set $x_2 = s$ and $x_4 = t$ where s and t are arbitrary, so these equations become

$$x_1 - 2s + t = 2 \quad \text{and} \quad x_3 - 2t = 1$$

Finally the solutions are given by

$$\begin{aligned} x_1 &= 2 + 2s - t \\ x_2 &= s \\ x_3 &= 1 + 2t \\ x_4 &= t \end{aligned}$$

where s and t are arbitrary.

The solution of Example 1.2.3 is typical of the general case. To solve a linear system, the augmented matrix is carried to reduced row-echelon form, and the variables corresponding to the leading ones are called **leading variables**. Because the matrix is in reduced form, each leading variable occurs in exactly one equation, so that equation can be solved to give a formula for the leading variable in terms of the nonleading variables. It is customary to call the nonleading variables "free" variables, and to label them by new variables s, t, $\dots$, called **parameters**. Hence, as in Example 1.2.3, every variable x_i is given by a formula in terms of the parameters s and t. Moreover, every choice of these parameters leads to a solution to the system, and every solution arises in this way. This procedure works in general, and has come to be called

Theorem: Gaussian Elimination

To solve a system of linear equations proceed as follows:

1. *Carry the augmented matrix to a reduced row-echelon matrix using elementary row operations.*
2. *If a row* $\left[\begin{array}{cccccc} 0 & 0 & 0 & \cdots & 0 & 1 \end{array}\right]$ *occurs, the system is inconsistent.*
3. *Otherwise, assign the nonleading variables (if any) as parameters, and use the equations*

corresponding to the reduced row-echelon matrix to solve for the leading variables in terms of the parameters.

There is a variant of this procedure, wherein the augmented matrix is carried only to row-echelon form. The nonleading variables are assigned as parameters as before. Then the last equation (corresponding to the row-echelon form) is used to solve for the last leading variable in terms of the parameters. This last leading variable is then substituted into all the preceding equations. Then, the second last equation yields the second last leading variable, which is also substituted back. The process continues to give the general solution. This procedure is called **back-substitution**. This procedure can be shown to be numerically more efficient and so is important when solving very large systems.[4]

Example 1.2.4

Find a condition on the numbers a, b, and c such that the following system of equations is consistent. When that condition is satisfied, find all solutions (in terms of a, b, and c).

$$\begin{aligned} x_1 + 3x_2 + x_3 &= a \\ -x_1 - 2x_2 + x_3 &= b \\ 3x_1 + 7x_2 - x_3 &= c \end{aligned}$$

Solution. We use gaussian elimination except that now the augmented matrix

$$\left[\begin{array}{rrr|c} 1 & 3 & 1 & a \\ -1 & -2 & 1 & b \\ 3 & 7 & -1 & c \end{array}\right]$$

has entries a, b, and c as well as known numbers. The first leading one is in place, so we create zeros below it in column 1:

$$\left[\begin{array}{rrr|c} 1 & 3 & 1 & a \\ 0 & 1 & 2 & a+b \\ 0 & -2 & -4 & c-3a \end{array}\right]$$

The second leading 1 has appeared, so use it to create zeros in the rest of column 2:

$$\left[\begin{array}{rrr|c} 1 & 0 & -5 & -2a-3b \\ 0 & 1 & 2 & a+b \\ 0 & 0 & 0 & c-a+2b \end{array}\right]$$

Now the whole solution depends on the number $c-a+2b = c-(a-2b)$. The last row corresponds to an equation $0 = c-(a-2b)$. If $c \neq a-2b$, there is *no* solution (just as in Example 1.2.2). Hence:

The system is consistent if and only if $c = a-2b$.

[4]With n equations where n is large, gaussian elimination requires roughly $n^3/2$ multiplications and divisions, whereas this number is roughly $n^3/3$ if back substitution is used.

In this case the last matrix becomes

$$\left[\begin{array}{rrr|c} 1 & 0 & -5 & -2a-3b \\ 0 & 1 & 2 & a+b \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Thus, if $c = a - 2b$, taking $x_3 = t$ where t is a parameter gives the solutions

$$x_1 = 5t - (2a+3b) \quad x_2 = (a+b) - 2t \quad x_3 = t.$$

Rank

It can be proven that the *reduced* row-echelon form of a matrix A is uniquely determined by A. That is, no matter which series of row operations is used to carry A to a reduced row-echelon matrix, the result will always be the same matrix. (A proof is given at the end of Section 2.5.) By contrast, this is not true for row-echelon matrices: Different series of row operations can carry the same matrix A to *different* row-echelon matrices. Indeed, the matrix $A = \left[\begin{array}{rrr} 1 & -1 & 4 \\ 2 & -1 & 2 \end{array}\right]$ can be carried (by one row operation) to the row-echelon matrix $\left[\begin{array}{rrr} 1 & -1 & 4 \\ 0 & 1 & -6 \end{array}\right]$, and then by another row operation to the (reduced) row-echelon matrix $\left[\begin{array}{rrr} 1 & 0 & -2 \\ 0 & 1 & -6 \end{array}\right]$. However, it *is* true that the number r of leading 1s must be the same in each of these row-echelon matrices (this will be proved in Chapter 5). Hence, the number r depends only on A and not on the way in which A is carried to row-echelon form.

Definition 1.4 Rank of a Matrix

The ***rank*** *of matrix A is the number of leading 1s in any row-echelon matrix to which A can be carried by row operations.*

Example 1.2.5

Compute the rank of $A = \left[\begin{array}{rrrr} 1 & 1 & -1 & 4 \\ 2 & 1 & 3 & 0 \\ 0 & 1 & -5 & 8 \end{array}\right]$.

Solution. The reduction of A to row-echelon form is

$$A = \left[\begin{array}{rrrr} 1 & 1 & -1 & 4 \\ 2 & 1 & 3 & 0 \\ 0 & 1 & -5 & 8 \end{array}\right] \rightarrow \left[\begin{array}{rrrr} 1 & 1 & -1 & 4 \\ 0 & -1 & 5 & -8 \\ 0 & 1 & -5 & 8 \end{array}\right] \rightarrow \left[\begin{array}{rrrr} 1 & 1 & -1 & 4 \\ 0 & 1 & -5 & 8 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Because this row-echelon matrix has two leading 1s, rank $A = 2$.

Suppose that rank $A = r$, where A is a matrix with m rows and n columns. Then $r \leq m$ because the leading 1s lie in different rows, and $r \leq n$ because the leading 1s lie in different columns. Moreover, the rank has a useful application to equations. Recall that a system of linear equations is called consistent if it has at least one solution.

Theorem 1.2.2

Suppose a system of m equations in n variables is **consistent**, *and that the rank of the augmented matrix is r.*

1. *The set of solutions involves exactly $n-r$ parameters.*
2. *If $r < n$, the system has infinitely many solutions.*
3. *If $r = n$, the system has a unique solution.*

Proof. The fact that the rank of the augmented matrix is r means there are exactly r leading variables, and hence exactly $n-r$ nonleading variables. These nonleading variables are all assigned as parameters in the gaussian algorithm, so the set of solutions involves exactly $n-r$ parameters. Hence if $r < n$, there is at least one parameter, and so infinitely many solutions. If $r = n$, there are no parameters and so a unique solution. □

Theorem 1.2.2 shows that, for any system of linear equations, exactly three possibilities exist:

1. *No solution. This occurs when a row* $\left[\begin{array}{ccccc} 0 & 0 & \cdots & 0 & 1 \end{array}\right]$ *occurs in the row-echelon form. This is the case where the system is inconsistent.*
2. *Unique solution. This occurs when* every *variable is a leading variable.*
3. *Infinitely many solutions. This occurs when the system is consistent and there is at least one nonleading variable, so at least one parameter is involved.*

Example 1.2.6

Suppose the matrix A in Example 1.2.5 is the augmented matrix of a system of $m = 3$ linear equations in $n = 3$ variables. As rank $A = r = 2$, the set of solutions will have $n - r = 1$ parameter. The reader can verify this fact directly.

Many important problems involve **linear inequalities** rather than **linear equations**. For example, a condition on the variables x and y might take the form of an inequality $2x - 5y \leq 4$ rather than an equality $2x - 5y = 4$. There is a technique (called the **simplex algorithm**) for finding solutions to a system of such inequalities that maximizes a function of the form $p = ax + by$ where a and b are fixed constants.

1.3 Homogeneous Equations

A system of equations in the variables $x_1, x_2, \dots, x_n$ is called **homogeneous** if all the constant terms are zero—that is, if each equation of the system has the form

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = 0$$

Clearly $x_1 = 0, x_2 = 0, \dots, x_n = 0$ is a solution to such a system; it is called the **trivial solution**. Any solution in which at least one variable has a nonzero value is called a **nontrivial solution**. Our chief goal in this section is to give a useful condition for a homogeneous system to have nontrivial solutions. The following example is instructive.

Example 1.3.1

Show that the following homogeneous system has nontrivial solutions.

$$\begin{array}{rcrcrcrcl} x_1 & - & x_2 & + & 2x_3 & - & x_4 & = & 0 \\ 2x_1 & + & 2x_2 & & & + & x_4 & = & 0 \\ 3x_1 & + & x_2 & + & 2x_3 & - & x_4 & = & 0 \end{array}$$

Solution. The reduction of the augmented matrix to reduced row-echelon form is outlined below.

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

The leading variables are x_1, x_2, and x_4, so x_3 is assigned as a parameter—say $x_3 = t$. Then the general solution is $x_1 = -t$, $x_2 = t$, $x_3 = t$, $x_4 = 0$. Hence, taking $t = 1$ (say), we get a nontrivial solution: $x_1 = -1$, $x_2 = 1$, $x_3 = 1$, $x_4 = 0$.

The existence of a nontrivial solution in Example 1.3.1 is ensured by the presence of a parameter in the solution. This is due to the fact that there is a *nonleading* variable (x_3 in this case). But there *must* be a nonleading variable here because there are four variables and only three equations (and hence at *most* three leading variables). This discussion generalizes to a proof of the following fundamental theorem.

Theorem 1.3.1

If a homogeneous system of linear equations has more variables than equations, then it has a nontrivial solution (in fact, infinitely many).

Proof. Suppose there are m equations in n variables where $n > m$, and let R denote the reduced row-echelon form of the augmented matrix. If there are r leading variables, there are $n - r$ nonleading variables, and so $n - r$ parameters. Hence, it suffices to show that $r < n$. But $r \leq m$ because R has r leading 1s and m rows, and $m < n$ by hypothesis. So $r \leq m < n$, which gives $r < n$. □

Note that the converse of Theorem 1.3.1 is not true: if a homogeneous system has nontrivial solutions, it need not have more variables than equations (the system $x_1 + x_2 = 0$, $2x_1 + 2x_2 = 0$ has nontrivial solutions but $m = 2 = n$.)

Theorem 1.3.1 is very useful in applications. The next example provides an illustration from geometry.

Example 1.3.2

We call the graph of an equation $ax^2 + bxy + cy^2 + dx + ey + f = 0$ a **conic** if the numbers a, b, and c are not all zero. Show that there is at least one conic through any five points in the plane that are not all on a line.

Solution. Let the coordinates of the five points be (p_1, q_1), (p_2, q_2), (p_3, q_3), (p_4, q_4), and (p_5, q_5). The graph of $ax^2 + bxy + cy^2 + dx + ey + f = 0$ passes through (p_i, q_i) if

$$ap_i^2 + bp_iq_i + cq_i^2 + dp_i + eq_i + f = 0$$

This gives five equations, one for each i, linear in the six variables a, b, c, d, e, and f. Hence, there is a nontrivial solution by Theorem 1.3.1. If $a = b = c = 0$, the five points all lie on the line with equation $dx + ey + f = 0$, contrary to assumption. Hence, one of a, b, c is nonzero.

Linear Combinations and Basic Solutions

As for rows, two columns are regarded as **equal** if they have the same number of entries and corresponding entries are the same. Let $\mathbf{x}$ and $\mathbf{y}$ be columns with the same number of entries. As for elementary row operations, their **sum** $\mathbf{x}+\mathbf{y}$ is obtained by adding corresponding entries and, if k is a number, the **scalar product** $k\mathbf{x}$ is defined by multiplying each entry of $\mathbf{x}$ by k. More precisely:

$$\text{If } \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \text{ then } \mathbf{x}+\mathbf{y} = \begin{bmatrix} x_1+y_1 \\ x_2+y_2 \\ \vdots \\ x_n+y_n \end{bmatrix} \text{ and } k\mathbf{x} = \begin{bmatrix} kx_1 \\ kx_2 \\ \vdots \\ kx_n \end{bmatrix}.$$

A sum of scalar multiples of several columns is called a **linear combination** of these columns. For example, $s\mathbf{x}+t\mathbf{y}$ is a linear combination of $\mathbf{x}$ and $\mathbf{y}$ for any choice of numbers s and t.

Example 1.3.3

If $\mathbf{x} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ then $2\mathbf{x}+5\mathbf{y} = \begin{bmatrix} 6 \\ -4 \end{bmatrix} + \begin{bmatrix} -5 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

Example 1.3.4

Let $\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix}$. If $\mathbf{v} = \begin{bmatrix} 0 \\ -1 \\ 2 \end{bmatrix}$ and $\mathbf{w} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, determine whether $\mathbf{v}$ and $\mathbf{w}$ are linear combinations of $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$.

Solution. For $\mathbf{v}$, we must determine whether numbers r, s, and t exist such that $\mathbf{v} = r\mathbf{x}+s\mathbf{y}+t\mathbf{z}$, that is, whether

$$\begin{bmatrix} 0 \\ -1 \\ 2 \end{bmatrix} = r\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + s\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} + t\begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} r+2s+3t \\ s+t \\ r+t \end{bmatrix}$$

Equating corresponding entries gives a system of linear equations $r+2s+3t=0$, $s+t=-1$, and $r+t=2$ for r, s, and t. By gaussian elimination, the solution is $r=2-k$, $s=-1-k$, and $t=k$ where k is a parameter. Taking $k=0$, we see that $\mathbf{v}=2\mathbf{x}-\mathbf{y}$ is a linear combination of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. Turning to $\mathbf{w}$, we again look for r, s, and t such that $\mathbf{w} = r\mathbf{x}+s\mathbf{y}+t\mathbf{z}$; that is,

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = r\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + s\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} + t\begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} r+2s+3t \\ s+t \\ r+t \end{bmatrix}$$

leading to equations $r+2s+3t=1$, $s+t=1$, and $r+t=1$ for real numbers r, s, and t. But this time there is *no* solution as the reader can verify, so $\mathbf{w}$ is *not* a linear combination of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$.

Our interest in linear combinations comes from the fact that they provide one of the best ways to describe the general solution of a homogeneous system of linear equations. When solving such a system with n variables $x_1, x_2, \dots, x_n$, write the variables as a column[5] matrix: $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$. The trivial solution is denoted $\mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$. As an illustration, the general solution in Example 1.3.1 is $x_1 = -t$, $x_2 = t$, $x_3 = t$, and $x_4 = 0$, where t is a parameter, and we would now express this by saying that the general solution is $\mathbf{x} = \begin{bmatrix} -t \\ t \\ t \\ 0 \end{bmatrix}$, where t is arbitrary.

Now let $\mathbf{x}$ and $\mathbf{y}$ be two solutions to a homogeneous system with n variables. Then any linear combination $s\mathbf{x} + t\mathbf{y}$ of these solutions turns out to be again a solution to the system. More generally:

$$\textit{Any linear combination of solutions to a homogeneous system is again a solution.} \tag{1.1}$$

In fact, suppose that a typical equation in the system is $a_1x_1 + a_2x_2 + \cdots + a_nx_n = 0$, and suppose that $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ are solutions. Then $a_1x_1 + a_2x_2 + \cdots + a_nx_n = 0$ and $a_1y_1 + a_2y_2 + \cdots + a_ny_n = 0$. Hence $s\mathbf{x} + t\mathbf{y} = \begin{bmatrix} sx_1 + ty_1 \\ sx_2 + ty_2 \\ \vdots \\ sx_n + ty_n \end{bmatrix}$ is also a solution because

$$\begin{aligned}
a_1(sx_1 + ty_1) + a_2(sx_2 + ty_2) + \cdots + a_n(sx_n + ty_n) & \\
&= [a_1(sx_1) + a_2(sx_2) + \cdots + a_n(sx_n)] + [a_1(ty_1) + a_2(ty_2) + \cdots + a_n(ty_n)] \\
&= s(a_1x_1 + a_2x_2 + \cdots + a_nx_n) + t(a_1y_1 + a_2y_2 + \cdots + a_ny_n) \\
&= s(0) + t(0) \\
&= 0
\end{aligned}$$

A similar argument shows that Statement 1.1 is true for linear combinations of more than two solutions.

The remarkable thing is that *every* solution to a homogeneous system is a linear combination of certain particular solutions and, in fact, these solutions are easily computed using the gaussian algorithm. Here is an example.

[5]The reason for using columns will be apparent later.

Example 1.3.5

Solve the homogeneous system with coefficient matrix

$$A = \begin{bmatrix} 1 & -2 & 3 & -2 \\ -3 & 6 & 1 & 0 \\ -2 & 4 & 4 & -2 \end{bmatrix}$$

Solution. The reduction of the augmented matrix to reduced form is

$$\left[\begin{array}{rrrr|r} 1 & -2 & 3 & -2 & 0 \\ -3 & 6 & 1 & 0 & 0 \\ -2 & 4 & 4 & -2 & 0 \end{array}\right] \rightarrow \left[\begin{array}{rrrr|r} 1 & -2 & 0 & -\frac{1}{5} & 0 \\ 0 & 0 & 1 & -\frac{3}{5} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

so the solutions are $x_1 = 2s + \frac{1}{5}t$, $x_2 = s$, $x_3 = \frac{3}{5}t$, and $x_4 = t$ by gaussian elimination. Hence we can write the general solution $\mathbf{x}$ in the matrix form

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 2s + \frac{1}{5}t \\ s \\ \frac{3}{5}t \\ t \end{bmatrix} = s \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} \frac{1}{5} \\ 0 \\ \frac{3}{5} \\ 1 \end{bmatrix} = s\mathbf{x}_1 + t\mathbf{x}_2.$$

Here $\mathbf{x}_1 = \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ and $\mathbf{x}_2 = \begin{bmatrix} \frac{1}{5} \\ 0 \\ \frac{3}{5} \\ 1 \end{bmatrix}$ are particular solutions determined by the gaussian algorithm.

The solutions $\mathbf{x}_1$ and $\mathbf{x}_2$ in Example 1.3.5 are denoted as follows:

Definition 1.5 Basic Solutions

The gaussian algorithm systematically produces solutions to any homogeneous linear system, called ***basic solutions****, one for every parameter.*

Moreover, the algorithm gives a routine way to express *every* solution as a linear combination of basic solutions as in Example 1.3.5, where the general solution $\mathbf{x}$ becomes

$$\mathbf{x} = s \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} \frac{1}{5} \\ 0 \\ \frac{3}{5} \\ 1 \end{bmatrix} = s \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \frac{1}{5}t \begin{bmatrix} 1 \\ 0 \\ 3 \\ 5 \end{bmatrix}$$

Hence by introducing a new parameter $r = t/5$ we can multiply the original basic solution $\mathbf{x}_2$ by 5 and so eliminate fractions. For this reason:

Convention:
Any nonzero scalar multiple of a basic solution will still be called a basic solution.

In the same way, the gaussian algorithm produces basic solutions to *every* homogeneous system, one for each parameter (there are *no* basic solutions if the system has only the trivial solution). Moreover every solution is given by the algorithm as a linear combination of these basic solutions (as in Example 1.3.5). If A has rank r, Theorem 1.2.2 shows that there are exactly $n-r$ parameters, and so $n-r$ basic solutions. This proves:

Theorem 1.3.2

Let A be an $m \times n$ matrix of rank r, and consider the homogeneous system in n variables with A as coefficient matrix. Then:

1. *The system has exactly $n-r$ basic solutions, one for each parameter.*
2. *Every solution is a linear combination of these basic solutions.*

Example 1.3.6

Find basic solutions of the homogeneous system with coefficient matrix A, and express every solution as a linear combination of the basic solutions, where

$$A = \begin{bmatrix} 1 & -3 & 0 & 2 & 2 \\ -2 & 6 & 1 & 2 & -5 \\ 3 & -9 & -1 & 0 & 7 \\ -3 & 9 & 2 & 6 & -8 \end{bmatrix}$$

Solution. The reduction of the augmented matrix to reduced row-echelon form is

$$\left[\begin{array}{rrrrr|r} 1 & -3 & 0 & 2 & 2 & 0 \\ -2 & 6 & 1 & 2 & -5 & 0 \\ 3 & -9 & -1 & 0 & 7 & 0 \\ -3 & 9 & 2 & 6 & -8 & 0 \end{array}\right] \rightarrow \left[\begin{array}{rrrrr|r} 1 & -3 & 0 & 2 & 2 & 0 \\ 0 & 0 & 1 & 6 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

so the general solution is $x_1 = 3r-2s-2t$, $x_2 = r$, $x_3 = -6s+t$, $x_4 = s$, and $x_5 = t$ where r, s, and t are parameters. In matrix form this is

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 3r-2s-2t \\ r \\ -6s+t \\ s \\ t \end{bmatrix} = r\begin{bmatrix} 3 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + s\begin{bmatrix} -2 \\ 0 \\ -6 \\ 1 \\ 0 \end{bmatrix} + t\begin{bmatrix} -2 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

Hence basic solutions are

$$\mathbf{x}_1 = \begin{bmatrix} 3 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} -2 \\ 0 \\ -6 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix} -2 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

1.4 An Application to Network Flow

There are many types of problems that concern a network of conductors along which some sort of flow is observed. Examples of these include an irrigation network and a network of streets or freeways. There are often points in the system at which a net flow either enters or leaves the system. The basic principle behind the analysis of such systems is that the total flow into the system must equal the total flow out. In fact, we apply this principle at every junction in the system.

Theorem: Junction Rule

At each of the junctions in the network, the total flow into that junction must equal the total flow out.

This requirement gives a linear equation relating the flows in conductors emanating from the junction.

Example 1.4.1

A network of one-way streets is shown in the accompanying diagram. The rate of flow of cars into intersection A is 500 cars per hour, and 400 and 100 cars per hour emerge from B and C, respectively. Find the possible flows along each street.

Solution. Suppose the flows along the streets are f_1, f_2, f_3, f_4, f_5, and f_6 cars per hour in the directions shown.

Then, equating the flow in with the flow out at each intersection, we get

$$\begin{array}{ll} \text{Intersection } A & 500 = f_1 + f_2 + f_3 \\ \text{Intersection } B & f_1 + f_4 + f_6 = 400 \\ \text{Intersection } C & f_3 + f_5 = f_6 + 100 \\ \text{Intersection } D & f_2 = f_4 + f_5 \end{array}$$

These give four equations in the six variables $f_1, f_2, \dots, f_6$.

$$\begin{array}{rcl} f_1 + f_2 + f_3 & & = 500 \\ f_1 \qquad\quad + f_4 \qquad + f_6 & & = 400 \\ f_3 \qquad + f_5 - f_6 & & = 100 \\ f_2 \qquad - f_4 - f_5 \qquad & & = 0 \end{array}$$

The reduction of the augmented matrix is

$$\left[\begin{array}{rrrrrr|r} 1 & 1 & 1 & 0 & 0 & 0 & 500 \\ 1 & 0 & 0 & 1 & 0 & 1 & 400 \\ 0 & 0 & 1 & 0 & 1 & -1 & 100 \\ 0 & 1 & 0 & -1 & -1 & 0 & 0 \end{array}\right] \rightarrow \left[\begin{array}{rrrrrr|r} 1 & 0 & 0 & 1 & 0 & 1 & 400 \\ 0 & 1 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & -1 & 100 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

Hence, when we use f_4, f_5, and f_6 as parameters, the general solution is

$$f_1 = 400 - f_4 - f_6 \qquad f_2 = f_4 + f_5 \qquad f_3 = 100 - f_5 + f_6$$

This gives all solutions to the system of equations and hence all the possible flows.
Of course, not all these solutions may be acceptable in the real situation. For example, the flows $f_1, f_2, \dots, f_6$ are all *positive* in the present context (if one came out negative, it would mean traffic flowed in the opposite direction). This imposes constraints on the flows: $f_1 \geq 0$ and $f_3 \geq 0$ become

$$f_4 + f_6 \leq 400 \qquad f_5 - f_6 \leq 100$$

Further constraints might be imposed by insisting on maximum values on the flow in each street.

1.5 An Application to Electrical Networks[6]

In an electrical network it is often necessary to find the current in amperes (A) flowing in various parts of the network. These networks usually contain resistors that retard the current. The resistors are indicated by a symbol ( ), and the resistance is measured in ohms (Ω). Also, the current is increased at various points by voltage sources (for example, a battery). The voltage of these sources is measured in volts (V), and they are represented by the symbol ( ). We assume these voltage sources have no resistance. The flow of current is governed by the following principles.

Theorem: Ohm's Law

The current I and the voltage drop V across a resistance R are related by the equation $V = RI$.

Theorem: Kirchhoff's Laws

1. *(Junction Rule) The current flow into a junction equals the current flow out of that junction.*
2. *(Circuit Rule) The algebraic sum of the voltage drops (due to resistances) around any closed circuit of the network must equal the sum of the voltage increases around the circuit.*

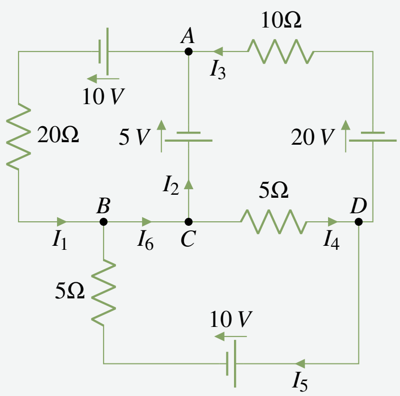
[6]This section is independent of Section 1.4

When applying rule 2, select a direction (clockwise or counterclockwise) around the closed circuit and then consider all voltages and currents positive when in this direction and negative when in the opposite direction. This is why the term *algebraic sum* is used in rule 2. Here is an example.

Example 1.5.1

Find the various currents in the circuit shown.

Solution.



First apply the junction rule at junctions A, B, C, and D to obtain

$$\begin{array}{ll} \text{Junction } A & I_1 = I_2 + I_3 \\ \text{Junction } B & I_6 = I_1 + I_5 \\ \text{Junction } C & I_2 + I_4 = I_6 \\ \text{Junction } D & I_3 + I_5 = I_4 \end{array}$$

Note that these equations are not independent (in fact, the third is an easy consequence of the other three). Next, the circuit rule insists that the sum of the voltage increases (due to the sources) around a closed circuit must equal the sum of the voltage drops (due to resistances). By Ohm's law, the voltage loss across a resistance R (in the direction of the current I) is RI. Going counterclockwise around three closed circuits yields

$$\begin{array}{lrl} \text{Upper left} & 10 + \quad 5 = & 20I_1 \\ \text{Upper right} & -5 + \quad 20 = & 10I_3 + 5I_4 \\ \text{Lower} & -10 = & -5I_5 - 5I_4 \end{array}$$

Hence, disregarding the redundant equation obtained at junction C, we have six equations in the six unknowns $I_1, \dots, I_6$. The solution is

$$\begin{array}{ll} I_1 = \frac{15}{20} & I_4 = \frac{28}{20} \\ I_2 = \frac{-1}{20} & I_5 = \frac{12}{20} \\ I_3 = \frac{16}{20} & I_6 = \frac{27}{20} \end{array}$$

The fact that I_2 is negative means, of course, that this current is in the opposite direction, with a magnitude of $\frac{1}{20}$ amperes.

1.6 An Application to Chemical Reactions

When a chemical reaction takes place a number of molecules combine to produce new molecules. Hence, when hydrogen H_2 and oxygen O_2 molecules combine, the result is water H_2O. We express this as

$$H_2 + O_2 \to H_2O$$

Individual atoms are neither created nor destroyed, so the number of hydrogen and oxygen atoms going into the reaction must equal the number coming out (in the form of water). In this case the reaction is said to be *balanced*. Note that each hydrogen molecule H_2 consists of two atoms as does each oxygen molecule O_2, while a water molecule H_2O consists of two hydrogen atoms and one oxygen atom. In the above reaction, this requires that twice as many hydrogen molecules enter the reaction; we express this as follows:

$$2H_2 + O_2 \to 2H_2O$$

This is now balanced because there are 4 hydrogen atoms and 2 oxygen atoms on each side of the reaction.

Example 1.6.1

Balance the following reaction for burning octane C_8H_{18} in oxygen O_2:

$$C_8H_{18} + O_2 \to CO_2 + H_2O$$

where CO_2 represents carbon dioxide. We must find positive integers x, y, z, and w such that

$$xC_8H_{18} + yO_2 \rightarrow zCO_2 + wH_2O$$

Equating the number of carbon, hydrogen, and oxygen atoms on each side gives $8x = z$, $18x = 2w$ and $2y = 2z + w$, respectively. These can be written as a homogeneous linear system

$$\begin{array}{rcrcrcrcl} 8x & & & - & z & & & = & 0 \\ 18x & & & & & - & 2w & = & 0 \\ & & 2y & - & 2z & - & w & = & 0 \end{array}$$

which can be solved by gaussian elimination. In larger systems this is necessary but, in such a simple situation, it is easier to solve directly. Set $w = t$, so that $x = \frac{1}{9}t$, $z = \frac{8}{9}t$, $2y = \frac{16}{9}t + t = \frac{25}{9}t$. But x, y, z, and w must be positive integers, so the smallest value of t that eliminates fractions is 18. Hence, $x = 2$, $y = 25$, $z = 16$, and $w = 18$, and the balanced reaction is

$$2C_8H_{18} + 25O_2 \rightarrow 16CO_2 + 18H_2O$$

The reader can verify that this is indeed balanced.

It is worth noting that this problem introduces a new element into the theory of linear equations: the insistence that the solution must consist of positive integers.

Chapter 2

Matrix Algebra

In the study of systems of linear equations in Chapter 1, we found it convenient to manipulate the augmented matrix of the system. Our aim was to reduce it to row-echelon form (using elementary row operations) and hence to write down all solutions to the system. In the present chapter we consider matrices for their own sake. While some of the motivation comes from linear equations, it turns out that matrices can be multiplied and added and so form an algebraic system somewhat analogous to the real numbers. This "matrix algebra" is useful in ways that are quite different from the study of linear equations. For example, the geometrical transformations obtained by rotating the euclidean plane about the origin can be viewed as multiplications by certain 2×2 matrices. These "matrix transformations" are an important tool in geometry and, in turn, the geometry provides a "picture" of the matrices. Furthermore, matrix algebra has many other applications, some of which will be explored in this chapter. This subject is quite old and was first studied systematically in 1858 by Arthur Cayley.[1]

[1]Arthur Cayley (1821-1895) showed his mathematical talent early and graduated from Cambridge in 1842 as senior wrangler. With no employment in mathematics in view, he took legal training and worked as a lawyer while continuing to do mathematics, publishing nearly 300 papers in fourteen years. Finally, in 1863, he accepted the Sadlerian professorship in Cambridge and remained there for the rest of his life, valued for his administrative and teaching skills as well as for his scholarship. His mathematical achievements were of the first rank. In addition to originating matrix theory and the theory of determinants, he did fundamental work in group theory, in higher-dimensional geometry, and in the theory of invariants. He was one of the most prolific mathematicians of all time and produced 966 papers.

2.1 Matrix Addition, Scalar Multiplication, and Transposition

A rectangular array of numbers is called a **matrix** (the plural is **matrices**), and the numbers are called the **entries** of the matrix. Matrices are usually denoted by uppercase letters: A, B, C, and so on. Hence,

$$A = \left[\begin{array}{rrr} 1 & 2 & -1 \\ 0 & 5 & 6 \end{array} \right] \quad B = \left[\begin{array}{rr} 1 & -1 \\ 0 & 2 \end{array} \right] \quad C = \left[\begin{array}{r} 1 \\ 3 \\ 2 \end{array} \right]$$

are matrices. Clearly matrices come in various shapes depending on the number of **rows** and **columns**. For example, the matrix A shown has 2 rows and 3 columns. In general, a matrix with m rows and n columns is referred to as an $\boldsymbol{m \times n}$ ***matrix*** or as having **size** $\boldsymbol{m \times n}$. Thus matrices A, B, and C above have sizes 2×3, 2×2, and 3×1, respectively. A matrix of size $1 \times n$ is called a **row matrix**, whereas one of size $m \times 1$ is called a **column matrix**. Matrices of size $n \times n$ for some n are called **square** matrices.

Each entry of a matrix is identified by the row and column in which it lies. The rows are numbered from the top down, and the columns are numbered from left to right. Then the $\boldsymbol{(i,\ j)}$**-entry** of a matrix is the number lying simultaneously in row i and column j. For example,

$$\text{The } (1,\ 2)\text{-entry of } \left[\begin{array}{rr} 1 & -1 \\ 0 & 1 \end{array} \right] \text{ is } -1.$$

$$\text{The } (2,\ 3)\text{-entry of } \left[\begin{array}{rrr} 1 & 2 & -1 \\ 0 & 5 & 6 \end{array} \right] \text{ is } 6.$$

A special notation is commonly used for the entries of a matrix. If A is an $m \times n$ matrix, and if the $(i,\ j)$-entry of A is denoted as a_{ij}, then A is displayed as follows:

$$A = \left[\begin{array}{ccccc} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{array} \right]$$

This is usually denoted simply as $A = \left[a_{ij} \right]$. Thus a_{ij} is the entry in row i and column j of A. For example, a 3×4 matrix in this notation is written

$$A = \left[\begin{array}{cccc} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{array} \right]$$

It is worth pointing out a convention regarding rows and columns: *Rows are mentioned before columns.* For example:

- *If a matrix has size $m \times n$, it has m rows and n columns.*
- *If we speak of the $(i,\ j)$-entry of a matrix, it lies in row i and column j.*
- *If an entry is denoted a_{ij}, the first subscript i refers to the row and the second subscript j to the column in which a_{ij} lies.*

Two points (x_1, y_1) and (x_2, y_2) in the plane are equal if and only if[2] they have the same coordinates, that is $x_1 = x_2$ and $y_1 = y_2$. Similarly, two matrices A and B are called **equal** (written $A = B$) if and only if:

1. *They have the same size.*
2. *Corresponding entries are equal.*

If the entries of A and B are written in the form $A = \left[a_{ij}\right]$, $B = \left[b_{ij}\right]$, described earlier, then the second condition takes the following form:

$$A = \left[a_{ij}\right] = \left[b_{ij}\right] \text{ means } a_{ij} = b_{ij} \text{ for all } i \text{ and } j$$

Example 2.1.1

Given $A = \left[\begin{array}{cc} a & b \\ c & d \end{array}\right]$, $B = \left[\begin{array}{rrr} 1 & 2 & -1 \\ 3 & 0 & 1 \end{array}\right]$ and $C = \left[\begin{array}{rr} 1 & 0 \\ -1 & 2 \end{array}\right]$ discuss the possibility that $A = B$, $B = C$, $A = C$.

Solution. $A = B$ is impossible because A and B are of different sizes: A is 2×2 whereas B is 2×3. Similarly, $B = C$ is impossible. But $A = C$ is possible provided that corresponding entries are equal: $\left[\begin{array}{cc} a & b \\ c & d \end{array}\right] = \left[\begin{array}{rr} 1 & 0 \\ -1 & 2 \end{array}\right]$ means $a = 1$, $b = 0$, $c = -1$, and $d = 2$.

Matrix Addition

Definition 2.1 Matrix Addition

If A and B are matrices of the same size, their **sum** *$A + B$ is the matrix formed by adding corresponding entries.*

If $A = \left[a_{ij}\right]$ and $B = \left[b_{ij}\right]$, this takes the form

$$A + B = \left[a_{ij} + b_{ij}\right]$$

Note that addition is *not* defined for matrices of different sizes.

Example 2.1.2

If $A = \left[\begin{array}{rrr} 2 & 1 & 3 \\ -1 & 2 & 0 \end{array}\right]$ and $B = \left[\begin{array}{rrr} 1 & 1 & -1 \\ 2 & 0 & 6 \end{array}\right]$, compute $A + B$.

Solution.

$$A + B = \left[\begin{array}{rrr} 2+1 & 1+1 & 3-1 \\ -1+2 & 2+0 & 0+6 \end{array}\right] = \left[\begin{array}{rrr} 3 & 2 & 2 \\ 1 & 2 & 6 \end{array}\right]$$

[2]If p and q are statements, we say that p implies q if q is true whenever p is true. Then "p if and only if q" means that both p implies q and q implies p. See Appendix B for more on this.

Example 2.1.3

Find a, b, and c if $\begin{bmatrix} a & b & c \end{bmatrix} + \begin{bmatrix} c & a & b \end{bmatrix} = \begin{bmatrix} 3 & 2 & -1 \end{bmatrix}$.

Solution. Add the matrices on the left side to obtain

$$\begin{bmatrix} a+c & b+a & c+b \end{bmatrix} = \begin{bmatrix} 3 & 2 & -1 \end{bmatrix}$$

Because corresponding entries must be equal, this gives three equations: $a+c=3$, $b+a=2$, and $c+b=-1$. Solving these yields $a=3$, $b=-1$, $c=0$.

If A, B, and C are any matrices *of the same size*, then

$$A+B=B+A \qquad \text{(commutative law)}$$
$$A+(B+C)=(A+B)+C \qquad \text{(associative law)}$$

In fact, if $A=\left[a_{ij}\right]$ and $B=\left[b_{ij}\right]$, then the $(i,\ j)$-entries of $A+B$ and $B+A$ are, respectively, $a_{ij}+b_{ij}$ and $b_{ij}+a_{ij}$. Since these are equal for all i and j, we get

$$A+B=\begin{bmatrix} a_{ij}+b_{ij} \end{bmatrix} = \begin{bmatrix} b_{ij}+a_{ij} \end{bmatrix} = B+A$$

The associative law is verified similarly.

The $m \times n$ matrix in which every entry is zero is called the $m \times n$ **zero matrix** and is denoted as 0 (or 0_{mn} if it is important to emphasize the size). Hence,

$$0+X=X$$

holds for all $m \times n$ matrices X. The **negative** of an $m \times n$ matrix A (written $-A$) is defined to be the $m \times n$ matrix obtained by multiplying each entry of A by -1. If $A=\left[a_{ij}\right]$, this becomes $-A=\left[-a_{ij}\right]$. Hence,

$$A+(-A)=0$$

holds for all matrices A where, of course, 0 is the zero matrix of the same size as A.

A closely related notion is that of subtracting matrices. If A and B are two $m \times n$ matrices, their **difference** $A-B$ is defined by

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

Note that if $A=\left[a_{ij}\right]$ and $B=\left[b_{ij}\right]$, then

$$A-B=\left[a_{ij}\right]+\left[-b_{ij}\right]=\left[a_{ij}-b_{ij}\right]$$

is the $m \times n$ matrix formed by *subtracting* corresponding entries.

Example 2.1.4

Let $A=\begin{bmatrix} 3 & -1 & 0 \\ 1 & 2 & -4 \end{bmatrix}$, $B=\begin{bmatrix} 1 & -1 & 1 \\ -2 & 0 & 6 \end{bmatrix}$, $C=\begin{bmatrix} 1 & 0 & -2 \\ 3 & 1 & 1 \end{bmatrix}$. Compute $-A$, $A-B$, and $A+B-C$.

Solution.

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

$$A - B = \begin{bmatrix} 3-1 & -1-(-1) & 0-1 \\ 1-(-2) & 2-0 & -4-6 \end{bmatrix} = \begin{bmatrix} 2 & 0 & -1 \\ 3 & 2 & -10 \end{bmatrix}$$

$$A + B - C = \begin{bmatrix} 3+1-1 & -1-1-0 & 0+1-(-2) \\ 1-2-3 & 2+0-1 & -4+6-1 \end{bmatrix} = \begin{bmatrix} 3 & -2 & 3 \\ -4 & 1 & 1 \end{bmatrix}$$

Example 2.1.5

Solve $\begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix} + X = \begin{bmatrix} 1 & 0 \\ -1 & 2 \end{bmatrix}$ where X is a matrix.

Solution. We solve a numerical equation $a + x = b$ by subtracting the number a from both sides to obtain $x = b - a$. This also works for matrices. To solve $\begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix} + X = \begin{bmatrix} 1 & 0 \\ -1 & 2 \end{bmatrix}$ simply subtract the matrix $\begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix}$ from both sides to get

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

The reader should verify that this matrix X does indeed satisfy the original equation.

The solution in Example 2.1.5 solves the single matrix equation $A + X = B$ directly via matrix subtraction: $X = B - A$. This ability to work with matrices as entities lies at the heart of matrix algebra.

It is important to note that the sizes of matrices involved in some calculations are often determined by the context. For example, if

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

then A and C must be the same size (so that $A + C$ makes sense), and that size must be 2×3 (so that the sum is 2×3). For simplicity we shall often omit reference to such facts when they are clear from the context.

Scalar Multiplication

In gaussian elimination, multiplying a row of a matrix by a number k means multiplying *every* entry of that row by k.

Definition 2.2 Matrix Scalar Multiplication

More generally, if A is any matrix and k is any number, the ***scalar multiple*** *kA is the matrix obtained from A by multiplying each entry of A by k.*

If $A = \left[a_{ij}\right]$, this is

$$kA = \left[ka_{ij}\right]$$

Thus $1A = A$ and $(-1)A = -A$ for any matrix A.

The term *scalar* arises here because the set of numbers from which the entries are drawn is usually referred to as the set of scalars. We have been using real numbers as scalars, but we could equally well have been using complex numbers.

Example 2.1.6

If $A = \begin{bmatrix} 3 & -1 & 4 \\ 2 & 0 & 6 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 3 & 2 \end{bmatrix}$ compute $5A$, $\frac{1}{2}B$, and $3A - 2B$.

Solution.

$$5A = \begin{bmatrix} 15 & -5 & 20 \\ 10 & 0 & 30 \end{bmatrix}, \quad \frac{1}{2}B = \begin{bmatrix} \frac{1}{2} & 1 & -\frac{1}{2} \\ 0 & \frac{3}{2} & 1 \end{bmatrix}$$

$$3A - 2B = \begin{bmatrix} 9 & -3 & 12 \\ 6 & 0 & 18 \end{bmatrix} - \begin{bmatrix} 2 & 4 & -2 \\ 0 & 6 & 4 \end{bmatrix} = \begin{bmatrix} 7 & -7 & 14 \\ 6 & -6 & 14 \end{bmatrix}$$

If A is any matrix, note that kA is the same size as A for all scalars k. We also have

$$0A = 0 \quad \text{and} \quad k0 = 0$$

because the zero matrix has every entry zero. In other words, $kA = 0$ if either $k = 0$ or $A = 0$. The converse of this statement is also true, as Example 2.1.7 shows.

Example 2.1.7

If $kA = 0$, show that either $k = 0$ or $A = 0$.

Solution. Write $A = \left[a_{ij}\right]$ so that $kA = 0$ means $ka_{ij} = 0$ for all i and j. If $k = 0$, there is nothing to do. If $k \neq 0$, then $ka_{ij} = 0$ implies that $a_{ij} = 0$ for all i and j; that is, $A = 0$.

For future reference, the basic properties of matrix addition and scalar multiplication are listed in Theorem 2.1.1.

Theorem 2.1.1

Let A, B, and C denote arbitrary $m \times n$ matrices where m and n are fixed. Let k and p denote arbitrary real numbers. Then

1. $A + B = B + A$.
2. $A + (B + C) = (A + B) + C$.
3. *There is an $m \times n$ matrix 0, such that $0 + A = A$ for each A.*
4. *For each A there is an $m \times n$ matrix, $-A$, such that $A + (-A) = 0$.*

5. $k(A+B) = kA+kB$.
6. $(k+p)A = kA+pA$.
7. $(kp)A = k(pA)$.
8. $1A = A$.

Proof. Properties 1–4 were given previously. To check Property 5, let $A = \left[a_{ij}\right]$ and $B = \left[b_{ij}\right]$ denote matrices of the same size. Then $A+B = \left[a_{ij}+b_{ij}\right]$, as before, so the $(i,\ j)$-entry of $k(A+B)$ is

$$k(a_{ij}+b_{ij}) = ka_{ij}+kb_{ij}$$

But this is just the $(i,\ j)$-entry of $kA+kB$, and it follows that $k(A+B) = kA+kB$. The other Properties can be similarly verified; the details are left to the reader. □

The Properties in Theorem 2.1.1 enable us to do calculations with matrices in much the same way that numerical calculations are carried out. To begin, Property 2 implies that the sum

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

is the same no matter how it is formed and so is written as $A+B+C$. Similarly, the sum

$$A+B+C+D$$

is independent of how it is formed; for example, it equals both $(A+B)+(C+D)$ and $A+[B+(C+D)]$. Furthermore, property 1 ensures that, for example,

$$B+D+A+C = A+B+C+D$$

In other words, the *order* in which the matrices are added does not matter. A similar remark applies to sums of five (or more) matrices.

Properties 5 and 6 in Theorem 2.1.1 are called **distributive laws** for scalar multiplication, and they extend to sums of more than two terms. For example,

$$k(A+B-C) = kA+kB-kC$$

$$(k+p-m)A = kA+pA-mA$$

Similar observations hold for more than three summands. These facts, together with properties 7 and 8, enable us to simplify expressions by collecting like terms, expanding, and taking common factors in exactly the same way that algebraic expressions involving variables and real numbers are manipulated. The following example illustrates these techniques.

Example 2.1.8

Simplify $2(A+3C)-3(2C-B)-3\left[2(2A+B-4C)-4(A-2C)\right]$ where A, B, and C are all matrices of the same size.

Solution. The reduction proceeds as though A, B, and C were variables.

$$\begin{aligned} & 2(A+3C)-3(2C-B)-3\left[2(2A+B-4C)-4(A-2C)\right] \\ & \quad = 2A+6C-6C+3B-3\left[4A+2B-8C-4A+8C\right] \\ & \quad = 2A+3B-3\left[2B\right] \\ & \quad = 2A-3B \end{aligned}$$

Transpose of a Matrix

Many results about a matrix A involve the *rows* of A, and the corresponding result for columns is derived in an analogous way, essentially by replacing the word *row* by the word *column* throughout. The following definition is made with such applications in mind.

Definition 2.3 Transpose of a Matrix

If A is an $m \times n$ matrix, the **transpose** *of A, written A^T, is the $n \times m$ matrix whose rows are just the columns of A in the same order.*

In other words, the first row of A^T is the first column of A (that is it consists of the entries of column 1 in order). Similarly the second row of A^T is the second column of A, and so on.

Example 2.1.9

Write down the transpose of each of the following matrices.

$$A = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} \quad B = \begin{bmatrix} 5 & 2 & 6 \end{bmatrix} \quad C = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \quad D = \begin{bmatrix} 3 & 1 & -1 \\ 1 & 3 & 2 \\ -1 & 2 & 1 \end{bmatrix}$$

Solution.

$$A^T = \begin{bmatrix} 1 & 3 & 2 \end{bmatrix}, \ B^T = \begin{bmatrix} 5 \\ 2 \\ 6 \end{bmatrix}, \ C^T = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}, \ \text{and } D^T = D.$$

If $A = \left[a_{ij}\right]$ is a matrix, write $A^T = \left[b_{ij}\right]$. Then b_{ij} is the jth element of the ith row of A^T and so is the jth element of the ith *column* of A. This means $b_{ij} = a_{ji}$, so the definition of A^T can be stated as follows:

$$\text{If } A = \left[a_{ij}\right], \text{ then } A^T = \left[a_{ji}\right]. \tag{2.1}$$

This is useful in verifying the following properties of transposition.

Theorem 2.1.2

Let A and B denote matrices of the same size, and let k denote a scalar.

1. *If A is an $m \times n$ matrix, then A^T is an $n \times m$ matrix.*
2. $(A^T)^T = A$.
3. $(kA)^T = kA^T$.
4. $(A+B)^T = A^T + B^T$.

Proof. Property 1 is part of the definition of A^T, and Property 2 follows from (2.1). As to Property 3: If $A = \left[a_{ij}\right]$, then $kA = \left[ka_{ij}\right]$, so (2.1) gives

$$(kA)^T = \left[ka_{ji}\right] = k\left[a_{ji}\right] = kA^T$$

Finally, if $B = \left[b_{ij}\right]$, then $A+B = \left[c_{ij}\right]$ where $c_{ij} = a_{ij} + b_{ij}$ Then (2.1) gives Property 4:

$$(A+B)^T = \left[c_{ij}\right]^T = \left[c_{ji}\right] = \left[a_{ji} + b_{ji}\right] = \left[a_{ji}\right] + \left[b_{ji}\right] = A^T + B^T$$

□

There is another useful way to think of transposition. If $A = [a_{ij}]$ is an $m \times n$ matrix, the elements $a_{11}, a_{22}, a_{33}, \ldots$ are called the **main diagonal** of A. Hence the main diagonal extends down and to the right from the upper left corner of the matrix A; it is outlined in the following examples:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix}$$

Thus forming the transpose of a matrix A can be viewed as "flipping" A about its main diagonal, or as "rotating" A through $180°$ about the line containing the main diagonal. This makes Property 2 in Theorem 2.1.2 transparent.

Example 2.1.10

Solve for A if $\left(2A^T - 3\begin{bmatrix} 1 & 2 \\ -1 & 1 \end{bmatrix}\right)^T = \begin{bmatrix} 2 & 3 \\ -1 & 2 \end{bmatrix}$.

Solution. Using Theorem 2.1.2, the left side of the equation is

$$\left(2A^T - 3\begin{bmatrix} 1 & 2 \\ -1 & 1 \end{bmatrix}\right)^T = 2\left(A^T\right)^T - 3\begin{bmatrix} 1 & 2 \\ -1 & 1 \end{bmatrix}^T = 2A - 3\begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}$$

Hence the equation becomes

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

Thus $2A = \begin{bmatrix} 2 & 3 \\ -1 & 2 \end{bmatrix} + 3\begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 5 & 5 \end{bmatrix}$, so finally $A = \frac{1}{2}\begin{bmatrix} 5 & 0 \\ 5 & 5 \end{bmatrix} = \frac{5}{2}\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$.

Note that Example 2.1.10 can also be solved by first transposing both sides, then solving for A^T, and so obtaining $A = (A^T)^T$. The reader should do this.

The matrix $D = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}$ in Example 2.1.9 has the property that $D = D^T$. Such matrices are important; a matrix A is called **symmetric** if $A = A^T$. A symmetric matrix A is necessarily square (if A is $m \times n$, then A^T is $n \times m$, so $A = A^T$ forces $n = m$). The name comes from the fact that these matrices exhibit a symmetry about the main diagonal. That is, entries that are directly across the main diagonal from each other are equal.

For example, $\begin{bmatrix} a & b & c \\ b' & d & e \\ c' & e' & f \end{bmatrix}$ is symmetric when $b = b'$, $c = c'$, and $e = e'$.

Example 2.1.11

If A and B are symmetric $n \times n$ matrices, show that $A + B$ is symmetric.

Solution. We have $A^T = A$ and $B^T = B$, so, by Theorem 2.1.2, we have

$(A+B)^T = A^T + B^T = A + B$. Hence $A+B$ is symmetric.

Example 2.1.12

Suppose a square matrix A satisfies $A = 2A^T$. Show that necessarily $A = 0$.

Solution. If we iterate the given equation, Theorem 2.1.2 gives

$$A = 2A^T = 2\left[2A^T\right]^T = 2\left[2(A^T)^T\right] = 4A$$

Subtracting A from both sides gives $3A = 0$, so $A = \frac{1}{3}(0) = 0$.

2.2 Matrix-Vector Multiplication

Up to now we have used matrices to solve systems of linear equations by manipulating the rows of the augmented matrix. In this section we introduce a different way of describing linear systems that makes more use of the coefficient matrix of the system and leads to a useful way of "multiplying" matrices.

Vectors

It is a well-known fact in analytic geometry that two points in the plane with coordinates $(a_1,\ a_2)$ and $(b_1,\ b_2)$ are equal if and only if $a_1 = b_1$ and $a_2 = b_2$. Moreover, a similar condition applies to points $(a_1,\ a_2,\ a_3)$ in space. We extend this idea as follows.

An ordered sequence $(a_1,\ a_2,\ \dots,\ a_n)$ of real numbers is called an **ordered n-tuple**. The word "ordered" here reflects our insistence that two ordered n-tuples are equal if and only if corresponding entries are the same. In other words,

$$(a_1,\ a_2,\ \dots,\ a_n) = (b_1,\ b_2,\ \dots,\ b_n) \quad \text{if and only if} \quad a_1 = b_1,\ a_2 = b_2,\ \dots,\ \text{ and } a_n = b_n.$$

Thus the ordered 2-tuples and 3-tuples are just the ordered pairs and triples familiar from geometry.

Definition 2.4 The set $\mathbb{R}^n$ of ordered n-tuples of real numbers

Let $\mathbb{R}$ denote the set of all real numbers. The set of all *ordered n-tuples from $\mathbb{R}$ has a special notation:*

$$\mathbb{R}^n \text{ denotes the set of all ordered } n\text{-tuples of real numbers.}$$

While elements in $\mathbb{R}^n$ can be written as rows $(r_1,\ r_2,\ \dots,\ r_n)$, we will most often write them as $n \times 1$ column matrices $\left[\begin{array}{c} r_1 \\ r_2 \\ \vdots \\ r_n \end{array} \right]$ and use matrix algebra as previously seen in Section 2.1 on elements of $\mathbb{R}^n$. These are called **vectors** or n-**vectors** and will be denoted using bold type such as $\mathbf{x}$ or $\mathbf{v}$. This is indeed very convenient and powerful as we will see. For example, an $m \times n$ matrix A will be written as a row of n-vectors (its columns):

$$A = \left[\begin{array}{cccc} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{array} \right] \text{ where } \mathbf{a}_j \text{ denotes column } j \text{ of } A \text{ for each } j.$$

If $\mathbf{x}$ and $\mathbf{y}$ are two n-vectors in $\mathbb{R}^n$, it is clear that their matrix sum $\mathbf{x} + \mathbf{y}$ is also in $\mathbb{R}^n$ as is the scalar multiple $k\mathbf{x}$ for any real number k. We express this observation by saying that $\mathbb{R}^n$ is **closed** under addition and scalar multiplication. In particular, all the basic properties in Theorem 2.1.1 are true of these n-vectors. These properties are fundamental and will be used frequently below without comment. As for matrices in general, the $n \times 1$ zero matrix is called the **zero n-vector** in $\mathbb{R}^n$ and, if $\mathbf{x}$ is an n-vector, the n-vector $-\mathbf{x}$ is called the **negative $\mathbf{x}$**.

Of course, we have already encountered these n-vectors in Section 1.3 as the solutions to systems of linear equations with n variables. In particular we defined the notion of a linear combination of vectors and showed that a linear combination of solutions to a homogeneous system is again a solution. Clearly, a linear combination of n-vectors in $\mathbb{R}^n$ is again in $\mathbb{R}^n$, a fact that we will be using.

There is also a geometric interpretation that will be revisited in Chapter 4 and 5: elements in $\mathbb{R}^n$ can be viewed as points, such as the point $P(2,\ 3)$ in the plane $\mathbb{R}^2$, or as a vector $\vec{v} = \left[\begin{array}{c} 2 \\ 3 \end{array}\right]$ (as in an arrow) from the origin to the point $P(2,\ 3)$ and hence in anticipation the reason for introducing the name vector.

Matrix-Vector Multiplication

Given a system of linear equations, the left sides of the equations depend only on the coefficient matrix A and the column $\mathbf{x}$ of variables, and not on the constants. This observation leads to a fundamental idea in linear algebra: We view the left sides of the equations as the “product” $A\mathbf{x}$ of the matrix A and the vector $\mathbf{x}$. This simple change of perspective leads to a completely new way of viewing linear systems—one that is very useful and will occupy our attention throughout this book.

To motivate the definition of the “product” $A\mathbf{x}$, consider first the following system of two equations in three variables:

$$\begin{array}{r} ax_1 + \ bx_2 + \ cx_3 = b_1 \\ a'x_1 + b'x_2 + c'x_3 = b_1 \end{array} \tag{2.2}$$

and let $A = \left[\begin{array}{ccc} a & b & c \\ a' & b' & c' \end{array}\right]$, $\mathbf{x} = \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right]$, $\mathbf{b} = \left[\begin{array}{c} b_1 \\ b_2 \end{array}\right]$ denote the coefficient matrix, the variable matrix, and the constant matrix, respectively. The system (2.2) can be expressed as a single vector equation

$$\left[\begin{array}{c} ax_1 + \ bx_2 + \ cx_3 \\ a'x_1 + b'x_2 + c'x_3 \end{array}\right] = \left[\begin{array}{c} b_1 \\ b_2 \end{array}\right]$$

which in turn can be written as follows:

$$x_1 \left[\begin{array}{c} a \\ a' \end{array}\right] + x_2 \left[\begin{array}{c} b \\ b' \end{array}\right] + x_3 \left[\begin{array}{c} c \\ c' \end{array}\right] = \left[\begin{array}{c} b_1 \\ b_2 \end{array}\right]$$

Now observe that the vectors appearing on the left side are just the columns

$$\mathbf{a}_1 = \left[\begin{array}{c} a \\ a' \end{array}\right],\ \mathbf{a}_2 = \left[\begin{array}{c} b \\ b' \end{array}\right],\ \text{and } \mathbf{a}_3 = \left[\begin{array}{c} c \\ c' \end{array}\right]$$

of the coefficient matrix A. Hence the system (2.2) takes the form

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + x_3\mathbf{a}_3 = \mathbf{b} \tag{2.3}$$

This shows that the system (2.2) has a solution if and only if the constant matrix $\mathbf{b}$ is a linear combination[3] of the columns of A, and that in this case the entries of the solution are the coefficients x_1, x_2, and x_3 in this linear combination.

Moreover, this holds in general. If A is any $m \times n$ matrix, it is often convenient to view A as a row of columns. That is, if $\mathbf{a}_1,\ \mathbf{a}_2,\ \dots,\ \mathbf{a}_n$ are the columns of A, we write

$$A = \left[\begin{array}{cccc} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{array}\right]$$

and say that $A = \left[\begin{array}{cccc} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{array}\right]$ is *given in terms of its columns.*

[3] Linear combinations were introduced in Section 1.3 to describe the solutions of homogeneous systems of linear equations. They will be used extensively in what follows.

Now consider any system of linear equations with $m \times n$ coefficient matrix A. If $\mathbf{b}$ is the constant matrix of the system, and if $\mathbf{x} = \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \right]$ is the matrix of variables then, exactly as above, the system can be written as a single vector equation

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \mathbf{b} \tag{2.4}$$

Example 2.2.1

Write the system $\left\{ \begin{array}{rrrrrrr} 3x_1 & + & 2x_2 & - & 4x_3 & = & 0 \\ x_1 & - & 3x_2 & + & x_3 & = & 3 \\ & & x_2 & - & 5x_3 & = & -1 \end{array} \right.$ in the form given in (2.4).

Solution.

$$x_1 \left[\begin{array}{r} 3 \\ 1 \\ 0 \end{array} \right] + x_2 \left[\begin{array}{r} 2 \\ -3 \\ 1 \end{array} \right] + x_3 \left[\begin{array}{r} -4 \\ 1 \\ -5 \end{array} \right] = \left[\begin{array}{r} 0 \\ 3 \\ -1 \end{array} \right]$$

As mentioned above, we view the left side of (2.4) as the *product* of the matrix A and the vector $\mathbf{x}$. This basic idea is formalized in the following definition:

Definition 2.5 Matrix-Vector Multiplication

*Let $A = \left[\begin{array}{cccc} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{array} \right]$ be an $m \times n$ matrix, written in terms of its columns $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$. If $\mathbf{x} = \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \right]$ is any n-vector, the **product** $A\mathbf{x}$ is defined to be the m-vector given by:*

$$A\mathbf{x} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n$$

In other words, if A is $m \times n$ and $\mathbf{x}$ is an n-vector, the product $A\mathbf{x}$ is the linear combination of the columns of A where the coefficients are the entries of $\mathbf{x}$ (in order).

Note that if A is an $m \times n$ matrix, the product $A\mathbf{x}$ is only defined if $\mathbf{x}$ is an n-vector and then the vector $A\mathbf{x}$ is an m-vector because this is true of each column $\mathbf{a}_j$ of A. But in this case the *system* of linear equations with coefficient matrix A and constant vector $\mathbf{b}$ takes the form of a *single* matrix equation

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

The following theorem combines Definition 2.5 and equation (2.4) and summarizes the above discussion. Recall that a system of linear equations is said to be *consistent* if it has at least one solution.

Theorem 2.2.1

1. *Every system of linear equations has the form* $A\mathbf{x} = \mathbf{b}$ *where* A *is the coefficient matrix,* $\mathbf{b}$ *is the constant matrix, and* $\mathbf{x}$ *is the matrix of variables.*

2. *The system* $A\mathbf{x} = \mathbf{b}$ *is consistent if and only if* $\mathbf{b}$ *is a linear combination of the columns of* A.

3. *If* $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ *are the columns of* A *and if* $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, *then* $\mathbf{x}$ *is a solution to the linear system* $A\mathbf{x} = \mathbf{b}$ *if and only if* $x_1, x_2, \dots, x_n$ *are a solution of the vector equation*

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \mathbf{b}$$

A system of linear equations in the form $A\mathbf{x} = \mathbf{b}$ as in (1) of Theorem 2.2.1 is said to be written in **matrix form**. This is a useful way to view linear systems as we shall see.

Theorem 2.2.1 transforms the problem of solving the linear system $A\mathbf{x} = \mathbf{b}$ into the problem of expressing the constant matrix B as a linear combination of the columns of the coefficient matrix A. Such a change in perspective is very useful because one approach or the other may be better in a particular situation; the importance of the theorem is that there is a choice.

Example 2.2.2

If $A = \begin{bmatrix} 2 & -1 & 3 & 5 \\ 0 & 2 & -3 & 1 \\ -3 & 4 & 1 & 2 \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} 2 \\ 1 \\ 0 \\ -2 \end{bmatrix}$, compute $A\mathbf{x}$.

Solution. By Definition 2.5: $A\mathbf{x} = 2\begin{bmatrix} 2 \\ 0 \\ -3 \end{bmatrix} + 1\begin{bmatrix} -1 \\ 2 \\ 4 \end{bmatrix} + 0\begin{bmatrix} 3 \\ -3 \\ 1 \end{bmatrix} - 2\begin{bmatrix} 5 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -7 \\ 0 \\ -6 \end{bmatrix}$.

Example 2.2.3

Given columns $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, and $\mathbf{a}_4$ in $\mathbb{R}^3$, write $2\mathbf{a}_1 - 3\mathbf{a}_2 + 5\mathbf{a}_3 + \mathbf{a}_4$ in the form $A\mathbf{x}$ where A is a matrix and $\mathbf{x}$ is a vector.

Solution. Here the column of coefficients is $\mathbf{x} = \begin{bmatrix} 2 \\ -3 \\ 5 \\ 1 \end{bmatrix}$. Hence Definition 2.5 gives

$$A\mathbf{x} = 2\mathbf{a}_1 - 3\mathbf{a}_2 + 5\mathbf{a}_3 + \mathbf{a}_4$$

where $A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 & \mathbf{a}_4 \end{bmatrix}$ is the matrix with $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, and $\mathbf{a}_4$ as its columns.

Example 2.2.4

Let $A = \left[\begin{array}{cccc} \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 & \mathbf{a}_4 \end{array} \right]$ be the 3×4 matrix given in terms of its columns $\mathbf{a}_1 = \left[\begin{array}{r} 2 \\ 0 \\ -1 \end{array} \right]$, $\mathbf{a}_2 = \left[\begin{array}{r} 1 \\ 1 \\ 1 \end{array} \right]$, $\mathbf{a}_3 = \left[\begin{array}{r} 3 \\ -1 \\ -3 \end{array} \right]$, and $\mathbf{a}_4 = \left[\begin{array}{r} 3 \\ 1 \\ 0 \end{array} \right]$. In each case below, either express $\mathbf{b}$ as a linear combination of $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, and $\mathbf{a}_4$, or show that it is not such a linear combination. Explain what your answer means for the corresponding system $A\mathbf{x} = \mathbf{b}$ of linear equations.

a. $\mathbf{b} = \left[\begin{array}{r} 1 \\ 2 \\ 3 \end{array} \right]$

b. $\mathbf{b} = \left[\begin{array}{r} 4 \\ 2 \\ 1 \end{array} \right]$

Solution. By Theorem 2.2.1, $\mathbf{b}$ is a linear combination of $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, and $\mathbf{a}_4$ if and only if the system $A\mathbf{x} = \mathbf{b}$ is consistent (that is, it has a solution). So in each case we carry the augmented matrix $[A|\mathbf{b}]$ of the system $A\mathbf{x} = \mathbf{b}$ to reduced form.

a. Here $\left[\begin{array}{rrrr|r} 2 & 1 & 3 & 3 & 1 \\ 0 & 1 & -1 & 1 & 2 \\ -1 & 1 & -3 & 0 & 3 \end{array} \right] \rightarrow \left[\begin{array}{rrrr|r} 1 & 0 & 2 & 1 & 0 \\ 0 & 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array} \right]$, so the system $A\mathbf{x} = \mathbf{b}$ has no solution in this case. Hence $\mathbf{b}$ is *not* a linear combination of $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, and $\mathbf{a}_4$.

b. Now $\left[\begin{array}{rrrr|r} 2 & 1 & 3 & 3 & 4 \\ 0 & 1 & -1 & 1 & 2 \\ -1 & 1 & -3 & 0 & 1 \end{array} \right] \rightarrow \left[\begin{array}{rrrr|r} 1 & 0 & 2 & 1 & 1 \\ 0 & 1 & -1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right]$, so the system $A\mathbf{x} = \mathbf{b}$ is consistent.

Thus $\mathbf{b}$ is a linear combination of $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, and $\mathbf{a}_4$ in this case. In fact the general solution is $x_1 = 1 - 2s - t$, $x_2 = 2 + s - t$, $x_3 = s$, and $x_4 = t$ where s and t are arbitrary parameters. Hence $x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + x_3\mathbf{a}_3 + x_4\mathbf{a}_4 = \mathbf{b} = \left[\begin{array}{r} 4 \\ 2 \\ 1 \end{array} \right]$ for *any* choice of s and t. If we take $s = 0$ and $t = 0$, this becomes $\mathbf{a}_1 + 2\mathbf{a}_2 = \mathbf{b}$, whereas taking $s = 1 = t$ gives $-2\mathbf{a}_1 + 2\mathbf{a}_2 + \mathbf{a}_3 + \mathbf{a}_4 = \mathbf{b}$.

Example 2.2.5

Taking A to be the zero matrix, we have $0\mathbf{x} = \mathbf{0}$ for all vectors $\mathbf{x}$ by Definition 2.5 because every column of the zero matrix is zero. Similarly, $A\mathbf{0} = \mathbf{0}$ for all matrices A because every entry of the zero vector is zero.

Example 2.2.6

If $I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, show that $I\mathbf{x} = \mathbf{x}$ for any vector $\mathbf{x}$ in $\mathbb{R}^3$.

Solution. If $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ then Definition 2.5 gives

$$I\mathbf{x} = x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} x_1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ x_2 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{x}$$

The matrix I in Example 2.2.6 is called the 3×3 **identity matrix**, and we will encounter such matrices again in Example 2.2.11 below. Before proceeding, we develop some algebraic properties of matrix-vector multiplication that are used extensively throughout linear algebra.

Theorem 2.2.2

Let A and B be $m \times n$ matrices, and let $\mathbf{x}$ and $\mathbf{y}$ be n-vectors in $\mathbb{R}^n$. Then:

1. $A(\mathbf{x} + \mathbf{y}) = A\mathbf{x} + A\mathbf{y}$.
2. $A(a\mathbf{x}) = a(A\mathbf{x}) = (aA)\mathbf{x}$ *for all scalars a.*
3. $(A + B)\mathbf{x} = A\mathbf{x} + B\mathbf{x}$.

Proof. We prove (3); the other verifications are similar and are left as exercises. Let $A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix}$ and $B = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_n \end{bmatrix}$ be given in terms of their columns. Since adding two matrices is the same as adding their columns, we have

$$A + B = \begin{bmatrix} \mathbf{a}_1 + \mathbf{b}_1 & \mathbf{a}_2 + \mathbf{b}_2 & \cdots & \mathbf{a}_n + \mathbf{b}_n \end{bmatrix}$$

If we write $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ Definition 2.5 gives

$$\begin{aligned} (A + B)\mathbf{x} &= x_1(\mathbf{a}_1 + \mathbf{b}_1) + x_2(\mathbf{a}_2 + \mathbf{b}_2) + \cdots + x_n(\mathbf{a}_n + \mathbf{b}_n) \\ &= (x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n) + (x_1\mathbf{b}_1 + x_2\mathbf{b}_2 + \cdots + x_n\mathbf{b}_n) \\ &= A\mathbf{x} + B\mathbf{x} \end{aligned}$$

□

Theorem 2.2.2 allows matrix-vector computations to be carried out much as in ordinary arithmetic. For example, for any $m \times n$ matrices A and B and any n-vectors $\mathbf{x}$ and $\mathbf{y}$, we have:

$$A(2\mathbf{x} - 5\mathbf{y}) = 2A\mathbf{x} - 5A\mathbf{y} \quad \text{and} \quad (3A - 7B)\mathbf{x} = 3A\mathbf{x} - 7B\mathbf{x}$$

We will use such manipulations throughout the book, often without mention.

Linear Equations

Theorem 2.2.2 also gives a useful way to describe the solutions to a system

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

of linear equations. There is a related system

$$A\mathbf{x} = \mathbf{0}$$

called the **associated homogeneous system**, obtained from the original system $A\mathbf{x} = \mathbf{b}$ by replacing all the constants by zeros. Suppose $\mathbf{x}_1$ is a solution to $A\mathbf{x} = \mathbf{b}$ and $\mathbf{x}_0$ is a solution to $A\mathbf{x} = \mathbf{0}$ (that is $A\mathbf{x}_1 = \mathbf{b}$ and $A\mathbf{x}_0 = \mathbf{0}$). Then $\mathbf{x}_1 + \mathbf{x}_0$ is another solution to $A\mathbf{x} = \mathbf{b}$. Indeed, Theorem 2.2.2 gives

$$A(\mathbf{x}_1 + \mathbf{x}_0) = A\mathbf{x}_1 + A\mathbf{x}_0 = \mathbf{b} + \mathbf{0} = \mathbf{b}$$

This observation has a useful converse.

Theorem 2.2.3

Suppose $\mathbf{x}_1$ is any particular solution to the system $A\mathbf{x} = \mathbf{b}$ of linear equations. Then every solution $\mathbf{x}_2$ to $A\mathbf{x} = \mathbf{b}$ has the form

$$\mathbf{x}_2 = \mathbf{x}_0 + \mathbf{x}_1$$

for some solution $\mathbf{x}_0$ of the associated homogeneous system $A\mathbf{x} = \mathbf{0}$.

Proof. Suppose $\mathbf{x}_2$ is also a solution to $A\mathbf{x} = \mathbf{b}$, so that $A\mathbf{x}_2 = \mathbf{b}$. Write $\mathbf{x}_0 = \mathbf{x}_2 - \mathbf{x}_1$. Then $\mathbf{x}_2 = \mathbf{x}_0 + \mathbf{x}_1$ and, using Theorem 2.2.2, we compute

$$A\mathbf{x}_0 = A(\mathbf{x}_2 - \mathbf{x}_1) = A\mathbf{x}_2 - A\mathbf{x}_1 = \mathbf{b} - \mathbf{b} = \mathbf{0}$$

Hence $\mathbf{x}_0$ is a solution to the associated homogeneous system $A\mathbf{x} = \mathbf{0}$. □

Note that gaussian elimination provides one such representation.

Example 2.2.7

Express every solution to the following system as the sum of a specific solution plus a solution to the associated homogeneous system.

$$\begin{aligned} x_1 - x_2 - x_3 + 3x_4 &= 2 \\ 2x_1 - x_2 - 3x_3 + 4x_4 &= 6 \\ x_1 \phantom{{}- x_2} - 2x_3 + x_4 &= 4 \end{aligned}$$

Solution. Gaussian elimination gives $x_1 = 4 + 2s - t$, $x_2 = 2 + s + 2t$, $x_3 = s$, and $x_4 = t$ where s and t are arbitrary parameters. Hence the general solution can be written

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 + 2s - t \\ 2 + s + 2t \\ s \\ t \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \\ 0 \\ 0 \end{bmatrix} + \left(s \begin{bmatrix} 2 \\ 1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 \\ 2 \\ 0 \\ 1 \end{bmatrix} \right)$$

Thus $\mathbf{x}_1 = \left[\begin{array}{r} 4 \\ 2 \\ 0 \\ 0 \end{array} \right]$ is a particular solution (where $s = 0 = t$), and $\mathbf{x}_0 = s \left[\begin{array}{r} 2 \\ 1 \\ 1 \\ 0 \end{array} \right] + t \left[\begin{array}{r} -1 \\ 2 \\ 0 \\ 1 \end{array} \right]$ gives *all* solutions to the associated homogeneous system. (To see why this is so, carry out the gaussian elimination again but with all the constants set equal to zero.)

The following useful result is included with no proof.

Theorem 2.2.4

Let $A\mathbf{x} = \mathbf{b}$ be a system of equations with augmented matrix $\left[\begin{array}{c|c} A & \mathbf{b} \end{array} \right]$. Write $\text{rank } A = r$.

1. $\text{rank} \left[\begin{array}{c|c} A & \mathbf{b} \end{array} \right]$ *is either r or $r+1$.*
2. *The system is consistent if and only if* $\text{rank} \left[\begin{array}{c|c} A & \mathbf{b} \end{array} \right] = r$.
3. *The system is inconsistent if and only if* $\text{rank} \left[\begin{array}{c|c} A & \mathbf{b} \end{array} \right] = r+1$.

The Dot Product

Definition 2.5 is not always the easiest way to compute a matrix-vector product $A\mathbf{x}$ because it requires that the columns of A be explicitly identified. There is another way to find such a product which uses the matrix A as a whole with no reference to its columns, and hence is useful in practice. The method depends on the following notion.

Definition 2.6 Dot Product in $\mathbb{R}^n$

If $(a_1, a_2, \dots, a_n)$ and $(b_1, b_2, \dots, b_n)$ are two ordered n-tuples, their ***dot product*** *is defined to be the number*

$$a_1b_1 + a_2b_2 + \cdots + a_nb_n$$

obtained by multiplying corresponding entries and adding the results.

To see how this relates to matrix products, let A denote a 3×4 matrix and let $\mathbf{x}$ be a 4-vector. Writing

$$\mathbf{x} = \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \right] \quad \text{and} \quad A = \left[\begin{array}{cccc} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{array} \right]$$

in the notation of Section 2.1, we compute

$$A\mathbf{x} = \left[\begin{array}{cccc} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{array} \right] \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \right] = x_1 \left[\begin{array}{c} a_{11} \\ a_{21} \\ a_{31} \end{array} \right] + x_2 \left[\begin{array}{c} a_{12} \\ a_{22} \\ a_{32} \end{array} \right] + x_3 \left[\begin{array}{c} a_{13} \\ a_{23} \\ a_{33} \end{array} \right] + x_4 \left[\begin{array}{c} a_{14} \\ a_{24} \\ a_{34} \end{array} \right]$$

$$= \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 \end{bmatrix}$$

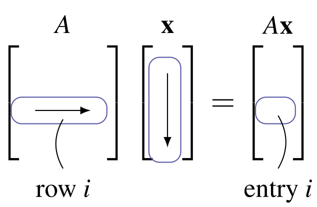
From this we see that each entry of $A\mathbf{x}$ is the dot product of the corresponding row of A with $\mathbf{x}$. This computation goes through in general, and we record the result in Theorem 2.2.5.

Theorem 2.2.5: Dot Product Rule

Let A be an $m \times n$ matrix and let $\mathbf{x}$ be an n-vector. Then each entry of the vector $A\mathbf{x}$ is the dot product of the corresponding row of A with $\mathbf{x}$.

This result is used extensively throughout linear algebra.

If A is $m \times n$ and $\mathbf{x}$ is an n-vector, the computation of $A\mathbf{x}$ by the dot product rule is simpler than using Definition 2.5 because the computation can be carried out directly with no explicit reference to the columns of A (as in Definition 2.5). The first entry of $A\mathbf{x}$ is the dot product of row 1 of A with $\mathbf{x}$. In hand calculations this is computed by going *across* row one of A, going *down* the column $\mathbf{x}$, multiplying corresponding entries, and adding the results. The other entries of $A\mathbf{x}$ are computed in the same way using the other rows of A with the column $\mathbf{x}$.



In general, compute entry i of $A\mathbf{x}$ as follows (see the diagram):

> Go *across* row i of A and *down* column $\mathbf{x}$, multiply corresponding entries, and add the results.

As an illustration, we rework Example 2.2.2 using the dot product rule instead of Definition 2.5.

Example 2.2.8

If $A = \begin{bmatrix} 2 & -1 & 3 & 5 \\ 0 & 2 & -3 & 1 \\ -3 & 4 & 1 & 2 \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} 2 \\ 1 \\ 0 \\ -2 \end{bmatrix}$, compute $A\mathbf{x}$.

Solution. The entries of $A\mathbf{x}$ are the dot products of the rows of A with $\mathbf{x}$:

$$A\mathbf{x} = \begin{bmatrix} 2 & -1 & 3 & 5 \\ 0 & 2 & -3 & 1 \\ -3 & 4 & 1 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 0 \\ -2 \end{bmatrix} = \begin{bmatrix} 2 \cdot 2 & + & (-1)1 & + & 3 \cdot 0 & + & 5(-2) \\ 0 \cdot 2 & + & 2 \cdot 1 & + & (-3)0 & + & 1(-2) \\ (-3)2 & + & 4 \cdot 1 & + & 1 \cdot 0 & + & 2(-2) \end{bmatrix} = \begin{bmatrix} -7 \\ 0 \\ -6 \end{bmatrix}$$

Of course, this agrees with the outcome in Example 2.2.2.

Example 2.2.9

Write the following system of linear equations in the form $A\mathbf{x} = \mathbf{b}$.

$$\begin{aligned} 5x_1 - x_2 + 2x_3 + x_4 - 3x_5 &= 8 \\ x_1 + x_2 + 3x_3 - 5x_4 + 2x_5 &= -2 \\ -x_1 + x_2 - 2x_3 + - 3x_5 &= 0 \end{aligned}$$

Solution. Write $A = \begin{bmatrix} 5 & -1 & 2 & 1 & -3 \\ 1 & 1 & 3 & -5 & 2 \\ -1 & 1 & -2 & 0 & -3 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} 8 \\ -2 \\ 0 \end{bmatrix}$, and $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}$. Then the dot product rule gives $A\mathbf{x} = \begin{bmatrix} 5x_1 - x_2 + 2x_3 + x_4 - 3x_5 \\ x_1 + x_2 + 3x_3 - 5x_4 + 2x_5 \\ -x_1 + x_2 - 2x_3 - 3x_5 \end{bmatrix}$, so the entries of $A\mathbf{x}$ are the left sides of the equations in the linear system. Hence the system becomes $A\mathbf{x} = \mathbf{b}$ because matrices are equal if and only corresponding entries are equal.

Example 2.2.10

If A is the zero $m \times n$ matrix, then $A\mathbf{x} = \mathbf{0}$ for each n-vector $\mathbf{x}$.

Solution. For each k, entry k of $A\mathbf{x}$ is the dot product of row k of A with $\mathbf{x}$, and this is zero because row k of A consists of zeros.

Definition 2.7 The Identity Matrix

For each $n > 2$, the ***identity matrix*** *I_n is the $n \times n$ matrix with 1s on the main diagonal (upper left to lower right), and zeros elsewhere.*

The first few identity matrices are

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \dots$$

In Example 2.2.6 we showed that $I_3\mathbf{x} = \mathbf{x}$ for each 3-vector $\mathbf{x}$ using Definition 2.5. The following result shows that this holds in general, and is the reason for the name.

Example 2.2.11

For each $n \geq 2$ we have $I_n\mathbf{x} = \mathbf{x}$ for each n-vector $\mathbf{x}$ in $\mathbb{R}^n$.

Solution. We verify the case $n = 4$. Given the 4-vector $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$ the dot product rule gives

$$I_4\mathbf{x} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_1+0+0+0 \\ 0+x_2+0+0 \\ 0+0+x_3+0 \\ 0+0+0+x_4 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \mathbf{x}$$

In general, $I_n\mathbf{x} = \mathbf{x}$ because entry k of $I_n\mathbf{x}$ is the dot product of row k of I_n with $\mathbf{x}$, and row k of I_n has 1 in position k and zeros elsewhere.

Example 2.2.12

Let $A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix}$ be any $m \times n$ matrix with columns $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$. If $\mathbf{e}_j$ denotes column j of the $n \times n$ identity matrix I_n, then $A\mathbf{e}_j = \mathbf{a}_j$ for each $j = 1, 2, \dots, n$.

Solution. Write $\mathbf{e}_j = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}$ where $t_j = 1$, but $t_i = 0$ for all $i \neq j$. Then Theorem 2.2.5 gives

$$A\mathbf{e}_j = t_1\mathbf{a}_1 + \cdots + t_j\mathbf{a}_j + \cdots + t_n\mathbf{a}_n = 0 + \cdots + \mathbf{a}_j + \cdots + 0 = \mathbf{a}_j$$

Example 2.2.12 will be referred to later; for now we use it to prove:

Theorem 2.2.6

Let A and B be $m \times n$ matrices. If $A\mathbf{x} = B\mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^n$, then $A = B$.

Proof. Write $A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix}$ and $B = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_n \end{bmatrix}$ and in terms of their columns. It is enough to show that $\mathbf{a}_k = \mathbf{b}_k$ holds for all k. But we are assuming that $A\mathbf{e}_k = B\mathbf{e}_k$, which gives $\mathbf{a}_k = \mathbf{b}_k$ by Example 2.2.12. □

We have introduced matrix-vector multiplication as a new way to think about systems of linear equations. But it has several other uses as well. It turns out that many geometric operations can be described using matrix multiplication, and we now investigate how this happens. As a bonus, this description provides a geometric "picture" of a matrix by revealing the effect on a vector when it is multiplied by A. This "geometric view" of matrices is a fundamental tool in understanding them.

Transformations

The set $\mathbb{R}^2$ has a geometrical interpretation as the euclidean plane where a vector $\left[\begin{array}{c} a_1 \\ a_2 \end{array}\right]$ in $\mathbb{R}^2$ represents the point $(a_1,\ a_2)$ in the plane (see Figure 2.2.1). In this way we regard $\mathbb{R}^2$ as the set of all points in the plane. Accordingly, we will refer to vectors in $\mathbb{R}^2$ as points, and denote their coordinates as a column rather than a row. To enhance this geometrical interpretation of the vector $\left[\begin{array}{c} a_1 \\ a_2 \end{array}\right]$, it is denoted graphically by an arrow from the origin $\left[\begin{array}{c} 0 \\ 0 \end{array}\right]$ to the vector as in Figure 2.2.1.

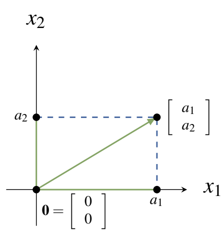


Figure 2.2.1

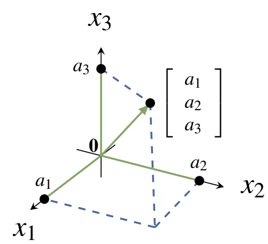


Figure 2.2.2

Similarly we identify $\mathbb{R}^3$ with 3-dimensional space by writing a point $(a_1,\ a_2,\ a_3)$ as the vector $\left[\begin{array}{c} a_1 \\ a_2 \\ a_3 \end{array}\right]$ in $\mathbb{R}^3$, again represented by an arrow[4] from the origin to the point as in Figure 2.2.2. In this way the terms "point" and "vector" mean the same thing in the plane or in space.

We begin by describing a particular geometrical transformation of the plane $\mathbb{R}^2$.

Example 2.2.13

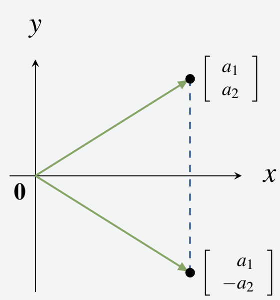


Figure 2.2.3

Consider the transformation of $\mathbb{R}^2$ given by *reflection* in the x axis. This operation carries the vector $\left[\begin{array}{c} a_1 \\ a_2 \end{array}\right]$ to its reflection $\left[\begin{array}{c} a_1 \\ -a_2 \end{array}\right]$ as in Figure 2.2.3. Now observe that

$$\left[\begin{array}{c} a_1 \\ -a_2 \end{array}\right] = \left[\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right] \left[\begin{array}{c} a_1 \\ a_2 \end{array}\right]$$

so reflecting $\left[\begin{array}{c} a_1 \\ a_2 \end{array}\right]$ in the x axis can be achieved by multiplying by the matrix $\left[\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right]$.

[4]This "arrow" representation of vectors in $\mathbb{R}^2$ and $\mathbb{R}^3$ will be used extensively in Chapter 4.

If we write $A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$, Example 2.2.13 shows that reflection in the x axis carries each vector $\mathbf{x}$ in $\mathbb{R}^2$ to the vector $A\mathbf{x}$ in $\mathbb{R}^2$. It is thus an example of a function

$$T : \mathbb{R}^2 \to \mathbb{R}^2 \quad \text{where} \quad T(\mathbf{x}) = A\mathbf{x} \text{ for all } \mathbf{x} \text{ in } \mathbb{R}^2$$

As such it is a generalization of the familiar functions $f : \mathbb{R} \to \mathbb{R}$ that carry a *number* x to another real *number* $f(x)$.

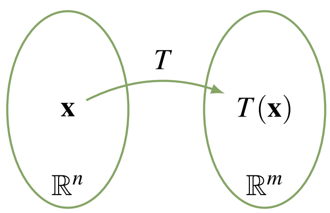


Figure 2.2.4

More generally, functions $T : \mathbb{R}^n \to \mathbb{R}^m$ are called **transformations** from $\mathbb{R}^n$ to $\mathbb{R}^m$. Such a transformation T is a rule that assigns to every vector $\mathbf{x}$ in $\mathbb{R}^n$ a uniquely determined vector $T(\mathbf{x})$ in $\mathbb{R}^m$ called the **image** of $\mathbf{x}$ under T. We denote this state of affairs by writing

$$T : \mathbb{R}^n \to \mathbb{R}^m \quad \text{or} \quad \mathbb{R}^n \xrightarrow{T} \mathbb{R}^m$$

The transformation T can be visualized as in Figure 2.2.4.

To describe a transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ we must specify the vector $T(\mathbf{x})$ in $\mathbb{R}^m$ for every $\mathbf{x}$ in $\mathbb{R}^n$. This is referred to as **defining** T, or as specifying the **action** of T. Saying that the action *defines* the transformation means that we regard two transformations $S : \mathbb{R}^n \to \mathbb{R}^m$ and $T : \mathbb{R}^n \to \mathbb{R}^m$ as **equal** if they have the **same action**; more formally

$$S = T \quad \text{if and only if} \quad S(\mathbf{x}) = T(\mathbf{x}) \text{ for all } \mathbf{x} \text{ in } \mathbb{R}^n.$$

Again, this what we mean by $f = g$ where $f,\ g : \mathbb{R} \to \mathbb{R}$ are ordinary functions.

Functions $f : \mathbb{R} \to \mathbb{R}$ are often described by a formula, examples being $f(x) = x^2 + 1$ and $f(x) = \sin x$. The same is true of transformations; here is an example.

Example 2.2.14

The formula $T \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 \\ x_2 + x_3 \\ x_3 + x_4 \end{bmatrix}$ defines a transformation $\mathbb{R}^4 \to \mathbb{R}^3$.

Example 2.2.13 suggests that matrix multiplication is an important way of defining transformations $\mathbb{R}^n \to \mathbb{R}^m$. If A is any $m \times n$ matrix, multiplication by A gives a transformation

$$T_A : \mathbb{R}^n \to \mathbb{R}^m \quad \text{defined by} \quad T_A(\mathbf{x}) = A\mathbf{x} \text{ for every } \mathbf{x} \text{ in } \mathbb{R}^n$$

Definition 2.8 Matrix Transformation T_A

*T_A is called the **matrix transformation induced** by A.*

Thus Example 2.2.13 shows that reflection in the x axis is the matrix transformation $\mathbb{R}^2 \to \mathbb{R}^2$ induced by the matrix $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$. Also, the transformation $R : \mathbb{R}^4 \to \mathbb{R}^3$ in Example 2.2.13 is the matrix transformation induced by the matrix

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \text{because} \quad \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 \\ x_2 + x_3 \\ x_3 + x_4 \end{bmatrix}$$

Example 2.2.15

Let $R_{\frac{\pi}{2}} : \mathbb{R}^2 \to \mathbb{R}^2$ denote counterclockwise rotation about the origin through $\frac{\pi}{2}$ radians (that is, $90°$)[5]. Show that $R_{\frac{\pi}{2}}$ is induced by the matrix $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$.

Solution.

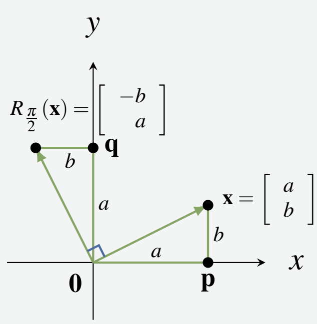


Figure 2.2.5

The effect of $R_{\frac{\pi}{2}}$ is to rotate the vector $\mathbf{x} = \begin{bmatrix} a \\ b \end{bmatrix}$ counterclockwise through $\frac{\pi}{2}$ to produce the vector $R_{\frac{\pi}{2}}(\mathbf{x})$ shown in Figure 2.2.5. Since triangles $\mathbf{0px}$ and $\mathbf{0q}R_{\frac{\pi}{2}}(\mathbf{x})$ are identical, we obtain $R_{\frac{\pi}{2}}(\mathbf{x}) = \begin{bmatrix} -b \\ a \end{bmatrix}$. But $\begin{bmatrix} -b \\ a \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$, so we obtain $R_{\frac{\pi}{2}}(\mathbf{x}) = A\mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^2$ where $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. In other words, $R_{\frac{\pi}{2}}$ is the matrix transformation induced by A.

If A is the $m \times n$ zero matrix, then A induces the transformation

$$T : \mathbb{R}^n \to \mathbb{R}^m \quad \text{given by} \quad T(\mathbf{x}) = A\mathbf{x} = \mathbf{0} \text{ for all } \mathbf{x} \text{ in } \mathbb{R}^n$$

This is called the **zero transformation**, and is denoted $T = 0$.

Another important example is the **identity transformation**

$$1_{\mathbb{R}^n} : \mathbb{R}^n \to \mathbb{R}^n \quad \text{given by} \quad 1_{\mathbb{R}^n}(\mathbf{x}) = \mathbf{x} \text{ for all } \mathbf{x} \text{ in } \mathbb{R}^n$$

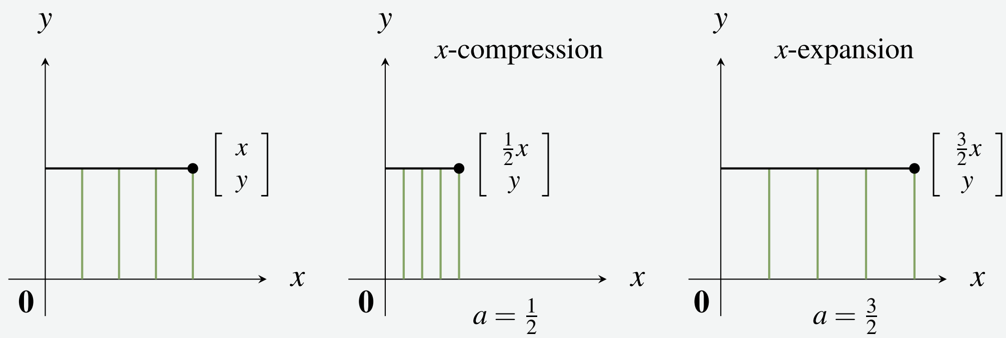
That is, the action of $1_{\mathbb{R}^n}$ on $\mathbf{x}$ is to do nothing to it. If I_n denotes the $n \times n$ identity matrix, we showed in Example 2.2.11 that $I_n\mathbf{x} = \mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^n$. Hence $1_{\mathbb{R}^n}(\mathbf{x}) = I_n\mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^n$; that is, the identity matrix I_n induces the identity transformation.

Here are two more examples of matrix transformations with a clear geometric description.

[5] *Radian measure* for angles is based on the fact that $360°$ equals 2π radians. Hence π radians $= 180°$ and $\frac{\pi}{2}$ radians $= 90°$.

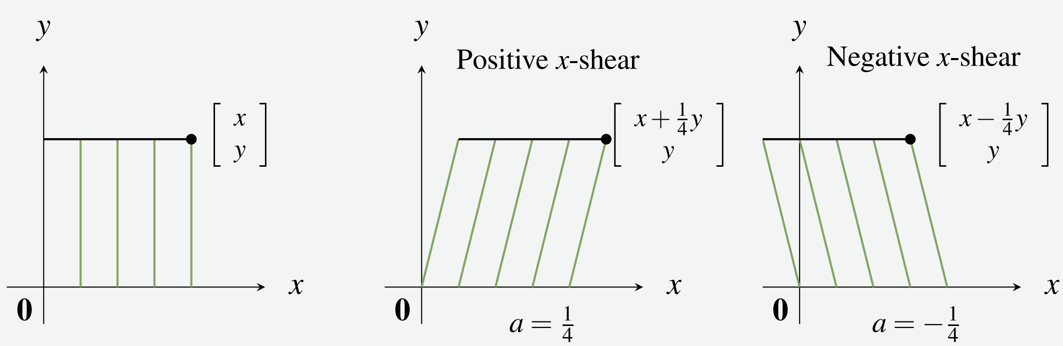
Example 2.2.16

If $a > 0$, the matrix transformation $T\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax \\ y \end{bmatrix}$ induced by the matrix $A = \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}$ is called an ***x*-expansion** of $\mathbb{R}^2$ if $a > 1$, and an ***x*-compression** if $0 < a < 1$. The reason for the names is clear in the diagram below. Similarly, if $b > 0$ the matrix $A = \begin{bmatrix} 1 & 0 \\ 0 & b \end{bmatrix}$ gives rise to ***y*-expansions** and ***y*-compressions**.



Example 2.2.17

If a is a number, the matrix transformation $T\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x + ay \\ y \end{bmatrix}$ induced by the matrix $A = \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix}$ is called an ***x*-shear** of $\mathbb{R}^2$ (**positive** if $a > 0$ and **negative** if $a < 0$). Its effect is illustrated below when $a = \frac{1}{4}$ and $a = -\frac{1}{4}$.



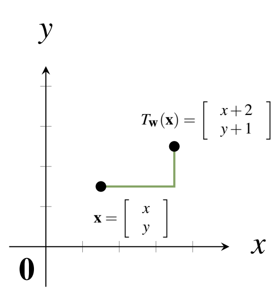


Figure 2.2.6

We hasten to note that there are important geometric transformations that are *not* matrix transformations. For example, if $\mathbf{w}$ is a fixed column in $\mathbb{R}^n$, define the transformation $T_{\mathbf{w}} : \mathbb{R}^n \to \mathbb{R}^n$ by

$$T_{\mathbf{w}}(\mathbf{x}) = \mathbf{x} + \mathbf{w} \quad \text{for all } \mathbf{x} \text{ in } \mathbb{R}^n$$

Then $T_{\mathbf{w}}$ is called **translation** by $\mathbf{w}$. In particular, if $\mathbf{w} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ in $\mathbb{R}^2$, the effect of $T_\mathbf{w}$ on $\left[\begin{array}{c} x \\ y \end{array}\right]$ is to translate it two units to the right and one unit up (see Figure 2.2.6).

The translation $T_\mathbf{w}$ is not a matrix transformation unless $\mathbf{w} = \mathbf{0}$. Indeed, if $T_\mathbf{w}$ were induced by a matrix A, then $A\mathbf{x} = T_\mathbf{w}(\mathbf{x}) = \mathbf{x} + \mathbf{w}$ would hold for every $\mathbf{x}$ in $\mathbb{R}^n$. In particular, taking $\mathbf{x} = \mathbf{0}$ gives $\mathbf{w} = A\mathbf{0} = \mathbf{0}$.

2.3 Matrix Multiplication

In Section 2.2 matrix-vector products were introduced. If A is an $m \times n$ matrix, the product $A\mathbf{x}$ was defined for any n-column $\mathbf{x}$ in $\mathbb{R}^n$ as follows: If $A = \left[\begin{array}{cccc} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{array}\right]$ where the $\mathbf{a}_j$ are the columns of A, and if $\mathbf{x} = \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array}\right]$, Definition 2.5 reads

$$A\mathbf{x} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n \tag{2.5}$$

This was motivated as a way of describing systems of linear equations with coefficient matrix A. Indeed every such system has the form $A\mathbf{x} = \mathbf{b}$ where $\mathbf{b}$ is the column of constants.

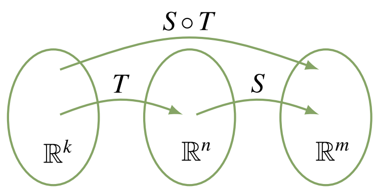
In this section we extend this matrix-vector multiplication to a way of multiplying matrices in general, and then investigate matrix algebra for its own sake. While it shares several properties of ordinary arithmetic, it will soon become clear that matrix arithmetic is different in a number of ways.

Matrix multiplication is closely related to composition of transformations.

Composition and Matrix Multiplication

Sometimes two transformations "link" together as follows:

$$\mathbb{R}^k \xrightarrow{T} \mathbb{R}^n \xrightarrow{S} \mathbb{R}^m$$



In this case we can apply T first and then apply S, and the result is a new transformation

$$S \circ T : \mathbb{R}^k \to \mathbb{R}^m$$

called the **composite** of S and T, defined by

$$(S \circ T)(\mathbf{x}) = S[T(\mathbf{x})] \quad \text{for all } \mathbf{x} \text{ in } \mathbb{R}^k$$

The action of $S \circ T$ can be described as "first T then S" (note the order!)[6]. This new transformation is described in the diagram. The reader will have encountered composition of ordinary functions: For example, consider $\mathbb{R} \xrightarrow{g} \mathbb{R} \xrightarrow{f} \mathbb{R}$ where $f(x) = x^2$ and $g(x) = x+1$ for all x in $\mathbb{R}$. Then

$$\begin{aligned}(f \circ g)(x) &= f[g(x)] = f(x+1) = (x+1)^2 \\ (g \circ f)(x) &= g[f(x)] = g(x^2) = x^2 + 1\end{aligned}$$

for all x in $\mathbb{R}$.

Our concern here is with matrix transformations. Suppose that A is an $m \times n$ matrix and B is an $n \times k$ matrix, and let $\mathbb{R}^k \xrightarrow{T_B} \mathbb{R}^n \xrightarrow{T_A} \mathbb{R}^m$ be the matrix transformations induced by B and A respectively, that is:

$$T_B(\mathbf{x}) = B\mathbf{x} \text{ for all } \mathbf{x} \text{ in } \mathbb{R}^k \quad \text{and} \quad T_A(\mathbf{y}) = A\mathbf{y} \text{ for all } \mathbf{y} \text{ in } \mathbb{R}^n$$

Write $B = \left[\begin{array}{cccc} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_k \end{array} \right]$ where $\mathbf{b}_j$ denotes column j of B for each j. Hence each $\mathbf{b}_j$ is an n-vector (B is $n \times k$) so we can form the matrix-vector product $A\mathbf{b}_j$. In particular, we obtain an $m \times k$ matrix

$$\left[\begin{array}{cccc} A\mathbf{b}_1 & A\mathbf{b}_2 & \cdots & A\mathbf{b}_k \end{array} \right]$$

with columns $A\mathbf{b}_1, A\mathbf{b}_2, \cdots, A\mathbf{b}_k$. Now compute $(T_A \circ T_B)(\mathbf{x})$ for any $\mathbf{x} = \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_k \end{array} \right]$ in $\mathbb{R}^k$:

$$\begin{array}{rcll} (T_A \circ T_B)(\mathbf{x}) & = & T_A[T_B(\mathbf{x})] & \text{Definition of } T_A \circ T_B \\ & = & A(B\mathbf{x}) & A \text{ and } B \text{ induce } T_A \text{ and } T_B \\ & = & A(x_1\mathbf{b}_1 + x_2\mathbf{b}_2 + \cdots + x_k\mathbf{b}_k) & \text{Equation 2.5 above} \\ & = & A(x_1\mathbf{b}_1) + A(x_2\mathbf{b}_2) + \cdots + A(x_k\mathbf{b}_k) & \text{Theorem 2.2.2} \\ & = & x_1(A\mathbf{b}_1) + x_2(A\mathbf{b}_2) + \cdots + x_k(A\mathbf{b}_k) & \text{Theorem 2.2.2} \\ & = & \left[\begin{array}{cccc} A\mathbf{b}_1 & A\mathbf{b}_2 & \cdots & A\mathbf{b}_k \end{array} \right] \mathbf{x} & \text{Equation 2.5 above} \end{array}$$

Because $\mathbf{x}$ was an arbitrary vector in $\mathbb{R}^n$, this shows that $T_A \circ T_B$ is the matrix transformation induced by the matrix $\left[\begin{array}{cccc} A\mathbf{b}_1 & A\mathbf{b}_2 & \cdots & A\mathbf{b}_n \end{array} \right]$. This motivates the following definition.

[6]When reading the notation $S \circ T$, we read S first and then T even though the action is "first T then S". This annoying state of affairs results because we write $T(\mathbf{x})$ for the effect of the transformation T on $\mathbf{x}$, with T on the left. If we wrote this instead as $(\mathbf{x})T$, the confusion would not occur. However the notation $T(\mathbf{x})$ is well established.

Definition 2.9 Matrix Multiplication

Let A be an $m \times n$ matrix, let B be an $n \times k$ matrix, and write $B = \left[\begin{array}{cccc} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_k \end{array} \right]$ where $\mathbf{b}_j$ is column j of B for each j. The product matrix AB is the $m \times k$ matrix defined as follows:

$$AB = A \left[\begin{array}{cccc} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_k \end{array} \right] = \left[\begin{array}{cccc} A\mathbf{b}_1 & A\mathbf{b}_2 & \cdots & A\mathbf{b}_k \end{array} \right]$$

Thus the product matrix AB is given in terms of its columns $A\mathbf{b}_1, A\mathbf{b}_2, \dots, A\mathbf{b}_n$: Column j of AB is the matrix-vector product $A\mathbf{b}_j$ of A and the corresponding column $\mathbf{b}_j$ of B. Note that each such product $A\mathbf{b}_j$ makes sense by Definition 2.5 because A is $m \times n$ and each $\mathbf{b}_j$ is in $\mathbb{R}^n$ (since B has n rows). Note also that if B is a column matrix, this definition reduces to Definition 2.5 for matrix-vector multiplication.

Given matrices A and B, Definition 2.9 and the above computation give

$$A(B\mathbf{x}) = \left[\begin{array}{cccc} A\mathbf{b}_1 & A\mathbf{b}_2 & \cdots & A\mathbf{b}_n \end{array} \right] \mathbf{x} = (AB)\mathbf{x}$$

for all $\mathbf{x}$ in $\mathbb{R}^k$. We record this for reference.

Theorem 2.3.1

Let A be an $m \times n$ matrix and let B be an $n \times k$ matrix. Then the product matrix AB is $m \times k$ and satisfies

$$A(B\mathbf{x}) = (AB)\mathbf{x} \quad \text{for all } \mathbf{x} \text{ in } \mathbb{R}^k$$

Here is an example of how to compute the product AB of two matrices using Definition 2.9.

Example 2.3.1

Compute AB if $A = \left[\begin{array}{ccc} 2 & 3 & 5 \\ 1 & 4 & 7 \\ 0 & 1 & 8 \end{array} \right]$ and $B = \left[\begin{array}{cc} 8 & 9 \\ 7 & 2 \\ 6 & 1 \end{array} \right]$.

Solution. The columns of B are $\mathbf{b}_1 = \left[\begin{array}{c} 8 \\ 7 \\ 6 \end{array} \right]$ and $\mathbf{b}_2 = \left[\begin{array}{c} 9 \\ 2 \\ 1 \end{array} \right]$, so Definition 2.5 gives

$$A\mathbf{b}_1 = \left[\begin{array}{ccc} 2 & 3 & 5 \\ 1 & 4 & 7 \\ 0 & 1 & 8 \end{array} \right] \left[\begin{array}{c} 8 \\ 7 \\ 6 \end{array} \right] = \left[\begin{array}{c} 67 \\ 78 \\ 55 \end{array} \right] \text{ and } A\mathbf{b}_2 = \left[\begin{array}{ccc} 2 & 3 & 5 \\ 1 & 4 & 7 \\ 0 & 1 & 8 \end{array} \right] \left[\begin{array}{c} 9 \\ 2 \\ 1 \end{array} \right] = \left[\begin{array}{c} 29 \\ 24 \\ 10 \end{array} \right]$$

Hence Definition 2.9 above gives $AB = \left[\begin{array}{cc} A\mathbf{b}_1 & A\mathbf{b}_2 \end{array} \right] = \left[\begin{array}{cc} 67 & 29 \\ 78 & 24 \\ 55 & 10 \end{array} \right]$.

Example 2.3.2

If A is $m \times n$ and B is $n \times k$, Theorem 2.3.1 gives a simple formula for the composite of the matrix transformations T_A and T_B:

$$T_A \circ T_B = T_{AB}$$

Solution. Given any $\mathbf{x}$ in $\mathbb{R}^k$,

$$\begin{aligned}(T_A \circ T_B)(\mathbf{x}) &= T_A[T_B(\mathbf{x})] \\ &= A[B\mathbf{x}] \\ &= (AB)\mathbf{x} \\ &= T_{AB}(\mathbf{x})\end{aligned}$$

While Definition 2.9 is important, there is another way to compute the matrix product AB that gives a way to calculate each individual entry. In Section 2.2 we defined the dot product of two n-tuples to be the sum of the products of corresponding entries. We went on to show (Theorem 2.2.5) that if A is an $m \times n$ matrix and $\mathbf{x}$ is an n-vector, then entry j of the product $A\mathbf{x}$ is the dot product of row j of A with $\mathbf{x}$. This observation was called the "dot product rule" for matrix-vector multiplication, and the next theorem shows that it extends to matrix multiplication in general.

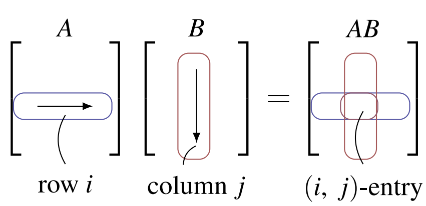
Theorem 2.3.2: Dot Product Rule

Let A and B be matrices of sizes $m \times n$ and $n \times k$, respectively. Then the (i, j)-entry of AB is the dot product of row i of A with column j of B.

Proof. Write $B = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_n \end{bmatrix}$ in terms of its columns. Then $A\mathbf{b}_j$ is column j of AB for each j. Hence the (i, j)-entry of AB is entry i of $A\mathbf{b}_j$, which is the dot product of row i of A with $\mathbf{b}_j$. This proves the theorem. □

Thus to compute the (i, j)-entry of AB, proceed as follows (see the diagram):

Go *across* row i of A, and *down* column j of B, multiply corresponding entries, and add the results.



Note that this requires that the rows of A must be the same length as the columns of B. The following rule is useful for remembering this and for deciding the size of the product matrix AB.

Compatibility Rule

$$\begin{array}{cc} A & B \\ m \times n & n' \times k \end{array}$$

Let A and B denote matrices. If A is $m \times n$ and B is $n' \times k$, the product AB can be formed if and only if $n = n'$. In this case the size of the product matrix AB is $m \times k$, and we say that AB is **defined**, or that A and B are **compatible** for multiplication.

The diagram provides a useful mnemonic for remembering this. We adopt the following convention:

Convention

Whenever a product of matrices is written, it is tacitly assumed that the sizes of the factors are such that the product is defined.

To illustrate the dot product rule, we recompute the matrix product in Example 2.3.1.

Example 2.3.3

Compute AB if $A = \begin{bmatrix} 2 & 3 & 5 \\ 1 & 4 & 7 \\ 0 & 1 & 8 \end{bmatrix}$ and $B = \begin{bmatrix} 8 & 9 \\ 7 & 2 \\ 6 & 1 \end{bmatrix}$.

Solution. Here A is 3×3 and B is 3×2, so the product matrix AB is defined and will be of size 3×2. Theorem 2.3.2 gives each entry of AB as the dot product of the corresponding row of A with the corresponding column of B_j that is,

$$AB = \begin{bmatrix} 2 & 3 & 5 \\ 1 & 4 & 7 \\ 0 & 1 & 8 \end{bmatrix} \begin{bmatrix} 8 & 9 \\ 7 & 2 \\ 6 & 1 \end{bmatrix} = \begin{bmatrix} 2\cdot 8+3\cdot 7+5\cdot 6 & 2\cdot 9+3\cdot 2+5\cdot 1 \\ 1\cdot 8+4\cdot 7+7\cdot 6 & 1\cdot 9+4\cdot 2+7\cdot 1 \\ 0\cdot 8+1\cdot 7+8\cdot 6 & 0\cdot 9+1\cdot 2+8\cdot 1 \end{bmatrix} = \begin{bmatrix} 67 & 29 \\ 78 & 24 \\ 55 & 10 \end{bmatrix}$$

Of course, this agrees with Example 2.3.1.

Example 2.3.4

Compute the $(1,\ 3)$- and $(2,\ 4)$-entries of AB where

$$A = \begin{bmatrix} 3 & -1 & 2 \\ 0 & 1 & 4 \end{bmatrix} \text{ and } B = \begin{bmatrix} 2 & 1 & 6 & 0 \\ 0 & 2 & 3 & 4 \\ -1 & 0 & 5 & 8 \end{bmatrix}.$$

Then compute AB.

Solution. The $(1,\ 3)$-entry of AB is the dot product of row 1 of A and column 3 of B (highlighted in the following display), computed by multiplying corresponding entries and adding the results.

$$\begin{bmatrix} 3 & -1 & 2 \\ 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} 2 & 1 & 6 & 0 \\ 0 & 2 & 3 & 4 \\ -1 & 0 & 5 & 8 \end{bmatrix} \quad (1,\ 3)\text{-entry} = 3\cdot 6+(-1)\cdot 3+2\cdot 5 = 25$$

Similarly, the $(2,\ 4)$-entry of AB involves row 2 of A and column 4 of B.

$$\begin{bmatrix} 3 & -1 & 2 \\ 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} 2 & 1 & 6 & 0 \\ 0 & 2 & 3 & 4 \\ -1 & 0 & 5 & 8 \end{bmatrix} \quad (2,\ 4)\text{-entry} = 0\cdot 0+1\cdot 4+4\cdot 8 = 36$$

Since A is 2×3 and B is 3×4, the product is 2×4.

$$AB = \begin{bmatrix} 3 & -1 & 2 \\ 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} 2 & 1 & 6 & 0 \\ 0 & 2 & 3 & 4 \\ -1 & 0 & 5 & 8 \end{bmatrix} = \begin{bmatrix} 4 & 1 & 25 & 12 \\ -4 & 2 & 23 & 36 \end{bmatrix}$$

Example 2.3.5

If $A = \begin{bmatrix} 1 & 3 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 \\ 6 \\ 4 \end{bmatrix}$, compute A^2, AB, BA, and B^2 when they are defined.[7]

Solution. Here, A is a 1×3 matrix and B is a 3×1 matrix, so A^2 and B^2 are not defined. However, the compatibility rule reads

$$\begin{array}{cc} A & B \\ 1\times 3 & 3\times 1 \end{array} \quad \text{and} \quad \begin{array}{cc} B & A \\ 3\times 1 & 1\times 3 \end{array}$$

so both AB and BA can be formed and these are 1×1 and 3×3 matrices, respectively.

$$AB = \begin{bmatrix} 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 1\cdot 5 + 3\cdot 6 + 2\cdot 4 \end{bmatrix} = \begin{bmatrix} 31 \end{bmatrix}$$

$$BA = \begin{bmatrix} 5 \\ 6 \\ 4 \end{bmatrix} \begin{bmatrix} 1 & 3 & 2 \end{bmatrix} = \begin{bmatrix} 5\cdot 1 & 5\cdot 3 & 5\cdot 2 \\ 6\cdot 1 & 6\cdot 3 & 6\cdot 2 \\ 4\cdot 1 & 4\cdot 3 & 4\cdot 2 \end{bmatrix} = \begin{bmatrix} 5 & 15 & 10 \\ 6 & 18 & 12 \\ 4 & 12 & 8 \end{bmatrix}$$

Unlike numerical multiplication, matrix products AB and BA *need not be equal*. In fact they need not even be the same size, as Example 2.3.5 shows. It turns out to be rare that $AB = BA$ (although it is by no means impossible), and A and B are said to **commute** when this happens.

Example 2.3.6

Let $A = \begin{bmatrix} 6 & 9 \\ -4 & -6 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix}$. Compute A^2, AB, BA.

Solution. $A^2 = \begin{bmatrix} 6 & 9 \\ -4 & -6 \end{bmatrix} \begin{bmatrix} 6 & 9 \\ -4 & -6 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, so $A^2 = 0$ can occur even if $A \neq 0$. Next,

$$AB = \begin{bmatrix} 6 & 9 \\ -4 & -6 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} -3 & 12 \\ 2 & -8 \end{bmatrix}$$
$$BA = \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 6 & 9 \\ -4 & -6 \end{bmatrix} = \begin{bmatrix} -2 & -3 \\ -6 & -9 \end{bmatrix}$$

Hence $AB \neq BA$, even though AB and BA are the same size.

[7] As for numbers, we write $A^2 = A \cdot A$, $A^3 = A \cdot A \cdot A$, etc. Note that A^2 is defined if and only if A is of size $n \times n$ for some n.

Example 2.3.7

If A is any matrix, then $IA = A$ and $AI = A$, and where I denotes an identity matrix of a size so that the multiplications are defined.

Solution. These both follow from the dot product rule as the reader should verify. For a more formal proof, write $A = \left[\begin{array}{cccc} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{array} \right]$ where $\mathbf{a}_j$ is column j of A. Then Definition 2.9 and Example 2.2.11 give

$$IA = \left[\begin{array}{cccc} I\mathbf{a}_1 & I\mathbf{a}_2 & \cdots & I\mathbf{a}_n \end{array} \right] = \left[\begin{array}{cccc} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{array} \right] = A$$

If $\mathbf{e}_j$ denotes column j of I, then $A\mathbf{e}_j = \mathbf{a}_j$ for each j by Example 2.2.12. Hence Definition 2.9 gives:

$$AI = A \left[\begin{array}{cccc} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_n \end{array} \right] = \left[\begin{array}{cccc} A\mathbf{e}_1 & A\mathbf{e}_2 & \cdots & A\mathbf{e}_n \end{array} \right] = \left[\begin{array}{cccc} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{array} \right] = A$$

The following theorem collects several results about matrix multiplication that are used everywhere in linear algebra.

Theorem 2.3.3

Assume that a is any scalar, and that A, B, and C are matrices of sizes such that the indicated matrix products are defined. Then:

1. *$IA = A$ and $AI = A$ where I denotes an identity matrix.*
2. *$A(BC) = (AB)C$.*
3. *$A(B+C) = AB + AC$.*
4. *$(B+C)A = BA + CA$.*
5. *$a(AB) = (aA)B = A(aB)$.*
6. *$(AB)^T = B^T A^T$.*

Proof. Condition (1) is Example 2.3.7; we prove (2), (4), and (6) and leave (3) and (5) as exercises.

2. If $C = \left[\begin{array}{cccc} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_k \end{array} \right]$ in terms of its columns, then $BC = \left[\begin{array}{cccc} B\mathbf{c}_1 & B\mathbf{c}_2 & \cdots & B\mathbf{c}_k \end{array} \right]$ by Definition 2.9, so

$$\begin{array}{rcll} A(BC) & = & \left[\begin{array}{cccc} A(B\mathbf{c}_1) & A(B\mathbf{c}_2) & \cdots & A(B\mathbf{c}_k) \end{array} \right] & \text{Definition 2.9} \\ & = & \left[\begin{array}{cccc} (AB)\mathbf{c}_1 & (AB)\mathbf{c}_2 & \cdots & (AB)\mathbf{c}_k) \end{array} \right] & \text{Theorem 2.3.1} \\ & = & (AB)C & \text{Definition 2.9} \end{array}$$

4. We know (Theorem 2.2.2) that $(B+C)\mathbf{x} = B\mathbf{x} + C\mathbf{x}$ holds for every column $\mathbf{x}$. If we write

$A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix}$ in terms of its columns, we get

$$\begin{aligned}
(B+C)A &= \begin{bmatrix} (B+C)\mathbf{a}_1 & (B+C)\mathbf{a}_2 & \cdots & (B+C)\mathbf{a}_n \end{bmatrix} && \text{Definition 2.9} \\
&= \begin{bmatrix} B\mathbf{a}_1 + C\mathbf{a}_1 & B\mathbf{a}_2 + C\mathbf{a}_2 & \cdots & B\mathbf{a}_n + C\mathbf{a}_n \end{bmatrix} && \text{Theorem 2.2.2} \\
&= \begin{bmatrix} B\mathbf{a}_1 & B\mathbf{a}_2 & \cdots & B\mathbf{a}_n \end{bmatrix} + \begin{bmatrix} C\mathbf{a}_1 & C\mathbf{a}_2 & \cdots & C\mathbf{a}_n \end{bmatrix} && \text{Adding Columns} \\
&= BA + CA && \text{Definition 2.9}
\end{aligned}$$

6. As in Section 2.1, write $A = [a_{ij}]$ and $B = [b_{ij}]$, so that $A^T = [a'_{ij}]$ and $B^T = [b'_{ij}]$ where $a'_{ij} = a_{ji}$ and $b'_{ji} = b_{ij}$ for all i and j. If c_{ij} denotes the $(i,\ j)$-entry of $B^T A^T$, then c_{ij} is the dot product of row i of B^T with column j of A^T. Hence

$$\begin{aligned}
c_{ij} = b'_{i1}a'_{1j} + b'_{i2}a'_{2j} + \cdots + b'_{im}a'_{mj} &= b_{1i}a_{j1} + b_{2i}a_{j2} + \cdots + b_{mi}a_{jm} \\
&= a_{j1}b_{1i} + a_{j2}b_{2i} + \cdots + a_{jm}b_{mi}
\end{aligned}$$

But this is the dot product of row j of A with column i of B; that is, the $(j,\ i)$-entry of AB; that is, the $(i,\ j)$-entry of $(AB)^T$. This proves (6). □

Property 2 in Theorem 2.3.3 is called the **associative law** of matrix multiplication. It asserts that the equation $A(BC) = (AB)C$ holds for all matrices (if the products are defined). Hence this product is the same no matter how it is formed, and so is written simply as ABC. This extends: The product $ABCD$ of four matrices can be formed several ways—for example, $(AB)(CD)$, $[A(BC)]D$, and $A[B(CD)]$—but the associative law implies that they are all equal and so are written as $ABCD$. A similar remark applies in general: Matrix products can be written unambiguously with no parentheses.

However, a note of caution about matrix multiplication must be taken: The fact that AB and BA need *not* be equal means that the *order* of the factors is important in a product of matrices. For example $ABCD$ and $ADCB$ may *not* be equal.

Warning

If the order of the factors in a product of matrices is changed, the product matrix may change (or may not be defined). Ignoring this warning is a source of many errors by students of linear algebra!

Properties 3 and 4 in Theorem 2.3.3 are called **distributive laws**. They assert that $A(B+C) = AB+AC$ and $(B+C)A = BA+CA$ hold whenever the sums and products are defined. These rules extend to more than two terms and, together with Property 5, ensure that many manipulations familiar from ordinary algebra extend to matrices. For example

$$\begin{gathered}
A(2B - 3C + D - 5E) = 2AB - 3AC + AD - 5AE \\
(A + 3C - 2D)B = AB + 3CB - 2DB
\end{gathered}$$

Note again that the warning is in effect: For example $A(B-C)$ need *not* equal $AB - CA$. These rules make possible a lot of simplification of matrix expressions.

Example 2.3.8

Simplify the expression $A(BC-CD)+A(C-B)D-AB(C-D)$.

Solution.

$$\begin{aligned} A(BC-CD)+A(C-B)D-AB(C-D) &= A(BC)-A(CD)+(AC-AB)D-(AB)C+(AB)D \\ &= ABC-ACD+ACD-ABD-ABC+ABD \\ &= 0 \end{aligned}$$

Example 2.3.9 and Example 2.3.10 below show how we can use the properties in Theorem 2.3.2 to deduce other facts about matrix multiplication. Matrices A and B are said to **commute** if $AB = BA$.

Example 2.3.9

Suppose that A, B, and C are $n \times n$ matrices and that both A and B commute with C; that is, $AC = CA$ and $BC = CB$. Show that AB commutes with C.

Solution. Showing that AB commutes with C means verifying that $(AB)C = C(AB)$. The computation uses the associative law several times, as well as the given facts that $AC = CA$ and $BC = CB$.

$$(AB)C = A(BC) = A(CB) = (AC)B = (CA)B = C(AB)$$

Example 2.3.10

Show that $AB = BA$ if and only if $(A-B)(A+B) = A^2 - B^2$.

Solution. The following *always* holds:

$$(A-B)(A+B) = A(A+B) - B(A+B) = A^2 + AB - BA - B^2 \tag{2.6}$$

Hence if $AB = BA$, then $(A-B)(A+B) = A^2 - B^2$ follows. Conversely, if this last equation holds, then equation (2.6) becomes

$$A^2 - B^2 = A^2 + AB - BA - B^2$$

This gives $0 = AB - BA$, and $AB = BA$ follows.

In Section 2.2 we saw (in Theorem 2.2.1) that every system of linear equations has the form

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

where A is the coefficient matrix, $\mathbf{x}$ is the column of variables, and $\mathbf{b}$ is the constant matrix. Thus the *system* of linear equations becomes a single matrix equation. Matrix multiplication can yield information about such a system.

Example 2.3.11

Consider a system $A\mathbf{x} = \mathbf{b}$ of linear equations where A is an $m \times n$ matrix. Assume that a matrix C exists such that $CA = I_n$. If the system $A\mathbf{x} = \mathbf{b}$ *has* a solution, show that this solution must be $C\mathbf{b}$. Give a condition guaranteeing that $C\mathbf{b}$ *is in fact* a solution.

Solution. Suppose that $\mathbf{x}$ is any solution to the system, so that $A\mathbf{x} = \mathbf{b}$. Multiply both sides of this matrix equation by C to obtain, successively,

$$C(A\mathbf{x}) = C\mathbf{b}, \quad (CA)\mathbf{x} = C\mathbf{b}, \quad I_n\mathbf{x} = C\mathbf{b}, \quad \mathbf{x} = C\mathbf{b}$$

This shows that *if* the system has a solution $\mathbf{x}$, then that solution must be $\mathbf{x} = C\mathbf{b}$, as required. But it does *not* guarantee that the system *has* a solution. However, if we write $\mathbf{x}_1 = C\mathbf{b}$, then

$$A\mathbf{x}_1 = A(C\mathbf{b}) = (AC)\mathbf{b}$$

Thus $\mathbf{x}_1 = C\mathbf{b}$ will be a solution if the condition $AC = I_m$ is satisfied.

The ideas in Example 2.3.11 lead to important information about matrices; this will be pursued in the next section.

Block Multiplication

Definition 2.10 Block Partition of a Matrix

*It is often useful to consider matrices whose entries are themselves matrices (called **blocks**). A matrix viewed in this way is said to be **partitioned into blocks**.*

For example, writing a matrix B in the form

$$B = \left[\begin{array}{cccc} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_k \end{array} \right] \text{ where the } \mathbf{b}_j \text{ are the columns of } B$$

is such a block partition of B. Here is another example.

Consider the matrices

$$A = \left[\begin{array}{rr|rrr} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline 2 & -1 & 4 & 2 & 1 \\ 3 & 1 & -1 & 7 & 5 \end{array} \right] = \left[\begin{array}{cc} I_2 & 0_{23} \\ P & Q \end{array} \right] \quad \text{and} \quad B = \left[\begin{array}{rr} 4 & -2 \\ 5 & 6 \\ \hline 7 & 3 \\ -1 & 0 \\ 1 & 6 \end{array} \right] = \left[\begin{array}{c} X \\ Y \end{array} \right]$$

where the blocks have been labelled as indicated. This is a natural way to partition A into blocks in view of the blocks I_2 and 0_{23} that occur. This notation is particularly useful when we are multiplying the matrices A and B because the product AB can be computed in block form as follows:

$$AB = \left[\begin{array}{cc} I & 0 \\ P & Q \end{array} \right] \left[\begin{array}{c} X \\ Y \end{array} \right] = \left[\begin{array}{c} IX + 0Y \\ PX + QY \end{array} \right] = \left[\begin{array}{c} X \\ PX + QY \end{array} \right] = \left[\begin{array}{rr} 4 & -2 \\ 5 & 6 \\ \hline 30 & 8 \\ 8 & 27 \end{array} \right]$$

This is easily checked to be the product AB, computed in the conventional manner.

In other words, *we can compute the product AB by ordinary matrix multiplication, using blocks as entries*. The only requirement is that the blocks be **compatible**. That is, *the sizes of the blocks must be such that all (matrix) products of blocks that occur make sense*. This means that the number of columns in each block of A must equal the number of rows in the corresponding block of B.

Theorem 2.3.4: Block Multiplication

If matrices A and B are partitioned compatibly into blocks, the product AB can be computed by matrix multiplication using blocks as entries.

We omit the proof.

We have been using two cases of block multiplication. If $B = \left[\begin{array}{cccc} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_k \end{array}\right]$ is a matrix where the $\mathbf{b}_j$ are the columns of B, and if the matrix product AB is defined, then we have

$$AB = A \left[\begin{array}{cccc} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_k \end{array}\right] = \left[\begin{array}{cccc} A\mathbf{b}_1 & A\mathbf{b}_2 & \cdots & A\mathbf{b}_k \end{array}\right]$$

This is Definition 2.9 and is a block multiplication where $A = [A]$ has only one block. As another illustration,

$$B\mathbf{x} = \left[\begin{array}{cccc} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_k \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_k \end{array}\right] = x_1\mathbf{b}_1 + x_2\mathbf{b}_2 + \cdots + x_k\mathbf{b}_k$$

where $\mathbf{x}$ is any $k \times 1$ column matrix (this is Definition 2.5).

It is not our intention to pursue block multiplication in detail here. However, we give one more example because it will be used below.

Theorem 2.3.5

Suppose matrices $A = \left[\begin{array}{cc} B & X \\ 0 & C \end{array}\right]$ and $A_1 = \left[\begin{array}{cc} B_1 & X_1 \\ 0 & C_1 \end{array}\right]$ are partitioned as shown where B and B_1 are square matrices of the same size, and C and C_1 are also square of the same size. These are compatible partitionings and block multiplication gives

$$AA_1 = \left[\begin{array}{cc} B & X \\ 0 & C \end{array}\right] \left[\begin{array}{cc} B_1 & X_1 \\ 0 & C_1 \end{array}\right] = \left[\begin{array}{cc} BB_1 & BX_1 + XC_1 \\ 0 & CC_1 \end{array}\right]$$

Example 2.3.12

Obtain a formula for A^k where $A = \left[\begin{array}{cc} I & X \\ 0 & 0 \end{array}\right]$ is square and I is an identity matrix.

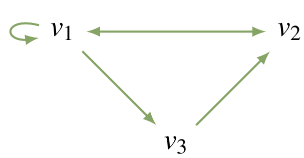
Solution. We have $A^2 = \left[\begin{array}{cc} I & X \\ 0 & 0 \end{array}\right] \left[\begin{array}{cc} I & X \\ 0 & 0 \end{array}\right] = \left[\begin{array}{cc} I^2 & IX + X0 \\ 0 & 0^2 \end{array}\right] = \left[\begin{array}{cc} I & X \\ 0 & 0 \end{array}\right] = A$. Hence $A^3 = AA^2 = AA = A^2 = A$. Continuing in this way, we see that $A^k = A$ for every $k \geq 1$.

Block multiplication has theoretical uses as we shall see. However, it is also useful in computing products of matrices in a computer with limited memory capacity. The matrices are partitioned into blocks in such a way that each product of blocks can be handled. Then the blocks are stored in auxiliary memory and their products are computed one by one.

Directed Graphs

The study of directed graphs illustrates how matrix multiplication arises in ways other than the study of linear equations or matrix transformations.

A **directed graph** consists of a set of points (called **vertices**) connected by arrows (called **edges**). For example, the vertices could represent cities and the edges available flights. If the graph has n vertices $v_1, v_2, \dots, v_n$, the **adjacency** matrix $A = \left[a_{ij}\right]$ is the $n \times n$ matrix whose (i, j)-entry a_{ij} is 1 if there is an edge from v_j to v_i (note the order), and zero otherwise. For example, the adjacency matrix of the directed graph shown is $A = \left[\begin{array}{ccc} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{array}\right]$.



A **path of length** r (or an r**-path**) from vertex j to vertex i is a sequence of r edges leading from v_j to v_i. Thus $v_1 \to v_2 \to v_1 \to v_1 \to v_3$ is a 4-path from v_1 to v_3 in the given graph. The edges are just the paths of length 1, so the (i, j)-entry a_{ij} of the adjacency matrix A is the number of 1-paths from v_j to v_i. This observation has an important extension:

Theorem 2.3.6

If A is the adjacency matrix of a directed graph with n vertices, then the (i, j)-entry of A^r is the number of r-paths $v_j \to v_i$.

As an illustration, consider the adjacency matrix A in the graph shown. Then

$$A = \left[\begin{array}{ccc} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{array}\right], \quad A^2 = \left[\begin{array}{ccc} 2 & 1 & 1 \\ 2 & 1 & 0 \\ 1 & 1 & 0 \end{array}\right], \quad \text{and} \quad A^3 = \left[\begin{array}{ccc} 4 & 2 & 1 \\ 3 & 2 & 1 \\ 2 & 1 & 1 \end{array}\right]$$

Hence, since the $(2, 1)$-entry of A^2 is 2, there are two 2-paths $v_1 \to v_2$ (in fact they are $v_1 \to v_1 \to v_2$ and $v_1 \to v_3 \to v_2$). Similarly, the $(2, 3)$-entry of A^2 is zero, so there are *no* 2-paths $v_3 \to v_2$, as the reader can verify. The fact that no entry of A^3 is zero shows that it is possible to go from any vertex to any other vertex in exactly three steps.

To see why Theorem 2.3.6 is true, observe that it asserts that

$$\text{the } (i, j)\text{-entry of } A^r \text{ equals the number of } r\text{-paths } v_j \to v_i \tag{2.7}$$

holds for each $r \geq 1$. We proceed by induction on r (see Appendix C). The case $r = 1$ is the definition of the adjacency matrix. So assume inductively that (2.7) is true for some $r \geq 1$; we must prove that (2.7) also holds for $r+1$. But every $(r+1)$-path $v_j \to v_i$ is the result of an r-path $v_j \to v_k$ for some k, followed by a 1-path $v_k \to v_i$. Writing $A = \left[a_{ij}\right]$ and $A^r = \left[b_{ij}\right]$, there are b_{kj} paths of the former type (by induction) and a_{ik} of the latter type, and so there are $a_{ik}b_{kj}$ such paths in all. Summing over k, this shows that there are

$$a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} \quad (r+1)\text{-paths } v_j \to v_i$$

But this sum is the dot product of the ith row $\begin{bmatrix} a_{i1} & a_{i2} & \cdots & a_{in} \end{bmatrix}$ of A with the jth column $\begin{bmatrix} b_{1j} & b_{2j} & \cdots & b_{nj} \end{bmatrix}^T$ of A^r. As such, it is the $(i,\ j)$-entry of the matrix product $A^rA = A^{r+1}$. This shows that (2.7) holds for $r+1$, as required.

2.4 Matrix Inverses

Three basic operations on matrices, addition, multiplication, and subtraction, are analogs for matrices of the same operations for numbers. In this section we introduce the matrix analog of numerical division.

To begin, consider how a numerical equation $ax = b$ is solved when a and b are known numbers. If $a = 0$, there is no solution (unless $b = 0$). But if $a \neq 0$, we can multiply both sides by the inverse $a^{-1} = \frac{1}{a}$ to obtain the solution $x = a^{-1}b$. Of course multiplying by a^{-1} is just dividing by a, and the property of a^{-1} that makes this work is that $a^{-1}a = 1$. Moreover, we saw in Section 2.2 that the role that 1 plays in arithmetic is played in matrix algebra by the identity matrix I. This suggests the following definition.

Definition 2.11 Matrix Inverses

If A is a square matrix, a matrix B is called an ***inverse*** *of A if and only if*

$$AB = I \quad \text{and} \quad BA = I$$

A matrix A that has an inverse is called an ***invertible matrix***.[8]

Example 2.4.1

Show that $B = \begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix}$ is an inverse of $A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$.

Solution. Compute AB and BA.

$$AB = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad BA = \begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Hence $AB = I = BA$, so B is indeed an inverse of A.

Example 2.4.2

Show that $A = \begin{bmatrix} 0 & 0 \\ 1 & 3 \end{bmatrix}$ has no inverse.

Solution. Let $B = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ denote an arbitrary 2×2 matrix. Then

$$AB = \begin{bmatrix} 0 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ a+3c & b+3d \end{bmatrix}$$

so AB has a row of zeros. Hence AB cannot equal I for any B.

The argument in Example 2.4.2 shows that no zero matrix has an inverse. But Example 2.4.2 also shows that, unlike arithmetic, *it is possible for a nonzero matrix to have no inverse*. However, if a matrix *does* have an inverse, it has only one.

Theorem 2.4.1

If B and C are both inverses of A, then $B = C$.

Proof. Since B and C are both inverses of A, we have $CA = I = AB$. Hence

$$B = IB = (CA)B = C(AB) = CI = C$$

□

If A is an invertible matrix, the (unique) inverse of A is denoted A^{-1}. Hence A^{-1} (when it exists) is a square matrix of the same size as A with the property that

$$AA^{-1} = I \quad \text{and} \quad A^{-1}A = I$$

These equations characterize A^{-1} in the following sense:

> **Inverse Criterion:** *If somehow a matrix B can be found such that $AB = I$ and $BA = I$, then A is invertible and B is the inverse of A; in symbols, $B = A^{-1}$.*

[8]Only square matrices have inverses. Even though it is plausible that nonsquare matrices A and B could exist such that $AB = I_m$ and $BA = I_n$, where A is $m \times n$ and B is $n \times m$, we claim that this forces $n = m$. Indeed, if $m < n$ there exists a nonzero column $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{0}$ (by Theorem 1.3.1), so $\mathbf{x} = I_n\mathbf{x} = (BA)\mathbf{x} = B(A\mathbf{x}) = B(\mathbf{0}) = \mathbf{0}$, a contradiction. Hence $m \geq n$. Similarly, the condition $AB = I_m$ implies that $n \geq m$. Hence $m = n$ so A is square.

This is a way to verify that the inverse of a matrix exists. Example 2.4.3 and Example 2.4.4 offer illustrations.

Example 2.4.3

If $A = \left[\begin{array}{rr} 0 & -1 \\ 1 & -1 \end{array} \right]$, show that $A^3 = I$ and so find A^{-1}.

Solution. We have $A^2 = \left[\begin{array}{rr} 0 & -1 \\ 1 & -1 \end{array} \right] \left[\begin{array}{rr} 0 & -1 \\ 1 & -1 \end{array} \right] = \left[\begin{array}{rr} -1 & 1 \\ -1 & 0 \end{array} \right]$, and so

$$A^3 = A^2A = \left[\begin{array}{rr} -1 & 1 \\ -1 & 0 \end{array} \right] \left[\begin{array}{rr} 0 & -1 \\ 1 & -1 \end{array} \right] = \left[\begin{array}{rr} 1 & 0 \\ 0 & 1 \end{array} \right] = I$$

Hence $A^3 = I$, as asserted. This can be written as $A^2A = I = AA^2$, so it shows that A^2 is the inverse of A. That is, $A^{-1} = A^2 = \left[\begin{array}{rr} -1 & 1 \\ -1 & 0 \end{array} \right]$.

The next example presents a useful formula for the inverse of a 2×2 matrix $A = \left[\begin{array}{cc} a & b \\ c & d \end{array} \right]$ when it exists. To state it, we define the **determinant** $\det A$ and the **adjugate** $\text{adj } A$ of the matrix A as follows:

$$\det \left[\begin{array}{cc} a & b \\ c & d \end{array} \right] = ad - bc, \quad \text{and} \quad \text{adj} \left[\begin{array}{cc} a & b \\ c & d \end{array} \right] = \left[\begin{array}{rr} d & -b \\ -c & a \end{array} \right]$$

Example 2.4.4

If $A = \left[\begin{array}{cc} a & b \\ c & d \end{array} \right]$, show that A has an inverse if and only if $\det A \neq 0$, and in this case

$$A^{-1} = \frac{1}{\det A} \text{adj } A$$

Solution. For convenience, write $e = \det A = ad - bc$ and $B = \text{adj } A = \left[\begin{array}{rr} d & -b \\ -c & a \end{array} \right]$. Then $AB = eI = BA$ as the reader can verify. So if $e \neq 0$, scalar multiplication by $\frac{1}{e}$ gives

$$A(\tfrac{1}{e}B) = I = (\tfrac{1}{e}B)A$$

Hence A is invertible and $A^{-1} = \frac{1}{e}B$. Thus it remains only to show that if A^{-1} exists, then $e \neq 0$. We prove this by showing that assuming $e = 0$ leads to a contradiction. In fact, if $e = 0$, then $AB = eI = 0$, so left multiplication by A^{-1} gives $A^{-1}AB = A^{-1}0$; that is, $IB = 0$, so $B = 0$. But this implies that a, b, c, and d are *all* zero, so $A = 0$, contrary to the assumption that A^{-1} exists.

As an illustration, if $A = \left[\begin{array}{rr} 2 & 4 \\ -3 & 8 \end{array} \right]$ then $\det A = 2 \cdot 8 - 4 \cdot (-3) = 28 \neq 0$. Hence A is invertible and $A^{-1} = \frac{1}{\det A} \text{adj } A = \frac{1}{28} \left[\begin{array}{rr} 8 & -4 \\ 3 & 2 \end{array} \right]$, as the reader is invited to verify.

The determinant and adjugate will be defined in Chapter 3 for any square matrix, and the conclusions in Example 2.4.4 will be proved in full generality.

Inverses and Linear Systems

Matrix inverses can be used to solve certain systems of linear equations. Recall that a *system* of linear equations can be written as a *single* matrix equation

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

where A and $\mathbf{b}$ are known and $\mathbf{x}$ is to be determined. If A is invertible, we multiply each side of the equation on the left by A^{-1} to get

$$\begin{aligned} A^{-1}A\mathbf{x} &= A^{-1}\mathbf{b} \\ I\mathbf{x} &= A^{-1}\mathbf{b} \\ \mathbf{x} &= A^{-1}\mathbf{b} \end{aligned}$$

This gives the solution to the system of equations (the reader should verify that $\mathbf{x} = A^{-1}\mathbf{b}$ really does satisfy $A\mathbf{x} = \mathbf{b}$). Furthermore, the argument shows that if $\mathbf{x}$ is *any* solution, then necessarily $\mathbf{x} = A^{-1}\mathbf{b}$, so the solution is unique. Of course the technique works only when the coefficient matrix A has an inverse. This proves Theorem 2.4.2.

Theorem 2.4.2

Suppose a system of n equations in n variables is written in matrix form as

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

If the $n \times n$ coefficient matrix A is invertible, the system has the unique solution

$$\mathbf{x} = A^{-1}\mathbf{b}$$

Example 2.4.5

Use Example 2.4.4 to solve the system $\left\{\begin{array}{rl} 5x_1 - 3x_2 = & -4 \\ 7x_1 + 4x_2 = & 8 \end{array}\right.$.

Solution. In matrix form this is $A\mathbf{x} = \mathbf{b}$ where $A = \left[\begin{array}{rr} 5 & -3 \\ 7 & 4 \end{array}\right]$, $\mathbf{x} = \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right]$, and $\mathbf{b} = \left[\begin{array}{r} -4 \\ 8 \end{array}\right]$. Then $\det A = 5 \cdot 4 - (-3) \cdot 7 = 41$, so A is invertible and $A^{-1} = \frac{1}{41}\left[\begin{array}{rr} 4 & 3 \\ -7 & 5 \end{array}\right]$ by Example 2.4.4. Thus Theorem 2.4.2 gives

$$\mathbf{x} = A^{-1}\mathbf{b} = \frac{1}{41}\left[\begin{array}{rr} 4 & 3 \\ -7 & 5 \end{array}\right]\left[\begin{array}{r} -4 \\ 8 \end{array}\right] = \frac{1}{41}\left[\begin{array}{r} 8 \\ 68 \end{array}\right]$$

so the solution is $x_1 = \frac{8}{41}$ and $x_2 = \frac{68}{41}$.

An Inversion Method

If a matrix A is $n \times n$ and invertible, it is desirable to have an efficient technique for finding the inverse. The following procedure will be justified in Section 2.5.

Theorem: Matrix Inversion Algorithm

If A is an invertible (square) matrix, there exists a sequence of elementary row operations that carry A to the identity matrix I of the same size, written $A \to I$. This same series of row operations carries I to A^{-1}; that is, $I \to A^{-1}$. The algorithm can be summarized as follows:

$$\left[\begin{array}{cc} A & I \end{array}\right] \to \left[\begin{array}{cc} I & A^{-1} \end{array}\right]$$

where the row operations on A and I are carried out simultaneously.

Example 2.4.6

Use the inversion algorithm to find the inverse of the matrix

$$A = \left[\begin{array}{rrr} 2 & 7 & 1 \\ 1 & 4 & -1 \\ 1 & 3 & 0 \end{array}\right]$$

Solution. Apply elementary row operations to the double matrix

$$\left[\begin{array}{cc} A & I \end{array}\right] = \left[\begin{array}{rrr|rrr} 2 & 7 & 1 & 1 & 0 & 0 \\ 1 & 4 & -1 & 0 & 1 & 0 \\ 1 & 3 & 0 & 0 & 0 & 1 \end{array}\right]$$

so as to carry A to I. First interchange rows 1 and 2.

$$\left[\begin{array}{rrr|rrr} 1 & 4 & -1 & 0 & 1 & 0 \\ 2 & 7 & 1 & 1 & 0 & 0 \\ 1 & 3 & 0 & 0 & 0 & 1 \end{array}\right]$$

Next subtract 2 times row 1 from row 2, and subtract row 1 from row 3.

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

Continue to reduced row-echelon form.

$$\left[\begin{array}{rrr|rrr} 1 & 0 & 11 & 4 & -7 & 0 \\ 0 & 1 & -3 & -1 & 2 & 0 \\ 0 & 0 & -2 & -1 & 1 & 1 \end{array}\right]$$

$$\left[\begin{array}{rrr|rrr} 1 & 0 & 0 & \frac{-3}{2} & \frac{-3}{2} & \frac{11}{2} \\ 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{-3}{2} \\ 0 & 0 & 1 & \frac{1}{2} & \frac{-1}{2} & \frac{-1}{2} \end{array}\right]$$

Hence $A^{-1} = \frac{1}{2}\begin{bmatrix} -3 & -3 & 11 \\ 1 & 1 & -3 \\ 1 & -1 & -1 \end{bmatrix}$, as is readily verified.

Given any $n \times n$ matrix A, Theorem 1.2.1 shows that A can be carried by elementary row operations to a matrix R in reduced row-echelon form. If $R = I$, the matrix A is invertible (this will be proved in the next section), so the algorithm produces A^{-1}. If $R \neq I$, then R has a row of zeros (it is square), so no system of linear equations $A\mathbf{x} = \mathbf{b}$ can have a unique solution. But then A is not invertible by Theorem 2.4.2. Hence, the algorithm is effective in the sense conveyed in Theorem 2.4.3.

Theorem 2.4.3

If A is an $n \times n$ matrix, either A can be reduced to I by elementary row operations or it cannot. In the first case, the algorithm produces A^{-1}; in the second case, A^{-1} does not exist.

Properties of Inverses

The following properties of an invertible matrix are used everywhere.

Example 2.4.7: Cancellation Laws

Let A be an invertible matrix. Show that:

1. If $AB = AC$, then $B = C$.
2. If $BA = CA$, then $B = C$.

Solution. Given the equation $AB = AC$, left multiply both sides by A^{-1} to obtain $A^{-1}AB = A^{-1}AC$. Thus $IB = IC$, that is $B = C$. This proves (1) and the proof of (2) is left to the reader.

Properties (1) and (2) in Example 2.4.7 are described by saying that an invertible matrix can be "left cancelled" and "right cancelled", respectively. Note however that "mixed" cancellation does not hold in general: If A is invertible and $AB = CA$, then B and C may *not* be equal, even if both are 2×2. Here is a specific example:

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \; B = \begin{bmatrix} 0 & 0 \\ 1 & 2 \end{bmatrix}, C = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Sometimes the inverse of a matrix is given by a formula. Example 2.4.4 is one illustration; Example 2.4.8 and Example 2.4.9 provide two more. The idea is the *Inverse Criterion*: If a matrix B can be found such that $AB = I = BA$, then A is invertible and $A^{-1} = B$.

Example 2.4.8

If A is an invertible matrix, show that the transpose A^T is also invertible. Show further that the inverse of A^T is just the transpose of A^{-1}; in symbols, $(A^T)^{-1} = (A^{-1})^T$.

Solution. A^{-1} exists (by assumption). Its transpose $(A^{-1})^T$ is the candidate proposed for the inverse of A^T. Using the inverse criterion, we test it as follows:

$$\begin{aligned} A^T(A^{-1})^T &= (A^{-1}A)^T = I^T = I \\ (A^{-1})^T A^T &= (AA^{-1})^T = I^T = I \end{aligned}$$

Hence $(A^{-1})^T$ is indeed the inverse of A^T; that is, $(A^T)^{-1} = (A^{-1})^T$.

Example 2.4.9

If A and B are invertible $n \times n$ matrices, show that their product AB is also invertible and $(AB)^{-1} = B^{-1}A^{-1}$.

Solution. We are given a candidate for the inverse of AB, namely $B^{-1}A^{-1}$. We test it as follows:

$$\begin{aligned} (B^{-1}A^{-1})(AB) &= B^{-1}(A^{-1}A)B = B^{-1}IB = B^{-1}B = I \\ (AB)(B^{-1}A^{-1}) &= A(BB^{-1})A^{-1} = AIA^{-1} = AA^{-1} = I \end{aligned}$$

Hence $B^{-1}A^{-1}$ is the inverse of AB; in symbols, $(AB)^{-1} = B^{-1}A^{-1}$.

We now collect several basic properties of matrix inverses for reference.

Theorem 2.4.4

All the following matrices are square matrices of the same size.

1. *I is invertible and $I^{-1} = I$.*
2. *If A is invertible, so is A^{-1}, and $(A^{-1})^{-1} = A$.*
3. *If A and B are invertible, so is AB, and $(AB)^{-1} = B^{-1}A^{-1}$.*
4. *If $A_1, A_2, \dots, A_k$ are all invertible, so is their product $A_1A_2\cdots A_k$, and*
$$(A_1A_2\cdots A_k)^{-1} = A_k^{-1}\cdots A_2^{-1}A_1^{-1}.$$
5. *If A is invertible, so is A^k for any $k \geq 1$, and $(A^k)^{-1} = (A^{-1})^k$.*
6. *If A is invertible and $a \neq 0$ is a number, then aA is invertible and $(aA)^{-1} = \frac{1}{a}A^{-1}$.*
7. *If A is invertible, so is its transpose A^T, and $(A^T)^{-1} = (A^{-1})^T$.*

Proof.

1. This is an immediate consequence of the fact that $I^2 = I$.
2. The equations $AA^{-1} = I = A^{-1}A$ show that A is the inverse of A^{-1}; in symbols, $(A^{-1})^{-1} = A$.

3. This is Example 2.4.9.

4. Use induction on k. If $k=1$, there is nothing to prove, and if $k=2$, the result is property 3. If $k>2$, assume inductively that $(A_1A_2\cdots A_{k-1})^{-1}=A_{k-1}^{-1}\cdots A_2^{-1}A_1^{-1}$. We apply this fact together with property 3 as follows:

$$\begin{aligned}[A_1A_2\cdots A_{k-1}A_k]^{-1} &= [(A_1A_2\cdots A_{k-1})A_k]^{-1}\\ &= A_k^{-1}(A_1A_2\cdots A_{k-1})^{-1}\\ &= A_k^{-1}\left(A_{k-1}^{-1}\cdots A_2^{-1}A_1^{-1}\right)\end{aligned}$$

   So the proof by induction is complete.

5. This is property 4 with $A_1=A_2=\cdots=A_k=A$.

6. This is left as Exercise **??**.

7. This is Example 2.4.8. □

The reversal of the order of the inverses in properties 3 and 4 of Theorem 2.4.4 is a consequence of the fact that matrix multiplication is not commutative. Another manifestation of this comes when matrix equations are dealt with. If a matrix equation $B=C$ is given, it can be *left-multiplied* by a matrix A to yield $AB=AC$. Similarly, *right-multiplication* gives $BA=CA$. However, we cannot mix the two: If $B=C$, it need *not* be the case that $AB=CA$ even if A is invertible, for example, $A=\left[\begin{array}{cc}1&1\\0&1\end{array}\right]$, $B=\left[\begin{array}{cc}0&0\\1&0\end{array}\right]=C$.

Part 7 of Theorem 2.4.4 together with the fact that $(A^T)^T=A$ gives

Corollary 2.4.1

A square matrix A is invertible if and only if A^T is invertible.

Example 2.4.10

Find A if $(A^T-2I)^{-1}=\left[\begin{array}{rr}2&1\\-1&0\end{array}\right]$.

Solution. By Theorem 2.4.4(2) and Example 2.4.4, we have

$$(A^T-2I)=\left[\left(A^T-2I\right)^{-1}\right]^{-1}=\left[\begin{array}{rr}2&1\\-1&0\end{array}\right]^{-1}=\left[\begin{array}{rr}0&-1\\1&2\end{array}\right]$$

Hence $A^T=2I+\left[\begin{array}{rr}0&-1\\1&2\end{array}\right]=\left[\begin{array}{rr}2&-1\\1&4\end{array}\right]$, so $A=\left[\begin{array}{rr}2&1\\-1&4\end{array}\right]$ by Theorem 2.4.4(7).

The following important theorem collects a number of conditions all equivalent[9] to invertibility. It will be referred to frequently below.

[9]If p and q are statements, we say that p **implies** q (written $p\Rightarrow q$) if q is true whenever p is true. The statements are called **equivalent** if both $p\Rightarrow q$ and $q\Rightarrow p$ (written $p\Leftrightarrow q$, spoken "p if and only if q"). See Appendix B.

Theorem 2.4.5: Inverse Theorem

The following conditions are equivalent for an $n \times n$ matrix A:

1. *A is invertible.*
2. *The homogeneous system $A\mathbf{x} = \mathbf{0}$ has only the trivial solution $\mathbf{x} = \mathbf{0}$.*
3. *A can be carried to the identity matrix I_n by elementary row operations.*
4. *The system $A\mathbf{x} = \mathbf{b}$ has at least one solution $\mathbf{x}$ for every choice of column $\mathbf{b}$.*
5. *There exists an $n \times n$ matrix C such that $AC = I_n$.*

Proof. We show that each of these conditions implies the next, and that (5) implies (1).

(1) $\Rightarrow$ (2). If A^{-1} exists, then $A\mathbf{x} = \mathbf{0}$ gives $\mathbf{x} = I_n\mathbf{x} = A^{-1}A\mathbf{x} = A^{-1}\mathbf{0} = \mathbf{0}$.

(2) $\Rightarrow$ (3). Assume that (2) is true. Certainly $A \to R$ by row operations where R is a reduced, row-echelon matrix. It suffices to show that $R = I_n$. Suppose that this is not the case. Then R has a row of zeros (being square). Now consider the augmented matrix $\left[\begin{array}{c|c} A & \mathbf{0} \end{array}\right]$ of the system $A\mathbf{x} = \mathbf{0}$. Then $\left[\begin{array}{c|c} A & \mathbf{0} \end{array}\right] \to \left[\begin{array}{c|c} R & \mathbf{0} \end{array}\right]$ is the reduced form, and $\left[\begin{array}{c|c} R & \mathbf{0} \end{array}\right]$ also has a row of zeros. Since R is square there must be at least one nonleading variable, and hence at least one parameter. Hence the system $A\mathbf{x} = \mathbf{0}$ has infinitely many solutions, contrary to (2). So $R = I_n$ after all.

(3) $\Rightarrow$ (4). Consider the augmented matrix $\left[\begin{array}{c|c} A & \mathbf{b} \end{array}\right]$ of the system $A\mathbf{x} = \mathbf{b}$. Using (3), let $A \to I_n$ by a sequence of row operations. Then these same operations carry $\left[\begin{array}{c|c} A & \mathbf{b} \end{array}\right] \to \left[\begin{array}{c|c} I_n & \mathbf{c} \end{array}\right]$ for some column $\mathbf{c}$. Hence the system $A\mathbf{x} = \mathbf{b}$ has a solution (in fact unique) by gaussian elimination. This proves (4).

(4) $\Rightarrow$ (5). Write $I_n = \left[\begin{array}{cccc} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_n \end{array}\right]$ where $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ are the columns of I_n. For each $j = 1, 2, \dots, n$, the system $A\mathbf{x} = \mathbf{e}_j$ has a solution $\mathbf{c}_j$ by (4), so $A\mathbf{c}_j = \mathbf{e}_j$. Now let $C = \left[\begin{array}{cccc} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_n \end{array}\right]$ be the $n \times n$ matrix with these matrices $\mathbf{c}_j$ as its columns. Then Definition 2.9 gives (5):

$$AC = A\left[\begin{array}{cccc} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_n \end{array}\right] = \left[\begin{array}{cccc} A\mathbf{c}_1 & A\mathbf{c}_2 & \cdots & A\mathbf{c}_n \end{array}\right] = \left[\begin{array}{cccc} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_n \end{array}\right] = I_n$$

(5) $\Rightarrow$ (1). Assume that (5) is true so that $AC = I_n$ for some matrix C. Then $C\mathbf{x} = 0$ implies $\mathbf{x} = \mathbf{0}$ (because $\mathbf{x} = I_n\mathbf{x} = AC\mathbf{x} = A\mathbf{0} = \mathbf{0}$). Thus condition (2) holds for the matrix C rather than A. Hence the argument above that (2) $\Rightarrow$ (3) $\Rightarrow$ (4) $\Rightarrow$ (5) (with A replaced by C) shows that a matrix C' exists such that $CC' = I_n$. But then

$$A = AI_n = A(CC') = (AC)C' = I_nC' = C'$$

Thus $CA = CC' = I_n$ which, together with $AC = I_n$, shows that C is the inverse of A. This proves (1). □

The proof of (5) $\Rightarrow$ (1) in Theorem 2.4.5 shows that if $AC = I$ for square matrices, then necessarily $CA = I$, and hence that C and A are inverses of each other. We record this important fact for reference.

Corollary 2.4.2

If A and C are square matrices such that $AC = I$, then also $CA = I$. In particular, both A and C are invertible, $C = A^{-1}$, and $A = C^{-1}$.

Here is a quick way to remember Corollary 2.4.2. If A is a square matrix, then

1. If $AC = I$ then $C = A^{-1}$.
2. If $CA = I$ then $C = A^{-1}$.

Observe that Corollary 2.4.2 is false if A and C are not square matrices. For example, we have

$$\begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} = I_2 \quad \text{but} \quad \begin{bmatrix} -1 & 1 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \neq I_3$$

In fact, it is verified in the footnote on page 68 that if $AB = I_m$ and $BA = I_n$, where A is $m \times n$ and B is $n \times m$, then $m = n$ and A and B are (square) inverses of each other.

An $n \times n$ matrix A has rank n if and only if (3) of Theorem 2.4.5 holds. Hence

Corollary 2.4.3

An $n \times n$ matrix A is invertible if and only if rank $A = n$.

Here is a useful fact about inverses of block matrices.

Example 2.4.11

Let $P = \begin{bmatrix} A & X \\ 0 & B \end{bmatrix}$ and $Q = \begin{bmatrix} A & 0 \\ Y & B \end{bmatrix}$ be block matrices where A is $m \times m$ and B is $n \times n$ (possibly $m \neq n$).

a. Show that P is invertible if and only if A and B are both invertible. In this case, show that

$$P^{-1} = \begin{bmatrix} A^{-1} & -A^{-1}XB^{-1} \\ 0 & B^{-1} \end{bmatrix}$$

b. Show that Q is invertible if and only if A and B are both invertible. In this case, show that

$$Q^{-1} = \begin{bmatrix} A^{-1} & 0 \\ -B^{-1}YA^{-1} & B^{-1} \end{bmatrix}$$

Solution. We do (a.) and leave (b.) for the reader.

a. If A^{-1} and B^{-1} both exist, write $R = \begin{bmatrix} A^{-1} & -A^{-1}XB^{-1} \\ 0 & B^{-1} \end{bmatrix}$. Using block multiplication, one verifies that $PR = I_{m+n} = RP$, so P is invertible, and $P^{-1} = R$. Conversely, suppose that P is invertible, and write $P^{-1} = \begin{bmatrix} C & V \\ W & D \end{bmatrix}$ in block form, where C is $m \times m$ and D is $n \times n$.

Then the equation $PP^{-1} = I_{n+m}$ becomes

$$\begin{bmatrix} A & X \\ 0 & B \end{bmatrix} \begin{bmatrix} C & V \\ W & D \end{bmatrix} = \begin{bmatrix} AC + XW & AV + XD \\ BW & BD \end{bmatrix} = I_{m+n} = \begin{bmatrix} I_m & 0 \\ 0 & I_n \end{bmatrix}$$

using block notation. Equating corresponding blocks, we find

$$AC + XW = I_m, \quad BW = 0, \quad \text{and } BD = I_n$$

Hence B is invertible because $BD = I_n$ (by Corollary 2.4.1), then $W = 0$ because $BW = 0$, and finally, $AC = I_m$ (so A is invertible, again by Corollary 2.4.1).

Inverses of Matrix Transformations

Let $T = T_A : \mathbb{R}^n \to \mathbb{R}^n$ denote the matrix transformation induced by the $n \times n$ matrix A. Since A is square, it may very well be invertible, and this leads to the question:

What does it mean geometrically for T that A is invertible?

To answer this, let $T' = T_{A^{-1}} : \mathbb{R}^n \to \mathbb{R}^n$ denote the transformation induced by A^{-1}. Then

$$\begin{aligned} T'\left[T(\mathbf{x})\right] &= A^{-1}\left[A\mathbf{x}\right] = I\mathbf{x} = \mathbf{x} \\ T\left[T'(\mathbf{x})\right] &= A\left[A^{-1}\mathbf{x}\right] = I\mathbf{x} = \mathbf{x} \end{aligned} \quad \text{for all } \mathbf{x} \text{ in } \mathbb{R}^n \tag{2.8}$$

The first of these equations asserts that, if T carries $\mathbf{x}$ to a vector $T(\mathbf{x})$, then T' carries $T(\mathbf{x})$ right back to $\mathbf{x}$; that is T' "reverses" the action of T. Similarly T "reverses" the action of T'. Conditions (2.8) can be stated compactly in terms of composition:

$$T' \circ T = 1_{\mathbb{R}^n} \quad \text{and} \quad T \circ T' = 1_{\mathbb{R}^n} \tag{2.9}$$

When these conditions hold, we say that the matrix transformation T' is an **inverse** of T, and we have shown that if the matrix A of T is invertible, then T has an inverse (induced by A^{-1}).

The converse is also true: If T has an inverse, then its matrix A must be invertible. Indeed, suppose $S : \mathbb{R}^n \to \mathbb{R}^n$ is any inverse of T, so that $S \circ T = 1_{\mathbb{R}_n}$ and $T \circ S = 1_{\mathbb{R}_n}$. It can be shown that S is also a matrix transformation. If B is the matrix of S, we have

$$BA\mathbf{x} = S\left[T(\mathbf{x})\right] = (S \circ T)(\mathbf{x}) = 1_{\mathbb{R}^n}(\mathbf{x}) = \mathbf{x} = I_n\mathbf{x} \quad \text{for all } \mathbf{x} \text{ in } \mathbb{R}^n$$

It follows by Theorem 2.2.6 that $BA = I_n$, and a similar argument shows that $AB = I_n$. Hence A is invertible with $A^{-1} = B$. Furthermore, the inverse transformation S has matrix A^{-1}, so $S = T'$ using the earlier notation. This proves the following important theorem.

Theorem 2.4.6

Let $T : \mathbb{R}^n \to \mathbb{R}^n$ denote the matrix transformation induced by an $n \times n$ matrix A. Then

A is invertible if and only if T has an inverse.

In this case, T has exactly one inverse (which we denote as T^{-1}), and $T^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ is the transformation induced by the matrix A^{-1}. In other words

$$(T_A)^{-1} = T_{A^{-1}}$$

The geometrical relationship between T and T^{-1} is embodied in equations (2.8) above:

$$T^{-1}\left[T(\mathbf{x})\right] = \mathbf{x} \quad \text{and} \quad T\left[T^{-1}(\mathbf{x})\right] = \mathbf{x} \quad \text{for all } \mathbf{x} \text{ in } \mathbb{R}^n$$

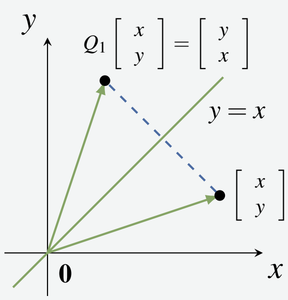
These equations are called the **fundamental identities** relating T and T^{-1}. Loosely speaking, they assert that each of T and T^{-1} "reverses" or "undoes" the action of the other.

This geometric view of the inverse of a linear transformation provides a new way to find the inverse of a matrix A. More precisely, if A is an invertible matrix, we proceed as follows:

1. *Let T be the linear transformation induced by A.*
2. *Obtain the linear transformation T^{-1} which "reverses" the action of T.*
3. *Then A^{-1} is the matrix of T^{-1}.*

Here is an example.

Example 2.4.12



Find the inverse of $A = \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array}\right]$ by viewing it as a linear transformation $\mathbb{R}^2 \to \mathbb{R}^2$.

Solution. If $\mathbf{x} = \left[\begin{array}{c} x \\ y \end{array}\right]$ the vector $A\mathbf{x} = \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array}\right]\left[\begin{array}{c} x \\ y \end{array}\right] = \left[\begin{array}{c} y \\ x \end{array}\right]$ is the result of reflecting $\mathbf{x}$ in the line $y = x$ (see the diagram). Hence, if $Q_1 : \mathbb{R}^2 \to \mathbb{R}^2$ denotes reflection in the line $y = x$, then A is the matrix of Q_1. Now observe that Q_1 reverses itself because reflecting a vector $\mathbf{x}$ twice results in $\mathbf{x}$. Consequently $Q_1^{-1} = Q_1$. Since A^{-1} is the matrix of Q_1^{-1} and A is the matrix of Q, it follows that $A^{-1} = A$. Of course this conclusion is clear by simply observing directly that $A^2 = I$, but the geometric method can often work where these other methods may be less straightforward.

2.5 Elementary Matrices

It is now clear that elementary row operations are important in linear algebra: They are essential in solving linear systems (using the gaussian algorithm) and in inverting a matrix (using the matrix inversion algorithm). It turns out that they can be performed by left multiplying by certain invertible matrices. These matrices are the subject of this section.

Definition 2.1 Elementary Matrices

An $n \times n$ matrix E is called an ***elementary matrix*** *if it can be obtained from the identity matrix I_n by a single elementary row operation (called the operation* ***corresponding*** *to E). We say that E is of type I, II, or III if the operation is of that type (see Definition 1.2).*

Hence

$$E_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad E_2 = \begin{bmatrix} 1 & 0 \\ 0 & 9 \end{bmatrix}, \quad \text{and} \quad E_3 = \begin{bmatrix} 1 & 5 \\ 0 & 1 \end{bmatrix}$$

are elementary of types I, II, and III, respectively, obtained from the 2×2 identity matrix by interchanging rows 1 and 2, multiplying row 2 by 9, and adding 5 times row 2 to row 1.

Suppose now that the matrix $A = \begin{bmatrix} a & b & c \\ p & q & r \end{bmatrix}$ is left multiplied by the above elementary matrices E_1,

E_2, and E_3. The results are:

$$E_1A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a & b & c \\ p & q & r \end{bmatrix} = \begin{bmatrix} p & q & r \\ a & b & c \end{bmatrix}$$
$$E_2A = \begin{bmatrix} 1 & 0 \\ 0 & 9 \end{bmatrix} \begin{bmatrix} a & b & c \\ p & q & r \end{bmatrix} = \begin{bmatrix} a & b & c \\ 9p & 9q & 9r \end{bmatrix}$$
$$E_3A = \begin{bmatrix} 1 & 5 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a & b & c \\ p & q & r \end{bmatrix} = \begin{bmatrix} a+5p & b+5q & c+5r \\ p & q & r \end{bmatrix}$$

In each case, left multiplying A by the elementary matrix has the *same* effect as doing the corresponding row operation to A. This works in general.

Lemma 2.5.1: [10]

If an elementary row operation is performed on an $m \times n$ matrix A, the result is EA where E is the elementary matrix obtained by performing the same operation on the $m \times m$ identity matrix.

Proof. We prove it for operations of type III; the proofs for types I and II are left as exercises. Let E be the elementary matrix corresponding to the operation that adds k times row p to row $q \neq p$. The proof depends on the fact that each row of EA is equal to the corresponding row of E times A. Let $K_1, K_2, \dots, K_m$ denote the rows of I_m. Then row i of E is K_i if $i \neq q$, while row q of E is $K_q + kK_p$. Hence:

$$\begin{aligned} \text{If } i \neq q \text{ then row } i \text{ of } EA = {} & K_iA = (\text{row } i \text{ of } A). \\ \text{Row } q \text{ of } EA = (K_q + kK_p)A = {} & K_qA + k(K_pA) \\ = {} & (\text{row } q \text{ of } A) \text{ plus } k \text{ (row } p \text{ of } A). \end{aligned}$$

Thus EA is the result of adding k times row p of A to row q, as required. □

The effect of an elementary row operation can be reversed by another such operation (called its inverse) which is also elementary of the same type (see the discussion following (Example 1.1.3). It follows that each elementary matrix E is invertible. In fact, if a row operation on I produces E, then the inverse operation carries E back to I. If F is the elementary matrix corresponding to the inverse operation, this means $FE = I$ (by Lemma 2.5.1). Thus $F = E^{-1}$ and we have proved

Lemma 2.5.2

Every elementary matrix E is invertible, and E^{-1} is also a elementary matrix (of the same type). Moreover, E^{-1} corresponds to the inverse of the row operation that produces E.

The following table gives the inverse of each type of elementary row operation:

Type	Operation	Inverse Operation
I	Interchange rows p and q	Interchange rows p and q
II	Multiply row p by $k \neq 0$	Multiply row p by $1/k$, $k \neq 0$
III	Add k times row p to row $q \neq p$	Subtract k times row p from row q, $q \neq p$

Note that elementary matrices of type I are self-inverse.

[10] A *lemma* is an auxiliary theorem used in the proof of other theorems.

Example 2.5.1

Find the inverse of each of the elementary matrices

$$E_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad E_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 9 \end{bmatrix}, \quad \text{and} \quad E_3 = \begin{bmatrix} 1 & 0 & 5 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Solution. E_1, E_2, and E_3 are of type I, II, and III respectively, so the table gives

$$E_1^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = E_1, \quad E_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{9} \end{bmatrix}, \quad \text{and} \quad E_3^{-1} = \begin{bmatrix} 1 & 0 & -5 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Inverses and Elementary Matrices

Suppose that an $m \times n$ matrix A is carried to a matrix B (written $A \to B$) by a series of k elementary row operations. Let $E_1, E_2, \dots, E_k$ denote the corresponding elementary matrices. By Lemma 2.5.1, the reduction becomes

$$A \to E_1A \to E_2E_1A \to E_3E_2E_1A \to \cdots \to E_kE_{k-1}\cdots E_2E_1A = B$$

In other words,

$$A \to UA = B \quad \text{where } U = E_kE_{k-1}\cdots E_2E_1$$

The matrix $U = E_kE_{k-1}\cdots E_2E_1$ is invertible, being a product of invertible matrices by Lemma 2.5.2. Moreover, U can be computed without finding the E_i as follows: If the above series of operations carrying $A \to B$ is performed on I_m in place of A, the result is $I_m \to UI_m = U$. Hence this series of operations carries the block matrix $\begin{bmatrix} A & I_m \end{bmatrix} \to \begin{bmatrix} B & U \end{bmatrix}$. This, together with the above discussion, proves

Theorem 2.5.1

Suppose A is $m \times n$ and $A \to B$ by elementary row operations.

1. *$B = UA$ where U is an $m \times m$ invertible matrix.*
2. *U can be computed by $\begin{bmatrix} A & I_m \end{bmatrix} \to \begin{bmatrix} B & U \end{bmatrix}$ using the operations carrying $A \to B$.*
3. *$U = E_kE_{k-1}\cdots E_2E_1$ where $E_1, E_2, \dots, E_k$ are the elementary matrices corresponding (in order) to the elementary row operations carrying A to B.*

Example 2.5.2

If $A = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 1 \end{bmatrix}$, express the reduced row-echelon form R of A as $R = UA$ where U is invertible.

Solution. Reduce the double matrix $\left[\begin{array}{cc} A & I \end{array}\right] \to \left[\begin{array}{cc} R & U \end{array}\right]$ as follows:

$$\left[\begin{array}{cc} A & I \end{array}\right] = \left[\begin{array}{ccc|cc} 2 & 3 & 1 & 1 & 0 \\ 1 & 2 & 1 & 0 & 1 \end{array}\right] \to \left[\begin{array}{ccc|cc} 1 & 2 & 1 & 0 & 1 \\ 2 & 3 & 1 & 1 & 0 \end{array}\right] \to \left[\begin{array}{ccc|cc} 1 & 2 & 1 & 0 & 1 \\ 0 & -1 & -1 & 1 & -2 \end{array}\right]$$
$$\to \left[\begin{array}{ccc|cc} 1 & 0 & -1 & 2 & -3 \\ 0 & 1 & 1 & -1 & 2 \end{array}\right]$$

Hence $R = \left[\begin{array}{ccc} 1 & 0 & -1 \\ 0 & 1 & 1 \end{array}\right]$ and $U = \left[\begin{array}{cc} 2 & -3 \\ -1 & 2 \end{array}\right]$.

Now suppose that A is invertible. We know that $A \to I$ by Theorem 2.4.5, so taking $B = I$ in Theorem 2.5.1 gives $\left[\begin{array}{cc} A & I \end{array}\right] \to \left[\begin{array}{cc} I & U \end{array}\right]$ where $I = UA$. Thus $U = A^{-1}$, so we have $\left[\begin{array}{cc} A & I \end{array}\right] \to \left[\begin{array}{cc} I & A^{-1} \end{array}\right]$. This is the matrix inversion algorithm in Section 2.4. However, more is true: Theorem 2.5.1 gives $A^{-1} = U = E_k E_{k-1} \cdots E_2 E_1$ where $E_1, E_2, \dots, E_k$ are the elementary matrices corresponding (in order) to the row operations carrying $A \to I$. Hence

$$A = \left(A^{-1}\right)^{-1} = (E_k E_{k-1} \cdots E_2 E_1)^{-1} = E_1^{-1} E_2^{-1} \cdots E_{k-1}^{-1} E_k^{-1} \tag{2.10}$$

By Lemma 2.5.2, this shows that every invertible matrix A is a product of elementary matrices. Since elementary matrices are invertible (again by Lemma 2.5.2), this proves the following important characterization of invertible matrices.

Theorem 2.5.2

A square matrix is invertible if and only if it is a product of elementary matrices.

It follows from Theorem 2.5.1 that $A \to B$ by row operations if and only if $B = UA$ for some invertible matrix U. In this case we say that A and B are **row-equivalent**. (See Exercise **??**.)

Example 2.5.3

Express $A = \left[\begin{array}{cc} -2 & 3 \\ 1 & 0 \end{array}\right]$ as a product of elementary matrices.

Solution. Using Lemma 2.5.1, the reduction of $A \to I$ is as follows:

$$A = \left[\begin{array}{cc} -2 & 3 \\ 1 & 0 \end{array}\right] \to E_1 A = \left[\begin{array}{cc} 1 & 0 \\ -2 & 3 \end{array}\right] \to E_2 E_1 A = \left[\begin{array}{cc} 1 & 0 \\ 0 & 3 \end{array}\right] \to E_3 E_2 E_1 A = \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right]$$

where the corresponding elementary matrices are

$$E_1 = \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array}\right], \quad E_2 = \left[\begin{array}{cc} 1 & 0 \\ 2 & 1 \end{array}\right], \quad E_3 = \left[\begin{array}{cc} 1 & 0 \\ 0 & \frac{1}{3} \end{array}\right]$$

Hence $(E_3\ E_2\ E_1)A = I$, so:

$$A = (E_3 E_2 E_1)^{-1} = E_1^{-1} E_2^{-1} E_3^{-1} = \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array}\right] \left[\begin{array}{cc} 1 & 0 \\ -2 & 1 \end{array}\right] \left[\begin{array}{cc} 1 & 0 \\ 0 & 3 \end{array}\right]$$

Smith Normal Form

Let A be an $m \times n$ matrix of rank r, and let R be the reduced row-echelon form of A. Theorem 2.5.1 shows that $R = UA$ where U is invertible, and that U can be found from $\left[\begin{array}{cc} A & I_m \end{array}\right] \to \left[\begin{array}{cc} R & U \end{array}\right]$.

The matrix R has r leading ones (since $\text{rank } A = r$) so, as R is reduced, the $n \times m$ matrix R^T contains each row of I_r in the first r columns. Thus row operations will carry $R^T \to \left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right]_{n \times m}$. Hence Theorem 2.5.1 (again) shows that $\left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right]_{n \times m} = U_1 R^T$ where U_1 is an $n \times n$ invertible matrix. Writing $V = U_1^T$, we obtain

$$UAV = RV = RU_1^T = \left(U_1 R^T\right)^T = \left(\left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right]_{n \times m}\right)^T = \left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right]_{m \times n}$$

Moreover, the matrix $U_1 = V^T$ can be computed by $\left[\begin{array}{cc} R^T & I_n \end{array}\right] \to \left[\begin{array}{cc} \left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right]_{n \times m} & V^T \end{array}\right]$. This proves

Theorem 2.5.3

Let A be an $m \times n$ matrix of rank r. There exist invertible matrices U and V of size $m \times m$ and $n \times n$, respectively, such that

$$UAV = \left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right]_{m \times n}$$

Moreover, if R is the reduced row-echelon form of A, then:

1. *U can be computed by $\left[\begin{array}{cc} A & I_m \end{array}\right] \to \left[\begin{array}{cc} R & U \end{array}\right]$;*
2. *V can be computed by $\left[\begin{array}{cc} R^T & I_n \end{array}\right] \to \left[\begin{array}{cc} \left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right]_{n \times m} & V^T \end{array}\right]$.*

If A is an $m \times n$ matrix of rank r, the matrix $\left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right]$ is called the **Smith normal form**[11] of A. Whereas the reduced row-echelon form of A is the "nicest" matrix to which A can be carried by row operations, the Smith canonical form is the "nicest" matrix to which A can be carried by *row and column* operations. This is because doing row operations to R^T amounts to doing *column* operations to R and then transposing.

Example 2.5.4

Given $A = \left[\begin{array}{rrrr} 1 & -1 & 1 & 2 \\ 2 & -2 & 1 & -1 \\ -1 & 1 & 0 & 3 \end{array}\right]$, find invertible matrices U and V such that $UAV = \left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right]$, where $r = \text{rank } A$.

Solution. The matrix U and the reduced row-echelon form R of A are computed by the row

[11] Named after Henry John Stephen Smith (1826–83).

reduction $\begin{bmatrix} A & I_3 \end{bmatrix} \to \begin{bmatrix} R & U \end{bmatrix}$:

$$\left[\begin{array}{rrrr|rrr} 1 & -1 & 1 & 2 & 1 & 0 & 0 \\ 2 & -2 & 1 & -1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 3 & 0 & 0 & 1 \end{array}\right] \to \left[\begin{array}{rrrr|rrr} 1 & -1 & 0 & -3 & -1 & 1 & 0 \\ 0 & 0 & 1 & 5 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 1 \end{array}\right]$$

Hence

$$R = \left[\begin{array}{rrrr} 1 & -1 & 0 & -3 \\ 0 & 0 & 1 & 5 \\ 0 & 0 & 0 & 0 \end{array}\right] \quad \text{and} \quad U = \left[\begin{array}{rrr} -1 & 1 & 0 \\ 2 & -1 & 0 \\ -1 & 1 & 1 \end{array}\right]$$

In particular, $r = \operatorname{rank} R = 2$. Now row-reduce $\begin{bmatrix} R^T & I_4 \end{bmatrix} \to \left[\begin{array}{cc} \left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right] & V^T \end{array}\right]$:

$$\left[\begin{array}{rrr|rrrr} 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ -3 & 5 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \to \left[\begin{array}{rrr|rrrr} 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & -5 & 1 \end{array}\right]$$

whence

$$V^T = \left[\begin{array}{rrrr} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 3 & 0 & -5 & -1 \end{array}\right] \quad \text{so} \quad V = \left[\begin{array}{rrrr} 1 & 0 & 1 & 3 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & -5 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

Then $UAV = \left[\begin{array}{cc} I_2 & 0 \\ 0 & 0 \end{array}\right]$ as is easily verified.

Uniqueness of the Reduced Row-echelon Form

In this short subsection, Theorem 2.5.1 is used to prove the following important theorem.

Theorem 2.5.4

If a matrix A is carried to reduced row-echelon matrices R and S by row operations, then $R = S$.

Proof. Observe first that $UR = S$ for some invertible matrix U (by Theorem 2.5.1 there exist invertible matrices P and Q such that $R = PA$ and $S = QA$; take $U = QP^{-1}$). We show that $R = S$ by induction on the number m of rows of R and S. The case $m = 1$ is left to the reader. If R_j and S_j denote column j in R and S respectively, the fact that $UR = S$ gives

$$UR_j = S_j \quad \text{for each } j \tag{2.11}$$

Since U is invertible, this shows that R and S have the same zero columns. Hence, by passing to the matrices obtained by deleting the zero columns from R and S, we may assume that R and S have no zero columns.

But then the first column of R and S is the first column of I_m because R and S are row-echelon, so (2.11) shows that the first column of U is column 1 of I_m. Now write U, R, and S in block form as follows.

$$U = \begin{bmatrix} 1 & X \\ 0 & V \end{bmatrix}, \quad R = \begin{bmatrix} 1 & Y \\ 0 & R' \end{bmatrix}, \quad \text{and} \quad S = \begin{bmatrix} 1 & Z \\ 0 & S' \end{bmatrix}$$

Since $UR = S$, block multiplication gives $VR' = S'$ so, since V is invertible (U is invertible) and both R' and S' are reduced row-echelon, we obtain $R' = S'$ by induction. Hence R and S have the same number (say r) of leading 1s, and so both have m–r zero rows.

In fact, R and S have leading ones in the same columns, say r of them. Applying (2.11) to these columns shows that the first r columns of U are the first r columns of I_m. Hence we can write U, R, and S in block form as follows:

$$U = \begin{bmatrix} I_r & M \\ 0 & W \end{bmatrix}, \quad R = \begin{bmatrix} R_1 & R_2 \\ 0 & 0 \end{bmatrix}, \quad \text{and} \quad S = \begin{bmatrix} S_1 & S_2 \\ 0 & 0 \end{bmatrix}$$

where R_1 and S_1 are $r \times r$. Then using $UR = S$ block multiplication gives $R_1 = S_1$ and $R_2 = S_2$; that is, $S = R$. This completes the proof. □

2.6 Linear Transformations

If A is an $m \times n$ matrix, recall that the transformation $T_A : \mathbb{R}^n \to \mathbb{R}^m$ defined by

$$T_A(\mathbf{x}) = A\mathbf{x} \quad \text{for all } \mathbf{x} \text{ in } \mathbb{R}^n$$

is called the *matrix transformation induced* by A. In Section 2.2, we saw that many important geometric transformations were in fact matrix transformations. These transformations can be characterized in a different way. The new idea is that of a linear transformation, one of the basic notions in linear algebra. We define these transformations in this section, and show that they are really just the matrix transformations looked at in another way. Having these two ways to view them turns out to be useful because, in a given situation, one perspective or the other may be preferable.

Linear Transformations

Definition 2.2 Linear Transformations $\mathbb{R}^n \to \mathbb{R}^m$

A transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ is called a ***linear transformation*** *if it satisfies the following two conditions for all vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$ and all scalars a:*

T1 $\quad T(\mathbf{x}+\mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y})$

T2 $\quad T(a\mathbf{x}) = aT(\mathbf{x})$

Of course, $\mathbf{x}+\mathbf{y}$ and $a\mathbf{x}$ here are computed in $\mathbb{R}^n$, while $T(\mathbf{x}) + T(\mathbf{y})$ and $aT(\mathbf{x})$ are in $\mathbb{R}^m$. We say that T *preserves addition* if T1 holds, and that T *preserves scalar multiplication* if T2 holds. Moreover, taking $a = 0$ and $a = -1$ in T2 gives

$$T(\mathbf{0}) = \mathbf{0} \quad \text{and} \quad T(-\mathbf{x}) = -T(\mathbf{x}) \quad \text{for all } \mathbf{x}$$

Hence T preserves the zero vector and the negative of a vector. Even more is true.

Recall that a vector $\mathbf{y}$ in $\mathbb{R}^n$ is called a **linear combination** of vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ if $\mathbf{y}$ has the form

$$\mathbf{y} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_k\mathbf{x}_k$$

for some scalars $a_1, a_2, \dots, a_k$. Conditions T1 and T2 combine to show that every linear transformation T *preserves linear combinations* in the sense of the following theorem. This result is used repeatedly in linear algebra.

Theorem 2.6.1: Linearity Theorem

If $T : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation, then for each $k = 1, 2, \dots$

$$T(a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_k\mathbf{x}_k) = a_1T(\mathbf{x}_1) + a_2T(\mathbf{x}_2) + \cdots + a_kT(\mathbf{x}_k)$$

for all scalars a_i and all vectors $\mathbf{x}_i$ in $\mathbb{R}^n$.

Proof. If $k = 1$, it reads $T(a_1\mathbf{x}_1) = a_1T(\mathbf{x}_1)$ which is Condition T1. If $k = 2$, we have

$$\begin{aligned} T(a_1\mathbf{x}_1 + a_2\mathbf{x}_2) &= T(a_1\mathbf{x}_1) + T(a_2\mathbf{x}_2) && \text{by Condition T1} \\ &= a_1T(\mathbf{x}_1) + a_2T(\mathbf{x}_2) && \text{by Condition T2} \end{aligned}$$

If $k = 3$, we use the case $k = 2$ to obtain

$$\begin{array}{rcll} T(a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + a_3\mathbf{x}_3) & = & T\left[(a_1\mathbf{x}_1 + a_2\mathbf{x}_2) + a_3\mathbf{x}_3\right] & \text{collect terms} \\ & = & T(a_1\mathbf{x}_1 + a_2\mathbf{x}_2) + T(a_3\mathbf{x}_3) & \text{by Condition T1} \\ & = & [a_1T(\mathbf{x}_1) + a_2T(\mathbf{x}_2)] + T(a_3\mathbf{x}_3) & \text{by the case } k = 2 \\ & = & [a_1T(\mathbf{x}_1) + a_2T(\mathbf{x}_2)] + a_3T(\mathbf{x}_3) & \text{by Condition T2} \end{array}$$

The proof for any k is similar, using the previous case $k-1$ and Conditions T1 and T2. □

The method of proof in Theorem 2.6.1 is called *mathematical induction* (Appendix C).

Theorem 2.6.1 shows that if T is a linear transformation and $T(\mathbf{x}_1)$, $T(\mathbf{x}_2)$, $\dots$, $T(\mathbf{x}_k)$ are all known, then $T(\mathbf{y})$ can be easily computed for any linear combination $\mathbf{y}$ of $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$. This is a very useful property of linear transformations, and is illustrated in the next two examples.

Example 2.6.1

If $T : \mathbb{R}^3 \to \mathbb{R}^2$ is a linear transformation, $T\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$, $T\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$, and $T\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -4 \\ 4 \end{bmatrix}$, find $T\begin{bmatrix} 3 \\ -4 \\ 2 \end{bmatrix}$.

Solution. Write $\mathbf{w} = \begin{bmatrix} 3 \\ -4 \\ 2 \end{bmatrix}$, $\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $\mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, and $\mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$. Then we know $T(\mathbf{e}_1)$, $T(\mathbf{e}_2)$, and $T(\mathbf{e}_3)$ and we want $T(\mathbf{w})$, so it is enough by Theorem 2.6.1 to express $\mathbf{w}$ as a linear combination of $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$. But clearly $\mathbf{w} = 3\mathbf{e}_1 - 4\mathbf{e}_2 + 2\mathbf{e}_3$. Thus Theorem 2.6.1 gives

$$T(\mathbf{w}) = 3T(\mathbf{e}_1) - 4T(\mathbf{e}_2) + 2T(\mathbf{e}_3) = 3\begin{bmatrix} 2 \\ -1 \end{bmatrix} - 4\begin{bmatrix} 1 \\ -2 \end{bmatrix} + 2\begin{bmatrix} -4 \\ 4 \end{bmatrix} = \begin{bmatrix} -6 \\ 13 \end{bmatrix}$$

The vectors $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ form what is called the standard basis of $\mathbb{R}^3$, more on this below in Definition 2.3. Here is possibly a more subtle example.

Example 2.6.2

If $T : \mathbb{R}^2 \to \mathbb{R}^2$ is a linear transformation, $T\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$ and $T\begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$, find $T\begin{bmatrix} 4 \\ 3 \end{bmatrix}$.

Solution. Write $\mathbf{z} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, and $\mathbf{y} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$ for convenience. Then we know $T(\mathbf{x})$ and $T(\mathbf{y})$ and we want $T(\mathbf{z})$, so it is enough by Theorem 2.6.1 to express $\mathbf{z}$ as a linear combination of $\mathbf{x}$ and $\mathbf{y}$. That is, we want to find numbers a and b such that $\mathbf{z} = a\mathbf{x} + b\mathbf{y}$. Equating entries gives two equations $4 = a + b$ and $3 = a - 2b$. The solution is, $a = \frac{11}{3}$ and $b = \frac{1}{3}$, so $\mathbf{z} = \frac{11}{3}\mathbf{x} + \frac{1}{3}\mathbf{y}$. Thus

Theorem 2.6.1 gives

$$T(\mathbf{z}) = \tfrac{11}{3}T(\mathbf{x}) + \tfrac{1}{3}T(\mathbf{y}) = \tfrac{11}{3}\left[\begin{array}{r} 2 \\ -3 \end{array}\right] + \tfrac{1}{3}\left[\begin{array}{r} 5 \\ 1 \end{array}\right] = \tfrac{1}{3}\left[\begin{array}{r} 27 \\ -32 \end{array}\right]$$

This is what we wanted.

We now show that any matrix transformation is a linear transformation.

Example 2.6.3

If A is $m \times n$, the matrix transformation $T_A : \mathbb{R}^n \to \mathbb{R}^m$, is a linear transformation.

Solution. We have $T_A(\mathbf{x}) = A\mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^n$, so Theorem 2.2.2 gives

$$T_A(\mathbf{x}+\mathbf{y}) = A(\mathbf{x}+\mathbf{y}) = A\mathbf{x} + A\mathbf{y} = T_A(\mathbf{x}) + T_A(\mathbf{y})$$

and

$$T_A(a\mathbf{x}) = A(a\mathbf{x}) = a(A\mathbf{x}) = aT_A(\mathbf{x})$$

hold for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$ and all scalars a. Hence T_A satisfies T1 and T2, and so is linear.

The remarkable thing is that the *converse* of Example 2.6.3 is true: Every linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ is actually a matrix transformation. To see why, we define the **standard basis** of $\mathbb{R}^n$.

Definition 2.3 Standard Basis of $\mathbb{R}^n$

The standard basis of $\mathbb{R}^n$ is the set of columns $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ of the identity matrix I_n. That is:

$$\mathbf{e}_1 = \left[\begin{array}{c} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{array}\right], \ \mathbf{e}_2 = \left[\begin{array}{c} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{array}\right], \ \cdots, \ \mathbf{e}_n = \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{array}\right].$$

Then each $\mathbf{e}_i$ is in $\mathbb{R}^n$ and every vector $\mathbf{x} = \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array}\right]$ in $\mathbb{R}^n$ is a linear combination of the $\mathbf{e}_i$. In fact:

$$\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n$$

as the reader can verify. Hence Theorem 2.6.1 shows that

$$T(\mathbf{x}) = T(x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n) = x_1T(\mathbf{e}_1) + x_2T(\mathbf{e}_2) + \cdots + x_nT(\mathbf{e}_n)$$

Now observe that each $T(\mathbf{e}_i)$ is a column in $\mathbb{R}^m$, so

$$A = \left[\begin{array}{cccc} T(\mathbf{e}_1) & T(\mathbf{e}_2) & \cdots & T(\mathbf{e}_n) \end{array} \right]$$

is an $m \times n$ matrix. Hence we can apply Definition 2.5 to get

$$T(\mathbf{x}) = x_1T(\mathbf{e}_1) + x_2T(\mathbf{e}_2) + \cdots + x_nT(\mathbf{e}_n) = \left[\begin{array}{cccc} T(\mathbf{e}_1) & T(\mathbf{e}_2) & \cdots & T(\mathbf{e}_n) \end{array} \right] \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \right] = A\mathbf{x}$$

Since this holds for every $\mathbf{x}$ in $\mathbb{R}^n$, it shows that T is the matrix transformation induced by A, and so proves most of the following theorem.

Theorem 2.6.2

Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a transformation.

1. *T is linear if and only if it is a matrix transformation.*
2. *In this case $T = T_A$ is the matrix transformation induced by a unique $m \times n$ matrix A, given in terms of its columns by*

$$A = \left[\begin{array}{cccc} T(\mathbf{e}_1) & T(\mathbf{e}_2) & \cdots & T(\mathbf{e}_n) \end{array} \right]$$

where $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ is the standard basis of $\mathbb{R}^n$.

Proof. It remains to verify that the matrix A is unique. Suppose that T is induced by another matrix B. Then $T(\mathbf{x}) = B\mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^n$. But $T(\mathbf{x}) = A\mathbf{x}$ for each $\mathbf{x}$, so $B\mathbf{x} = A\mathbf{x}$ for every $\mathbf{x}$. Hence $A = B$ by Theorem 2.2.6. □

Hence we can speak of *the* matrix of a linear transformation. Because of Theorem 2.6.2 we may (and shall) use the phrases "linear transformation" and "matrix transformation" interchangeably.

Example 2.6.4

Define $T : \mathbb{R}^3 \to \mathbb{R}^2$ by $T \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} \right] = \left[\begin{array}{c} x_1 \\ x_2 \end{array} \right]$ for all $\left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} \right]$ in $\mathbb{R}^3$. Show that T is a linear transformation and use Theorem 2.6.2 to find its matrix.

Solution. Write $\mathbf{x} = \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} \right]$ and $\mathbf{y} = \left[\begin{array}{c} y_1 \\ y_2 \\ y_3 \end{array} \right]$, so that $\mathbf{x} + \mathbf{y} = \left[\begin{array}{c} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \end{array} \right]$. Hence

$$T(\mathbf{x} + \mathbf{y}) = \left[\begin{array}{c} x_1 + y_1 \\ x_2 + y_2 \end{array} \right] = \left[\begin{array}{c} x_1 \\ x_2 \end{array} \right] + \left[\begin{array}{c} y_1 \\ y_2 \end{array} \right] = T(\mathbf{x}) + T(\mathbf{y})$$

Similarly, the reader can verify that $T(a\mathbf{x}) = aT(\mathbf{x})$ for all a in $\mathbb{R}$, so T is a linear transformation.

Now the standard basis of $\mathbb{R}^3$ is

$$\mathbf{e}_1 = \left[\begin{array}{c} 1 \\ 0 \\ 0 \end{array}\right], \quad \mathbf{e}_2 = \left[\begin{array}{c} 0 \\ 1 \\ 0 \end{array}\right], \quad \text{and} \quad \mathbf{e}_3 = \left[\begin{array}{c} 0 \\ 0 \\ 1 \end{array}\right]$$

so, by Theorem 2.6.2, the matrix of T is

$$A = \left[\begin{array}{ccc} T(\mathbf{e}_1) & T(\mathbf{e}_2) & T(\mathbf{e}_3) \end{array}\right] = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \end{array}\right]$$

Of course, the fact that $T\left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right] = \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \end{array}\right]\left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right]$ shows directly that T is a matrix transformation (hence linear) and reveals the matrix.

To illustrate how Theorem 2.6.2 is used, we rederive the matrices of the transformations in Examples 2.2.13 and 2.2.15.

Example 2.6.5

Let $Q_0 : \mathbb{R}^2 \to \mathbb{R}^2$ denote reflection in the x axis (as in Example 2.2.13) and let $R_{\frac{\pi}{2}} : \mathbb{R}^2 \to \mathbb{R}^2$ denote counterclockwise rotation through $\frac{\pi}{2}$ about the origin (as in Example 2.2.15). Use Theorem 2.6.2 to find the matrices of Q_0 and $R_{\frac{\pi}{2}}$.

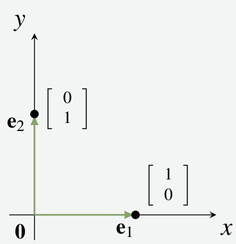


Figure 2.6.1

Solution. Observe that Q_0 and $R_{\frac{\pi}{2}}$ are linear by Example 2.6.3 (they are matrix transformations), so Theorem 2.6.2 applies to them. The standard basis of $\mathbb{R}^2$ is $\{\mathbf{e}_1, \mathbf{e}_2\}$ where $\mathbf{e}_1 = \left[\begin{array}{c} 1 \\ 0 \end{array}\right]$ points along the positive x axis, and $\mathbf{e}_2 = \left[\begin{array}{c} 0 \\ 1 \end{array}\right]$ points along the positive y axis (see Figure 2.6.1).

The reflection of $\mathbf{e}_1$ in the x axis is $\mathbf{e}_1$ itself because $\mathbf{e}_1$ points along the x axis, and the reflection of $\mathbf{e}_2$ in the x axis is $-\mathbf{e}_2$ because $\mathbf{e}_2$ is perpendicular to the x axis. In other words, $Q_0(\mathbf{e}_1) = \mathbf{e}_1$ and $Q_0(\mathbf{e}_2) = -\mathbf{e}_2$. Hence Theorem 2.6.2 shows that the matrix of Q_0 is

$$\left[\begin{array}{cc} Q_0(\mathbf{e}_1) & Q_0(\mathbf{e}_2) \end{array}\right] = \left[\begin{array}{cc} \mathbf{e}_1 & -\mathbf{e}_2 \end{array}\right] = \left[\begin{array}{rr} 1 & 0 \\ 0 & -1 \end{array}\right]$$

which agrees with Example 2.2.13.

Similarly, rotating $\mathbf{e}_1$ through $\frac{\pi}{2}$ counterclockwise about the origin produces $\mathbf{e}_2$, and rotating $\mathbf{e}_2$ through $\frac{\pi}{2}$ counterclockwise about the origin gives $-\mathbf{e}_1$. That is, $R_{\frac{\pi}{2}}(\mathbf{e}_1) = \mathbf{e}_2$ and $R_{\frac{\pi}{2}}(\mathbf{e}_2) = -\mathbf{e}_2$. Hence, again by Theorem 2.6.2, the matrix of $R_{\frac{\pi}{2}}$ is

$$\left[\begin{array}{cc} R_{\frac{\pi}{2}}(\mathbf{e}_1) & R_{\frac{\pi}{2}}(\mathbf{e}_2) \end{array}\right] = \left[\begin{array}{cc} \mathbf{e}_2 & -\mathbf{e}_1 \end{array}\right] = \left[\begin{array}{rr} 0 & -1 \\ 1 & 0 \end{array}\right]$$

agreeing with Example 2.2.15.

Example 2.6.6

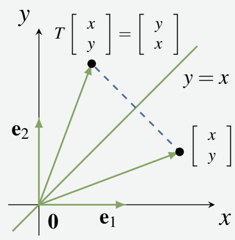


Figure 2.6.2

Let $Q_1 : \mathbb{R}^2 \to \mathbb{R}^2$ denote reflection in the line $y = x$. Show that Q_1 is a matrix transformation, find its matrix, and use it to illustrate Theorem 2.6.2.

Solution. Figure 2.6.2 shows that $Q_1\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} y \\ x \end{bmatrix}$. Hence $Q_1\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} y \\ x \end{bmatrix}$, so Q_1 is the matrix transformation induced by the matrix $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Hence Q_1 is linear (by Example 2.6.3) and so Theorem 2.6.2 applies. If $\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ are the standard basis of $\mathbb{R}^2$, then it is clear geometrically that $Q_1(\mathbf{e}_1) = \mathbf{e}_2$ and $Q_1(\mathbf{e}_2) = \mathbf{e}_1$. Thus (by Theorem 2.6.2) the matrix of Q_1 is $\begin{bmatrix} Q_1(\mathbf{e}_1) & Q_1(\mathbf{e}_2) \end{bmatrix} = \begin{bmatrix} \mathbf{e}_2 & \mathbf{e}_1 \end{bmatrix} = A$ as before.

Recall that, given two "linked" transformations

$$\mathbb{R}^k \xrightarrow{T} \mathbb{R}^n \xrightarrow{S} \mathbb{R}^m$$

we can apply T first and then apply S, and so obtain a new transformation

$$S \circ T : \mathbb{R}^k \to \mathbb{R}^m$$

called the **composite** of S and T, defined by

$$(S \circ T)(\mathbf{x}) = S[T(\mathbf{x})] \text{ for all } \mathbf{x} \text{ in } \mathbb{R}^k$$

If S and T are linear, the action of $S \circ T$ can be computed by multiplying their matrices.

Theorem 2.6.3

Let $\mathbb{R}^k \xrightarrow{T} \mathbb{R}^n \xrightarrow{S} \mathbb{R}^m$ be linear transformations, and let A and B be the matrices of S and T respectively. Then $S \circ T$ is linear with matrix AB.

Proof. $(S \circ T)(\mathbf{x}) = S[T(\mathbf{x})] = A[B\mathbf{x}] = (AB)\mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^k$. □

Theorem 2.6.3 shows that the action of the composite $S \circ T$ is determined by the matrices of S and T. But it also provides a very useful interpretation of matrix multiplication. If A and B are matrices, the product matrix AB induces the transformation resulting from first applying B and then applying A. Thus the study of matrices can cast light on geometrical transformations and vice-versa. Here is an example.

Example 2.6.7

Show that reflection in the x axis followed by rotation through $\frac{\pi}{2}$ is reflection in the line $y = x$.

Solution. The composite in question is $R_{\frac{\pi}{2}} \circ Q_0$ where Q_0 is reflection in the x axis and $R_{\frac{\pi}{2}}$ is rotation through $\frac{\pi}{2}$. By Example 2.6.5, $R_{\frac{\pi}{2}}$ has matrix $A = \left[\begin{array}{rr} 0 & -1 \\ 1 & 0 \end{array}\right]$ and Q_0 has matrix $B = \left[\begin{array}{rr} 1 & 0 \\ 0 & -1 \end{array}\right]$. Hence Theorem 2.6.3 shows that the matrix of $R_{\frac{\pi}{2}} \circ Q_0$ is $AB = \left[\begin{array}{rr} 0 & -1 \\ 1 & 0 \end{array}\right]\left[\begin{array}{rr} 1 & 0 \\ 0 & -1 \end{array}\right] = \left[\begin{array}{rr} 0 & 1 \\ 1 & 0 \end{array}\right]$, which is the matrix of reflection in the line $y = x$ by Example 2.6.4.

This conclusion can also be seen geometrically. Let $\mathbf{x}$ be a typical point in $\mathbb{R}^2$, and assume that $\mathbf{x}$ makes an angle α with the positive x axis. The effect of first applying Q_0 and then applying $R_{\frac{\pi}{2}}$ is shown in Figure 2.6.3. The fact that $R_{\frac{\pi}{2}}\left[Q_0(\mathbf{x})\right]$ makes the angle α with the positive y axis shows that $R_{\frac{\pi}{2}}\left[Q_0(\mathbf{x})\right]$ is the reflection of $\mathbf{x}$ in the line $y = x$.

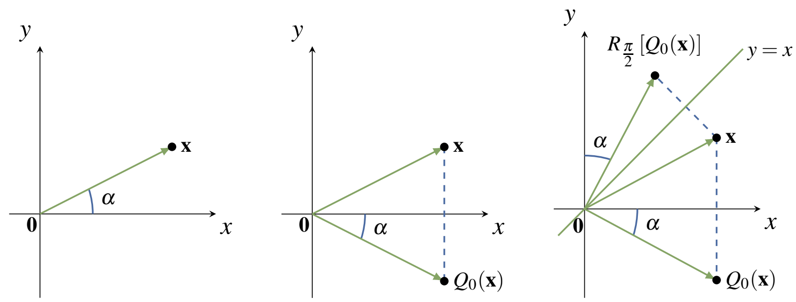


Figure 2.6.3

In Theorem 2.6.3, we saw that the matrix of the composite of two linear transformations is the product of their matrices (in fact, matrix products were defined so that this is the case). We are going to apply this fact to rotations, reflections, and projections in the plane. Before proceeding, we pause to present useful geometrical descriptions of vector addition and scalar multiplication in the plane, and to give a short review of angles and the trigonometric functions.

Some Geometry

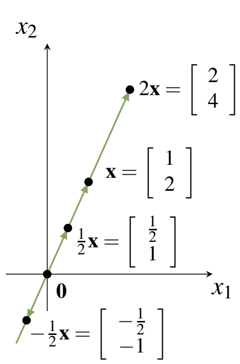


Figure 2.6.4

As we have seen, it is convenient to view a vector $\mathbf{x}$ in $\mathbb{R}^2$ as an arrow from the origin to the point $\mathbf{x}$ (see Section 2.2). This enables us to visualize what sums and scalar multiples mean geometrically. For example consider $\mathbf{x} = \left[\begin{array}{c} 1 \\ 2 \end{array}\right]$ in $\mathbb{R}^2$. Then $2\mathbf{x} = \left[\begin{array}{c} 2 \\ 4 \end{array}\right]$, $\frac{1}{2}\mathbf{x} = \left[\begin{array}{c} \frac{1}{2} \\ 1 \end{array}\right]$ and $-\frac{1}{2}\mathbf{x} = \left[\begin{array}{c} -\frac{1}{2} \\ -1 \end{array}\right]$, and these are shown as arrows in Figure 2.6.4.

Observe that the arrow for $2\mathbf{x}$ is twice as long as the arrow for $\mathbf{x}$ and in the same direction, and that the arrows for $\frac{1}{2}\mathbf{x}$ is also in the same direction as the arrow for $\mathbf{x}$, but only half as long. On the other hand, the arrow for $-\frac{1}{2}\mathbf{x}$ is half as long as the arrow for $\mathbf{x}$, but in the *opposite* direction.

More generally, we have the following geometrical description of scalar multiplication in $\mathbb{R}^2$:

Theorem: Scalar Multiple Law

Let $\mathbf{x}$ *be a vector in* $\mathbb{R}^2$*. The arrow for* $k\mathbf{x}$ *is* $|k|$ *times*[12] *as long as the arrow for* $\mathbf{x}$*, and is in the same direction as the arrow for* $\mathbf{x}$ *if* $k > 0$*, and in the opposite direction if* $k < 0$*.*

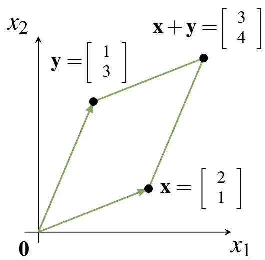


Figure 2.6.5

Now consider two vectors $\mathbf{x} = \left[\begin{array}{c} 2 \\ 1 \end{array}\right]$ and $\mathbf{y} = \left[\begin{array}{c} 1 \\ 3 \end{array}\right]$ in $\mathbb{R}^2$. They are plotted in Figure 2.6.5 along with their sum $\mathbf{x}+\mathbf{y} = \left[\begin{array}{c} 3 \\ 4 \end{array}\right]$. It is a routine matter to verify that the four points $\mathbf{0}$, $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{x}+\mathbf{y}$ form the vertices of a **parallelogram**–that is opposite sides are parallel and of the same length. (The reader should verify that the side from $\mathbf{0}$ to $\mathbf{x}$ has slope of $\frac{1}{2}$, as does the side from $\mathbf{y}$ to $\mathbf{x}+\mathbf{y}$, so these sides are parallel.) We state this as follows:

Theorem: Parallelogram Law

Consider vectors $\mathbf{x}$ *and* $\mathbf{y}$ *in* $\mathbb{R}^2$*. If the arrows for* $\mathbf{x}$ *and* $\mathbf{y}$ *are drawn (see Figure 2.6.6), the arrow for* $\mathbf{x}+\mathbf{y}$ *corresponds to the fourth vertex of the parallelogram determined by the points* $\mathbf{x}$*,* $\mathbf{y}$*, and* $\mathbf{0}$*.*

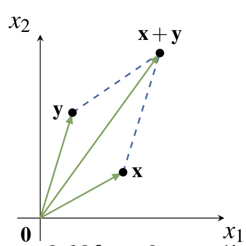


Figure 2.6.6

We will have more to say about this in Chapter 4.

Before proceeding we turn to a brief review of angles and the trigonometric functions. Recall that an angle θ is said to be in **standard position** if it is measured counterclockwise from the positive x axis (as in Figure 2.6.7). Then θ uniquely determines a point $\mathbf{p}$ on the **unit circle** (radius 1, centre at the origin). The **radian** measure of θ is the length of the arc on the unit circle from the positive x axis to $\mathbf{p}$. Thus $360^\circ = 2\pi$ radians, $180^\circ = \pi$, $90^\circ = \frac{\pi}{2}$, and so on.

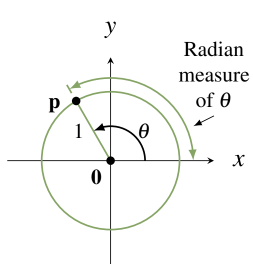


Figure 2.6.7

The point $\mathbf{p}$ in Figure 2.6.7 is also closely linked to the trigonometric functions **cosine** and **sine**, written $\cos\theta$ and $\sin\theta$ respectively. In fact these functions are *defined* to be the x and y coordinates of $\mathbf{p}$; that is $\mathbf{p} = \left[\begin{array}{c} \cos\theta \\ \sin\theta \end{array}\right]$. This defines $\cos\theta$ and $\sin\theta$ for the arbitrary angle θ (possibly negative), and agrees with the usual values when θ is an acute angle $\left(0 \leq \theta \leq \frac{\pi}{2}\right)$ as the reader should verify. For more discussion of this, see Appendix A.

[12] If k is a real number, $|k|$ denotes the **absolute value** of k; that is, $|k| = k$ if $k \geq 0$ and $|k| = -k$ if $k < 0$.

Rotations

We can now describe rotations in the plane. Given an angle θ, let

$$R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$$

denote counterclockwise rotation of $\mathbb{R}^2$ about the origin through the angle θ. The action of R_θ is depicted in Figure 2.6.8. We have already looked at $R_{\frac{\pi}{2}}$ (in Example 2.2.15) and found it to be a matrix transformation. It turns out that R_θ is a matrix transformation for *every* angle θ (with a simple formula for the matrix), but it is not clear how to find the matrix. Our approach is to first establish the (somewhat surprising) fact that R_θ is *linear*, and then obtain the matrix from Theorem 2.6.2.

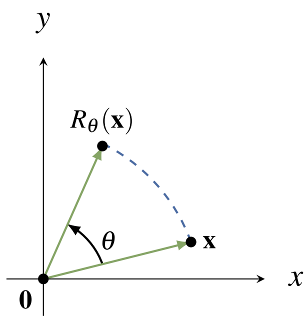


Figure 2.6.8

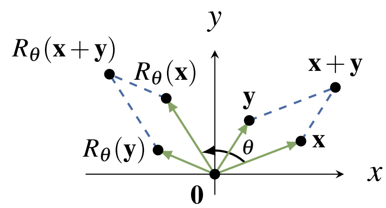


Figure 2.6.9

Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors in $\mathbb{R}^2$. Then $\mathbf{x}+\mathbf{y}$ is the diagonal of the parallelogram determined by $\mathbf{x}$ and $\mathbf{y}$ as in Figure 2.6.9.

The effect of R_θ is to rotate the *entire* parallelogram to obtain the new parallelogram determined by $R_\theta(\mathbf{x})$ and $R_\theta(\mathbf{y})$, with diagonal $R_\theta(\mathbf{x}+\mathbf{y})$. But this diagonal is $R_\theta(\mathbf{x})+R_\theta(\mathbf{y})$ by the parallelogram law (applied to the new parallelogram). It follows that

$$R_\theta(\mathbf{x}+\mathbf{y}) = R_\theta(\mathbf{x})+R_\theta(\mathbf{y})$$

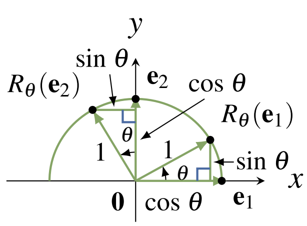


Figure 2.6.10

A similar argument shows that $R_\theta(a\mathbf{x}) = aR_\theta(\mathbf{x})$ for any scalar a, so $R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ is indeed a linear transformation.

With linearity established we can find the matrix of R_θ. Let $\mathbf{e}_1 = \left[\begin{array}{c} 1 \\ 0 \end{array}\right]$ and $\mathbf{e}_2 = \left[\begin{array}{c} 0 \\ 1 \end{array}\right]$ denote the standard basis of $\mathbb{R}^2$. By Figure 2.6.10 we see that

$$R_\theta(\mathbf{e}_1) = \left[\begin{array}{c} \cos\theta \\ \sin\theta \end{array}\right] \quad \text{and} \quad R_\theta(\mathbf{e}_2) = \left[\begin{array}{c} -\sin\theta \\ \cos\theta \end{array}\right]$$

Hence Theorem 2.6.2 shows that R_θ is induced by the matrix

$$\left[\begin{array}{cc} R_\theta(\mathbf{e}_1) & R_\theta(\mathbf{e}_2) \end{array}\right] = \left[\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right]$$

We record this as

Theorem 2.6.4

The rotation $R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ *is the linear transformation with matrix* $\left[\begin{array}{rr} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right]$.

For example, $R_{\frac{\pi}{2}}$ and R_π have matrices $\left[\begin{array}{rr} 0 & -1 \\ 1 & 0 \end{array}\right]$ and $\left[\begin{array}{rr} -1 & 0 \\ 0 & -1 \end{array}\right]$, respectively, by Theorem 2.6.4. The first of these confirms the result in Example 2.2.15. The second shows that rotating a vector $\mathbf{x} = \left[\begin{array}{c} x \\ y \end{array}\right]$ through the angle π results in $R_\pi(\mathbf{x}) = \left[\begin{array}{rr} -1 & 0 \\ 0 & -1 \end{array}\right]\left[\begin{array}{c} x \\ y \end{array}\right] = \left[\begin{array}{c} -x \\ -y \end{array}\right] = -\mathbf{x}$. Thus applying R_π is the same as negating $\mathbf{x}$, a fact that is evident without Theorem 2.6.4.

Example 2.6.8

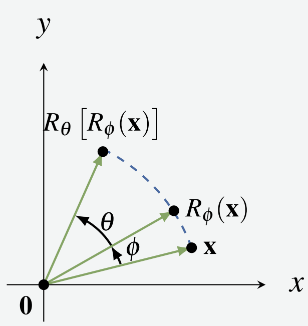


Figure 2.6.11

Let θ and ϕ be angles. By finding the matrix of the composite $R_\theta \circ R_\phi$, obtain expressions for $\cos(\theta+\phi)$ and $\sin(\theta+\phi)$.

Solution. Consider the transformations $\mathbb{R}^2 \xrightarrow{R_\phi} \mathbb{R}^2 \xrightarrow{R_\theta} \mathbb{R}^2$. Their composite $R_\theta \circ R_\phi$ is the transformation that first rotates the plane through ϕ and then rotates it through θ, and so is the rotation through the angle $\theta+\phi$ (see Figure 2.6.11).
In other words

$$R_{\theta+\phi} = R_\theta \circ R_\phi$$

Theorem 2.6.3 shows that the corresponding equation holds for the matrices of these transformations, so Theorem 2.6.4 gives:

$$\left[\begin{array}{rr} \cos(\theta+\phi) & -\sin(\theta+\phi) \\ \sin(\theta+\phi) & \cos(\theta+\phi) \end{array}\right] = \left[\begin{array}{rr} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right]\left[\begin{array}{rr} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{array}\right]$$

If we perform the matrix multiplication on the right, and then compare first column entries, we obtain

$$\begin{aligned} \cos(\theta+\phi) &= \cos\theta\cos\phi - \sin\theta\sin\phi \\ \sin(\theta+\phi) &= \sin\theta\cos\phi + \cos\theta\sin\phi \end{aligned}$$

These are the two basic identities from which most of trigonometry can be derived.

Reflections

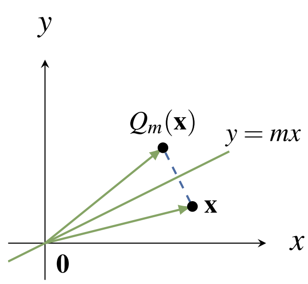


Figure 2.6.12

The line through the origin with slope m has equation $y = mx$, and we let $Q_m : \mathbb{R}^2 \to \mathbb{R}^2$ denote reflection in the line $y = mx$.

This transformation is described geometrically in Figure 2.6.12. In words, $Q_m(\mathbf{x})$ is the "mirror image" of $\mathbf{x}$ in the line $y = mx$. If $m = 0$ then Q_0 is reflection in the x axis, so we already know Q_0 is linear. While we could show directly that Q_m is linear (with an argument like that for R_θ), we prefer to do it another way that is instructive and derives the matrix of Q_m directly without using Theorem 2.6.2.

Let θ denote the angle between the positive x axis and the line $y = mx$. The key observation is that the transformation Q_m can be accomplished in three steps: First rotate through $-\theta$ (so our line coincides with the x axis), then reflect in the x axis, and finally rotate back through θ. In other words:

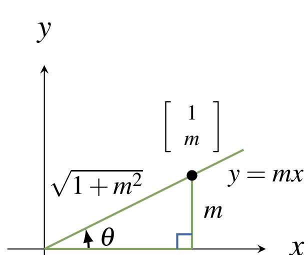
$$Q_m = R_\theta \circ Q_0 \circ R_{-\theta}$$

Since $R_{-\theta}$, Q_0, and R_θ are all linear, this (with Theorem 2.6.3) shows that Q_m is linear and that its matrix is the product of the matrices of R_θ, Q_0, and $R_{-\theta}$. If we write $c = \cos\theta$ and $s = \sin\theta$ for simplicity, then the matrices of R_θ, $R_{-\theta}$, and Q_0 are

$$\left[\begin{array}{rr} c & -s \\ s & c \end{array}\right], \quad \left[\begin{array}{rr} c & s \\ -s & c \end{array}\right], \quad \text{and} \quad \left[\begin{array}{rr} 1 & 0 \\ 0 & -1 \end{array}\right] \text{ respectively.}^{13}$$

Hence, by Theorem 2.6.3, the matrix of $Q_m = R_\theta \circ Q_0 \circ R_{-\theta}$ is

$$\left[\begin{array}{rr} c & -s \\ s & c \end{array}\right] \left[\begin{array}{rr} 1 & 0 \\ 0 & -1 \end{array}\right] \left[\begin{array}{rr} c & s \\ -s & c \end{array}\right] = \left[\begin{array}{cc} c^2 - s^2 & 2sc \\ 2sc & s^2 - c^2 \end{array}\right]$$



We can obtain this matrix in terms of m alone. Figure 2.6.13 shows that

$$\cos\theta = \frac{1}{\sqrt{1+m^2}} \text{ and } \sin\theta = \frac{m}{\sqrt{1+m^2}}$$

so the matrix $\left[\begin{array}{cc} c^2 - s^2 & 2sc \\ 2sc & s^2 - c^2 \end{array}\right]$ of Q_m becomes $\frac{1}{1+m^2} \left[\begin{array}{cc} 1 - m^2 & 2m \\ 2m & m^2 - 1 \end{array}\right]$.

Theorem 2.6.5

Let Q_m denote reflection in the line $y = mx$. Then Q_m is a linear transformation with matrix $\frac{1}{1+m^2} \left[\begin{array}{cc} 1 - m^2 & 2m \\ 2m & m^2 - 1 \end{array}\right]$.

Note that if $m = 0$, the matrix in Theorem 2.6.5 becomes $\left[\begin{array}{rr} 1 & 0 \\ 0 & -1 \end{array}\right]$, as expected. Of course this analysis fails for reflection

[13] The matrix of $R_{-\theta}$ comes from the matrix of R_θ using the fact that, for all angles θ, $\cos(-\theta) = \cos\theta$ and $\sin(-\theta) = -\sin(\theta)$.

in the y axis because vertical lines have no slope. However it is an easy exercise to verify directly that reflection in the y axis is indeed linear with matrix $\left[\begin{array}{rr} -1 & 0 \\ 0 & 1 \end{array}\right]$.[14]

Example 2.6.9

Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be rotation through $-\frac{\pi}{2}$ followed by reflection in the y axis. Show that T is a reflection in a line through the origin and find the line.

Solution. The matrix of $R_{-\frac{\pi}{2}}$ is $\left[\begin{array}{cc} \cos(-\frac{\pi}{2}) & -\sin(-\frac{\pi}{2}) \\ \sin(-\frac{\pi}{2}) & \cos(-\frac{\pi}{2}) \end{array}\right] = \left[\begin{array}{rr} 0 & 1 \\ -1 & 0 \end{array}\right]$ and the matrix of reflection in the y axis is $\left[\begin{array}{rr} -1 & 0 \\ 0 & 1 \end{array}\right]$. Hence the matrix of T is $\left[\begin{array}{rr} -1 & 0 \\ 0 & 1 \end{array}\right]\left[\begin{array}{rr} 0 & 1 \\ -1 & 0 \end{array}\right] = \left[\begin{array}{rr} 0 & -1 \\ -1 & 0 \end{array}\right]$ and this is reflection in the line $y = -x$ (take $m = -1$ in Theorem 2.6.5).

Projections

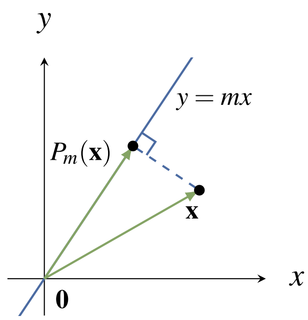


Figure 2.6.14

The method in the proof of Theorem 2.6.5 works more generally. Let $P_m : \mathbb{R}^2 \to \mathbb{R}^2$ denote projection on the line $y = mx$. This transformation is described geometrically in Figure 2.6.14.

If $m = 0$, then $P_0 \left[\begin{array}{c} x \\ y \end{array}\right] = \left[\begin{array}{c} x \\ 0 \end{array}\right]$ for all $\left[\begin{array}{c} x \\ y \end{array}\right]$ in $\mathbb{R}^2$, so P_0 is linear with matrix $\left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array}\right]$. Hence the argument above for Q_m goes through for P_m. First observe that

$$P_m = R_\theta \circ P_0 \circ R_{-\theta}$$

as before. So, P_m is linear with matrix

$$\left[\begin{array}{rr} c & -s \\ s & c \end{array}\right]\left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array}\right]\left[\begin{array}{rr} c & s \\ -s & c \end{array}\right] = \left[\begin{array}{cc} c^2 & sc \\ sc & s^2 \end{array}\right]$$

where $c = \cos\theta = \frac{1}{\sqrt{1+m^2}}$ and $s = \sin\theta = \frac{m}{\sqrt{1+m^2}}$.

This gives:

Theorem 2.6.6

Let $P_m : \mathbb{R}^2 \to \mathbb{R}^2$ be projection on the line $y = mx$. Then P_m is a linear transformation with matrix $\frac{1}{1+m^2}\left[\begin{array}{cc} 1 & m \\ m & m^2 \end{array}\right]$.

[14] Note that $\left[\begin{array}{rr} -1 & 0 \\ 0 & 1 \end{array}\right] = \lim\limits_{m\to\infty} \frac{1}{1+m^2}\left[\begin{array}{cc} 1-m^2 & 2m \\ 2m & m^2-1 \end{array}\right]$.

Again, if $m=0$, then the matrix in Theorem 2.6.6 reduces to $\left[\begin{array}{cc}1 & 0\\0 & 0\end{array}\right]$ as expected. As the y axis has no slope, the analysis fails for projection on the y axis, but this transformation is indeed linear with matrix $\left[\begin{array}{cc}0 & 0\\0 & 1\end{array}\right]$ as is easily verified directly.

Note that the formula for the matrix of Q_m in Theorem 2.6.5 can be derived from the above formula for the matrix of P_m. Using Figure 2.6.12, observe that $Q_m(\mathbf{x}) = \mathbf{x} + 2[P_m(\mathbf{x}) - \mathbf{x}]$ so $Q_m(x) = 2P_m(\mathbf{x}) - \mathbf{x}$. Substituting the matrices for $P_m(\mathbf{x})$ and $1_{\mathbb{R}^2}(\mathbf{x})$ gives the desired formula.

Example 2.6.10

Given $\mathbf{x}$ in $\mathbb{R}^2$, write $\mathbf{y} = P_m(\mathbf{x})$. The fact that $\mathbf{y}$ lies on the line $y = mx$ means that $P_m(\mathbf{y}) = \mathbf{y}$. But then

$$(P_m \circ P_m)(\mathbf{x}) = P_m(\mathbf{y}) = \mathbf{y} = P_m(\mathbf{x}) \text{ for all } \mathbf{x} \text{ in } \mathbb{R}^2, \quad \text{that is, } P_m \circ P_m = P_m.$$

In particular, if we write the matrix of P_m as $A = \frac{1}{1+m^2}\left[\begin{array}{cc}1 & m\\m & m^2\end{array}\right]$, then $A^2 = A$. The reader should verify this directly.

2.7 LU-Factorization[15]

The solution to a system $A\mathbf{x} = \mathbf{b}$ of linear equations can be solved quickly if A can be factored as $A = LU$ where L and U are of a particularly nice form. In this section we show that gaussian elimination can be used to find such factorizations.

Triangular Matrices

As for square matrices, if $A = \left[a_{ij}\right]$ is an $m \times n$ matrix, the elements $a_{11}, a_{22}, a_{33}, \dots$ form the **main diagonal** of A. Then A is called **upper triangular** if every entry below and to the left of the main diagonal is zero. Every row-echelon matrix is upper triangular, as are the matrices

$$\left[\begin{array}{rrrr} 1 & -1 & 0 & 3 \\ 0 & 2 & 1 & 1 \\ 0 & 0 & -3 & 0 \end{array}\right] \quad \left[\begin{array}{rrrrr} 0 & 2 & 1 & 0 & 5 \\ 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{array}\right] \quad \left[\begin{array}{rrr} 1 & 1 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

By analogy, a matrix A is called **lower triangular** if its transpose is upper triangular, that is if each entry above and to the right of the main diagonal is zero. A matrix is called **triangular** if it is upper or lower triangular.

Example 2.7.1

Solve the system

$$\begin{aligned} x_1 + 2x_2 - 3x_3 - x_4 + 5x_5 &= 3 \\ 5x_3 + x_4 + x_5 &= 8 \\ 2x_5 &= 6 \end{aligned}$$

where the coefficient matrix is upper triangular.

Solution. As in gaussian elimination, let the "non-leading" variables be parameters: $x_2 = s$ and $x_4 = t$. Then solve for x_5, x_3, and x_1 in that order as follows. The last equation gives

$$x_5 = \frac{6}{2} = 3$$

Substitution into the second last equation gives

$$x_3 = 1 - \frac{1}{5}t$$

Finally, substitution of both x_5 and x_3 into the first equation gives

$$x_1 = -9 - 2s + \frac{2}{5}t$$

The method used in Example 2.7.1 is called **back substitution** because later variables are substituted into earlier equations. It works because the coefficient matrix is upper triangular. Similarly, if the coefficient matrix is lower triangular the system can be solved by **forward substitution** where earlier variables are substituted into later equations. As observed in Section 1.2, these procedures are more numerically efficient than gaussian elimination.

[15] This section is not used later and so may be omitted with no loss of continuity.

Now consider a system $A\mathbf{x} = \mathbf{b}$ where A can be factored as $A = LU$ where L is lower triangular and U is upper triangular. Then the system $A\mathbf{x} = \mathbf{b}$ can be solved in two stages as follows:

1. *First solve* $L\mathbf{y} = \mathbf{b}$ *for* $\mathbf{y}$ *by forward substitution.*

2. *Then solve* $U\mathbf{x} = \mathbf{y}$ *for* $\mathbf{x}$ *by back substitution.*

Then $\mathbf{x}$ is a solution to $A\mathbf{x} = \mathbf{b}$ because $A\mathbf{x} = LU\mathbf{x} = L\mathbf{y} = \mathbf{b}$. Moreover, every solution $\mathbf{x}$ arises this way (take $\mathbf{y} = U\mathbf{x}$). Furthermore the method adapts easily for use in a computer.

This focuses attention on efficiently obtaining such factorizations $A = LU$. The following result will be needed; the proof is straightforward and is left as Exercises **??** and **??**.

Lemma 2.7.1

Let A and B denote matrices.

1. *If A and B are both lower (upper) triangular, the same is true of AB.*

2. *If A is $n \times n$ and lower (upper) triangular, then A is invertible if and only if every main diagonal entry is nonzero. In this case A^{-1} is also lower (upper) triangular.*

LU-Factorization

Let A be an $m \times n$ matrix. Then A can be carried to a row-echelon matrix U (that is, upper triangular). As in Section 2.5, the reduction is

$$A \to E_1A \to E_2E_1A \to E_3E_2E_1A \to \cdots \to E_kE_{k-1}\cdots E_2E_1A = U$$

where $E_1, E_2, \dots, E_k$ are elementary matrices corresponding to the row operations used. Hence

$$A = LU$$

where $L = (E_kE_{k-1}\cdots E_2E_1)^{-1} = E_1^{-1}E_2^{-1}\cdots E_{k-1}^{-1}E_k^{-1}$. If we do not insist that U is reduced then, except for row interchanges, none of these row operations involve adding a row to a row *above* it. Thus, if no row interchanges are used, all the E_i are *lower* triangular, and so L is lower triangular (and invertible) by Lemma 2.7.1. This proves the following theorem. For convenience, let us say that A can be **lower reduced** if it can be carried to row-echelon form using no row interchanges.

Theorem 2.7.1

If A can be lower reduced to a row-echelon matrix U, then

$$A = LU$$

where L is lower triangular and invertible and U is upper triangular and row-echelon.

Definition 2.4 LU-factorization

A factorization $A = LU$ as in Theorem 2.7.1 is called an ***LU-factorization*** *of A.*

Such a factorization may not exist (Exercise **??**) because A cannot be carried to row-echelon form using no row interchange. A procedure for dealing with this situation will be outlined later. However, if an LU-factorization $A = LU$ does exist, then the gaussian algorithm gives U and also leads to a procedure for finding L. Example 2.7.2 provides an illustration. For convenience, the first nonzero column from the left in a matrix A is called the **leading column** of A.

Example 2.7.2

Find an LU-factorization of $A = \left[\begin{array}{rrrrr} 0 & 2 & -6 & -2 & 4 \\ 0 & -1 & 3 & 3 & 2 \\ 0 & -1 & 3 & 7 & 10 \end{array}\right]$.

Solution. We lower reduce A to row-echelon form as follows:

$$A = \left[\begin{array}{rrrrr} 0 & 2 & -6 & -2 & 4 \\ 0 & -1 & 3 & 3 & 2 \\ 0 & -1 & 3 & 7 & 10 \end{array}\right] \to \left[\begin{array}{rrrrr} 0 & 1 & -3 & -1 & 2 \\ 0 & 0 & 0 & 2 & 4 \\ 0 & 0 & 0 & 6 & 12 \end{array}\right] \to \left[\begin{array}{rrrrr} 0 & 1 & -3 & -1 & 2 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] = U$$

The circled columns are determined as follows: The first is the leading column of A, and is used (by lower reduction) to create the first leading 1 and create zeros below it. This completes the work on row 1, and we repeat the procedure on the matrix consisting of the remaining rows. Thus the second circled column is the leading column of this smaller matrix, which we use to create the second leading 1 and the zeros below it. As the remaining row is zero here, we are finished. Then $A = LU$ where

$$L = \left[\begin{array}{rrr} 2 & 0 & 0 \\ -1 & 2 & 0 \\ -1 & 6 & 1 \end{array}\right]$$

This matrix L is obtained from I_3 by replacing the bottom of the first two columns by the circled columns in the reduction. Note that the rank of A is 2 here, and this is the number of circled columns.

The calculation in Example 2.7.2 works in general. There is no need to calculate the elementary matrices E_i, and the method is suitable for use in a computer because the circled columns can be stored in memory as they are created. The procedure can be formally stated as follows:

Theorem: LU-Algorithm

Let A be an $m \times n$ matrix of rank r, and suppose that A can be lower reduced to a row-echelon matrix U. Then $A = LU$ where the lower triangular, invertible matrix L is constructed as follows:

1. *If $A = 0$, take $L = I_m$ and $U = 0$.*

2. *If* $A \neq 0$, *write* $A_1 = A$ *and let* $\mathbf{c}_1$ *be the leading column of* A_1. *Use* $\mathbf{c}_1$ *to create the first leading* 1 *and create zeros below it (using lower reduction). When this is completed, let* A_2 *denote the matrix consisting of rows 2 to* m *of the matrix just created.*

3. *If* $A_2 \neq 0$, *let* $\mathbf{c}_2$ *be the leading column of* A_2 *and repeat Step 2 on* A_2 *to create* A_3.

4. *Continue in this way until* U *is reached, where all rows below the last leading* 1 *consist of zeros. This will happen after* r *steps.*

5. *Create* L *by placing* $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_r$ *at the bottom of the first* r *columns of* I_m.

A proof of the LU-algorithm is given at the end of this section.

LU-factorization is particularly important if, as often happens in business and industry, a series of equations $A\mathbf{x} = B_1$, $A\mathbf{x} = B_2$, $\dots$, $A\mathbf{x} = B_k$, must be solved, each with the same coefficient matrix A. It is very efficient to solve the first system by gaussian elimination, simultaneously creating an LU-factorization of A, and then using the factorization to solve the remaining systems by forward and back substitution.

Example 2.7.3

Find an LU-factorization for $A = \left[\begin{array}{rrrrr} 5 & -5 & 10 & 0 & 5 \\ -3 & 3 & 2 & 2 & 1 \\ -2 & 2 & 0 & -1 & 0 \\ 1 & -1 & 10 & 2 & 5 \end{array} \right]$.

Solution. The reduction to row-echelon form is

$$
\begin{aligned}
\left[\begin{array}{rrrrr} 5 & -5 & 10 & 0 & 5 \\ -3 & 3 & 2 & 2 & 1 \\ -2 & 2 & 0 & -1 & 0 \\ 1 & -1 & 10 & 2 & 5 \end{array} \right]
&\rightarrow \left[\begin{array}{rrrrr} 1 & -1 & 2 & 0 & 1 \\ 0 & 0 & 8 & 2 & 4 \\ 0 & 0 & 4 & -1 & 2 \\ 0 & 0 & 8 & 2 & 4 \end{array} \right] \\
&\rightarrow \left[\begin{array}{rrrrr} 1 & -1 & 2 & 0 & 1 \\ 0 & 0 & 1 & \frac{1}{4} & \frac{1}{2} \\ 0 & 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right] \\
&\rightarrow \left[\begin{array}{rrrrr} 1 & -1 & 2 & 0 & 1 \\ 0 & 0 & 1 & \frac{1}{4} & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right] = U
\end{aligned}
$$

If U denotes this row-echelon matrix, then $A = LU$, where

$$
L = \left[\begin{array}{rrrr} 5 & 0 & 0 & 0 \\ -3 & 8 & 0 & 0 \\ -2 & 4 & -2 & 0 \\ 1 & 8 & 0 & 1 \end{array} \right]
$$

The next example deals with a case where no row of zeros is present in U (in fact, A is invertible).

Example 2.7.4

Find an LU-factorization for $A = \begin{bmatrix} 2 & 4 & 2 \\ 1 & 1 & 2 \\ -1 & 0 & 2 \end{bmatrix}$.

Solution. The reduction to row-echelon form is

$$\begin{bmatrix} \boxed{\begin{matrix} 2 \\ 1 \\ -1 \end{matrix}} & \begin{matrix} 4 \\ 1 \\ 0 \end{matrix} & \begin{matrix} 2 \\ 2 \\ 2 \end{matrix} \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 1 \\ 0 & \boxed{\begin{matrix} -1 \\ 2 \end{matrix}} & \begin{matrix} 1 \\ 3 \end{matrix} \\ 0 & & \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & \boxed{5} \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} = U$$

Hence $A = LU$ where $L = \begin{bmatrix} 2 & 0 & 0 \\ 1 & -1 & 0 \\ -1 & 2 & 5 \end{bmatrix}$.

There are matrices (for example $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$) that have no LU-factorization and so require at least one row interchange when being carried to row-echelon form via the gaussian algorithm. However, it turns out that, if all the row interchanges encountered in the algorithm are carried out first, the resulting matrix requires no interchanges and so has an LU-factorization. Here is the precise result.

Theorem 2.7.2

Suppose an $m \times n$ matrix A is carried to a row-echelon matrix U via the gaussian algorithm. Let $P_1, P_2, \dots, P_s$ be the elementary matrices corresponding (in order) to the row interchanges used, and write $P = P_s \cdots P_2 P_1$. (If no interchanges are used take $P = I_m$.) Then:

1. *PA is the matrix obtained from A by doing these interchanges (in order) to A.*
2. *PA has an LU-factorization.*

The proof is given at the end of this section.

A matrix P that is the product of elementary matrices corresponding to row interchanges is called a **permutation matrix**. Such a matrix is obtained from the identity matrix by arranging the rows in a different order, so it has exactly one 1 in each row and each column, and has zeros elsewhere. We regard the identity matrix as a permutation matrix. The elementary permutation matrices are those obtained from I by a single row interchange, and every permutation matrix is a product of elementary ones.

Example 2.7.5

If $A = \begin{bmatrix} 0 & 0 & -1 & 2 \\ -1 & -1 & 1 & 2 \\ 2 & 1 & -3 & 6 \\ 0 & 1 & -1 & 4 \end{bmatrix}$, find a permutation matrix P such that PA has an LU-factorization, and then find the factorization.

Solution. Apply the gaussian algorithm to A:

$$A \xrightarrow{*} \begin{bmatrix} -1 & -1 & 1 & 2 \\ 0 & 0 & -1 & 2 \\ 2 & 1 & -3 & 6 \\ 0 & 1 & -1 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & -1 & -2 \\ 0 & 0 & -1 & 2 \\ 0 & -1 & -1 & 10 \\ 0 & 1 & -1 & 4 \end{bmatrix} \xrightarrow{*} \begin{bmatrix} 1 & 1 & -1 & -2 \\ 0 & -1 & -1 & 10 \\ 0 & 0 & -1 & 2 \\ 0 & 1 & -1 & 4 \end{bmatrix}$$
$$\rightarrow \begin{bmatrix} 1 & 1 & -1 & -2 \\ 0 & 1 & 1 & -10 \\ 0 & 0 & -1 & 2 \\ 0 & 0 & -2 & 14 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & -1 & -2 \\ 0 & 1 & 1 & -10 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 10 \end{bmatrix}$$

Two row interchanges were needed (marked with $*$), first rows 1 and 2 and then rows 2 and 3. Hence, as in Theorem 2.7.2,

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

If we do these interchanges (in order) to A, the result is PA. Now apply the LU-algorithm to PA:

$$PA = \begin{bmatrix} -1 & -1 & 1 & 2 \\ 2 & 1 & -3 & 6 \\ 0 & 0 & -1 & 2 \\ 0 & 1 & -1 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & -1 & -2 \\ 0 & -1 & -1 & 10 \\ 0 & 0 & -1 & 2 \\ 0 & 1 & -1 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & -1 & -2 \\ 0 & 1 & 1 & -10 \\ 0 & 0 & -1 & 2 \\ 0 & 0 & -2 & 14 \end{bmatrix}$$
$$\rightarrow \begin{bmatrix} 1 & 1 & -1 & -2 \\ 0 & 1 & 1 & -10 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & \boxed{10} \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & -1 & -2 \\ 0 & 1 & 1 & -10 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \end{bmatrix} = U$$

Hence, $PA = LU$, where $L = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & -2 & 10 \end{bmatrix}$ and $U = \begin{bmatrix} 1 & 1 & -1 & -2 \\ 0 & 1 & 1 & -10 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$.

Theorem 2.7.2 provides an important general factorization theorem for matrices. If A is any $m \times n$

matrix, it asserts that there exists a permutation matrix P and an LU-factorization $PA = LU$. Moreover, it shows that either $P = I$ or $P = P_s \cdots P_2 P_1$, where $P_1, P_2, \dots, P_s$ are the elementary permutation matrices arising in the reduction of A to row-echelon form. Now observe that $P_i^{-1} = P_i$ for each i (they are elementary row interchanges). Thus, $P^{-1} = P_1 P_2 \cdots P_s$, so the matrix A can be factored as

$$A = P^{-1}LU$$

where P^{-1} is a permutation matrix, L is lower triangular and invertible, and U is a row-echelon matrix. This is called a **PLU-factorization** of A.

The LU-factorization in Theorem 2.7.1 is not unique. For example,

$$\left[\begin{array}{cc} 1 & 0 \\ 3 & 2 \end{array}\right]\left[\begin{array}{rrr} 1 & -2 & 3 \\ 0 & 0 & 0 \end{array}\right] = \left[\begin{array}{cc} 1 & 0 \\ 3 & 1 \end{array}\right]\left[\begin{array}{rrr} 1 & -2 & 3 \\ 0 & 0 & 0 \end{array}\right]$$

However, it is necessary here that the row-echelon matrix has a row of zeros. Recall that the rank of a matrix A is the number of nonzero rows in any row-echelon matrix U to which A can be carried by row operations. Thus, if A is $m \times n$, the matrix U has no row of zeros if and only if A has rank m.

Theorem 2.7.3

Let A be an $m \times n$ matrix that has an LU-factorization

$$A = LU$$

If A has rank m *(that is, U has no row of zeros), then L and U are uniquely determined by A.*

Proof. Suppose $A = MV$ is another LU-factorization of A, so M is lower triangular and invertible and V is row-echelon. Hence $LU = MV$, and we must show that $L = M$ and $U = V$. We write $N = M^{-1}L$. Then N is lower triangular and invertible (Lemma 2.7.1) and $NU = V$, so it suffices to prove that $N = I$. If N is $m \times m$, we use induction on m. The case $m = 1$ is left to the reader. If $m > 1$, observe first that column 1 of V is N times column 1 of U. Thus if either column is zero, so is the other (N is invertible). Hence, we can assume (by deleting zero columns) that the $(1,\ 1)$-entry is 1 in both U and V.

Now we write $N = \left[\begin{array}{cc} a & 0 \\ X & N_1 \end{array}\right]$, $U = \left[\begin{array}{cc} 1 & Y \\ 0 & U_1 \end{array}\right]$, and $V = \left[\begin{array}{cc} 1 & Z \\ 0 & V_1 \end{array}\right]$ in block form. Then $NU = V$ becomes $\left[\begin{array}{cc} a & aY \\ X & XY + N_1U_1 \end{array}\right] = \left[\begin{array}{cc} 1 & Z \\ 0 & V_1 \end{array}\right]$. Hence $a = 1$, $Y = Z$, $X = 0$, and $N_1U_1 = V_1$. But $N_1U_1 = V_1$ implies $N_1 = I$ by induction, whence $N = I$. □

If A is an $m \times m$ invertible matrix, then A has rank m by Theorem 2.4.5. Hence, we get the following important special case of Theorem 2.7.3.

Corollary 2.7.1

If an invertible matrix A has an LU-factorization $A = LU$, then L and U are uniquely determined by A.

Of course, in this case U is an upper triangular matrix with 1s along the main diagonal.

Proofs of Theorems

Proof of the LU-Algorithm. If $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_r$ are columns of lengths $m, m-1, \dots, m-r+1$, respectively, write $L^{(m)}[\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_r]$ for the lower triangular $m \times m$ matrix obtained from I_m by placing $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_r$ at the bottom of the first r columns of I_m.

Proceed by induction on n. If $A = 0$ or $n = 1$, it is left to the reader. If $n > 1$, let $\mathbf{c}_1$ denote the leading column of A and let $\mathbf{k}_1$ denote the first column of the $m \times m$ identity matrix. There exist elementary matrices $E_1, \dots, E_k$ such that, in block form,

$$(E_k \cdots E_2 E_1)A = \left[\begin{array}{c|c|c} 0 & \mathbf{k}_1 & \begin{array}{c} X_1 \\ \hline A_1 \end{array} \end{array}\right] \quad \text{where } (E_k \cdots E_2 E_1)\mathbf{c}_1 = \mathbf{k}_1$$

Moreover, each E_j can be taken to be lower triangular (by assumption). Write

$$G = (E_k \cdots E_2 E_1)^{-1} = E_1^{-1} E_2^{-1} \cdots E_k^{-1}$$

Then G is lower triangular, and $G\mathbf{k}_1 = \mathbf{c}_1$. Also, each E_j (and so each E_j^{-1}) is the result of either multiplying row 1 of I_m by a constant or adding a multiple of row 1 to another row. Hence,

$$G = (E_1^{-1} E_2^{-1} \cdots E_k^{-1})I_m = \left[\begin{array}{c|c} \mathbf{c}_1 & \begin{array}{c} 0 \\ \hline I_{m-1} \end{array} \end{array}\right]$$

in block form. Now, by induction, let $A_1 = L_1 U_1$ be an LU-factorization of A_1, where $L_1 = L^{(m-1)}[\mathbf{c}_2, \dots, \mathbf{c}_r]$ and U_1 is row-echelon. Then block multiplication gives

$$G^{-1}A = \left[\begin{array}{c|c|c} 0 & \mathbf{k}_1 & \begin{array}{c} X_1 \\ \hline L_1 U_1 \end{array} \end{array}\right] = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & L_1 \end{array}\right] \left[\begin{array}{c|c|c} 0 & 1 & X_1 \\ \hline 0 & 0 & U_1 \end{array}\right]$$

Hence $A = LU$, where $U = \left[\begin{array}{c|c|c} 0 & 1 & X_1 \\ \hline 0 & 0 & U_1 \end{array}\right]$ is row-echelon and

$$L = \left[\begin{array}{c|c} \mathbf{c}_1 & \begin{array}{c} 0 \\ \hline I_{m-1} \end{array} \end{array}\right] \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & L_1 \end{array}\right] = \left[\begin{array}{c|c} \mathbf{c}_1 & \begin{array}{c} 0 \\ \hline L_1 \end{array} \end{array}\right] = L^{(m)}[\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_r]$$

This completes the proof. □

Proof of Theorem 2.7.2. Let A be a nonzero $m \times n$ matrix and let $\mathbf{k}_j$ denote column j of I_m. There is a permutation matrix P_1 (where either P_1 is elementary or $P_1 = I_m$) such that the first nonzero column $\mathbf{c}_1$ of $P_1 A$ has a nonzero entry on top. Hence, as in the LU-algorithm,

$$L^{(m)}[\mathbf{c}_1]^{-1} \cdot P_1 \cdot A = \left[\begin{array}{c|c|c} 0 & 1 & X_1 \\ \hline 0 & 0 & A_1 \end{array}\right]$$

in block form. Then let P_2 be a permutation matrix (either elementary or I_m) such that

$$P_2 \cdot L^{(m)}[\mathbf{c}_1]^{-1} \cdot P_1 \cdot A = \left[\begin{array}{c|c|c} 0 & 1 & X_1 \\ \hline 0 & 0 & A_1' \end{array}\right]$$

and the first nonzero column $\mathbf{c}_2$ of A_1' has a nonzero entry on top. Thus,

$$L^{(m)}[\mathbf{k}_1, \mathbf{c}_2]^{-1} \cdot P_2 \cdot L^{(m)}[\mathbf{c}_1]^{-1} \cdot P_1 \cdot A = \left[\begin{array}{c|c|c} 0 & 1 & X_1 \\ \hline 0 & 0 & \begin{array}{c|c|c} 0 & 1 & X_2 \\ \hline 0 & 0 & A_2 \end{array} \end{array}\right]$$

in block form. Continue to obtain elementary permutation matrices $P_1, P_2, \dots, P_r$ and columns $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_r$ of lengths $m, m-1, \dots,$ such that

$$(L_r P_r L_{r-1} P_{r-1} \cdots L_2 P_2 L_1 P_1)A = U$$

where U is a row-echelon matrix and $L_j = L^{(m)}\left[\mathbf{k}_1, \dots, \mathbf{k}_{j-1}, \mathbf{c}_j\right]^{-1}$ for each j, where the notation means the first $j-1$ columns are those of I_m. It is not hard to verify that each L_j has the form $L_j = L^{(m)}\left[\mathbf{k}_1, \dots, \mathbf{k}_{j-1}, \mathbf{c}'_j\right]$ where $\mathbf{c}'_j$ is a column of length $m-j+1$. We now claim that each permutation matrix P_k can be "moved past" each matrix L_j to the right of it, in the sense that

$$P_k L_j = L'_j P_k$$

where $L'_j = L^{(m)}\left[\mathbf{k}_1, \dots, \mathbf{k}_{j-1}, \mathbf{c}''_j\right]$ for some column $\mathbf{c}''_j$ of length $m-j+1$. Given that this is true, we obtain a factorization of the form

$$(L_r L'_{r-1} \cdots L'_2 L'_1)(P_r P_{r-1} \cdots P_2 P_1)A = U$$

If we write $P = P_r P_{r-1} \cdots P_2 P_1$, this shows that PA has an LU-factorization because $L_r L'_{r-1} \cdots L'_2 L'_1$ is lower triangular and invertible. All that remains is to prove the following rather technical result. □

Lemma 2.7.2

Let P_k result from interchanging row k of I_m with a row below it. If $j < k$, let c_j be a column of length $m-j+1$. Then there is another column $\mathbf{c}'_j$ of length $m-j+1$ such that

$$P_k \cdot L^{(m)}\left[\mathbf{k}_1, \dots, \mathbf{k}_{j-1}, \mathbf{c}_j\right] = L^{(m)}\left[\mathbf{k}_1, \dots, \mathbf{k}_{j-1}, \mathbf{c}'_j\right] \cdot P_k$$

The proof is left as Exercise **??**.

2.8 An Application to Input-Output Economic Models[16]

In 1973 Wassily Leontief was awarded the Nobel prize in economics for his work on mathematical models.[17] Roughly speaking, an economic system in this model consists of several industries, each of which produces a product and each of which uses some of the production of the other industries. The following example is typical.

Example 2.8.1

A primitive society has three basic needs: food, shelter, and clothing. There are thus three industries in the society—the farming, housing, and garment industries—that produce these commodities. Each of these industries consumes a certain proportion of the total output of each commodity according to the following table.

		OUTPUT		
		Farming	**Housing**	**Garment**
CONSUMPTION	**Farming**	0.4	0.2	0.3
	Housing	0.2	0.6	0.4
	Garment	0.4	0.2	0.3

Find the annual prices that each industry must charge for its income to equal its expenditures.

[16]The applications in this section and the next are independent and may be taken in any order.
[17]See W. W. Leontief, "The world economy of the year 2000," *Scientific American,* Sept. 1980.

Solution. Let p_1, p_2, and p_3 be the prices charged per year by the farming, housing, and garment industries, respectively, for their total output. To see how these prices are determined, consider the farming industry. It receives p_1 for its production in any year. But it *consumes* products from all these industries in the following amounts (from row 1 of the table): 40% of the food, 20% of the housing, and 30% of the clothing. Hence, the expenditures of the farming industry are $0.4p_1 + 0.2p_2 + 0.3p_3$, so

$$0.4p_1 + 0.2p_2 + 0.3p_3 = p_1$$

A similar analysis of the other two industries leads to the following system of equations.

$$\begin{aligned} 0.4p_1 + 0.2p_2 + 0.3p_3 &= p_1 \\ 0.2p_1 + 0.6p_2 + 0.4p_3 &= p_2 \\ 0.4p_1 + 0.2p_2 + 0.3p_3 &= p_3 \end{aligned}$$

This has the matrix form $E\mathbf{p} = \mathbf{p}$, where

$$E = \begin{bmatrix} 0.4 & 0.2 & 0.3 \\ 0.2 & 0.6 & 0.4 \\ 0.4 & 0.2 & 0.3 \end{bmatrix} \quad \text{and} \quad \mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}$$

The equations can be written as the homogeneous system

$$(I - E)\mathbf{p} = \mathbf{0}$$

where I is the 3×3 identity matrix, and the solutions are

$$\mathbf{p} = \begin{bmatrix} 2t \\ 3t \\ 2t \end{bmatrix}$$

where t is a parameter. Thus, the pricing must be such that the total output of the farming industry has the same value as the total output of the garment industry, whereas the total value of the housing industry must be $\frac{3}{2}$ as much.

In general, suppose an economy has n industries, each of which uses some (possibly none) of the production of every industry. We assume first that the economy is **closed** (that is, no product is exported or imported) and that all product is used. Given two industries i and j, let e_{ij} denote the proportion of the total annual output of industry j that is consumed by industry i. Then $E = \left[e_{ij}\right]$ is called the **input-output** matrix for the economy. Clearly,

$$0 \leq e_{ij} \leq 1 \quad \text{for all } i \text{ and } j \tag{2.12}$$

Moreover, all the output from industry j is used by *some* industry (the model is closed), so

$$e_{1j} + e_{2j} + \cdots + e_{ij} = 1 \quad \text{for each } j \tag{2.13}$$

This condition asserts that each column of E sums to 1. Matrices satisfying conditions (2.12) and (2.13) are called **stochastic matrices**.

As in Example 2.8.1, let p_i denote the price of the total annual production of industry i. Then p_i is the annual revenue of industry i. On the other hand, industry i spends $e_{i1}p_1 + e_{i2}p_2 + \cdots + e_{in}p_n$ annually for the product it uses ($e_{ij}p_j$ is the cost for product from industry j). The closed economic system is said to be in **equilibrium** if the annual expenditure equals the annual revenue for each industry—that is, if

$$e_{i1}p_1 + e_{i2}p_2 + \cdots + e_{in}p_n = p_i \quad \text{for each } i = 1, 2, \dots, n$$

If we write $\mathbf{p} = \left[\begin{array}{c} p_1 \\ p_2 \\ \vdots \\ p_n \end{array} \right]$, these equations can be written as the matrix equation

$$E\mathbf{p} = \mathbf{p}$$

This is called the **equilibrium condition**, and the solutions $\mathbf{p}$ are called **equilibrium price structures**. The equilibrium condition can be written as

$$(I - E)\mathbf{p} = \mathbf{0}$$

which is a system of homogeneous equations for $\mathbf{p}$. Moreover, there is always a nontrivial solution $\mathbf{p}$. Indeed, the column sums of $I - E$ are all 0 (because E is stochastic), so the row-echelon form of $I - E$ has a row of zeros. In fact, more is true:

Theorem 2.8.1

Let E be any $n \times n$ stochastic matrix. Then there is a nonzero $n \times 1$ vector $\mathbf{p}$ with nonnegative entries such that $E\mathbf{p} = \mathbf{p}$. If all the entries of E are positive, the matrix $\mathbf{p}$ can be chosen with all entries positive.

Theorem 2.8.1 guarantees the existence of an equilibrium price structure for any closed input-output system of the type discussed here. The proof is beyond the scope of this book.[18]

Example 2.8.2

Find the equilibrium price structures for four industries if the input-output matrix is

$$E = \left[\begin{array}{rrrr} 0.6 & 0.2 & 0.1 & 0.1 \\ 0.3 & 0.4 & 0.2 & 0 \\ 0.1 & 0.3 & 0.5 & 0.2 \\ 0 & 0.1 & 0.2 & 0.7 \end{array} \right]$$

Find the prices if the total value of business is \$1000.

Solution. If $\mathbf{p} = \left[\begin{array}{c} p_1 \\ p_2 \\ p_3 \\ p_4 \end{array} \right]$ is the equilibrium price structure, then the equilibrium condition reads

[18] The interested reader is referred to P. Lancaster's *Theory of Matrices* (New York: Academic Press, 1969) or to E. Seneta's *Non-negative Matrices* (New York: Wiley, 1973).

$E\mathbf{p} = \mathbf{p}$. When we write this as $(I - E)\mathbf{p} = \mathbf{0}$, the methods of Chapter 1 yield the following family of solutions:

$$\mathbf{p} = \left[\begin{array}{c} 44t \\ 39t \\ 51t \\ 47t \end{array} \right]$$

where t is a parameter. If we insist that $p_1 + p_2 + p_3 + p_4 = 1000$, then $t = 5.525$. Hence

$$\mathbf{p} = \left[\begin{array}{c} 243.09 \\ 215.47 \\ 281.76 \\ 259.67 \end{array} \right]$$

to five figures.

The Open Model

We now assume that there is a demand for products in the **open sector** of the economy, which is the part of the economy other than the producing industries (for example, consumers). Let d_i denote the total value of the demand for product i in the open sector. If p_i and e_{ij} are as before, the value of the annual demand for product i by the producing industries themselves is $e_{i1}p_1 + e_{i2}p_2 + \cdots + e_{in}p_n$, so the total annual revenue p_i of industry i breaks down as follows:

$$p_i = (e_{i1}p_1 + e_{i2}p_2 + \cdots + e_{in}p_n) + d_i \quad \text{for each } i = 1, 2, \dots, n$$

The column $\mathbf{d} = \left[\begin{array}{c} d_1 \\ \vdots \\ d_n \end{array} \right]$ is called the **demand matrix**, and this gives a matrix equation

$$\mathbf{p} = E\mathbf{p} + \mathbf{d}$$

or

$$(I - E)\mathbf{p} = \mathbf{d} \tag{2.14}$$

This is a system of linear equations for $\mathbf{p}$, and we ask for a solution $\mathbf{p}$ with every entry nonnegative. Note that every entry of E is between 0 and 1, but the column sums of E need not equal 1 as in the closed model.

Before proceeding, it is convenient to introduce a useful notation. If $A = \left[a_{ij}\right]$ and $B = \left[b_{ij}\right]$ are matrices of the same size, we write $A > B$ if $a_{ij} > b_{ij}$ for all i and j, and we write $A \geq B$ if $a_{ij} \geq b_{ij}$ for all i and j. Thus $P \geq 0$ means that every entry of P is nonnegative. Note that $A \geq 0$ and $B \geq 0$ implies that $AB \geq 0$.

Now, given a demand matrix $\mathbf{d} \geq \mathbf{0}$, we look for a production matrix $\mathbf{p} \geq \mathbf{0}$ satisfying equation (2.14). This certainly exists if $I - E$ is invertible and $(I - E)^{-1} \geq 0$. On the other hand, the fact that $\mathbf{d} \geq \mathbf{0}$ means any solution $\mathbf{p}$ to equation (2.14) satisfies $\mathbf{p} \geq E\mathbf{p}$. Hence, the following theorem is not too surprising.

Theorem 2.8.2

Let $E \geq 0$ be a square matrix. Then $I-E$ is invertible and $(I-E)^{-1} \geq 0$ if and only if there exists a column $\mathbf{p} > \mathbf{0}$ such that $\mathbf{p} > E\mathbf{p}$.

Heuristic Proof.
If $(I-E)^{-1} \geq 0$, the existence of $\mathbf{p} > \mathbf{0}$ with $\mathbf{p} > E\mathbf{p}$ is left as Exercise **??**. Conversely, suppose such a column $\mathbf{p}$ exists. Observe that

$$(I-E)(I+E+E^2+\cdots+E^{k-1}) = I-E^k$$

holds for all $k \geq 2$. If we can show that every entry of E^k approaches 0 as k becomes large then, intuitively, the infinite matrix sum

$$U = I+E+E^2+\cdots$$

exists and $(I-E)U = I$. Since $U \geq 0$, this does it. To show that E^k approaches 0, it suffices to show that $EP < \mu P$ for some number μ with $0 < \mu < 1$ (then $E^kP < \mu^kP$ for all $k \geq 1$ by induction). The existence of μ is left as Exercise **??**. □

The condition $\mathbf{p} > E\mathbf{p}$ in Theorem 2.8.2 has a simple economic interpretation. If $\mathbf{p}$ is a production matrix, entry i of $E\mathbf{p}$ is the total value of all product used by industry i in a year. Hence, the condition $\mathbf{p} > E\mathbf{p}$ means that, for each i, the value of product produced by industry i exceeds the value of the product it uses. In other words, each industry runs at a profit.

Example 2.8.3

If $E = \begin{bmatrix} 0.6 & 0.2 & 0.3 \\ 0.1 & 0.4 & 0.2 \\ 0.2 & 0.5 & 0.1 \end{bmatrix}$, show that $I-E$ is invertible and $(I-E)^{-1} \geq 0$.

Solution. Use $\mathbf{p} = (3,\ 2,\ 2)^T$ in Theorem 2.8.2.

If $\mathbf{p}_0 = (1,\ 1,\ 1)^T$, the entries of $E\mathbf{p}_0$ are the row sums of E. Hence $\mathbf{p}_0 > E\mathbf{p}_0$ holds if the row sums of E are all less than 1. This proves the first of the following useful facts (the second is Exercise **??**).

Corollary 2.8.1

Let $E \geq 0$ be a square matrix. In each case, $I-E$ is invertible and $(I-E)^{-1} \geq 0$:

1. *All row sums of E are less than 1.*
2. *All column sums of E are less than 1.*

2.9 An Application to Markov Chains

Many natural phenomena progress through various stages and can be in a variety of states at each stage. For example, the weather in a given city progresses day by day and, on any given day, may be sunny or rainy. Here the states are "sun" and "rain," and the weather progresses from one state to another in daily stages. Another example might be a football team: The stages of its evolution are the games it plays, and the possible states are "win," "draw," and "loss."

The general setup is as follows: A real conceptual "system" is run generating a sequence of outcomes. The system evolves through a series of "stages," and at any stage it can be in any one of a finite number of "states." At any given stage, the state to which it will go at the next stage depends on the past and present history of the system—that is, on the sequence of states it has occupied to date.

Definition 2.5 Markov Chain

A ***Markov chain*** *is such an evolving system wherein the state to which it will go next depends only on its present state and does not depend on the earlier history of the system.*[19]

Even in the case of a Markov chain, the state the system will occupy at any stage is determined only in terms of probabilities. In other words, chance plays a role. For example, if a football team wins a

[19]The name honours Andrei Andreyevich Markov (1856–1922) who was a professor at the university in St. Petersburg, Russia.

particular game, we do not know whether it will win, draw, or lose the next game. On the other hand, we may know that the team tends to persist in winning streaks; for example, if it wins one game it may win the next game $\frac{1}{2}$ of the time, lose $\frac{4}{10}$ of the time, and draw $\frac{1}{10}$ of the time. These fractions are called the **probabilities** of these various possibilities. Similarly, if the team loses, it may lose the next game with probability $\frac{1}{2}$ (that is, half the time), win with probability $\frac{1}{4}$, and draw with probability $\frac{1}{4}$. The probabilities of the various outcomes after a drawn game will also be known.

We shall treat probabilities informally here: *The probability that a given event will occur is the long-run proportion of the time that the event does indeed occur*. Hence, all probabilities are numbers between 0 and 1. A probability of 0 means the event is impossible and never occurs; events with probability 1 are certain to occur.

If a Markov chain is in a particular state, the probabilities that it goes to the various states at the next stage of its evolution are called the **transition probabilities** for the chain, and they are assumed to be known quantities. To motivate the general conditions that follow, consider the following simple example. Here the system is a man, the stages are his successive lunches, and the states are the two restaurants he chooses.

Example 2.9.1

A man always eats lunch at one of two restaurants, A and B. He never eats at A twice in a row. However, if he eats at B, he is three times as likely to eat at B next time as at A. Initially, he is equally likely to eat at either restaurant.

a. What is the probability that he eats at A on the third day after the initial one?

b. What proportion of his lunches does he eat at A?

Solution. The table of transition probabilities follows. The A column indicates that if he eats at A on one day, he never eats there again on the next day and so is certain to go to B.

		Present Lunch	
		A	B
Next	A	0	0.25
Lunch	B	1	0.75

The B column shows that, if he eats at B on one day, he will eat there on the next day $\frac{3}{4}$ of the time and switches to A only $\frac{1}{4}$ of the time.

The restaurant he visits on a given day is not determined. The most that we can expect is to know the probability that he will visit A or B on that day.

Let $\mathbf{s}_m = \left[\begin{array}{c} s_1^{(m)} \\ s_2^{(m)} \end{array} \right]$ denote the *state vector* for day m. Here $s_1^{(m)}$ denotes the probability that he eats at A on day m, and $s_2^{(m)}$ is the probability that he eats at B on day m. It is convenient to let $\mathbf{s}_0$ correspond to the initial day. Because he is equally likely to eat at A or B on that initial day, $s_1^{(0)} = 0.5$ and $s_2^{(0)} = 0.5$, so $\mathbf{s}_0 = \left[\begin{array}{c} 0.5 \\ 0.5 \end{array} \right]$. Now let

$$P = \left[\begin{array}{cc} 0 & 0.25 \\ 1 & 0.75 \end{array} \right]$$

denote the *transition matrix*. We claim that the relationship

$$\mathbf{s}_{m+1} = P\mathbf{s}_m$$

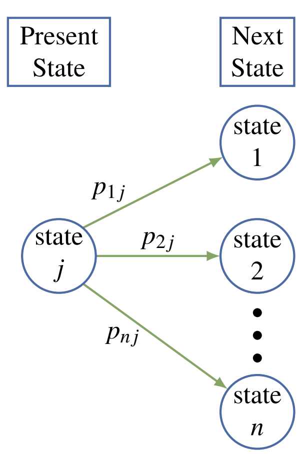
holds for all integers $m \geq 0$. This will be derived later; for now, we use it as follows to successively compute $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \dots$.

$$\mathbf{s}_1 = P\mathbf{s}_0 = \left[\begin{array}{cc} 0 & 0.25 \\ 1 & 0.75 \end{array}\right]\left[\begin{array}{c} 0.5 \\ 0.5 \end{array}\right] = \left[\begin{array}{c} 0.125 \\ 0.875 \end{array}\right]$$
$$\mathbf{s}_2 = P\mathbf{s}_1 = \left[\begin{array}{cc} 0 & 0.25 \\ 1 & 0.75 \end{array}\right]\left[\begin{array}{c} 0.125 \\ 0.875 \end{array}\right] = \left[\begin{array}{c} 0.21875 \\ 0.78125 \end{array}\right]$$
$$\mathbf{s}_3 = P\mathbf{s}_2 = \left[\begin{array}{cc} 0 & 0.25 \\ 1 & 0.75 \end{array}\right]\left[\begin{array}{c} 0.21875 \\ 0.78125 \end{array}\right] = \left[\begin{array}{c} 0.1953125 \\ 0.8046875 \end{array}\right]$$

Hence, the probability that his third lunch (after the initial one) is at A is approximately 0.195, whereas the probability that it is at B is 0.805. If we carry these calculations on, the next state vectors are (to five figures):

$$\mathbf{s}_4 = \left[\begin{array}{c} 0.20117 \\ 0.79883 \end{array}\right] \quad \mathbf{s}_5 = \left[\begin{array}{c} 0.19971 \\ 0.80029 \end{array}\right]$$
$$\mathbf{s}_6 = \left[\begin{array}{c} 0.20007 \\ 0.79993 \end{array}\right] \quad \mathbf{s}_7 = \left[\begin{array}{c} 0.19998 \\ 0.80002 \end{array}\right]$$

Moreover, as m increases the entries of $\mathbf{s}_m$ get closer and closer to the corresponding entries of $\left[\begin{array}{c} 0.2 \\ 0.8 \end{array}\right]$. Hence, in the long run, he eats 20% of his lunches at A and 80% at B.



Example 2.9.1 incorporates most of the essential features of all Markov chains. The general model is as follows: The system evolves through various stages and at each stage can be in exactly one of n distinct states. It progresses through a sequence of states as time goes on. If a Markov chain is in state j at a particular stage of its development, the probability p_{ij} that it goes to state i at the next stage is called the **transition probability**. The $n \times n$ matrix $P = \left[p_{ij}\right]$ is called the **transition matrix** for the Markov chain. The situation is depicted graphically in the diagram.

We make one important assumption about the transition matrix $P = \left[p_{ij}\right]$: It does *not* depend on which stage the process is in. This assumption means that the transition probabilities are *independent of time*—that is, they do not change as time goes on. It is this assumption that distinguishes Markov chains in the literature of this subject.

Example 2.9.2

Suppose the transition matrix of a three-state Markov chain is

$$P = \left[\begin{array}{ccc} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{array}\right] = \begin{array}{c} \text{Present state} \\ \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \left[\begin{array}{ccc} 0.3 & 0.1 & 0.6 \\ 0.5 & 0.9 & 0.2 \\ 0.2 & 0.0 & 0.2 \end{array}\right] \end{array} \begin{array}{l} \\ \\ 1 \\ 2 \quad \text{Next state} \\ 3 \end{array}$$

If, for example, the system is in state 2, then column 2 lists the probabilities of where it goes next. Thus, the probability is $p_{12} = 0.1$ that it goes from state 2 to state 1, and the probability is $p_{22} = 0.9$ that it goes from state 2 to state 2. The fact that $p_{32} = 0$ means that it is impossible for it to go from state 2 to state 3 at the next stage.

Consider the jth column of the transition matrix P.

$$\left[\begin{array}{c} p_{1j} \\ p_{2j} \\ \vdots \\ p_{nj} \end{array}\right]$$

If the system is in state j at some stage of its evolution, the transition probabilities $p_{1j},\ p_{2j},\ \dots,\ p_{nj}$ represent the fraction of the time that the system will move to state 1, state 2, ..., state n, respectively, at the next stage. We assume that it has to go to *some* state at each transition, so the sum of these probabilities is 1:

$$p_{1j} + p_{2j} + \cdots + p_{nj} = 1 \quad \text{for each } j$$

Thus, the columns of P all sum to 1 and the entries of P lie between 0 and 1. Hence P is called a **stochastic matrix**.

As in Example 2.9.1, we introduce the following notation: Let $s_i^{(m)}$ denote the probability that the system is in state i after m transitions. The $n \times 1$ matrices

$$\mathbf{s}_m = \left[\begin{array}{c} s_1^{(m)} \\ s_2^{(m)} \\ \vdots \\ s_n^{(m)} \end{array}\right] \qquad m = 0,\ 1,\ 2,\ \dots$$

are called the **state vectors** for the Markov chain. Note that the sum of the entries of $\mathbf{s}_m$ must equal 1 because the system must be in *some* state after m transitions. The matrix $\mathbf{s}_0$ is called the **initial state vector** for the Markov chain and is given as part of the data of the particular chain. For example, if the chain has only two states, then an initial vector $\mathbf{s}_0 = \left[\begin{array}{c} 1 \\ 0 \end{array}\right]$ means that it started in state 1. If it started in state 2, the initial vector would be $\mathbf{s}_0 = \left[\begin{array}{c} 0 \\ 1 \end{array}\right]$. If $\mathbf{s}_0 = \left[\begin{array}{c} 0.5 \\ 0.5 \end{array}\right]$, it is equally likely that the system started in state 1 or in state 2.

Theorem 2.9.1

Let P be the transition matrix for an n-state Markov chain. If $\mathbf{s}_m$ is the state vector at stage m, then

$$\mathbf{s}_{m+1} = P\mathbf{s}_m$$

for each $m = 0, 1, 2, \dots$.

Heuristic Proof. Suppose that the Markov chain has been run N times, each time starting with the same initial state vector. Recall that p_{ij} is the proportion of the time the system goes from state j at some stage to state i at the next stage, whereas $s_i^{(m)}$ is the proportion of the time it is in state i at stage m. Hence

$$s_i^{m+1}N$$

is (approximately) the number of times the system is in state i at stage $m+1$. We are going to calculate this number another way. The system got to state i at stage $m+1$ through *some* other state (say state j) at stage m. The number of times it was *in* state j at that stage is (approximately) $s_j^{(m)}N$, so the number of times it got to state i via state j is $p_{ij}(s_j^{(m)}N)$. Summing over j gives the number of times the system is in state i (at stage $m+1$). This is the number we calculated before, so

$$s_i^{(m+1)}N = p_{i1}s_1^{(m)}N + p_{i2}s_2^{(m)}N + \cdots + p_{in}s_n^{(m)}N$$

Dividing by N gives $s_i^{(m+1)} = p_{i1}s_1^{(m)} + p_{i2}s_2^{(m)} + \cdots + p_{in}s_n^{(m)}$ for each i, and this can be expressed as the matrix equation $\mathbf{s}_{m+1} = P\mathbf{s}_m$. □

If the initial probability vector $\mathbf{s}_0$ and the transition matrix P are given, Theorem 2.9.1 gives $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \dots,$ one after the other, as follows:

$$\begin{aligned}\mathbf{s}_1 &= P\mathbf{s}_0\\ \mathbf{s}_2 &= P\mathbf{s}_1\\ \mathbf{s}_3 &= P\mathbf{s}_2\\ &\vdots\end{aligned}$$

Hence, the state vector $\mathbf{s}_m$ is completely determined for each $m = 0, 1, 2, \dots$ by P and $\mathbf{s}_0$.

Example 2.9.3

A wolf pack always hunts in one of three regions R_1, R_2, and R_3. Its hunting habits are as follows:

1. If it hunts in some region one day, it is as likely as not to hunt there again the next day.
2. If it hunts in R_1, it never hunts in R_2 the next day.
3. If it hunts in R_2 or R_3, it is equally likely to hunt in each of the other regions the next day.

If the pack hunts in R_1 on Monday, find the probability that it hunts there on Thursday.

Solution. The stages of this process are the successive days; the states are the three regions. The transition matrix P is determined as follows (see the table): The first habit asserts that $p_{11} = p_{22} = p_{33} = \frac{1}{2}$. Now column 1 displays what happens when the pack starts in R_1: It never goes to state 2, so $p_{21} = 0$ and, because the column must sum to 1, $p_{31} = \frac{1}{2}$. Column 2 describes what happens if it starts in R_2: $p_{22} = \frac{1}{2}$ and p_{12} and p_{32} are equal (by habit 3), so $p_{12} = p_{32} = \frac{1}{2}$ because the column sum must equal 1. Column 3 is filled in a similar way.

	R_1	R_2	R_3
R_1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
R_2	0	$\frac{1}{2}$	$\frac{1}{4}$
R_3	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$

Now let Monday be the initial stage. Then $\mathbf{s}_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ because the pack hunts in R_1 on that day.

Then $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ describe Tuesday, Wednesday, and Thursday, respectively, and we compute them using Theorem 2.9.1.

$$\mathbf{s}_1 = P\mathbf{s}_0 = \begin{bmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{bmatrix} \quad \mathbf{s}_2 = P\mathbf{s}_1 = \begin{bmatrix} \frac{3}{8} \\ \frac{1}{8} \\ \frac{4}{8} \end{bmatrix} \quad \mathbf{s}_3 = P\mathbf{s}_2 = \begin{bmatrix} \frac{11}{32} \\ \frac{6}{32} \\ \frac{15}{32} \end{bmatrix}$$

Hence, the probability that the pack hunts in Region R_1 on Thursday is $\frac{11}{32}$.

Steady State Vector

Another phenomenon that was observed in Example 2.9.1 can be expressed in general terms. The state vectors $\mathbf{s}_0$, $\mathbf{s}_1$, $\mathbf{s}_2$, $\dots$ were calculated in that example and were found to "approach" $\mathbf{s} = \begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix}$. This means that the first component of $\mathbf{s}_m$ becomes and remains very close to 0.2 as m becomes large, whereas the second component gets close to 0.8 as m increases. When this is the case, we say that $\mathbf{s}_m$ **converges** to $\mathbf{s}$. For large m, then, there is very little error in taking $\mathbf{s}_m = \mathbf{s}$, so the long-term probability that the system is in state 1 is 0.2, whereas the probability that it is in state 2 is 0.8. In Example 2.9.1, enough state vectors were computed for the limiting vector $\mathbf{s}$ to be apparent. However, there is a better way to do this that works in most cases.

Suppose P is the transition matrix of a Markov chain, and assume that the state vectors $\mathbf{s}_m$ converge to a limiting vector $\mathbf{s}$. Then $\mathbf{s}_m$ is very close to $\mathbf{s}$ for sufficiently large m, so $\mathbf{s}_{m+1}$ is also very close to $\mathbf{s}$. Thus, the equation $\mathbf{s}_{m+1} = P\mathbf{s}_m$ from Theorem 2.9.1 is closely approximated by

$$\mathbf{s} = P\mathbf{s}$$

so it is not surprising that $\mathbf{s}$ should be a solution to this matrix equation. Moreover, it is easily solved because it can be written as a system of homogeneous linear equations

$$(I - P)\mathbf{s} = \mathbf{0}$$

with the entries of $\mathbf{s}$ as variables.

In Example 2.9.1, where $P = \begin{bmatrix} 0 & 0.25 \\ 1 & 0.75 \end{bmatrix}$, the general solution to $(I - P)\mathbf{s} = \mathbf{0}$ is $\mathbf{s} = \begin{bmatrix} t \\ 4t \end{bmatrix}$, where t is a parameter. But if we insist that the entries of S sum to 1 (as must be true of all state vectors), we find $t = 0.2$ and so $\mathbf{s} = \begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix}$ as before.

All this is predicated on the existence of a limiting vector for the sequence of state vectors of the Markov chain, and such a vector may not always exist. However, it does exist in one commonly occurring situation. A stochastic matrix P is called **regular** if some power P^m of P has every entry greater than zero. The matrix $P = \begin{bmatrix} 0 & 0.25 \\ 1 & 0.75 \end{bmatrix}$ of Example 2.9.1 is regular (in this case, each entry of P^2 is positive), and the general theorem is as follows:

Theorem 2.9.2

Let P be the transition matrix of a Markov chain and assume that P is regular. Then there is a unique column matrix $\mathbf{s}$ satisfying the following conditions:

1. $P\mathbf{s} = \mathbf{s}$.
2. *The entries of $\mathbf{s}$ are positive and sum to 1.*

Moreover, condition 1 can be written as

$$(I - P)\mathbf{s} = \mathbf{0}$$

and so gives a homogeneous system of linear equations for $\mathbf{s}$. Finally, the sequence of state vectors $\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \dots$ converges to $\mathbf{s}$ in the sense that if m is large enough, each entry of $\mathbf{s}_m$ is closely approximated by the corresponding entry of $\mathbf{s}$.

This theorem will not be proved here.[20]

If P is the regular transition matrix of a Markov chain, the column $\mathbf{s}$ satisfying conditions 1 and 2 of Theorem 2.9.2 is called the **steady-state vector** for the Markov chain. The entries of $\mathbf{s}$ are the long-term probabilities that the chain will be in each of the various states.

Example 2.9.4

A man eats one of three soups—beef, chicken, and vegetable—each day. He never eats the same soup two days in a row. If he eats beef soup on a certain day, he is equally likely to eat each of the others the next day; if he does not eat beef soup, he is twice as likely to eat it the next day as the alternative.

a. If he has beef soup one day, what is the probability that he has it again two days later?

b. What are the long-run probabilities that he eats each of the three soups?

[20] The interested reader can find an elementary proof in J. Kemeny, H. Mirkil, J. Snell, and G. Thompson, *Finite Mathematical Structures* (Englewood Cliffs, N.J.: Prentice-Hall, 1958).

Solution. The states here are B, C, and V, the three soups. The transition matrix P is given in the table. (Recall that, for each state, the corresponding column lists the probabilities for the next state.)

	B	C	V
B	0	$\frac{2}{3}$	$\frac{2}{3}$
C	$\frac{1}{2}$	0	$\frac{1}{3}$
V	$\frac{1}{2}$	$\frac{1}{3}$	0

If he has beef soup initially, then the initial state vector is

$$\mathbf{s}_0 = \left[\begin{array}{c} 1 \\ 0 \\ 0 \end{array}\right]$$

Then two days later the state vector is $\mathbf{s}_2$. If P is the transition matrix, then

$$\mathbf{s}_1 = P\mathbf{s}_0 = \frac{1}{2}\left[\begin{array}{c} 0 \\ 1 \\ 1 \end{array}\right], \quad \mathbf{s}_2 = P\mathbf{s}_1 = \frac{1}{6}\left[\begin{array}{c} 4 \\ 1 \\ 1 \end{array}\right]$$

so he eats beef soup two days later with probability $\frac{2}{3}$. This answers (a.) and also shows that he eats chicken and vegetable soup each with probability $\frac{1}{6}$.

To find the long-run probabilities, we must find the steady-state vector $\mathbf{s}$. Theorem 2.9.2 applies because P is regular (P^2 has positive entries), so $\mathbf{s}$ satisfies $P\mathbf{s} = \mathbf{s}$. That is, $(I - P)\mathbf{s} = \mathbf{0}$ where

$$I - P = \frac{1}{6}\left[\begin{array}{rrr} 6 & -4 & -4 \\ -3 & 6 & -2 \\ -3 & -2 & 6 \end{array}\right]$$

The solution is $\mathbf{s} = \left[\begin{array}{c} 4t \\ 3t \\ 3t \end{array}\right]$, where t is a parameter, and we use $\mathbf{s} = \left[\begin{array}{c} 0.4 \\ 0.3 \\ 0.3 \end{array}\right]$ because the entries of $\mathbf{s}$ must sum to 1. Hence, in the long run, he eats beef soup 40% of the time and eats chicken soup and vegetable soup each 30% of the time.

Chapter 3

Determinants and Diagonalization

With each square matrix we can calculate a number, called the determinant of the matrix, which tells us whether or not the matrix is invertible. In fact, determinants can be used to give a formula for the inverse of a matrix. They also arise in calculating certain numbers (called eigenvalues) associated with the matrix. These eigenvalues are essential to a technique called diagonalization that is used in many applications where it is desired to predict the future behaviour of a system. For example, we use it to predict whether a species will become extinct.

Determinants were first studied by Leibnitz in 1696, and the term "determinant" was first used in 1801 by Gauss is his *Disquisitiones Arithmeticae*. Determinants are much older than matrices (which were introduced by Cayley in 1878) and were used extensively in the eighteenth and nineteenth centuries, primarily because of their significance in geometry (see Section 4.4). Although they are somewhat less important today, determinants still play a role in the theory and application of matrix algebra.

3.1 The Cofactor Expansion

In Section 2.4 we defined the determinant of a 2×2 matrix $A = \left[\begin{array}{cc} a & b \\ c & d \end{array} \right]$ as follows:[1]

$$\det A = \left| \begin{array}{cc} a & b \\ c & d \end{array} \right| = ad - bc$$

and showed (in Example 2.4.4) that A has an inverse if and only if $\det A \neq 0$. One objective of this chapter is to do this for *any* square matrix A. There is no difficulty for 1×1 matrices: If $A = [a]$, we define $\det A = \det [a] = a$ and note that A is invertible if and only if $a \neq 0$.

If A is 3×3 and invertible, we look for a suitable definition of $\det A$ by trying to carry A to the identity matrix by row operations. The first column is not zero (A is invertible); suppose the (1, 1)-entry a is not zero. Then row operations give

$$A = \left[\begin{array}{ccc} a & b & c \\ d & e & f \\ g & h & i \end{array} \right] \rightarrow \left[\begin{array}{ccc} a & b & c \\ ad & ae & af \\ ag & ah & ai \end{array} \right] \rightarrow \left[\begin{array}{ccc} a & b & c \\ 0 & ae-bd & af-cd \\ 0 & ah-bg & ai-cg \end{array} \right] = \left[\begin{array}{ccc} a & b & c \\ 0 & u & af-cd \\ 0 & v & ai-cg \end{array} \right]$$

where $u = ae - bd$ and $v = ah - bg$. Since A is invertible, one of u and v is nonzero (by Example 2.4.11); suppose that $u \neq 0$. Then the reduction proceeds

$$A \rightarrow \left[\begin{array}{ccc} a & b & c \\ 0 & u & af-cd \\ 0 & v & ai-cg \end{array} \right] \rightarrow \left[\begin{array}{ccc} a & b & c \\ 0 & u & af-cd \\ 0 & uv & u(ai-cg) \end{array} \right] \rightarrow \left[\begin{array}{ccc} a & b & c \\ 0 & u & af-cd \\ 0 & 0 & w \end{array} \right]$$

where $w = u(ai - cg) - v(af - cd) = a(aei + bfg + cdh - ceg - afh - bdi)$. We define

$$\det A = aei + bfg + cdh - ceg - afh - bdi \tag{3.1}$$

and observe that $\det A \neq 0$ because $a \det A = w \neq 0$ (is invertible).

To motivate the definition below, collect the terms in Equation 3.1 involving the entries a, b, and c in row 1 of A:

$$\begin{aligned} \det A = \left| \begin{array}{ccc} a & b & c \\ d & e & f \\ g & h & i \end{array} \right| &= aei + bfg + cdh - ceg - afh - bdi \\ &= a(ei - fh) - b(di - fg) + c(dh - eg) \\ &= a \left| \begin{array}{cc} e & f \\ h & i \end{array} \right| - b \left| \begin{array}{cc} d & f \\ g & i \end{array} \right| + c \left| \begin{array}{cc} d & e \\ g & h \end{array} \right| \end{aligned}$$

This last expression can be described as follows: To compute the determinant of a 3×3 matrix A, multiply each entry in row 1 by a sign times the determinant of the 2×2 matrix obtained by deleting the row and column of that entry, and add the results. The signs alternate down row 1, starting with $+$. It is this observation that we generalize below.

[1]Determinants are commonly written $|A| = \det A$ using vertical bars. We will use both notations.

Example 3.1.1

$$\det \begin{bmatrix} 2 & 3 & 7 \\ -4 & 0 & 6 \\ 1 & 5 & 0 \end{bmatrix} = 2 \begin{vmatrix} 0 & 6 \\ 5 & 0 \end{vmatrix} - 3 \begin{vmatrix} -4 & 6 \\ 1 & 0 \end{vmatrix} + 7 \begin{vmatrix} -4 & 0 \\ 1 & 5 \end{vmatrix}$$
$$= 2(-30) - 3(-6) + 7(-20)$$
$$= -182$$

This suggests an inductive method of defining the determinant of any square matrix in terms of determinants of matrices one size smaller. The idea is to define determinants of 3×3 matrices in terms of determinants of 2×2 matrices, then we do 4×4 matrices in terms of 3×3 matrices, and so on.

To describe this, we need some terminology.

Definition 3.1 Cofactors of a Matrix

Assume that determinants of $(n-1) \times (n-1)$ matrices have been defined. Given the $n \times n$ matrix A, let

A_{ij} denote the $(n-1) \times (n-1)$ matrix obtained from A by deleting row i and column j.

*Then the $(i,\ j)$-**cofactor** $c_{ij}(A)$ is the scalar defined by*

$$c_{ij}(A) = (-1)^{i+j} \det(A_{ij})$$

*Here $(-1)^{i+j}$ is called the **sign** of the $(i,\ j)$-position.*

The sign of a position is clearly 1 or -1, and the following diagram is useful for remembering it:

$$\begin{bmatrix} + & - & + & - & \cdots \\ - & + & - & + & \cdots \\ + & - & + & - & \cdots \\ - & + & - & + & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix}$$

Note that the signs alternate along each row and column with $+$ in the upper left corner.

Example 3.1.2

Find the cofactors of positions $(1,\ 2)$, $(3,\ 1)$, and $(2,\ 3)$ in the following matrix.

$$A = \begin{bmatrix} 3 & -1 & 6 \\ 5 & 2 & 7 \\ 8 & 9 & 4 \end{bmatrix}$$

Solution. Here A_{12} is the matrix $\begin{bmatrix} 5 & 7 \\ 8 & 4 \end{bmatrix}$ that remains when row 1 and column 2 are deleted. The sign of position $(1, 2)$ is $(-1)^{1+2} = -1$ (this is also the $(1, 2)$-entry in the sign diagram), so the $(1, 2)$-cofactor is

$$c_{12}(A) = (-1)^{1+2}\begin{vmatrix} 5 & 7 \\ 8 & 4 \end{vmatrix} = (-1)(5 \cdot 4 - 7 \cdot 8) = (-1)(-36) = 36$$

Turning to position $(3, 1)$, we find

$$c_{31}(A) = (-1)^{3+1}A_{31} = (-1)^{3+1}\begin{vmatrix} -1 & 6 \\ 2 & 7 \end{vmatrix} = (+1)(-7-12) = -19$$

Finally, the $(2, 3)$-cofactor is

$$c_{23}(A) = (-1)^{2+3}A_{23} = (-1)^{2+3}\begin{vmatrix} 3 & -1 \\ 8 & 9 \end{vmatrix} = (-1)(27+8) = -35$$

Clearly other cofactors can be found—there are nine in all, one for each position in the matrix.

We can now define $\det A$ for any square matrix A

Definition 3.2 Cofactor expansion of a Matrix

Assume that determinants of $(n-1) \times (n-1)$ matrices have been defined. If $A = \left[a_{ij}\right]$ is $n \times n$ define

$$\det A = a_{11}c_{11}(A) + a_{12}c_{12}(A) + \cdots + a_{1n}c_{1n}(A)$$

*This is called the **cofactor expansion** of $\det A$ along row 1.*

It asserts that $\det A$ can be computed by multiplying the entries of row 1 by the corresponding cofactors, and adding the results. The astonishing thing is that $\det A$ can be computed by taking the cofactor expansion along *any row or column*: Simply multiply each entry of that row or column by the corresponding cofactor and add.

Theorem 3.1.1: Cofactor Expansion Theorem[2]

The determinant of an $n \times n$ matrix A can be computed by using the cofactor expansion along any row or column of A. That is $\det A$ can be computed by multiplying each entry of the row or column by the corresponding cofactor and adding the results.

The proof will be given in Section 3.6.

[2]The cofactor expansion is due to Pierre Simon de Laplace (1749–1827), who discovered it in 1772 as part of a study of linear differential equations. Laplace is primarily remembered for his work in astronomy and applied mathematics.

Example 3.1.3

Compute the determinant of $A = \begin{bmatrix} 3 & 4 & 5 \\ 1 & 7 & 2 \\ 9 & 8 & -6 \end{bmatrix}$.

Solution. The cofactor expansion along the first row is as follows:

$$\begin{aligned} \det A &= 3c_{11}(A) + 4c_{12}(A) + 5c_{13}(A) \\ &= 3\begin{vmatrix} 7 & 2 \\ 8 & -6 \end{vmatrix} - 4\begin{vmatrix} 1 & 2 \\ 9 & -6 \end{vmatrix} + 5\begin{vmatrix} 1 & 7 \\ 9 & 8 \end{vmatrix} \\ &= 3(-58) - 4(-24) + 5(-55) \\ &= -353 \end{aligned}$$

Note that the signs alternate along the row (indeed along *any* row or column). Now we compute $\det A$ by expanding along the first column.

$$\begin{aligned} \det A &= 3c_{11}(A) + 1c_{21}(A) + 9c_{31}(A) \\ &= 3\begin{vmatrix} 7 & 2 \\ 8 & -6 \end{vmatrix} - \begin{vmatrix} 4 & 5 \\ 8 & -6 \end{vmatrix} + 9\begin{vmatrix} 4 & 5 \\ 7 & 2 \end{vmatrix} \\ &= 3(-58) - (-64) + 9(-27) \\ &= -353 \end{aligned}$$

The reader is invited to verify that $\det A$ can be computed by expanding along any other row or column.

The fact that the cofactor expansion along *any row or column* of a matrix A always gives the same result (the determinant of A) is remarkable, to say the least. The choice of a particular row or column can simplify the calculation.

Example 3.1.4

Compute $\det A$ where $A = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 5 & 1 & 2 & 0 \\ 2 & 6 & 0 & -1 \\ -6 & 3 & 1 & 0 \end{bmatrix}$.

Solution. The first choice we must make is which row or column to use in the cofactor expansion. The expansion involves multiplying entries by cofactors, so the work is minimized when the row or column contains as many zero entries as possible. Row 1 is a best choice in this matrix (column 4 would do as well), and the expansion is

$$\begin{aligned} \det A &= 3c_{11}(A) + 0c_{12}(A) + 0c_{13}(A) + 0c_{14}(A) \\ &= 3\begin{vmatrix} 1 & 2 & 0 \\ 6 & 0 & -1 \\ 3 & 1 & 0 \end{vmatrix} \end{aligned}$$

This is the first stage of the calculation, and we have succeeded in expressing the determinant of the 4×4 matrix A in terms of the determinant of a 3×3 matrix. The next stage involves this 3×3 matrix. Again, we can use any row or column for the cofactor expansion. The third column is preferred (with two zeros), so

$$\begin{aligned}
\det A &= 3\left(0\begin{vmatrix} 6 & 0 \\ 3 & 1 \end{vmatrix} - (-1)\begin{vmatrix} 1 & 2 \\ 3 & 1 \end{vmatrix} + 0\begin{vmatrix} 1 & 2 \\ 6 & 0 \end{vmatrix}\right) \\
&= 3[0 + 1(-5) + 0] \\
&= -15
\end{aligned}$$

This completes the calculation.

Computing the determinant of a matrix A can be tedious. For example, if A is a 4×4 matrix, the cofactor expansion along any row or column involves calculating four cofactors, each of which involves the determinant of a 3×3 matrix. And if A is 5×5, the expansion involves five determinants of 4×4 matrices! There is a clear need for some techniques to cut down the work.[3]

The motivation for the method is the observation (see Example 3.1.4) that calculating a determinant is simplified a great deal when a row or column consists mostly of zeros. (In fact, when a row or column consists *entirely* of zeros, the determinant is zero—simply expand along that row or column.)

Recall next that one method of *creating* zeros in a matrix is to apply elementary row operations to it. Hence, a natural question to ask is what effect such a row operation has on the determinant of the matrix. It turns out that the effect is easy to determine and that elementary *column* operations can be used in the same way. These observations lead to a technique for evaluating determinants that greatly reduces the labour involved. The necessary information is given in Theorem 3.1.2.

Theorem 3.1.2

Let A denote an $n \times n$ matrix.

1. *If A has a row or column of zeros,* $\det A = 0$.
2. *If two distinct rows (or columns) of A are interchanged, the determinant of the resulting matrix is* $-\det A$.
3. *If a row (or column) of A is multiplied by a constant u, the determinant of the resulting matrix is* $u(\det A)$.
4. *If two distinct rows (or columns) of A are identical,* $\det A = 0$.
5. *If a multiple of one row of A is added to a different row (or if a multiple of a column is added to a different column), the determinant of the resulting matrix is* $\det A$.

[3] If $A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$ we can calculate $\det A$ by considering $\begin{bmatrix} a & b & c & a & b \\ d & e & f & d & e \\ g & h & i & g & h \end{bmatrix}$ obtained from A by adjoining columns 1 and 2 on the right. Then $\det A = aei + bfg + cdh - ceg - afh - bdi$, where the positive terms aei, bfg, and cdh are the products down and to the right starting at a, b, and c, and the negative terms ceg, afh, and bdi are the products down and to the left starting at c, a, and b. **Warning**: This rule does **not** apply to $n \times n$ matrices where $n > 3$ or $n = 2$.

Proof. We prove properties 2, 4, and 5 and leave the rest as exercises.

Property 2. If A is $n \times n$, this follows by induction on n. If $n = 2$, the verification is left to the reader. If $n > 2$ and two rows are interchanged, let B denote the resulting matrix. Expand $\det A$ and $\det B$ along a row *other than* the two that were interchanged. The entries in this row are the same for both A and B, but the cofactors in B are the negatives of those in A (by induction) because the corresponding $(n-1) \times (n-1)$ matrices have two rows interchanged. Hence, $\det B = -\det A$, as required. A similar argument works if two columns are interchanged.

Property 4. If two rows of A are equal, let B be the matrix obtained by interchanging them. Then $B = A$, so $\det B = detA$. But $\det B = -\det A$ by property 2, so $\det A = \det B = 0$. Again, the same argument works for columns.

Property 5. Let B be obtained from $A = \left[a_{ij}\right]$ by adding u times row p to row q. Then row q of B is

$$(a_{q1} + ua_{p1},\ a_{q2} + ua_{p2},\ \dots,\ a_{qn} + ua_{pn})$$

The cofactors of these elements in B are the same as in A (they do not involve row q): in symbols, $c_{qj}(B) = c_{qj}(A)$ for each j. Hence, expanding B along row q gives

$$\begin{aligned}
\det B &= (a_{q1} + ua_{p1})c_{q1}(A) + (a_{q2} + ua_{p2})c_{q2}(A) + \cdots + (a_{qn} + ua_{pn})c_{qn}(A) \\
&= [a_{q1}c_{q1}(A) + a_{q2}c_{q2}(A) + \cdots + a_{qn}c_{qn}(A)] + u[a_{p1}c_{q1}(A) + a_{p2}c_{q2}(A) + \cdots + a_{pn}c_{qn}(A)] \\
&= \det A + u \det C
\end{aligned}$$

where C is the matrix obtained from A by replacing row q by row p (and both expansions are along row q). Because rows p and q of C are equal, $\det C = 0$ by property 4. Hence, $\det B = \det A$, as required. As before, a similar proof holds for columns. □

To illustrate Theorem 3.1.2, consider the following determinants.

$$\left|\begin{array}{rrr} 3 & -1 & 2 \\ 2 & 5 & 1 \\ 0 & 0 & 0 \end{array}\right| = 0 \qquad \text{(because the last row consists of zeros)}$$

$$\left|\begin{array}{rrr} 3 & -1 & 5 \\ 2 & 8 & 7 \\ 1 & 2 & -1 \end{array}\right| = -\left|\begin{array}{rrr} 5 & -1 & 3 \\ 7 & 8 & 2 \\ -1 & 2 & 1 \end{array}\right| \qquad \text{(because two columns are interchanged)}$$

$$\left|\begin{array}{rrr} 8 & 1 & 2 \\ 3 & 0 & 9 \\ 1 & 2 & -1 \end{array}\right| = 3\left|\begin{array}{rrr} 8 & 1 & 2 \\ 1 & 0 & 3 \\ 1 & 2 & -1 \end{array}\right| \qquad \text{(because the second row of the matrix on the left is 3 times the second row of the matrix on the right)}$$

$$\left|\begin{array}{rrr} 2 & 1 & 2 \\ 4 & 0 & 4 \\ 1 & 3 & 1 \end{array}\right| = 0 \qquad \text{(because two columns are identical)}$$

$$\left|\begin{array}{rrr} 2 & 5 & 2 \\ -1 & 2 & 9 \\ 3 & 1 & 1 \end{array}\right| = \left|\begin{array}{rrr} 0 & 9 & 20 \\ -1 & 2 & 9 \\ 3 & 1 & 1 \end{array}\right| \qquad \text{(because twice the second row of the matrix on the left was added to the first row)}$$

The following four examples illustrate how Theorem 3.1.2 is used to evaluate determinants.

Example 3.1.5

Evaluate $\det A$ when $A = \begin{bmatrix} 1 & -1 & 3 \\ 1 & 0 & -1 \\ 2 & 1 & 6 \end{bmatrix}$.

Solution. The matrix does have zero entries, so expansion along (say) the second row would involve somewhat less work. However, a column operation can be used to get a zero in position (2, 3)—namely, add column 1 to column 3. Because this does not change the value of the determinant, we obtain

$$\det A = \begin{vmatrix} 1 & -1 & 3 \\ 1 & 0 & -1 \\ 2 & 1 & 6 \end{vmatrix} = \begin{vmatrix} 1 & -1 & 4 \\ 1 & 0 & 0 \\ 2 & 1 & 8 \end{vmatrix} = - \begin{vmatrix} -1 & 4 \\ 1 & 8 \end{vmatrix} = 12$$

where we expanded the second 3×3 matrix along row 2.

Example 3.1.6

If $\det \begin{bmatrix} a & b & c \\ p & q & r \\ x & y & z \end{bmatrix} = 6$, evaluate $\det A$ where $A = \begin{bmatrix} a+x & b+y & c+z \\ 3x & 3y & 3z \\ -p & -q & -r \end{bmatrix}$.

Solution. First take common factors out of rows 2 and 3.

$$\det A = 3(-1) \det \begin{bmatrix} a+x & b+y & c+z \\ x & y & z \\ p & q & r \end{bmatrix}$$

Now subtract the second row from the first and interchange the last two rows.

$$\det A = -3 \det \begin{bmatrix} a & b & c \\ x & y & z \\ p & q & r \end{bmatrix} = 3 \det \begin{bmatrix} a & b & c \\ p & q & r \\ x & y & z \end{bmatrix} = 3 \cdot 6 = 18$$

The determinant of a matrix is a sum of products of its entries. In particular, if these entries are polynomials in x, then the determinant itself is a polynomial in x. It is often of interest to determine which values of x make the determinant zero, so it is very useful if the determinant is given in factored form. Theorem 3.1.2 can help.

Example 3.1.7

Find the values of x for which $\det A = 0$, where $A = \begin{bmatrix} 1 & x & x \\ x & 1 & x \\ x & x & 1 \end{bmatrix}$.

Solution. To evaluate $\det A$, first subtract x times row 1 from rows 2 and 3.

$$\det A = \left| \begin{array}{ccc} 1 & x & x \\ x & 1 & x \\ x & x & 1 \end{array} \right| = \left| \begin{array}{ccc} 1 & x & x \\ 0 & 1-x^2 & x-x^2 \\ 0 & x-x^2 & 1-x^2 \end{array} \right| = \left| \begin{array}{cc} 1-x^2 & x-x^2 \\ x-x^2 & 1-x^2 \end{array} \right|$$

At this stage we could simply evaluate the determinant (the result is $2x^3 - 3x^2 + 1$). But then we would have to factor this polynomial to find the values of x that make it zero. However, this factorization can be obtained directly by first factoring each entry in the determinant and taking a common factor of $(1-x)$ from each row.

$$\begin{aligned} \det A = \left| \begin{array}{cc} (1-x)(1+x) & x(1-x) \\ x(1-x) & (1-x)(1+x) \end{array} \right| &= (1-x)^2 \left| \begin{array}{cc} 1+x & x \\ x & 1+x \end{array} \right| \\ &= (1-x)^2(2x+1) \end{aligned}$$

Hence, $\det A = 0$ means $(1-x)^2(2x+1) = 0$, that is $x = 1$ or $x = -\frac{1}{2}$.

Example 3.1.8

If a_1, a_2, and a_3 are given show that

$$\det \left[\begin{array}{ccc} 1 & a_1 & a_1^2 \\ 1 & a_2 & a_2^2 \\ 1 & a_3 & a_3^2 \end{array} \right] = (a_3 - a_1)(a_3 - a_2)(a_2 - a_1)$$

Solution. Begin by subtracting row 1 from rows 2 and 3, and then expand along column 1:

$$\det \left[\begin{array}{ccc} 1 & a_1 & a_1^2 \\ 1 & a_2 & a_2^2 \\ 1 & a_3 & a_3^2 \end{array} \right] = \det \left[\begin{array}{ccc} 1 & a_1 & a_1^2 \\ 0 & a_2 - a_1 & a_2^2 - a_1^2 \\ 0 & a_3 - a_1 & a_3^2 - a_1^2 \end{array} \right] = \det \left[\begin{array}{cc} a_2 - a_1 & a_2^2 - a_1^2 \\ a_3 - a_1 & a_3^2 - a_1^2 \end{array} \right]$$

Now $(a_2 - a_1)$ and $(a_3 - a_1)$ are common factors in rows 1 and 2, respectively, so

$$\begin{aligned} \det \left[\begin{array}{ccc} 1 & a_1 & a_1^2 \\ 1 & a_2 & a_2^2 \\ 1 & a_3 & a_3^2 \end{array} \right] &= (a_2 - a_1)(a_3 - a_1) \det \left[\begin{array}{cc} 1 & a_2 + a_1 \\ 1 & a_3 + a_1 \end{array} \right] \\ &= (a_2 - a_1)(a_3 - a_1)(a_3 - a_2) \end{aligned}$$

The matrix in Example 3.1.8 is called a Vandermonde matrix, and the formula for its determinant can be generalized to the $n \times n$ case (see Theorem 3.2.7).

If A is an $n \times n$ matrix, forming uA means multiplying *every* row of A by u. Applying property 3 of Theorem 3.1.2, we can take the common factor u out of each row and so obtain the following useful result.

Theorem 3.1.3

If A is an $n \times n$ matrix, then $\det(uA) = u^n \det A$ for any number u.

The next example displays a type of matrix whose determinant is easy to compute.

Example 3.1.9

Evaluate $\det A$ if $A = \begin{bmatrix} a & 0 & 0 & 0 \\ u & b & 0 & 0 \\ v & w & c & 0 \\ x & y & z & d \end{bmatrix}$.

Solution. Expand along row 1 to get $\det A = a \begin{vmatrix} b & 0 & 0 \\ w & c & 0 \\ y & z & d \end{vmatrix}$. Now expand this along the top row to get $\det A = ab \begin{vmatrix} c & 0 \\ z & d \end{vmatrix} = abcd$, the product of the main diagonal entries.

A square matrix is called a **lower triangular matrix** if all entries above the main diagonal are zero (as in Example 3.1.9). Similarly, an **upper triangular matrix** is one for which all entries below the main diagonal are zero. A **triangular matrix** is one that is either upper or lower triangular. Theorem 3.1.4 gives an easy rule for calculating the determinant of any triangular matrix. The proof is like the solution to Example 3.1.9.

Theorem 3.1.4

If A is a square triangular matrix, then $\det A$ is the product of the entries on the main diagonal.

Theorem 3.1.4 is useful in computer calculations because it is a routine matter to carry a matrix to triangular form using row operations.

Block matrices such as those in the next theorem arise frequently in practice, and the theorem gives an easy method for computing their determinants. This dovetails with Example 2.4.11.

Theorem 3.1.5

Consider matrices $\begin{bmatrix} A & X \\ 0 & B \end{bmatrix}$ and $\begin{bmatrix} A & 0 \\ Y & B \end{bmatrix}$ in block form, where A and B are square matrices. Then

$$\det \begin{bmatrix} A & X \\ 0 & B \end{bmatrix} = \det A \det B \text{ and } \det \begin{bmatrix} A & 0 \\ Y & B \end{bmatrix} = \det A \det B$$

Proof. Write $T = \det \begin{bmatrix} A & X \\ 0 & B \end{bmatrix}$ and proceed by induction on k where A is $k \times k$. If $k = 1$, it is the cofactor expansion along column 1. In general let $S_i(T)$ denote the matrix obtained from T by deleting row i and column 1. Then the cofactor expansion of $\det T$ along the first column is

$$\det T = a_{11} \det(S_1(T)) - a_{21} \det(S_2(T)) + \cdots \pm a_{k1} \det(S_k(T)) \tag{3.2}$$

where a_{11}, a_{21}, $\cdots$, a_{k1} are the entries in the first column of A. But $S_i(T) = \left[\begin{array}{cc} S_i(A) & X_i \\ 0 & B \end{array} \right]$ for each $i = 1,\ 2,\ \cdots,\ k$, so $\det(S_i(T)) = \det(S_i(A)) \cdot \det B$ by induction. Hence, Equation 3.2 becomes

$$\begin{aligned} \det T &= \{a_{11} \det(S_1(T)) - a_{21} \det(S_2(T)) + \cdots \pm a_{k1} \det(S_k(T))\} \det B \\ &= \{\det A\} \det B \end{aligned}$$

as required. The lower triangular case is similar. □

Example 3.1.10

$$\det \left[\begin{array}{rrrr} 2 & 3 & 1 & 3 \\ 1 & -2 & -1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 4 & 0 & 1 \end{array} \right] = - \left| \begin{array}{rrrr} 2 & 1 & 3 & 3 \\ 1 & -1 & -2 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 4 & 1 \end{array} \right| = - \left| \begin{array}{rr} 2 & 1 \\ 1 & -1 \end{array} \right| \left| \begin{array}{rr} 1 & 1 \\ 4 & 1 \end{array} \right| = -(-3)(-3) = -9$$

The next result shows that $\det A$ is a linear transformation when regarded as a function of a fixed column of A. The proof is Exercise **??**.

Theorem 3.1.6

Given columns $\mathbf{c}_1,\ \cdots,\ \mathbf{c}_{j-1},\ \mathbf{c}_{j+1},\ \cdots,\ \mathbf{c}_n$ in $\mathbb{R}^n$, define $T : \mathbb{R}^n \to \mathbb{R}$ by

$$T(\mathbf{x}) = \det \left[\begin{array}{ccccccc} \mathbf{c}_1 & \cdots & \mathbf{c}_{j-1} & \mathbf{x} & \mathbf{c}_{j+1} & \cdots & \mathbf{c}_n \end{array} \right] \text{ for all } \mathbf{x} \text{ in } \mathbb{R}^n$$

Then, for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$ and all a in $\mathbb{R}$,

$$T(\mathbf{x} + \mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y}) \quad \text{and} \quad T(a\mathbf{x}) = aT(\mathbf{x})$$

3.2 Determinants and Matrix Inverses

In this section, several theorems about determinants are derived. One consequence of these theorems is that a square matrix A is invertible if and only if $\det A \neq 0$. Moreover, determinants are used to give a formula for A^{-1} which, in turn, yields a formula (called Cramer's rule) for the solution of any system of linear equations with an invertible coefficient matrix.

We begin with a remarkable theorem (due to Cauchy in 1812) about the determinant of a product of matrices. The proof is given at the end of this section.

Theorem 3.2.1: Product Theorem

If A and B are $n \times n$ matrices, then $\det(AB) = \det A \det B$.

The complexity of matrix multiplication makes the product theorem quite unexpected. Here is an example where it reveals an important numerical identity.

Example 3.2.1

If $A = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}$ and $B = \begin{bmatrix} c & d \\ -d & c \end{bmatrix}$ then $AB = \begin{bmatrix} ac-bd & ad+bc \\ -(ad+bc) & ac-bd \end{bmatrix}$.

Hence $\det A \det B = \det(AB)$ gives the identity

$$(a^2+b^2)(c^2+d^2) = (ac-bd)^2 + (ad+bc)^2$$

Theorem 3.2.1 extends easily to $\det(ABC) = \det A \det B \det C$. In fact, induction gives

$$\det(A_1A_2\cdots A_{k-1}A_k) = \det A_1 \det A_2 \cdots \det A_{k-1} \det A_k$$

for any square matrices $A_1, \dots, A_k$ of the same size. In particular, if each $A_i = A$, we obtain

$$\det(A^k) = (detA)^k, \quad \text{for any } k \geq 1$$

We can now give the invertibility condition.

Theorem 3.2.2

An $n \times n$ matrix A is invertible if and only if $\det A \neq 0$. When this is the case, $\det(A^{-1}) = \frac{1}{\det A}$

Proof. If A is invertible, then $AA^{-1} = I$; so the product theorem gives

$$1 = \det I = \det(AA^{-1}) = \det A \det A^{-1}$$

Hence, $\det A \neq 0$ and also $\det A^{-1} = \frac{1}{\det A}$.

Conversely, if $\det A \neq 0$, we show that A can be carried to I by elementary row operations (and invoke Theorem 2.4.5). Certainly, A can be carried to its reduced row-echelon form R, so $R = E_k \cdots E_2E_1A$ where the E_i are elementary matrices (Theorem 2.5.1). Hence the product theorem gives

$$\det R = \det E_k \cdots \det E_2 \det E_1 \det A$$

Since $\det E \neq 0$ for all elementary matrices E, this shows $\det R \neq 0$. In particular, R has no row of zeros, so $R = I$ because R is square and reduced row-echelon. This is what we wanted. □

Example 3.2.2

For which values of c does $A = \left[\begin{array}{rrr} 1 & 0 & -c \\ -1 & 3 & 1 \\ 0 & 2c & -4 \end{array}\right]$ have an inverse?

Solution. Compute $\det A$ by first adding c times column 1 to column 3 and then expanding along row 1.

$$\det A = \det \left[\begin{array}{rrr} 1 & 0 & -c \\ -1 & 3 & 1 \\ 0 & 2c & -4 \end{array}\right] = \det \left[\begin{array}{rrc} 1 & 0 & 0 \\ -1 & 3 & 1-c \\ 0 & 2c & -4 \end{array}\right] = 2(c+2)(c-3)$$

Hence, $\det A = 0$ if $c = -2$ or $c = 3$, and A has an inverse if $c \neq -2$ and $c \neq 3$.

Example 3.2.3

If a product $A_1A_2\cdots A_k$ of square matrices is invertible, show that each A_i is invertible.

Solution. We have $\det A_1 \det A_2 \cdots \det A_k = \det(A_1A_2\cdots A_k)$ by the product theorem, and $\det(A_1A_2\cdots A_k) \neq 0$ by Theorem 3.2.2 because $A_1A_2\cdots A_k$ is invertible. Hence

$$\det A_1 \det A_2 \cdots \det A_k \neq 0$$

so $\det A_i \neq 0$ for each i. This shows that each A_i is invertible, again by Theorem 3.2.2.

Theorem 3.2.3

If A is any square matrix, $\det A^T = \det A$.

Proof. Consider first the case of an elementary matrix E. If E is of type I or II, then $E^T = E$; so certainly $\det E^T = \det E$. If E is of type III, then E^T is also of type III; so $\det E^T = 1 = \det E$ by Theorem 3.1.2. Hence, $\det E^T = \det E$ for every elementary matrix E.

Now let A be any square matrix. If A is not invertible, then neither is A^T; so $\det A^T = 0 = \det A$ by Theorem 3.2.2. On the other hand, if A is invertible, then $A = E_k\cdots E_2E_1$, where the E_i are elementary matrices (Theorem 2.5.2). Hence, $A^T = E_1^TE_2^T\cdots E_k^T$ so the product theorem gives

$$\begin{aligned}\det A^T = \det E_1^T \det E_2^T \cdots \det E_k^T &= \det E_1 \det E_2 \cdots \det E_k \\ &= \det E_k \cdots \det E_2 \det E_1 \\ &= \det A\end{aligned}$$

This completes the proof. □

Example 3.2.4

If $\det A = 2$ and $\det B = 5$, calculate $\det(A^3B^{-1}A^TB^2)$.

Solution. We use several of the facts just derived.

$$\begin{aligned}\det(A^3B^{-1}A^TB^2) &= \det(A^3)\det(B^{-1})\det(A^T)\det(B^2) \\ &= (\det A)^3 \tfrac{1}{\det B} \det A (\det B)^2 \\ &= 2^3 \cdot \tfrac{1}{5} \cdot 2 \cdot 5^2 \\ &= 80\end{aligned}$$

Example 3.2.5

A square matrix is called **orthogonal** if $A^{-1} = A^T$. What are the possible values of $\det A$ if A is orthogonal?

Solution. If A is orthogonal, we have $I = AA^T$. Take determinants to obtain

$$1 = \det I = \det\left(AA^T\right) = \det A \det A^T = (\det A)^2$$

Since $\det A$ is a number, this means $\det A = \pm 1$.

Hence Theorems 2.6.4 and 2.6.5 imply that rotation about the origin and reflection about a line through the origin in $\mathbb{R}^2$ have orthogonal matrices with determinants 1 and -1 respectively. In fact they are the *only* such transformations of $\mathbb{R}^2$. We have more to say about this in Section 8.2.

Adjugates

In Section 2.4 we defined the adjugate of a 2×2 matrix $A = \left[\begin{array}{cc} a & b \\ c & d \end{array}\right]$ to be $\operatorname{adj}(A) = \left[\begin{array}{rr} d & -b \\ -c & a \end{array}\right]$. Then we verified that $A(\operatorname{adj} A) = (\det A)I = (\operatorname{adj} A)A$ and hence that, if $\det A \neq 0$, $A^{-1} = \frac{1}{\det A} \operatorname{adj} A$. We are now able to define the adjugate of an arbitrary square matrix and to show that this formula for the inverse remains valid (when the inverse exists).

Recall that the (i, j)-cofactor $c_{ij}(A)$ of a square matrix A is a number defined for each position (i, j) in the matrix. If A is a square matrix, the **cofactor matrix of** A is defined to be the matrix $\left[c_{ij}(A)\right]$ whose (i, j)-entry is the (i, j)-cofactor of A.

Definition 3.3 Adjugate of a Matrix

The ***adjugate***[4] *of* A, *denoted* $\operatorname{adj}(A)$, *is the transpose of this cofactor matrix; in symbols,*

$$\operatorname{adj}(A) = \left[c_{ij}(A)\right]^T$$

This agrees with the earlier definition for a 2×2 matrix A as the reader can verify.

Example 3.2.6

Compute the adjugate of $A = \left[\begin{array}{rrr} 1 & 3 & -2 \\ 0 & 1 & 5 \\ -2 & -6 & 7 \end{array}\right]$ and calculate $A(\operatorname{adj} A)$ and $(\operatorname{adj} A)A$.

Solution. We first find the cofactor matrix.

[4]This is also called the classical adjoint of A, but the term "adjoint" has another meaning.

$$\begin{bmatrix} c_{11}(A) & c_{12}(A) & c_{13}(A) \\ c_{21}(A) & c_{22}(A) & c_{23}(A) \\ c_{31}(A) & c_{32}(A) & c_{33}(A) \end{bmatrix} = \begin{bmatrix} \begin{vmatrix} 1 & 5 \\ -6 & 7 \end{vmatrix} & -\begin{vmatrix} 0 & 5 \\ -2 & 7 \end{vmatrix} & \begin{vmatrix} 0 & 1 \\ -2 & -6 \end{vmatrix} \\ -\begin{vmatrix} 3 & -2 \\ -6 & 7 \end{vmatrix} & \begin{vmatrix} 1 & -2 \\ -2 & 7 \end{vmatrix} & -\begin{vmatrix} 1 & 3 \\ -2 & -6 \end{vmatrix} \\ \begin{vmatrix} 3 & -2 \\ 1 & 5 \end{vmatrix} & -\begin{vmatrix} 1 & -2 \\ 0 & 5 \end{vmatrix} & \begin{vmatrix} 1 & 3 \\ 0 & 1 \end{vmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} 37 & -10 & 2 \\ -9 & 3 & 0 \\ 17 & -5 & 1 \end{bmatrix}$$

Then the adjugate of A is the transpose of this cofactor matrix.

$$\text{adj } A = \begin{bmatrix} 37 & -10 & 2 \\ -9 & 3 & 0 \\ 17 & -5 & 1 \end{bmatrix}^T = \begin{bmatrix} 37 & -9 & 17 \\ -10 & 3 & -5 \\ 2 & 0 & 1 \end{bmatrix}$$

The computation of $A(\text{adj } A)$ gives

$$A(\text{adj } A) = \begin{bmatrix} 1 & 3 & -2 \\ 0 & 1 & 5 \\ -2 & -6 & 7 \end{bmatrix} \begin{bmatrix} 37 & -9 & 17 \\ -10 & 3 & -5 \\ 2 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix} = 3I$$

and the reader can verify that also $(\text{adj } A)A = 3I$. Hence, analogy with the 2×2 case would indicate that $\det A = 3$; this is, in fact, the case.

The relationship $A(\text{adj } A) = (\det A)I$ holds for any square matrix A. To see why this is so, consider the general 3×3 case. Writing $c_{ij}(A) = c_{ij}$ for short, we have

$$\text{adj } A = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}^T = \begin{bmatrix} c_{11} & c_{21} & c_{31} \\ c_{12} & c_{22} & c_{32} \\ c_{13} & c_{23} & c_{33} \end{bmatrix}$$

If $A = \left[a_{ij}\right]$ in the usual notation, we are to verify that $A(\text{adj } A) = (\det A)I$. That is,

$$A(\text{adj } A) = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} c_{11} & c_{21} & c_{31} \\ c_{12} & c_{22} & c_{32} \\ c_{13} & c_{23} & c_{33} \end{bmatrix} = \begin{bmatrix} \det A & 0 & 0 \\ 0 & \det A & 0 \\ 0 & 0 & \det A \end{bmatrix}$$

Consider the $(1,\ 1)$-entry in the product. It is given by $a_{11}c_{11} + a_{12}c_{12} + a_{13}c_{13}$, and this is just the cofactor expansion of $\det A$ along the first row of A. Similarly, the $(2,\ 2)$-entry and the $(3,\ 3)$-entry are the cofactor expansions of $\det A$ along rows 2 and 3, respectively.

So it remains to be seen why the off-diagonal elements in the matrix product $A(\text{adj } A)$ are all zero. Consider the $(1,\ 2)$-entry of the product. It is given by $a_{11}c_{21} + a_{12}c_{22} + a_{13}c_{23}$. This *looks* like the cofactor expansion of the determinant of *some* matrix. To see which, observe that c_{21}, c_{22}, and c_{23} are all computed by *deleting* row 2 of A (and one of the columns), so they remain the same if row 2 of A is changed. In particular, if row 2 of A is replaced by row 1, we obtain

$$a_{11}c_{21}+a_{12}c_{22}+a_{13}c_{23} = \det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = 0$$

where the expansion is along row 2 and where the determinant is zero because two rows are identical. A similar argument shows that the other off-diagonal entries are zero.

This argument works in general and yields the first part of Theorem 3.2.4. The second assertion follows from the first by multiplying through by the scalar $\frac{1}{\det A}$.

Theorem 3.2.4: Adjugate Formula

If A is any square matrix, then

$$A(\operatorname{adj} A) = (\det A)I = (\operatorname{adj} A)A$$

In particular, if $\det A \neq 0$, the inverse of A is given by

$$A^{-1} = \frac{1}{\det A} \operatorname{adj} A$$

It is important to note that this theorem is *not* an efficient way to find the inverse of the matrix A. For example, if A were 10×10, the calculation of $\operatorname{adj} A$ would require computing $10^2 = 100$ determinants of 9×9 matrices! On the other hand, the matrix inversion algorithm would find A^{-1} with about the same effort as finding $\det A$. Clearly, Theorem 3.2.4 is not a *practical* result: its virtue is that it gives a formula for A^{-1} that is useful for *theoretical* purposes.

Example 3.2.7

Find the $(2,\ 3)$-entry of A^{-1} if $A = \begin{bmatrix} 2 & 1 & 3 \\ 5 & -7 & 1 \\ 3 & 0 & -6 \end{bmatrix}$.

Solution. First compute

$$\det A = \begin{vmatrix} 2 & 1 & 3 \\ 5 & -7 & 1 \\ 3 & 0 & -6 \end{vmatrix} = \begin{vmatrix} 2 & 1 & 7 \\ 5 & -7 & 11 \\ 3 & 0 & 0 \end{vmatrix} = 3\begin{vmatrix} 1 & 7 \\ -7 & 11 \end{vmatrix} = 180$$

Since $A^{-1} = \frac{1}{\det A} \operatorname{adj} A = \frac{1}{180}\left[c_{ij}(A)\right]^T$, the $(2,\ 3)$-entry of A^{-1} is the $(3,\ 2)$-entry of the matrix $\frac{1}{180}\left[c_{ij}(A)\right]$; that is, it equals $\frac{1}{180}c_{32}(A) = \frac{1}{180}\left(-\begin{vmatrix} 2 & 3 \\ 5 & 1 \end{vmatrix}\right) = \frac{13}{180}$.

Example 3.2.8

If A is $n \times n$, $n \geq 2$, show that $\det(\text{adj } A) = (\det A)^{n-1}$.

Solution. Write $d = \det A$; we must show that $\det(\text{adj } A) = d^{n-1}$. We have $A(\text{adj } A) = dI$ by Theorem 3.2.4, so taking determinants gives $d \det(\text{adj } A) = d^n$. Hence we are done if $d \neq 0$. Assume $d = 0$; we must show that $\det(\text{adj } A) = 0$, that is, $\text{adj } A$ is not invertible. If $A \neq 0$, this follows from $A(\text{adj } A) = dI = 0$; if $A = 0$, it follows because then $\text{adj } A = 0$.

Cramer's Rule

Theorem 3.2.4 has a nice application to linear equations. Suppose

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

is a system of n equations in n variables $x_1, x_2, \dots, x_n$. Here A is the $n \times n$ coefficient matrix, and $\mathbf{x}$ and $\mathbf{b}$ are the columns

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

of variables and constants, respectively. If $\det A \neq 0$, we left multiply by A^{-1} to obtain the solution $\mathbf{x} = A^{-1}\mathbf{b}$. When we use the adjugate formula, this becomes

$$\begin{aligned} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} &= \tfrac{1}{\det A}(\text{adj } A)\mathbf{b} \\ &= \tfrac{1}{\det A} \begin{bmatrix} c_{11}(A) & c_{21}(A) & \cdots & c_{n1}(A) \\ c_{12}(A) & c_{22}(A) & \cdots & c_{n2}(A) \\ \vdots & \vdots & & \vdots \\ c_{1n}(A) & c_{2n}(A) & \cdots & c_{nn}(A) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \end{aligned}$$

Hence, the variables $x_1, x_2, \dots, x_n$ are given by

$$\begin{aligned} x_1 &= \tfrac{1}{\det A}\left[b_1c_{11}(A) + b_2c_{21}(A) + \cdots + b_nc_{n1}(A)\right] \\ x_2 &= \tfrac{1}{\det A}\left[b_1c_{12}(A) + b_2c_{22}(A) + \cdots + b_nc_{n2}(A)\right] \\ &\vdots \qquad\qquad \vdots \\ x_n &= \tfrac{1}{\det A}\left[b_1c_{1n}(A) + b_2c_{2n}(A) + \cdots + b_nc_{nn}(A)\right] \end{aligned}$$

Now the quantity $b_1c_{11}(A) + b_2c_{21}(A) + \cdots + b_nc_{n1}(A)$ occurring in the formula for x_1 looks like the cofactor expansion of the determinant of a matrix. The cofactors involved are $c_{11}(A), c_{21}(A), \dots, c_{n1}(A)$, corresponding to the first column of A. If A_1 is obtained from A by replacing the first column of A by $\mathbf{b}$, then $c_{i1}(A_1) = c_{i1}(A)$ for each i because column 1 is deleted when computing them. Hence, expanding $\det(A_1)$ by the first column gives

$$\begin{aligned}\det A_1 &= b_1c_{11}(A_1) + b_2c_{21}(A_1) + \cdots + b_nc_{n1}(A_1)\\ &= b_1c_{11}(A) + b_2c_{21}(A) + \cdots + b_nc_{n1}(A)\\ &= (\det A)x_1\end{aligned}$$

Hence, $x_1 = \frac{\det A_1}{\det A}$ and similar results hold for the other variables.

Theorem 3.2.5: Cramer's Rule[5]

If A is an invertible $n \times n$ matrix, the solution to the system

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

of n equations in the variables $x_1, x_2, \ldots, x_n$ is given by

$$x_1 = \frac{\det A_1}{\det A}, \quad x_2 = \frac{\det A_2}{\det A}, \quad \cdots, \quad x_n = \frac{\det A_n}{\det A}$$

where, for each k, A_k is the matrix obtained from A by replacing column k by $\mathbf{b}$.

Example 3.2.9

Find x_1, given the following system of equations.

$$\begin{aligned}5x_1 + x_2 - x_3 &= 4\\ 9x_1 + x_2 - x_3 &= 1\\ x_1 - x_2 + 5x_3 &= 2\end{aligned}$$

Solution. Compute the determinants of the coefficient matrix A and the matrix A_1 obtained from it by replacing the first column by the column of constants.

$$\det A = \det \begin{bmatrix} 5 & 1 & -1 \\ 9 & 1 & -1 \\ 1 & -1 & 5 \end{bmatrix} = -16$$

$$\det A_1 = \det \begin{bmatrix} 4 & 1 & -1 \\ 1 & 1 & -1 \\ 2 & -1 & 5 \end{bmatrix} = 12$$

Hence, $x_1 = \frac{\det A_1}{\det A} = -\frac{3}{4}$ by Cramer's rule.

Cramer's rule is *not* an efficient way to solve linear systems or invert matrices. True, it enabled us to calculate x_1 here without computing x_2 or x_3. Although this might seem an advantage, the truth of the

[5]Gabriel Cramer (1704–1752) was a Swiss mathematician who wrote an introductory work on algebraic curves. He popularized the rule that bears his name, but the idea was known earlier.

matter is that, for large systems of equations, the number of computations needed to find *all* the variables by the gaussian algorithm is comparable to the number required to find *one* of the determinants involved in Cramer's rule. Furthermore, the algorithm works when the matrix of the system is not invertible and even when the coefficient matrix is not square. Like the adjugate formula, then, Cramer's rule is *not* a practical numerical technique; its virtue is theoretical.

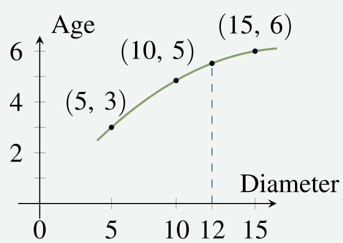
Polynomial Interpolation

Given a set of data, it is often the case that one is interested to understand a trend so to forecast other values. One such method is that of modeling the trend with a polynomial, here is an example.

Example 3.2.10

A forester wants to estimate the age (in years) of a tree by measuring the diameter of the trunk (in cm). She obtains the following data:

	Tree 1	Tree 2	Tree 3
Trunk Diameter	5	10	15
Age	3	5	6



Use this
date to estimate the age of a tree with a trunk diameter of 12 cm.

Solution. The forester decides to "fit" a quadratic polynomial

$$p(x) = r_0 + r_1x + r_2x^2$$

to the data, that is choose the coefficients r_0,
r_1, and r_2 so that $p(5) = 3$, $p(10) = 5$, and $p(15) = 6$, and then use $p(12)$ as the estimate. These conditions give three linear equations:

$$\begin{aligned} r_0 + 5r_1 + 25r_2 &= 3 \\ r_0 + 10r_1 + 100r_2 &= 5 \\ r_0 + 15r_1 + 225r_2 &= 6 \end{aligned}$$

The (unique) solution is $r_0 = 0$, $r_1 = \frac{7}{10}$, and $r_2 = -\frac{1}{50}$, so

$$p(x) = \tfrac{7}{10}x - \tfrac{1}{50}x^2 = \tfrac{1}{50}x(35 - x)$$

Hence the estimate is $p(12) = 5.52$.

As in Example 3.2.10, it often happens that two variables x and y are related but the actual functional form $y = f(x)$ of the relationship is unknown. Suppose that for certain values $x_1, x_2, \dots, x_n$ of x the corresponding values $y_1, y_2, \dots, y_n$ are known (say from experimental measurements). One way to estimate the value of y corresponding to some other value a of x is to find a polynomial[6]

$$p(x) = r_0 + r_1x + r_2x^2 + \cdots + r_{n-1}x^{n-1}$$

that "fits" the data, that is $p(x_i) = y_i$ holds for each $i = 1, 2, \ldots, n$. Then the estimate for y is $p(a)$. As we will see, such a polynomial always exists if the x_i are distinct.

The conditions that $p(x_i) = y_i$ are

$$\begin{array}{ccccccccccc} r_0 & + & r_1x_1 & + & r_2x_1^2 & + & \cdots & + & r_{n-1}x_1^{n-1} & = & y_1 \\ r_0 & + & r_1x_2 & + & r_2x_2^2 & + & \cdots & + & r_{n-1}x_2^{n-1} & = & y_2 \\ \vdots & & \vdots & & \vdots & & & & \vdots & & \vdots \\ r_0 & + & r_1x_n & + & r_2x_n^2 & + & \cdots & + & r_{n-1}x_n^{n-1} & = & y_n \end{array}$$

In matrix form, this is

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{bmatrix} \begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ r_{n-1} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{3.3}$$

It can be shown (see Theorem 3.2.7) that the determinant of the coefficient matrix equals the product of all terms $(x_i - x_j)$ with $i > j$ and so is nonzero (because the x_i are distinct). Hence the equations have a unique solution $r_0, r_1, \ldots, r_{n-1}$. This proves

Theorem 3.2.6

Let n data pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be given, and assume that the x_i are distinct. Then there exists a unique polynomial

$$p(x) = r_0 + r_1x + r_2x^2 + \cdots + r_{n-1}x^{n-1}$$

such that $p(x_i) = y_i$ for each $i = 1, 2, \ldots, n$.

The polynomial in Theorem 3.2.6 is called the **interpolating polynomial** for the data.

We conclude by evaluating the determinant of the coefficient matrix in Equation 3.3. If $a_1, a_2, \ldots, a_n$ are numbers, the determinant

$$\det \begin{bmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^{n-1} \\ 1 & a_2 & a_2^2 & \cdots & a_2^{n-1} \\ 1 & a_3 & a_3^2 & \cdots & a_3^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_n & a_n^2 & \cdots & a_n^{n-1} \end{bmatrix}$$

is called a **Vandermonde determinant**.[7] There is a simple formula for this determinant. If $n = 2$, it equals $(a_2 - a_1)$; if $n = 3$, it is $(a_3 - a_2)(a_3 - a_1)(a_2 - a_1)$ by Example 3.1.8. The general result is the product

$$\prod_{1 \le j < i \le n} (a_i - a_j)$$

[6]A **polynomial** is an expression of the form $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ where the a_i are numbers and x is a variable. If $a_n \neq 0$, the integer n is called the degree of the polynomial, and a_n is called the leading coefficient. See Appendix D.

[7]Alexandre Théophile Vandermonde (1735–1796) was a French mathematician who made contributions to the theory of equations.

of all factors $(a_i - a_j)$ where $1 \leq j < i \leq n$. For example, if $n = 4$, it is

$$(a_4 - a_3)(a_4 - a_2)(a_4 - a_1)(a_3 - a_2)(a_3 - a_1)(a_2 - a_1)$$

Theorem 3.2.7

Let $a_1, a_2, \dots, a_n$ be numbers where $n \geq 2$. Then the corresponding Vandermonde determinant is given by

$$\det \begin{bmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^{n-1} \\ 1 & a_2 & a_2^2 & \cdots & a_2^{n-1} \\ 1 & a_3 & a_3^2 & \cdots & a_3^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_n & a_n^2 & \cdots & a_n^{n-1} \end{bmatrix} = \prod_{1 \leq j < i \leq n} (a_i - a_j)$$

Proof. We may assume that the a_i are distinct; otherwise both sides are zero. We proceed by induction on $n \geq 2$; we have it for $n = 2, 3$. So assume it holds for $n - 1$. The trick is to replace a_n by a variable x, and consider the determinant

$$p(x) = \det \begin{bmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^{n-1} \\ 1 & a_2 & a_2^2 & \cdots & a_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_{n-1} & a_{n-1}^2 & \cdots & a_{n-1}^{n-1} \\ 1 & x & x^2 & \cdots & x^{n-1} \end{bmatrix}$$

Then $p(x)$ is a polynomial of degree at most $n - 1$ (expand along the last row), and $p(a_i) = 0$ for each $i = 1, 2, \dots, n - 1$ because in each case there are two identical rows in the determinant. In particular, $p(a_1) = 0$, so we have $p(x) = (x - a_1)p_1(x)$ by the factor theorem (see Appendix D). Since $a_2 \neq a_1$, we obtain $p_1(a_2) = 0$, and so $p_1(x) = (x - a_2)p_2(x)$. Thus $p(x) = (x - a_1)(x - a_2)p_2(x)$. As the a_i are distinct, this process continues to obtain

$$p(x) = (x - a_1)(x - a_2) \cdots (x - a_{n-1})d \tag{3.4}$$

where d is the coefficient of x^{n-1} in $p(x)$. By the cofactor expansion of $p(x)$ along the last row we get

$$d = (-1)^{n+n} \det \begin{bmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^{n-2} \\ 1 & a_2 & a_2^2 & \cdots & a_2^{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_{n-1} & a_{n-1}^2 & \cdots & a_{n-1}^{n-2} \end{bmatrix}$$

Because $(-1)^{n+n} = 1$ the induction hypothesis shows that d is the product of all factors $(a_i - a_j)$ where $1 \leq j < i \leq n - 1$. The result now follows from Equation 3.4 by substituting a_n for x in $p(x)$. □

Proof of Theorem 3.2.1. If A and B are $n \times n$ matrices we must show that

$$\det(AB) = \det A \det B \tag{3.5}$$

Recall that if E is an elementary matrix obtained by doing one row operation to I_n, then doing that operation to a matrix C (Lemma 2.5.1) results in EC. By looking at the three types of elementary matrices separately, Theorem 3.1.2 shows that

$$\det(EC) = \det E \det C \quad \text{for any matrix } C \tag{3.6}$$

Thus if $E_1, E_2, \dots, E_k$ are all elementary matrices, it follows by induction that

$$\det(E_k \cdots E_2 E_1 C) = \det E_k \cdots \det E_2 \det E_1 \det C \text{ for any matrix } C \tag{3.7}$$

Lemma. If A has no inverse, then $\det A = 0$.

Proof. Let $A \to R$ where R is reduced row-echelon, say $E_n \cdots E_2 E_1 A = R$. Then R has a row of zeros by Part (4) of Theorem 2.4.5, and hence $\det R = 0$. But then Equation 3.7 gives $\det A = 0$ because $\det E \neq 0$ for any elementary matrix E. This proves the Lemma.

Now we can prove Equation 3.5 by considering two cases.

Case 1. A has no inverse. Then AB also has no inverse (otherwise $A[B(AB)^{-1}] = I$ so A is invertible by Corollary 2.4.2 to Theorem 2.4.5). Hence the above Lemma (twice) gives

$$\det(AB) = 0 = 0 \det B = \det A \det B$$

proving Equation 3.5 in this case.

Case 2. A has an inverse. Then A is a product of elementary matrices by Theorem 2.5.2, say $A = E_1 E_2 \cdots E_k$. Then Equation 3.7 with $C = I$ gives

$$\det A = \det(E_1 E_2 \cdots E_k) = \det E_1 \det E_2 \cdots \det E_k$$

But then Equation 3.7 with $C = B$ gives

$$\det(AB) = \det[(E_1 E_2 \cdots E_k)B] = \det E_1 \det E_2 \cdots \det E_k \det B = \det A \det B$$

and Equation 3.5 holds in this case too. □

3.3 Eigenvalues and Eigenvectors

The world is filled with examples of systems that evolve in time—the weather in a region, the economy of a nation, the diversity of an ecosystem, etc. Describing such systems is difficult in general and various methods have been developed in special cases. In this section we describe one such method, called *diagonalization,* which is one of the most important techniques in linear algebra. A very fertile example of this procedure is in modelling the growth of the population of an animal species. This has attracted more attention in recent years with the ever increasing awareness that many species are endangered. To motivate the technique, we begin by setting up a simple model of a bird population in which we make assumptions about survival and reproduction rates.

Example 3.3.1

Consider the evolution of the population of a species of birds. Because the number of males and females are nearly equal, we count only females. We assume that each female remains a juvenile for one year and then becomes an adult, and that only adults have offspring. We make three assumptions about reproduction and survival rates:

1. The number of juvenile females hatched in any year is twice the number of adult females alive the year before (we say the **reproduction rate** is 2).
2. Half of the adult females in any year survive to the next year (the **adult survival rate** is $\frac{1}{2}$).
3. One quarter of the juvenile females in any year survive into adulthood (the **juvenile survival rate** is $\frac{1}{4}$).

If there were 100 adult females and 40 juvenile females alive initially, compute the population of females k years later.

Solution. Let a_k and j_k denote, respectively, the number of adult and juvenile females after k years, so that the total female population is the sum $a_k + j_k$. Assumption 1 shows that $j_{k+1} = 2a_k$, while assumptions 2 and 3 show that $a_{k+1} = \frac{1}{2}a_k + \frac{1}{4}j_k$. Hence the numbers a_k and j_k in successive years are related by the following equations:

$$\begin{aligned} a_{k+1} &= \tfrac{1}{2}a_k + \tfrac{1}{4}j_k \\ j_{k+1} &= 2a_k \end{aligned}$$

If we write $\mathbf{v}_k = \left[\begin{array}{c} a_k \\ j_k \end{array} \right]$ and $A = \left[\begin{array}{cc} \frac{1}{2} & \frac{1}{4} \\ 2 & 0 \end{array} \right]$ these equations take the matrix form

$$\mathbf{v}_{k+1} = A\mathbf{v}_k, \quad \text{for each } k = 0,\ 1,\ 2,\ \dots$$

Taking $k = 0$ gives $\mathbf{v}_1 = A\mathbf{v}_0$, then taking $k = 1$ gives $\mathbf{v}_2 = A\mathbf{v}_1 = A^2\mathbf{v}_0$, and taking $k = 2$ gives $\mathbf{v}_3 = A\mathbf{v}_2 = A^3\mathbf{v}_0$. Continuing in this way, we get

$$\mathbf{v}_k = A^k\mathbf{v}_0, \quad \text{for each } k = 0,\ 1,\ 2,\ \dots$$

Since $\mathbf{v}_0 = \left[\begin{array}{c} a_0 \\ j_0 \end{array} \right] = \left[\begin{array}{c} 100 \\ 40 \end{array} \right]$ is known, finding the population profile $\mathbf{v}_k$ amounts to computing A^k for all $k \geq 0$. We will complete this calculation in Example 3.5.1 after some new techniques have been developed.

Let A be a fixed $n \times n$ matrix. A sequence $\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \dots$ of column vectors in $\mathbb{R}^n$ is called a **linear dynamical system**[8] if $\mathbf{v}_0$ is known and the other $\mathbf{v}_k$ are determined (as in Example 3.3.1) by the conditions

$$\mathbf{v}_{k+1} = A\mathbf{v}_k \text{ for each } k = 0, 1, 2, \dots$$

These conditions are called a **matrix recurrence** for the vectors $\mathbf{v}_k$. As in Example 3.3.1, they imply that

$$\mathbf{v}_k = A^k\mathbf{v}_0 \text{ for all } k \geq 0$$

so finding the columns $\mathbf{v}_k$ amounts to calculating A^k for $k \geq 0$.

Direct computation of the powers A^k of a square matrix A can be time-consuming, so we adopt an indirect method that is commonly used. The idea is to first **diagonalize** the matrix A, that is, to find an invertible matrix P such that

$$P^{-1}AP = D \text{ is a diagonal matrix} \tag{3.8}$$

This works because the powers D^k of the diagonal matrix D are easy to compute, and Equation 3.8 enables us to compute powers A^k of the matrix A in terms of powers D^k of D. Indeed, we can solve Equation 3.8 for A to get $A = PDP^{-1}$. Squaring this gives

$$A^2 = (PDP^{-1})(PDP^{-1}) = PD^2P^{-1}$$

Using this we can compute A^3 as follows:

$$A^3 = AA^2 = (PDP^{-1})(PD^2P^{-1}) = PD^3P^{-1}$$

Continuing in this way we obtain Theorem 3.3.1 (even if D is not diagonal).

Theorem 3.3.1

If $A = PDP^{-1}$ then $A^k = PD^kP^{-1}$ for each $k = 1, 2, \dots$.

Hence computing A^k comes down to finding an invertible matrix P as in equation Equation 3.8. To do this it is necessary to first compute certain numbers (called eigenvalues) associated with the matrix A.

Eigenvalues and Eigenvectors

Definition 3.4 Eigenvalues and Eigenvectors of a Matrix

*If A is an $n \times n$ matrix, a number λ is called an **eigenvalue** of A if*

$$A\mathbf{x} = \lambda\mathbf{x} \text{ for some column } \mathbf{x} \neq \mathbf{0}.$$

*In this case, $\mathbf{x}$ is called an **eigenvector** of A corresponding to the eigenvalue λ, or a λ-**eigenvector** for short.*

[8]More precisely, this is *a linear discrete* dynamical system. Many models regard $\mathbf{v}_t$ as a continuous function of the time t, and replace our condition between $\mathbf{b}_{k+1}$ and $A\mathbf{v}_k$ with a differential relationship viewed as functions of time.

Example 3.3.2

If $A = \begin{bmatrix} 3 & 5 \\ 1 & -1 \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$ then $A\mathbf{x} = 4\mathbf{x}$ so $\lambda = 4$ is an eigenvalue of A with corresponding eigenvector $\mathbf{x}$.

The matrix A in Example 3.3.2 has another eigenvalue in addition to $\lambda = 4$. To find it, we develop a general procedure for *any* $n \times n$ matrix A.

By definition a number λ is an eigenvalue of the $n \times n$ matrix A if and only if $A\mathbf{x} = \lambda\mathbf{x}$ for some column $\mathbf{x} \neq \mathbf{0}$. This is equivalent to asking that the homogeneous system

$$(\lambda I - A)\mathbf{x} = \mathbf{0}$$

of linear equations has a nontrivial solution $\mathbf{x} \neq \mathbf{0}$. By Theorem 2.4.5 this happens if and only if the matrix $\lambda I - A$ is not invertible and this, in turn, holds if and only if the determinant of the coefficient matrix is zero:

$$\det(\lambda I - A) = 0$$

This last condition prompts the following definition:

Definition 3.5 Characteristic Polynomial of a Matrix

If A is an $n \times n$ matrix, the ***characteristic polynomial*** $c_A(x)$ *of A is defined by*

$$c_A(x) = \det(xI - A)$$

Note that $c_A(x)$ is indeed a polynomial in the variable x, and it has degree n when A is an $n \times n$ matrix (this is illustrated in the examples below). The above discussion shows that a number λ is an eigenvalue of A if and only if $c_A(\lambda) = 0$, that is if and only if λ is a **root** of the characteristic polynomial $c_A(x)$. We record these observations in

Theorem 3.3.2

Let A be an $n \times n$ matrix.

1. *The eigenvalues λ of A are the roots of the characteristic polynomial $c_A(x)$ of A.*
2. *The λ-eigenvectors $\mathbf{x}$ are the nonzero solutions to the homogeneous system*

$$(\lambda I - A)\mathbf{x} = \mathbf{0}$$

of linear equations with $\lambda I - A$ as coefficient matrix.

In practice, solving the equations in part 2 of Theorem 3.3.2 is a routine application of gaussian elimination, but finding the eigenvalues can be difficult, often requiring computers (see Section 8.5). For now, the examples and exercises will be constructed so that the roots of the characteristic polynomials are relatively easy to find (usually integers). However, the reader should not be misled by this into thinking that eigenvalues are so easily obtained for the matrices that occur in practical applications!

Example 3.3.3

Find the characteristic polynomial of the matrix $A = \begin{bmatrix} 3 & 5 \\ 1 & -1 \end{bmatrix}$ discussed in Example 3.3.2, and then find all the eigenvalues and their eigenvectors.

Solution. Since $xI - A = \begin{bmatrix} x & 0 \\ 0 & x \end{bmatrix} - \begin{bmatrix} 3 & 5 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} x-3 & -5 \\ -1 & x+1 \end{bmatrix}$ we get

$$c_A(x) = \det \begin{bmatrix} x-3 & -5 \\ -1 & x+1 \end{bmatrix} = x^2 - 2x - 8 = (x-4)(x+2)$$

Hence, the roots of $c_A(x)$ are $\lambda_1 = 4$ and $\lambda_2 = -2$, so these are the eigenvalues of A. Note that $\lambda_1 = 4$ was the eigenvalue mentioned in Example 3.3.2, but we have found a new one: $\lambda_2 = -2$. To find the eigenvectors corresponding to $\lambda_2 = -2$, observe that in this case

$$(\lambda_2 I - A)\mathbf{x} = \begin{bmatrix} \lambda_2 - 3 & -5 \\ -1 & \lambda_2 + 1 \end{bmatrix} = \begin{bmatrix} -5 & -5 \\ -1 & -1 \end{bmatrix}$$

so the general solution to $(\lambda_2 I - A)\mathbf{x} = \mathbf{0}$ is $\mathbf{x} = t \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ where t is an arbitrary real number. Hence, the eigenvectors $\mathbf{x}$ corresponding to λ_2 are $\mathbf{x} = t \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ where $t \neq 0$ is arbitrary. Similarly, $\lambda_1 = 4$ gives rise to the eigenvectors $\mathbf{x} = t \begin{bmatrix} 5 \\ 1 \end{bmatrix}$, $t \neq 0$ which includes the observation in Example 3.3.2.

Note that a square matrix A has *many* eigenvectors associated with any given eigenvalue λ. In fact *every* nonzero solution $\mathbf{x}$ of $(\lambda I - A)\mathbf{x} = \mathbf{0}$ is an eigenvector. Recall that these solutions are all linear combinations of certain basic solutions determined by the gaussian algorithm (see Theorem 1.3.2). Observe that any nonzero multiple of an eigenvector is again an eigenvector,[9] and such multiples are often more convenient.[10] Any set of nonzero multiples of the basic solutions of $(\lambda I - A)\mathbf{x} = \mathbf{0}$ will be called a set of **basic eigenvectors** corresponding to λ.

Example 3.3.4

Find the characteristic polynomial, eigenvalues, and basic eigenvectors for

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

[9] In fact, any nonzero linear combination of λ-eigenvectors is again a λ-eigenvector.

[10] Allowing nonzero multiples helps eliminate round-off error when the eigenvectors involve fractions.

Solution. Here the characteristic polynomial is given by

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

so the eigenvalues are $\lambda_1 = 2$, $\lambda_2 = 1$, and $\lambda_3 = -1$. To find all eigenvectors for $\lambda_1 = 2$, compute

$$\lambda_1 I - A = \begin{bmatrix} \lambda_1 - 2 & 0 & 0 \\ -1 & \lambda_1 - 2 & 1 \\ -1 & -3 & \lambda_1 + 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 1 \\ -1 & -3 & 4 \end{bmatrix}$$

We want the (nonzero) solutions to $(\lambda_1 I - A)\mathbf{x} = \mathbf{0}$. The augmented matrix becomes

$$\left[\begin{array}{rrr|r} 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & -3 & 4 & 0 \end{array}\right] \to \left[\begin{array}{rrr|r} 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

using row operations. Hence, the general solution $\mathbf{x}$ to $(\lambda_1 I - A)\mathbf{x} = \mathbf{0}$ is $\mathbf{x} = t \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ where t is arbitrary, so we can use $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ as the basic eigenvector corresponding to $\lambda_1 = 2$. As the reader can verify, the gaussian algorithm gives basic eigenvectors $\mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$ and $\mathbf{x}_3 = \begin{bmatrix} 0 \\ \frac{1}{3} \\ 1 \end{bmatrix}$ corresponding to $\lambda_2 = 1$ and $\lambda_3 = -1$, respectively. Note that to eliminate fractions, we could instead use $3\mathbf{x}_3 = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ as the basic λ_3-eigenvector.

Example 3.3.5

If A is a square matrix, show that A and A^T have the same characteristic polynomial, and hence the same eigenvalues.

Solution. We use the fact that $xI - A^T = (xI - A)^T$. Then

$$c_{A^T}(x) = \det\left(xI - A^T\right) = \det\left[(xI - A)^T\right] = \det(xI - A) = c_A(x)$$

by Theorem 3.2.3. Hence $c_{A^T}(x)$ and $c_A(x)$ have the same roots, and so A^T and A have the same eigenvalues (by Theorem 3.3.2).

The eigenvalues of a matrix need not be distinct. For example, if $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ the characteristic polynomial is $(x-1)^2$ so the eigenvalue 1 occurs twice. Furthermore, eigenvalues are usually not computed as the roots of the characteristic polynomial. There are iterative, numerical methods (for example the QR-algorithm in Section 8.5) that are much more efficient for large matrices.

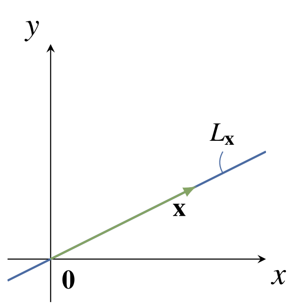
A-Invariance

If A is a 2×2 matrix, we can describe the eigenvectors of A geometrically using the following concept. A line L through the origin in $\mathbb{R}^2$ is called A-**invariant** if $A\mathbf{x}$ is in L whenever $\mathbf{x}$ is in L. If we think of A as a linear transformation $\mathbb{R}^2 \to \mathbb{R}^2$, this asks that A carries L into itself, that is the image $A\mathbf{x}$ of each vector $\mathbf{x}$ in L is again in L.

Example 3.3.6

The x axis $L = \left\{ \left[\begin{array}{c} x \\ 0 \end{array} \right] \mid x \text{ in } \mathbb{R} \right\}$ is A-invariant for any matrix of the form

$$A = \left[\begin{array}{cc} a & b \\ 0 & c \end{array} \right] \text{ because } \left[\begin{array}{cc} a & b \\ 0 & c \end{array} \right] \left[\begin{array}{c} x \\ 0 \end{array} \right] = \left[\begin{array}{c} ax \\ 0 \end{array} \right] \text{ is } L \text{ for all } \mathbf{x} = \left[\begin{array}{c} x \\ 0 \end{array} \right] \text{ in } L$$



To see the connection with eigenvectors, let $\mathbf{x} \neq \mathbf{0}$ be any nonzero vector in $\mathbb{R}^2$ and let $L_\mathbf{x}$ denote the unique line through the origin containing $\mathbf{x}$ (see the diagram). By the definition of scalar multiplication in Section 2.6, we see that $L_\mathbf{x}$ consists of all scalar multiples of $\mathbf{x}$, that is

$$L_\mathbf{x} = \mathbb{R}\mathbf{x} = \{t\mathbf{x} \mid t \text{ in } \mathbb{R}\}$$

Now suppose that $\mathbf{x}$ is an eigenvector of A, say $A\mathbf{x} = \lambda\mathbf{x}$ for some λ in $\mathbb{R}$. Then if $t\mathbf{x}$ is in $L_\mathbf{x}$ then

$$A(t\mathbf{x}) = t\,(A\mathbf{x}) = t(\lambda\mathbf{x}) = (t\lambda)\mathbf{x} \text{ is again in } L_\mathbf{x}$$

That is, $L_\mathbf{x}$ is A-invariant. On the other hand, if $L_\mathbf{x}$ is A-invariant then $A\mathbf{x}$ is in $L_\mathbf{x}$ (since $\mathbf{x}$ is in $L_\mathbf{x}$). Hence $A\mathbf{x} = t\mathbf{x}$ for some t in $\mathbb{R}$, so $\mathbf{x}$ is an eigenvector for A (with eigenvalue t). This proves:

Theorem 3.3.3

Let A be a 2×2 matrix, let $\mathbf{x} \neq \mathbf{0}$ be a vector in $\mathbb{R}^2$, and let $L_\mathbf{x}$ be the line through the origin in $\mathbb{R}^2$ containing $\mathbf{x}$. Then

$$\mathbf{x} \text{ is an eigenvector of } A \quad \text{if and only if} \quad L_\mathbf{x} \text{ is } A\text{-invariant}$$

Example 3.3.7

1. If θ is not a multiple of π, show that $A = \left[\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array} \right]$ has no real eigenvalue.

2. If m is real show that $B = \frac{1}{1+m^2}\begin{bmatrix} 1-m^2 & 2m \\ 2m & m^2-1 \end{bmatrix}$ has a 1 as an eigenvalue.

Solution.

1. A induces rotation about the origin through the angle θ (Theorem 2.6.4). Since θ is not a multiple of π, this shows that no line through the origin is A-invariant. Hence A has no eigenvector by Theorem 3.3.3, and so has no eigenvalue.

2. B induces reflection Q_m in the line through the origin with slope m by Theorem 2.6.5. If $\mathbf{x}$ is any nonzero point on this line then it is clear that $Q_m\mathbf{x} = \mathbf{x}$, that is $Q_m\mathbf{x} = 1\mathbf{x}$. Hence 1 is an eigenvalue (with eigenvector $\mathbf{x}$).

If $\theta = \frac{\pi}{2}$ in Example 3.3.7, then $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ so $c_A(x) = x^2 + 1$. This polynomial has no root in $\mathbb{R}$, so A has no (real) eigenvalue, and hence no eigenvector. In fact its eigenvalues are the complex numbers i and $-i$, with corresponding eigenvectors $\begin{bmatrix} 1 \\ -i \end{bmatrix}$ and $\begin{bmatrix} 1 \\ i \end{bmatrix}$ In other words, A *has* eigenvalues and eigenvectors, just not real ones.

Note that *every* polynomial has complex roots,[11] so every matrix has complex eigenvalues. While these eigenvalues may very well be real, this suggests that we really should be doing linear algebra over the complex numbers. Indeed, everything we have done (gaussian elimination, matrix algebra, determinants, etc.) works if all the scalars are complex.

[11]This is called the *Fundamental Theorem of Algebra* and was first proved by Gauss in his doctoral dissertation.

3.4 Diagonalization

An $n \times n$ matrix D is called a **diagonal matrix** if all its entries off the main diagonal are zero, that is if D has the form

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} = \text{diag}\,(\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_n)$$

where $\lambda_1, \lambda_2, \dots, \lambda_n$ are numbers. Calculations with diagonal matrices are very easy. Indeed, if $D = \text{diag}\,(\lambda_1, \lambda_2, \dots, \lambda_n)$ and $E = \text{diag}\,(\mu_1, \mu_2, \dots, \mu_n)$ are two diagonal matrices, their product DE and sum $D+E$ are again diagonal, and are obtained by doing the same operations to corresponding diagonal elements:

$$\begin{aligned} DE &= \text{diag}\,(\lambda_1\mu_1,\ \lambda_2\mu_2,\ \dots,\ \lambda_n\mu_n) \\ D+E &= \text{diag}\,(\lambda_1+\mu_1,\ \lambda_2+\mu_2,\ \dots,\ \lambda_n+\mu_n) \end{aligned}$$

Because of the simplicity of these formulas, and with an eye on Theorem 3.3.1 and the discussion preceding it, we make another definition:

Definition 3.6 Diagonalizable Matrices

An $n \times n$ matrix A is called ***diagonalizable*** *if*

$$P^{-1}AP \text{ is diagonal for some invertible } n \times n \text{ matrix } P$$

Here the invertible matrix P is called a ***diagonalizing matrix*** *for A.*

To discover when such a matrix P exists, we let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ denote the columns of P and look for ways to determine when such $\mathbf{x}_i$ exist and how to compute them. To this end, write P in terms of its columns as follows:

$$P = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$$

Observe that $P^{-1}AP = D$ for some diagonal matrix D holds if and only if

$$AP = PD$$

If we write $D = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$, where the λ_i are numbers to be determined, the equation $AP = PD$ becomes

$$A[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

By the definition of matrix multiplication, each side simplifies as follows

$$\begin{bmatrix} A\mathbf{x}_1 & A\mathbf{x}_2 & \cdots & A\mathbf{x}_n \end{bmatrix} = \begin{bmatrix} \lambda_1\mathbf{x}_1 & \lambda_2\mathbf{x}_2 & \cdots & \lambda_n\mathbf{x}_n \end{bmatrix}$$

Comparing columns shows that $A\mathbf{x}_i = \lambda_i\mathbf{x}_i$ for each i, so

$$P^{-1}AP = D \quad \text{if and only if } A\mathbf{x}_i = \lambda_i\mathbf{x}_i \text{ for each } i$$

In other words, $P^{-1}AP = D$ holds if and only if the diagonal entries of D are eigenvalues of A and the columns of P are corresponding eigenvectors. This proves the following fundamental result.

Theorem 3.4.1

Let A be an $n \times n$ matrix.

1. *A is diagonalizable if and only if it has eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ such that the matrix $P = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{bmatrix}$ is invertible.*

2. *When this is the case, $P^{-1}AP = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ where, for each i, λ_i is the eigenvalue of A corresponding to $\mathbf{x}_i$.*

Example 3.4.1

Diagonalize the matrix $A = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & -1 \\ 1 & 3 & -2 \end{bmatrix}$ in Example 3.3.4.

Solution. By Example 3.3.4, the eigenvalues of A are $\lambda_1 = 2$, $\lambda_2 = 1$, and $\lambda_3 = -1$, with corresponding basic eigenvectors $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, $\mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$, and $\mathbf{x}_3 = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ respectively. Since the matrix $P = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 3 \end{bmatrix}$ is invertible, Theorem 3.4.1 guarantees that

$$P^{-1}AP = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} = D$$

The reader can verify this directly—easier to check $AP = PD$.

In Example 3.4.1, suppose we let $Q = \begin{bmatrix} \mathbf{x}_2 & \mathbf{x}_1 & \mathbf{x}_3 \end{bmatrix}$ be the matrix formed from the eigenvectors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ of A, but in a *different order* than that used to form P. Then $Q^{-1}AQ = \text{diag}\,(\lambda_2, \lambda_1, \lambda_3)$ is diagonal by Theorem 3.4.1, but the eigenvalues are in the *new* order. Hence we can choose the diagonalizing matrix P so that the eigenvalues λ_i appear in any order we want along the main diagonal of D.

In every example above each eigenvalue has had only one basic eigenvector. Here is a diagonalizable matrix where this is not the case.

Example 3.4.2

Diagonalize the matrix $A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$

Solution. To compute the characteristic polynomial of A first add rows 2 and 3 of $xI - A$ to row 1:

$$\begin{aligned} c_A(x) &= \det \begin{bmatrix} x & -1 & -1 \\ -1 & x & -1 \\ -1 & -1 & x \end{bmatrix} = \det \begin{bmatrix} x-2 & x-2 & x-2 \\ -1 & x & -1 \\ -1 & -1 & x \end{bmatrix} \\ &= \det \begin{bmatrix} x-2 & 0 & 0 \\ -1 & x+1 & 0 \\ -1 & 0 & x+1 \end{bmatrix} = (x-2)(x+1)^2 \end{aligned}$$

Hence the eigenvalues are $\lambda_1 = 2$ and $\lambda_2 = -1$, with λ_2 repeated twice (we say that λ_2 has *multiplicity* two). However, A *is* diagonalizable. For $\lambda_1 = 2$, the system of equations $(\lambda_1 I - A)\mathbf{x} = \mathbf{0}$ has general solution $\mathbf{x} = t \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ as the reader can verify, so a basic λ_1-eigenvector is $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

Turning to the repeated eigenvalue $\lambda_2 = -1$, we must solve $(\lambda_2 I - A)\mathbf{x} = \mathbf{0}$. By gaussian elimination, the general solution is $\mathbf{x} = s \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$ where s and t are arbitrary. Hence the gaussian algorithm produces *two* basic λ_2-eigenvectors $\mathbf{x}_2 = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$ and $\mathbf{y}_2 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$ If we take $P = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$ we find that P is invertible. Hence $P^{-1}AP = \text{diag}(2, -1, -1)$ by Theorem 3.4.1.

Example 3.4.2 typifies every diagonalizable matrix. To describe the general case, we need some terminology.

Definition 3.7 Multiplicity of an Eigenvalue

*An eigenvalue λ of a square matrix A is said to have **multiplicity** m if it occurs m times as a root of the characteristic polynomial $c_A(x)$.*

For example, the eigenvalue $\lambda_2 = -1$ in Example 3.4.2 has multiplicity 2. In that example the gaussian algorithm yields two basic λ_2-eigenvectors, the same number as the multiplicity. This works in general.

Theorem 3.4.2

A square matrix A is diagonalizable if and only if every eigenvalue λ of multiplicity m yields exactly m basic eigenvectors; that is, if and only if the general solution of the system $(\lambda I - A)\mathbf{x} = \mathbf{0}$ has exactly m parameters.

One case of Theorem 3.4.2 deserves mention.

Theorem 3.4.3

An $n \times n$ matrix with n distinct eigenvalues is diagonalizable.

The proofs of Theorem 3.4.2 and Theorem 3.4.3 require more advanced techniques and are given in Chapter 5. The following procedure summarizes the method.

Theorem: Diagonalization Algorithm

To diagonalize an $n \times n$ matrix A:

Step 1. Find the distinct eigenvalues λ of A.

Step 2. Compute a set of basic eigenvectors corresponding to each of these eigenvalues λ as basic solutions of the homogeneous system $(\lambda I - A)\mathbf{x} = \mathbf{0}$.

Step 3. The matrix A is diagonalizable if and only if there are n basic eigenvectors in all.

Step 4. If A is diagonalizable, the $n \times n$ matrix P with these basic eigenvectors as its columns is a diagonalizing matrix for A, that is, P is invertible and $P^{-1}AP$ is diagonal.

The diagonalization algorithm is valid even if the eigenvalues are nonreal complex numbers. In this case the eigenvectors will also have complex entries, but we will not pursue this here.

Example 3.4.3

Show that $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ is not diagonalizable.

Solution 1. The characteristic polynomial is $c_A(x) = (x-1)^2$, so A has only one eigenvalue $\lambda_1 = 1$ of multiplicity 2. But the system of equations $(\lambda_1 I - A)\mathbf{x} = \mathbf{0}$ has general solution $t\begin{bmatrix} 1 \\ 0 \end{bmatrix}$, so there is only one parameter, and so only one basic eigenvector $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Hence A is not diagonalizable.

Solution 2. We have $c_A(x) = (x-1)^2$ so the only eigenvalue of A is $\lambda = 1$. Hence, if A were diagonalizable, Theorem 3.4.1 would give $P^{-1}AP = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$ for some invertible matrix P. But then $A = PIP^{-1} = I$, which is not the case. So A cannot be diagonalizable.

Diagonalizable matrices share many properties of their eigenvalues. The following example illustrates why.

Example 3.4.4

If $\lambda^3 = 5\lambda$ for every eigenvalue of the diagonalizable matrix A, show that $A^3 = 5A$.

Solution. Let $P^{-1}AP = D = \text{diag}(\lambda_1, \dots, \lambda_n)$. Because $\lambda_i^3 = 5\lambda_i$ for each i, we obtain

$$D^3 = \text{diag}(\lambda_1^3, \dots, \lambda_n^3) = \text{diag}(5\lambda_1, \dots, 5\lambda_n) = 5D$$

Hence $A^3 = (PDP^{-1})^3 = PD^3P^{-1} = P(5D)P^{-1} = 5(PDP^{-1}) = 5A$ using Theorem 3.3.1. This is what we wanted.

If $p(x)$ is any polynomial and $p(\lambda) = 0$ for every eigenvalue of the diagonalizable matrix A, an argument similar to that in Example 3.4.4 shows that $p(A) = 0$. Thus Example 3.4.4 deals with the case $p(x) = x^3 - 5x$. In general, $p(A)$ is called the *evaluation* of the polynomial $p(x)$ at the matrix A. For example, if $p(x) = 2x^3 - 3x + 5$, then $p(A) = 2A^3 - 3A + 5I$—note the use of the identity matrix.

In particular, if $c_A(x)$ denotes the characteristic polynomial of A, we certainly have $c_A(\lambda) = 0$ for each eigenvalue λ of A (Theorem 3.3.2). Hence $c_A(A) = 0$ for every diagonalizable matrix A. This is, in fact, true for *any* square matrix, diagonalizable or not, and the general result is called the Cayley-Hamilton theorem. It is proved in Section 8.7 and again in Section 11.1.

3.5 Linear Dynamical Systems

We began Section 3.3 with an example from ecology which models the evolution of the population of a species of birds as time goes on. As promised, we now complete the example—Example 3.5.1 below.

The bird population was described by computing the female population profile $\mathbf{v}_k = \left[\begin{array}{c} a_k \\ j_k \end{array} \right]$ of the species, where a_k and j_k represent the number of adult and juvenile females present k years after the initial values a_0 and j_0 were observed. The model assumes that these numbers are related by the following equations:

$$\begin{aligned} a_{k+1} &= \tfrac{1}{2}a_k + \tfrac{1}{4}j_k \\ j_{k+1} &= 2a_k \end{aligned}$$

If we write $A = \left[\begin{array}{cc} \frac{1}{2} & \frac{1}{4} \\ 2 & 0 \end{array} \right]$ the columns $\mathbf{v}_k$ satisfy $\mathbf{v}_{k+1} = A\mathbf{v}_k$ for each $k = 0, 1, 2, \dots$.

Hence $\mathbf{v}_k = A^k\mathbf{v}_0$ for each $k = 1, 2, \dots$. We can now use our diagonalization techniques to determine the population profile $\mathbf{v}_k$ for all values of k in terms of the initial values.

Example 3.5.1

Assuming that the initial values were $a_0 = 100$ adult females and $j_0 = 40$ juvenile females, compute a_k and j_k for $k = 1, 2, \dots$.

Solution. The characteristic polynomial of the matrix $A = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} \\ 2 & 0 \end{bmatrix}$ is $c_A(x) = x^2 - \frac{1}{2}x - \frac{1}{2} = (x-1)(x+\frac{1}{2})$, so the eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = -\frac{1}{2}$ and gaussian elimination gives corresponding basic eigenvectors $\begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -\frac{1}{4} \\ 1 \end{bmatrix}$. For convenience, we can use multiples $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\mathbf{x}_2 = \begin{bmatrix} -1 \\ 4 \end{bmatrix}$ respectively. Hence a diagonalizing matrix is $P = \begin{bmatrix} 1 & -1 \\ 2 & 4 \end{bmatrix}$ and we obtain

$$P^{-1}AP = D \text{ where } D = \begin{bmatrix} 1 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}$$

This gives $A = PDP^{-1}$ so, for each $k \geq 0$, we can compute A^k explicitly:

$$\begin{aligned} A^k = PD^kP^{-1} &= \begin{bmatrix} 1 & -1 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (-\frac{1}{2})^k \end{bmatrix} \frac{1}{6} \begin{bmatrix} 4 & 1 \\ -2 & 1 \end{bmatrix} \\ &= \frac{1}{6} \begin{bmatrix} 4+2(-\frac{1}{2})^k & 1-(-\frac{1}{2})^k \\ 8-8(-\frac{1}{2})^k & 2+4(-\frac{1}{2})^k \end{bmatrix} \end{aligned}$$

Hence we obtain

$$\begin{aligned} \begin{bmatrix} a_k \\ j_k \end{bmatrix} = \mathbf{v}_k = A^k\mathbf{v}_0 &= \frac{1}{6} \begin{bmatrix} 4+2(-\frac{1}{2})^k & 1-(-\frac{1}{2})^k \\ 8-8(-\frac{1}{2})^k & 2+4(-\frac{1}{2})^k \end{bmatrix} \begin{bmatrix} 100 \\ 40 \end{bmatrix} \\ &= \frac{1}{6} \begin{bmatrix} 440+160(-\frac{1}{2})^k \\ 880-640(-\frac{1}{2})^k \end{bmatrix} \end{aligned}$$

Equating top and bottom entries, we obtain exact formulas for a_k and j_k:

$$a_k = \frac{220}{3} + \frac{80}{3}\left(-\frac{1}{2}\right)^k \text{ and } j_k = \frac{440}{3} + \frac{320}{3}\left(-\frac{1}{2}\right)^k \text{ for } k = 1, 2, \cdots$$

In practice, the exact values of a_k and j_k are not usually required. What is needed is a measure of how these numbers behave for large values of k. This is easy to obtain here. Since $(-\frac{1}{2})^k$ is nearly zero for large k, we have the following approximate values

$$a_k \approx \frac{220}{3} \text{ and } j_k \approx \frac{440}{3} \text{ if } k \text{ is large}$$

Hence, in the long term, the female population stabilizes with approximately twice as many juveniles as adults.

Definition 3.8 Linear Dynamical System

If A is an $n \times n$ matrix, a sequence $\mathbf{v}_0,\ \mathbf{v}_1,\ \mathbf{v}_2,\ \dots$ of columns in $\mathbb{R}^n$ is called a ***linear dynamical system*** *if $\mathbf{v}_0$ is specified and $\mathbf{v}_1,\ \mathbf{v}_2,\ \dots$ are given by the matrix recurrence $\mathbf{v}_{k+1} = A\mathbf{v}_k$ for each $k \geq 0$. We call A the* ***migration*** *matrix of the system.*

We have $\mathbf{v}_1 = A\mathbf{v}_0$, then $\mathbf{v}_2 = A\mathbf{v}_1 = A^2\mathbf{v}_0$, and continuing we find

$$\mathbf{v}_k = A^k\mathbf{v}_0 \text{ for each } k = 1,\ 2,\ \cdots \tag{3.9}$$

Hence the columns $\mathbf{v}_k$ are determined by the powers A^k of the matrix A and, as we have seen, these powers can be efficiently computed if A is diagonalizable. In fact Equation 3.9 can be used to give a nice "formula" for the columns $\mathbf{v}_k$ in this case.

Assume that A is diagonalizable with eigenvalues $\lambda_1,\ \lambda_2,\ \dots,\ \lambda_n$ and corresponding basic eigenvectors $\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_n$. If $P = \left[\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{array}\right]$ is a diagonalizing matrix with the $\mathbf{x}_i$ as columns, then P is invertible and

$$P^{-1}AP = D = \operatorname{diag}(\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_n)$$

by Theorem 3.4.1. Hence $A = PDP^{-1}$ so Equation 3.9 and Theorem 3.3.1 give

$$\mathbf{v}_k = A^k\mathbf{v}_0 = (PDP^{-1})^k\mathbf{v}_0 = (PD^kP^{-1})\mathbf{v}_0 = PD^k(P^{-1}\mathbf{v}_0)$$

for each $k = 1,\ 2,\ \dots$. For convenience, we denote the column $P^{-1}\mathbf{v}_0$ arising here as follows:

$$\mathbf{b} = P^{-1}\mathbf{v}_0 = \left[\begin{array}{c} b_1 \\ b_2 \\ \vdots \\ b_n \end{array}\right]$$

Then matrix multiplication gives

$$\begin{aligned}
\mathbf{v}_k &= PD^k(P^{-1}\mathbf{v}_0) \\
&= \left[\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{array}\right] \left[\begin{array}{cccc} \lambda_1^k & 0 & \cdots & 0 \\ 0 & \lambda_2^k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n^k \end{array}\right] \left[\begin{array}{c} b_1 \\ b_2 \\ \vdots \\ b_n \end{array}\right] \\
&= \left[\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{array}\right] \left[\begin{array}{c} b_1\lambda_1^k \\ b_2\lambda_2^k \\ \vdots \\ b_3\lambda_n^k \end{array}\right] \\
&= b_1\lambda_1^k\mathbf{x}_1 + b_2\lambda_2^k\mathbf{x}_2 + \cdots + b_n\lambda_n^k\mathbf{x}_n
\end{aligned} \tag{3.10}$$

for each $k \geq 0$. This is a useful **exact formula** for the columns $\mathbf{v}_k$. Note that, in particular,

$$\mathbf{v}_0 = b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \cdots + b_n\mathbf{x}_n$$

However, such an exact formula for $\mathbf{v}_k$ is often not required in practice; all that is needed is to *estimate* $\mathbf{v}_k$ for large values of k (as was done in Example 3.5.1). This can be easily done if A has a largest eigenvalue. An eigenvalue λ of a matrix A is called a **dominant eigenvalue** of A if it has multiplicity 1 and

$$|\lambda| > |\mu| \text{ for all eigenvalues } \mu \neq \lambda$$

where $|\lambda|$ denotes the absolute value of the number λ. For example, $\lambda_1 = 1$ is dominant in Example 3.5.1.

Returning to the above discussion, suppose that A has a dominant eigenvalue. By choosing the order in which the columns $\mathbf{x}_i$ are placed in P, we may assume that λ_1 is dominant among the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ of A (see the discussion following Example 3.4.1). Now recall the exact expression for $\mathbf{v}_k$ in Equation 3.10 above:

$$\mathbf{v}_k = b_1\lambda_1^k\mathbf{x}_1 + b_2\lambda_2^k\mathbf{x}_2 + \cdots + b_n\lambda_n^k\mathbf{x}_n$$

Take λ_1^k out as a common factor in this equation to get

$$\mathbf{v}_k = \lambda_1^k\left[b_1\mathbf{x}_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\mathbf{x}_2 + \cdots + b_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\mathbf{x}_n\right]$$

for each $k \geq 0$. Since λ_1 is dominant, we have $|\lambda_i| < |\lambda_1|$ for each $i \geq 2$, so each of the numbers $(\lambda_i/\lambda_1)^k$ become small in absolute value as k increases. Hence $\mathbf{v}_k$ is approximately equal to the first term $\lambda_1^k b_1\mathbf{x}_1$, and we write this as $\mathbf{v}_k \approx \lambda_1^k b_1\mathbf{x}_1$. These observations are summarized in the following theorem (together with the above exact formula for $\mathbf{v}_k$).

Theorem 3.5.1

Consider the dynamical system $\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \dots$ with matrix recurrence

$$\mathbf{v}_{k+1} = A\mathbf{v}_k \text{ for } k \geq 0$$

where A and $\mathbf{v}_0$ are given. Assume that A is a diagonalizable $n \times n$ matrix with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ and corresponding basic eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, and let $P = \left[\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{array}\right]$ be the diagonalizing matrix. Then an exact formula for $\mathbf{v}_k$ is

$$\mathbf{v}_k = b_1\lambda_1^k\mathbf{x}_1 + b_2\lambda_2^k\mathbf{x}_2 + \cdots + b_n\lambda_n^k\mathbf{x}_n \text{ for each } k \geq 0$$

where the coefficients b_i come from

$$\mathbf{b} = P^{-1}\mathbf{v}_0 = \left[\begin{array}{c} b_1 \\ b_2 \\ \vdots \\ b_n \end{array}\right]$$

Moreover, if A has dominant[12] eigenvalue λ_1, then $\mathbf{v}_k$ is approximated by

$$\mathbf{v}_k = b_1\lambda_1^k\mathbf{x}_1 \text{ for sufficiently large } k.$$

[12]Similar results can be found in other situations. If for example, eigenvalues λ_1 and λ_2 (possibly equal) satisfy $|\lambda_1| = |\lambda_2| > |\lambda_i|$ for all $i > 2$, then we obtain $\mathbf{v}_k \approx b_1\lambda_1^k x_1 + b_2\lambda_2^k x_2$ for large k.

Example 3.5.2

Returning to Example 3.5.1, we see that $\lambda_1 = 1$ is the dominant eigenvalue, with eigenvector $\mathbf{x}_1 = \left[\begin{array}{c} 1 \\ 2 \end{array}\right]$. Here $P = \left[\begin{array}{rr} 1 & -1 \\ 2 & 4 \end{array}\right]$ and $\mathbf{v}_0 = \left[\begin{array}{c} 100 \\ 40 \end{array}\right]$ so $P^{-1}\mathbf{v}_0 = \frac{1}{3}\left[\begin{array}{r} 220 \\ -80 \end{array}\right]$. Hence $b_1 = \frac{220}{3}$ in the notation of Theorem 3.5.1, so

$$\left[\begin{array}{c} a_k \\ j_k \end{array}\right] = \mathbf{v}_k \approx b_1\lambda_1^k\mathbf{x}_1 = \tfrac{220}{3}1^k\left[\begin{array}{c} 1 \\ 2 \end{array}\right]$$

where k is large. Hence $a_k \approx \frac{220}{3}$ and $j_k \approx \frac{440}{3}$ as in Example 3.5.1.

This next example uses Theorem 3.5.1 to solve a "linear recurrence." See also Section 3.4.

Example 3.5.3

Suppose a sequence $x_0,\ x_1,\ x_2,\ \dots$ is determined by insisting that

$$x_0 = 1,\ x_1 = -1,\ \text{ and } x_{k+2} = 2x_k - x_{k+1} \text{ for every } k \geq 0$$

Find a formula for x_k in terms of k.

Solution. Using the linear recurrence $x_{k+2} = 2x_k - x_{k+1}$ repeatedly gives

$$x_2 = 2x_0 - x_1 = 3, \quad x_3 = 2x_1 - x_2 = -5, \quad x_4 = 11, \quad x_5 = -21,\ \dots$$

so the x_i are determined but no pattern is apparent. The idea is to find $\mathbf{v}_k = \left[\begin{array}{c} x_k \\ x_{k+1} \end{array}\right]$ for each k instead, and then retrieve x_k as the top component of $\mathbf{v}_k$. The reason this works is that the linear recurrence guarantees that these $\mathbf{v}_k$ are a dynamical system:

$$\mathbf{v}_{k+1} = \left[\begin{array}{c} x_{k+1} \\ x_{k+2} \end{array}\right] = \left[\begin{array}{c} x_{k+1} \\ 2x_k - x_{k+1} \end{array}\right] = A\mathbf{v}_k \text{ where } A = \left[\begin{array}{rr} 0 & 1 \\ 2 & -1 \end{array}\right]$$

The eigenvalues of A are $\lambda_1 = -2$ and $\lambda_2 = 1$ with eigenvectors $\mathbf{x}_1 = \left[\begin{array}{r} 1 \\ -2 \end{array}\right]$ and $\mathbf{x}_2 = \left[\begin{array}{c} 1 \\ 1 \end{array}\right]$, so the diagonalizing matrix is $P = \left[\begin{array}{rr} 1 & 1 \\ -2 & 1 \end{array}\right]$.

Moreover, $\mathbf{b} = P_0^{-1}\mathbf{v}_0 = \frac{1}{3}\left[\begin{array}{c} 2 \\ 1 \end{array}\right]$ so the exact formula for $\mathbf{v}_k$ is

$$\left[\begin{array}{c} x_k \\ x_{k+1} \end{array}\right] = \mathbf{v}_k = b_1\lambda_1^k\mathbf{x}_1 + b_2\lambda_2^k\mathbf{x}_2 = \tfrac{2}{3}(-2)^k\left[\begin{array}{r} 1 \\ -2 \end{array}\right] + \tfrac{1}{3}1^k\left[\begin{array}{c} 1 \\ 1 \end{array}\right]$$

Equating top entries gives the desired formula for x_k:

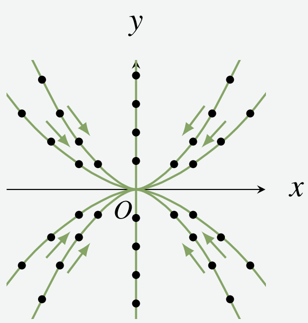
$$x_k = \tfrac{1}{3}\left[2(-2)^k + 1\right] \text{ for all } k = 0,\ 1,\ 2,\ \dots$$

The reader should check this for the first few values of k.

Graphical Description of Dynamical Systems

If a dynamical system $\mathbf{v}_{k+1} = A\mathbf{v}_k$ is given, the sequence $\mathbf{v}_0,\ \mathbf{v}_1,\ \mathbf{v}_2,\ \dots$ is called the **trajectory** of the system starting at $\mathbf{v}_0$. It is instructive to obtain a graphical plot of the system by writing $\mathbf{v}_k = \left[\begin{array}{c} x_k \\ y_k \end{array} \right]$ and plotting the successive values as points in the plane, identifying $\mathbf{v}_k$ with the point $(x_k,\ y_k)$ in the plane. We give several examples which illustrate properties of dynamical systems. For ease of calculation we assume that the matrix A is simple, usually diagonal.

Example 3.5.4

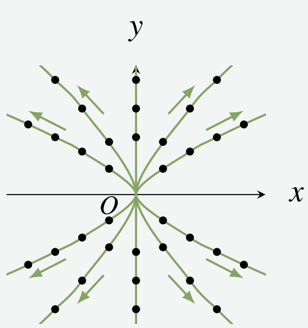


Let $A = \left[\begin{array}{cc} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{array} \right]$ Then the eigenvalues are $\frac{1}{2}$ and $\frac{1}{3}$, with corresponding eigenvectors $\mathbf{x}_1 = \left[\begin{array}{c} 1 \\ 0 \end{array} \right]$ and $\mathbf{x}_2 = \left[\begin{array}{c} 0 \\ 1 \end{array} \right]$. The exact formula is

$$\mathbf{v}_k = b_1 \left(\tfrac{1}{2}\right)^k \left[\begin{array}{c} 1 \\ 0 \end{array} \right] + b_2 \left(\tfrac{1}{3}\right)^k \left[\begin{array}{c} 0 \\ 1 \end{array} \right]$$

for $k = 0,\ 1,\ 2,\ \dots$ by Theorem 3.5.1, where the coefficients b_1 and b_2 depend on the initial point $\mathbf{v}_0$. Several trajectories are plotted in the diagram and, for each choice of $\mathbf{v}_0$, the trajectories converge toward the origin because both eigenvalues are less than 1 in absolute value. For this reason, the origin is called an **attractor** for the system.

Example 3.5.5

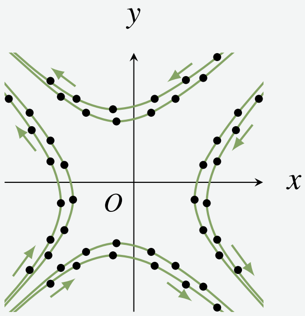


Let $A = \left[\begin{array}{cc} \frac{3}{2} & 0 \\ 0 & \frac{4}{3} \end{array} \right]$. Here the eigenvalues are $\frac{3}{2}$ and $\frac{4}{3}$, with corresponding eigenvectors $\mathbf{x}_1 = \left[\begin{array}{c} 1 \\ 0 \end{array} \right]$ and $\mathbf{x}_2 = \left[\begin{array}{c} 0 \\ 1 \end{array} \right]$ as before. The exact formula is

$$\mathbf{v}_k = b_1 \left(\tfrac{3}{2}\right)^k \left[\begin{array}{c} 1 \\ 0 \end{array} \right] + b_2 \left(\tfrac{4}{3}\right)^k \left[\begin{array}{c} 0 \\ 1 \end{array} \right]$$

for $k = 0,\ 1,\ 2,\ \dots$. Since both eigenvalues are greater than 1 in absolute value, the trajectories diverge away from the origin for every choice of initial point V_0. For this reason, the origin is called a **repellor** for the system.

Example 3.5.6



Let $A = \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}$. Now the eigenvalues are $\frac{3}{2}$ and $\frac{1}{2}$, with corresponding eigenvectors $\mathbf{x}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and $\mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ The exact formula is

$$\mathbf{v}_k = b_1 \left(\frac{3}{2}\right)^k \begin{bmatrix} -1 \\ 1 \end{bmatrix} + b_2 \left(\frac{1}{2}\right)^k \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

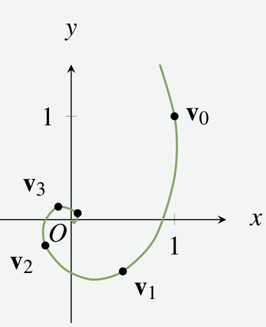
for $k = 0, 1, 2, \dots$. In this case $\frac{3}{2}$ is the dominant eigenvalue so, if $b_1 \neq 0$, we have $\mathbf{v}_k \approx b_1 \left(\frac{3}{2}\right)^k \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ for large k and $\mathbf{v}_k$ is approaching the line $y = -x$.

However, if $b_1 = 0$, then $\mathbf{v}_k = b_2 \left(\frac{1}{2}\right)^k \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and so approaches the origin along the line $y = x$. In general the trajectories appear as in the diagram, and the origin is called a **saddle point** for the dynamical system in this case.

Example 3.5.7

Let $A = \begin{bmatrix} 0 & \frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix}$. Now the characteristic polynomial is $c_A(x) = x^2 + \frac{1}{4}$, so the eigenvalues are the complex numbers $\frac{i}{2}$ and $-\frac{i}{2}$ where $i^2 = -1$. Hence A is not diagonalizable as a real matrix. However, the trajectories are not difficult to describe. If we start with $\mathbf{v}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ then the trajectory begins as

$$\mathbf{v}_1 = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} -\frac{1}{4} \\ -\frac{1}{4} \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} -\frac{1}{8} \\ \frac{1}{8} \end{bmatrix}, \mathbf{v}_4 = \begin{bmatrix} \frac{1}{16} \\ \frac{1}{16} \end{bmatrix}, \mathbf{v}_5 = \begin{bmatrix} \frac{1}{32} \\ -\frac{1}{32} \end{bmatrix}, \mathbf{v}_6 = \begin{bmatrix} -\frac{1}{64} \\ -\frac{1}{64} \end{bmatrix}, \dots$$



The first five of these points are plotted in the diagram. Here each trajectory spirals in toward the origin, so the origin is an attractor. Note that the two (complex) eigenvalues have absolute value less than 1 here. If they had absolute value greater than 1, the trajectories would spiral out from the origin.

Google PageRank

Dominant eigenvalues are useful to the Google search engine for finding information on the Web. If an information query comes in from a client, Google has a sophisticated method of establishing the "relevance" of each site to that query. When the relevant sites have been determined, they are placed in order of importance using a ranking of *all* sites called the PageRank. The relevant sites with the highest PageRank are the ones presented to the client. It is the construction of the PageRank that is our interest here.

The Web contains many links from one site to another. Google interprets a link from site j to site i as a "vote" for the importance of site i. Hence if site i has more links to it than does site j, then i is regarded as more "important" and assigned a higher PageRank. One way to look at this is to view the sites as vertices in a huge directed graph (see Section 2.2). Then if site j links to site i there is an edge from j to i, and hence the (i, j)-entry is a 1 in the associated adjacency matrix (called the *connectivity* matrix in this context). Thus a large number of 1s in row i of this matrix is a measure of the PageRank of site i.[13]

However this does not take into account the PageRank of the sites that link to i. Intuitively, the higher the rank of these sites, the higher the rank of site i. One approach is to compute a dominant eigenvector $\mathbf{x}$ for the connectivity matrix. In most cases the entries of $\mathbf{x}$ can be chosen to be positive with sum 1. Each site corresponds to an entry of $\mathbf{x}$, so the sum of the entries of sites linking to a given site i is a measure of the rank of site i. In fact, Google chooses the PageRank of a site so that it is proportional to this sum.[14]

[13] For more on PageRank, visit https://en.wikipedia.org/wiki/PageRank.

[14] See the articles "Searching the web with eigenvectors" by Herbert S. Wilf, UMAP Journal 23(2), 2002, pages 101–103, and "The worlds largest matrix computation: Google's PageRank is an eigenvector of a matrix of order 2.7 billion" by Cleve Moler, Matlab News and Notes, October 2002, pages 12–13.

3.6 An Application to Linear Recurrences

It often happens that a problem can be solved by finding a sequence of numbers $x_0, x_1, x_2, \dots$ where the first few are known, and subsequent numbers are given in terms of earlier ones. Here is a combinatorial example where the object is to count the number of ways to do something.

Example 3.6.1

An urban planner wants to determine the number x_k of ways that a row of k parking spaces can be filled with cars and trucks if trucks take up two spaces each. Find the first few values of x_k.

Solution. Clearly, $x_0 = 1$ and $x_1 = 1$, while $x_2 = 2$ since there can be two cars or one truck. We have $x_3 = 3$ (the 3 configurations are ccc, cT, and Tc) and $x_4 = 5$ ($cccc$, ccT, cTc, Tcc, and TT). The key to this method is to find a way to express each subsequent x_k in terms of earlier values. In this case we claim that

$$x_{k+2} = x_k + x_{k+1} \text{ for every } k \geq 0 \tag{3.11}$$

Indeed, every way to fill $k+2$ spaces falls into one of two categories: Either a car is parked in the first space (and the remaining $k+1$ spaces are filled in x_{k+1} ways), or a truck is parked in the first two spaces (with the other k spaces filled in x_k ways). Hence, there are $x_{k+1} + x_k$ ways to fill the $k+2$ spaces. This is Equation 3.11.

The recurrence in Equation 3.11 determines x_k for every $k \geq 2$ since x_0 and x_1 are given. In fact, the first few values are

$$\begin{array}{lllll} x_0 & = & 1 & & \\ x_1 & = & 1 & & \\ x_2 & = & x_0 + x_1 & = & 2 \\ x_3 & = & x_1 + x_2 & = & 3 \\ x_4 & = & x_2 + x_3 & = & 5 \\ x_5 & = & x_3 + x_4 & = & 8 \\ \vdots & & \vdots & & \vdots \end{array}$$

Clearly, we can find x_k for any value of k, but one wishes for a "formula" for x_k as a function of k. It turns out that such a formula can be found using diagonalization. We will return to this example later.

A sequence $x_0, x_1, x_2, \dots$ of numbers is said to be given **recursively** if each number in the sequence is completely determined by those that come before it. Such sequences arise frequently in mathematics and computer science, and also occur in other parts of science. The formula $x_{k+2} = x_{k+1} + x_k$ in Example 3.6.1 is an example of a **linear recurrence relation** of length 2 because x_{k+2} is the sum of the two preceding terms x_{k+1} and x_k; in general, the **length** is m if x_{k+m} is a sum of multiples of $x_k, x_{k+1}, \dots, x_{k+m-1}$.

The simplest linear recursive sequences are of length 1, that is x_{k+1} is a fixed multiple of x_k for each k, say $x_{k+1} = ax_k$. If x_0 is specified, then $x_1 = ax_0$, $x_2 = ax_1 = a^2x_0$, and $x_3 = ax_2 = a^3x_0$, ... Continuing, we obtain $x_k = a^kx_0$ for each $k \geq 0$, which is an explicit formula for x_k as a function of k (when x_0 is given).

Such formulas are not always so easy to find for all choices of the initial values. Here is an example where diagonalization helps.

Example 3.6.2

Suppose the numbers $x_0, x_1, x_2, \dots$ are given by the linear recurrence relation

$$x_{k+2} = x_{k+1} + 6x_k \text{ for } k \geq 0$$

where x_0 and x_1 are specified. Find a formula for x_k when $x_0 = 1$ and $x_1 = 3$, and also when $x_0 = 1$ and $x_1 = 1$.

Solution. If $x_0 = 1$ and $x_1 = 3$, then

$$x_2 = x_1 + 6x_0 = 9, \quad x_3 = x_2 + 6x_1 = 27, \quad x_4 = x_3 + 6x_2 = 81$$

and it is apparent that

$$x_k = 3^k \text{ for } k = 0, 1, 2, 3, \text{ and } 4$$

This formula holds for all k because it is true for $k = 0$ and $k = 1$, and it satisfies the recurrence $x_{k+2} = x_{k+1} + 6x_k$ for each k as is readily checked.

However, if we begin instead with $x_0 = 1$ and $x_1 = 1$, the sequence continues

$$x_2 = 7, \quad x_3 = 13, \quad x_4 = 55, \quad x_5 = 133, \quad \dots$$

In this case, the sequence is uniquely determined but no formula is apparent. Nonetheless, a simple device transforms the recurrence into a matrix recurrence to which our diagonalization techniques apply.

The idea is to compute the sequence $\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \dots$ of columns instead of the numbers $x_0, x_1, x_2, \dots,$ where

$$\mathbf{v}_k = \left[\begin{array}{c} x_k \\ x_{k+1} \end{array} \right] \text{ for each } k \geq 0$$

Then $\mathbf{v}_0 = \left[\begin{array}{c} x_0 \\ x_1 \end{array} \right] = \left[\begin{array}{c} 1 \\ 1 \end{array} \right]$ is specified, and the numerical recurrence $x_{k+2} = x_{k+1} + 6x_k$ transforms into a matrix recurrence as follows:

$$\mathbf{v}_{k+1} = \left[\begin{array}{c} x_{k+1} \\ x_{k+2} \end{array} \right] = \left[\begin{array}{c} x_{k+1} \\ 6x_k + x_{k+1} \end{array} \right] = \left[\begin{array}{cc} 0 & 1 \\ 6 & 1 \end{array} \right] \left[\begin{array}{c} x_k \\ x_{k+1} \end{array} \right] = A\mathbf{v}_k$$

where $A = \left[\begin{array}{cc} 0 & 1 \\ 6 & 1 \end{array} \right]$. Thus these columns $\mathbf{v}_k$ are a linear dynamical system, so Theorem 3.5.1 applies provided the matrix A is diagonalizable.

We have $c_A(x) = (x-3)(x+2)$ so the eigenvalues are $\lambda_1 = 3$ and $\lambda_2 = -2$ with corresponding eigenvectors $\mathbf{x}_1 = \left[\begin{array}{c} 1 \\ 3 \end{array} \right]$ and $\mathbf{x}_2 = \left[\begin{array}{r} -1 \\ 2 \end{array} \right]$ as the reader can check. Since $P = \left[\begin{array}{cc} \mathbf{x}_1 & \mathbf{x}_2 \end{array} \right] = \left[\begin{array}{rr} 1 & -1 \\ 3 & 2 \end{array} \right]$ is invertible, it is a diagonalizing matrix for A. The coefficients b_i in Theorem 3.5.1 are given by $\left[\begin{array}{c} b_1 \\ b_2 \end{array} \right] = P^{-1}\mathbf{v}_0 = \left[\begin{array}{c} \frac{3}{5} \\ \frac{-2}{5} \end{array} \right]$, so that the theorem gives

$$\left[\begin{array}{c} x_k \\ x_{k+1} \end{array} \right] = \mathbf{v}_k = b_1\lambda_1^k\mathbf{x}_1 + b_2\lambda_2^k\mathbf{x}_2 = \frac{3}{5}3^k \left[\begin{array}{c} 1 \\ 3 \end{array} \right] + \frac{-2}{5}(-2)^k \left[\begin{array}{r} -1 \\ 2 \end{array} \right]$$

Equating top entries yields

$$x_k = \frac{1}{5}\left[3^{k+1} - (-2)^{k+1}\right] \text{ for } k \geq 0$$

This gives $x_0 = 1 = x_1$, and it satisfies the recurrence $x_{k+2} = x_{k+1} + 6x_k$ as is easily verified. Hence, it is the desired formula for the x_k.

Returning to Example 3.6.1, these methods give an exact formula and a good approximation for the numbers x_k in that problem.

Example 3.6.3

In Example 3.6.1, an urban planner wants to determine x_k, the number of ways that a row of k parking spaces can be filled with cars and trucks if trucks take up two spaces each. Find a formula for x_k and estimate it for large k.

Solution. We saw in Example 3.6.1 that the numbers x_k satisfy a linear recurrence

$$x_{k+2} = x_k + x_{k+1} \text{ for every } k \geq 0$$

If we write $\mathbf{v}_k = \left[\begin{array}{c} x_k \\ x_{k+1} \end{array}\right]$ as before, this recurrence becomes a matrix recurrence for the $\mathbf{v}_k$:

$$\mathbf{v}_{k+1} = \left[\begin{array}{c} x_{k+1} \\ x_{k+2} \end{array}\right] = \left[\begin{array}{c} x_{k+1} \\ x_k + x_{k+1} \end{array}\right] = \left[\begin{array}{cc} 0 & 1 \\ 1 & 1 \end{array}\right]\left[\begin{array}{c} x_k \\ x_{k+1} \end{array}\right] = A\mathbf{v}_k$$

for all $k \geq 0$ where $A = \left[\begin{array}{cc} 0 & 1 \\ 1 & 1 \end{array}\right]$. Moreover, A is diagonalizable here. The characteristic polynomial is $c_A(x) = x^2 - x - 1$ with roots $\frac{1}{2}\left[1 \pm \sqrt{5}\right]$ by the quadratic formula, so A has eigenvalues

$$\lambda_1 = \frac{1}{2}\left(1 + \sqrt{5}\right) \text{ and } \lambda_2 = \frac{1}{2}\left(1 - \sqrt{5}\right)$$

Corresponding eigenvectors are $\mathbf{x}_1 = \left[\begin{array}{c} 1 \\ \lambda_1 \end{array}\right]$ and $\mathbf{x}_2 = \left[\begin{array}{c} 1 \\ \lambda_2 \end{array}\right]$ respectively as the reader can verify. As the matrix $P = \left[\begin{array}{cc} \mathbf{x}_1 & \mathbf{x}_2 \end{array}\right] = \left[\begin{array}{cc} 1 & 1 \\ \lambda_1 & \lambda_2 \end{array}\right]$ is invertible, it is a diagonalizing matrix for A. We compute the coefficients b_1 and b_2 (in Theorem 3.5.1) as follows:

$$\left[\begin{array}{c} b_1 \\ b_2 \end{array}\right] = P^{-1}\mathbf{v}_0 = \frac{1}{-\sqrt{5}}\left[\begin{array}{cc} \lambda_2 & -1 \\ -\lambda_1 & 1 \end{array}\right]\left[\begin{array}{c} 1 \\ 1 \end{array}\right] = \frac{1}{\sqrt{5}}\left[\begin{array}{c} \lambda_1 \\ -\lambda_2 \end{array}\right]$$

where we used the fact that $\lambda_1 + \lambda_2 = 1$. Thus Theorem 3.5.1 gives

$$\left[\begin{array}{c} x_k \\ x_{k+1} \end{array}\right] = \mathbf{v}_k = b_1\lambda_1^k\mathbf{x}_1 + b_2\lambda_2^k\mathbf{x}_2 = \frac{\lambda_1}{\sqrt{5}}\lambda_1^k\left[\begin{array}{c} 1 \\ \lambda_1 \end{array}\right] - \frac{\lambda_2}{\sqrt{5}}\lambda_2^k\left[\begin{array}{c} 1 \\ \lambda_2 \end{array}\right]$$

Comparing top entries gives an exact formula for the numbers x_k:

$$x_k = \frac{1}{\sqrt{5}}\left[\lambda_1^{k+1} - \lambda_2^{k+1}\right] \text{ for } k \geq 0$$

Finally, observe that λ_1 is dominant here (in fact, $\lambda_1 = 1.618$ and $\lambda_2 = -0.618$ to three decimal places) so λ_2^{k+1} is negligible compared with λ_1^{k+1} is large. Thus,

$$x_k \approx \frac{1}{\sqrt{5}}\lambda_1^{k+1} \text{ for each } k \geq 0.$$

This is a good approximation, even for as small a value as $k = 12$. Indeed, repeated use of the recurrence $x_{k+2} = x_k + x_{k+1}$ gives the exact value $x_{12} = 233$, while the approximation is $x_{12} \approx \frac{(1.618)^{13}}{\sqrt{5}} = 232.94$.

The sequence $x_0, x_1, x_2, \dots$ in Example 3.6.3 was first discussed in 1202 by Leonardo Pisano of Pisa, also known as Fibonacci,[15] and is now called the **Fibonacci sequence**. It is completely determined by the conditions $x_0 = 1$, $x_1 = 1$ and the recurrence $x_{k+2} = x_k + x_{k+1}$ for each $k \geq 0$. These numbers have been studied for centuries and have many interesting properties (there is even a journal, the *Fibonacci Quarterly*, devoted exclusively to them). For example, biologists have discovered that the arrangement of leaves around the stems of some plants follow a Fibonacci pattern. The formula $x_k = \frac{1}{\sqrt{5}}\left[\lambda_1^{k+1} - \lambda_2^{k+1}\right]$ in Example 3.6.3 is called the **Binet formula**. It is remarkable in that the x_k are integers but λ_1 and λ_2 are not. This phenomenon can occur even if the eigenvalues λ_i are nonreal complex numbers.

We conclude with an example showing that *nonlinear* recurrences can be very complicated.

Example 3.6.4

Suppose a sequence $x_0, x_1, x_2, \dots$ satisfies the following recurrence:

$$x_{k+1} = \begin{cases} \frac{1}{2}x_k & \text{if } x_k \text{ is even} \\ 3x_k + 1 & \text{if } x_k \text{ is odd} \end{cases}$$

If $x_0 = 1$, the sequence is $1, 4, 2, 1, 4, 2, 1, \dots$ and so continues to cycle indefinitely. The same thing happens if $x_0 = 7$. Then the sequence is

$$7, 22, 11, 34, 17, 52, 26, 13, 40, 20, 10, 5, 16, 8, 4, 2, 1, \dots$$

and it again cycles. However, it is not known whether every choice of x_0 will lead eventually to 1. It is quite possible that, for some x_0, the sequence will continue to produce different values indefinitely, or will repeat a value and cycle without reaching 1. No one knows for sure.

[15]Fibonacci was born in Italy. As a young man he travelled to India where he encountered the "Fibonacci" sequence. He returned to Italy and published this in his book *Liber Abaci* in 1202. In the book he is the first to bring the Hindu decimal system for representing numbers to Europe.

3.7 An Application to Systems of Differential Equations

A function f of a real variable is said to be **differentiable** if its derivative exists and, in this case, we let f' denote the derivative. If f and g are differentiable functions, a system

$$\begin{aligned} f' &= 3f + 5g \\ g' &= -f + 2g \end{aligned}$$

is called a *system of first order differential equations,* or a *differential system* for short. Solving many practical problems often comes down to finding sets of functions that satisfy such a system (often involving more than two functions). In this section we show how diagonalization can help. Of course an acquaintance with calculus is required.

The Exponential Function

The simplest differential system is the following single equation:

$$f' = af \text{ where } a \text{ is constant} \tag{3.12}$$

It is easily verified that $f(x) = e^{ax}$ is one solution; in fact, Equation 3.12 is simple enough for us to find *all* solutions. Suppose that f is any solution, so that $f'(x) = af(x)$ for all x. Consider the new function g given by $g(x) = f(x)e^{-ax}$. Then the product rule of differentiation gives

$$g'(x) = f(x)\left[-ae^{-ax}\right] + f'(x)e^{-ax}$$

$$= -af(x)e^{-ax} + [af(x)]\, e^{-ax}$$
$$= 0$$

for all x. Hence the function $g(x)$ has zero derivative and so must be a constant, say $g(x) = c$. Thus $c = g(x) = f(x)e^{-ax}$, that is

$$f(x) = ce^{ax}$$

In other words, every solution $f(x)$ of Equation 3.12 is just a scalar multiple of e^{ax}. Since every such scalar multiple is easily seen to be a solution of Equation 3.12, we have proved

Theorem 3.7.1

The set of solutions to $f' = af$ is $\{ce^{ax} \mid c \text{ any constant}\} = \mathbb{R}e^{ax}$.

Remarkably, this result together with diagonalization enables us to solve a wide variety of differential systems.

Example 3.7.1

Assume that the number $n(t)$ of bacteria in a culture at time t has the property that the rate of change of n is proportional to n itself. If there are n_0 bacteria present when $t = 0$, find the number at time t.

Solution. Let k denote the proportionality constant. The rate of change of $n(t)$ is its time-derivative $n'(t)$, so the given relationship is $n'(t) = kn(t)$. Thus Theorem 3.7.1 shows that all solutions n are given by $n(t) = ce^{kt}$, where c is a constant. In this case, the constant c is determined by the requirement that there be n_0 bacteria present when $t = 0$. Hence $n_0 = n(0) = ce^{k0} = c$, so

$$n(t) = n_0 e^{kt}$$

gives the number at time t. Of course the constant k depends on the strain of bacteria.

The condition that $n(0) = n_0$ in Example 3.7.1 is called an **initial condition** or a **boundary condition** and serves to select one solution from the available solutions.

General Differential Systems

Solving a variety of problems, particularly in science and engineering, comes down to solving a system of linear differential equations. Diagonalization enters into this as follows. The general problem is to find differentiable functions $f_1, f_2, \ldots, f_n$ that satisfy a system of equations of the form

$$\begin{array}{ccccccccc}
f_1' & = & a_{11}f_1 & + & a_{12}f_2 & + & \cdots & + & a_{1n}f_n \\
f_2' & = & a_{21}f_1 & + & a_{22}f_2 & + & \cdots & + & a_{2n}f_n \\
\vdots & & \vdots & & \vdots & & & & \vdots \\
f_n' & = & a_{n1}f_1 & + & a_{n2}f_2 & + & \cdots & + & a_{nn}f_n
\end{array}$$

where the a_{ij} are constants. This is called a **linear system of differential equations** or simply a **differential system**. The first step is to put it in matrix form. Write

$$\mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix} \quad \mathbf{f}' = \begin{bmatrix} f_1' \\ f_2' \\ \vdots \\ f_n' \end{bmatrix} \quad A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

Then the system can be written compactly using matrix multiplication:

$$\mathbf{f}' = A\mathbf{f}$$

Hence, given the matrix A, the problem is to find a column $\mathbf{f}$ of differentiable functions that satisfies this condition. This can be done if A is diagonalizable. Here is an example.

Example 3.7.2

Find a solution to the system

$$\begin{aligned} f_1' &= f_1 + 3f_2 \\ f_2' &= 2f_1 + 2f_2 \end{aligned}$$

that satisfies $f_1(0) = 0$, $f_2(0) = 5$.

Solution. This is $\mathbf{f}' = A\mathbf{f}$, where $\mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$ and $A = \begin{bmatrix} 1 & 3 \\ 2 & 2 \end{bmatrix}$. The reader can verify that $c_A(x) = (x-4)(x+1)$, and that $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\mathbf{x}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$ are eigenvectors corresponding to the eigenvalues 4 and -1, respectively. Hence the diagonalization algorithm gives $P^{-1}AP = \begin{bmatrix} 4 & 0 \\ 0 & -1 \end{bmatrix}$, where $P = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 1 & -2 \end{bmatrix}$. Now consider new functions g_1 and g_2 given by $\mathbf{f} = P\mathbf{g}$ (equivalently, $\mathbf{g} = P^{-1}\mathbf{f}$), where $\mathbf{g} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$ Then

$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \quad \text{that is,} \quad \begin{aligned} f_1 &= g_1 + 3g_2 \\ f_2 &= g_1 - 2g_2 \end{aligned}$$

Hence $f_1' = g_1' + 3g_2'$ and $f_2' = g_1' - 2g_2'$ so that

$$\mathbf{f}' = \begin{bmatrix} f_1' \\ f_2' \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} g_1' \\ g_2' \end{bmatrix} = P\mathbf{g}'$$

If this is substituted in $\mathbf{f}' = A\mathbf{f}$, the result is $P\mathbf{g}' = AP\mathbf{g}$, whence

$$\mathbf{g}' = P^{-1}AP\mathbf{g}$$

But this means that

$$\begin{bmatrix} g_1' \\ g_2' \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}, \quad \text{so} \quad \begin{aligned} g_1' &= 4g_1 \\ g_2' &= -g_2 \end{aligned}$$

Hence Theorem 3.7.1 gives $g_1(x) = ce^{4x}$, $g_2(x) = de^{-x}$, where c and d are constants. Finally, then,

$$\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = P \begin{bmatrix} g_1(x) \\ g_2(x) \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} ce^{4x} \\ de^{-x} \end{bmatrix} = \begin{bmatrix} ce^{4x} + 3de^{-x} \\ ce^{4x} - 2de^{-x} \end{bmatrix}$$

so the *general solution* is

$$\begin{array}{rcl} f_1(x) & = & ce^{4x} + 3de^{-x} \\ f_2(x) & = & ce^{4x} - 2de^{-x} \end{array} \quad c \text{ and } d \text{ constants}$$

It is worth observing that this can be written in matrix form as

$$\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = c \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{4x} + d \begin{bmatrix} 3 \\ -2 \end{bmatrix} e^{-x}$$

That is,

$$\mathbf{f}(x) = c\mathbf{x}_1 e^{4x} + d\mathbf{x}_2 e^{-x}$$

This form of the solution works more generally, as will be shown.
Finally, the requirement that $f_1(0) = 0$ and $f_2(0) = 5$ in this example determines the constants c and d:

$$0 = f_1(0) = ce^0 + 3de^0 = c + 3d$$
$$5 = f_2(0) = ce^0 - 2de^0 = c - 2d$$

These equations give $c = 3$ and $d = -1$, so

$$f_1(x) = 3e^{4x} - 3e^{-x}$$
$$f_2(x) = 3e^{4x} + 2e^{-x}$$

satisfy all the requirements.

The technique in this example works in general.

Theorem 3.7.2

Consider a linear system

$$\mathbf{f}' = A\mathbf{f}$$

of differential equations, where A is an $n \times n$ diagonalizable matrix. Let $P^{-1}AP$ be diagonal, where P is given in terms of its columns

$$P = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$$

and $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ are eigenvectors of A. If $\mathbf{x}_i$ corresponds to the eigenvalue λ_i for each i, then every solution $\mathbf{f}$ of $\mathbf{f}' = A\mathbf{f}$ has the form

$$\mathbf{f}(x) = c_1\mathbf{x}_1 e^{\lambda_1 x} + c_2\mathbf{x}_2 e^{\lambda_2 x} + \cdots + c_n\mathbf{x}_n e^{\lambda_n x}$$

where $c_1, c_2, \ldots, c_n$ are arbitrary constants.

Proof. By Theorem 3.4.1, the matrix $P = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{bmatrix}$ is invertible and

$$P^{-1}AP = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

As in Example 3.7.2, write $\mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}$ and define $\mathbf{g} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{bmatrix}$ by $\mathbf{g} = P^{-1}\mathbf{f}$; equivalently, $\mathbf{f} = P\mathbf{g}$. If $P = \begin{bmatrix} p_{ij} \end{bmatrix}$, this gives

$$f_i = p_{i1}g_1 + p_{i2}g_2 + \cdots + p_{in}g_n$$

Since the p_{ij} are constants, differentiation preserves this relationship:

$$f_i' = p_{i1}g_1' + p_{i2}g_2' + \cdots + p_{in}g_n'$$

so $\mathbf{f}' = P\mathbf{g}'$. Substituting this into $\mathbf{f}' = A\mathbf{f}$ gives $P\mathbf{g}' = AP\mathbf{g}$. But then left multiplication by P^{-1} gives $\mathbf{g}' = P^{-1}AP\mathbf{g}$, so the original system of equations $\mathbf{f}' = A\mathbf{f}$ for $\mathbf{f}$ becomes much simpler in terms of $\mathbf{g}$:

$$\begin{bmatrix} g_1' \\ g_2' \\ \vdots \\ g_n' \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{bmatrix}$$

Hence $g_i' = \lambda_i g_i$ holds for each i, and Theorem 3.7.1 implies that the only solutions are

$$g_i(x) = c_i e^{\lambda_i x} \quad c_i \text{ some constant}$$

Then the relationship $\mathbf{f} = P\mathbf{g}$ gives the functions $f_1, f_2, \dots, f_n$ as follows:

$$\mathbf{f}(x) = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \begin{bmatrix} c_1 e^{\lambda_1 x} \\ c_2 e^{\lambda_2 x} \\ \vdots \\ c_n e^{\lambda_n x} \end{bmatrix} = c_1\mathbf{x}_1 e^{\lambda_1 x} + c_2\mathbf{x}_2 e^{\lambda_2 x} + \cdots + c_n\mathbf{x}_n e^{\lambda_n x}$$

This is what we wanted. □

The theorem shows that *every* solution to $\mathbf{f}' = A\mathbf{f}$ is a linear combination

$$\mathbf{f}(x) = c_1\mathbf{x}_1 e^{\lambda_1 x} + c_2\mathbf{x}_2 e^{\lambda_2 x} + \cdots + c_n\mathbf{x}_n e^{\lambda_n x}$$

where the coefficients c_i are arbitrary. Hence this is called the **general solution** to the system of differential equations. In most cases the solution functions $f_i(x)$ are required to satisfy boundary conditions, often of the form $f_i(a) = b_i$, where $a, b_1, \dots, b_n$ are prescribed numbers. These conditions determine the constants c_i. The following example illustrates this and displays a situation where one eigenvalue has multiplicity greater than 1.

Example 3.7.3

Find the general solution to the system

$$\begin{aligned} f_1' &= 5f_1 + 8f_2 + 16f_3 \\ f_2' &= 4f_1 + f_2 + 8f_3 \\ f_3' &= -4f_1 - 4f_2 - 11f_3 \end{aligned}$$

Then find a solution satisfying the boundary conditions $f_1(0) = f_2(0) = f_3(0) = 1$.

Solution. The system has the form $\mathbf{f}' = A\mathbf{f}$, where $A = \left[\begin{array}{rrr} 5 & 8 & 16 \\ 4 & 1 & 8 \\ -4 & -4 & -11 \end{array} \right]$. In this case $c_A(x) = (x+3)^2(x-1)$ and eigenvectors corresponding to the eigenvalues -3, -3, and 1 are, respectively,

$$\mathbf{x}_1 = \left[\begin{array}{r} -1 \\ 1 \\ 0 \end{array} \right] \quad \mathbf{x}_2 = \left[\begin{array}{r} -2 \\ 0 \\ 1 \end{array} \right] \quad \mathbf{x}_3 = \left[\begin{array}{r} 2 \\ 1 \\ -1 \end{array} \right]$$

Hence, by Theorem 3.7.2, the general solution is

$$\mathbf{f}(x) = c_1 \left[\begin{array}{r} -1 \\ 1 \\ 0 \end{array} \right] e^{-3x} + c_2 \left[\begin{array}{r} -2 \\ 0 \\ 1 \end{array} \right] e^{-3x} + c_3 \left[\begin{array}{r} 2 \\ 1 \\ -1 \end{array} \right] e^{x}, \quad c_i \text{ constants.}$$

The boundary conditions $f_1(0) = f_2(0) = f_3(0) = 1$ determine the constants c_i.

$$\begin{aligned} \left[\begin{array}{r} 1 \\ 1 \\ 1 \end{array} \right] = \mathbf{f}(0) &= c_1 \left[\begin{array}{r} -1 \\ 1 \\ 0 \end{array} \right] + c_2 \left[\begin{array}{r} -2 \\ 0 \\ 1 \end{array} \right] + c_3 \left[\begin{array}{r} 2 \\ 1 \\ -1 \end{array} \right] \\ &= \left[\begin{array}{rrr} -1 & -2 & 2 \\ 1 & 0 & 1 \\ 0 & 1 & -1 \end{array} \right] \left[\begin{array}{c} c_1 \\ c_2 \\ c_3 \end{array} \right] \end{aligned}$$

The solution is $c_1 = -3$, $c_2 = 5$, $c_3 = 4$, so the required specific solution is

$$\begin{aligned} f_1(x) &= -7e^{-3x} + 8e^{x} \\ f_2(x) &= -3e^{-3x} + 4e^{x} \\ f_3(x) &= 5e^{-3x} - 4e^{x} \end{aligned}$$

3.8 Proof of the Cofactor Expansion Theorem

Recall that our definition of the term *determinant* is inductive: The determinant of any 1×1 matrix is defined first; then it is used to define the determinants of 2×2 matrices. Then that is used for the 3×3 case, and so on. The case of a 1×1 matrix $[a]$ poses no problem. We simply define

$$\det [a] = a$$

as in Section 3.1. Given an $n \times n$ matrix A, define A_{ij} to be the $(n-1) \times (n-1)$ matrix obtained from A by deleting row i and column j. Now assume that the determinant of any $(n-1) \times (n-1)$ matrix has been defined. Then the determinant of A is *defined* to be

$$\begin{aligned}\det A &= a_{11} \det A_{11} - a_{21} \det A_{21} + \cdots + (-1)^{n+1} a_{n1} \det A_{n1} \\ &= \sum_{i=1}^{n} (-1)^{i+1} a_{i1} \det A_{i1}\end{aligned}$$

where summation notation has been introduced for convenience.[16] Observe that, in the terminology of Section 3.1, this is just the cofactor expansion of $\det A$ along the first column, and that $(-1)^{i+j} \det A_{ij}$ is the $(i,\ j)$-cofactor (previously denoted as $c_{ij}(A)$).[17] To illustrate the definition, consider the 2×2 matrix

[16] Summation notation is a convenient shorthand way to write sums of similar expressions. For example $a_1 + a_2 + a_3 + a_4 = \sum_{i=1}^{4} a_i$, $a_5 b_5 + a_6 b_6 + a_7 b_7 + a_8 b_8 = \sum_{k=5}^{8} a_k b_k$, and $1^2 + 2^2 + 3^2 + 4^2 + 5^2 = \sum_{j=1}^{5} j^2$.

[17] Note that we used the expansion along *row* 1 at the beginning of Section 3.1. The column 1 expansion definition is more convenient here.

$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$. Then the definition gives

$$\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11} \det [a_{22}] - a_{21} \det [a_{12}] = a_{11}a_{22} - a_{21}a_{12}$$

and this is the same as the definition in Section 3.1.

Of course, the task now is to use this definition to *prove* that the cofactor expansion along *any* row or column yields $\det A$ (this is Theorem 3.1.1). The proof proceeds by first establishing the properties of determinants stated in Theorem 3.1.2 but for *rows* only (see Lemma 3.8.2). This being done, the full proof of Theorem 3.1.1 is not difficult. The proof of Lemma 3.8.2 requires the following preliminary result.

Lemma 3.8.1

Let A, B, and C be $n \times n$ matrices that are identical except that the pth row of A is the sum of the pth rows of B and C. Then

$$\det A = \det B + \det C$$

Proof. We proceed by induction on n, the cases $n = 1$ and $n = 2$ being easily checked. Consider a_{i1} and A_{i1}:

Case 1: If $i \neq p$,

$$a_{i1} = b_{i1} = c_{i1} \quad \text{and} \quad \det A_{i1} = \det B_{i1} = \det C_{i1}$$

by induction because A_{i1}, B_{i1}, C_{i1} are identical except that one row of A_{i1} is the sum of the corresponding rows of B_{i1} and C_{i1}.

Case 2: If $i = p$,

$$a_{p1} = b_{p1} + c_{p1} \quad \text{and} \quad A_{p1} = B_{p1} = C_{p1}$$

Now write out the defining sum for $\det A$, splitting off the pth term for special attention.

$$\begin{aligned} \det A &= \sum_{i \neq p} a_{i1}(-1)^{i+1} \det A_{i1} + a_{p1}(-1)^{p+1} \det A_{p1} \\ &= \sum_{i \neq p} a_{i1}(-1)^{i+1} [\det B_{i1} + \det B_{i1}] + (b_{p1} + c_{p1})(-1)^{p+1} \det A_{p1} \end{aligned}$$

where $\det A_{i1} = \det B_{i1} + \det C_{i1}$ by induction. But the terms here involving B_{i1} and b_{p1} add up to $\det B$ because $a_{i1} = b_{i1}$ if $i \neq p$ and $A_{p1} = B_{p1}$. Similarly, the terms involving C_{i1} and c_{p1} add up to $\det C$. Hence $\det A = \det B + \det C$, as required. □

Lemma 3.8.2

Let $A = [a_{ij}]$ denote an $n \times n$ matrix.

1. *If $B = [b_{ij}]$ is formed from A by multiplying a row of A by a number u, then $\det B = u \det A$.*
2. *If A contains a row of zeros, then $\det A = 0$.*
3. *If $B = [b_{ij}]$ is formed by interchanging two rows of A, then $\det B = -\det A$.*
4. *If A contains two identical rows, then $\det A = 0$.*

5. *If* $B = [b_{ij}]$ *is formed by adding a multiple of one row of* A *to a different row, then* $\det B = \det A$.

Proof. For later reference the defining sums for $\det A$ and $\det B$ are as follows:

$$\det A = \sum_{i=1}^{n} a_{i1}(-1)^{i+1} \det A_{i1} \tag{3.13}$$

$$\det B = \sum_{i=1}^{n} b_{i1}(-1)^{i+1} \det B_{i1} \tag{3.14}$$

Property 1. The proof is by induction on n, the cases $n = 1$ and $n = 2$ being easily verified. Consider the ith term in the sum 3.14 for $\det B$ where B is the result of multiplying row p of A by u.

a. If $i \neq p$, then $b_{i1} = a_{i1}$ and $\det B_{i1} = u \det A_{i1}$ by induction because B_{i1} comes from A_{i1} by multiplying a row by u.

b. If $i = p$, then $b_{p1} = ua_{p1}$ and $B_{p1} = A_{p1}$.

In either case, each term in Equation 3.14 is u times the corresponding term in Equation 3.13, so it is clear that $\det B = u \det A$.

Property 2. This is clear by property 1 because the row of zeros has a common factor $u = 0$.

Property 3. Observe first that it suffices to prove property 3 for interchanges of adjacent rows. (Rows p and q $(q > p)$ can be interchanged by carrying out $2(q-p)-1$ adjacent changes, which results in an *odd* number of sign changes in the determinant.) So suppose that rows p and $p+1$ of A are interchanged to obtain B. Again consider the ith term in Equation 3.14.

a. If $i \neq p$ and $i \neq p+1$, then $b_{i1} = a_{i1}$ and $\det B_{i1} = -\det A_{i1}$ by induction because B_{i1} results from interchanging adjacent rows in A_{i1}. Hence the ith term in Equation 3.14 is the negative of the ith term in Equation 3.13. Hence $\det B = -\det A$ in this case.

b. If $i = p$ or $i = p+1$, then $b_{p1} = a_{p+1,\,1}$ and $B_{p1} = A_{p+1,\,1}$, whereas $b_{p+1,\,1} = a_{p1}$ and $B_{p+1,\,1} = A_{p1}$. Hence terms p and $p+1$ in Equation 3.14 are

$$b_{p1}(-1)^{p+1} \det B_{p1} = -a_{p+1,\,1}(-1)^{(p+1)+1} \det (A_{p+1,\,1})$$

$$b_{p+1,\,1}(-1)^{(p+1)+1} \det B_{p+1,\,1} = -a_{p1}(-1)^{p+1} \det (A_{p1})$$

This means that terms p and $p+1$ in Equation 3.14 are the same as these terms in Equation 3.13, except that the order is reversed and the signs are changed. Thus the sum 3.14 is the negative of the sum 3.13; that is, $\det B = -\det A$.

Property 4. If rows p and q in A are identical, let B be obtained from A by interchanging these rows. Then $B = A$ so $\det A = \det B$. But $\det B = -\det A$ by property 3 so $\det A = -\det A$. This implies that $\det A = 0$.

Property 5. Suppose B results from adding u times row q of A to row p. Then Lemma 3.8.1 applies to B to show that $\det B = \det A + \det C$, where C is obtained from A by replacing row p by u times row q. It now follows from properties 1 and 4 that $\det C = 0$ so $\det B = \det A$, as asserted. □

These facts are enough to enable us to prove Theorem 3.1.1. For convenience, it is restated here in the notation of the foregoing lemmas. The only difference between the notations is that the $(i,\ j)$-cofactor of an $n \times n$ matrix A was denoted earlier by

$$c_{ij}(A) = (-1)^{i+j} \det A_{ij}$$

Theorem 3.8.1

If $A = \left[a_{ij}\right]$ is an $n \times n$ matrix, then

1. $\det A = \sum_{i=1}^{n} a_{ij}(-1)^{i+j} \det A_{ij}$ *(cofactor expansion along column j).*
2. $\det A = \sum_{j=1}^{n} a_{ij}(-1)^{i+j} \det A_{ij}$ *(cofactor expansion along row i).*

Here A_{ij} denotes the matrix obtained from A by deleting row i and column j.

Proof. Lemma 3.8.2 establishes the truth of Theorem 3.1.2 for *rows*. With this information, the arguments in Section 3.2 proceed exactly as written to establish that $\det A = \det A^T$ holds for any $n \times n$ matrix A. Now suppose B is obtained from A by interchanging two columns. Then B^T is obtained from A^T by interchanging two rows so, by property 3 of Lemma 3.8.2,

$$\det B = \det B^T = -\det A^T = -\det A$$

Hence property 3 of Lemma 3.8.2 holds for *columns* too.

This enables us to prove the cofactor expansion for columns. Given an $n \times n$ matrix $A = \left[a_{ij}\right]$, let $B = \left[b_{ij}\right]$ be obtained by moving column j to the left side, using $j-1$ interchanges of adjacent columns. Then $\det B = (-1)^{j-1} \det A$ and, because $B_{i1} = A_{ij}$ and $b_{i1} = a_{ij}$ for all i, we obtain

$$\begin{aligned} \det A &= (-1)^{j-1} \det B = (-1)^{j-1} \sum_{i=1}^{n} b_{i1}(-1)^{i+1} \det B_{i1} \\ &= \sum_{i=1}^{n} a_{ij}(-1)^{i+j} \det A_{ij} \end{aligned}$$

This is the cofactor expansion of $\det A$ along column j.

Finally, to prove the row expansion, write $B = A^T$. Then $B_{ij} = (A_{ij}^T)$ and $b_{ij} = a_{ji}$ for all i and j. Expanding $\det B$ along column j gives

$$\begin{aligned} \det A = \det A^T = \det B &= \sum_{i=1}^{n} b_{ij}(-1)^{i+j} \det B_{ij} \\ &= \sum_{i=1}^{n} a_{ji}(-1)^{j+i} \det\left[(A_{ji}^T)\right] = \sum_{i=1}^{n} a_{ji}(-1)^{j+i} \det A_{ji} \end{aligned}$$

This is the required expansion of $\det A$ along row j. □

Chapter 4

Vector Geometry

4.1 Vectors and Lines

In this chapter we study the geometry of 3-dimensional space. We view a point in 3-space as an arrow from the origin to that point. Doing so provides a "picture" of the point that is truly worth a thousand words. We used this idea earlier, in Section 2.6, to describe rotations, reflections, and projections of the plane $\mathbb{R}^2$. We now apply the same techniques to 3-space to examine similar transformations of $\mathbb{R}^3$. Moreover, the method enables us to completely describe all lines and planes in space.

Vectors in $\mathbb{R}^3$

Introduce a coordinate system in 3-dimensional space in the usual way. First choose a point O called the *origin*, then choose three mutually perpendicular lines through O, called the x, y, and z *axes*, and establish a number scale on each axis with zero at the origin. Given a point P in 3-space we associate three numbers x, y, and z with P, as described in Figure 4.1.1. These numbers are called the *coordinates* of P, and we denote the point as (x, y, z), or $P(x, y, z)$ to emphasize the label P. The result is called a *cartesian*[1] coordinate system for 3-space, and the resulting description of 3-space is called *cartesian geometry*.

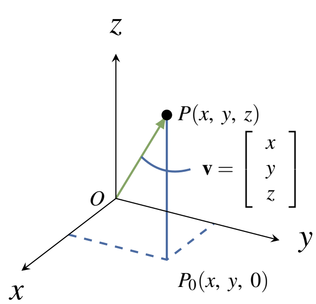


Figure 4.1.1

As in the plane, we introduce vectors by identifying each point $P(x, y, z)$ with the vector $\mathbf{v} = \left[\begin{array}{c} x \\ y \\ z \end{array}\right]$ in $\mathbb{R}^3$, represented by the **arrow** from the origin to P as in Figure 4.1.1. Informally, we say that the point P *has vector* $\mathbf{v}$, and that vector $\mathbf{v}$ *has point* P. In this way 3-space is identified with $\mathbb{R}^3$, and this identification will be made throughout this chapter, often without comment. In particular, the terms "vector" and "point" are interchangeable.[2] The resulting description of 3-space is called **vector geometry**. Note that the origin is $\mathbf{0} = \left[\begin{array}{c} 0 \\ 0 \\ 0 \end{array}\right]$.

Length and Direction

We are going to discuss two fundamental geometric properties of vectors in $\mathbb{R}^3$: length and direction. First, if $\mathbf{v}$ is a vector with point P, the **length** $\|\mathbf{v}\|$ of vector $\mathbf{v}$ is defined to be the distance from the origin to P, that is the length of the arrow representing $\mathbf{v}$. The following properties of length will be used frequently.

[1]Named after René Descartes who introduced the idea in 1637.

[2]Recall that we defined $\mathbb{R}^n$ as the set of all ordered n-tuples of real numbers, and reserved the right to denote them as rows or as columns.

Theorem 4.1.1

Let $\mathbf{v} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ *be a vector.*

1. $\|\mathbf{v}\| = \sqrt{x^2+y^2+z^2}$.[3]
2. $\mathbf{v} = \mathbf{0}$ *if and only if* $\|\mathbf{v}\| = 0$
3. $\|a\mathbf{v}\| = |a|\|\mathbf{v}\|$ *for all scalars* a.[4]

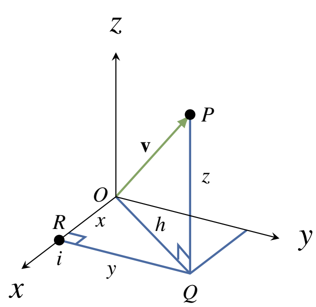


Figure 4.1.2

Proof. Let $\mathbf{v}$ have point $P(x, y, z)$.

1. In Figure 4.1.2, $\|\mathbf{v}\|$ is the hypotenuse of the right triangle OQP, and so $\|\mathbf{v}\|^2 = h^2+z^2$ by Pythagoras' theorem.[5] But h is the hypotenuse of the right triangle ORQ, so $h^2 = x^2+y^2$. Now (1) follows by eliminating h^2 and taking positive square roots.

2. If $\|\mathbf{v}\| = 0$, then $x^2+y^2+z^2 = 0$ by (1). Because squares of real numbers are nonnegative, it follows that $x = y = z = 0$, and hence that $\mathbf{v} = \mathbf{0}$. The converse is because $\|\mathbf{0}\| = 0$.

3. We have $a\mathbf{v} = \begin{bmatrix} ax & ay & az \end{bmatrix}^T$ so (1) gives

$$\|a\mathbf{v}\|^2 = (ax)^2+(ay)^2+(az)^2 = a^2\|\mathbf{v}\|^2$$

Hence $\|a\mathbf{v}\| = \sqrt{a^2}\|\mathbf{v}\|$, and we are done because $\sqrt{a^2} = |a|$ for any real number a. □

Of course the $\mathbb{R}^2$-version of Theorem 4.1.1 also holds.

Example 4.1.1

If $\mathbf{v} = \begin{bmatrix} 2 \\ -1 \\ 3 \end{bmatrix}$ then $\|\mathbf{v}\| = \sqrt{4+1+9} = \sqrt{14}$. Similarly if $\mathbf{v} = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$ in 2-space then $\|\mathbf{v}\| = \sqrt{9+16} = 5$.

When we view two nonzero vectors as arrows emanating from the origin, it is clear geometrically what we mean by saying that they have the same or opposite **direction**. This leads to a fundamental new description of vectors.

[3]When we write $\sqrt{p}$ we mean the positive square root of p.

[4]Recall that the absolute value $|a|$ of a real number is defined by $|a| = \begin{cases} a \text{ if } a \geq 0 \\ -a \text{ if } a < 0 \end{cases}$.

[5]Pythagoras' theorem states that if a and b are sides of right triangle with hypotenuse c, then $a^2+b^2=c^2$. A proof is given at the end of this section.

Theorem 4.1.2

Let $\mathbf{v} \neq \mathbf{0}$ and $\mathbf{w} \neq \mathbf{0}$ be vectors in $\mathbb{R}^3$. Then $\mathbf{v} = \mathbf{w}$ as matrices if and only if $\mathbf{v}$ and $\mathbf{w}$ have the same direction and the same length.[6]

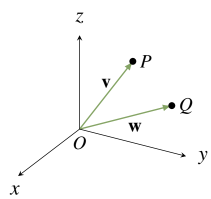


Figure 4.1.3

Proof. If $\mathbf{v} = \mathbf{w}$, they clearly have the same direction and length. Conversely, let $\mathbf{v}$ and $\mathbf{w}$ be vectors with points $P(x, y, z)$ and $Q(x_1, y_1, z_1)$ respectively. If $\mathbf{v}$ and $\mathbf{w}$ have the same length and direction then, geometrically, P and Q must be the same point (see Figure 4.1.3). Hence $x = x_1$, $y = y_1$, and $z = z_1$, that is

$$\mathbf{v} = \left[\begin{array}{c} x \\ y \\ z \end{array}\right] = \left[\begin{array}{c} x_1 \\ y_1 \\ z_1 \end{array}\right] = \mathbf{w}. \qquad \square$$

A characterization of a vector in terms of its length and direction only is called an **intrinsic** description of the vector. The point to note is that such a description does *not* depend on the choice of coordinate system in $\mathbb{R}^3$. Such descriptions are important in applications because physical laws are often stated in terms of vectors, and these laws cannot depend on the particular coordinate system used to describe the situation.

Geometric Vectors

If A and B are distinct points in space, the arrow from A to B has length and direction.

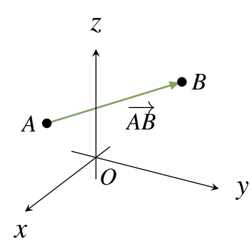


Figure 4.1.4

Hence:

Definition 4.1 Geometric Vectors

*Suppose that A and B are any two points in $\mathbb{R}^3$. In Figure 4.1.4 the line segment from A to B is denoted $\overrightarrow{AB}$ and is called the **geometric vector** from A to B. Point A is called the **tail** of $\overrightarrow{AB}$, B is called the **tip** of $\overrightarrow{AB}$, and the **length** of $\overrightarrow{AB}$ is denoted $\|\overrightarrow{AB}\|$.*

Note that if $\mathbf{v}$ is any vector in $\mathbb{R}^3$ with point P then $\mathbf{v} = \overrightarrow{OP}$ is itself a geometric vector where O is the origin. Referring to $\overrightarrow{AB}$ as a "vector" seems justified by Theorem 4.1.2 because it has a direction (from A to B) and a length $\|\overrightarrow{AB}\|$.

[6] It is Theorem 4.1.2 that gives vectors their power in science and engineering because many physical quantities are determined by their length and magnitude (and are called **vector quantities**). For example, saying that an airplane is flying at 200 km/h does not describe where it is going; the direction must also be specified. The speed and direction comprise the **velocity** of the airplane, a vector quantity.

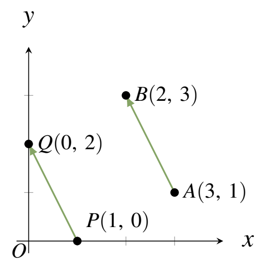


Figure 4.1.5

However there appears to be a problem because two geometric vectors can have the same length and direction even if the tips and tails are different. For example $\overrightarrow{AB}$ and $\overrightarrow{PQ}$ in Figure 4.1.5 have the same length $\sqrt{5}$ and the same direction (1 unit left and 2 units up) so, by Theorem 4.1.2, they are the same vector! The best way to understand this apparent paradox is to see $\overrightarrow{AB}$ and $\overrightarrow{PQ}$ as different *representations* of the same[7] underlying vector $\left[\begin{array}{r} -1 \\ 2 \end{array}\right]$. Once it is clarified, this phenomenon is a great benefit because, thanks to Theorem 4.1.2, it means that the same geometric vector can be positioned anywhere in space; what is important is the length and direction, not the location of the tip and tail. This ability to move geometric vectors about is very useful as we shall soon see.

The Parallelogram Law

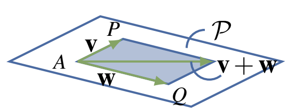


Figure 4.1.6

We now give an intrinsic description of the sum of two vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^3$, that is a description that depends only on the lengths and directions of $\mathbf{v}$ and $\mathbf{w}$ and not on the choice of coordinate system. Using Theorem 4.1.2 we can think of these vectors as having a common tail A. If their tips are P and Q respectively, then they both lie in a plane $\mathcal{P}$ containing A, P, and Q, as shown in Figure 4.1.6. The vectors $\mathbf{v}$ and $\mathbf{w}$ create a parallelogram[8] in $\mathcal{P}$, shaded in Figure 4.1.6, called the parallelogram **determined** by $\mathbf{v}$ and $\mathbf{w}$.

If we now choose a coordinate system in the plane $\mathcal{P}$ with A as origin, then the parallelogram law in the plane (Section 2.6) shows that their sum $\mathbf{v}+\mathbf{w}$ is the diagonal of the parallelogram they determine with tail A. This is an intrinsic description of the sum $\mathbf{v}+\mathbf{w}$ because it makes no reference to coordinates. This discussion proves:

Theorem: The Parallelogram Law

In the parallelogram determined by two vectors $\mathbf{v}$ and $\mathbf{w}$, the vector $\mathbf{v}+\mathbf{w}$ is the diagonal with the same tail as $\mathbf{v}$ and $\mathbf{w}$.

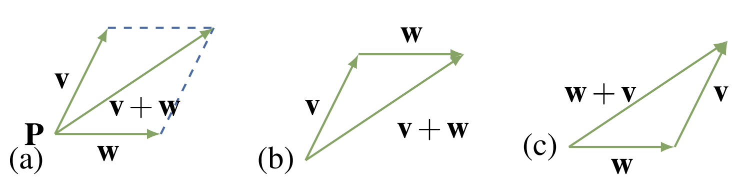


Figure 4.1.7

Because a vector can be positioned with its tail at any point, the parallelogram law leads to another way to view vector addition. In Figure 4.1.7(a) the sum $\mathbf{v}+\mathbf{w}$ of two vectors $\mathbf{v}$ and $\mathbf{w}$ is shown as given by the parallelogram law. If $\mathbf{w}$ is moved so its tail coincides with the tip of $\mathbf{v}$ (Figure 4.1.7(b)) then the sum $\mathbf{v}+\mathbf{w}$ is seen as "first $\mathbf{v}$ and then $\mathbf{w}$. Similarly, moving the tail of $\mathbf{v}$ to the tip of $\mathbf{w}$ shows in Figure 4.1.7(c) that $\mathbf{v}+\mathbf{w}$ is "first $\mathbf{w}$ and then $\mathbf{v}$." This will be referred to as the

[7]Fractions provide another example of quantities that can be the same but *look* different. For example $\frac{6}{9}$ and $\frac{14}{21}$ certainly appear different, but they are equal fractions—both equal $\frac{2}{3}$ in "lowest terms".

[8]Recall that a parallelogram is a four-sided figure whose opposite sides are parallel and of equal length.

tip-to-tail rule, and it gives a graphic illustration of why $\mathbf{v}+\mathbf{w}=\mathbf{w}+\mathbf{v}$.

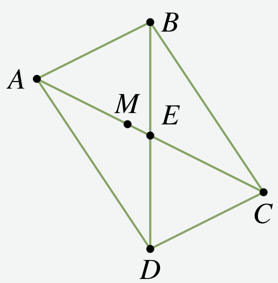
Since $\overrightarrow{AB}$ denotes the vector from a point A to a point B, the tip-to-tail rule takes the easily remembered form

$$\overrightarrow{AB}+\overrightarrow{BC}=\overrightarrow{AC}$$

for any points A, B, and C. The next example uses this to derive a theorem in geometry without using coordinates.

Example 4.1.2

Show that the diagonals of a parallelogram bisect each other.



Solution. Let the parallelogram have vertices A, B, C, and D, as shown; let E denote the intersection of the two diagonals; and let M denote the midpoint of diagonal AC. We must show that $M=E$ and that this is the midpoint of diagonal BD. This is accomplished by showing that $\overrightarrow{BM}=\overrightarrow{MD}$. (Then the fact that these vectors have the same direction means that $M=E$, and the fact that they have the same length means that $M=E$ is the midpoint of BD.) Now $\overrightarrow{AM}=\overrightarrow{MC}$ because M is the midpoint of AC, and $\overrightarrow{BA}=\overrightarrow{CD}$ because the figure is a parallelogram. Hence

$$\overrightarrow{BM}=\overrightarrow{BA}+\overrightarrow{AM}=\overrightarrow{CD}+\overrightarrow{MC}=\overrightarrow{MC}+\overrightarrow{CD}=\overrightarrow{MD}$$

where the first and last equalities use the tip-to-tail rule of vector addition.

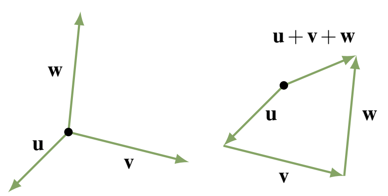


Figure 4.1.8

One reason for the importance of the tip-to-tail rule is that it means two or more vectors can be added by placing them tip-to-tail in sequence. This gives a useful "picture" of the sum of several vectors, and is illustrated for three vectors in Figure 4.1.8 where $\mathbf{u}+\mathbf{v}+\mathbf{w}$ is viewed as first $\mathbf{u}$, then $\mathbf{v}$, then $\mathbf{w}$.

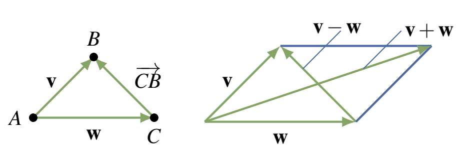


Figure 4.1.9

There is also a simple geometrical way to visualize the (matrix) **difference** $\mathbf{v}-\mathbf{w}$ of two vectors. If $\mathbf{v}$ and $\mathbf{w}$ are positioned so that they have a common tail A (see Figure 4.1.9), and if B and C are their respective tips, then the tip-to-tail rule gives $\mathbf{w}+\overrightarrow{CB}=\mathbf{v}$. Hence $\mathbf{v}-\mathbf{w}=\overrightarrow{CB}$ is the vector from the tip of $\mathbf{w}$ to the tip of $\mathbf{v}$. Thus both $\mathbf{v}-\mathbf{w}$ and $\mathbf{v}+\mathbf{w}$ appear as diagonals in the parallelogram determined by $\mathbf{v}$ and $\mathbf{w}$ (see Figure 4.1.9). We record this for reference.

Theorem 4.1.3

If $\mathbf{v}$ and $\mathbf{w}$ have a common tail, then $\mathbf{v}-\mathbf{w}$ is the vector from the tip of $\mathbf{w}$ to the tip of $\mathbf{v}$.

One of the most useful applications of vector subtraction is that it gives a simple formula for the vector from one point to another, and for the distance between the points.

Theorem 4.1.4

Let $P_1(x_1, y_1, z_1)$ and $P_2(x_2, y_2, z_2)$ be two points. Then:

1. $\overrightarrow{P_1P_2} = \left[\begin{array}{c} x_2 - x_1 \\ y_2 - y_1 \\ z_2 - z_1 \end{array} \right]$.

2. *The distance between P_1 and P_2 is* $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$.

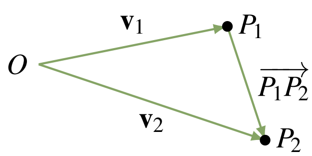


Figure 4.1.10

Proof. If O is the origin, write

$$\mathbf{v}_1 = \overrightarrow{OP}_1 = \left[\begin{array}{c} x_1 \\ y_1 \\ z_1 \end{array} \right] \text{ and } \mathbf{v}_2 = \overrightarrow{OP}_2 = \left[\begin{array}{c} x_2 \\ y_2 \\ z_2 \end{array} \right]$$

as in Figure 4.1.10.

Then Theorem 4.1.3 gives $\overrightarrow{P_1P_2} = \mathbf{v}_2 - \mathbf{v}_1$, and (1) follows. But the distance between P_1 and P_2 is $\|\overrightarrow{P_1P_2}\|$, so (2) follows from (1) and Theorem 4.1.1. □

Of course the $\mathbb{R}^2$-version of Theorem 4.1.4 is also valid: If $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ are points in $\mathbb{R}^2$, then $\overrightarrow{P_1P_2} = \left[\begin{array}{c} x_2 - x_1 \\ y_2 - y_1 \end{array} \right]$, and the distance between P_1 and P_2 is $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$.

Example 4.1.3

The distance between $P_1(2, -1, 3)$ and $P_2(1, 1, 4)$ is $\sqrt{(-1)^2 + (2)^2 + (1)^2} = \sqrt{6}$, and the vector from P_1 to P_2 is $\overrightarrow{P_1P_2} = \left[\begin{array}{r} -1 \\ 2 \\ 1 \end{array} \right]$.

As for the parallelogram law, the intrinsic rule for finding the length and direction of a scalar multiple of a vector in $\mathbb{R}^3$ follows easily from the same situation in $\mathbb{R}^2$.

Theorem: Scalar Multiple Law

If a is a real number and $\mathbf{v} \neq \mathbf{0}$ is a vector then:

1. *The length of $a\mathbf{v}$ is $\|a\mathbf{v}\| = |a|\|\mathbf{v}\|$.*

2. *If[9] $a\mathbf{v} \neq \mathbf{0}$, the direction of $a\mathbf{v}$ is* $\left\{ \begin{array}{l} \textit{the same as } \mathbf{v} \textit{ if } a > 0, \\ \textit{opposite to } \mathbf{v} \textit{ if } a < 0. \end{array} \right.$

Proof.

1. This is part of Theorem 4.1.1.

2. Let O denote the origin in $\mathbb{R}^3$, let $\mathbf{v}$ have point P, and choose any plane containing O and P. If we set up a coordinate system in this plane with O as origin, then $\mathbf{v} = \overrightarrow{OP}$ so the result in (2) follows from the scalar multiple law in the plane (Section 2.6). □

Figure 4.1.11 gives several examples of scalar multiples of a vector $\mathbf{v}$.

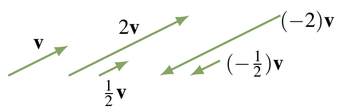


Figure 4.1.11

Consider a line L through the origin, let P be any point on L other than the origin O, and let $\mathbf{p} = \overrightarrow{OP}$. If $t \neq 0$, then $t\mathbf{p}$ is a point on L because it has direction the same or opposite as that of $\mathbf{p}$. Moreover $t > 0$ or $t < 0$ according as the point $t\mathbf{p}$ lies on the same or opposite side of the origin as P. This is illustrated in Figure 4.1.12.

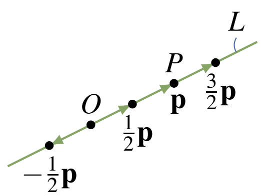


Figure 4.1.12

A vector $\mathbf{u}$ is called a **unit vector** if $\|\mathbf{u}\| = 1$. Then $\mathbf{i} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $\mathbf{j} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, and $\mathbf{k} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ are unit vectors, called the **coordinate** vectors. We discuss them in more detail in Section 4.2.

Example 4.1.4

If $\mathbf{v} \neq \mathbf{0}$ show that $\frac{1}{\|\mathbf{v}\|}\mathbf{v}$ is the unique unit vector in the same direction as $\mathbf{v}$.

Solution. The vectors in the same direction as $\mathbf{v}$ are the scalar multiples $a\mathbf{v}$ where $a > 0$. But $\|a\mathbf{v}\| = |a|\|\mathbf{v}\| = a\|\mathbf{v}\|$ when $a > 0$, so $a\mathbf{v}$ is a unit vector if and only if $a = \frac{1}{\|\mathbf{v}\|}$.

The next example shows how to find the coordinates of a point on the line segment between two given points. The technique is important and will be used again below.

Example 4.1.5

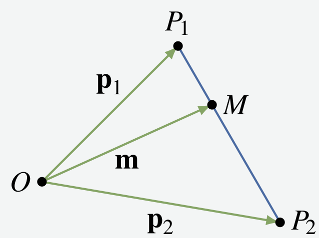
Let $\mathbf{p}_1$ and $\mathbf{p}_2$ be the vectors of two points P_1 and P_2. If M is the point one third the way from P_1 to P_2, show that the vector $\mathbf{m}$ of M is given by

$$\mathbf{m} = \frac{2}{3}\mathbf{p}_1 + \frac{1}{3}\mathbf{p}_2$$

Conclude that if $P_1 = P_1(x_1, y_1, z_1)$ and $P_2 = P_2(x_2, y_2, z_2)$, then M has coordinates

$$M = M\left(\frac{2}{3}x_1 + \frac{1}{3}x_2,\ \frac{2}{3}y_1 + \frac{1}{3}y_2,\ \frac{2}{3}z_1 + \frac{1}{3}z_2\right)$$

[9]Since the zero vector has no direction, we deal only with the case $a\mathbf{v} \neq \mathbf{0}$.



Solution. The vectors $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{m}$ are shown in the diagram. We have $\overrightarrow{P_1M} = \frac{1}{3}\overrightarrow{P_1P_2}$ because $\overrightarrow{P_1M}$ is in the same direction as $\overrightarrow{P_1P_2}$ and $\frac{1}{3}$ as long. By Theorem 4.1.3 we have $\overrightarrow{P_1P_2} = \mathbf{p}_2 - \mathbf{p}_1$, so tip-to-tail addition gives

$$\mathbf{m} = \mathbf{p}_1 + \overrightarrow{P_1M} = \mathbf{p}_1 + \tfrac{1}{3}(\mathbf{p}_2 - \mathbf{p}_1) = \tfrac{2}{3}\mathbf{p}_1 + \tfrac{1}{3}\mathbf{p}_2$$

as required. For the coordinates, we have $\mathbf{p}_1 = \left[\begin{array}{c} x_1 \\ y_1 \\ z_1 \end{array}\right]$ and $\mathbf{p}_2 = \left[\begin{array}{c} x_2 \\ y_2 \\ z_2 \end{array}\right]$, so

$$\mathbf{m} = \tfrac{2}{3}\left[\begin{array}{c} x_1 \\ y_1 \\ z_1 \end{array}\right] + \tfrac{1}{3}\left[\begin{array}{c} x_2 \\ y_2 \\ z_2 \end{array}\right] = \left[\begin{array}{c} \frac{2}{3}x_1 + \frac{1}{3}x_2 \\ \frac{2}{3}y_1 + \frac{1}{3}y_2 \\ \frac{2}{3}z_1 + \frac{1}{3}z_2 \end{array}\right]$$

by matrix addition. The last statement follows.

Note that in Example 4.1.5 $\mathbf{m} = \frac{2}{3}\mathbf{p}_1 + \frac{1}{3}\mathbf{p}_2$ is a "weighted average" of $\mathbf{p}_1$ and $\mathbf{p}_2$ with more weight on $\mathbf{p}_1$ because $\mathbf{m}$ is closer to $\mathbf{p}_1$.

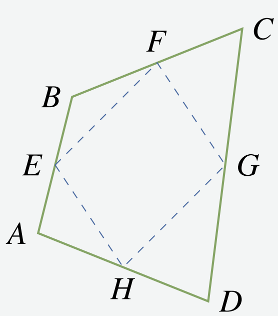
The point M halfway between points P_1 and P_2 is called the **midpoint** between these points. In the same way, the vector $\mathbf{m}$ of M is

$$\mathbf{m} = \tfrac{1}{2}\mathbf{p}_1 + \tfrac{1}{2}\mathbf{p}_2 = \tfrac{1}{2}(\mathbf{p}_1 + \mathbf{p}_2)$$

as the reader can verify, so $\mathbf{m}$ is the "average" of $\mathbf{p}_1$ and $\mathbf{p}_2$ in this case.

Example 4.1.6

Show that the midpoints of the four sides of any quadrilateral are the vertices of a parallelogram. Here a quadrilateral is any figure with four vertices and straight sides.



Solution. Suppose that the vertices of the quadrilateral are A, B, C, and D (in that order) and that E, F, G, and H are the midpoints of the sides as shown in the diagram. It suffices to show $\overrightarrow{EF} = \overrightarrow{HG}$ (because then sides EF and HG are parallel and of equal length). Now the fact that E is the midpoint of AB means that $\overrightarrow{EB} = \frac{1}{2}\overrightarrow{AB}$. Similarly, $\overrightarrow{BF} = \frac{1}{2}\overrightarrow{BC}$, so

$$\overrightarrow{EF} = \overrightarrow{EB} + \overrightarrow{BF} = \tfrac{1}{2}\overrightarrow{AB} + \tfrac{1}{2}\overrightarrow{BC} = \tfrac{1}{2}(\overrightarrow{AB} + \overrightarrow{BC}) = \tfrac{1}{2}\overrightarrow{AC}$$

A similar argument shows that $\overrightarrow{HG} = \frac{1}{2}\overrightarrow{AC}$ too, so $\overrightarrow{EF} = \overrightarrow{HG}$ as required.

Definition 4.2 Parallel Vectors in $\mathbb{R}^3$

Two nonzero vectors are called ***parallel*** *if they have the same or opposite direction.*

Many geometrical propositions involve this notion, so the following theorem will be referred to repeatedly.

Theorem 4.1.5

Two nonzero vectors $\mathbf{v}$ *and* $\mathbf{w}$ *are parallel if and only if one is a scalar multiple of the other.*

Proof. If one of them is a scalar multiple of the other, they are parallel by the scalar multiple law.

Conversely, assume that $\mathbf{v}$ and $\mathbf{w}$ are parallel and write $d = \frac{\|\mathbf{v}\|}{\|\mathbf{w}\|}$ for convenience. Then $\mathbf{v}$ and $\mathbf{w}$ have the same or opposite direction. If they have the same direction we show that $\mathbf{v} = d\mathbf{w}$ by showing that $\mathbf{v}$ and $d\mathbf{w}$ have the same length and direction. In fact, $\|d\mathbf{w}\| = |d|\|\mathbf{w}\| = \|\mathbf{v}\|$ by Theorem 4.1.1; as to the direction, $d\mathbf{w}$ and $\mathbf{w}$ have the same direction because $d > 0$, and this is the direction of $\mathbf{v}$ by assumption. Hence $\mathbf{v} = d\mathbf{w}$ in this case by Theorem 4.1.2. In the other case, $\mathbf{v}$ and $\mathbf{w}$ have opposite direction and a similar argument shows that $\mathbf{v} = -d\mathbf{w}$. We leave the details to the reader. □

Example 4.1.7

Given points $P(2,\ -1,\ 4)$, $Q(3,\ -1,\ 3)$, $A(0,\ 2,\ 1)$, and $B(1,\ 3,\ 0)$, determine if $\overrightarrow{PQ}$ and $\overrightarrow{AB}$ are parallel.

Solution. By Theorem 4.1.3, $\overrightarrow{PQ} = (1,\ 0,\ -1)$ and $\overrightarrow{AB} = (1,\ 1,\ -1)$. If $\overrightarrow{PQ} = t\overrightarrow{AB}$ then $(1,\ 0,\ -1) = (t,\ t,\ -t)$, so $1 = t$ and $0 = t$, which is impossible. Hence $\overrightarrow{PQ}$ is *not* a scalar multiple of $\overrightarrow{AB}$, so these vectors are not parallel by Theorem 4.1.5.

Lines in Space

These vector techniques can be used to give a very simple way of describing straight lines in space. In order to do this, we first need a way to specify the orientation of such a line, much as the slope does in the plane.

Definition 4.3 Direction Vector of a Line

With this in mind, we call a nonzero vector $\mathbf{d} \neq \mathbf{0}$ *a* ***direction vector*** *for the line if it is parallel to* $\overrightarrow{AB}$ *for some pair of distinct points* A *and* B *on the line.*

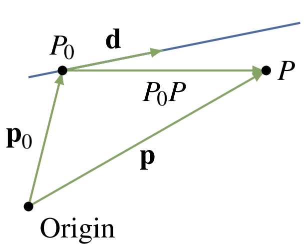


Figure 4.1.13

Of course it is then parallel to $\overrightarrow{CD}$ for *any* distinct points C and D on the line. In particular, any nonzero scalar multiple of $\mathbf{d}$ will also serve as a direction vector of the line.

We use the fact that there is exactly one line that passes through a particular point $P_0(x_0,\ y_0,\ z_0)$ and has a given direction vector $\mathbf{d} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$. We want to describe this line by giving a condition on x, y, and z that the point $P(x,\ y,\ z)$ lies on this line. Let $\mathbf{p}_0 = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix}$ and $\mathbf{p} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ denote the vectors of P_0 and P, respectively (see Figure 4.1.13). Then

$$\mathbf{p} = \mathbf{p}_0 + \overrightarrow{P_0P}$$

Hence P lies on the line if and only if $\overrightarrow{P_0P}$ is parallel to $\mathbf{d}$—that is, if and only if $\overrightarrow{P_0P} = t\mathbf{d}$ for some scalar t by Theorem 4.1.5. Thus $\mathbf{p}$ is the vector of a point on the line if and only if $\mathbf{p} = \mathbf{p}_0 + t\mathbf{d}$ for some scalar t. This discussion is summed up as follows.

Theorem: Vector Equation of a Line

The line parallel to $\mathbf{d} \neq \mathbf{0}$ through the point with vector $\mathbf{p}_0$ is given by

$$\mathbf{p} = \mathbf{p}_0 + t\mathbf{d} \quad t \text{ any scalar}$$

In other words, the point P with vector $\mathbf{p}$ is on this line if and only if a real number t exists such that $\mathbf{p} = \mathbf{p}_0 + t\mathbf{d}$.

In component form the vector equation becomes

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} + t \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

Equating components gives a different description of the line.

Theorem: Parametric Equations of a Line

The line through $P_0(x_0,\ y_0,\ z_0)$ with direction vector $\mathbf{d} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \neq \mathbf{0}$ is given by

$$\begin{aligned} x &= x_0 + ta \\ y &= y_0 + tb \quad t \text{ any scalar} \\ z &= z_0 + tc \end{aligned}$$

In other words, the point $P(x,\ y,\ z)$ is on this line if and only if a real number t exists such that $x = x_0 + ta$, $y = y_0 + tb$, and $z = z_0 + tc$.

Example 4.1.8

Find the equations of the line through the points $P_0(2,\ 0,\ 1)$ and $P_1(4,\ -1,\ 1)$.

Solution. Let $\mathbf{d} = \overrightarrow{P_0P_1} = \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}$ denote the vector from P_0 to P_1. Then $\mathbf{d}$ is parallel to the line (P_0 and P_1 are *on* the line), so $\mathbf{d}$ serves as a direction vector for the line. Using P_0 as the point on the line leads to the parametric equations

$$\begin{aligned} x &= 2 + 2t \\ y &= -t \qquad t \text{ a parameter} \\ z &= 1 \end{aligned}$$

Note that if P_1 is used (rather than P_0), the equations are

$$\begin{aligned} x &= 4 + 2s \\ y &= -1 - s \quad s \text{ a parameter} \\ z &= 1 \end{aligned}$$

These are different from the preceding equations, but this is merely the result of a change of parameter. In fact, $s = t - 1$.

Example 4.1.9

Find the equations of the line through $P_0(3,\ -1,\ 2)$ parallel to the line with equations

$$\begin{aligned} x &= -1 + 2t \\ y &= 1 + t \\ z &= -3 + 4t \end{aligned}$$

Solution. The coefficients of t give a direction vector $\mathbf{d} = \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}$ of the given line. Because the line we seek is parallel to this line, $\mathbf{d}$ also serves as a direction vector for the new line. It passes through P_0, so the parametric equations are

$$\begin{aligned} x &= 3 + 2t \\ y &= -1 + t \\ z &= 2 + 4t \end{aligned}$$

Example 4.1.10

Determine whether the following lines intersect and, if so, find the point of intersection.

$$\begin{aligned} x &= 1 - 3t & \qquad x &= -1 + s \\ y &= 2 + 5t & \qquad y &= 3 - 4s \\ z &= 1 + t & \qquad z &= 1 - s \end{aligned}$$

Solution. Suppose $P(x, y, z)$ with vector $\mathbf{p}$ lies on both lines. Then

$$\begin{bmatrix} 1-3t \\ 2+5t \\ 1+t \end{bmatrix} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -1+s \\ 3-4s \\ 1-s \end{bmatrix} \text{ for some } t \text{ and } s,$$

where the first (second) equation is because P lies on the first (second) line. Hence the lines intersect if and only if the three equations

$$\begin{aligned} 1-3t &= -1+s \\ 2+5t &= 3-4s \\ 1+t &= 1-s \end{aligned}$$

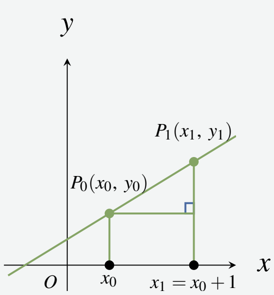
have a solution. In this case, $t = 1$ and $s = -1$ satisfy all three equations, so the lines *do* intersect and the point of intersection is

$$\mathbf{p} = \begin{bmatrix} 1-3t \\ 2+5t \\ 1+t \end{bmatrix} = \begin{bmatrix} -2 \\ 7 \\ 2 \end{bmatrix}$$

using $t = 1$. Of course, this point can also be found from $\mathbf{p} = \begin{bmatrix} -1+s \\ 3-4s \\ 1-s \end{bmatrix}$ using $s = -1$.

Example 4.1.11

Show that the line through $P_0(x_0, y_0)$ with slope m has direction vector $\mathbf{d} = \begin{bmatrix} 1 \\ m \end{bmatrix}$ and equation $y - y_0 = m(x - x_0)$. This equation is called the *point-slope* formula.



Solution. Let $P_1(x_1, y_1)$ be the point on the line one unit to the right of P_0 (see the diagram). Hence $x_1 = x_0 + 1$. Then $\mathbf{d} = \overrightarrow{P_0P_1}$ serves as direction vector of the line, and $\mathbf{d} = \begin{bmatrix} x_1 - x_0 \\ y_1 - y_0 \end{bmatrix} = \begin{bmatrix} 1 \\ y_1 - y_0 \end{bmatrix}$. But the slope m can be computed as follows:

$$m = \frac{y_1 - y_0}{x_1 - x_0} = \frac{y_1 - y_0}{1} = y_1 - y_0$$

Hence $\mathbf{d} = \begin{bmatrix} 1 \\ m \end{bmatrix}$ and the parametric equations are $x = x_0 + t$, $y = y_0 + mt$. Eliminating t gives $y - y_0 = mt = m(x - x_0)$, as asserted.

Note that the vertical line through $P_0(x_0, y_0)$ has a direction vector $\mathbf{d} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ that is *not* of the form $\begin{bmatrix} 1 \\ m \end{bmatrix}$ for any m. This result confirms that the notion of slope makes no sense in this case. However, the vector method gives parametric equations for the line:

$$x = x_0$$
$$y = y_0 + t$$

Because y is arbitrary here (t is arbitrary), this is usually written simply as $x = x_0$.

Pythagoras' Theorem

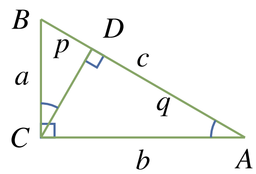


Figure 4.1.14

The Pythagorean theorem was known earlier, but Pythagoras (c. 550 B.C.) is credited with giving the first rigorous, logical, deductive proof of the result. The proof we give depends on a basic property of similar triangles: ratios of corresponding sides are equal.

Theorem 4.1.6: Pythagoras' Theorem

Given a right-angled triangle with hypotenuse c and sides a and b, then $a^2 + b^2 = c^2$.

Proof. Let A, B, and C be the vertices of the triangle as in Figure 4.1.14. Draw a perpendicular line from C to the point D on the hypotenuse, and let p and q be the lengths of BD and DA respectively. Then DBC and CBA are similar triangles so $\frac{p}{a} = \frac{a}{c}$. This means $a^2 = pc$. In the same way, the similarity of DCA and CBA gives $\frac{q}{b} = \frac{b}{c}$, whence $b^2 = qc$. But then

$$a^2 + b^2 = pc + qc = (p+q)c = c^2$$

because $p + q = c$. This proves Pythagoras' theorem[10]. □

[10] There is an intuitive geometrical proof of Pythagoras' theorem in Example B.3.

4.2 Projections and Planes

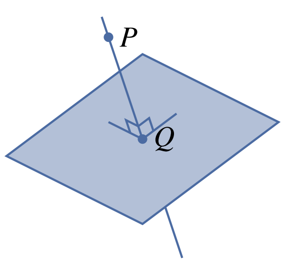


Figure 4.2.1

Any student of geometry soon realizes that the notion of perpendicular lines is fundamental. As an illustration, suppose a point P and a plane are given and it is desired to find the point Q that lies in the plane and is closest to P, as shown in Figure 4.2.1. Clearly, what is required is to find the line through P that is perpendicular to the plane and then to obtain Q as the point of intersection of this line with the plane. Finding the line *perpendicular* to the plane requires a way to determine when two vectors are perpendicular. This can be done using the idea of the dot product of two vectors.

The Dot Product and Angles

Definition 4.4 Dot Product in $\mathbb{R}^3$

Given vectors $\mathbf{v} = \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix}$ *and* $\mathbf{w} = \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix}$, *their* ***dot product*** $\mathbf{v} \cdot \mathbf{w}$ *is a number defined*

$$\mathbf{v} \cdot \mathbf{w} = x_1x_2 + y_1y_2 + z_1z_2 = \mathbf{v}^T \mathbf{w}$$

Because $\mathbf{v}\cdot\mathbf{w}$ is a number, it is sometimes called the **scalar product** of $\mathbf{v}$ and $\mathbf{w}$.[11]

Example 4.2.1

If $\mathbf{v} = \left[\begin{array}{r} 2 \\ -1 \\ 3 \end{array}\right]$ and $\mathbf{w} = \left[\begin{array}{r} 1 \\ 4 \\ -1 \end{array}\right]$, then $\mathbf{v}\cdot\mathbf{w} = 2\cdot 1 + (-1)\cdot 4 + 3\cdot(-1) = -5$.

The next theorem lists several basic properties of the dot product.

Theorem 4.2.1

Let $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ denote vectors in $\mathbb{R}^3$ (or $\mathbb{R}^2$).

1. $\mathbf{v}\cdot\mathbf{w}$ *is a real number.*
2. $\mathbf{v}\cdot\mathbf{w} = \mathbf{w}\cdot\mathbf{v}$.
3. $\mathbf{v}\cdot\mathbf{0} = 0 = \mathbf{0}\cdot\mathbf{v}$.
4. $\mathbf{v}\cdot\mathbf{v} = \|\mathbf{v}\|^2$.
5. $(k\mathbf{v})\cdot\mathbf{w} = k(\mathbf{w}\cdot\mathbf{v}) = \mathbf{v}\cdot(k\mathbf{w})$ *for all scalars k.*
6. $\mathbf{u}\cdot(\mathbf{v}\pm\mathbf{w}) = \mathbf{u}\cdot\mathbf{v}\pm\mathbf{u}\cdot\mathbf{w}$

Proof. (1), (2), and (3) are easily verified, and (4) comes from Theorem 4.1.1. The rest are properties of matrix arithmetic (because $\mathbf{w}\cdot\mathbf{v} = \mathbf{v}^T\mathbf{w}$), and are left to the reader. □

The properties in Theorem 4.2.1 enable us to do calculations like

$$3\mathbf{u}\cdot(2\mathbf{v} - 3\mathbf{w} + 4\mathbf{z}) = 6(\mathbf{u}\cdot\mathbf{v}) - 9(\mathbf{u}\cdot\mathbf{w}) + 12(\mathbf{u}\cdot\mathbf{z})$$

and such computations will be used without comment below. Here is an example.

Example 4.2.2

Verify that $\|\mathbf{v} - 3\mathbf{w}\|^2 = 1$ when $\|\mathbf{v}\| = 2$, $\|\mathbf{w}\| = 1$, and $\mathbf{v}\cdot\mathbf{w} = 2$.

Solution. We apply Theorem 4.2.1 several times:

$$\begin{aligned}
\|\mathbf{v} - 3\mathbf{w}\|^2 &= (\mathbf{v} - 3\mathbf{w})\cdot(\mathbf{v} - 3\mathbf{w}) \\
&= \mathbf{v}\cdot(\mathbf{v} - 3\mathbf{w}) - 3\mathbf{w}\cdot(\mathbf{v} - 3\mathbf{w}) \\
&= \mathbf{v}\cdot\mathbf{v} - 3(\mathbf{v}\cdot\mathbf{w}) - 3(\mathbf{w}\cdot\mathbf{v}) + 9(\mathbf{w}\cdot\mathbf{w}) \\
&= \|\mathbf{v}\|^2 - 6(\mathbf{v}\cdot\mathbf{w}) + 9\|\mathbf{w}\|^2 \\
&= 4 - 12 + 9 = 1
\end{aligned}$$

[11] Similarly, if $\mathbf{v} = \left[\begin{array}{c} x_1 \\ y_1 \end{array}\right]$ and $\mathbf{w} = \left[\begin{array}{c} x_2 \\ y_2 \end{array}\right]$ in $\mathbb{R}^2$, then $\mathbf{v}\cdot\mathbf{w} = x_1x_2 + y_1y_2$.

There is an intrinsic description of the dot product of two nonzero vectors in $\mathbb{R}^3$. To understand it we require the following result from trigonometry.

Theorem: Law of Cosines

If a triangle has sides a, b, and c, and if θ is the interior angle opposite c then

$$c^2 = a^2 + b^2 - 2ab\cos\theta$$

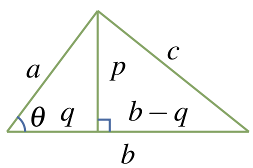


Figure 4.2.2

<u>**Proof.**</u> We prove it when is θ acute, that is $0 \leq \theta < \frac{\pi}{2}$; the obtuse case is similar. In Figure 4.2.2 we have $p = a\sin\theta$ and $q = a\cos\theta$. Hence Pythagoras' theorem gives

$$\begin{aligned} c^2 = p^2 + (b-q)^2 &= a^2\sin^2\theta + (b - a\cos\theta)^2 \\ &= a^2(\sin^2\theta + \cos^2\theta) + b^2 - 2ab\cos\theta \end{aligned}$$

The law of cosines follows because $\sin^2\theta + \cos^2\theta = 1$ for any angle θ. □

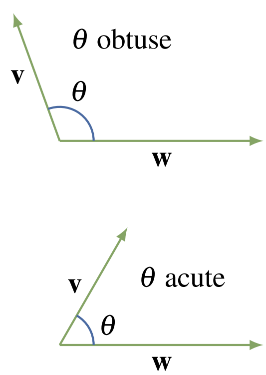


Figure 4.2.3

Note that the law of cosines reduces to Pythagoras' theorem if θ is a right angle (because $\cos\frac{\pi}{2} = 0$).

Now let $\mathbf{v}$ and $\mathbf{w}$ be nonzero vectors positioned with a common tail as in Figure 4.2.3. Then they determine a unique angle θ in the range

$$0 \leq \theta \leq \pi$$

This angle θ will be called the **angle between v** and **w**. Figure 4.2.3 illustrates when θ is acute (less than $\frac{\pi}{2}$) and obtuse (greater than $\frac{\pi}{2}$). Clearly $\mathbf{v}$ and $\mathbf{w}$ are parallel if θ is either 0 or π. Note that we do not define the angle between $\mathbf{v}$ and $\mathbf{w}$ if one of these vectors is $\mathbf{0}$.

The next result gives an easy way to compute the angle between two nonzero vectors using the dot product.

Theorem 4.2.2

Let $\mathbf{v}$ and $\mathbf{w}$ be nonzero vectors. If θ is the angle between $\mathbf{v}$ and $\mathbf{w}$, then

$$\mathbf{v}\cdot\mathbf{w} = \|\mathbf{v}\|\|\mathbf{w}\|\cos\theta$$

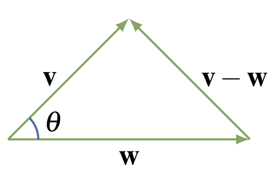


Figure 4.2.4

<u>**Proof.**</u> We calculate $\|\mathbf{v}-\mathbf{w}\|^2$ in two ways. First apply the law of cosines to the triangle in Figure 4.2.4 to obtain:

$$\|\mathbf{v}-\mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - 2\|\mathbf{v}\|\|\mathbf{w}\|\cos\theta$$

On the other hand, we use Theorem 4.2.1:

$$\|\mathbf{v}-\mathbf{w}\|^2 = (\mathbf{v}-\mathbf{w})\cdot(\mathbf{v}-\mathbf{w})$$

$$= \mathbf{v} \cdot \mathbf{v} - \mathbf{v} \cdot \mathbf{w} - \mathbf{w} \cdot \mathbf{v} + \mathbf{w} \cdot \mathbf{w}$$
$$= \|\mathbf{v}\|^2 - 2(\mathbf{v} \cdot \mathbf{w}) + \|\mathbf{w}\|^2$$

Comparing these we see that $-2\|\mathbf{v}\|\|\mathbf{w}\|\cos\theta = -2(\mathbf{v} \cdot \mathbf{w})$, and the result follows. □

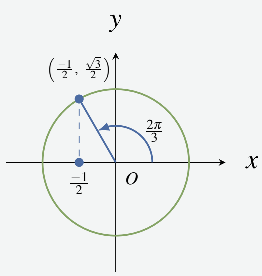
If $\mathbf{v}$ and $\mathbf{w}$ are nonzero vectors, Theorem 4.2.2 gives an intrinsic description of $\mathbf{v} \cdot \mathbf{w}$ because $\|\mathbf{v}\|$, $\|\mathbf{w}\|$, and the angle θ between $\mathbf{v}$ and $\mathbf{w}$ do not depend on the choice of coordinate system. Moreover, since $\|\mathbf{v}\|$ and $\|\mathbf{w}\|$ are nonzero ($\mathbf{v}$ and $\mathbf{w}$ are nonzero vectors), it gives a formula for the cosine of the angle θ:

$$\cos\theta = \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\|\|\mathbf{w}\|} \tag{4.1}$$

Since $0 \le \theta \le \pi$, this can be used to find θ.

Example 4.2.3

Compute the angle between $\mathbf{u} = \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}$.



Solution. Compute $\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} = \frac{-2+1-2}{\sqrt{6}\sqrt{6}} = -\frac{1}{2}$. Now recall that $\cos\theta$ and $\sin\theta$ are defined so that $(\cos\theta, \sin\theta)$ is the point on the unit circle determined by the angle θ (drawn counterclockwise, starting from the positive x axis). In the present case, we know that $\cos\theta = -\frac{1}{2}$ and that $0 \le \theta \le \pi$. Because $\cos\frac{\pi}{3} = \frac{1}{2}$, it follows that $\theta = \frac{2\pi}{3}$ (see the diagram).

If $\mathbf{v}$ and $\mathbf{w}$ are nonzero, equation (4.1) shows that $\cos\theta$ has the same sign as $\mathbf{v} \cdot \mathbf{w}$, so

$$\begin{array}{lll} \mathbf{v} \cdot \mathbf{w} > 0 & \text{if and only if} & \theta \text{ is acute } (0 \le \theta < \frac{\pi}{2}) \\ \mathbf{v} \cdot \mathbf{w} < 0 & \text{if and only if} & \theta \text{ is obtuse } (\frac{\pi}{2} < \theta \le 0) \\ \mathbf{v} \cdot \mathbf{w} = 0 & \text{if and only if} & \theta = \frac{\pi}{2} \end{array}$$

In this last case, the (nonzero) vectors are perpendicular. The following terminology is used in linear algebra:

Definition 4.5 Orthogonal Vectors in $\mathbb{R}^3$

*Two vectors $\mathbf{v}$ and $\mathbf{w}$ are said to be **orthogonal** if $\mathbf{v} = \mathbf{0}$ or $\mathbf{w} = \mathbf{0}$ or the angle between them is $\frac{\pi}{2}$.*

Since $\mathbf{v} \cdot \mathbf{w} = 0$ if either $\mathbf{v} = \mathbf{0}$ or $\mathbf{w} = \mathbf{0}$, we have the following theorem:

Theorem 4.2.3

Two vectors $\mathbf{v}$ and $\mathbf{w}$ are orthogonal if and only if $\mathbf{v} \cdot \mathbf{w} = 0$.

Example 4.2.4

Show that the points $P(3, -1, 1)$, $Q(4, 1, 4)$, and $R(6, 0, 4)$ are the vertices of a right triangle.

Solution. The vectors along the sides of the triangle are

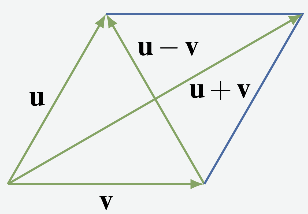
$$\overrightarrow{PQ} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \ \overrightarrow{PR} = \begin{bmatrix} 3 \\ 1 \\ 3 \end{bmatrix}, \ \text{and } \overrightarrow{QR} = \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}$$

Evidently $\overrightarrow{PQ} \cdot \overrightarrow{QR} = 2 - 2 + 0 = 0$, so $\overrightarrow{PQ}$ and $\overrightarrow{QR}$ are orthogonal vectors. This means sides PQ and QR are perpendicular—that is, the angle at Q is a right angle.

Example 4.2.5 demonstrates how the dot product can be used to verify geometrical theorems involving perpendicular lines.

Example 4.2.5

A parallelogram with sides of equal length is called a **rhombus**. Show that the diagonals of a rhombus are perpendicular.



Solution. Let $\mathbf{u}$ and $\mathbf{v}$ denote vectors along two adjacent sides of a rhombus, as shown in the diagram. Then the diagonals are $\mathbf{u} - \mathbf{v}$ and $\mathbf{u} + \mathbf{v}$, and we compute

$$\begin{aligned}(\mathbf{u} - \mathbf{v}) \cdot (\mathbf{u} + \mathbf{v}) &= \mathbf{u} \cdot (\mathbf{u} + \mathbf{v}) - \mathbf{v} \cdot (\mathbf{u} + \mathbf{v}) \\ &= \mathbf{u} \cdot \mathbf{u} + \mathbf{u} \cdot \mathbf{v} - \mathbf{v} \cdot \mathbf{u} - \mathbf{v} \cdot \mathbf{v} \\ &= \|\mathbf{u}\|^2 - \|\mathbf{v}\|^2 \\ &= 0\end{aligned}$$

because $\|\mathbf{u}\| = \|\mathbf{v}\|$ (it is a rhombus). Hence $\mathbf{u} - \mathbf{v}$ and $\mathbf{u} + \mathbf{v}$ are orthogonal.

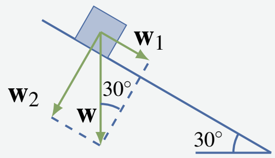
Projections

In applications of vectors, it is frequently useful to write a vector as the sum of two orthogonal vectors. Here is an example.

Example 4.2.6

Suppose a ten-kilogram block is placed on a flat surface inclined 30° to the horizontal as in the diagram. Neglecting friction, how much force is required to keep the block from sliding down the surface?

Solution. Let $\mathbf{w}$ denote the weight (force due to gravity) exerted on the block. Then $\|\mathbf{w}\| = 10$ kilograms and the direction of $\mathbf{w}$ is vertically down as in the diagram.



The idea is to write $\mathbf{w}$ as a sum $\mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2$ where $\mathbf{w}_1$ is parallel to the inclined surface and $\mathbf{w}_2$ is perpendicular to the surface. Since there is no friction, the force required is $-\mathbf{w}_1$ because the force $\mathbf{w}_2$ has no effect parallel to the surface. As the angle between $\mathbf{w}$ and $\mathbf{w}_2$ is 30° in the diagram, we have $\frac{\|\mathbf{w}_1\|}{\|\mathbf{w}\|} = \sin 30^\circ = \frac{1}{2}$. Hence $\|\mathbf{w}_1\| = \frac{1}{2}\|\mathbf{w}\| = \frac{1}{2}10 = 5$. Thus the required force has a magnitude of 5 kilograms weight directed up the surface.

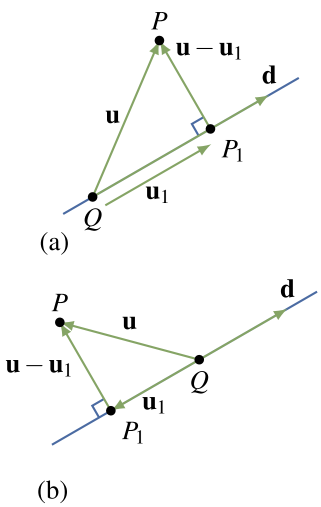


Figure 4.2.5

If a nonzero vector $\mathbf{d}$ is specified, the key idea in Example 4.2.6 is to be able to write an arbitrary vector $\mathbf{u}$ as a sum of two vectors,

$$\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$$

where $\mathbf{u}_1$ is parallel to $\mathbf{d}$ and $\mathbf{u}_2 = \mathbf{u} - \mathbf{u}_1$ is orthogonal to $\mathbf{d}$. Suppose that $\mathbf{u}$ and $\mathbf{d} \neq \mathbf{0}$ emanate from a common tail Q (see Figure 4.2.5). Let P be the tip of $\mathbf{u}$, and let P_1 denote the foot of the perpendicular from P to the line through Q parallel to $\mathbf{d}$.

Then $\mathbf{u}_1 = \overrightarrow{QP}_1$ has the required properties:

1. $\mathbf{u}_1$ *is parallel to* $\mathbf{d}$.
2. $\mathbf{u}_2 = \mathbf{u} - \mathbf{u}_1$ *is orthogonal to* $\mathbf{d}$.
3. $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$.

Definition 4.6 Projection in $\mathbb{R}^3$

The vector $\boldsymbol{u}_1 = \overrightarrow{QP}_1$ *in Figure 4.2.5 is called* ***the projection*** *of* $\boldsymbol{u}$ *on* $\boldsymbol{d}$. *It is denoted*

$$\boldsymbol{u}_1 = \operatorname{proj}_{\boldsymbol{d}} \boldsymbol{u}$$

In Figure 4.2.5(a) the vector $\mathbf{u}_1 = \operatorname{proj}_{\mathbf{d}} \mathbf{u}$ has the same direction as $\mathbf{d}$; however, $\mathbf{u}_1$ and $\mathbf{d}$ have opposite directions if the angle between $\mathbf{u}$ and $\mathbf{d}$ is greater than $\frac{\pi}{2}$ (Figure 4.2.5(b)). Note that the projection $\mathbf{u}_1 = \operatorname{proj}_{\mathbf{d}} \mathbf{u}$ is zero if and only if $\mathbf{u}$ and $\mathbf{d}$ are orthogonal.

Calculating the projection of $\mathbf{u}$ on $\mathbf{d} \neq \mathbf{0}$ is remarkably easy.

Theorem 4.2.4

Let $\boldsymbol{u}$ *and* $\boldsymbol{d} \neq \boldsymbol{0}$ *be vectors.*

1. *The projection of* $\boldsymbol{u}$ *on* $\boldsymbol{d}$ *is given by* $\operatorname{proj}_{\boldsymbol{d}} \boldsymbol{u} = \frac{\boldsymbol{u} \cdot \boldsymbol{d}}{\|\boldsymbol{d}\|^2} \boldsymbol{d}$.
2. *The vector* $\boldsymbol{u} - \operatorname{proj}_{\boldsymbol{d}} \boldsymbol{u}$ *is orthogonal to* $\boldsymbol{d}$.

Proof. The vector $\mathbf{u}_1 = \mathrm{proj}_{\mathbf{d}}\, \mathbf{u}$ is parallel to $\mathbf{d}$ and so has the form $\mathbf{u}_1 = t\mathbf{d}$ for some scalar t. The requirement that $\mathbf{u} - \mathbf{u}_1$ and $\mathbf{d}$ are orthogonal determines t. In fact, it means that $(\mathbf{u} - \mathbf{u}_1) \cdot \mathbf{d} = 0$ by Theorem 4.2.3. If $\mathbf{u}_1 = t\mathbf{d}$ is substituted here, the condition is

$$0 = (\mathbf{u} - t\mathbf{d}) \cdot \mathbf{d} = \mathbf{u} \cdot \mathbf{d} - t(\mathbf{d} \cdot \mathbf{d}) = \mathbf{u} \cdot \mathbf{d} - t\|\mathbf{d}\|^2$$

It follows that $t = \frac{\mathbf{u} \cdot \mathbf{d}}{\|\mathbf{d}\|^2}$, where the assumption that $\mathbf{d} \neq \mathbf{0}$ guarantees that $\|\mathbf{d}\|^2 \neq 0$. □

Example 4.2.7

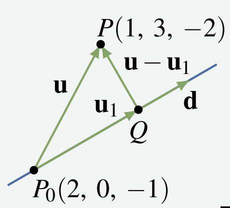
Find the projection of $\mathbf{u} = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$ on $\mathbf{d} = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix}$ and express $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$ where $\mathbf{u}_1$ is parallel to $\mathbf{d}$ and $\mathbf{u}_2$ is orthogonal to $\mathbf{d}$.

Solution. The projection $\mathbf{u}_1$ of $\mathbf{u}$ on $\mathbf{d}$ is

$$\mathbf{u}_1 = \mathrm{proj}_{\mathbf{d}}\, \mathbf{u} = \frac{\mathbf{u} \cdot \mathbf{d}}{\|\mathbf{d}\|^2}\mathbf{d} = \frac{2+3+3}{1^2+(-1)^2+3^2}\begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix} = \frac{8}{11}\begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix}$$

Hence $\mathbf{u}_2 = \mathbf{u} - \mathbf{u}_1 = \frac{1}{11}\begin{bmatrix} 14 \\ -25 \\ -13 \end{bmatrix}$, and this is orthogonal to $\mathbf{d}$ by Theorem 4.2.4 (alternatively, observe that $\mathbf{d} \cdot \mathbf{u}_2 = 0$). Since $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$, we are done.

Example 4.2.8



Find the shortest distance (see diagram) from the point $P(1,\ 3,\ -2)$ to the line through $P_0(2,\ 0,\ -1)$ with direction vector $\mathbf{d} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$. Also find the point Q that lies on the line and is closest to P.

Solution. Let $\mathbf{u} = \begin{bmatrix} 1 \\ 3 \\ -2 \end{bmatrix} - \begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 3 \\ -1 \end{bmatrix}$ denote the vector from P_0 to P, and let $\mathbf{u}_1$ denote the projection of $\mathbf{u}$ on $\mathbf{d}$. Thus

$$\mathbf{u}_1 = \frac{\mathbf{u} \cdot \mathbf{d}}{\|\mathbf{d}\|^2}\mathbf{d} = \frac{-1-3+0}{1^2+(-1)^2+0^2}\mathbf{d} = -2\mathbf{d} = \begin{bmatrix} -2 \\ 2 \\ 0 \end{bmatrix}$$

by Theorem 4.2.4. We see geometrically that the point Q on the line is closest to P, so the distance is

$$\|\overrightarrow{QP}\| = \|\mathbf{u} - \mathbf{u}_1\| = \left\| \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} \right\| = \sqrt{3}$$

To find the coordinates of Q, let $\mathbf{p}_0$ and $\mathbf{q}$ denote the vectors of P_0 and Q, respectively. Then $\mathbf{p}_0 = \begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix}$ and $\mathbf{q} = \mathbf{p}_0 + \mathbf{u}_1 = \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix}$. Hence $Q(0,\ 2,\ -1)$ is the required point. It can be checked that the distance from Q to P is $\sqrt{3}$, as expected.

Planes

It is evident geometrically that among all planes that are perpendicular to a given straight line there is exactly one containing any given point. This fact can be used to give a very simple description of a plane. To do this, it is necessary to introduce the following notion:

Definition 4.7 Normal Vector in a Plane

A nonzero vector $\mathbf{n}$ *is called a* ***normal*** *for a plane if it is orthogonal to every vector in the plane.*

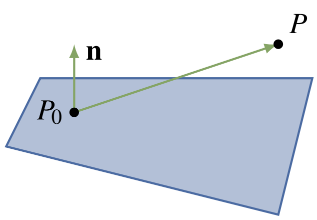


Figure 4.2.6

For example, the coordinate vector $\mathbf{k}$ is a normal for the x-y plane.

Given a point $P_0 = P_0(x_0,\ y_0,\ z_0)$ and a nonzero vector $\mathbf{n}$, there is a unique plane through P_0 with normal $\mathbf{n}$, shaded in Figure 4.2.6. A point $P = P(x,\ y,\ z)$ lies on this plane if and only if the vector $\overrightarrow{P_0P}$ is orthogonal to $\mathbf{n}$—that is, if and only if $\mathbf{n} \cdot \overrightarrow{P_0P} = 0$. Because $\overrightarrow{P_0P} = \begin{bmatrix} x - x_0 \\ y - y_0 \\ z - z_0 \end{bmatrix}$ this gives the following result:

Theorem: Scalar Equation of a Plane

The plane through $P_0(x_0,\ y_0,\ z_0)$ *with normal* $\mathbf{n} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \neq \mathbf{0}$ *as a normal vector is given by*

$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0$$

In other words, a point $P(x,\ y,\ z)$ *is on this plane if and only if* x, y, *and* z *satisfy this equation.*

Example 4.2.9

Find an equation of the plane through $P_0(1,\ -1,\ 3)$ with $\mathbf{n} = \begin{bmatrix} 3 \\ -1 \\ 2 \end{bmatrix}$ as normal.

Solution. Here the general scalar equation becomes

$$3(x - 1) - (y + 1) + 2(z - 3) = 0$$

This simplifies to $3x - y + 2z = 10$.

If we write $d = ax_0 + by_0 + cz_0$, the scalar equation shows that every plane with normal $\mathbf{n} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ has a linear equation of the form

$$ax + by + cz = d \tag{4.2}$$

for some constant d. Conversely, the graph of this equation is a plane with $\mathbf{n} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ as a normal vector (assuming that a, b, and c are not all zero).

Example 4.2.10

Find an equation of the plane through $P_0(3,\ -1,\ 2)$ that is parallel to the plane with equation $2x - 3y = 6$.

Solution. The plane with equation $2x - 3y = 6$ has normal $\mathbf{n} = \begin{bmatrix} 2 \\ -3 \\ 0 \end{bmatrix}$. Because the two planes are parallel, $\mathbf{n}$ serves as a normal for the plane we seek, so the equation is $2x - 3y = d$ for some d by Equation 4.2. Insisting that $P_0(3,\ -1,\ 2)$ lies on the plane determines d; that is, $d = 2 \cdot 3 - 3(-1) = 9$. Hence, the equation is $2x - 3y = 9$.

Consider points $P_0(x_0,\ y_0,\ z_0)$ and $P(x,\ y,\ z)$ with vectors $\mathbf{p}_0 = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix}$ and $\mathbf{p} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$. Given a nonzero vector $\mathbf{n}$, the scalar equation of the plane through $P_0(x_0,\ y_0,\ z_0)$ with normal $\mathbf{n} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ takes the vector form:

Theorem: Vector Equation of a Plane

The plane with normal $\mathbf{n} \neq \mathbf{0}$ through the point with vector $\mathbf{p}_0$ is given by

$$\mathbf{n} \cdot (\mathbf{p} - \mathbf{p}_0) = 0$$

In other words, the point with vector $\mathbf{p}$ is on the plane if and only if $\mathbf{p}$ satisfies this condition.

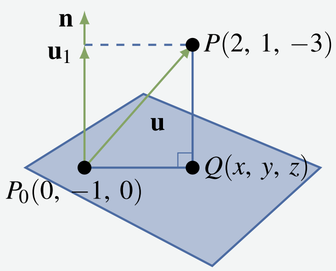
Moreover, Equation 4.2 translates as follows:

Every plane with normal $\mathbf{n}$ has vector equation $\mathbf{n} \cdot \mathbf{p} = d$ for some number d.

This is useful in the second solution of Example 4.2.11.

Example 4.2.11

Find the shortest distance from the point $P(2, 1, -3)$ to the plane with equation $3x - y + 4z = 1$. Also find the point Q on this plane closest to P.



Solution 1. The plane in question has normal $\mathbf{n} = \begin{bmatrix} 3 \\ -1 \\ 4 \end{bmatrix}$.

Choose any point P_0 on the plane—say $P_0(0, -1, 0)$—and let $Q(x, y, z)$ be the point on the plane closest to P (see the diagram).

The vector from P_0 to P is $\mathbf{u} = \begin{bmatrix} 2 \\ 2 \\ -3 \end{bmatrix}$. Now erect $\mathbf{n}$ with its tail at P_0. Then $\overrightarrow{QP} = \mathbf{u}_1$ and $\mathbf{u}_1$ is the projection of $\mathbf{u}$ on $\mathbf{n}$:

$$\mathbf{u}_1 = \frac{\mathbf{n}\cdot\mathbf{u}}{\|\mathbf{n}\|^2}\mathbf{n} = \frac{-8}{26}\begin{bmatrix} 3 \\ -1 \\ 4 \end{bmatrix} = \frac{-4}{13}\begin{bmatrix} 3 \\ -1 \\ 4 \end{bmatrix}$$

Hence the distance is $\|\overrightarrow{QP}\| = \|\mathbf{u}_1\| = \frac{4\sqrt{26}}{13}$. To calculate the point Q, let $\mathbf{q} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ and $\mathbf{p}_0 = \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}$ be the vectors of Q and P_0. Then

$$\mathbf{q} = \mathbf{p}_0 + \mathbf{u} - \mathbf{u}_1 = \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} + \begin{bmatrix} 2 \\ 2 \\ -3 \end{bmatrix} + \frac{4}{13}\begin{bmatrix} 3 \\ -1 \\ 4 \end{bmatrix} = \begin{bmatrix} \frac{38}{13} \\ \frac{9}{13} \\ \frac{-23}{13} \end{bmatrix}$$

This gives the coordinates of $Q(\frac{38}{13}, \frac{9}{13}, \frac{-23}{13})$.

Solution 2. Let $\mathbf{q} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ and $\mathbf{p} = \begin{bmatrix} 2 \\ 1 \\ -3 \end{bmatrix}$ be the vectors of Q and P. Then Q is on the line through P with direction vector $\mathbf{n}$, so $\mathbf{q} = \mathbf{p} + t\mathbf{n}$ for some scalar t. In addition, Q lies on the plane, so $\mathbf{n}\cdot\mathbf{q} = 1$. This determines t:

$$1 = \mathbf{n}\cdot\mathbf{q} = \mathbf{n}\cdot(\mathbf{p} + t\mathbf{n}) = \mathbf{n}\cdot\mathbf{p} + t\|\mathbf{n}\|^2 = -7 + t(26)$$

This gives $t = \frac{8}{26} = \frac{4}{13}$, so

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{q} = \mathbf{p} + t\mathbf{n} = \begin{bmatrix} 2 \\ 1 \\ -3 \end{bmatrix} + \frac{4}{13}\begin{bmatrix} 3 \\ -1 \\ 4 \end{bmatrix} = \frac{1}{13}\begin{bmatrix} 38 \\ 9 \\ -23 \end{bmatrix}$$

as before. This determines Q (in the diagram), and the reader can verify that the required distance is $\|\overrightarrow{QP}\| = \frac{4}{13}\sqrt{26}$, as before.

The Cross Product

If P, Q, and R are three distinct points in $\mathbb{R}^3$ that are not all on some line, it is clear geometrically that there is a unique plane containing all three. The vectors $\overrightarrow{PQ}$ and $\overrightarrow{PR}$ both lie in this plane, so finding a normal amounts to finding a nonzero vector orthogonal to both $\overrightarrow{PQ}$ and $\overrightarrow{PR}$. The cross product provides a systematic way to do this.

Definition 4.8 Cross Product

Given vectors $\mathbf{v}_1 = \left[\begin{array}{c} x_1 \\ y_1 \\ z_1 \end{array}\right]$ *and* $\mathbf{v}_2 = \left[\begin{array}{c} x_2 \\ y_2 \\ z_2 \end{array}\right]$, *define the* ***cross product*** $\mathbf{v}_1 \times \mathbf{v}_2$ *by*

$$\mathbf{v}_1 \times \mathbf{v}_2 = \left[\begin{array}{c} y_1z_2 - z_1y_2 \\ -(x_1z_2 - z_1x_2) \\ x_1y_2 - y_1x_2 \end{array}\right]$$

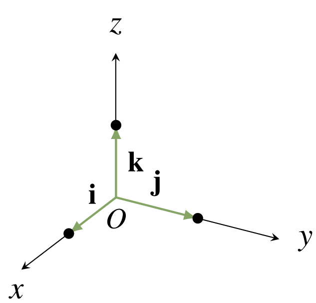


Figure 4.2.7

(Because it is a vector, $\mathbf{v}_1 \times \mathbf{v}_2$ is often called the **vector product**.) There is an easy way to remember this definition using the **coordinate vectors**:

$$\mathbf{i} = \left[\begin{array}{c} 1 \\ 0 \\ 0 \end{array}\right], \ \mathbf{j} = \left[\begin{array}{c} 0 \\ 1 \\ 0 \end{array}\right], \ \text{and } \mathbf{k} = \left[\begin{array}{c} 0 \\ 0 \\ 1 \end{array}\right]$$

They are vectors of length 1 pointing along the positive x, y, and z axes, respectively, as in Figure 4.2.7. The reason for the name is that any vector can be written as

$$\left[\begin{array}{c} x \\ y \\ z \end{array}\right] = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$$

With this, the cross product can be described as follows:

Theorem: Determinant Form of the Cross Product

If $\mathbf{v}_1 = \left[\begin{array}{c} x_1 \\ y_1 \\ z_1 \end{array}\right]$ *and* $\mathbf{v}_2 = \left[\begin{array}{c} x_2 \\ y_2 \\ z_2 \end{array}\right]$ *are two vectors, then*

$$\mathbf{v}_1 \times \mathbf{v}_2 = \det \left[\begin{array}{ccc} \boldsymbol{i} & x_1 & x_2 \\ \boldsymbol{j} & y_1 & y_2 \\ \boldsymbol{k} & z_1 & z_2 \end{array}\right] = \left|\begin{array}{cc} y_1 & y_2 \\ z_1 & z_2 \end{array}\right| \boldsymbol{i} - \left|\begin{array}{cc} x_1 & x_2 \\ z_1 & z_2 \end{array}\right| \boldsymbol{j} + \left|\begin{array}{cc} x_1 & x_2 \\ y_1 & y_2 \end{array}\right| \boldsymbol{k}$$

where the determinant is expanded along the first column.

Example 4.2.12

If $\mathbf{v} = \begin{bmatrix} 2 \\ -1 \\ 4 \end{bmatrix}$ and $\mathbf{w} = \begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix}$, then

$$\begin{aligned} \mathbf{v}_1 \times \mathbf{v}_2 = \det \begin{bmatrix} \mathbf{i} & 2 & 1 \\ \mathbf{j} & -1 & 3 \\ \mathbf{k} & 4 & 7 \end{bmatrix} &= \begin{vmatrix} -1 & 3 \\ 4 & 7 \end{vmatrix} \mathbf{i} - \begin{vmatrix} 2 & 1 \\ 4 & 7 \end{vmatrix} \mathbf{j} + \begin{vmatrix} 2 & 1 \\ -1 & 3 \end{vmatrix} \mathbf{k} \\ &= -19\mathbf{i} - 10\mathbf{j} + 7\mathbf{k} \\ &= \begin{bmatrix} -19 \\ -10 \\ 7 \end{bmatrix} \end{aligned}$$

Observe that $\mathbf{v} \times \mathbf{w}$ is orthogonal to both $\mathbf{v}$ and $\mathbf{w}$ in Example 4.2.12. This holds in general as can be verified directly by computing $\mathbf{v} \cdot (\mathbf{v} \times \mathbf{w})$ and $\mathbf{w} \cdot (\mathbf{v} \times \mathbf{w})$, and is recorded as the first part of the following theorem. It will follow from a more general result which, together with the second part, will be proved in Section 4.3 where a more detailed study of the cross product will be undertaken.

Theorem 4.2.5

Let $\mathbf{v}$ and $\mathbf{w}$ be vectors in $\mathbb{R}^3$.

1. *$\mathbf{v} \times \mathbf{w}$ is a vector orthogonal to both $\mathbf{v}$ and $\mathbf{w}$.*
2. *If $\mathbf{v}$ and $\mathbf{w}$ are nonzero, then $\mathbf{v} \times \mathbf{w} = \mathbf{0}$ if and only if $\mathbf{v}$ and $\mathbf{w}$ are parallel.*

It is interesting to contrast Theorem 4.2.5(2) with the assertion (in Theorem 4.2.3) that

$$\mathbf{v} \cdot \mathbf{w} = 0 \quad \text{if and only if } \mathbf{v} \text{ and } \mathbf{w} \text{ are orthogonal.}$$

Example 4.2.13

Find the equation of the plane through $P(1,\ 3,\ -2)$, $Q(1,\ 1,\ 5)$, and $R(2,\ -2,\ 3)$.

Solution. The vectors $\overrightarrow{PQ} = \begin{bmatrix} 0 \\ -2 \\ 7 \end{bmatrix}$ and $\overrightarrow{PR} = \begin{bmatrix} 1 \\ -5 \\ 5 \end{bmatrix}$ lie in the plane, so

$$\overrightarrow{PQ} \times \overrightarrow{PR} = \det \begin{bmatrix} \mathbf{i} & 0 & 1 \\ \mathbf{j} & -2 & -5 \\ \mathbf{k} & 7 & 5 \end{bmatrix} = 25\mathbf{i} + 7\mathbf{j} + 2\mathbf{k} = \begin{bmatrix} 25 \\ 7 \\ 2 \end{bmatrix}$$

is a normal for the plane (being orthogonal to both $\overrightarrow{PQ}$ and $\overrightarrow{PR}$). Hence the plane has equation

$$25x + 7y + 2z = d \quad \text{for some number } d.$$

Since $P(1,\ 3,\ -2)$ lies in the plane we have $25 \cdot 1 + 7 \cdot 3 + 2(-2) = d$. Hence $d = 42$ and the equation is $25x + 7y + 2z = 42$. Incidentally, the same equation is obtained (verify) if $\overrightarrow{QP}$ and $\overrightarrow{QR}$, or $\overrightarrow{RP}$ and $\overrightarrow{RQ}$, are used as the vectors in the plane.

Example 4.2.14

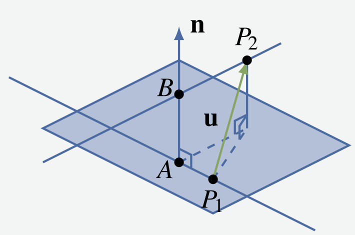
Find the shortest distance between the nonparallel lines

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} + t \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix} + s \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$$

Then find the points A and B on the lines that are closest together.

Solution. Direction vectors for the two lines are $\mathbf{d}_1 = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}$ and $\mathbf{d}_2 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$, so

$$\mathbf{n} = \mathbf{d}_1 \times \mathbf{d}_2 = \det \begin{bmatrix} \mathbf{i} & 2 & 1 \\ \mathbf{j} & 0 & 1 \\ \mathbf{k} & 1 & -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 3 \\ 2 \end{bmatrix}$$



is perpendicular to both lines. Consider the plane shaded in the diagram containing the first line with $\mathbf{n}$ as normal. This plane contains $P_1(1,\ 0,\ -1)$ and is parallel to the second line. Because $P_2(3,\ 1,\ 0)$ is on the second line, the distance in question is just the shortest distance between $P_2(3,\ 1,\ 0)$ and this plane. The vector $\mathbf{u}$ from P_1 to P_2 is $\mathbf{u} = \overrightarrow{P_1P_2} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$ and so, as in Example 4.2.11, the distance is the length of the projection of $\mathbf{u}$ on $\mathbf{n}$.

$$\text{distance} = \left\| \frac{\mathbf{u} \cdot \mathbf{n}}{\|\mathbf{n}\|^2} \mathbf{n} \right\| = \frac{|\mathbf{u} \cdot \mathbf{n}|}{\|\mathbf{n}\|} = \frac{3}{\sqrt{14}} = \frac{3\sqrt{14}}{14}$$

Note that it is necessary that $\mathbf{n} = \mathbf{d}_1 \times \mathbf{d}_2$ be nonzero for this calculation to be possible. As is shown later (Theorem 4.3.4), this is guaranteed by the fact that $\mathbf{d}_1$ and $\mathbf{d}_2$ are *not* parallel.
The points A and B have coordinates $A(1 + 2t,\ 0,\ t - 1)$ and $B(3 + s,\ 1 + s,\ -s)$ for some s and t, so $\overrightarrow{AB} = \begin{bmatrix} 2 + s - 2t \\ 1 + s \\ 1 - s - t \end{bmatrix}$. This vector is orthogonal to both $\mathbf{d}_1$ and $\mathbf{d}_2$, and the conditions $\overrightarrow{AB} \cdot \mathbf{d}_1 = 0$ and $\overrightarrow{AB} \cdot \mathbf{d}_2 = 0$ give equations $5t - s = 5$ and $t - 3s = 2$. The solution is $s = \frac{-5}{14}$ and $t = \frac{13}{14}$, so the points are $A(\frac{40}{14},\ 0,\ \frac{-1}{14})$ and $B(\frac{37}{14},\ \frac{9}{14},\ \frac{5}{14})$. We have $\|\overrightarrow{AB}\| = \frac{3\sqrt{14}}{14}$, as before.

4.3 More on the Cross Product

The cross product $\mathbf{v} \times \mathbf{w}$ of two $\mathbb{R}^3$-vectors $\mathbf{v} = \left[\begin{array}{c} x_1 \\ y_1 \\ z_1 \end{array} \right]$ and $\mathbf{w} = \left[\begin{array}{c} x_2 \\ y_2 \\ z_2 \end{array} \right]$ was defined in Section 4.2 where we observed that it can be best remembered using a determinant:

$$\mathbf{v} \times \mathbf{w} = \det \left[\begin{array}{ccc} \mathbf{i} & x_1 & x_2 \\ \mathbf{j} & y_1 & y_2 \\ \mathbf{k} & z_1 & z_2 \end{array} \right] = \left| \begin{array}{cc} y_1 & y_2 \\ z_1 & z_2 \end{array} \right| \mathbf{i} - \left| \begin{array}{cc} x_1 & x_2 \\ z_1 & z_2 \end{array} \right| \mathbf{j} + \left| \begin{array}{cc} x_1 & x_2 \\ y_1 & y_2 \end{array} \right| \mathbf{k} \tag{4.3}$$

Here $\mathbf{i} = \left[\begin{array}{c} 1 \\ 0 \\ 0 \end{array} \right]$, $\mathbf{j} = \left[\begin{array}{c} 0 \\ 1 \\ 0 \end{array} \right]$, and $\mathbf{k} = \left[\begin{array}{c} 1 \\ 0 \\ 0 \end{array} \right]$ are the coordinate vectors, and the determinant is expanded along the first column. We observed (but did not prove) in Theorem 4.2.5 that $\mathbf{v} \times \mathbf{w}$ is orthogonal to both $\mathbf{v}$ and $\mathbf{w}$. This follows easily from the next result.

Theorem 4.3.1

If $\mathbf{u} = \left[\begin{array}{c} x_0 \\ y_0 \\ z_0 \end{array} \right]$, $\mathbf{v} = \left[\begin{array}{c} x_1 \\ y_1 \\ z_1 \end{array} \right]$, *and* $\mathbf{w} = \left[\begin{array}{c} x_2 \\ y_2 \\ z_2 \end{array} \right]$, *then* $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = \det \left[\begin{array}{ccc} x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \\ z_0 & z_1 & z_2 \end{array} \right]$.

Proof. Recall that $\mathbf{u}\cdot(\mathbf{v}\times\mathbf{w})$ is computed by multiplying corresponding components of $\mathbf{u}$ and $\mathbf{v}\times\mathbf{w}$ and then adding. Using equation (4.3), the result is:

$$\mathbf{u}\cdot(\mathbf{v}\times\mathbf{w}) = x_0\left(\left|\begin{array}{cc} y_1 & y_2 \\ z_1 & z_2 \end{array}\right|\right) + y_0\left(-\left|\begin{array}{cc} x_1 & x_2 \\ z_1 & z_2 \end{array}\right|\right) + z_0\left(\left|\begin{array}{cc} x_1 & x_2 \\ y_1 & y_2 \end{array}\right|\right) = \det\left[\begin{array}{ccc} x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \\ z_0 & z_1 & z_2 \end{array}\right]$$

where the last determinant is expanded along column 1. □

The result in Theorem 4.3.1 can be succinctly stated as follows: If $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ are three vectors in $\mathbb{R}^3$, then

$$\mathbf{u}\cdot(\mathbf{v}\times\mathbf{w}) = \det\left[\begin{array}{ccc} \mathbf{u} & \mathbf{v} & \mathbf{w} \end{array}\right]$$

where $\left[\begin{array}{ccc} \mathbf{u} & \mathbf{v} & \mathbf{w} \end{array}\right]$ denotes the matrix with $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ as its columns. Now it is clear that $\mathbf{v}\times\mathbf{w}$ is orthogonal to both $\mathbf{v}$ and $\mathbf{w}$ because the determinant of a matrix is zero if two columns are identical.

Because of (4.3) and Theorem 4.3.1, several of the following properties of the cross product follow from properties of determinants (they can also be verified directly).

Theorem 4.3.2

Let $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$ denote arbitrary vectors in $\mathbb{R}^3$.

1. *$\boldsymbol{u}\times\boldsymbol{v}$ is a vector.*
2. *$\boldsymbol{u}\times\boldsymbol{v}$ is orthogonal to both $\boldsymbol{u}$ and $\boldsymbol{v}$.*
3. *$\boldsymbol{u}\times\boldsymbol{0} = \boldsymbol{0} = \boldsymbol{0}\times\boldsymbol{u}$.*
4. *$\boldsymbol{u}\times\boldsymbol{u} = \boldsymbol{0}$.*
5. *$\boldsymbol{u}\times\boldsymbol{v} = -(\boldsymbol{v}\times\boldsymbol{u})$.*
6. *$(k\boldsymbol{u})\times\boldsymbol{v} = k(\boldsymbol{u}\times\boldsymbol{v}) = \boldsymbol{u}\times(k\boldsymbol{v})$ for any scalar k.*
7. *$\boldsymbol{u}\times(\boldsymbol{v}+\boldsymbol{w}) = (\boldsymbol{u}\times\boldsymbol{v}) + (\boldsymbol{u}\times\boldsymbol{w})$.*
8. *$(\boldsymbol{v}+\boldsymbol{w})\times\boldsymbol{u} = (\boldsymbol{v}\times\boldsymbol{u}) + (\boldsymbol{w}\times\boldsymbol{u})$.*

Proof. (1) is clear; (2) follows from Theorem 4.3.1; and (3) and (4) follow because the determinant of a matrix is zero if one column is zero or if two columns are identical. If two columns are interchanged, the determinant changes sign, and this proves (5). The proofs of (6), (7), and (8) are left as Exercise **??**. □

We now come to a fundamental relationship between the dot and cross products.

Theorem 4.3.3: Lagrange Identity[12]

If $\boldsymbol{u}$ and $\boldsymbol{v}$ are any two vectors in $\mathbb{R}^3$, then

$$\|\boldsymbol{u}\times\boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2\|\boldsymbol{v}\|^2 - (\boldsymbol{u}\cdot\boldsymbol{v})^2$$

[12]Joseph Louis Lagrange (1736–1813) was born in Italy and spent his early years in Turin. At the age of 19 he solved a famous problem by inventing an entirely new method, known today as the calculus of variations, and went on to become one of the greatest mathematicians of all time. His work brought a new level of rigour to analysis and his *Mécanique Analytique* is a masterpiece in which he introduced methods still in use. In 1766 he was appointed to the Berlin Academy by Frederik the Great who asserted that the "greatest mathematician in Europe" should be at the court of the "greatest king in Europe." After the death of Frederick, Lagrange went to Paris at the invitation of Louis XVI. He remained there throughout the revolution and was made a count by Napoleon.

Proof. Given $\mathbf{u}$ and $\mathbf{v}$, introduce a coordinate system and write $\mathbf{u} = \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix}$ in component form. Then all the terms in the identity can be computed in terms of the components. The detailed proof is left as Exercise **??**. □

An expression for the magnitude of the vector $\mathbf{u} \times \mathbf{v}$ can be easily obtained from the Lagrange identity. If θ is the angle between $\mathbf{u}$ and $\mathbf{v}$, substituting $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| \cos\theta$ into the Lagrange identity gives

$$\|\mathbf{u} \times \mathbf{v}\|^2 = \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 - \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 \cos^2\theta = \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 \sin^2\theta$$

using the fact that $1 - \cos^2\theta = \sin^2\theta$. But $\sin\theta$ is nonnegative on the range $0 \leq \theta \leq \pi$, so taking the positive square root of both sides gives

$$\|\mathbf{u} \times \mathbf{v}\| = \|\mathbf{u}\| \|\mathbf{v}\| \sin\theta$$

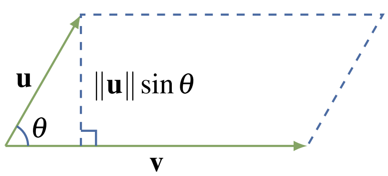


Figure 4.3.1

This expression for $\|\mathbf{u} \times \mathbf{v}\|$ makes no reference to a coordinate system and, moreover, it has a nice geometrical interpretation. The parallelogram determined by the vectors $\mathbf{u}$ and $\mathbf{v}$ has base length $\|\mathbf{v}\|$ and altitude $\|\mathbf{u}\| \sin\theta$ (see Figure 4.3.1). Hence the area of the parallelogram formed by $\mathbf{u}$ and $\mathbf{v}$ is

$$(\|\mathbf{u}\| \sin\theta)\|\mathbf{v}\| = \|\mathbf{u} \times \mathbf{v}\|$$

This proves the first part of Theorem 4.3.4.

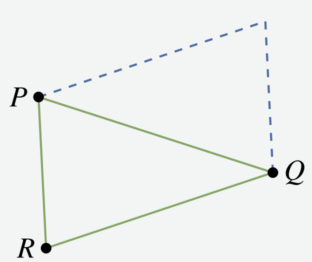
Theorem 4.3.4

If $\mathbf{u}$ and $\mathbf{v}$ are two nonzero vectors and θ is the angle between $\mathbf{u}$ and $\mathbf{v}$, then

1. $\|\mathbf{u} \times \mathbf{v}\| = \|\mathbf{u}\| \|\mathbf{v}\| \sin\theta =$ *the area of the parallelogram determined by $\mathbf{u}$ and $\mathbf{v}$.*
2. *$\mathbf{u}$ and $\mathbf{v}$ are parallel if and only if $\mathbf{u} \times \mathbf{v} = \mathbf{0}$.*

Proof of (2). By (1), $\mathbf{u} \times \mathbf{v} = \mathbf{0}$ if and only if the area of the parallelogram is zero. By Figure 4.3.1 the area vanishes if and only if $\mathbf{u}$ and $\mathbf{v}$ have the same or opposite direction—that is, if and only if they are parallel. □

Example 4.3.1



Find the area of the triangle with vertices $P(2,\ 1,\ 0)$, $Q(3,\ -1,\ 1)$, and $R(1,\ 0,\ 1)$.

Solution. We have $\overrightarrow{RP} = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$ and $\overrightarrow{RQ} = \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}$. The area of the triangle is half the area of the parallelogram (see the diagram), and so equals $\frac{1}{2}\|\overrightarrow{RP} \times \overrightarrow{RQ}\|$. We have

$$\overrightarrow{RP} \times \overrightarrow{RQ} = \det \begin{bmatrix} \mathbf{i} & 1 & 2 \\ \mathbf{j} & 1 & -1 \\ \mathbf{k} & -1 & 0 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \\ -3 \end{bmatrix}$$

so the area of the triangle is $\frac{1}{2}\|\overrightarrow{RP} \times \overrightarrow{RQ}\| = \frac{1}{2}\sqrt{1+4+9} = \frac{1}{2}\sqrt{14}$.

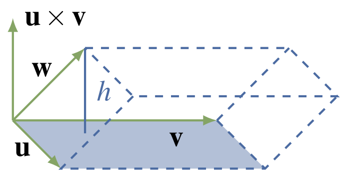


Figure 4.3.2

If three vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ are given, they determine a "squashed" rectangular solid called a **parallelepiped** (Figure 4.3.2), and it is often useful to be able to find the volume of such a solid. The base of the solid is the parallelogram determined by $\mathbf{u}$ and $\mathbf{v}$, so it has area $A = \|\mathbf{u} \times \mathbf{v}\|$ by Theorem 4.3.4. The height of the solid is the length h of the projection of $\mathbf{w}$ on $\mathbf{u} \times \mathbf{v}$. Hence

$$h = \left| \frac{\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})}{\|\mathbf{u} \times \mathbf{v}\|^2} \right| \|\mathbf{u} \times \mathbf{v}\| = \frac{|\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})|}{\|\mathbf{u} \times \mathbf{v}\|} = \frac{|\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})|}{A}$$

Thus the volume of the parallelepiped is $hA = |\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})|$. This proves

Theorem 4.3.5

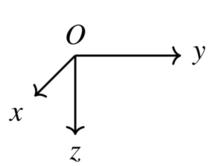
The volume of the parallelepiped determined by three vectors $\mathbf{w}$, $\mathbf{u}$, and $\mathbf{v}$ (Figure 4.3.2) is given by $|\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})|$.

Example 4.3.2

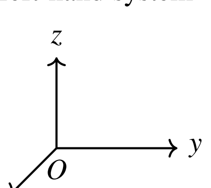
Find the volume of the parallelepiped determined by the vectors

$$\mathbf{w} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}, \mathbf{u} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \mathbf{v} = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$$

Solution. By Theorem 4.3.1, $\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v}) = \det \begin{bmatrix} 1 & 1 & -2 \\ 2 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} = -3$. Hence the volume is $|\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})| = |-3| = 3$ by Theorem 4.3.5.



Left-hand system



We can now give an intrinsic description of the cross product $\mathbf{u} \times \mathbf{v}$. Its magnitude $\|\mathbf{u} \times \mathbf{v}\| = \|\mathbf{u}\|\|\mathbf{v}\| \sin\theta$ is coordinate-free. If $\mathbf{u} \times \mathbf{v} \neq \mathbf{0}$, its direction is very nearly determined by the fact that it is orthogonal to both $\mathbf{u}$ and $\mathbf{v}$ and so points along the line normal to the plane determined by $\mathbf{u}$ and $\mathbf{v}$. It remains only to decide which of the two possible directions is correct.

Before this can be done, the basic issue of how coordinates are assigned must be clarified. When coordinate axes are chosen in space, the procedure is as follows: An origin is selected, two perpendicular lines (the x and y axes) are chosen through the origin, and a positive direction on each of these axes is selected quite arbitrarily. Then the line through the origin normal to this x-y plane is called the z axis, but there is a choice of which direction on this axis is the positive one. The two possibilities are shown in Figure 4.3.3, and it is a standard convention that cartesian coordinates are always **right-hand coordinate systems**. The reason for this terminology is that, in such a system, if the z axis is grasped in the right hand with the thumb pointing in the positive z direction, then the fingers curl around from the positive x axis to the positive y axis (through a right angle).

Suppose now that $\mathbf{u}$ and $\mathbf{v}$ are given and that θ is the angle between them (so $0 \leq \theta \leq \pi$). Then the direction of $\|\mathbf{u} \times \mathbf{v}\|$ is given by the right-hand rule.

Theorem: Right-hand Rule

If the vector $\mathbf{u} \times \mathbf{v}$ is grasped in the right hand and the fingers curl around from $\mathbf{u}$ to $\mathbf{v}$ through the angle θ, the thumb points in the direction for $\mathbf{u} \times \mathbf{v}$.

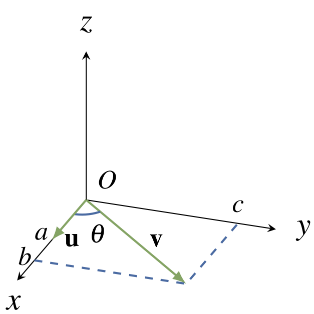


Figure 4.3.4

To indicate why this is true, introduce coordinates in $\mathbb{R}^3$ as follows: Let $\mathbf{u}$ and $\mathbf{v}$ have a common tail O, choose the origin at O, choose the x axis so that $\mathbf{u}$ points in the positive x direction, and then choose the y axis so that $\mathbf{v}$ is in the x-y plane and the positive y axis is on the same side of the x axis as $\mathbf{v}$. Then, in this system, $\mathbf{u}$ and $\mathbf{v}$ have component form $\mathbf{u} = \begin{bmatrix} a \\ 0 \\ 0 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} b \\ c \\ 0 \end{bmatrix}$ where $a > 0$ and $c > 0$. The situation is depicted in Figure 4.3.4. The right-hand rule asserts that $\mathbf{u} \times \mathbf{v}$ should point in the positive z direction. But our definition of $\mathbf{u} \times \mathbf{v}$ gives

$$\mathbf{u} \times \mathbf{v} = \det \begin{bmatrix} \mathbf{i} & a & b \\ \mathbf{j} & 0 & c \\ \mathbf{k} & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ ac \end{bmatrix} = (ac)\mathbf{k}$$

and $(ac)\mathbf{k}$ has the positive z direction because $ac > 0$.

4.4 Linear Operators on $\mathbb{R}^3$

Recall that a transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ is called *linear* if $T(\mathbf{x}+\mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y})$ and $T(a\mathbf{x}) = aT(\mathbf{x})$ holds for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$ and all scalars a. In this case we showed (in Theorem 2.6.2) that there exists an $m \times n$ matrix A such that $T(\mathbf{x}) = A\mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^n$, and we say that T is the **matrix transformation induced** by A.

Definition 4.9 Linear Operator on $\mathbb{R}^n$

A linear transformation

$$T : \mathbb{R}^n \to \mathbb{R}^n$$

is called a ***linear operator*** *on* $\mathbb{R}^n$*.*

In Section 2.6 we investigated three important linear operators on $\mathbb{R}^2$: rotations about the origin, reflections in a line through the origin, and projections on this line.

In this section we investigate the analogous operators on $\mathbb{R}^3$: Rotations about a line through the origin, reflections in a plane through the origin, and projections onto a plane or line through the origin in $\mathbb{R}^3$. In every case we show that the operator is linear, and we find the matrices of all the reflections and projections.

To do this we must prove that these reflections, projections, and rotations are actually *linear* operators on $\mathbb{R}^3$. In the case of reflections and rotations, it is convenient to examine a more general situation. A transformation $T : \mathbb{R}^3 \to \mathbb{R}^3$ is said to be **distance preserving** if the distance between $T(\mathbf{v})$ and $T(\mathbf{w})$ is the same as the distance between $\mathbf{v}$ and $\mathbf{w}$ for all $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^3$; that is,

$$\|T(\mathbf{v}) - T(\mathbf{w})\| = \|\mathbf{v} - \mathbf{w}\| \text{ for all } \mathbf{v} \text{ and } \mathbf{w} \text{ in } \mathbb{R}^3 \tag{4.4}$$

Clearly reflections and rotations are distance preserving, and both carry $\mathbf{0}$ to $\mathbf{0}$, so the following theorem shows that they are both linear.

Theorem 4.4.1

If $T : \mathbb{R}^3 \to \mathbb{R}^3$ is distance preserving, and if $T(\mathbf{0}) = \mathbf{0}$, then T is linear.

Proof. Since $T(\mathbf{0}) = \mathbf{0}$, taking $\mathbf{w} = \mathbf{0}$ in (4.4) shows that $\|T(\mathbf{v})\| = \|\mathbf{v}\|$ for all $\mathbf{v}$ in $\mathbb{R}^3$, that is T preserves length. Also, $\|T(\mathbf{v}) - T(\mathbf{w})\|^2 = \|\mathbf{v} - \mathbf{w}\|^2$ by (4.4). Since $\|\mathbf{v} - \mathbf{w}\|^2 = \|\mathbf{v}\|^2 - 2\mathbf{v} \cdot \mathbf{w} + \|\mathbf{w}\|^2$ always holds, it follows that $T(\mathbf{v}) \cdot T(\mathbf{w}) = \mathbf{v} \cdot \mathbf{w}$ for all $\mathbf{v}$ and $\mathbf{w}$. Hence (by Theorem 4.2.2) the angle between $T(\mathbf{v})$ and $T(\mathbf{w})$ is the same as the angle between $\mathbf{v}$ and $\mathbf{w}$ for all (nonzero) vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^3$.

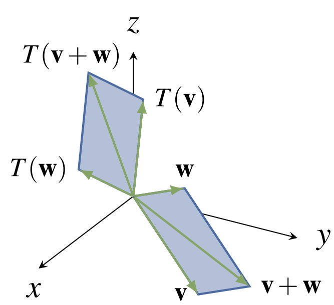


Figure 4.4.1

With this we can show that T is linear. Given nonzero vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^3$, the vector $\mathbf{v} + \mathbf{w}$ is the diagonal of the parallelogram determined by $\mathbf{v}$ and $\mathbf{w}$. By the preceding paragraph, the effect of T is to carry this *entire parallelogram* to the parallelogram determined by $T(\mathbf{v})$ and $T(\mathbf{w})$, with diagonal $T(\mathbf{v} + \mathbf{w})$. But this diagonal is $T(\mathbf{v}) + T(\mathbf{w})$ by the parallelogram law (see Figure 4.4.1).

In other words, $T(\mathbf{v} + \mathbf{w}) = T(\mathbf{v}) + T(\mathbf{w})$. A similar argument shows that $T(a\mathbf{v}) = aT(\mathbf{v})$ for all scalars a, proving that T is indeed linear. □

Distance-preserving linear operators are called **isometries**, and we return to them in Section 10.4.

Reflections and Projections

In Section 2.6 we studied the reflection $Q_m : \mathbb{R}^2 \to \mathbb{R}^2$ in the line $y = mx$ and projection $P_m : \mathbb{R}^2 \to \mathbb{R}^2$ on the same line. We found (in Theorems 2.6.5 and 2.6.6) that they are both linear and

$$Q_m \text{ has matrix } \frac{1}{1+m^2}\begin{bmatrix} 1-m^2 & 2m \\ 2m & m^2-1 \end{bmatrix} \quad \text{and} \quad P_m \text{ has matrix } \frac{1}{1+m^2}\begin{bmatrix} 1 & m \\ m & m^2 \end{bmatrix}.$$

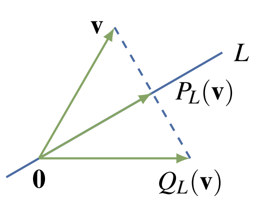


Figure 4.4.2

We now look at the analogues in $\mathbb{R}^3$.

Let L denote a line through the origin in $\mathbb{R}^3$. Given a vector $\mathbf{v}$ in $\mathbb{R}^3$, the reflection $Q_L(\mathbf{v})$ of $\mathbf{v}$ in L and the projection $P_L(\mathbf{v})$ of $\mathbf{v}$ on L are defined in Figure 4.4.2. In the same figure, we see that

$$P_L(\mathbf{v}) = \mathbf{v} + \tfrac{1}{2}[Q_L(\mathbf{v}) - \mathbf{v}] = \tfrac{1}{2}[Q_L(\mathbf{v}) + \mathbf{v}] \tag{4.5}$$

so the fact that Q_L is linear (by Theorem 4.4.1) shows that P_L is also linear.[13]

[13] Note that Theorem 4.4.1 does *not* apply to P_L since it does not preserve distance.

However, Theorem 4.2.4 gives us the matrix of P_L directly. In fact, if $\mathbf{d} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \neq \mathbf{0}$ is a direction vector for L, and we write $\mathbf{v} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$, then

$$P_L(\mathbf{v}) = \frac{\mathbf{v}\cdot\mathbf{d}}{\|\mathbf{d}\|^2}\mathbf{d} = \frac{ax+by+cz}{a^2+b^2+c^2}\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \frac{1}{a^2+b^2+c^2}\begin{bmatrix} a^2 & ab & ac \\ ab & b^2 & bc \\ ac & bc & c^2 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

as the reader can verify. Note that this shows directly that P_L is a matrix transformation and so gives another proof that it is linear.

Theorem 4.4.2

Let L denote the line through the origin in $\mathbb{R}^3$ with direction vector $\mathbf{d} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \neq \mathbf{0}$. Then P_L and Q_L are both linear and

$$P_L \text{ has matrix } \frac{1}{a^2+b^2+c^2}\begin{bmatrix} a^2 & ab & ac \\ ab & b^2 & bc \\ ac & bc & c^2 \end{bmatrix}$$

$$Q_L \text{ has matrix } \frac{1}{a^2+b^2+c^2}\begin{bmatrix} a^2-b^2-c^2 & 2ab & 2ac \\ 2ab & b^2-a^2-c^2 & 2bc \\ 2ac & 2bc & c^2-a^2-b^2 \end{bmatrix}$$

Proof. It remains to find the matrix of Q_L. But (4.5) implies that $Q_L(\mathbf{v}) = 2P_L(\mathbf{v}) - \mathbf{v}$ for each $\mathbf{v}$ in $\mathbb{R}^3$, so if $\mathbf{v} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ we obtain (with some matrix arithmetic):

$$\begin{aligned} Q_L(\mathbf{v}) &= \left\{ \frac{2}{a^2+b^2+c^2}\begin{bmatrix} a^2 & ab & ac \\ ab & b^2 & bc \\ ac & bc & c^2 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right\} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \\ &= \frac{1}{a^2+b^2+c^2}\begin{bmatrix} a^2-b^2-c^2 & 2ab & 2ac \\ 2ab & b^2-a^2-c^2 & 2bc \\ 2ac & 2bc & c^2-a^2-b^2 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} \end{aligned}$$

as required. □

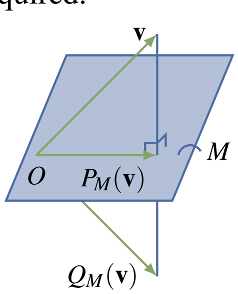


Figure 4.4.3

In $\mathbb{R}^3$ we can reflect in planes as well as lines. Let M denote a plane through the origin in $\mathbb{R}^3$. Given a vector $\mathbf{v}$ in $\mathbb{R}^3$, the reflection $Q_M(\mathbf{v})$ of $\mathbf{v}$ in M and the projection $P_M(\mathbf{v})$ of $\mathbf{v}$ on M are defined in Figure 4.4.3. As above, we have

$$P_M(\mathbf{v}) = \mathbf{v} + \tfrac{1}{2}[Q_M(\mathbf{v}) - \mathbf{v}] = \tfrac{1}{2}[Q_M(\mathbf{v}) + \mathbf{v}]$$

so the fact that Q_M is linear (again by Theorem 4.4.1) shows that P_M is also linear.

Again we can obtain the matrix directly. If $\mathbf{n}$ is a normal for the plane M, then Figure 4.4.3 shows that

$$P_M(\mathbf{v}) = \mathbf{v} - \text{proj}_{\mathbf{n}}\, \mathbf{v} = \mathbf{v} - \frac{\mathbf{v}\cdot\mathbf{n}}{\|\mathbf{n}\|^2}\mathbf{n} \text{ for all vectors } \mathbf{v}.$$

If $\mathbf{n} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \neq \mathbf{0}$ and $\mathbf{v} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$, a computation like the above gives

$$\begin{aligned} P_M(\mathbf{v}) &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} - \frac{ax+by+cz}{a^2+b^2+c^2}\begin{bmatrix} a \\ b \\ c \end{bmatrix} \\ &= \frac{1}{a^2+b^2+c^2}\begin{bmatrix} b^2+c^2 & -ab & -ac \\ -ab & a^2+c^2 & -bc \\ -ac & -bc & b^2+c^2 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} \end{aligned}$$

This proves the first part of

Theorem 4.4.3

Let M denote the plane through the origin in $\mathbb{R}^3$ with normal $\mathbf{n} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \neq \mathbf{0}$. Then P_M and Q_M are both linear and

$$P_M \text{ has matrix } \frac{1}{a^2+b^2+c^2}\begin{bmatrix} b^2+c^2 & -ab & -ac \\ -ab & a^2+c^2 & -bc \\ -ac & -bc & a^2+b^2 \end{bmatrix}$$

$$Q_M \text{ has matrix } \frac{1}{a^2+b^2+c^2}\begin{bmatrix} b^2+c^2-a^2 & -2ab & -2ac \\ -2ab & a^2+c^2-b^2 & -2bc \\ -2ac & -2bc & a^2+b^2-c^2 \end{bmatrix}$$

Proof. It remains to compute the matrix of Q_M. Since $Q_M(\mathbf{v}) = 2P_M(\mathbf{v}) - \mathbf{v}$ for each $\mathbf{v}$ in $\mathbb{R}^3$, the computation is similar to the above and is left as an exercise for the reader. □

Rotations

In Section 2.6 we studied the rotation $R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ counterclockwise about the origin through the angle θ. Moreover, we showed in Theorem 2.6.4 that R_θ is linear and has matrix $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$. One extension of this is given in the following example.

Example 4.4.1

Let $R_{z,\,\theta} : \mathbb{R}^3 \to \mathbb{R}^3$ denote rotation of $\mathbb{R}^3$ about the z axis through an angle θ from the positive x axis toward the positive y axis. Show that $R_{z,\,\theta}$ is linear and find its matrix.

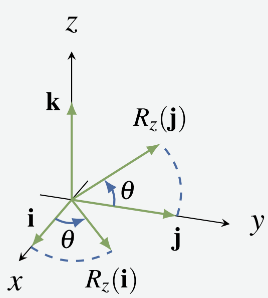


Figure 4.4.4

Solution.
First R is distance preserving and so is linear by Theorem 4.4.1. Hence we apply Theorem 2.6.2 to obtain the matrix of $R_{z,\ \theta}$.

Let $\mathbf{i} = \left[\begin{array}{c} 1 \\ 0 \\ 0 \end{array}\right]$, $\mathbf{j} = \left[\begin{array}{c} 0 \\ 1 \\ 0 \end{array}\right]$, and $\mathbf{k} = \left[\begin{array}{c} 0 \\ 0 \\ 1 \end{array}\right]$ denote the standard basis of $\mathbb{R}^3$; we must find $R_{z,\ \theta}(\mathbf{i})$, $R_{z,\ \theta}(\mathbf{j})$, and $R_{z,\ \theta}(\mathbf{k})$. Clearly $R_{z,\ \theta}(\mathbf{k}) = \mathbf{k}$. The effect of $R_{z,\ \theta}$ on the x-y plane is to rotate it counterclockwise through the angle θ. Hence Figure 4.4.4 gives

$$R_{z,\ \theta}(\mathbf{i}) = \left[\begin{array}{c} \cos\theta \\ \sin\theta \\ 0 \end{array}\right], \quad R_{z,\ \theta}(\mathbf{j}) = \left[\begin{array}{c} -\sin\theta \\ \cos\theta \\ 0 \end{array}\right]$$

so, by Theorem 2.6.2, $R_{z,\ \theta}$ has matrix

$$\left[\begin{array}{ccc} R_{z,\ \theta}(\mathbf{i}) & R_{z,\ \theta}(\mathbf{j}) & R_{z,\ \theta}(\mathbf{k}) \end{array}\right] = \left[\begin{array}{ccc} \cos\theta & -\sin\theta & 0 \\ sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{array}\right]$$

Example 4.4.1 begs to be generalized. Given a line L through the origin in $\mathbb{R}^3$, every rotation about L through a fixed angle is clearly distance preserving, and so is a linear operator by Theorem 4.4.1. However, giving a precise description of the matrix of this rotation is not easy and will have to wait until more techniques are available.

Transformations of Areas and Volumes

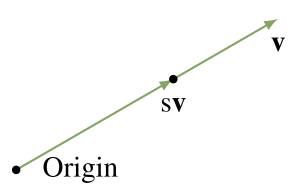


Figure 4.4.5

Let $\mathbf{v}$ be a nonzero vector in $\mathbb{R}^3$. Each vector in the same direction as $\mathbf{v}$ whose length is a fraction s of the length of $\mathbf{v}$ has the form $s\mathbf{v}$ (see Figure 4.4.5).

With this, scrutiny of Figure 4.4.6 shows that a vector $\mathbf{u}$ is in the parallelogram determined by $\mathbf{v}$ and $\mathbf{w}$ if and only if it has the form $\mathbf{u} = s\mathbf{v} + t\mathbf{w}$ where $0 \leq s \leq 1$ and $0 \leq t \leq 1$. But then, if $T : \mathbb{R}^3 \to \mathbb{R}^3$ is a linear transformation, we have

$$T(s\mathbf{v} + t\mathbf{w}) = T(s\mathbf{v}) + T(t\mathbf{w}) = sT(\mathbf{v}) + tT(\mathbf{w})$$

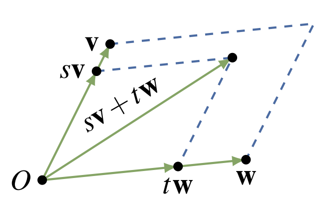


Figure 4.4.6

Hence $T(s\mathbf{v} + t\mathbf{w})$ is in the parallelogram determined by $T(\mathbf{v})$ and $T(\mathbf{w})$. Conversely, every vector in this parallelogram has the form $T(s\mathbf{v} + t\mathbf{w})$ where $s\mathbf{v} + t\mathbf{w}$ is in the parallelogram determined by $\mathbf{v}$ and $\mathbf{w}$. For this reason, the parallelogram determined by $T(\mathbf{v})$ and $T(\mathbf{w})$ is called the **image** of the parallelogram determined by $\mathbf{v}$ and $\mathbf{w}$. We record this discussion as:

Theorem 4.4.4

If $T : \mathbb{R}^3 \to \mathbb{R}^3$ (or $\mathbb{R}^2 \to \mathbb{R}^2$) is a linear operator, the image of the parallelogram determined by vectors $\mathbf{v}$ and $\mathbf{w}$ is the parallelogram determined by $T(\mathbf{v})$ and $T(\mathbf{w})$.

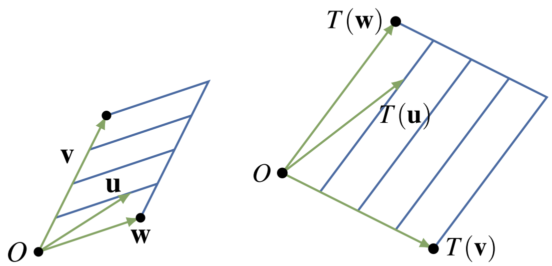


Figure 4.4.7

This result is illustrated in Figure 4.4.7, and was used in Examples 2.2.15 and 2.2.16 to reveal the effect of expansion and shear transformations.

We now describe the effect of a linear transformation $T : \mathbb{R}^3 \to \mathbb{R}^3$ on the parallelepiped determined by three vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ in $\mathbb{R}^3$ (see the discussion preceding Theorem 4.3.5). If T has matrix A, Theorem 4.4.4 shows that this parallelepiped is carried to the parallelepiped determined by $T(\mathbf{u}) = A\mathbf{u}$, $T(\mathbf{v}) = A\mathbf{v}$, and $T(\mathbf{w}) = A\mathbf{w}$. In particular, we want to discover how the volume changes, and it turns out to be closely related to the determinant of the matrix A.

Theorem 4.4.5

Let $\text{vol}(\mathbf{u}, \mathbf{v}, \mathbf{w})$ denote the volume of the parallelepiped determined by three vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ in $\mathbb{R}^3$, and let $\text{area}(\mathbf{p}, \mathbf{q})$ denote the area of the parallelogram determined by two vectors $\mathbf{p}$ and $\mathbf{q}$ in $\mathbb{R}^2$. Then:

1. *If A is a 3×3 matrix, then $\text{vol}(A\mathbf{u}, A\mathbf{v}, A\mathbf{w}) = |\det(A)| \cdot \text{vol}(\mathbf{u}, \mathbf{v}, \mathbf{w})$.*
2. *If A is a 2×2 matrix, then $\text{area}(A\mathbf{p}, A\mathbf{q}) = |\det(A)| \cdot \text{area}(\mathbf{p}, \mathbf{q})$.*

Proof.

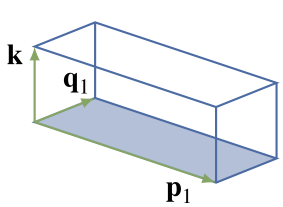
1. Let $\begin{bmatrix} \mathbf{u} & \mathbf{v} & \mathbf{w} \end{bmatrix}$ denote the 3×3 matrix with columns $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$. Then

$$\text{vol}(A\mathbf{u}, A\mathbf{v}, A\mathbf{w}) = |A\mathbf{u} \cdot (A\mathbf{v} \times A\mathbf{w})|$$

by Theorem 4.3.5. Now apply Theorem 4.3.1 twice to get

$$\begin{aligned} A\mathbf{u} \cdot (A\mathbf{v} \times A\mathbf{w}) = \det \begin{bmatrix} A\mathbf{u} & A\mathbf{v} & A\mathbf{w} \end{bmatrix} &= \det(A \begin{bmatrix} \mathbf{u} & \mathbf{v} & \mathbf{w} \end{bmatrix}) \\ &= \det(A) \det \begin{bmatrix} \mathbf{u} & \mathbf{v} & \mathbf{w} \end{bmatrix} \\ &= \det(A)(\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})) \end{aligned}$$

where we used Definition 2.9 and the product theorem for determinants. Finally (1) follows from Theorem 4.3.5 by taking absolute values.



2. Given $\mathbf{p} = \begin{bmatrix} x \\ y \end{bmatrix}$ in $\mathbb{R}^2$, $\mathbf{p}_1 = \begin{bmatrix} x \\ y \\ 0 \end{bmatrix}$ in $\mathbb{R}^3$.

By the diagram, $\text{area}(\mathbf{p}, \mathbf{q}) = \text{vol}(\mathbf{p}_1, \mathbf{q}_1, \mathbf{k})$ where $\mathbf{k}$ is the (length 1) coordinate vector along the z axis. If A is a 2×2

matrix, write $A_1 = \begin{bmatrix} A & 0 \\ 0 & 1 \end{bmatrix}$ in block form, and observe that $(A\mathbf{v})_1 = (A_1\mathbf{v}_1)$ for all $\mathbf{v}$ in $\mathbb{R}^2$ and $A_1\mathbf{k} = \mathbf{k}$.

Hence part (1) of this theorem shows

$$\begin{aligned}\text{area}\,(A\mathbf{p},\ A\mathbf{q}) &= \text{vol}\,(A_1\mathbf{p}_1,\ A_1\mathbf{q}_1,\ A_1\mathbf{k}) \\ &= |\det(A_1)|\ \text{vol}\,(\mathbf{p}_1,\ \mathbf{q}_1,\ \mathbf{k}) \\ &= |\det(A)|\ \text{area}\,(\mathbf{p},\ \mathbf{q})\end{aligned}$$

as required.

□

Define the **unit square** and **unit cube** to be the square and cube corresponding to the coordinate vectors in $\mathbb{R}^2$ and $\mathbb{R}^3$, respectively. Then Theorem 4.4.5 gives a geometrical meaning to the determinant of a matrix A:

- If A is a 2×2 matrix, then $|\det(A)|$ is the area of the image of the unit square under multiplication by A;
- If A is a 3×3 matrix, then $|\det(A)|$ is the volume of the image of the unit cube under multiplication by A.

These results, together with the importance of areas and volumes in geometry, were among the reasons for the initial development of determinants.

4.5 An Application to Computer Graphics

Computer graphics deals with images displayed on a computer screen, and so arises in a variety of applications, ranging from word processors, to *Star Wars* animations, to video games, to wire-frame images of an airplane. These images consist of a number of points on the screen, together with instructions on how to fill in areas bounded by lines and curves. Often curves are approximated by a set of short straight-line segments, so that the curve is specified by a series of points on the screen at the end of these segments. Matrix transformations are important here because matrix images of straight line segments are again line segments.[14] Note that a colour image requires that three images are sent, one to each of the red, green, and blue phosphorus dots on the screen, in varying intensities.

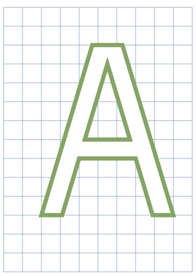


Figure 4.5.1

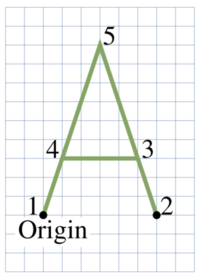


Figure 4.5.2

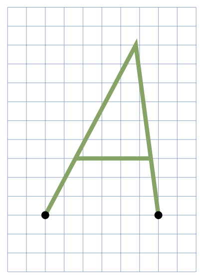


Figure 4.5.3

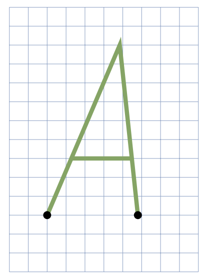


Figure 4.5.4

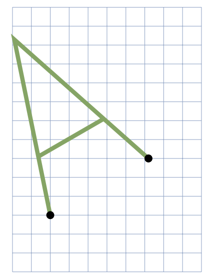


Figure 4.5.5

Consider displaying the letter A. In reality, it is depicted on the screen, as in Figure 4.5.1, by specifying the coordinates of the 11 corners and filling in the interior. For simplicity, we will disregard the thickness of the letter, so we require only five coordinates as in Figure 4.5.2.

This simplified letter can then be stored as a data matrix

$$\begin{array}{cc} \textit{Vertex} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ D = & \left[\begin{array}{ccccc} 0 & 6 & 5 & 1 & 3 \\ 0 & 0 & 3 & 3 & 9 \end{array}\right] \end{array}$$

where the columns are the coordinates of the vertices in order. Then if we want to transform the letter by a 2×2 matrix A, we left-multiply this data matrix by A (the effect is to multiply each column by A and so transform each vertex).

For example, we can slant the letter to the right by multiplying by an x-shear matrix $A = \left[\begin{array}{cc} 1 & 0.2 \\ 0 & 1 \end{array}\right]$ —see Section 2.2. The result is the letter with data matrix

$$A = \left[\begin{array}{cc} 1 & 0.2 \\ 0 & 1 \end{array}\right] \left[\begin{array}{ccccc} 0 & 6 & 5 & 1 & 3 \\ 0 & 0 & 3 & 3 & 9 \end{array}\right] = \left[\begin{array}{ccccc} 0 & 6 & 5.6 & 1.6 & 4.8 \\ 0 & 0 & 3 & 3 & 9 \end{array}\right]$$

which is shown in Figure 4.5.3.

If we want to make this slanted matrix narrower, we can now apply an x-scale matrix $B = \left[\begin{array}{cc} 0.8 & 0 \\ 0 & 1 \end{array}\right]$ that shrinks the x-coordinate by 0.8. The result is the composite transformation

$$BAD = \left[\begin{array}{cc} 0.8 & 0 \\ 0 & 1 \end{array}\right] \left[\begin{array}{cc} 1 & 0.2 \\ 0 & 1 \end{array}\right] \left[\begin{array}{ccccc} 0 & 6 & 5 & 1 & 3 \\ 0 & 0 & 3 & 3 & 9 \end{array}\right]$$

[14]If $\mathbf{v}_0$ and $\mathbf{v}_1$ are vectors, the vector from $\mathbf{v}_0$ to $\mathbf{v}_1$ is $\mathbf{d} = \mathbf{v}_1 - \mathbf{v}_0$. So a vector $\mathbf{v}$ lies on the line segment between $\mathbf{v}_0$ and $\mathbf{v}_1$ if and only if $\mathbf{v} = \mathbf{v}_0 + t\mathbf{d}$ for some number t in the range $0 \leq t \leq 1$. Thus the image of this segment is the set of vectors $A\mathbf{v} = A\mathbf{v}_0 + tA\mathbf{d}$ with $0 \leq t \leq 1$, that is the image is the segment between $A\mathbf{v}_0$ and $A\mathbf{v}_1$.

$$= \begin{bmatrix} 0 & 4.8 & 4.48 & 1.28 & 3.84 \\ 0 & 0 & 3 & 3 & 9 \end{bmatrix}$$

which is drawn in Figure 4.5.4.

On the other hand, we can rotate the letter about the origin through $\frac{\pi}{6}$ (or 30°) by multiplying by the matrix $R_{\frac{\pi}{2}} = \begin{bmatrix} \cos(\frac{\pi}{6}) & -\sin(\frac{\pi}{6}) \\ \sin(\frac{\pi}{6}) & \cos(\frac{\pi}{6}) \end{bmatrix} = \begin{bmatrix} 0.866 & -0.5 \\ 0.5 & 0.866 \end{bmatrix}$. This gives

$$\begin{aligned} R_{\frac{\pi}{2}} &= \begin{bmatrix} 0.866 & -0.5 \\ 0.5 & 0.866 \end{bmatrix} \begin{bmatrix} 0 & 6 & 5 & 1 & 3 \\ 0 & 0 & 3 & 3 & 9 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 5.196 & 2.83 & -0.634 & -1.902 \\ 0 & 3 & 5.098 & 3.098 & 9.294 \end{bmatrix} \end{aligned}$$

and is plotted in Figure 4.5.5.

This poses a problem: How do we rotate at a point other than the origin? It turns out that we can do this when we have solved another more basic problem. It is clearly important to be able to translate a screen image by a fixed vector $\mathbf{w}$, that is apply the transformation $T_w : \mathbb{R}^2 \to \mathbb{R}^2$ given by $T_w(\mathbf{v}) = \mathbf{v} + \mathbf{w}$ for all $\mathbf{v}$ in $\mathbb{R}^2$. The problem is that these translations are not matrix transformations $\mathbb{R}^2 \to \mathbb{R}^2$ because they do not carry $\mathbf{0}$ to $\mathbf{0}$ (unless $\mathbf{w} = \mathbf{0}$). However, there is a clever way around this.

The idea is to represent a point $\mathbf{v} = \begin{bmatrix} x \\ y \end{bmatrix}$ as a 3×1 column $\begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$, called the **homogeneous coordinates** of $\mathbf{v}$. Then translation by $\mathbf{w} = \begin{bmatrix} p \\ q \end{bmatrix}$ can be achieved by multiplying by a 3×3 matrix:

$$\begin{bmatrix} 1 & 0 & p \\ 0 & 1 & q \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} x+p \\ y+q \\ 1 \end{bmatrix} = \begin{bmatrix} T_{\mathbf{w}}(\mathbf{v}) \\ 1 \end{bmatrix}$$

Thus, by using homogeneous coordinates we can implement the translation T_w in the top two coordinates. On the other hand, the matrix transformation induced by $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is also given by a 3×3 matrix:

$$\begin{bmatrix} a & b & 0 \\ c & d & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} ax+by \\ cx+dy \\ 1 \end{bmatrix} = \begin{bmatrix} A\mathbf{v} \\ 1 \end{bmatrix}$$

So everything can be accomplished at the expense of using 3×3 matrices and homogeneous coordinates.

Example 4.5.1

Rotate the letter A in Figure 4.5.2 through $\frac{\pi}{6}$ about the point $\begin{bmatrix} 4 \\ 5 \end{bmatrix}$.

Solution.

Using homogeneous coordinates for the vertices of the letter results in a data matrix with three rows:

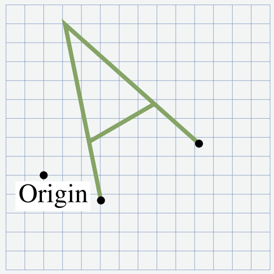


Figure 4.5.6

$$K_d = \begin{bmatrix} 0 & 6 & 5 & 1 & 3 \\ 0 & 0 & 3 & 3 & 9 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

If we write $\mathbf{w} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$,
the idea is to use a composite of transformations: First translate the letter by $-\mathbf{w}$ so that the point $\mathbf{w}$ moves to the origin, then rotate this translated letter, and then translate it by $\mathbf{w}$ back to its original position. The matrix arithmetic is as follows (remember the order of composition!):

$$\begin{bmatrix} 1 & 0 & 4 \\ 0 & 1 & 5 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.866 & -0.5 & 0 \\ 0.5 & 0.866 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -4 \\ 0 & 1 & -5 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 6 & 5 & 1 & 3 \\ 0 & 0 & 3 & 3 & 9 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$
$$= \begin{bmatrix} 3.036 & 8.232 & 5.866 & 2.402 & 1.134 \\ -1.33 & 1.67 & 3.768 & 1.768 & 7.964 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

This is plotted in Figure 4.5.6.

This discussion merely touches the surface of computer graphics, and the reader is referred to specialized books on the subject. Realistic graphic rendering requires an enormous number of matrix calculations. In fact, matrix multiplication algorithms are now embedded in microchip circuits, and can perform over 100 million matrix multiplications per second. This is particularly important in the field of three-dimensional graphics where the homogeneous coordinates have four components and 4×4 matrices are required.

Chapter 5

Vector Space $\mathbb{R}^n$

5.1 Subspaces and Spanning

In Section 2.2 we introduced the set $\mathbb{R}^n$ of all n-tuples (called *vectors*), and began our investigation of the matrix transformations $\mathbb{R}^n \to \mathbb{R}^m$ given by matrix multiplication by an $m \times n$ matrix. Particular attention was paid to the euclidean plane $\mathbb{R}^2$ where certain simple geometric transformations were seen to be matrix transformations. Then in Section 2.6 we introduced linear transformations, showed that they are all matrix transformations, and found the matrices of rotations and reflections in $\mathbb{R}^2$. We returned to this in Section 4.4 where we showed that projections, reflections, and rotations of $\mathbb{R}^2$ and $\mathbb{R}^3$ were all linear, and where we related areas and volumes to determinants.

In this chapter we investigate $\mathbb{R}^n$ in full generality, and introduce some of the most important concepts and methods in linear algebra. The n-tuples in $\mathbb{R}^n$ will continue to be denoted $\mathbf{x}$, $\mathbf{y}$, and so on, and will be written as rows or columns depending on the context.

Subspaces of $\mathbb{R}^n$

Definition 5.1 Subspace of $\mathbb{R}^n$

A set[1] U of vectors in $\mathbb{R}^n$ is called a **subspace** *of $\mathbb{R}^n$ if it satisfies the following properties:*

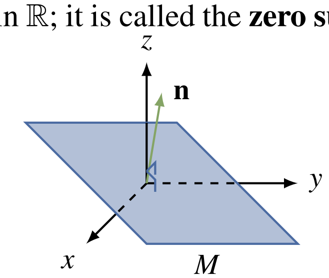
S1. The zero vector $\mathbf{0} \in U$.

S2. If $\mathbf{x} \in U$ and $\mathbf{y} \in U$, then $\mathbf{x}+\mathbf{y} \in U$.

S3. If $\mathbf{x} \in U$, then $a\mathbf{x} \in U$ for every real number a.

We say that the subset U is **closed under addition** if S2 holds, and that U is **closed under scalar multiplication** if S3 holds.

Clearly $\mathbb{R}^n$ is a subspace of itself, and this chapter is about these subspaces and their properties. The set $U = \{\mathbf{0}\}$, consisting of only the zero vector, is also a subspace because $\mathbf{0}+\mathbf{0}=\mathbf{0}$ and $a\mathbf{0}=\mathbf{0}$ for each a in $\mathbb{R}$; it is called the **zero subspace**. Any subspace of $\mathbb{R}^n$ other than $\{\mathbf{0}\}$ or $\mathbb{R}^n$ is called a **proper** subspace.



We saw in Section 4.2 that every plane M through the origin in $\mathbb{R}^3$ has equation $ax+by+cz=0$ where a, b, and c are not all zero. Here $\mathbf{n} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ is a normal for the plane and

$$M = \{\mathbf{v} \text{ in } \mathbb{R}^3 \mid \mathbf{n}\cdot\mathbf{v} = 0\}$$

[1] We use the language of sets. Informally, a **set** X is a collection of objects, called the **elements** of the set. The fact that x is an element of X is denoted $x \in X$. Two sets X and Y are called equal (written $X = Y$) if they have the same elements. If every element of X is in the set Y, we say that X is a **subset** of Y, and write $X \subseteq Y$. Hence $X \subseteq Y$ and $Y \subseteq X$ both hold if and only if $X = Y$.

where $\mathbf{v} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ and $\mathbf{n} \cdot \mathbf{v}$ denotes the dot product introduced in Section 2.2 (see the diagram).[2] Then M is a subspace of $\mathbb{R}^3$. Indeed we show that M satisfies S1, S2, and S3 as follows:

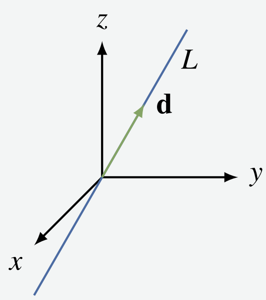
S1. $\mathbf{0} \in M$ *because* $\mathbf{n} \cdot \mathbf{0} = 0$*;*

S2. If $\mathbf{v} \in M$ *and* $\mathbf{v}_1 \in M$ *, then* $\mathbf{n} \cdot (\mathbf{v} + \mathbf{v}_1) = \mathbf{n} \cdot \mathbf{v} + \mathbf{n} \cdot \mathbf{v}_1 = 0 + 0 = 0$ *, so* $\mathbf{v} + \mathbf{v}_1 \in M$*;*

S3. If $\mathbf{v} \in M$ *, then* $\mathbf{n} \cdot (a\mathbf{v}) = a(\mathbf{n} \cdot \mathbf{v}) = a(0) = 0$ *, so* $a\mathbf{v} \in M$.

This proves the first part of

Example 5.1.1



Planes and lines through the origin in $\mathbb{R}^3$ are all subspaces of $\mathbb{R}^3$.

Solution. We dealt with planes above. If L is a line through the origin with direction vector $\mathbf{d}$, then $L = \{t\mathbf{d} \mid t \in \mathbb{R}\}$ (see the diagram). We leave it as an exercise to verify that L satisfies S1, S2, and S3.

Example 5.1.1 shows that lines through the origin in $\mathbb{R}^2$ are subspaces; in fact, they are the *only* proper subspaces of $\mathbb{R}^2$ (Exercise **??**). Indeed, we shall see in Example 5.2.14 that lines and planes through the origin in $\mathbb{R}^3$ are the only proper subspaces of $\mathbb{R}^3$. Thus the geometry of lines and planes through the origin is captured by the subspace concept. (Note that *every* line or plane is just a translation of one of these.)

Subspaces can also be used to describe important features of an $m \times n$ matrix A. The **null space** of A, denoted $\text{null } A$, and the **image space** of A, denoted $\text{im } A$, are defined by

$$\text{null } A = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{0}\} \quad \text{and} \quad \text{im } A = \{A\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n\}$$

In the language of Chapter 2, $\text{null } A$ consists of all solutions $\mathbf{x}$ in $\mathbb{R}^n$ of the homogeneous system $A\mathbf{x} = \mathbf{0}$, and $\text{im } A$ is the set of all vectors $\mathbf{y}$ in $\mathbb{R}^m$ such that $A\mathbf{x} = \mathbf{y}$ *has* a solution $\mathbf{x}$. Note that $\mathbf{x}$ is in $\text{null } A$ if it satisfies the *condition* $A\mathbf{x} = \mathbf{0}$, while $\text{im } A$ consists of vectors of the *form* $A\mathbf{x}$ for some $\mathbf{x}$ in $\mathbb{R}^n$. These two ways to describe subsets occur frequently.

Example 5.1.2

If A is an $m \times n$ matrix, then:

1. $\text{null } A$ is a subspace of $\mathbb{R}^n$.
2. $\text{im } A$ is a subspace of $\mathbb{R}^m$.

[2] We are using set notation here. In general $\{q \mid p\}$ means the set of all objects q with property p.

Solution.

1. The zero vector $\mathbf{0} \in \mathbb{R}^n$ lies in $\operatorname{null} A$ because $A\mathbf{0} = \mathbf{0}$.[3] If $\mathbf{x}$ and $\mathbf{x}_1$ are in $\operatorname{null} A$, then $\mathbf{x} + \mathbf{x}_1$ and $a\mathbf{x}$ are in $\operatorname{null} A$ because they satisfy the required condition:

$$A(\mathbf{x} + \mathbf{x}_1) = A\mathbf{x} + A\mathbf{x}_1 = \mathbf{0} + \mathbf{0} = \mathbf{0} \quad \text{and} \quad A(a\mathbf{x}) = a(A\mathbf{x}) = a\mathbf{0} = \mathbf{0}$$

Hence $\operatorname{null} A$ satisfies S1, S2, and S3, and so is a subspace of $\mathbb{R}^n$.

2. The zero vector $\mathbf{0} \in \mathbb{R}^m$ lies in $\operatorname{im} A$ because $\mathbf{0} = A\mathbf{0}$. Suppose that $\mathbf{y}$ and $\mathbf{y}_1$ are in $\operatorname{im} A$, say $\mathbf{y} = A\mathbf{x}$ and $\mathbf{y}_1 = A\mathbf{x}_1$ where $\mathbf{x}$ and $\mathbf{x}_1$ are in $\mathbb{R}^n$. Then

$$\mathbf{y} + \mathbf{y}_1 = A\mathbf{x} + A\mathbf{x}_1 = A(\mathbf{x} + \mathbf{x}_1) \quad \text{and} \quad a\mathbf{y} = a(A\mathbf{x}) = A(a\mathbf{x})$$

show that $\mathbf{y} + \mathbf{y}_1$ and $a\mathbf{y}$ are both in $\operatorname{im} A$ (they have the required form). Hence $\operatorname{im} A$ is a subspace of $\mathbb{R}^m$.

There are other important subspaces associated with a matrix A that clarify basic properties of A. If A is an $n \times n$ matrix and λ is any number, let

$$E_\lambda(A) = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \lambda\mathbf{x}\}$$

A vector $\mathbf{x}$ is in $E_\lambda(A)$ if and only if $(\lambda I - A)\mathbf{x} = \mathbf{0}$, so Example 5.1.2 gives:

Example 5.1.3

$E_\lambda(A) = \operatorname{null}(\lambda I - A)$ is a subspace of $\mathbb{R}^n$ for each $n \times n$ matrix A and number λ.

$E_\lambda(A)$ is called the **eigenspace** of A corresponding to λ. The reason for the name is that, in the terminology of Section 3.3, λ is an **eigenvalue** of A if $E_\lambda(A) \neq \{\mathbf{0}\}$. In this case the nonzero vectors in $E_\lambda(A)$ are called the **eigenvectors** of A corresponding to λ.

The reader should not get the impression that *every* subset of $\mathbb{R}^n$ is a subspace. For example:

$$U_1 = \left\{ \left[\begin{array}{c} x \\ y \end{array} \right] \middle| x \geq 0 \right\} \text{ satisfies S1 and S2, but not S3;}$$

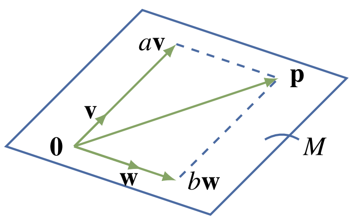
$$U_2 = \left\{ \left[\begin{array}{c} x \\ y \end{array} \right] \middle| x^2 = y^2 \right\} \text{ satisfies S1 and S3, but not S2;}$$

Hence neither U_1 nor U_2 is a subspace of $\mathbb{R}^2$. (However, see Exercise **??**.)

[3] We are using $\mathbf{0}$ to represent the zero vector in both $\mathbb{R}^m$ and $\mathbb{R}^n$. This abuse of notation is common and causes no confusion once everybody knows what is going on.

Spanning Sets

Let $\mathbf{v}$ and $\mathbf{w}$ be two nonzero, nonparallel vectors in $\mathbb{R}^3$ with their tails at the origin. The plane M through the origin containing these vectors is described in Section 4.2 by saying that $\mathbf{n} = \mathbf{v} \times \mathbf{w}$ is a *normal* for M, and that M consists of all vectors $\mathbf{p}$ such that $\mathbf{n} \cdot \mathbf{p} = 0$.[4] While this is a very useful way to look at planes, there is another approach that is at least as useful in $\mathbb{R}^3$ and, more importantly, works for all subspaces of $\mathbb{R}^n$ for any $n \geq 1$.



The idea is as follows: Observe that, by the diagram, a vector $\mathbf{p}$ is in M if and only if it has the form

$$\mathbf{p} = a\mathbf{v} + b\mathbf{w}$$

for certain real numbers a and b (we say that $\mathbf{p}$ is a *linear combination* of $\mathbf{v}$ and $\mathbf{w}$). Hence we can describe M as

$$M = \{a\mathbf{v} + b\mathbf{w} \mid a,\ b \in \mathbb{R}\}.$$[5]

and we say that $\{\mathbf{v},\ \mathbf{w}\}$ is a *spanning set* for M. It is this notion of a spanning set that provides a way to describe all subspaces of $\mathbb{R}^n$.

As in Section 1.3, given vectors $\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_k$ in $\mathbb{R}^n$, a vector of the form

$$t_1\mathbf{x}_1 + t_2\mathbf{x}_2 + \cdots + t_k\mathbf{x}_k \quad \text{where the } t_i \text{ are scalars}$$

is called a **linear combination** of the $\mathbf{x}_i$, and t_i is called the **coefficient** of $\mathbf{x}_i$ in the linear combination.

Definition 5.2 Linear Combinations and Span in $\mathbb{R}^n$

*The set of all such linear combinations is called the **span** of the $\mathbf{x}_i$ and is denoted*

$$\operatorname{span}\{\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_k\} = \{t_1\mathbf{x}_1 + t_2\mathbf{x}_2 + \cdots + t_k\mathbf{x}_k \mid t_i \text{ in } \mathbb{R}\}$$

*If $V = \operatorname{span}\{\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_k\}$, we say that V is **spanned** by the vectors $\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_k$, and that the vectors $\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_k$ **span** the space V.*

Here are two examples:

$$\operatorname{span}\{\mathbf{x}\} = \{t\mathbf{x} \mid t \in \mathbb{R}\}$$

which we write as $\operatorname{span}\{\mathbf{x}\} = \mathbb{R}\mathbf{x}$ for simplicity.

$$\operatorname{span}\{\mathbf{x},\ \mathbf{y}\} = \{r\mathbf{x} + s\mathbf{y} \mid r,\ s \in \mathbb{R}\}$$

In particular, the above discussion shows that, if $\mathbf{v}$ and $\mathbf{w}$ are two nonzero, nonparallel vectors in $\mathbb{R}^3$, then

$$M = \operatorname{span}\{\mathbf{v},\ \mathbf{w}\}$$

is the plane in $\mathbb{R}^3$ containing $\mathbf{v}$ and $\mathbf{w}$. Moreover, if $\mathbf{d}$ is any nonzero vector in $\mathbb{R}^3$ (or $\mathbb{R}^2$), then

$$L = \operatorname{span}\{\mathbf{v}\} = \{t\mathbf{d} \mid t \in \mathbb{R}\} = \mathbb{R}\mathbf{d}$$

is the line with direction vector $\mathbf{d}$. Hence lines and planes can both be described in terms of spanning sets.

[4]The vector $\mathbf{n} = \mathbf{v} \times \mathbf{w}$ is nonzero because $\mathbf{v}$ and $\mathbf{w}$ are not parallel.

[5]In particular, this implies that any vector $\mathbf{p}$ orthogonal to $\mathbf{v} \times \mathbf{w}$ must be a linear combination $\mathbf{p} = a\mathbf{v} + b\mathbf{w}$ of $\mathbf{v}$ and $\mathbf{w}$ for some a and b. Can you prove this directly?

Example 5.1.4

Let $\mathbf{x} = (2,\ -1,\ 2,\ 1)$ and $\mathbf{y} = (3,\ 4,\ -1,\ 1)$ in $\mathbb{R}^4$. Determine whether $\mathbf{p} = (0,\ -11,\ 8,\ 1)$ or $\mathbf{q} = (2,\ 3,\ 1,\ 2)$ are in $U = \operatorname{span}\{\mathbf{x},\ \mathbf{y}\}$.

Solution. The vector $\mathbf{p}$ is in U if and only if $\mathbf{p} = s\mathbf{x} + t\mathbf{y}$ for scalars s and t. Equating components gives equations

$$2s + 3t = 0, \quad -s + 4t = -11, \quad 2s - t = 8, \quad \text{and} \quad s + t = 1$$

This linear system has solution $s = 3$ and $t = -2$, so $\mathbf{p}$ is in U. On the other hand, asking that $\mathbf{q} = s\mathbf{x} + t\mathbf{y}$ leads to equations

$$2s + 3t = 2, \quad -s + 4t = 3, \quad 2s - t = 1, \quad \text{and} \quad s + t = 2$$

and this system has *no* solution. So $\mathbf{q}$ does *not* lie in U.

Theorem 5.1.1: Span Theorem

Let $U = \operatorname{span}\{\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_k\}$ in $\mathbb{R}^n$. Then:

1. *U is a subspace of $\mathbb{R}^n$ containing each $\mathbf{x}_i$.*
2. *If W is a subspace of $\mathbb{R}^n$ and each $\mathbf{x}_i \in W$, then $U \subseteq W$.*

Proof.

1. The zero vector $\mathbf{0}$ is in U because $\mathbf{0} = 0\mathbf{x}_1 + 0\mathbf{x}_2 + \cdots + 0\mathbf{x}_k$ is a linear combination of the $\mathbf{x}_i$. If $\mathbf{x} = t_1\mathbf{x}_1 + t_2\mathbf{x}_2 + \cdots + t_k\mathbf{x}_k$ and $\mathbf{y} = s_1\mathbf{x}_1 + s_2\mathbf{x}_2 + \cdots + s_k\mathbf{x}_k$ are in U, then $\mathbf{x} + \mathbf{y}$ and $a\mathbf{x}$ are in U because

$$\begin{array}{rcl} \mathbf{x} + \mathbf{y} & = & (t_1 + s_1)\mathbf{x}_1 + (t_2 + s_2)\mathbf{x}_2 + \cdots + (t_k + s_k)\mathbf{x}_k, \ \text{and} \\ a\mathbf{x} & = & (at_1)\mathbf{x}_1 + (at_2)\mathbf{x}_2 + \cdots + (at_k)\mathbf{x}_k \end{array}$$

   Finally each $\mathbf{x}_i$ is in U (for example, $\mathbf{x}_2 = 0\mathbf{x}_1 + 1\mathbf{x}_2 + \cdots + 0\mathbf{x}_k$) so S1, S2, and S3 are satisfied for U, proving (1).

2. Let $\mathbf{x} = t_1\mathbf{x}_1 + t_2\mathbf{x}_2 + \cdots + t_k\mathbf{x}_k$ where the t_i are scalars and each $\mathbf{x}_i \in W$. Then each $t_i\mathbf{x}_i \in W$ because W satisfies S3. But then $\mathbf{x} \in W$ because W satisfies S2 (verify). This proves (2). □

Condition (2) in Theorem 5.1.1 can be expressed by saying that $\operatorname{span}\{\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_k\}$ is the *smallest* subspace of $\mathbb{R}^n$ that contains each $\mathbf{x}_i$. This is useful for showing that two subspaces U and W are equal, since this amounts to showing that both $U \subseteq W$ and $W \subseteq U$. Here is an example of how it is used.

Example 5.1.5

If $\mathbf{x}$ and $\mathbf{y}$ are in $\mathbb{R}^n$, show that $\operatorname{span}\{\mathbf{x},\ \mathbf{y}\} = \operatorname{span}\{\mathbf{x} + \mathbf{y},\ \mathbf{x} - \mathbf{y}\}$.

Solution. Since both $\mathbf{x}+\mathbf{y}$ and $\mathbf{x}-\mathbf{y}$ are in $\text{span}\{\mathbf{x},\ \mathbf{y}\}$, Theorem 5.1.1 gives

$$\text{span}\{\mathbf{x}+\mathbf{y},\ \mathbf{x}-\mathbf{y}\} \subseteq \text{span}\{\mathbf{x},\ \mathbf{y}\}$$

But $\mathbf{x} = \frac{1}{2}(\mathbf{x}+\mathbf{y}) + \frac{1}{2}(\mathbf{x}-\mathbf{y})$ and $\mathbf{y} = \frac{1}{2}(\mathbf{x}+\mathbf{y}) - \frac{1}{2}(\mathbf{x}-\mathbf{y})$ are both in $\text{span}\{\mathbf{x}+\mathbf{y},\ \mathbf{x}-\mathbf{y}\}$, so

$$\text{span}\{\mathbf{x},\ \mathbf{y}\} \subseteq \text{span}\{\mathbf{x}+\mathbf{y},\ \mathbf{x}-\mathbf{y}\}$$

again by Theorem 5.1.1. Thus $\text{span}\{\mathbf{x},\ \mathbf{y}\} = \text{span}\{\mathbf{x}+\mathbf{y},\ \mathbf{x}-\mathbf{y}\}$, as desired.

It turns out that many important subspaces are best described by giving a spanning set. Here are three examples, beginning with an important spanning set for $\mathbb{R}^n$ itself.

Recall from Definition 2.3 the standard basis $\{\mathbf{e}_1,\ \mathbf{e}_2,\ \dots,\ \mathbf{e}_n\}$ of $\mathbb{R}^n$ as the set of columns of the $n \times n$ identity matrix. If $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ is any vector in $\mathbb{R}^n$, then $\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n$, as the reader can verify. This proves:

Example 5.1.6

$\mathbb{R}^n = \text{span}\{\mathbf{e}_1,\ \mathbf{e}_2,\ \dots,\ \mathbf{e}_n\}$ where $\mathbf{e}_1,\ \mathbf{e}_2,\ \dots,\ \mathbf{e}_n$ are the columns of I_n.

If A is an $m \times n$ matrix A, the next two examples show that it is a routine matter to find spanning sets for $\text{null}\, A$ and $\text{im}\, A$.

Example 5.1.7

Given an $m \times n$ matrix A, let $\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_k$ denote the basic solutions to the system $A\mathbf{x} = \mathbf{0}$ given by the gaussian algorithm. Then

$$\text{null}\, A = \text{span}\{\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_k\}$$

Solution. If $\mathbf{x} \in \text{null}\, A$, then $A\mathbf{x} = \mathbf{0}$ so Theorem 1.3.2 shows that $\mathbf{x}$ is a linear combination of the basic solutions; that is, $\text{null}\, A \subseteq \text{span}\{\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_k\}$. On the other hand, if $\mathbf{x}$ is in $\text{span}\{\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_k\}$, then $\mathbf{x} = t_1\mathbf{x}_1 + t_2\mathbf{x}_2 + \cdots + t_k\mathbf{x}_k$ for scalars t_i, so

$$A\mathbf{x} = t_1A\mathbf{x}_1 + t_2A\mathbf{x}_2 + \cdots + t_kA\mathbf{x}_k = t_1\mathbf{0} + t_2\mathbf{0} + \cdots + t_k\mathbf{0} = \mathbf{0}$$

This shows that $\mathbf{x} \in \text{null}\, A$, and hence that $\text{span}\{\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_k\} \subseteq \text{null}\, A$. Thus we have equality.

Example 5.1.8

Let $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n$ denote the columns of the $m \times n$ matrix A. Then

$$\operatorname{im} A = \operatorname{span}\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\}$$

Solution. If $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ is the standard basis of $\mathbb{R}^n$, observe that

$$\left[\begin{array}{cccc} A\mathbf{e}_1 & A\mathbf{e}_2 & \cdots & A\mathbf{e}_n \end{array}\right] = A\left[\begin{array}{cccc} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_n \end{array}\right] = AI_n = A = \left[\begin{array}{cccc} \mathbf{c}_1 & \mathbf{c}_2 & \cdots \mathbf{c}_n \end{array}\right].$$

Hence $\mathbf{c}_i = A\mathbf{e}_i$ is in $\operatorname{im} A$ for each i, so $\operatorname{span}\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\} \subseteq \operatorname{im} A$.

Conversely, let $\mathbf{y}$ be in $\operatorname{im} A$, say $\mathbf{y} = A\mathbf{x}$ for some $\mathbf{x}$ in $\mathbb{R}^n$. If $\mathbf{x} = \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array}\right]$, then Definition 2.5 gives

$$\mathbf{y} = A\mathbf{x} = x_1\mathbf{c}_1 + x_2\mathbf{c}_2 + \cdots + x_n\mathbf{c}_n \text{ is in } \operatorname{span}\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\}$$

This shows that $\operatorname{im} A \subseteq \operatorname{span}\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\}$, and the result follows.

5.2 Independence and Dimension

Some spanning sets are better than others. If $U = \text{span}\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ is a subspace of $\mathbb{R}^n$, then every vector in U can be written as a linear combination of the $\mathbf{x}_i$ in at least one way. Our interest here is in spanning sets where each vector in U has *exactly one* representation as a linear combination of these vectors.

Linear Independence

Given $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ in $\mathbb{R}^n$, suppose that two linear combinations are equal:

$$r_1\mathbf{x}_1 + r_2\mathbf{x}_2 + \cdots + r_k\mathbf{x}_k = s_1\mathbf{x}_1 + s_2\mathbf{x}_2 + \cdots + s_k\mathbf{x}_k$$

We are looking for a condition on the set $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ of vectors that guarantees that this representation is *unique*; that is, $r_i = s_i$ for each i. Taking all terms to the left side gives

$$(r_1 - s_1)\mathbf{x}_1 + (r_2 - s_2)\mathbf{x}_2 + \cdots + (r_k - s_k)\mathbf{x}_k = \mathbf{0}$$

so the required condition is that this equation forces all the coefficients $r_i - s_i$ to be zero.

Definition 5.3 Linear Independence in $\mathbb{R}^n$

With this in mind, we call a set $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ of vectors ***linearly independent*** *(or simply* ***independent****) if it satisfies the following condition:*

$$\text{If } t_1\mathbf{x}_1 + t_2\mathbf{x}_2 + \cdots + t_k\mathbf{x}_k = \mathbf{0} \text{ then } t_1 = t_2 = \cdots = t_k = 0$$

We record the result of the above discussion for reference.

Theorem 5.2.1

If $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ is an independent set of vectors in $\mathbb{R}^n$, then every vector in span $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ *has a* ***unique*** *representation as a linear combination of the $\mathbf{x}_i$.*

It is useful to state the definition of independence in different language. Let us say that a linear combination **vanishes** if it equals the zero vector, and call a linear combination **trivial** if every coefficient is zero. Then the definition of independence can be compactly stated as follows:

> A set of vectors is independent if and only if the only linear combination that vanishes is the trivial one.

Hence we have a procedure for checking that a set of vectors is independent:

Theorem: Independence Test

To verify that a set $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ of vectors in $\mathbb{R}^n$ is independent, proceed as follows:

1. *Set a linear combination equal to zero:* $t_1\mathbf{x}_1 + t_2\mathbf{x}_2 + \cdots + t_k\mathbf{x}_k = \mathbf{0}$.

2. *Show that $t_i = 0$ for each i (that is, the linear combination is trivial).*

Of course, if some nontrivial linear combination vanishes, the vectors are not independent.

Example 5.2.1

Determine whether $\{(1, 0, -2, 5), (2, 1, 0, -1), (1, 1, 2, 1)\}$ is independent in $\mathbb{R}^4$.

Solution. Suppose a linear combination vanishes:

$$r(1, 0, -2, 5) + s(2, 1, 0, -1) + t(1, 1, 2, 1) = (0, 0, 0, 0)$$

Equating corresponding entries gives a system of four equations:

$$r + 2s + t = 0, \; s + t = 0, \; -2r + 2t = 0, \; \text{and } 5r - s + t = 0$$

The only solution is the trivial one $r = s = t = 0$ (verify), so these vectors are independent by the independence test.

Recall from Definition 2.3 that the standard basis of $\mathbb{R}^n$ is the set of columns of the identity matrix I_n.

Example 5.2.2

Show that the standard basis $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ of $\mathbb{R}^n$ is independent.

Solution. The components of $t_1\mathbf{e}_1 + t_2\mathbf{e}_2 + \cdots + t_n\mathbf{e}_n$ are $t_1, t_2, \dots, t_n$ (see the discussion preceding Example 5.1.6) So the linear combination vanishes if and only if each $t_i = 0$. Hence the independence test applies.

Example 5.2.3

If $\{\mathbf{x}, \mathbf{y}\}$ is independent, show that $\{2\mathbf{x} + 3\mathbf{y}, \mathbf{x} - 5\mathbf{y}\}$ is also independent.

Solution. If $s(2\mathbf{x} + 3\mathbf{y}) + t(\mathbf{x} - 5\mathbf{y}) = \mathbf{0}$, collect terms to get $(2s + t)\mathbf{x} + (3s - 5t)\mathbf{y} = \mathbf{0}$. Since $\{\mathbf{x}, \mathbf{y}\}$ is independent this combination must be trivial; that is, $2s + t = 0$ and $3s - 5t = 0$. These equations have only the trivial solution $s = t = 0$, as required.

Example 5.2.4

Show that the zero vector in $\mathbb{R}^n$ does not belong to any independent set.

Solution. No set $\{\mathbf{0}, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ of vectors is independent because we have a vanishing, nontrivial linear combination $1 \cdot \mathbf{0} + 0\mathbf{x}_1 + 0\mathbf{x}_2 + \cdots + 0\mathbf{x}_k = \mathbf{0}$.

Example 5.2.5

Given $\mathbf{x}$ in $\mathbb{R}^n$, show that $\{\mathbf{x}\}$ is independent if and only if $\mathbf{x} \neq \mathbf{0}$.

Solution. A vanishing linear combination from $\{\mathbf{x}\}$ takes the form $t\mathbf{x} = \mathbf{0}$, t in $\mathbb{R}$. This implies that $t = 0$ because $\mathbf{x} \neq \mathbf{0}$.

The next example will be needed later.

Example 5.2.6

Show that the nonzero rows of a row-echelon matrix R are independent.

Solution. We illustrate the case with 3 leading 1s; the general case is analogous. Suppose R has the form $R = \left[\begin{array}{cccccc} 0 & 1 & * & * & * & * \\ 0 & 0 & 0 & 1 & * & * \\ 0 & 0 & 0 & 0 & 1 & * \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right]$ where $*$ indicates a nonspecified number. Let R_1, R_2, and R_3 denote the nonzero rows of R. If $t_1R_1 + t_2R_2 + t_3R_3 = 0$ we show that $t_1 = 0$, then $t_2 = 0$, and finally $t_3 = 0$. The condition $t_1R_1 + t_2R_2 + t_3R_3 = 0$ becomes

$$(0,\ t_1,\ *,\ *,\ *,\ *) + (0,\ 0,\ 0,\ t_2,\ *,\ *) + (0,\ 0,\ 0,\ 0,\ t_3,\ *) = (0,\ 0,\ 0,\ 0,\ 0,\ 0)$$

Equating second entries show that $t_1 = 0$, so the condition becomes $t_2R_2 + t_3R_3 = 0$. Now the same argument shows that $t_2 = 0$. Finally, this gives $t_3R_3 = 0$ and we obtain $t_3 = 0$.

A set of vectors in $\mathbb{R}^n$ is called **linearly dependent** (or simply **dependent**) if it is *not* linearly independent, equivalently if some nontrivial linear combination vanishes.

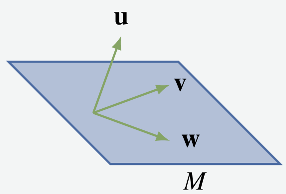
Example 5.2.7

If $\mathbf{v}$ and $\mathbf{w}$ are nonzero vectors in $\mathbb{R}^3$, show that $\{\mathbf{v},\ \mathbf{w}\}$ is dependent if and only if $\mathbf{v}$ and $\mathbf{w}$ are parallel.

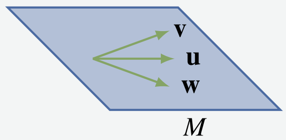
Solution. If $\mathbf{v}$ and $\mathbf{w}$ are parallel, then one is a scalar multiple of the other (Theorem 4.1.5), say $\mathbf{v} = a\mathbf{w}$ for some scalar a. Then the nontrivial linear combination $\mathbf{v} - a\mathbf{w} = \mathbf{0}$ vanishes, so $\{\mathbf{v},\ \mathbf{w}\}$ is dependent.
Conversely, if $\{\mathbf{v},\ \mathbf{w}\}$ is dependent, let $s\mathbf{v} + t\mathbf{w} = \mathbf{0}$ be nontrivial, say $s \neq 0$. Then $\mathbf{v} = -\frac{t}{s}\mathbf{w}$ so $\mathbf{v}$ and $\mathbf{w}$ are parallel (by Theorem 4.1.5). A similar argument works if $t \neq 0$.

With this we can give a geometric description of what it means for a set $\{\mathbf{u},\ \mathbf{v},\ \mathbf{w}\}$ in $\mathbb{R}^3$ to be independent. Note that this requirement means that $\{\mathbf{v},\ \mathbf{w}\}$ is also independent ($a\mathbf{v} + b\mathbf{w} = \mathbf{0}$ means that $0\mathbf{u} + a\mathbf{v} + b\mathbf{w} = \mathbf{0}$), so $M = \text{span}\,\{\mathbf{v},\ \mathbf{w}\}$ is the plane containing $\mathbf{v}$, $\mathbf{w}$, and $\mathbf{0}$ (see the discussion preceding Example 5.1.4). So we assume that $\{\mathbf{v},\ \mathbf{w}\}$ is independent in the following example.

Example 5.2.8



$\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$ independent



$\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$ not independent

Let $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ be nonzero vectors in $\mathbb{R}^3$ where $\{\mathbf{v}, \mathbf{w}\}$ independent. Show that $\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$ is independent if and only if $\mathbf{u}$ is not in the plane $M = \text{span}\{\mathbf{v}, \mathbf{w}\}$. This is illustrated in the diagrams.

Solution. If $\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$ is independent, suppose $\mathbf{u}$ is in the plane $M = \text{span}\{\mathbf{v}, \mathbf{w}\}$, say $\mathbf{u} = a\mathbf{v} + b\mathbf{w}$, where a and b are in $\mathbb{R}$. Then $1\mathbf{u} - a\mathbf{v} - b\mathbf{w} = \mathbf{0}$, contradicting the independence of $\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$. On the other hand, suppose that $\mathbf{u}$ is not in M; we must show that $\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$ is independent. If $r\mathbf{u} + s\mathbf{v} + t\mathbf{w} = \mathbf{0}$ where r, s, and t are in $\mathbb{R}^3$, then $r = 0$ since otherwise $\mathbf{u} = -\frac{s}{r}\mathbf{v} + \frac{-t}{r}\mathbf{w}$ is in M. But then $s\mathbf{v} + t\mathbf{w} = \mathbf{0}$, so $s = t = 0$ by our assumption. This shows that $\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$ is independent, as required.

By the inverse theorem, the following conditions are equivalent for an $n \times n$ matrix A:

1. *A is invertible.*
2. *If $A\mathbf{x} = \mathbf{0}$ where $\mathbf{x}$ is in $\mathbb{R}^n$, then $\mathbf{x} = \mathbf{0}$.*
3. *$A\mathbf{x} = \mathbf{b}$ has a solution $\mathbf{x}$ for every vector $\mathbf{b}$ in $\mathbb{R}^n$.*

While condition 1 makes no sense if A is not square, conditions 2 and 3 are meaningful for any matrix A and, in fact, are related to independence and spanning. Indeed, if $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n$ are the columns of A, and if we write $\mathbf{x} = \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array}\right]$, then

$$A\mathbf{x} = x_1\mathbf{c}_1 + x_2\mathbf{c}_2 + \cdots + x_n\mathbf{c}_n$$

by Definition 2.5. Hence the definitions of independence and spanning show, respectively, that condition 2 is equivalent to the independence of $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\}$ and condition 3 is equivalent to the requirement that $\text{span}\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\} = \mathbb{R}^m$. This discussion is summarized in the following theorem:

Theorem 5.2.2

If A is an $m \times n$ matrix, let $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\}$ denote the columns of A.

1. *$\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\}$ is independent in $\mathbb{R}^m$ if and only if $A\mathbf{x} = \mathbf{0}$, $\mathbf{x}$ in $\mathbb{R}^n$, implies $\mathbf{x} = \mathbf{0}$.*
2. *$\mathbb{R}^m = \text{span}\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\}$ if and only if $A\mathbf{x} = \mathbf{b}$ has a solution $\mathbf{x}$ for every vector $\mathbf{b}$ in $\mathbb{R}^m$.*

For a *square* matrix A, Theorem 5.2.2 characterizes the invertibility of A in terms of the spanning and independence of its columns (see the discussion preceding Theorem 5.2.2). It is important to be able to discuss these notions for *rows*. If $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ are $1 \times n$ rows, we define $\text{span}\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ to be the set of all linear combinations of the $\mathbf{x}_i$ (as matrices), and we say that $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ is linearly independent if the only vanishing linear combination is the trivial one (that is, if $\{\mathbf{x}_1^T, \mathbf{x}_2^T, \dots, \mathbf{x}_k^T\}$ is independent in $\mathbb{R}^n$, as the reader can verify).[6]

Theorem 5.2.3

The following are equivalent for an $n \times n$ matrix A:

1. *A is invertible.*
2. *The columns of A are linearly independent.*
3. *The columns of A span $\mathbb{R}^n$.*
4. *The rows of A are linearly independent.*
5. *The rows of A span the set of all $1 \times n$ rows.*

Proof. Let $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n$ denote the columns of A.

(1) $\Leftrightarrow$ (2). By Theorem 2.4.5, A is invertible if and only if $A\mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$; this holds if and only if $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\}$ is independent by Theorem 5.2.2.

(1) $\Leftrightarrow$ (3). Again by Theorem 2.4.5, A is invertible if and only if $A\mathbf{x} = \mathbf{b}$ has a solution for every column B in $\mathbb{R}^n$; this holds if and only if span $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\} = \mathbb{R}^n$ by Theorem 5.2.2.

(1) $\Leftrightarrow$ (4). The matrix A is invertible if and only if A^T is invertible (by Corollary 2.4.1 to Theorem 2.4.4); this in turn holds if and only if A^T has independent columns (by (1) $\Leftrightarrow$ (2)); finally, this last statement holds if and only if A has independent rows (because the rows of A are the transposes of the columns of A^T).

(1) $\Leftrightarrow$ (5). The proof is similar to (1) $\Leftrightarrow$ (4). □

Example 5.2.9

Show that $S = \{(2, -2, 5), (-3, 1, 1), (2, 7, -4)\}$ is independent in $\mathbb{R}^3$.

Solution. Consider the matrix $A = \left[\begin{array}{rrr} 2 & -2 & 5 \\ -3 & 1 & 1 \\ 2 & 7 & -4 \end{array} \right]$ with the vectors in S as its rows. A routine computation shows that $\det A = -117 \neq 0$, so A is invertible. Hence S is independent by Theorem 5.2.3. Note that Theorem 5.2.3 also shows that $\mathbb{R}^3 = \text{span } S$.

[6]It is best to view columns and rows as just two different *notations* for ordered n-tuples. This discussion will become redundant in Chapter 6 where we define the general notion of a vector space.

Dimension

It is common geometrical language to say that $\mathbb{R}^3$ is 3-dimensional, that planes are 2-dimensional and that lines are 1-dimensional. The next theorem is a basic tool for clarifying this idea of "dimension". Its importance is difficult to exaggerate.

Theorem 5.2.4: Fundamental Theorem

Let U be a subspace of $\mathbb{R}^n$. If U is spanned by m vectors, and if U contains k linearly independent vectors, then $k \leq m$.

This proof is given in Theorem 6.3.2 in much greater generality.

Definition 5.4 Basis of a Subspace of $\mathbb{R}^n$

*If U is a subspace of $\mathbb{R}^n$, a set $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ of vectors in U is called a **basis** of U if it satisfies the following two conditions:*

1. *$\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ is linearly independent.*
2. *$U = \text{span}\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$.*

The most remarkable result about bases[7] is:

Theorem 5.2.5: Invariance Theorem

If $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ and $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k\}$ are bases of a subspace U of $\mathbb{R}^n$, then $m = k$.

Proof. We have $k \leq m$ by the fundamental theorem because $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ spans U, and $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k\}$ is independent. Similarly, by interchanging $\mathbf{x}$'s and $\mathbf{y}$'s we get $m \leq k$. Hence $m = k$. □

The invariance theorem guarantees that there is no ambiguity in the following definition:

Definition 5.5 Dimension of a Subspace of $\mathbb{R}^n$

*If U is a subspace of $\mathbb{R}^n$ and $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ is any basis of U, the number, m, of vectors in the basis is called the **dimension** of U, denoted*

$$\dim U = m$$

The importance of the invariance theorem is that the dimension of U can be determined by counting the number of vectors in *any* basis.[8]

Recall from Definition 2.3 the standard basis of $\mathbb{R}^n$ $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$, that is the set of columns of the identity matrix. Then $\mathbb{R}^n = \text{span}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ by Example 5.1.6, and $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ is independent by Example 5.2.2. Hence it is indeed a basis of $\mathbb{R}^n$ in the present terminology, and we have

[7] The plural of "basis" is "bases".

[8] We will show in Theorem 5.2.6 that every subspace of $\mathbb{R}^n$ does indeed *have* a basis.

Example 5.2.10

$\dim(\mathbb{R}^n) = n$ and $\{\mathbf{e}_1,\ \mathbf{e}_2,\ \dots,\ \mathbf{e}_n\}$ is a basis.

This agrees with our geometric sense that $\mathbb{R}^2$ is two-dimensional and $\mathbb{R}^3$ is three-dimensional. It also says that $\mathbb{R}^1 = \mathbb{R}$ is one-dimensional, and $\{1\}$ is a basis. Returning to subspaces of $\mathbb{R}^n$, we define

$$\dim\{\mathbf{0}\} = 0$$

This amounts to saying $\{\mathbf{0}\}$ has a basis containing *no* vectors. This makes sense because $\mathbf{0}$ cannot belong to *any* independent set (Example 5.2.4).

Example 5.2.11

Let $U = \left\{ \left[\begin{array}{c} r \\ s \\ r \end{array} \right] \middle| \ r,\ s \text{ in } \mathbb{R} \right\}$. Show that U is a subspace of $\mathbb{R}^3$, find a basis, and calculate $\dim U$.

Solution. Clearly, $\left[\begin{array}{c} r \\ s \\ r \end{array} \right] = r\mathbf{u} + s\mathbf{v}$ where $\mathbf{u} = \left[\begin{array}{c} 1 \\ 0 \\ 1 \end{array} \right]$ and $\mathbf{v} = \left[\begin{array}{c} 0 \\ 1 \\ 0 \end{array} \right]$. It follows that $U = \text{span}\{\mathbf{u},\ \mathbf{v}\}$, and hence that U is a subspace of $\mathbb{R}^3$. Moreover, if $r\mathbf{u} + s\mathbf{v} = \mathbf{0}$, then $\left[\begin{array}{c} r \\ s \\ r \end{array} \right] = \left[\begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right]$ so $r = s = 0$. Hence $\{\mathbf{u},\ \mathbf{v}\}$ is independent, and so a **basis** of U. This means $\dim U = 2$.

Example 5.2.12

Let $B = \{\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_n\}$ be a basis of $\mathbb{R}^n$. If A is an invertible $n \times n$ matrix, then $D = \{A\mathbf{x}_1,\ A\mathbf{x}_2,\ \dots,\ A\mathbf{x}_n\}$ is also a basis of $\mathbb{R}^n$.

Solution. Let $\mathbf{x}$ be a vector in $\mathbb{R}^n$. Then $A^{-1}\mathbf{x}$ is in $\mathbb{R}^n$ so, since B is a basis, we have $A^{-1}\mathbf{x} = t_1\mathbf{x}_1 + t_2\mathbf{x}_2 + \cdots + t_n\mathbf{x}_n$ for t_i in $\mathbb{R}$. Left multiplication by A gives $\mathbf{x} = t_1(A\mathbf{x}_1) + t_2(A\mathbf{x}_2) + \cdots + t_n(A\mathbf{x}_n)$, and it follows that D spans $\mathbb{R}^n$. To show independence, let $s_1(A\mathbf{x}_1) + s_2(A\mathbf{x}_2) + \cdots + s_n(A\mathbf{x}_n) = \mathbf{0}$, where the s_i are in $\mathbb{R}$. Then $A(s_1\mathbf{x}_1 + s_2\mathbf{x}_2 + \cdots + s_n\mathbf{x}_n) = \mathbf{0}$ so left multiplication by A^{-1} gives $s_1\mathbf{x}_1 + s_2\mathbf{x}_2 + \cdots + s_n\mathbf{x}_n = \mathbf{0}$. Now the independence of B shows that each $s_i = 0$, and so proves the independence of D. Hence D is a basis of $\mathbb{R}^n$.

While we have found bases in many subspaces of $\mathbb{R}^n$, we have not yet shown that *every* subspace *has* a basis. This is part of the next theorem, the proof of which is deferred to Section 6.4 (Theorem 6.4.1) where it will be proved in more generality.

Theorem 5.2.6

Let $U \neq \{\mathbf{0}\}$ be a subspace of $\mathbb{R}^n$. Then:

1. *U has a basis and* $\dim U \leq n$.
2. *Any independent set in U can be enlarged (by adding vectors from any fixed basis of U) to a basis of U, if not already so.*
3. *Any spanning set for U can be cut down (by deleting vectors) to a basis of U, if not already so.*

Example 5.2.13

Find a basis of $\mathbb{R}^4$ containing $S = \{\mathbf{u},\ \mathbf{v}\}$ where $\mathbf{u} = (0,\ 1,\ 2,\ 3)$ and $\mathbf{v} = (2,\ -1,\ 0,\ 1)$.

Solution. By Theorem 5.2.6 we can find such a basis by adding vectors from the standard basis of $\mathbb{R}^4$ to S. If we try $\mathbf{e}_1 = (1,\ 0,\ 0,\ 0)$, we find easily that $\{\mathbf{e}_1,\ \mathbf{u},\ \mathbf{v}\}$ is independent.
Now add another vector from the standard basis, say $\mathbf{e}_2$. Again we find that $B = \{\mathbf{e}_1,\ \mathbf{e}_2,\ \mathbf{u},\ \mathbf{v}\}$ is independent. Since B has $4 = \dim \mathbb{R}^4$ vectors, then B must span $\mathbb{R}^4$ by Theorem 5.2.7 below (or simply verify it directly). Hence B is a basis of $\mathbb{R}^4$.

Theorem 5.2.6 has a number of useful consequences. Here is the first.

Theorem 5.2.7

Let U be a subspace of $\mathbb{R}^n$ where $\dim U = m$ and let $B = \{\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_m\}$ be a set of m vectors in U. Then B is independent if and only if B spans U.

Proof. Suppose B is independent. If B does not span U then, by Theorem 5.2.6, B can be enlarged to a basis of U containing more than m vectors. This contradicts the invariance theorem because $\dim U = m$, so B spans U. Conversely, if B spans U but is not independent, then B can be cut down to a basis of U containing fewer than m vectors, again a contradiction. So B is independent, as required. □

As we saw in Example 5.2.13, Theorem 5.2.7 is a "labour-saving" result. It asserts that, given a subspace U of dimension m and a set B of exactly m vectors in U, to prove that B is a basis of U it suffices to show either that B spans U or that B is independent. It is not necessary to verify both properties.

Theorem 5.2.8

Let $U \subseteq W$ be subspaces of $\mathbb{R}^n$. Then:

1. $\dim U \leq \dim W$.
2. *If* $\dim U = \dim W$, *then* $U = W$.

Proof. Write $\dim W = k$, and let B be a basis of U.

1. If $\dim U > k$, then B is an independent set in W containing more than k vectors, contradicting the fundamental theorem. So $\dim U \leq k = \dim W$.

2. If $\dim U = k$, then B is an independent set in W containing $k = \dim W$ vectors, so B spans W by Theorem 5.2.7. Hence $W = \text{span } B = U$, proving (2). □

It follows from Theorem 5.2.8 that if U is a subspace of $\mathbb{R}^n$, then $\dim U$ is one of the integers $0, 1, 2, \dots, n$, and that:

$$\begin{array}{lll} \dim U = 0 & \text{if and only if} & U = \{\mathbf{0}\}, \\ \dim U = n & \text{if and only if} & U = \mathbb{R}^n \end{array}$$

The other subspaces of $\mathbb{R}^n$ are called **proper**. The following example uses Theorem 5.2.8 to show that the proper subspaces of $\mathbb{R}^2$ are the lines through the origin, while the proper subspaces of $\mathbb{R}^3$ are the lines and planes through the origin.

Example 5.2.14

1. If U is a subspace of $\mathbb{R}^2$ or $\mathbb{R}^3$, then $\dim U = 1$ if and only if U is a line through the origin.

2. If U is a subspace of $\mathbb{R}^3$, then $\dim U = 2$ if and only if U is a plane through the origin.

Proof.

1. Since $\dim U = 1$, let $\{\mathbf{u}\}$ be a basis of U. Then $U = \text{span}\{\mathbf{u}\} = \{t\mathbf{u} \mid t \text{ in } \mathbb{R}\}$, so U is the line through the origin with direction vector $\mathbf{u}$. Conversely each line L with direction vector $\mathbf{d} \neq \mathbf{0}$ has the form $L = \{t\mathbf{d} \mid t \text{ in } \mathbb{R}\}$. Hence $\{\mathbf{d}\}$ is a basis of U, so U has dimension 1.

2. If $U \subseteq \mathbb{R}^3$ has dimension 2, let $\{\mathbf{v}, \mathbf{w}\}$ be a basis of U. Then $\mathbf{v}$ and $\mathbf{w}$ are not parallel (by Example 5.2.7) so $\mathbf{n} = \mathbf{v} \times \mathbf{w} \neq \mathbf{0}$. Let $P = \{\mathbf{x} \text{ in } \mathbb{R}^3 \mid \mathbf{n} \cdot \mathbf{x} = 0\}$ denote the plane through the origin with normal $\mathbf{n}$. Then P is a subspace of $\mathbb{R}^3$ (Example 5.1.1) and both $\mathbf{v}$ and $\mathbf{w}$ lie in P (they are orthogonal to $\mathbf{n}$), so $U = \text{span}\{\mathbf{v}, \mathbf{w}\} \subseteq P$ by Theorem 5.1.1. Hence

$$U \subseteq P \subseteq \mathbb{R}^3$$

Since $\dim U = 2$ and $\dim(\mathbb{R}^3) = 3$, it follows from Theorem 5.2.8 that $\dim P = 2$ or 3, whence $P = U$ or $\mathbb{R}^3$. But $P \neq \mathbb{R}^3$ (for example, $\mathbf{n}$ is not in P) and so $U = P$ is a plane through the origin.

Conversely, if U is a plane through the origin, then $\dim U = 0, 1, 2$, or 3 by Theorem 5.2.8. But $\dim U \neq 0$ or 3 because $U \neq \{\mathbf{0}\}$ and $U \neq \mathbb{R}^3$, and $\dim U \neq 1$ by (1). So $\dim U = 2$. □

Note that this proof shows that if $\mathbf{v}$ and $\mathbf{w}$ are nonzero, nonparallel vectors in $\mathbb{R}^3$, then $\text{span}\{\mathbf{v}, \mathbf{w}\}$ is the plane with normal $\mathbf{n} = \mathbf{v} \times \mathbf{w}$. We gave a geometrical verification of this fact in Section 5.1.

5.3 Orthogonality

Length and orthogonality are basic concepts in geometry and, in $\mathbb{R}^2$ and $\mathbb{R}^3$, they both can be defined using the dot product. In this section we extend the dot product to vectors in $\mathbb{R}^n$, and so endow $\mathbb{R}^n$ with euclidean geometry. We then introduce the idea of an orthogonal basis—one of the most useful concepts in linear algebra, and begin exploring some of its applications.

Dot Product, Length, and Distance

If $\mathbf{x} = (x_1,\ x_2,\ \dots,\ x_n)$ and $\mathbf{y} = (y_1,\ y_2,\ \dots,\ y_n)$ are two n-tuples in $\mathbb{R}^n$, recall that their **dot product** was defined in Section 2.2 as follows:

$$\mathbf{x}\cdot\mathbf{y} = x_1y_1 + x_2y_2 + \cdots + x_ny_n$$

Observe that if $\mathbf{x}$ and $\mathbf{y}$ are written as columns then $\mathbf{x}\cdot\mathbf{y} = \mathbf{x}^T\mathbf{y}$ is a matrix product (and $\mathbf{x}\cdot\mathbf{y} = \mathbf{x}\mathbf{y}^T$ if they are written as rows). Here $\mathbf{x}\cdot\mathbf{y}$ is a 1×1 matrix, which we take to be a number.

Definition 5.6 Length in $\mathbb{R}^n$

As in $\mathbb{R}^3$, the ***length*** $\|\mathbf{x}\|$ *of the vector is defined by*

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}\cdot\mathbf{x}} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

Where $\sqrt{(\quad)}$ indicates the positive square root.

A vector $\mathbf{x}$ of length 1 is called a **unit vector**. If $\mathbf{x} \neq \mathbf{0}$, then $\|\mathbf{x}\| \neq 0$ and it follows easily that $\frac{1}{\|\mathbf{x}\|}\mathbf{x}$ is a unit vector (see Theorem 5.3.6 below), a fact that we shall use later.

Example 5.3.1

If $\mathbf{x} = (1,\ -1,\ -3,\ 1)$ and $\mathbf{y} = (2,\ 1,\ 1,\ 0)$ in $\mathbb{R}^4$, then $\mathbf{x} \cdot \mathbf{y} = 2 - 1 - 3 + 0 = -2$ and $\|\mathbf{x}\| = \sqrt{1+1+9+1} = \sqrt{12} = 2\sqrt{3}$. Hence $\frac{1}{2\sqrt{3}}\mathbf{x}$ is a unit vector; similarly $\frac{1}{\sqrt{6}}\mathbf{y}$ is a unit vector.

These definitions agree with those in $\mathbb{R}^2$ and $\mathbb{R}^3$, and many properties carry over to $\mathbb{R}^n$:

Theorem 5.3.1

Let $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ denote vectors in $\mathbb{R}^n$. Then:

1. $\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x}$.
2. $\mathbf{x} \cdot (\mathbf{y} + \mathbf{z}) = \mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{z}$.
3. $(a\mathbf{x}) \cdot \mathbf{y} = a(\mathbf{x} \cdot \mathbf{y}) = \mathbf{x} \cdot (a\mathbf{y})$ *for all scalars* a.
4. $\|\mathbf{x}\|^2 = \mathbf{x} \cdot \mathbf{x}$.
5. $\|\mathbf{x}\| \geq 0$, *and* $\|\mathbf{x}\| = 0$ *if and only if* $\mathbf{x} = \mathbf{0}$.
6. $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$ *for all scalars* a.

Proof. (1), (2), and (3) follow from matrix arithmetic because $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T\mathbf{y}$; (4) is clear from the definition; and (6) is a routine verification since $|a| = \sqrt{a^2}$. If $\mathbf{x} = (x_1,\ x_2,\ \dots,\ x_n)$, then $\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$ so $\|\mathbf{x}\| = 0$ if and only if $x_1^2 + x_2^2 + \cdots + x_n^2 = 0$. Since each x_i is a real number this happens if and only if $x_i = 0$ for each i; that is, if and only if $\mathbf{x} = \mathbf{0}$. This proves (5). □

Because of Theorem 5.3.1, computations with dot products in $\mathbb{R}^n$ are similar to those in $\mathbb{R}^3$. In particular, the dot product

$$(\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_m) \cdot (\mathbf{y}_1 + \mathbf{y}_2 + \cdots + \mathbf{y}_k)$$

equals the sum of mk terms, $\mathbf{x}_i \cdot \mathbf{y}_j$, one for each choice of i and j. For example:

$$\begin{aligned}(3\mathbf{x} - 4\mathbf{y}) \cdot (7\mathbf{x} + 2\mathbf{y}) &= 21(\mathbf{x} \cdot \mathbf{x}) + 6(\mathbf{x} \cdot \mathbf{y}) - 28(\mathbf{y} \cdot \mathbf{x}) - 8(\mathbf{y} \cdot \mathbf{y}) \\ &= 21\|\mathbf{x}\|^2 - 22(\mathbf{x} \cdot \mathbf{y}) - 8\|\mathbf{y}\|^2\end{aligned}$$

holds for all vectors $\mathbf{x}$ and $\mathbf{y}$.

Example 5.3.2

Show that $\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + 2(\mathbf{x} \cdot \mathbf{y}) + \|\mathbf{y}\|^2$ for any $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$.

Solution. Using Theorem 5.3.1 several times:

$$\|\mathbf{x} + \mathbf{y}\|^2 = (\mathbf{x} + \mathbf{y}) \cdot (\mathbf{x} + \mathbf{y}) = \mathbf{x} \cdot \mathbf{x} + \mathbf{x} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{x} + \mathbf{y} \cdot \mathbf{y}$$

$$= \|\mathbf{x}\|^2 + 2(\mathbf{x}\cdot\mathbf{y}) + \|\mathbf{y}\|^2$$

Example 5.3.3

Suppose that $\mathbb{R}^n = \text{span}\,\{\mathbf{f}_1,\ \mathbf{f}_2,\ \dots,\ \mathbf{f}_k\}$ for some vectors $\mathbf{f}_i$. If $\mathbf{x}\cdot\mathbf{f}_i = 0$ for each i where $\mathbf{x}$ is in $\mathbb{R}^n$, show that $\mathbf{x} = \mathbf{0}$.

Solution. We show $\mathbf{x} = \mathbf{0}$ by showing that $\|\mathbf{x}\| = 0$ and using (5) of Theorem 5.3.1. Since the $\mathbf{f}_i$ span $\mathbb{R}^n$, write $\mathbf{x} = t_1\mathbf{f}_1 + t_2\mathbf{f}_2 + \cdots + t_k\mathbf{f}_k$ where the t_i are in $\mathbb{R}$. Then

$$\begin{aligned}\|\mathbf{x}\|^2 &= \mathbf{x}\cdot\mathbf{x} = \mathbf{x}\cdot(t_1\mathbf{f}_1 + t_2\mathbf{f}_2 + \cdots + t_k\mathbf{f}_k)\\ &= t_1(\mathbf{x}\cdot\mathbf{f}_1) + t_2(\mathbf{x}\cdot\mathbf{f}_2) + \cdots + t_k(\mathbf{x}\cdot\mathbf{f}_k)\\ &= t_1(0) + t_2(0) + \cdots + t_k(0)\\ &= 0\end{aligned}$$

We saw in Section 4.2 that if $\mathbf{u}$ and $\mathbf{v}$ are nonzero vectors in $\mathbb{R}^3$, then $\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} = \cos\theta$ where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$. Since $|\cos\theta| \le 1$ for any angle θ, this shows that $|\mathbf{u}\cdot\mathbf{v}| \le \|\mathbf{u}\|\|\mathbf{v}\|$. In this form the result holds in $\mathbb{R}^n$.

Theorem 5.3.2: Cauchy Inequality[9]

If $\mathbf{x}$ and $\mathbf{y}$ are vectors in $\mathbb{R}^n$, then

$$|\mathbf{x}\cdot\mathbf{y}| \le \|\mathbf{x}\|\|\mathbf{y}\|$$

Moreover $|\mathbf{x}\cdot\mathbf{y}| = \|\mathbf{x}\|\|\mathbf{y}\|$ if and only if one of $\mathbf{x}$ and $\mathbf{y}$ is a multiple of the other.

Proof. The inequality holds if $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$ (in fact it is equality). Otherwise, write $\|\mathbf{x}\| = a > 0$ and $\|\mathbf{y}\| = b > 0$ for convenience. A computation like that preceding Example 5.3.2 gives

$$\|b\mathbf{x} - a\mathbf{y}\|^2 = 2ab(ab - \mathbf{x}\cdot\mathbf{y}) \text{ and } \|b\mathbf{x} + a\mathbf{y}\|^2 = 2ab(ab + \mathbf{x}\cdot\mathbf{y}) \tag{5.1}$$

It follows that $ab - \mathbf{x}\cdot\mathbf{y} \ge 0$ and $ab + \mathbf{x}\cdot\mathbf{y} \ge 0$, and hence that $-ab \le \mathbf{x}\cdot\mathbf{y} \le ab$. Hence $|\mathbf{x}\cdot\mathbf{y}| \le ab = \|\mathbf{x}\|\|\mathbf{y}\|$, proving the Cauchy inequality.

If equality holds, then $|\mathbf{x}\cdot\mathbf{y}| = ab$, so $\mathbf{x}\cdot\mathbf{y} = ab$ or $\mathbf{x}\cdot\mathbf{y} = -ab$. Hence Equation 5.1 shows that $b\mathbf{x} - a\mathbf{y} = 0$ or $b\mathbf{x} + a\mathbf{y} = 0$, so one of $\mathbf{x}$ and $\mathbf{y}$ is a multiple of the other (even if $a = 0$ or $b = 0$). □

The Cauchy inequality is equivalent to $(\mathbf{x}\cdot\mathbf{y})^2 \le \|\mathbf{x}\|^2\|\mathbf{y}\|^2$. In $\mathbb{R}^5$ this becomes

$$(x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4 + x_5y_5)^2 \le (x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2)(y_1^2 + y_2^2 + y_3^2 + y_4^2 + y_5^2)$$

for all x_i and y_i in $\mathbb{R}$.

[9] Augustin Louis Cauchy (1789–1857) was born in Paris and became a professor at the École Polytechnique at the age of 26. He was one of the great mathematicians, producing more than 700 papers, and is best remembered for his work in analysis in which he established new standards of rigour and founded the theory of functions of a complex variable. He was a devout Catholic with a long-term interest in charitable work, and he was a royalist, following King Charles X into exile in Prague after he was deposed in 1830. Theorem 5.3.2 first appeared in his 1812 memoir on determinants.

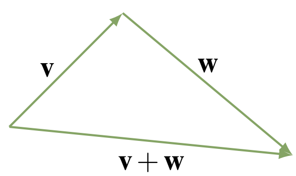
There is an important consequence of the Cauchy inequality. Given $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$, use Example 5.3.2 and the fact that $\mathbf{x} \cdot \mathbf{y} \leq \|\mathbf{x}\|\|\mathbf{y}\|$ to compute

$$\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + 2(\mathbf{x}\cdot\mathbf{y}) + \|\mathbf{y}\|^2 \leq \|\mathbf{x}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| + \|\mathbf{y}\|^2 = (\|\mathbf{x}\| + \|\mathbf{y}\|)^2$$

Taking positive square roots gives:

Corollary 5.3.1: Triangle Inequality

If $\mathbf{x}$ and $\mathbf{y}$ are vectors in $\mathbb{R}^n$, then $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

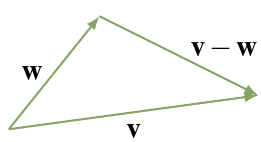


The reason for the name comes from the observation that in $\mathbb{R}^3$ the inequality asserts that the sum of the lengths of two sides of a triangle is not less than the length of the third side. This is illustrated in the diagram.

Definition 5.7 Distance in $\mathbb{R}^n$

If $\mathbf{x}$ and $\mathbf{y}$ are two vectors in $\mathbb{R}^n$, we define the ***distance*** *$d(\mathbf{x},\ \mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$ by*

$$d(\mathbf{x},\ \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$$



The motivation again comes from $\mathbb{R}^3$ as is clear in the diagram. This distance function has all the intuitive properties of distance in $\mathbb{R}^3$, including another version of the triangle inequality.

Theorem 5.3.3

If $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ are three vectors in $\mathbb{R}^n$ we have:

1. $d(\mathbf{x},\ \mathbf{y}) \geq 0$ *for all* $\mathbf{x}$ *and* $\mathbf{y}$.
2. $d(\mathbf{x},\ \mathbf{y}) = 0$ *if and only if* $\mathbf{x} = \mathbf{y}$.
3. $d(\mathbf{x},\ \mathbf{y}) = d(\mathbf{y},\ \mathbf{x})$ *for all* $\mathbf{x}$ *and* $\mathbf{y}$.
4. $d(\mathbf{x},\ \mathbf{z}) \leq d(\mathbf{x},\ \mathbf{y}) + d(\mathbf{y},\ \mathbf{z})$ *for all* $\mathbf{x}$, $\mathbf{y}$, *and* $\mathbf{z}$. *Triangle inequality.*

Proof. (1) and (2) restate part (5) of Theorem 5.3.1 because $d(\mathbf{x},\ \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$, and (3) follows because $\|\mathbf{u}\| = \|-\mathbf{u}\|$ for every vector $\mathbf{u}$ in $\mathbb{R}^n$. To prove (4) use the Corollary to Theorem 5.3.2:

$$\begin{aligned} d(\mathbf{x},\ \mathbf{z}) = \|\mathbf{x} - \mathbf{z}\| &= \|(\mathbf{x}-\mathbf{y}) + (\mathbf{y}-\mathbf{z})\| \\ &\leq \|(\mathbf{x}-\mathbf{y})\| + \|(\mathbf{y}-\mathbf{z})\| = d(\mathbf{x},\ \mathbf{y}) + d(\mathbf{y},\ \mathbf{z}) \end{aligned}$$

□

Orthogonal Sets and the Expansion Theorem

Definition 5.8 Orthogonal and Orthonormal Sets

We say that two vectors $\mathbf{x}$ *and* $\mathbf{y}$ *in* $\mathbb{R}^n$ *are* ***orthogonal*** *if* $\mathbf{x} \cdot \mathbf{y} = 0$*, extending the terminology in* $\mathbb{R}^3$ *(See Theorem 4.2.3). More generally, a set* $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ *of vectors in* $\mathbb{R}^n$ *is called an* ***orthogonal set*** *if*

$$\mathbf{x}_i \cdot \mathbf{x}_j = 0 \text{ for all } i \neq j \quad \text{ and } \quad \mathbf{x}_i \neq \mathbf{0} \text{ for all } i$$[10]

Note that $\{\mathbf{x}\}$ *is an orthogonal set if* $\mathbf{x} \neq \mathbf{0}$*. A set* $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ *of vectors in* $\mathbb{R}^n$ *is called* ***orthonormal*** *if it is orthogonal and, in addition, each* $\mathbf{x}_i$ *is a unit vector:*

$$\|\mathbf{x}_i\| = 1 \text{ for each } i.$$

Example 5.3.4

The standard basis $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ is an orthonormal set in $\mathbb{R}^n$.

The routine verification is left to the reader, as is the proof of:

Example 5.3.5

If $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ is orthogonal, so also is $\{a_1\mathbf{x}_1, a_2\mathbf{x}_2, \dots, a_k\mathbf{x}_k\}$ for any nonzero scalars a_i.

If $\mathbf{x} \neq \mathbf{0}$, it follows from item (6) of Theorem 5.3.1 that $\frac{1}{\|\mathbf{x}\|}\mathbf{x}$ is a unit vector, that is it has length 1.

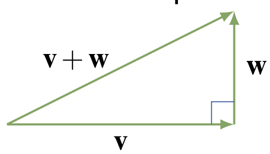
Definition 5.9 Normalizing an Orthogonal Set

Hence if $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ *is an orthogonal set, then* $\{\frac{1}{\|\mathbf{x}_1\|}\mathbf{x}_1, \frac{1}{\|\mathbf{x}_2\|}\mathbf{x}_2, \cdots, \frac{1}{\|\mathbf{x}_k\|}\mathbf{x}_k\}$ *is an orthonormal set, and we say that it is the result of* ***normalizing*** *the orthogonal set* $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k\}$*.*

Example 5.3.6

If $\mathbf{f}_1 = \left[\begin{array}{r} 1 \\ 1 \\ 1 \\ -1 \end{array}\right]$, $\mathbf{f}_2 = \left[\begin{array}{r} 1 \\ 0 \\ 1 \\ 2 \end{array}\right]$, $\mathbf{f}_3 = \left[\begin{array}{r} -1 \\ 0 \\ 1 \\ 0 \end{array}\right]$, and $\mathbf{f}_4 = \left[\begin{array}{r} -1 \\ 3 \\ -1 \\ 1 \end{array}\right]$ then $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{f}_4\}$ is an orthogonal set in $\mathbb{R}^4$ as is easily verified. After normalizing, the corresponding orthonormal set is $\{\frac{1}{2}\mathbf{f}_1, \frac{1}{\sqrt{6}}\mathbf{f}_2, \frac{1}{\sqrt{2}}\mathbf{f}_3, \frac{1}{2\sqrt{3}}\mathbf{f}_4\}$

[10]The reason for insisting that orthogonal sets consist of *nonzero* vectors is that we will be primarily concerned with orthogonal bases.



The most important result about orthogonality is Pythagoras' theorem. Given orthogonal vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^3$, it asserts that

$$\|\mathbf{v}+\mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2$$

as in the diagram. In this form the result holds for any orthogonal set in $\mathbb{R}^n$.

Theorem 5.3.4: Pythagoras' Theorem

If $\{\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_k\}$ is an orthogonal set in $\mathbb{R}^n$, then

$$\|\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_k\|^2 = \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2 + \cdots + \|\mathbf{x}_k\|^2.$$

Proof. The fact that $\mathbf{x}_i \cdot \mathbf{x}_j = 0$ whenever $i \neq j$ gives

$$\begin{aligned}\|\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_k\|^2 &= (\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_k)\cdot(\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_k)\\ &= (\mathbf{x}_1\cdot\mathbf{x}_1+\mathbf{x}_2\cdot\mathbf{x}_2+\cdots+\mathbf{x}_k\cdot\mathbf{x}_k)+\sum_{i\neq j}\mathbf{x}_i\cdot\mathbf{x}_j\\ &= \|\mathbf{x}_1\|^2+\|\mathbf{x}_2\|^2+\cdots+\|\mathbf{x}_k\|^2+0\end{aligned}$$

This is what we wanted. □

If $\mathbf{v}$ and $\mathbf{w}$ are orthogonal, nonzero vectors in $\mathbb{R}^3$, then they are certainly not parallel, and so are linearly independent by Example 5.2.7. The next theorem gives a far-reaching extension of this observation.

Theorem 5.3.5

Every orthogonal set in $\mathbb{R}^n$ is linearly independent.

Proof. Let $\{\mathbf{x}_1,\ \mathbf{x}_2,\ \dots,\ \mathbf{x}_k\}$ be an orthogonal set in $\mathbb{R}^n$ and suppose a linear combination vanishes, say: $t_1\mathbf{x}_1+t_2\mathbf{x}_2+\cdots+t_k\mathbf{x}_k=\mathbf{0}$. Then

$$\begin{aligned}0=\mathbf{x}_1\cdot\mathbf{0} &= \mathbf{x}_1\cdot(t_1\mathbf{x}_1+t_2\mathbf{x}_2+\cdots+t_k\mathbf{x}_k)\\ &= t_1(\mathbf{x}_1\cdot\mathbf{x}_1)+t_2(\mathbf{x}_1\cdot\mathbf{x}_2)+\cdots+t_k(\mathbf{x}_1\cdot\mathbf{x}_k)\\ &= t_1\|\mathbf{x}_1\|^2+t_2(0)+\cdots+t_k(0)\\ &= t_1\|\mathbf{x}_1\|^2\end{aligned}$$

Since $\|\mathbf{x}_1\|^2 \neq 0$, this implies that $t_1=0$. Similarly $t_i=0$ for each i. □

Theorem 5.3.5 suggests considering orthogonal bases for $\mathbb{R}^n$, that is orthogonal sets that span $\mathbb{R}^n$. These turn out to be the best bases in the sense that, when expanding a vector as a linear combination of the basis vectors, there are explicit formulas for the coefficients.

Theorem 5.3.6: Expansion Theorem

Let $\{\mathbf{f}_1,\ \mathbf{f}_2,\ \dots,\ \mathbf{f}_m\}$ be an orthogonal basis of a subspace U of $\mathbb{R}^n$. If $\mathbf{x}$ is any vector in U, we have

$$\mathbf{x} = \left(\frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\right)\mathbf{f}_1 + \left(\frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\right)\mathbf{f}_2 + \cdots + \left(\frac{\mathbf{x}\cdot\mathbf{f}_m}{\|\mathbf{f}_m\|^2}\right)\mathbf{f}_m$$

Proof. Since $\{\mathbf{f}_1,\ \mathbf{f}_2,\ \dots,\ \mathbf{f}_m\}$ spans U, we have $\mathbf{x} = t_1\mathbf{f}_1 + t_2\mathbf{f}_2 + \cdots + t_m\mathbf{f}_m$ where the t_i are scalars. To find t_1 we take the dot product of both sides with $\mathbf{f}_1$:

$$\begin{aligned}
\mathbf{x}\cdot\mathbf{f}_1 &= (t_1\mathbf{f}_1 + t_2\mathbf{f}_2 + \cdots + t_m\mathbf{f}_m)\cdot\mathbf{f}_1 \\
&= t_1(\mathbf{f}_1\cdot\mathbf{f}_1) + t_2(\mathbf{f}_2\cdot\mathbf{f}_1) + \cdots + t_m(\mathbf{f}_m\cdot\mathbf{f}_1) \\
&= t_1\|\mathbf{f}_1\|^2 + t_2(0) + \cdots + t_m(0) \\
&= t_1\|\mathbf{f}_1\|^2
\end{aligned}$$

Since $\mathbf{f}_1 \neq \mathbf{0}$, this gives $t_1 = \frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}$. Similarly, $t_i = \frac{\mathbf{x}\cdot\mathbf{f}_i}{\|\mathbf{f}_i\|^2}$ for each i. □

The expansion in Theorem 5.3.6 of $\mathbf{x}$ as a linear combination of the orthogonal basis $\{\mathbf{f}_1,\ \mathbf{f}_2,\ \dots,\ \mathbf{f}_m\}$ is called the **Fourier expansion** of $\mathbf{x}$, and the coefficients $t_1 = \frac{\mathbf{x}\cdot\mathbf{f}_i}{\|\mathbf{f}_i\|^2}$ are called the **Fourier coefficients**. Note that if $\{\mathbf{f}_1,\ \mathbf{f}_2,\ \dots,\ \mathbf{f}_m\}$ is actually orthonormal, then $t_i = \mathbf{x}\cdot\mathbf{f}_i$ for each i. We will have a great deal more to say about this in Section 10.5.

Example 5.3.7

Expand $\mathbf{x} = (a,\ b,\ c,\ d)$ as a linear combination of the orthogonal basis $\{\mathbf{f}_1,\ \mathbf{f}_2,\ \mathbf{f}_3,\ \mathbf{f}_4\}$ of $\mathbb{R}^4$ given in Example 5.3.6.

Solution. We have $\mathbf{f}_1 = (1,\ 1,\ 1,\ -1)$, $\mathbf{f}_2 = (1,\ 0,\ 1,\ 2)$, $\mathbf{f}_3 = (-1,\ 0,\ 1,\ 0)$, and $\mathbf{f}_4 = (-1,\ 3,\ -1,\ 1)$ so the Fourier coefficients are

$$\begin{aligned}
t_1 &= \frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2} = \frac{1}{4}(a+b+c-d) & t_3 &= \frac{\mathbf{x}\cdot\mathbf{f}_3}{\|\mathbf{f}_3\|^2} = \frac{1}{2}(-a+c) \\
t_2 &= \frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2} = \frac{1}{6}(a+c+2d) & t_4 &= \frac{\mathbf{x}\cdot\mathbf{f}_4}{\|\mathbf{f}_4\|^2} = \frac{1}{12}(-a+3b-c+d)
\end{aligned}$$

The reader can verify that indeed $\mathbf{x} = t_1\mathbf{f}_1 + t_2\mathbf{f}_2 + t_3\mathbf{f}_3 + t_4\mathbf{f}_4$.

A natural question arises here: Does every subspace U of $\mathbb{R}^n$ *have* an orthogonal basis? The answer is "yes"; in fact, there is a systematic procedure, called the Gram-Schmidt algorithm, for turning any basis of U into an orthogonal one. This leads to a definition of the projection onto a subspace U that generalizes the projection along a vector used in $\mathbb{R}^2$ and $\mathbb{R}^3$. All this is discussed in Section 8.1.

5.4 Rank of a Matrix

In this section we use the concept of dimension to clarify the definition of the rank of a matrix given in Section 1.2, and to study its properties. This requires that we deal with rows and columns in the same way. While it has been our custom to write the n-tuples in $\mathbb{R}^n$ as columns, in this section we will frequently write them as rows. Subspaces, independence, spanning, and dimension are defined for rows using matrix operations, just as for columns. If A is an $m \times n$ matrix, we define:

Definition 5.10 Column and Row Space of a Matrix

*The **column space**, $\text{col } A$, of A is the subspace of $\mathbb{R}^m$ spanned by the columns of A.*
*The **row space**, $\text{row } A$, of A is the subspace of $\mathbb{R}^n$ spanned by the rows of A.*

Much of what we do in this section involves these subspaces. We begin with:

Lemma 5.4.1

Let A and B denote $m \times n$ matrices.

1. *If $A \to B$ by elementary row operations, then $\text{row } A = \text{row } B$.*
2. *If $A \to B$ by elementary column operations, then $\text{col } A = \text{col } B$.*

Proof. We prove (1); the proof of (2) is analogous. It is enough to do it in the case when $A \to B$ by a single row operation. Let $R_1, R_2, \dots, R_m$ denote the rows of A. The row operation $A \to B$ either interchanges two rows, multiplies a row by a nonzero constant, or adds a multiple of a row to a different row. We leave the first two cases to the reader. In the last case, suppose that a times row p is added to row q where $p < q$. Then the rows of B are $R_1, \dots, R_p, \dots, R_q + aR_p, \dots, R_m$, and Theorem 5.1.1 shows that

$$\text{span}\,\{R_1, \dots, R_p, \dots, R_q, \dots, R_m\} = \text{span}\,\{R_1, \dots, R_p, \dots, R_q + aR_p, \dots, R_m\}$$

That is, $\text{row}\, A = \text{row}\, B$. □

If A is any matrix, we can carry $A \to R$ by elementary row operations where R is a row-echelon matrix. Hence $\text{row}\, A = \text{row}\, R$ by Lemma 5.4.1; so the first part of the following result is of interest.

Lemma 5.4.2

If R is a row-echelon matrix, then

1. *The nonzero rows of R are a basis of* $\text{row}\, R$.
2. *The columns of R containing leading ones are a basis of* $\text{col}\, R$.

Proof. The rows of R are independent by Example 5.2.6, and they span $\text{row}\, R$ by definition. This proves (1).

Let $\mathbf{c}_{j_1}, \mathbf{c}_{j_2}, \dots, \mathbf{c}_{j_r}$ denote the columns of R containing leading 1s. Then $\{\mathbf{c}_{j_1}, \mathbf{c}_{j_2}, \dots, \mathbf{c}_{j_r}\}$ is independent because the leading 1s are in different rows (and have zeros below and to the left of them). Let U denote the subspace of all columns in $\mathbb{R}^m$ in which the last $m-r$ entries are zero. Then $\dim U = r$ (it is just $\mathbb{R}^r$ with extra zeros). Hence the independent set $\{\mathbf{c}_{j_1}, \mathbf{c}_{j_2}, \dots, \mathbf{c}_{j_r}\}$ is a basis of U by Theorem 5.2.7. Since each $\mathbf{c}_{j_i}$ is in $\text{col}\, R$, it follows that $\text{col}\, R = U$, proving (2). □

With Lemma 5.4.2 we can fill a gap in the definition of the rank of a matrix given in Chapter 1. Let A be any matrix and suppose A is carried to some row-echelon matrix R by row operations. Note that R is not unique. In Section 1.2 we defined the **rank** of A, denoted $\text{rank}\, A$, to be the number of leading 1s in R, that is the number of nonzero rows of R. The fact that this number does not depend on the choice of R was not proved in Section 1.2. However part 1 of Lemma 5.4.2 shows that

$$\text{rank}\, A = \dim\,(\text{row}\, A)$$

and hence that $\text{rank}\, A$ is independent of R.

Lemma 5.4.2 can be used to find bases of subspaces of $\mathbb{R}^n$ (written as rows). Here is an example.

Example 5.4.1

Find a basis of $U = \text{span}\,\{(1, 1, 2, 3), (2, 4, 1, 0), (1, 5, -4, -9)\}$.

Solution. U is the row space of $\begin{bmatrix} 1 & 1 & 2 & 3 \\ 2 & 4 & 1 & 0 \\ 1 & 5 & -4 & -9 \end{bmatrix}$. This matrix has row-echelon form

$\begin{bmatrix} 1 & 1 & 2 & 3 \\ 0 & 1 & -\frac{3}{2} & -3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$, so $\{(1,\ 1,\ 2,\ 3),\ (0,\ 1,\ -\frac{3}{2},\ -3)\}$ is basis of U by Lemma 5.4.2.
Note that $\{(1,\ 1,\ 2,\ 3),\ (0,\ 2,\ -3,\ -6)\}$ is another basis that avoids fractions.

Lemmas 5.4.1 and 5.4.2 are enough to prove the following fundamental theorem.

Theorem 5.4.1: Rank Theorem

Let A denote any $m \times n$ matrix of rank r. Then

$$\dim(\operatorname{col} A) = \dim(\operatorname{row} A) = r$$

Moreover, if A is carried to a row-echelon matrix R by row operations, then

1. *The r nonzero rows of R are a basis of* $\operatorname{row} A$.
2. *If the leading 1s lie in columns $j_1,\ j_2,\ \dots,\ j_r$ of R, then columns $j_1,\ j_2,\ \dots,\ j_r$ of A are a basis of* $\operatorname{col} A$.

Proof. We have $\operatorname{row} A = \operatorname{row} R$ by Lemma 5.4.1, so (1) follows from Lemma 5.4.2. Moreover, $R = UA$ for some invertible matrix U by Theorem 2.5.1. Now write $A = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \dots & \mathbf{c}_n \end{bmatrix}$ where $\mathbf{c}_1,\ \mathbf{c}_2,\ \dots,\ \mathbf{c}_n$ are the columns of A. Then

$$R = UA = U\begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_n \end{bmatrix} = \begin{bmatrix} U\mathbf{c}_1 & U\mathbf{c}_2 & \cdots & U\mathbf{c}_n \end{bmatrix}$$

Thus, in the notation of (2), the set $B = \{U\mathbf{c}_{j_1},\ U\mathbf{c}_{j_2},\ \dots,\ U\mathbf{c}_{j_r}\}$ is a basis of $\operatorname{col} R$ by Lemma 5.4.2. So, to prove (2) and the fact that $\dim(\operatorname{col} A) = r$, it is enough to show that $D = \{\mathbf{c}_{j_1},\ \mathbf{c}_{j_2},\ \dots,\ \mathbf{c}_{j_r}\}$ is a basis of $\operatorname{col} A$. First, D is linearly independent because U is invertible (verify), so we show that, for each j, column $\mathbf{c}_j$ is a linear combination of the $\mathbf{c}_{j_i}$. But $U\mathbf{c}_j$ is column j of R, and so is a linear combination of the $U\mathbf{c}_{j_i}$, say $U\mathbf{c}_j = a_1U\mathbf{c}_{j_1} + a_2U\mathbf{c}_{j_2} + \cdots + a_rU\mathbf{c}_{j_r}$ where each a_i is a real number.

Since U is invertible, it follows that $\mathbf{c}_j = a_1\mathbf{c}_{j_1} + a_2\mathbf{c}_{j_2} + \cdots + a_r\mathbf{c}_{j_r}$ and the proof is complete. □

Example 5.4.2

Compute the rank of $A = \begin{bmatrix} 1 & 2 & 2 & -1 \\ 3 & 6 & 5 & 0 \\ 1 & 2 & 1 & 2 \end{bmatrix}$ and find bases for $\operatorname{row} A$ and $\operatorname{col} A$.

Solution. The reduction of A to row-echelon form is as follows:

$$\begin{bmatrix} 1 & 2 & 2 & -1 \\ 3 & 6 & 5 & 0 \\ 1 & 2 & 1 & 2 \end{bmatrix} \to \begin{bmatrix} 1 & 2 & 2 & -1 \\ 0 & 0 & -1 & 3 \\ 0 & 0 & -1 & 3 \end{bmatrix} \to \begin{bmatrix} 1 & 2 & 2 & -1 \\ 0 & 0 & -1 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Hence $\operatorname{rank} A = 2$, and $\left\{\begin{bmatrix} 1 & 2 & 2 & -1 \end{bmatrix},\ \begin{bmatrix} 0 & 0 & 1 & -3 \end{bmatrix}\right\}$ is a basis of $\operatorname{row} A$ by Lemma 5.4.2. Since the leading 1s are in columns 1 and 3 of the row-echelon matrix, Theorem 5.4.1 shows that columns 1 and 3 of A are a basis $\left\{ \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 5 \\ 1 \end{bmatrix} \right\}$ of $\text{col } A$.

Theorem 5.4.1 has several important consequences. The first, Corollary 5.4.1 below, follows because the rows of A are independent (respectively span $\text{row } A$) if and only if their transposes are independent (respectively span $\text{col } A$).

Corollary 5.4.1

If A is any matrix, then $\text{rank } A = \text{rank } (A^T)$.

If A is an $m \times n$ matrix, we have $\text{col } A \subseteq \mathbb{R}^m$ and $\text{row } A \subseteq \mathbb{R}^n$. Hence Theorem 5.2.8 shows that $\dim(\text{col } A) \leq \dim(\mathbb{R}^m) = m$ and $\dim(\text{row } A) \leq \dim(\mathbb{R}^n) = n$. Thus Theorem 5.4.1 gives:

Corollary 5.4.2

If A is an $m \times n$ matrix, then $\text{rank } A \leq m$ *and* $\text{rank } A \leq n$.

Corollary 5.4.3

$\text{rank } A = \text{rank } (UA) = \text{rank } (AV)$ *whenever U and V are invertible.*

Proof. Lemma 5.4.1 gives $\text{rank } A = \text{rank } (UA)$. Using this and Corollary 5.4.1 we get

$$\text{rank } (AV) = \text{rank } (AV)^T = \text{rank } (V^T A^T) = \text{rank } (A^T) = \text{rank } A$$

The next corollary requires a preliminary lemma. □

Lemma 5.4.3

Let A, U, and V be matrices of sizes $m \times n$, $p \times m$, and $n \times q$ respectively.

1. $\text{col } (AV) \subseteq \text{col } A$, *with equality if $VV' = I_n$ for some V'.*
2. $\text{row } (UA) \subseteq \text{row } A$, *with equality if $U'U = I_m$ for some U'.*

Proof. For (1), write $V = \begin{bmatrix} \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_q \end{bmatrix}$ where $\mathbf{v}_j$ is column j of V. Then we have $AV = \begin{bmatrix} A\mathbf{v}_1, A\mathbf{v}_2, \dots, A\mathbf{v}_q \end{bmatrix}$, and each $A\mathbf{v}_j$ is in $\text{col } A$ by Definition 2.4. It follows that $\text{col } (AV) \subseteq \text{col } A$. If $VV' = I_n$, we obtain $\text{col } A = \text{col } [(AV)V'] \subseteq \text{col } (AV)$ in the same way. This proves (1).

As to (2), we have $\text{col } \left[(UA)^T \right] = \text{col } (A^T U^T) \subseteq \text{col } (A^T)$ by (1), from which $\text{row } (UA) \subseteq \text{row } A$. If $U'U = I_m$, this is equality as in the proof of (1). □

Corollary 5.4.4

If A is $m \times n$ and B is $n \times m$, then $\operatorname{rank} AB \leq \operatorname{rank} A$ *and* $\operatorname{rank} AB \leq \operatorname{rank} B$.

Proof. By Lemma 5.4.3, $\operatorname{col}(AB) \subseteq \operatorname{col} A$ and $\operatorname{row}(BA) \subseteq \operatorname{row} A$, so Theorem 5.4.1 applies. □

In Section 5.1 we discussed two other subspaces associated with an $m \times n$ matrix A: the null space $\operatorname{null}(A)$ and the image space $\operatorname{im}(A)$

$$\operatorname{null}(A) = \{\mathbf{x} \text{ in } \mathbb{R}^n \mid A\mathbf{x} = \mathbf{0}\} \text{ and } \operatorname{im}(A) = \{A\mathbf{x} \mid \mathbf{x} \text{ in } \mathbb{R}^n\}$$

Before we proceed to an important theorem, we first define what is meant by the nullity of a matrix.

Definition 5.11 Nullity

The dimension of the null space of a matrix is called the nullity, denoted by $\dim[\operatorname{null}(A)]$.

We will see shortly that the rank and the nullity of an $m \times n$ matrix A add up to n (no matter m!), this is part (1) of the following theorem.

Recall that $\operatorname{im}(A) = \operatorname{col}(A)$ by Example 5.1.8. So if A has rank r, we have $\dim[\operatorname{im}(A)] = \dim[\operatorname{col}(A)] = r$. Hence Theorem 5.4.1 provides a method of finding a basis of $\operatorname{im}(A)$ and this is recorded as part (3) of the following theorem.

Theorem 5.4.2: Rank and Nullity

Let A denote an $m \times n$ matrix of rank r. Then

1. $\operatorname{rank}(A) + \dim[\operatorname{null}(A)] = n$.
2. *The $n - r$ basic solutions to the system $A\mathbf{x} = \mathbf{0}$ provided by the gaussian algorithm are a basis of* $\operatorname{null}(A)$*, so* $\dim[\operatorname{null}(A)] = n - r$.
3. *Theorem 5.4.1 provides a basis of* $\operatorname{im}(A) = \operatorname{col}(A)$*, and* $\dim[\operatorname{im}(A)] = r$.

Proof. Part (1) follows from part (2), which only remains to be proved. We already know (Theorem 2.2.1) that $\operatorname{null}(A)$ is spanned by the $n - r$ basic solutions of $A\mathbf{x} = \mathbf{0}$. Hence using Theorem 5.2.7, it suffices to show that $\dim[\operatorname{null}(A)] = n - r$. So let $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ be a basis of $\operatorname{null}(A)$, and extend it to a basis $\{\mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{x}_{k+1}, \dots, \mathbf{x}_n\}$ of $\mathbb{R}^n$ (by Theorem 5.2.6). It is enough to show that $\{A\mathbf{x}_{k+1}, \dots, A\mathbf{x}_n\}$ is a basis of $\operatorname{im}(A)$; then $n - k = r$ by the above and so $k = n - r$ as required.

Spanning. Choose $A\mathbf{x}$ in $\operatorname{im}(A)$, $\mathbf{x}$ in $\mathbb{R}^n$, and write $\mathbf{x} = a_1\mathbf{x}_1 + \cdots + a_k\mathbf{x}_k + a_{k+1}\mathbf{x}_{k+1} + \cdots + a_n\mathbf{x}_n$ where the a_i are in $\mathbb{R}$. Then $A\mathbf{x} = a_{k+1}A\mathbf{x}_{k+1} + \cdots + a_nA\mathbf{x}_n$ because $\{\mathbf{x}_1, \dots, \mathbf{x}_k\} \subseteq \operatorname{null}(A)$.

Independence. Let $t_{k+1}A\mathbf{x}_{k+1} + \cdots + t_nA\mathbf{x}_n = \mathbf{0}$, t_i in $\mathbb{R}$. Then $t_{k+1}\mathbf{x}_{k+1} + \cdots + t_n\mathbf{x}_n$ is in $\operatorname{null} A$, so $t_{k+1}\mathbf{x}_{k+1} + \cdots + t_n\mathbf{x}_n = t_1\mathbf{x}_1 + \cdots + t_k\mathbf{x}_k$ for some $t_1, \dots, t_k$ in $\mathbb{R}$. But then the independence of the $\mathbf{x}_i$ shows that $t_i = 0$ for every i. □

We can now see this result in practice.

Example 5.4.3

If $A = \begin{bmatrix} 1 & -2 & 1 & 1 \\ -1 & 2 & 0 & 1 \\ 2 & -4 & 1 & 0 \end{bmatrix}$, find bases of $\text{null}(A)$ and $\text{im}(A)$, and so find their dimensions.

Solution. If $\mathbf{x}$ is in $\text{null}(A)$, then $A\mathbf{x} = \mathbf{0}$, so $\mathbf{x}$ is given by solving the system $A\mathbf{x} = \mathbf{0}$. The reduction of the augmented matrix to reduced form is

$$\left[\begin{array}{rrrr|r} 1 & -2 & 1 & 1 & 0 \\ -1 & 2 & 0 & 1 & 0 \\ 2 & -4 & 1 & 0 & 0 \end{array}\right] \to \left[\begin{array}{rrrr|r} 1 & -2 & 0 & -1 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

Hence $r = \text{rank}(A) = 2$. Here, $\text{im}(A) = \text{col}(A)$ has basis $\left\{ \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \right\}$ by Theorem 5.4.1 because the leading 1s are in columns 1 and 3. In particular, $\dim[\text{im}(A)] = 2 = r$ as in Theorem 5.4.2.

Turning to $\text{null}(A)$, we use gaussian elimination. The leading variables are x_1 and x_3, so the nonleading variables become parameters: $x_2 = s$ and $x_4 = t$. It follows from the reduced matrix that $x_1 = 2s + t$ and $x_3 = -2t$, so the general solution is

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 2s+t \\ s \\ -2t \\ t \end{bmatrix} = s\mathbf{x}_1 + t\mathbf{x}_2 \text{ where } \mathbf{x}_1 = \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \text{ and } \mathbf{x}_2 = \begin{bmatrix} 1 \\ 0 \\ -2 \\ 1 \end{bmatrix}.$$

Hence $\text{null}(A)$. But $\mathbf{x}_1$ and $\mathbf{x}_2$ *are* solutions (basic), so

$$\text{null}(A) = \text{span}\{\mathbf{x}_1, \mathbf{x}_2\}$$

However Theorem 5.4.2 asserts that $\{\mathbf{x}_1, \mathbf{x}_2\}$ is a *basis* of $\text{null}(A)$. (In fact it is easy to verify directly that $\{\mathbf{x}_1, \mathbf{x}_2\}$ is independent in this case.) In particular, $\dim[\text{null}(A)] = 2 = n - r$, as Theorem 5.4.2 asserts.

Let A be an $m \times n$ matrix. Corollary 5.4.2 of Theorem 5.4.1 asserts that $\text{rank}\, A \leq m$ and $\text{rank}\, A \leq n$, and it is natural to ask when these extreme cases arise. If $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n$ are the columns of A, Theorem 5.2.2 shows that $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\}$ spans $\mathbb{R}^m$ if and only if the system $A\mathbf{x} = \mathbf{b}$ is consistent for every $\mathbf{b}$ in $\mathbb{R}^m$, and that $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\}$ is independent if and only if $A\mathbf{x} = \mathbf{0}$, $\mathbf{x}$ in $\mathbb{R}^n$, implies $\mathbf{x} = \mathbf{0}$. The next two useful theorems improve on both these results, and relate them to when the rank of A is n or m.

Theorem 5.4.3

The following are equivalent for an $m \times n$ matrix A:

1. $\text{rank}\, A = n$.

2. *The rows of A span $\mathbb{R}^n$.*

3. *The columns of A are linearly independent in $\mathbb{R}^m$.*

4. *The $n \times n$ matrix $A^T A$ is invertible.*

5. *$CA = I_n$ for some $n \times m$ matrix C.*

6. *If $A\mathbf{x} = \mathbf{0}$, $\mathbf{x}$ in $\mathbb{R}^n$, then $\mathbf{x} = \mathbf{0}$.*

Proof. (1) $\Rightarrow$ (2). We have row $A \subseteq \mathbb{R}^n$, and $\dim(\text{row } A) = n$ by (1), so row $A = \mathbb{R}^n$ by Theorem 5.2.8. This is (2).

(2) $\Rightarrow$ (3). By (2), row $A = \mathbb{R}^n$, so rank $A = n$. This means $\dim(\text{col } A) = n$. Since the n columns of A span col A, they are independent by Theorem 5.2.7.

(3) $\Rightarrow$ (4). If $(A^T A)\mathbf{x} = \mathbf{0}$, $\mathbf{x}$ in $\mathbb{R}^n$, we show that $\mathbf{x} = \mathbf{0}$ (Theorem 2.4.5). We have

$$||A\mathbf{x}||^2 = (A\mathbf{x})^T A\mathbf{x} = \mathbf{x}^T A^T A\mathbf{x} = \mathbf{x}^T \mathbf{0} = \mathbf{0}$$

Hence $A\mathbf{x} = \mathbf{0}$, so $\mathbf{x} = \mathbf{0}$ by (3) and Theorem 5.2.2.

(4) $\Rightarrow$ (5). Given (4), take $C = (A^T A)^{-1} A^T$.

(5) $\Rightarrow$ (6). If $A\mathbf{x} = \mathbf{0}$, then left multiplication by C (from (5)) gives $\mathbf{x} = \mathbf{0}$.

(6) $\Rightarrow$ (1). Given (6), the columns of A are independent by Theorem 5.2.2. Hence $\dim(\text{col } A) = n$, and (1) follows. □

Theorem 5.4.4

The following are equivalent for an $m \times n$ matrix A:

1. rank $A = m$.

2. *The columns of A span $\mathbb{R}^m$.*

3. *The rows of A are linearly independent in $\mathbb{R}^n$.*

4. *The $m \times m$ matrix AA^T is invertible.*

5. *$AC = I_m$ for some $n \times m$ matrix C.*

6. *The system $A\mathbf{x} = \mathbf{b}$ is consistent for every $\mathbf{b}$ in $\mathbb{R}^m$.*

Proof. (1) $\Rightarrow$ (2). By (1), $\dim(\text{col } A = m$, so col $A = \mathbb{R}^m$ by Theorem 5.2.8.

(2) $\Rightarrow$ (3). By (2), col $A = \mathbb{R}^m$, so rank $A = m$. This means $\dim(\text{row } A) = m$. Since the m rows of A span row A, they are independent by Theorem 5.2.7.

(3) $\Rightarrow$ (4). We have rank $A = m$ by (3), so the $n \times m$ matrix A^T has rank m. Hence applying Theorem 5.4.3 to A^T in place of A shows that $(A^T)^T A^T$ is invertible, proving (4).

(4) $\Rightarrow$ (5). Given (4), take $C = A^T (AA_T)^{-1}$ in (5).

(5) $\Rightarrow$ (6). Comparing columns in $AC = I_m$ gives $A\mathbf{c}_j = \mathbf{e}_j$ for each j, where $\mathbf{c}_j$ and $\mathbf{e}_j$ denote column j of C and I_m respectively. Given $\mathbf{b}$ in $\mathbb{R}^m$, write $\mathbf{b} = \sum_{j=1}^m r_j \mathbf{e}_j$, r_j in $\mathbb{R}$. Then $A\mathbf{x} = \mathbf{b}$ holds with $\mathbf{x} = \sum_{j=1}^m r_j \mathbf{c}_j$ as the reader can verify.

(6) $\Rightarrow$ (1). Given (6), the columns of A span $\mathbb{R}^m$ by Theorem 5.2.2. Thus $\text{col } A = \mathbb{R}^m$ and (1) follows. □

Example 5.4.4

Show that $\begin{bmatrix} 3 & x+y+z \\ x+y+z & x^2+y^2+z^2 \end{bmatrix}$ is invertible if x, y, and z are not all equal.

Solution. The given matrix has the form $A^T A$ where $A = \begin{bmatrix} 1 & x \\ 1 & y \\ 1 & z \end{bmatrix}$ has independent columns because x, y, and z are not all equal (verify). Hence Theorem 5.4.3 applies.

Theorem 5.4.3 and Theorem 5.4.4 relate several important properties of an $m \times n$ matrix A to the invertibility of the square, symmetric matrices $A^T A$ and AA^T. In fact, even if the columns of A are not independent or do not span $\mathbb{R}^m$, the matrices $A^T A$ and AA^T are both symmetric and, as such, have real eigenvalues as we shall see. We return to this in Chapter 7.

5.5 Similarity and Diagonalization

In Section 3.3 we studied diagonalization of a square matrix A, and found important applications (for example to linear dynamical systems). We can now utilize the concepts of subspace, basis, and dimension to clarify the diagonalization process, reveal some new results, and prove some theorems which could not be demonstrated in Section 3.3.

Before proceeding, we introduce a notion that simplifies the discussion of diagonalization, and is used throughout the book.

Similar Matrices

Definition 5.12 Similar Matrices

*If A and B are $n \times n$ matrices, we say that A and B are **similar**, and write $A \sim B$, if $B = P^{-1}AP$ for some invertible matrix P.*

Note that $A \sim B$ if and only if $B = QAQ^{-1}$ where Q is invertible (write $P^{-1} = Q$). The language of similarity is used throughout linear algebra. For example, a matrix A is diagonalizable if and only if it is similar to a diagonal matrix.

If $A \sim B$, then necessarily $B \sim A$. To see why, suppose that $B = P^{-1}AP$. Then $A = PBP^{-1} = Q^{-1}BQ$ where $Q = P^{-1}$ is invertible. This proves the second of the following properties of similarity (the others are left as an exercise):

$$\begin{aligned} &1.\ A \sim A \text{ for all square matrices } A.\\ &2.\ \text{If } A \sim B, \text{ then } B \sim A.\\ &3.\ \text{If } A \sim B \text{ and } B \sim C, \text{ then } A \sim C.\end{aligned} \tag{5.2}$$

These properties are often expressed by saying that the similarity relation $\sim$ is an **equivalence relation** on the set of $n \times n$ matrices. Here is an example showing how these properties are used.

Example 5.5.1

If A is similar to B and either A or B is diagonalizable, show that the other is also diagonalizable.

Solution. We have $A \sim B$. Suppose that A is diagonalizable, say $A \sim D$ where D is diagonal. Since $B \sim A$ by (2) of (5.2), we have $B \sim A$ and $A \sim D$. Hence $B \sim D$ by (3) of (5.2), so B is diagonalizable too. An analogous argument works if we assume instead that B is diagonalizable.

Similarity is compatible with inverses, transposes, and powers:

$$\text{If } A \sim B \text{ then} \quad A^{-1} \sim B^{-1}, \quad A^T \sim B^T, \quad \text{and} \quad A^k \sim B^k \text{ for all integers } k \geq 1.$$

The proofs are routine matrix computations using Theorem 3.3.1. Thus, for example, if A is diagonalizable, so also are A^T, A^{-1} (if it exists), and A^k (for each $k \geq 1$). Indeed, if $A \sim D$ where D is a diagonal matrix, we obtain $A^T \sim D^T$, $A^{-1} \sim D^{-1}$, and $A^k \sim D^k$, and each of the matrices D^T, D^{-1}, and D^k is diagonal.

We pause to introduce a simple matrix function that will be referred to later.

Definition 5.13 Trace of a Matrix

*The **trace** $\operatorname{tr} A$ of an $n \times n$ matrix A is defined to be the sum of the main diagonal elements of A.*

In other words:

$$\text{If } A = \left[a_{ij}\right], \text{ then } \operatorname{tr} A = a_{11} + a_{22} + \cdots + a_{nn}.$$

It is evident that $\operatorname{tr}(A+B) = \operatorname{tr} A + \operatorname{tr} B$ and that $\operatorname{tr}(cA) = c \operatorname{tr} A$ holds for all $n \times n$ matrices A and B and all scalars c. The following fact is more surprising.

Lemma 5.5.1

Let A and B be $n \times n$ matrices. Then $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.

Proof. Write $A = \left[a_{ij}\right]$ and $B = \left[b_{ij}\right]$. For each i, the $(i,\ i)$-entry d_i of the matrix AB is given as follows: $d_i = a_{i1}b_{1i} + a_{i2}b_{2i} + \cdots + a_{in}b_{ni} = \sum_j a_{ij}b_{ji}$. Hence

$$\operatorname{tr}(AB) = d_1 + d_2 + \cdots + d_n = \sum_i d_i = \sum_i \left(\sum_j a_{ij}b_{ji} \right)$$

Similarly we have $\operatorname{tr}(BA) = \sum_i (\sum_j b_{ij}a_{ji})$. Since these two double sums are the same, Lemma 5.5.1 is proved. □

As the name indicates, similar matrices share many properties, some of which are collected in the next theorem for reference.

Theorem 5.5.1

If A and B are similar $n \times n$ matrices, then A and B have the same determinant, rank, trace, characteristic polynomial, and eigenvalues.

Proof. Let $B = P^{-1}AP$ for some invertible matrix P. Then we have

$$\det B = \det(P^{-1}) \det A \det P = \det A \text{ because } \det(P^{-1}) = 1/\det P$$

Similarly, $\operatorname{rank} B = \operatorname{rank}(P^{-1}AP) = \operatorname{rank} A$ by Corollary 5.4.3. Next Lemma 5.5.1 gives

$$\operatorname{tr}(P^{-1}AP) = \operatorname{tr}\left[P^{-1}(AP)\right] = \operatorname{tr}\left[(AP)P^{-1}\right] = \operatorname{tr} A$$

As to the characteristic polynomial,

$$\begin{aligned} c_B(x) = \det(xI - B) &= \det\{x(P^{-1}IP) - P^{-1}AP\} \\ &= \det\{P^{-1}(xI - A)P\} \\ &= \det(xI - A) \\ &= c_A(x) \end{aligned}$$

Finally, this shows that A and B have the same eigenvalues because the eigenvalues of a matrix are the roots of its characteristic polynomial. □

Example 5.5.2

Sharing the five properties in Theorem 5.5.1 does not guarantee that two matrices are similar. The matrices $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ have the same determinant, rank, trace, characteristic polynomial, and eigenvalues, but they are not similar because $P^{-1}IP = I$ for any invertible matrix P.

Diagonalization Revisited

Recall that a square matrix A is **diagonalizable** if there exists an invertible matrix P such that $P^{-1}AP = D$ is a diagonal matrix, that is if A is similar to a diagonal matrix D. Unfortunately, not all matrices are diagonalizable, for example $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ (see Example 3.4.3). Determining whether A is diagonalizable is closely related to the eigenvalues and eigenvectors of A. Recall that a number λ is called an **eigenvalue** of A if $A\mathbf{x} = \lambda\mathbf{x}$ for some nonzero column $\mathbf{x}$ in $\mathbb{R}^n$, and any such nonzero vector $\mathbf{x}$ is called an **eigenvector** of A corresponding to λ (or simply a λ-eigenvector of A). The eigenvalues and eigenvectors of A are closely related to the **characteristic polynomial** $c_A(x)$ of A, defined by

$$c_A(x) = \det(xI - A)$$

If A is $n \times n$ this is a polynomial of degree n, and its relationship to the eigenvalues is given in the following theorem (a repeat of Theorem 3.3.2).

Theorem 5.5.2

Let A be an $n \times n$ matrix.

1. *The eigenvalues λ of A are the roots of the characteristic polynomial $c_A(x)$ of A.*
2. *The λ-eigenvectors $\mathbf{x}$ are the nonzero solutions to the homogeneous system*

$$(\lambda I - A)\mathbf{x} = \mathbf{0}$$

of linear equations with $\lambda I - A$ as coefficient matrix.

Example 5.5.3

Show that the eigenvalues of a triangular matrix are the main diagonal entries.

Solution. Assume that A is triangular. Then the matrix $xI - A$ is also triangular and has diagonal entries $(x - a_{11}), (x - a_{22}), \dots, (x - a_{nn})$ where $A = [a_{ij}]$. Hence Theorem 3.1.4 gives

$$c_A(x) = (x - a_{11})(x - a_{22}) \cdots (x - a_{nn})$$

and the result follows because the eigenvalues are the roots of $c_A(x)$.

Theorem 3.4.1 asserts (in part) that an $n \times n$ matrix A is diagonalizable if and only if it has n eigenvectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ such that the matrix $P = \left[\begin{array}{ccc} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{array} \right]$ with the $\mathbf{x}_i$ as columns is invertible. This is equivalent to requiring that $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is a basis of $\mathbb{R}^n$ consisting of eigenvectors of A. Hence we can restate Theorem 3.4.1 as follows:

Theorem 5.5.3

Let A be an $n \times n$ matrix.

1. *A is diagonalizable if and only if $\mathbb{R}^n$ has a basis $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ consisting of eigenvectors of A.*
2. *When this is the case, the matrix $P = \left[\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{array} \right]$ is invertible and $P^{-1}AP = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ where, for each i, λ_i is the eigenvalue of A corresponding to $\mathbf{x}_i$.*

The next result is a basic tool for determining when a matrix is diagonalizable. It reveals an important connection between eigenvalues and linear independence: Eigenvectors corresponding to distinct eigenvalues are necessarily linearly independent.

Theorem 5.5.4

Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ be eigenvectors corresponding to distinct eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_k$ of an $n \times n$ matrix A. Then $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ is a linearly independent set.

Proof. We use induction on k. If $k = 1$, then $\{\mathbf{x}_1\}$ is independent because $\mathbf{x}_1 \neq \mathbf{0}$. In general, suppose the theorem is true for some $k \geq 1$. Given eigenvectors $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{k+1}\}$, suppose a linear combination vanishes:

$$t_1\mathbf{x}_1 + t_2\mathbf{x}_2 + \cdots + t_{k+1}\mathbf{x}_{k+1} = \mathbf{0} \tag{5.3}$$

We must show that each $t_i = 0$. Left multiply (5.3) by A and use the fact that $A\mathbf{x}_i = \lambda_i\mathbf{x}_i$ to get

$$t_1\lambda_1\mathbf{x}_1 + t_2\lambda_2\mathbf{x}_2 + \cdots + t_{k+1}\lambda_{k+1}\mathbf{x}_{k+1} = \mathbf{0} \tag{5.4}$$

If we multiply (5.3) by λ_1 and subtract the result from (5.4), the first terms cancel and we obtain

$$t_2(\lambda_2 - \lambda_1)\mathbf{x}_2 + t_3(\lambda_3 - \lambda_1)\mathbf{x}_3 + \cdots + t_{k+1}(\lambda_{k+1} - \lambda_1)\mathbf{x}_{k+1} = \mathbf{0}$$

Since $\mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_{k+1}$ correspond to distinct eigenvalues $\lambda_2, \lambda_3, \dots, \lambda_{k+1}$, the set $\{\mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_{k+1}\}$ is independent by the induction hypothesis. Hence,

$$t_2(\lambda_2 - \lambda_1) = 0, \quad t_3(\lambda_3 - \lambda_1) = 0, \quad \dots, \quad t_{k+1}(\lambda_{k+1} - \lambda_1) = 0$$

and so $t_2 = t_3 = \cdots = t_{k+1} = 0$ because the λ_i are distinct. Hence (5.3) becomes $t_1\mathbf{x}_1 = \mathbf{0}$, which implies that $t_1 = 0$ because $\mathbf{x}_1 \neq \mathbf{0}$. This is what we wanted. □

Theorem 5.5.4 will be applied several times; we begin by using it to give a useful condition for when a matrix is diagonalizable.

Theorem 5.5.5

If A is an $n \times n$ matrix with n distinct eigenvalues, then A is diagonalizable.

Proof. Choose one eigenvector for each of the n distinct eigenvalues. Then these eigenvectors are independent by Theorem 5.5.4, and so are a basis of $\mathbb{R}^n$ by Theorem 5.2.7. Now use Theorem 5.5.3. □

Example 5.5.4

Show that $A = \left[\begin{array}{rrr} 1 & 0 & 0 \\ 1 & 2 & 3 \\ -1 & 1 & 0 \end{array} \right]$ is diagonalizable.

Solution. A routine computation shows that $c_A(x) = (x-1)(x-3)(x+1)$ and so has distinct eigenvalues 1, 3, and -1. Hence Theorem 5.5.5 applies.

However, a matrix can have multiple eigenvalues as we saw in Section 3.3. To deal with this situation, we prove an important lemma which formalizes a technique that is basic to diagonalization, and which will be used three times below.

Lemma 5.5.2

Let $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ be a linearly independent set of eigenvectors of an $n \times n$ matrix A, extend it to a basis $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k, \dots, \mathbf{x}_n\}$ of $\mathbb{R}^n$, and let

$$P = \left[\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{array} \right]$$

be the (invertible) $n \times n$ matrix with the $\mathbf{x}_i$ as its columns. If $\lambda_1, \lambda_2, \dots, \lambda_k$ are the (not necessarily distinct) eigenvalues of A corresponding to $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ respectively, then $P^{-1}AP$ has block form

$$P^{-1}AP = \left[\begin{array}{cc} \text{diag}\,(\lambda_1, \lambda_2, \dots, \lambda_k) & B \\ 0 & A_1 \end{array} \right]$$

where B has size $k \times (n-k)$ and A_1 has size $(n-k) \times (n-k)$.

Proof. If $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ is the standard basis of $\mathbb{R}^n$, then

$$\begin{aligned} \left[\begin{array}{cccc} \mathbf{e}_1 & \mathbf{e}_2 & \dots & \mathbf{e}_n \end{array} \right] = I_n = P^{-1}P = P^{-1} \left[\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{array} \right] \\ = \left[\begin{array}{cccc} P^{-1}\mathbf{x}_1 & P^{-1}\mathbf{x}_2 & \cdots & P^{-1}\mathbf{x}_n \end{array} \right] \end{aligned}$$

Comparing columns, we have $P^{-1}\mathbf{x}_i = \mathbf{e}_i$ for each $1 \leq i \leq n$. On the other hand, observe that

$$P^{-1}AP = P^{-1}A \left[\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{array} \right] = \left[\begin{array}{cccc} (P^{-1}A)\mathbf{x}_1 & (P^{-1}A)\mathbf{x}_2 & \cdots & (P^{-1}A)\mathbf{x}_n \end{array} \right]$$

Hence, if $1 \leq i \leq k$, column i of $P^{-1}AP$ is

$$(P^{-1}A)\mathbf{x}_i = P^{-1}(\lambda_i \mathbf{x}_i) = \lambda_i(P^{-1}\mathbf{x}_i) = \lambda_i \mathbf{e}_i$$

This describes the first k columns of $P^{-1}AP$, and Lemma 5.5.2 follows. □

Note that Lemma 5.5.2 (with $k = n$) shows that an $n \times n$ matrix A is diagonalizable if $\mathbb{R}^n$ has a basis of eigenvectors of A, as in (1) of Theorem 5.5.3.

Definition 5.14 Eigenspace of a Matrix

*If λ is an eigenvalue of an $n \times n$ matrix A, define the **eigenspace** of A corresponding to λ by*

$$E_\lambda(A) = \{\mathbf{x} \text{ in } \mathbb{R}^n \mid A\mathbf{x} = \lambda\mathbf{x}\}$$

This is a subspace of $\mathbb{R}^n$ and the eigenvectors corresponding to λ are just the nonzero vectors in $E_\lambda(A)$. In fact $E_\lambda(A)$ is the null space of the matrix $(\lambda I - A)$:

$$E_\lambda(A) = \{\mathbf{x} \mid (\lambda I - A)\mathbf{x} = \mathbf{0}\} = \text{null}\,(\lambda I - A)$$

Hence, by Theorem 5.4.2, the basic solutions of the homogeneous system $(\lambda I - A)\mathbf{x} = \mathbf{0}$ given by the gaussian algorithm form a basis for $E_\lambda(A)$. In particular

$$\dim E_\lambda(A) \text{ is the number of basic solutions } \mathbf{x} \text{ of } (\lambda I - A)\mathbf{x} = \mathbf{0} \tag{5.5}$$

Now recall (Definition 3.7) that the **multiplicity**[11] of an eigenvalue λ of A is the number of times λ occurs as a root of the characteristic polynomial $c_A(x)$ of A. In other words, the multiplicity of λ is the largest integer $m \geq 1$ such that

$$c_A(x) = (x - \lambda)^m g(x)$$

for some polynomial $g(x)$. Because of (5.5), the assertion (without proof) in Theorem 3.4.2 can be stated as follows: A square matrix is diagonalizable if and only if the multiplicity of each eigenvalue λ equals $\dim [E_\lambda(A)]$. We are going to prove this, and the proof requires the following result which is valid for *any* square matrix, diagonalizable or not.

Lemma 5.5.3

Let λ be an eigenvalue of multiplicity m of a square matrix A. Then $\dim [E_\lambda(A)] \leq m$.

Proof. Write $\dim [E_\lambda(A)] = d$. It suffices to show that $c_A(x) = (x - \lambda)^d g(x)$ for some polynomial $g(x)$, because m is the highest power of $(x - \lambda)$ that divides $c_A(x)$. To this end, let $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_d\}$ be a basis of $E_\lambda(A)$. Then Lemma 5.5.2 shows that an invertible $n \times n$ matrix P exists such that

$$P^{-1}AP = \left[\begin{array}{cc} \lambda I_d & B \\ 0 & A_1 \end{array} \right]$$

in block form, where I_d denotes the $d \times d$ identity matrix. Now write $A' = P^{-1}AP$ and observe that $c_{A'}(x) = c_A(x)$ by Theorem 5.5.1. But Theorem 3.1.5 gives

$$c_A(x) = c_{A'}(x) = \det(xI_n - A') = \det \left[\begin{array}{cc} (x - \lambda)I_d & -B \\ 0 & xI_{n-d} - A_1 \end{array} \right]$$

[11] This is often called the *algebraic* multiplicity of λ.

$$= \det\left[(x-\lambda)I_d\right] \det\left[(xI_{n-d} - A_1)\right]$$
$$= (x-\lambda)^d g(x)$$

where $g(x) = cA_1(x)$. This is what we wanted. □

It is impossible to ignore the question when equality holds in Lemma 5.5.3 for each eigenvalue λ. It turns out that this characterizes the diagonalizable $n \times n$ matrices A for which $c_A(x)$ **factors completely** over $\mathbb{R}$. By this we mean that $c_A(x) = (x-\lambda_1)(x-\lambda_2)\cdots(x-\lambda_n)$, where the λ_i are *real* numbers (not necessarily distinct); in other words, every eigenvalue of A is real. This need not happen (consider $A = \left[\begin{array}{rr} 0 & -1 \\ 1 & 0 \end{array}\right]$), and we investigate the general case below.

Theorem 5.5.6

The following are equivalent for a square matrix A for which $c_A(x)$ factors completely.

1. *A is diagonalizable.*
2. *$\dim\left[E_\lambda(A)\right]$ equals the multiplicity of λ for every eigenvalue λ of the matrix A.*

Proof. Let A be $n \times n$ and let $\lambda_1, \lambda_2, \dots, \lambda_k$ be the distinct eigenvalues of A. For each i, let m_i denote the multiplicity of λ_i and write $d_i = \dim\left[E_{\lambda_i}(A)\right]$. Then

$$c_A(x) = (x-\lambda_1)^{m_1}(x-\lambda_2)^{m_2}\dots(x-\lambda_k)^{m_k}$$

so $m_1 + \cdots + m_k = n$ because $c_A(x)$ has degree n. Moreover, $d_i \leq m_i$ for each i by Lemma 5.5.3.

(1) $\Rightarrow$ (2). By (1), $\mathbb{R}^n$ has a basis of n eigenvectors of A, so let t_i of them lie in $E_{\lambda_i}(A)$ for each i. Since the subspace spanned by these t_i eigenvectors has dimension t_i, we have $t_i \leq d_i$ for each i by Theorem 5.2.4. Hence

$$n = t_1 + \cdots + t_k \leq d_1 + \cdots + d_k \leq m_1 + \cdots + m_k = n$$

It follows that $d_1 + \cdots + d_k = m_1 + \cdots + m_k$ so, since $d_i \leq m_i$ for each i, we must have $d_i = m_i$. This is (2).

(2) $\Rightarrow$ (1). Let B_i denote a basis of $E_{\lambda_i}(A)$ for each i, and let $B = B_1 \cup \cdots \cup B_k$. Since each B_i contains m_i vectors by (2), and since the B_i are pairwise disjoint (the λ_i are distinct), it follows that B contains n vectors. So it suffices to show that B is linearly independent (then B is a basis of $\mathbb{R}^n$). Suppose a linear combination of the vectors in B vanishes, and let $\mathbf{y}_i$ denote the sum of all terms that come from B_i. Then $\mathbf{y}_i$ lies in $E_{\lambda_i}(A)$, so the nonzero $\mathbf{y}_i$ are independent by Theorem 5.5.4 (as the λ_i are distinct). Since the sum of the $\mathbf{y}_i$ is zero, it follows that $\mathbf{y}_i = \mathbf{0}$ for each i. Hence all coefficients of terms in $\mathbf{y}_i$ are zero (because B_i is independent). Since this holds for each i, it shows that B is independent. □

Example 5.5.5

If $A = \left[\begin{array}{rrr} 5 & 8 & 16 \\ 4 & 1 & 8 \\ -4 & -4 & -11 \end{array}\right]$ and $B = \left[\begin{array}{rrr} 2 & 1 & 1 \\ 2 & 1 & -2 \\ -1 & 0 & -2 \end{array}\right]$ show that A is diagonalizable but B is not.

Solution. We have $c_A(x) = (x+3)^2(x-1)$ so the eigenvalues are $\lambda_1 = -3$ and $\lambda_2 = 1$. The corresponding eigenspaces are $E_{\lambda_1}(A) = \text{span}\{\mathbf{x}_1, \mathbf{x}_2\}$ and $E_{\lambda_2}(A) = \text{span}\{\mathbf{x}_3\}$ where

$$\mathbf{x}_1 = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}$$

as the reader can verify. Since $\{\mathbf{x}_1, \mathbf{x}_2\}$ is independent, we have $\dim(E_{\lambda_1}(A)) = 2$ which is the multiplicity of λ_1. Similarly, $\dim(E_{\lambda_2}(A)) = 1$ equals the multiplicity of λ_2. Hence A is diagonalizable by Theorem 5.5.6, and a diagonalizing matrix is $P = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{bmatrix}$.
Turning to B, $c_B(x) = (x+1)^2(x-3)$ so the eigenvalues are $\lambda_1 = -1$ and $\lambda_2 = 3$. The corresponding eigenspaces are $E_{\lambda_1}(B) = \text{span}\{\mathbf{y}_1\}$ and $E_{\lambda_2}(B) = \text{span}\{\mathbf{y}_2\}$ where

$$\mathbf{y}_1 = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}, \mathbf{y}_2 = \begin{bmatrix} 5 \\ 6 \\ -1 \end{bmatrix}$$

Here $\dim(E_{\lambda_1}(B)) = 1$ is *smaller* than the multiplicity of λ_1, so the matrix B is *not* diagonalizable, again by Theorem 5.5.6. The fact that $\dim(E_{\lambda_1}(B)) = 1$ means that there is no possibility of finding *three* linearly independent eigenvectors.

Complex Eigenvalues

All the matrices we have considered have had real eigenvalues. But this need not be the case: The matrix $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ has characteristic polynomial $c_A(x) = x^2 + 1$ which has no real roots. Nonetheless, this matrix is diagonalizable; the only difference is that we must use a larger set of scalars, the complex numbers. The basic properties of these numbers are outlined in Appendix A.

Indeed, nearly everything we have done for real matrices can be done for complex matrices. The methods are the same; the only difference is that the arithmetic is carried out with complex numbers rather than real ones. For example, the gaussian algorithm works in exactly the same way to solve systems of linear equations with complex coefficients, matrix multiplication is defined the same way, and the matrix inversion algorithm works in the same way.

But the complex numbers are better than the real numbers in one respect: While there are polynomials like $x^2 + 1$ with real coefficients that have no real root, this problem does not arise with the complex numbers: *Every* nonconstant polynomial with complex coefficients has a complex root, and hence factors completely as a product of linear factors. This fact is known as the fundamental theorem of algebra.[12]

Example 5.5.6

Diagonalize the matrix $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$.

Solution. The characteristic polynomial of A is

$$c_A(x) = \det(xI - A) = x^2 + 1 = (x - i)(x + i)$$

[12]This was a famous open problem in 1799 when Gauss solved it at the age of 22 in his Ph.D. dissertation.

where $i^2 = -1$. Hence the eigenvalues are $\lambda_1 = i$ and $\lambda_2 = -i$, with corresponding eigenvectors $\mathbf{x}_1 = \begin{bmatrix} 1 \\ -i \end{bmatrix}$ and $\mathbf{x}_2 = \begin{bmatrix} 1 \\ i \end{bmatrix}$. Hence A is diagonalizable by the complex version of Theorem 5.5.5, and the complex version of Theorem 5.5.3 shows that $P = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix}$ is invertible and $P^{-1}AP = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}$. Of course, this can be checked directly.

We shall return to complex linear algebra in Section 8.7.

Symmetric Matrices[13]

On the other hand, many of the applications of linear algebra involve a real matrix A and, while A will have complex eigenvalues by the fundamental theorem of algebra, it is always of interest to know when the eigenvalues are, in fact, real. While this can happen in a variety of ways, it turns out to hold whenever A is symmetric. This important theorem will be used extensively later. Surprisingly, the theory of *complex* eigenvalues can be used to prove this useful result about *real* eigenvalues.

Let $\overline{z}$ denote the conjugate of a complex number z. If A is a complex matrix, the **conjugate matrix** $\overline{A}$ is defined to be the matrix obtained from A by conjugating every entry. Thus, if $A = \left[z_{ij}\right]$, then $\overline{A} = \left[\overline{z}_{ij}\right]$. For example,

$$\text{If } A = \begin{bmatrix} -i+2 & 5 \\ i & 3+4i \end{bmatrix} \text{ then } \overline{A} = \begin{bmatrix} i+2 & 5 \\ -i & 3-4i \end{bmatrix}$$

Recall that $\overline{z+w} = \overline{z} + \overline{w}$ and $\overline{zw} = \overline{z}\,\overline{w}$ hold for all complex numbers z and w. It follows that if A and B are two complex matrices, then

$$\overline{A+B} = \overline{A} + \overline{B}, \quad \overline{AB} = \overline{A}\,\overline{B} \quad \text{and } \overline{\lambda A} = \overline{\lambda}\,\overline{A}$$

hold for all complex scalars λ. These facts are used in the proof of the following theorem.

Theorem 5.5.7

Let A be a symmetric real matrix. If λ is any complex eigenvalue of A, then λ is real.[14]

Proof. Observe that $\overline{A} = A$ because A is real. If λ is an eigenvalue of A, we show that λ is real by showing that $\overline{\lambda} = \lambda$. Let $\mathbf{x}$ be a (possibly complex) eigenvector corresponding to λ, so that $\mathbf{x} \neq \mathbf{0}$ and $A\mathbf{x} = \lambda\mathbf{x}$. Define $c = \mathbf{x}^T\overline{\mathbf{x}}$.

If we write $\mathbf{x} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$ where the z_i are complex numbers, we have

$$c = \mathbf{x}^T\overline{\mathbf{x}} = z_1\overline{z_1} + z_2\overline{z_2} + \cdots + z_n\overline{z_n} = |\overline{z_1}|^2 + |\overline{z_2}|^2 + \cdots + |\overline{z_n}|^2$$

[13] This discussion uses complex conjugation and absolute value. These topics are discussed in Appendix A.

[14] This theorem was first proved in 1829 by the great French mathematician Augustin Louis Cauchy (1789–1857).

Thus c is a real number, and $c > 0$ because at least one of the $z_i \neq 0$ (as $\mathbf{x} \neq \mathbf{0}$). We show that $\overline{\lambda} = \lambda$ by verifying that $\lambda c = \overline{\lambda} c$. We have

$$\lambda c = \lambda(\mathbf{x}^T \overline{\mathbf{x}}) = (\lambda \mathbf{x})^T \overline{\mathbf{x}} = (A\mathbf{x})^T \overline{\mathbf{x}} = \mathbf{x}^T A^T \overline{\mathbf{x}}$$

At this point we use the hypothesis that A is symmetric and real. This means $A^T = A = \overline{A}$ so we continue the calculation:

$$\begin{aligned}\lambda c = \mathbf{x}^T A^T \overline{\mathbf{x}} = \mathbf{x}^T (\overline{A}\, \overline{\mathbf{x}}) = \mathbf{x}^T (\overline{A\mathbf{x}}) &= \mathbf{x}^T (\overline{\lambda \mathbf{x}}) \\ &= \mathbf{x}^T (\overline{\lambda}\, \overline{\mathbf{x}}) \\ &= \overline{\lambda} \mathbf{x}^T \overline{\mathbf{x}} \\ &= \overline{\lambda} c\end{aligned}$$

as required. □

The technique in the proof of Theorem 5.5.7 will be used again when we return to complex linear algebra in Section 8.7.

Example 5.5.7

Verify Theorem 5.5.7 for every real, symmetric 2×2 matrix A.

Solution. If $A = \left[\begin{array}{cc} a & b \\ b & c \end{array} \right]$ we have $c_A(x) = x^2 - (a+c)x + (ac - b^2)$, so the eigenvalues are given by $\lambda = \frac{1}{2}[(a+c) \pm \sqrt{(a+c)^2 - 4(ac-b^2)}]$. But here

$$(a+c)^2 - 4(ac - b^2) = (a-c)^2 + 4b^2 \geq 0$$

for any choice of a, b, and c. Hence, the eigenvalues are real numbers.

5.6 Best Approximation and Least Squares

Often an exact solution to a problem in applied mathematics is difficult to obtain. However, it is usually just as useful to find arbitrarily close approximations to a solution. In particular, finding "linear approximations" is a potent technique in applied mathematics. One basic case is the situation where a system of linear equations has no solution, and it is desirable to find a "best approximation" to a solution to the system. In this section best approximations are defined and a method for finding them is described. The result is then applied to "least squares" approximation of data.

Suppose A is an $m \times n$ matrix and $\mathbf{b}$ is a column in $\mathbb{R}^m$, and consider the system

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

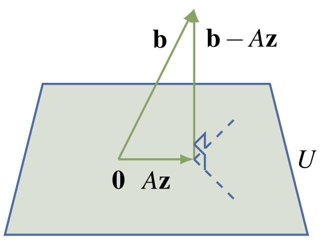
of m linear equations in n variables. This need not have a solution. However, given any column $\mathbf{z} \in \mathbb{R}^n$, the distance $\|\mathbf{b} - A\mathbf{z}\|$ is a measure of how far $A\mathbf{z}$ is from $\mathbf{b}$. Hence it is natural to ask whether there is a column $\mathbf{z}$ in $\mathbb{R}^n$ that is as close as possible to a solution in the sense that

$$\|\mathbf{b} - A\mathbf{z}\|$$

is the minimum value of $\|\mathbf{b} - A\mathbf{x}\|$ as $\mathbf{x}$ ranges over all columns in $\mathbb{R}^n$.

The answer is "yes", and to describe it define

$$U = \{A\mathbf{x} \mid \mathbf{x} \text{ lies in } \mathbb{R}^n\}$$



This is a subspace of $\mathbb{R}^m$ (verify) and we want a vector $A\mathbf{z}$ in U as close as possible to $\mathbf{b}$. That there is such a vector is clear geometrically if $m = 3$ by the diagram. In general such a vector $A\mathbf{z}$ exists by a general result called the *projection theorem* that will be proved in Chapter 8 (Theorem 8.1.3). Moreover, the projection theorem gives a simple way to compute $\mathbf{z}$ because it also shows that the vector $\mathbf{b} - A\mathbf{z}$ is *orthogonal* to every vector $A\mathbf{x}$ in U. Thus, for all $\mathbf{x}$ in $\mathbb{R}^n$,

$$\begin{aligned} 0 = (A\mathbf{x}) \cdot (\mathbf{b} - A\mathbf{z}) = (A\mathbf{x})^T(\mathbf{b} - A\mathbf{z}) &= \mathbf{x}^T A^T (\mathbf{b} - A\mathbf{z}) \\ &= \mathbf{x} \cdot [A^T(\mathbf{b} - A\mathbf{z})] \end{aligned}$$

In other words, the vector $A^T(\mathbf{b} - A\mathbf{z})$ in $\mathbb{R}^n$ is orthogonal to *every* vector in $\mathbb{R}^n$ and so must be zero (being orthogonal to itself). Hence $\mathbf{z}$ satisfies

$$(A^T A)\mathbf{z} = A^T \mathbf{b}$$

Definition 5.15 Normal Equations

This is a system of linear equations called the ***normal equations*** *for* $\mathbf{z}$.

Note that this system can have more than one solution (see Exercise **??**). However, the $n \times n$ matrix $A^T A$ is invertible if (and only if) the columns of A are linearly independent (Theorem 5.4.3); so, in this case, $\mathbf{z}$ is uniquely determined and is given explicitly by $\mathbf{z} = (A^T A)^{-1} A^T \mathbf{b}$. However, the most efficient way to find $\mathbf{z}$ is to apply gaussian elimination to the normal equations.

This discussion is summarized in the following theorem.

Theorem 5.6.1: Best Approximation Theorem

Let A be an $m \times n$ matrix, let $\mathbf{b}$ be any column in $\mathbb{R}^m$, and consider the system

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

of m equations in n variables.

1. *Any solution $\mathbf{z}$ to the normal equations*

   $$(A^T A)\mathbf{z} = A^T \mathbf{b}$$

   is a best approximation to a solution to $A\mathbf{x} = \mathbf{b}$ in the sense that $\|\mathbf{b} - A\mathbf{z}\|$ is the minimum value of $\|\mathbf{b} - A\mathbf{x}\|$ as $\mathbf{x}$ ranges over all columns in $\mathbb{R}^n$.

2. *If the columns of A are linearly independent, then $A^T A$ is invertible and $\mathbf{z}$ is given uniquely by $\mathbf{z} = (A^T A)^{-1} A^T \mathbf{b}$.*

We note in passing that if A is $n \times n$ and invertible, then

$$\mathbf{z} = (A^T A)^{-1} A^T \mathbf{b} = A^{-1}\mathbf{b}$$

is the solution to the system of equations, and $\|\mathbf{b} - A\mathbf{z}\| = 0$. Hence if A has independent columns, then $(A^TA)^{-1}A^T$ is playing the role of the inverse of the nonsquare matrix A. The matrix $A^T(AA^T)^{-1}$ plays a similar role when the rows of A are linearly independent. These are both special cases of the **generalized inverse** of a matrix A (see Exercise **??**). However, we shall not pursue this topic here.

Example 5.6.1

The system of linear equations

$$\begin{aligned} 3x - y &= 4 \\ x + 2y &= 0 \\ 2x + y &= 1 \end{aligned}$$

has no solution. Find the vector $\mathbf{z} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$ that best approximates a solution.

Solution. In this case,

$$A = \begin{bmatrix} 3 & -1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}, \text{ so } A^TA = \begin{bmatrix} 3 & 1 & 2 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 14 & 1 \\ 1 & 6 \end{bmatrix}$$

is invertible. The normal equations $(A^TA)\mathbf{z} = A^T\mathbf{b}$ are

$$\begin{bmatrix} 14 & 1 \\ 1 & 6 \end{bmatrix} \mathbf{z} = \begin{bmatrix} 14 \\ -3 \end{bmatrix}, \text{ so } \mathbf{z} = \frac{1}{83} \begin{bmatrix} 87 \\ -56 \end{bmatrix}$$

Thus $x_0 = \frac{87}{83}$ and $y_0 = \frac{-56}{83}$. With these values of x and y, the left sides of the equations are, approximately,

$$\begin{aligned} 3x_0 - y_0 &= \tfrac{317}{83} = 3.82 \\ x_0 + 2y_0 &= \tfrac{-25}{83} = -0.30 \\ 2x_0 + y_0 &= \tfrac{118}{83} = 1.42 \end{aligned}$$

This is as close as possible to a solution.

Example 5.6.2

The average number g of goals per game scored by a hockey player seems to be related linearly to two factors: the number x_1 of years of experience and the number x_2 of goals in the preceding 10 games. The data on the following page were collected on four players. Find the linear function $g = a_0 + a_1x_1 + a_2x_2$ that best fits these data.

g	x_1	x_2
0.8	5	3
0.8	3	4
0.6	1	5
0.4	2	1

Solution. If the relationship is given by $g = r_0 + r_1x_1 + r_2x_2$, then the data can be described as follows:

$$\begin{bmatrix} 1 & 5 & 3 \\ 1 & 3 & 4 \\ 1 & 1 & 5 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} r_0 \\ r_1 \\ r_2 \end{bmatrix} = \begin{bmatrix} 0.8 \\ 0.8 \\ 0.6 \\ 0.4 \end{bmatrix}$$

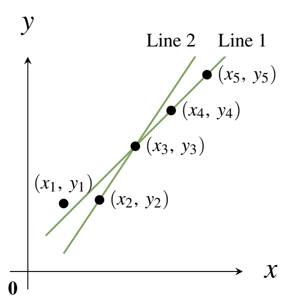
Using the notation in Theorem 5.6.1, we get

$$\begin{aligned} \mathbf{z} &= (A^TA)^{-1}A^T\mathbf{b} \\ &= \frac{1}{42}\begin{bmatrix} 119 & -17 & -19 \\ -17 & 5 & 1 \\ -19 & 1 & 5 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 5 & 3 & 1 & 2 \\ 3 & 4 & 5 & 1 \end{bmatrix} \begin{bmatrix} 0.8 \\ 0.8 \\ 0.6 \\ 0.4 \end{bmatrix} = \begin{bmatrix} 0.14 \\ 0.09 \\ 0.08 \end{bmatrix} \end{aligned}$$

Hence the best-fitting function is $g = 0.14 + 0.09x_1 + 0.08x_2$. The amount of computation would have been reduced if the normal equations had been constructed and then solved by gaussian elimination.

Least Squares Approximation

In many scientific investigations, data are collected that relate two variables. For example, if x is the number of dollars spent on advertising by a manufacturer and y is the value of sales in the region in question, the manufacturer could generate data by spending $x_1, x_2, \dots, x_n$ dollars at different times and measuring the corresponding sales values $y_1, y_2, \dots, y_n$.

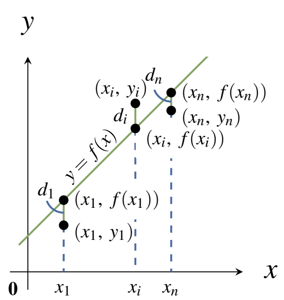


Suppose it is known that a linear relationship exists between the variables x and y—in other words, that $y = a + bx$ for some constants a and b. If the data are plotted, the points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ may appear to lie on a straight line and estimating a and b requires finding the “best-fitting” line through these data points. For example, if five data points occur as shown in the diagram, line 1 is clearly a better fit than line 2. In general, the problem is to find the values of the constants a and b such that the line $y = a + bx$ best approximates the data in question. Note that an exact fit would be obtained if a and b were such that $y_i = a + bx_i$ were true for each data point (x_i, y_i). But this is too much to expect. Experimental errors in measurement are bound to occur, so the choice of a and b should be made in such a way that the errors between the observed values y_i and the corresponding fitted values $a + bx_i$ are in some sense minimized. Least squares approximation is a way to do this.

The first thing we must do is explain exactly what we mean by the *best fit* of a line $y = a + bx$ to an observed set of data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. For convenience, write the linear function $r_0 + r_1x$ as

$$f(x) = r_0 + r_1x$$

so that the fitted points (on the line) have coordinates $(x_1, f(x_1)), \dots, (x_n, f(x_n))$.



The second diagram is a sketch of what the line $y = f(x)$ might look like. For each i the observed data point $(x_i,\ y_i)$ and the fitted point $(x_i,\ f(x_i))$ need not be the same, and the distance d_i between them measures how far the line misses the observed point. For this reason d_i is often called the **error** at x_i, and a natural measure of how close the line $y = f(x)$ is to the observed data points is the sum $d_1 + d_2 + \cdots + d_n$ of all these errors. However, it turns out to be better to use the sum of squares

$$S = d_1^2 + d_2^2 + \cdots + d_n^2$$

as the measure of error, and the line $y = f(x)$ is to be chosen so as to make this sum as small as possible. This line is said to be the **least squares approximating line** for the data points $(x_1,\ y_1),\ (x_2,\ y_2),\ \ldots,\ (x_n,\ y_n)$.

The square of the error d_i is given by $d_i^2 = [y_i - f(x_i)]^2$ for each i, so the quantity S to be minimized is the sum:

$$S = [y_1 - f(x_1)]^2 + [y_2 - f(x_2)]^2 + \cdots + [y_n - f(x_n)]^2$$

Note that all the numbers x_i and y_i are *given* here; what is required is that the *function* f be chosen in such a way as to minimize S. Because $f(x) = r_0 + r_1 x$, this amounts to choosing r_0 and r_1 to minimize S. This problem can be solved using Theorem 5.6.1. The following notation is convenient.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad f(\mathbf{x}) = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix} = \begin{bmatrix} r_0 + r_1 x_1 \\ r_0 + r_1 x_2 \\ \vdots \\ r_0 + r_1 x_n \end{bmatrix}$$

Then the problem takes the following form: Choose r_0 and r_1 such that

$$S = [y_1 - f(x_1)]^2 + [y_2 - f(x_2)]^2 + \cdots + [y_n - f(x_n)]^2 = \|\mathbf{y} - f(\mathbf{x})\|^2$$

is as small as possible. Now write

$$M = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{and} \quad \mathbf{r} = \begin{bmatrix} r_0 \\ r_1 \end{bmatrix}$$

Then $M\mathbf{r} = f(\mathbf{x})$, so we are looking for a column $\mathbf{r} = \begin{bmatrix} r_0 \\ r_1 \end{bmatrix}$ such that $\|\mathbf{y} - M\mathbf{r}\|^2$ is as small as possible. In other words, we are looking for a best approximation $\mathbf{z}$ to the system $M\mathbf{r} = \mathbf{y}$. Hence Theorem 5.6.1 applies directly, and we have

Theorem 5.6.2

Suppose that n data points $(x_1,\ y_1),\ (x_2,\ y_2),\ \ldots,\ (x_n,\ y_n)$ *are given, where at least two of*

$x_1, x_2, \dots, x_n$ are distinct. Put

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad M = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Then the least squares approximating line for these data points has equation

$$y = z_0 + z_1 x$$

where $\mathbf{z} = \begin{bmatrix} z_0 \\ z_1 \end{bmatrix}$ is found by gaussian elimination from the normal equations

$$(M^T M)\mathbf{z} = M^T \mathbf{y}$$

The condition that at least two of $x_1, x_2, \dots, x_n$ are distinct ensures that $M^T M$ is an invertible matrix, so $\mathbf{z}$ is unique:

$$\mathbf{z} = (M^T M)^{-1} M^T \mathbf{y}$$

Example 5.6.3

Let data points $(x_1, y_1), (x_2, y_2), \dots, (x_5, y_5)$ be given as in the accompanying table. Find the least squares approximating line for these data.

x	y
1	1
3	2
4	3
6	4
7	5

Solution. In this case we have

$$\begin{aligned} M^T M &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_5 \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_5 \end{bmatrix} \\ &= \begin{bmatrix} 5 & x_1 + \cdots + x_5 \\ x_1 + \cdots + x_5 & x_1^2 + \cdots + x_5^2 \end{bmatrix} = \begin{bmatrix} 5 & 21 \\ 21 & 111 \end{bmatrix} \end{aligned}$$

$$\text{and } M^T \mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_5 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_5 \end{bmatrix}$$

$$= \begin{bmatrix} y_1 + y_2 + \cdots + y_5 \\ x_1y_1 + x_2y_2 + \cdots + x_5y_5 \end{bmatrix} = \begin{bmatrix} 15 \\ 78 \end{bmatrix}$$

so the normal equations $(M^TM)\mathbf{z} = M^T\mathbf{y}$ for $\mathbf{z} = \begin{bmatrix} z_0 \\ z_1 \end{bmatrix}$ become

$$\begin{bmatrix} 5 & 21 \\ 21 & 111 \end{bmatrix} \begin{bmatrix} z_0 \\ z_1 \end{bmatrix} = \begin{bmatrix} 15 \\ 78 \end{bmatrix}$$

The solution (using gaussian elimination) is $\mathbf{z} = \begin{bmatrix} z_0 \\ z_1 \end{bmatrix} = \begin{bmatrix} 0.24 \\ 0.66 \end{bmatrix}$ to two decimal places, so the least squares approximating line for these data is $y = 0.24 + 0.66x$. Note that M^TM is indeed invertible here (the determinant is 114), and the exact solution is

$$\mathbf{z} = (M^TM)^{-1}M^T\mathbf{y} = \frac{1}{114}\begin{bmatrix} 111 & -21 \\ -21 & 5 \end{bmatrix} \begin{bmatrix} 15 \\ 78 \end{bmatrix} = \frac{1}{114}\begin{bmatrix} 27 \\ 75 \end{bmatrix} = \frac{1}{38}\begin{bmatrix} 9 \\ 25 \end{bmatrix}$$

Least Squares Approximating Polynomials

Suppose now that, rather than a straight line, we want to find a polynomial

$$y = f(x) = r_0 + r_1x + r_2x^2 + \cdots + r_mx^m$$

of degree m that best approximates the data pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. As before, write

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad f(\mathbf{x}) = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix}$$

For each x_i we have two values of the variable y, the observed value y_i, and the computed value $f(x_i)$. The problem is to choose $f(x)$—that is, choose $r_0, r_1, \dots, r_m$ —such that the $f(x_i)$ are as close as possible to the y_i. Again we define "as close as possible" by the least squares condition: We choose the r_i such that

$$\|\mathbf{y} - f(\mathbf{x})\|^2 = [y_1 - f(x_1)]^2 + [y_2 - f(x_2)]^2 + \cdots + [y_n - f(x_n)]^2$$

is as small as possible.

Definition 5.16 Least Squares Approximation

A polynomial $f(x)$ satisfying this condition is called a ***least squares approximating polynomial*** *of degree m for the given data pairs.*

If we write

$$M = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix} \quad \text{and} \quad \mathbf{r} = \begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ r_m \end{bmatrix}$$

we see that $f(\mathbf{x}) = M\mathbf{r}$. Hence we want to find $\mathbf{r}$ such that $\|\mathbf{y} - M\mathbf{r}\|^2$ is as small as possible; that is, we want a best approximation $\mathbf{z}$ to the system $M\mathbf{r} = \mathbf{y}$. Theorem 5.6.1 gives the first part of Theorem 5.6.3.

Theorem 5.6.3

Let n data pairs $(x_1,\ y_1),\ (x_2,\ y_2),\ \dots,\ (x_n,\ y_n)$ be given, and write

$$\mathbf{y} = \left[\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array}\right] \quad M = \left[\begin{array}{ccccc} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{array}\right] \quad \mathbf{z} = \left[\begin{array}{c} z_0 \\ z_1 \\ \vdots \\ z_m \end{array}\right]$$

1. *If $\mathbf{z}$ is any solution to the normal equations*

$$(M^T M)\mathbf{z} = M^T \mathbf{y}$$

   then the polynomial

$$z_0 + z_1 x + z_2 x^2 + \cdots + z_m x^m$$

   is a least squares approximating polynomial of degree m for the given data pairs.

2. *If at least $m+1$ of the numbers $x_1,\ x_2,\ \dots,\ x_n$ are distinct (so $n \geq m+1$), the matrix $M^T M$ is invertible and $\mathbf{z}$ is uniquely determined by*

$$\mathbf{z} = (M^T M)^{-1} M^T \mathbf{y}$$

Proof. It remains to prove (2), and for that we show that the columns of M are linearly independent (Theorem 5.4.3). Suppose a linear combination of the columns vanishes:

$$r_0 \left[\begin{array}{c} 1 \\ 1 \\ \vdots \\ 1 \end{array}\right] + r_1 \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array}\right] + \cdots + r_m \left[\begin{array}{c} x_1^m \\ x_2^m \\ \vdots \\ x_n^m \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ \vdots \\ 0 \end{array}\right]$$

If we write $q(x) = r_0 + r_1 x + \cdots + r_m x^m$, equating coefficients shows that

$$q(x_1) = q(x_2) = \cdots = q(x_n) = 0$$

Hence $q(x)$ is a polynomial of degree m with at least $m+1$ distinct roots, so $q(x)$ must be the zero polynomial (see Appendix D or Theorem 6.5.4). Thus $r_0 = r_1 = \cdots = r_m = 0$ as required. □

Example 5.6.4

Find the least squares approximating quadratic $y = z_0 + z_1 x + z_2 x^2$ for the following data points.

$$(-3,\ 3),\ (-1,\ 1),\ (0,\ 1),\ (1,\ 2),\ (3,\ 4)$$

Solution. This is an instance of Theorem 5.6.3 with $m = 2$. Here

$$\mathbf{y} = \begin{bmatrix} 3 \\ 1 \\ 1 \\ 2 \\ 4 \end{bmatrix} \quad M = \begin{bmatrix} 1 & -3 & 9 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \end{bmatrix}$$

Hence,

$$M^T M = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -3 & -1 & 0 & 1 & 3 \\ 9 & 1 & 0 & 1 & 9 \end{bmatrix} \begin{bmatrix} 1 & -3 & 9 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 20 \\ 0 & 20 & 0 \\ 20 & 0 & 164 \end{bmatrix}$$

$$M^T \mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -3 & -1 & 0 & 1 & 3 \\ 9 & 1 & 0 & 1 & 9 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 1 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 11 \\ 4 \\ 66 \end{bmatrix}$$

The normal equations for $\mathbf{z}$ are

$$\begin{bmatrix} 5 & 0 & 20 \\ 0 & 20 & 0 \\ 20 & 0 & 164 \end{bmatrix} \mathbf{z} = \begin{bmatrix} 11 \\ 4 \\ 66 \end{bmatrix} \quad \text{whence } \mathbf{z} = \begin{bmatrix} 1.15 \\ 0.20 \\ 0.26 \end{bmatrix}$$

This means that the least squares approximating quadratic for these data is $y = 1.15 + 0.20x + 0.26x^2$.

Other Functions

There is an extension of Theorem 5.6.3 that should be mentioned. Given data pairs (x_1, y_1), (x_2, y_2), $\dots$, (x_n, y_n), that theorem shows how to find a polynomial

$$f(x) = r_0 + r_1 x + \cdots + r_m x^m$$

such that $\|\mathbf{y} - f(\mathbf{x})\|^2$ is as small as possible, where $\mathbf{x}$ and $f(\mathbf{x})$ are as before. Choosing the appropriate polynomial $f(x)$ amounts to choosing the coefficients $r_0, r_1, \dots, r_m$, and Theorem 5.6.3 gives a formula for the optimal choices. Here $f(x)$ is a linear combination of the functions $1, x, x^2, \dots, x^m$ where the r_i are the coefficients, and this suggests applying the method to other functions. If $f_0(x), f_1(x), \dots, f_m(x)$ are given functions, write

$$f(x) = r_0 f_0(x) + r_1 f_1(x) + \cdots + r_m f_m(x)$$

where the r_i are real numbers. Then the more general question is whether $r_0, r_1, \dots, r_m$ can be found such that $\|\mathbf{y} - f(\mathbf{x})\|^2$ is as small as possible where

$$f(\mathbf{x}) = \left[\begin{array}{c} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \end{array}\right]$$

Such a function $f(\mathbf{x})$ is called a **least squares best approximation** for these data pairs of the form $r_0 f_0(x) + r_1 f_1(x) + \cdots + r_m f_m(x)$, r_i in $\mathbb{R}$. The proof of Theorem 5.6.3 goes through to prove

Theorem 5.6.4

Let n data pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be given, and suppose that $m+1$ functions $f_0(x), f_1(x), \dots, f_m(x)$ are specified. Write

$$\mathbf{y} = \left[\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array}\right] \quad M = \left[\begin{array}{cccc} f_0(x_1) & f_1(x_1) & \cdots & f_m(x_1) \\ f_0(x_2) & f_1(x_2) & \cdots & f_m(x_2) \\ \vdots & \vdots & & \vdots \\ f_0(x_n) & f_1(x_n) & \cdots & f_m(x_n) \end{array}\right] \quad \mathbf{z} = \left[\begin{array}{c} z_1 \\ z_2 \\ \vdots \\ z_m \end{array}\right]$$

1. *If $\mathbf{z}$ is any solution to the normal equations*

$$(M^T M)\mathbf{z} = M^T \mathbf{y}$$

   then the function

$$z_0 f_0(x) + z_1 f_1(x) + \cdots + z_m f_m(x)$$

   is the best approximation for these data among all functions of the form $r_0 f_0(x) + r_1 f_1(x) + \cdots + r_m f_m(x)$ where the r_i are in $\mathbb{R}$.

2. *If $M^T M$ is invertible (that is, if $\operatorname{rank}(M) = m+1$), then $\mathbf{z}$ is uniquely determined; in fact, $\mathbf{z} = (M^T M)^{-1}(M^T \mathbf{y})$.*

Clearly Theorem 5.6.4 contains Theorem 5.6.3 as a special case, but there is no simple test in general for whether $M^T M$ is invertible. Conditions for this to hold depend on the choice of the functions $f_0(x), f_1(x), \dots, f_m(x)$.

Example 5.6.5

Given the data pairs $(-1, 0)$, $(0, 1)$, and $(1, 4)$, find the least squares approximating function of the form $r_0 x + r_1 2^x$.

Solution. The functions are $f_0(x) = x$ and $f_1(x) = 2^x$, so the matrix M is

$$M = \left[\begin{array}{cc} f_0(x_1) & f_1(x_1) \\ f_0(x_2) & f_1(x_2) \\ f_0(x_3) & f_1(x_3) \end{array}\right] = \left[\begin{array}{cc} -1 & 2^{-1} \\ 0 & 2^0 \\ 1 & 2^1 \end{array}\right] = \frac{1}{2}\left[\begin{array}{cc} -2 & 1 \\ 0 & 2 \\ 2 & 4 \end{array}\right]$$

In this case $M^T M = \frac{1}{4}\begin{bmatrix} 8 & 6 \\ 6 & 21 \end{bmatrix}$ is invertible, so the normal equations

$$\frac{1}{4}\begin{bmatrix} 8 & 6 \\ 6 & 21 \end{bmatrix}\mathbf{z} = \begin{bmatrix} 4 \\ 9 \end{bmatrix}$$

have a unique solution $\mathbf{z} = \frac{1}{11}\begin{bmatrix} 10 \\ 16 \end{bmatrix}$. Hence the best-fitting function of the form $r_0 x + r_1 2^x$ is $\overline{f}(x) = \frac{10}{11}x + \frac{16}{11}2^x$. Note that $\overline{f}(\mathbf{x}) = \begin{bmatrix} \overline{f}(-1) \\ \overline{f}(0) \\ \overline{f}(1) \end{bmatrix} = \begin{bmatrix} \frac{-2}{11} \\ \frac{16}{11} \\ \frac{42}{11} \end{bmatrix}$, compared with $\mathbf{y} = \begin{bmatrix} 0 \\ 1 \\ 4 \end{bmatrix}$

5.7 An Application to Correlation and Variance

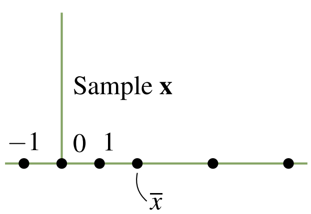
Suppose the heights $h_1, h_2, \dots, h_n$ of n men are measured. Such a data set is called a **sample** of the heights of all the men in the population under study, and various questions are often asked about such a sample: What is the average height in the sample? How much variation is there in the sample heights, and how can it be measured? What can be inferred from the sample about the heights of all men in the population? How do these heights compare to heights of men in neighbouring countries? Does the prevalence of smoking affect the height of a man?

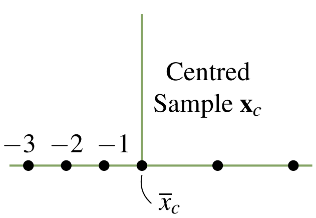
The analysis of samples, and of inferences that can be drawn from them, is a subject called *mathematical statistics*, and an extensive body of information has been developed to answer many such questions. In this section we will describe a few ways that linear algebra can be used.

It is convenient to represent a sample $\{x_1, x_2, \dots, x_n\}$ as a **sample vector**[15] $\mathbf{x} = \left[\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array} \right]$ in $\mathbb{R}^n$. This being done, the dot product in $\mathbb{R}^n$ provides a convenient tool to study the sample and describe some of the statistical concepts related to it. The most widely known statistic for describing a data set is the **sample mean** $\overline{x}$ defined by[16]

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The mean $\overline{x}$ is "typical" of the sample values x_i, but may not itself be one of them. The number $x_i - \overline{x}$ is called the **deviation** of x_i from the mean $\overline{x}$. The deviation is positive if $x_i > \overline{x}$ and it is negative if $x_i < \overline{x}$. Moreover, the sum of these deviations is zero:





$$\sum_{i=1}^{n}(x_i - \overline{x}) = \left(\sum_{i=1}^{n} x_i\right) - n\overline{x} = n\overline{x} - n\overline{x} = 0 \tag{5.6}$$

This is described by saying that the sample mean $\overline{x}$ is *central* to the sample values x_i.

If the mean $\overline{x}$ is subtracted from each data value x_i, the resulting data $x_i - \overline{x}$ are said to be **centred**. The corresponding data vector is

$$\mathbf{x}_c = \left[\begin{array}{cccc} x_1 - \overline{x} & x_2 - \overline{x} & \cdots & x_n - \overline{x} \end{array} \right]$$

and (5.6) shows that the mean $\overline{x}_c = 0$. For example, we have plotted the sample $\mathbf{x} = \left[\begin{array}{ccccc} -1 & 0 & 1 & 4 & 6 \end{array} \right]$ in the first diagram. The mean is $\overline{x} = 2$, and the centred sample $\mathbf{x}_c = \left[\begin{array}{ccccc} -3 & -2 & -1 & 2 & 4 \end{array} \right]$ is also plotted. Thus, the effect of centring is to shift the data by an amount $\overline{x}$ (to the left if $\overline{x}$ is positive) so that the mean moves to 0.

Another question that arises about samples is how much variability there is in the sample

$$\mathbf{x} = \left[\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array} \right]$$

that is, how widely are the data "spread out" around the sample mean $\overline{x}$. A natural measure of variability would be the sum of the deviations of the x_i about the mean, but this sum is zero by (5.6); these deviations

[15] We write vectors in $\mathbb{R}^n$ as row matrices, for convenience.

[16] The mean is often called the "average" of the sample values x_i, but statisticians use the term "mean".

cancel out. To avoid this cancellation, statisticians use the *squares* $(x_i - \overline{x})^2$ of the deviations as a measure of variability. More precisely, they compute a statistic called the **sample variance** s_x^2 defined[17] as follows:

$$s_x^2 = \frac{1}{n-1}[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

The sample variance will be large if there are many x_i at a large distance from the mean $\overline{x}$, and it will be small if all the x_i are tightly clustered about the mean. The variance is clearly nonnegative (hence the notation s_x^2), and the square root s_x of the variance is called the **sample standard deviation**.

The sample mean and variance can be conveniently described using the dot product. Let

$$\mathbf{1} = \left[\begin{array}{cccc} 1 & 1 & \cdots & 1 \end{array}\right]$$

denote the row with every entry equal to 1. If $\mathbf{x} = \left[\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array}\right]$, then $\mathbf{x} \cdot \mathbf{1} = x_1 + x_2 + \cdots + x_n$, so the sample mean is given by the formula

$$\overline{x} = \frac{1}{n}(\mathbf{x} \cdot \mathbf{1})$$

Moreover, remembering that $\overline{x}$ is a scalar, we have $\overline{x}\mathbf{1} = \left[\begin{array}{cccc} \overline{x} & \overline{x} & \cdots & \overline{x} \end{array}\right]$, so the centred sample vector $\mathbf{x}_c$ is given by

$$\mathbf{x}_c = \mathbf{x} - \overline{x}\mathbf{1} = \left[\begin{array}{cccc} x_1 - \overline{x} & x_2 - \overline{x} & \cdots & x_n - \overline{x} \end{array}\right]$$

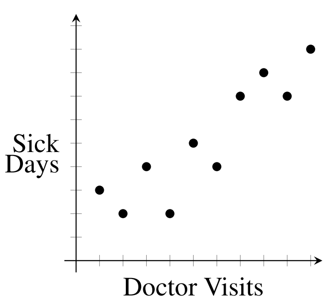
Thus we obtain a formula for the sample variance:

$$s_x^2 = \frac{1}{n-1}\|\mathbf{x}_c\|^2 = \frac{1}{n-1}\|\mathbf{x} - \overline{x}\mathbf{1}\|^2$$

Linear algebra is also useful for comparing two different samples. To illustrate how, consider two examples.

The following table represents the number of sick days at work per year and the yearly number of visits to a physician for 10 individuals.

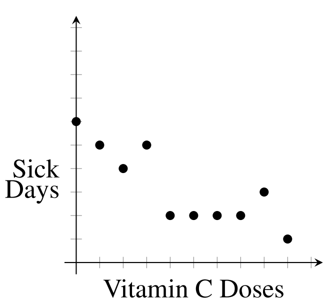
Individual	1	2	3	4	5	6	7	8	9	10
Doctor visits	2	6	8	1	5	10	3	9	7	4
Sick days	2	4	8	3	5	9	4	7	7	2



The data are plotted in the **scatter diagram** where it is evident that, roughly speaking, the more visits to the doctor the more sick days. This is an example of a *positive correlation* between sick days and doctor visits.

Now consider the following table representing the daily doses of vitamin C and the number of sick days.

Individual	1	2	3	4	5	6	7	8	9	10
Vitamin C	1	5	7	0	4	9	2	8	6	3
Sick days	5	2	2	6	2	1	4	3	2	5



The scatter diagram is plotted as shown and it appears that the more vitamin C taken, the fewer sick days. In this case there is a *negative correlation* between daily vitamin C and sick days.

[17] Since there are n sample values, it seems more natural to divide by n here, rather than by $n-1$. The reason for using $n-1$ is that then the sample variance s^2x provides a better estimate of the variance of the entire population from which the sample was drawn.

In both these situations, we have **paired samples**, that is observations of two variables are made for ten individuals: doctor visits and sick days in the first case; daily vitamin C and sick days in the second case. The scatter diagrams point to a relationship between these variables, and there is a way to use the sample to compute a number, called the correlation coefficient, that measures the degree to which the variables are associated.

To motivate the definition of the correlation coefficient, suppose two paired samples $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$, and $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}$ are given and consider the centred samples

$$\mathbf{x}_c = \begin{bmatrix} x_1 - \overline{x} & x_2 - \overline{x} & \cdots & x_n - \overline{x} \end{bmatrix} \text{ and } \mathbf{y}_c = \begin{bmatrix} y_1 - \overline{y} & y_2 - \overline{y} & \cdots & y_n - \overline{y} \end{bmatrix}$$

If x_k is large among the x_i's, then the deviation $x_k - \overline{x}$ will be positive; and $x_k - \overline{x}$ will be negative if x_k is small among the x_i's. The situation is similar for $\mathbf{y}$, and the following table displays the sign of the quantity $(x_i - \overline{x})(y_k - \overline{y})$ in all four cases:

Sign of $(x_i - \overline{x})(y_k - \overline{y})$:

	x_i large	x_i small
y_i large	positive	negative
y_i small	negative	positive

Intuitively, if $\mathbf{x}$ and $\mathbf{y}$ are positively correlated, then two things happen:

1. *Large values of the x_i tend to be associated with large values of the y_i, and*
2. *Small values of the x_i tend to be associated with small values of the y_i.*

It follows from the table that, if $\mathbf{x}$ and $\mathbf{y}$ are positively correlated, then the dot product

$$\mathbf{x}_c \cdot \mathbf{y}_c = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

is positive. Similarly $\mathbf{x}_c \cdot \mathbf{y}_c$ is negative if $\mathbf{x}$ and $\mathbf{y}$ are negatively correlated. With this in mind, the **sample correlation coefficient**[18] r is defined by

$$r = r(\mathbf{x},\ \mathbf{y}) = \frac{\mathbf{x}_c \cdot \mathbf{y}_c}{\|\mathbf{x}_c\|\ \|\mathbf{y}_c\|}$$

Bearing the situation in $\mathbb{R}^3$ in mind, r is the cosine of the "angle" between the vectors $\mathbf{x}_c$ and $\mathbf{y}_c$, and so we would expect it to lie between -1 and 1. Moreover, we would expect r to be near 1 (or -1) if these vectors were pointing in the same (opposite) direction, that is the "angle" is near zero (or π).

This is confirmed by Theorem 5.7.1 below, and it is also borne out in the examples above. If we compute the correlation between sick days and visits to the physician (in the first scatter diagram above) the result is $r = 0.90$ as expected. On the other hand, the correlation between daily vitamin C doses and sick days (second scatter diagram) is $r = -0.84$.

However, a word of caution is in order here. We *cannot* conclude from the second example that taking more vitamin C will reduce the number of sick days at work. The (negative) correlation may arise because

[18]The idea of using a single number to measure the degree of relationship between different variables was pioneered by Francis Galton (1822–1911). He was studying the degree to which characteristics of an offspring relate to those of its parents. The idea was refined by Karl Pearson (1857–1936) and r is often referred to as the Pearson correlation coefficient.

of some third factor that is related to both variables. For example, case it may be that less healthy people are inclined to take more vitamin C. Correlation does *not* imply causation. Similarly, the correlation between sick days and visits to the doctor does not mean that having many sick days *causes* more visits to the doctor. A correlation between two variables may point to the existence of other underlying factors, but it does not necessarily mean that there is a causality relationship between the variables.

Our discussion of the dot product in $\mathbb{R}^n$ provides the basic properties of the correlation coefficient:

Theorem 5.7.1

Let $\mathbf{x} = \left[\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array} \right]$ *and* $\mathbf{y} = \left[\begin{array}{cccc} y_1 & y_2 & \cdots & y_n \end{array} \right]$ *be (nonzero) paired samples, and let* $r = r(\mathbf{x},\ \mathbf{y})$ *denote the correlation coefficient. Then:*

1. $-1 \leq r \leq 1$.
2. $r = 1$ *if and only if there exist* a *and* $b > 0$ *such that* $y_i = a + bx_i$ *for each* i.
3. $r = -1$ *if and only if there exist* a *and* $b < 0$ *such that* $y_i = a + bx_i$ *for each* i.

Proof. The Cauchy inequality (Theorem 5.3.2) proves (1), and also shows that $r = \pm 1$ if and only if one of $\mathbf{x}_c$ and $\mathbf{y}_c$ is a scalar multiple of the other. This in turn holds if and only if $\mathbf{y}_c = b\mathbf{x}_c$ for some $b \neq 0$, and it is easy to verify that $r = 1$ when $b > 0$ and $r = -1$ when $b < 0$.

Finally, $\mathbf{y}_c = b\mathbf{x}_c$ means $y_i - \overline{y} = b(x_i - \overline{x})$ for each i; that is, $y_i = a + bx_i$ where $a = \overline{y} - b\overline{x}$. Conversely, if $y_i = a + bx_i$, then $\overline{y} = a + b\overline{x}$ (verify), so $y_1 - \overline{y} = (a + bx_i) - (a + b\overline{x}) = b(x_1 - \overline{x})$ for each i. In other words, $\mathbf{y}_c = b\mathbf{x}_c$. This completes the proof. □

Properties (2) and (3) in Theorem 5.7.1 show that $r(\mathbf{x},\ \mathbf{y}) = 1$ means that there is a linear relation with *positive* slope between the paired data (so large x values are paired with large y values). Similarly, $r(\mathbf{x},\ \mathbf{y}) = -1$ means that there is a linear relation with *negative* slope between the paired data (so small x values are paired with small y values). This is borne out in the two scatter diagrams above.

We conclude by using the dot product to derive some useful formulas for computing variances and correlation coefficients. Given samples $\mathbf{x} = \left[\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array} \right]$ and $\mathbf{y} = \left[\begin{array}{cccc} y_1 & y_2 & \cdots & y_n \end{array} \right]$, the key observation is the following formula:

$$\mathbf{x}_c \cdot \mathbf{y}_c = \mathbf{x} \cdot \mathbf{y} - n\overline{x}\,\overline{y} \tag{5.7}$$

Indeed, remembering that $\overline{x}$ and $\overline{y}$ are scalars:

$$\begin{aligned}
\mathbf{x}_c \cdot \mathbf{y}_c &= (\mathbf{x} - \overline{x}\mathbf{1}) \cdot (\mathbf{y} - \overline{y}\mathbf{1}) \\
&= \mathbf{x} \cdot \mathbf{y} - \mathbf{x} \cdot (\overline{y}\mathbf{1}) - (\overline{x}\mathbf{1}) \cdot \mathbf{y} + (\overline{x}\mathbf{1})(\overline{y}\mathbf{1}) \\
&= \mathbf{x} \cdot \mathbf{y} - \overline{y}(\mathbf{x} \cdot \mathbf{1}) - \overline{x}(\mathbf{1} \cdot \mathbf{y}) + \overline{x}\overline{y}(\mathbf{1} \cdot \mathbf{1}) \\
&= \mathbf{x} \cdot \mathbf{y} - \overline{y}(n\overline{x}) - \overline{x}(n\overline{y}) + \overline{x}\,\overline{y}(n) \\
&= \mathbf{x} \cdot \mathbf{y} - n\overline{x}\,\overline{y}
\end{aligned}$$

Taking $\mathbf{y} = \mathbf{x}$ in (5.7) gives a formula for the variance $s_x^2 = \frac{1}{n-1}\|\mathbf{x}_c\|^2$ of $\mathbf{x}$.

Theorem: Variance Formula

If x is a sample vector, then $s_x^2 = \frac{1}{n-1}\left(\|\mathbf{x}_c\|^2 - n\overline{x}^2\right)$.

We also get a convenient formula for the correlation coefficient, $r = r(\mathbf{x},\ \mathbf{y}) = \frac{\mathbf{x}_c \cdot \mathbf{y}_c}{\|\mathbf{x}_c\|\ \|\mathbf{y}_c\|}$. Moreover, (5.7) and the fact that $s_x^2 = \frac{1}{n-1}\|\mathbf{x}_c\|^2$ give:

Theorem: Correlation Formula

If $\mathbf{x}$ and $\mathbf{y}$ are sample vectors, then

$$r = r(\mathbf{x},\ \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y} - n\overline{x}\,\overline{y}}{(n-1)s_x s_y}$$

Finally, we give a method that simplifies the computations of variances and correlations.

Theorem: Data Scaling

Let $\mathbf{x} = \left[\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array}\right]$ and $\mathbf{y} = \left[\begin{array}{cccc} y_1 & y_2 & \cdots & y_n \end{array}\right]$ be sample vectors. Given constants a, b, c, and d, consider new samples $\mathbf{z} = \left[\begin{array}{cccc} z_1 & z_2 & \cdots & z_n \end{array}\right]$ and $\mathbf{w} = \left[\begin{array}{cccc} w_1 & w_2 & \cdots & w_n \end{array}\right]$ where $z_i = a + bx_i$, for each i and $w_i = c + dy_i$ for each i. Then:

a. $\overline{z} = a + b\overline{x}$

b. $s_z^2 = b^2 s_x^2$, so $s_z = |b| s_x$

c. *If b and d have the same sign, then $r(\mathbf{x},\ \mathbf{y}) = r(\mathbf{z},\ \mathbf{w})$.*

The verification is left as an exercise. For example, if $\mathbf{x} = \left[\begin{array}{cccccc} 101 & 98 & 103 & 99 & 100 & 97 \end{array}\right]$, subtracting 100 yields $\mathbf{z} = \left[\begin{array}{cccccc} 1 & -2 & 3 & -1 & 0 & -3 \end{array}\right]$. A routine calculation shows that $\overline{z} = -\frac{1}{3}$ and $s_z^2 = \frac{14}{3}$, so $\overline{x} = 100 - \frac{1}{3} = 99.67$, and $s_z^2 = \frac{14}{3} = 4.67$.

Chapter 6

Vector Spaces

In this chapter we introduce vector spaces in full generality. The reader will notice some similarity with the discussion of the space $\mathbb{R}^n$ in Chapter 5. In fact much of the present material has been developed in that context, and there is some repetition. However, Chapter 6 deals with the notion of an *abstract* vector space, a concept that will be new to most readers. It turns out that there are many systems in which a natural addition and scalar multiplication are defined and satisfy the usual rules familiar from $\mathbb{R}^n$. The study of abstract vector spaces is a way to deal with all these examples simultaneously. The new aspect is that we are dealing with an abstract system in which *all we know* about the vectors is that they are objects that can be added and multiplied by a scalar and satisfy rules familiar from $\mathbb{R}^n$.

The novel thing is the *abstraction*. Getting used to this new conceptual level is facilitated by the work done in Chapter 5: First, the vector manipulations are familiar, giving the reader more time to become accustomed to the abstract setting; and, second, the mental images developed in the concrete setting of $\mathbb{R}^n$ serve as an aid to doing many of the exercises in Chapter 6.

The concept of a vector space was first introduced in 1844 by the German mathematician Hermann Grassmann (1809-1877), but his work did not receive the attention it deserved. It was not until 1888 that the Italian mathematician Guiseppe Peano (1858-1932) clarified Grassmann's work in his book *Calcolo Geometrico* and gave the vector space axioms in their present form. Vector spaces became established with the work of the Polish mathematician Stephan Banach (1892-1945), and the idea was finally accepted in 1918 when Hermann Weyl (1885-1955) used it in his widely read book *Raum-Zeit-Materie* ("Space-Time-Matter"), an introduction to the general theory of relativity.

6.1 Examples and Basic Properties

Many mathematical entities have the property that they can be added and multiplied by a number. Numbers themselves have this property, as do $m \times n$ matrices: The sum of two such matrices is again $m \times n$ as is any scalar multiple of such a matrix. Polynomials are another familiar example, as are the geometric vectors in Chapter 4. It turns out that there are many other types of mathematical objects that can be added and multiplied by a scalar, and the general study of such systems is introduced in this chapter. Remarkably, much of what we could say in Chapter 5 about the dimension of subspaces in $\mathbb{R}^n$ can be formulated in this generality.

Definition 6.1 Vector Spaces

*A **vector space** consists of a nonempty set V of objects (called **vectors**) that can be added, that can be multiplied by a real number (called a **scalar** in this context), and for which certain axioms hold.*[1] *If $\mathbf{v}$ and $\mathbf{w}$ are two vectors in V, their sum is expressed as $\mathbf{v}+\mathbf{w}$, and the scalar product of $\mathbf{v}$ by a real number a is denoted as $a\mathbf{v}$. These operations are called **vector addition** and **scalar multiplication**, respectively, and the following axioms are assumed to hold.*

[1] The scalars will usually be real numbers, but they could be complex numbers, or elements of an algebraic system called a field. Another example is the field $\mathbb{Q}$ of rational numbers. We will look briefly at finite fields in Section 8.8.

Axioms for vector addition

A1. *If* $\mathbf{u}$ *and* $\mathbf{v}$ *are in* V*, then* $\mathbf{u}+\mathbf{v}$ *is in* V*.*

A2. $\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$ *for all* $\mathbf{u}$ *and* $\mathbf{v}$ *in* V*.*

A3. $\mathbf{u}+(\mathbf{v}+\mathbf{w})=(\mathbf{u}+\mathbf{v})+\mathbf{w}$ *for all* $\mathbf{u}$, $\mathbf{v}$, *and* $\mathbf{w}$ *in* V*.*

A4. *An element* $\mathbf{0}$ *in* V *exists such that* $\mathbf{v}+\mathbf{0}=\mathbf{v}=\mathbf{0}+\mathbf{v}$ *for every* $\mathbf{v}$ *in* V*.*

A5. *For each* $\mathbf{v}$ *in* V*, an element* $-\mathbf{v}$ *in* V *exists such that* $-\mathbf{v}+\mathbf{v}=\mathbf{0}$ *and* $\mathbf{v}+(-\mathbf{v})=\mathbf{0}$.

Axioms for scalar multiplication

S1. *If* $\mathbf{v}$ *is in* V*, then* $a\mathbf{v}$ *is in* V *for all* a *in* $\mathbb{R}$*.*

S2. $a(\mathbf{v}+\mathbf{w})=a\mathbf{v}+a\mathbf{w}$ *for all* $\mathbf{v}$ *and* $\mathbf{w}$ *in* V *and all* a *in* $\mathbb{R}$*.*

S3. $(a+b)\mathbf{v}=a\mathbf{v}+b\mathbf{v}$ *for all* $\mathbf{v}$ *in* V *and all* a *and* b *in* $\mathbb{R}$*.*

S4. $a(b\mathbf{v})=(ab)\mathbf{v}$ *for all* $\mathbf{v}$ *in* V *and all* a *and* b *in* $\mathbb{R}$*.*

S5. $1\mathbf{v}=\mathbf{v}$ *for all* $\mathbf{v}$ *in* V*.*

The content of axioms A1 and S1 is described by saying that V is **closed** under vector addition and scalar multiplication. The element $\mathbf{0}$ in axiom A4 is called the **zero vector**, and the vector $-\mathbf{v}$ in axiom A5 is called the **negative** of $\mathbf{v}$.

The rules of matrix arithmetic, when applied to $\mathbb{R}^n$, give

Example 6.1.1

$\mathbb{R}^n$ is a vector space using matrix addition and scalar multiplication.[2]

It is important to realize that, in a general vector space, the vectors need not be n-tuples as in $\mathbb{R}^n$. They can be any kind of objects at all as long as the addition and scalar multiplication are defined and the axioms are satisfied. The following examples illustrate the diversity of the concept.

The space $\mathbb{R}^n$ consists of special types of matrices. More generally, let $\mathbf{M}_{mn}$ denote the set of all $m\times n$ matrices with real entries. Then Theorem 2.1.1 gives:

Example 6.1.2

The set $\mathbf{M}_{mn}$ of all $m\times n$ matrices is a vector space using matrix addition and scalar multiplication. The zero element in this vector space is the zero matrix of size $m\times n$, and the vector space negative of a matrix (required by axiom A5) is the usual matrix negative discussed in Section 2.1. Note that $\mathbf{M}_{mn}$ is just $\mathbb{R}^{mn}$ in different notation.

In Chapter 5 we identified many important subspaces of $\mathbb{R}^n$ such as $\operatorname{im} A$ and $\operatorname{null} A$ for a matrix A. These are all vector spaces.

[2]We will usually write the vectors in $\mathbb{R}^n$ as n-tuples. However, if it is convenient, we will sometimes denote them as rows or columns.

Example 6.1.3

Show that every subspace of $\mathbb{R}^n$ is a vector space in its own right using the addition and scalar multiplication of $\mathbb{R}^n$.

Solution. Axioms A1 and S1 are two of the defining conditions for a subspace U of $\mathbb{R}^n$ (see Section 5.1). The other eight axioms for a vector space are inherited from $\mathbb{R}^n$. For example, if $\mathbf{x}$ and $\mathbf{y}$ are in U and a is a scalar, then $a(\mathbf{x}+\mathbf{y}) = a\mathbf{x}+a\mathbf{y}$ because $\mathbf{x}$ and $\mathbf{y}$ are in $\mathbb{R}^n$. This shows that axiom S2 holds for U; similarly, the other axioms also hold for U.

Example 6.1.4

Let V denote the set of all ordered pairs $(x,\ y)$ and define addition in V as in $\mathbb{R}^2$. However, define a new scalar multiplication in V by

$$a(x,\ y) = (ay,\ ax)$$

Determine if V is a vector space with these operations.

Solution. Axioms A1 to A5 are valid for V because they hold for matrices. Also $a(x,\ y) = (ay,\ ax)$ is again in V, so axiom S1 holds. To verify axiom S2, let $\mathbf{v} = (x,\ y)$ and $\mathbf{w} = (x_1,\ y_1)$ be typical elements in V and compute

$$\begin{aligned} a(\mathbf{v}+\mathbf{w}) &= a(x+x_1,\ y+y_1) = (a(y+y_1),\ a(x+x_1)) \\ a\mathbf{v}+a\mathbf{w} &= (ay,\ ax)+(ay_1,\ ax_1) = (ay+ay_1,\ ax+ax_1) \end{aligned}$$

Because these are equal, axiom S2 holds. Similarly, the reader can verify that axiom S3 holds. However, axiom S4 fails because

$$a(b(x,\ y)) = a(by,\ bx) = (abx,\ aby)$$

need not equal $ab(x,\ y) = (aby,\ abx)$. Hence, V is *not* a vector space. (In fact, axiom S5 also fails.)

Sets of polynomials provide another important source of examples of vector spaces, so we review some basic facts. A **polynomial** in an indeterminate x is an expression

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$$

where $a_0,\ a_1,\ a_2,\ \ldots,\ a_n$ are real numbers called the **coefficients** of the polynomial. If all the coefficients are zero, the polynomial is called the **zero polynomial** and is denoted simply as 0. If $p(x) \neq 0$, the highest power of x with a nonzero coefficient is called the **degree** of $p(x)$ denoted as $\deg p(x)$. The coefficient itself is called the **leading coefficient** of $p(x)$. Hence $\deg(3+5x) = 1$, $\deg(1+x+x^2) = 2$, and $\deg(4) = 0$. (The degree of the zero polynomial is not defined.)

Let $\mathbf{P}$ denote the set of all polynomials and suppose that

$$\begin{aligned} p(x) &= a_0 + a_1x + a_2x^2 + \cdots \\ q(x) &= b_0 + b_1x + b_2x^2 + \cdots \end{aligned}$$

are two polynomials in $\mathbf{P}$ (possibly of different degrees). Then $p(x)$ and $q(x)$ are called **equal** [written $p(x) = q(x)$] if and only if all the corresponding coefficients are equal—that is, $a_0 = b_0$, $a_1 = b_1$, $a_2 = b_2$, and so on. In particular, $a_0 + a_1x + a_2x^2 + \cdots = 0$ means that $a_0 = 0$, $a_1 = 0$, $a_2 = 0$, $\dots$, and this is the reason for calling x an **indeterminate**. The set $\mathbf{P}$ has an addition and scalar multiplication defined on it as follows: if $p(x)$ and $q(x)$ are as before and a is a real number,

$$\begin{aligned} p(x) + q(x) &= (a_0 + b_0) + (a_1 + b_1)x + (a_2 + b_2)x^2 + \cdots \\ ap(x) &= aa_0 + (aa_1)x + (aa_2)x^2 + \cdots \end{aligned}$$

Evidently, these are again polynomials, so $\mathbf{P}$ is closed under these operations, called **pointwise** addition and scalar multiplication. The other vector space axioms are easily verified, and we have

Example 6.1.5

The set $\mathbf{P}$ of all polynomials is a vector space with the foregoing addition and scalar multiplication. The zero vector is the zero polynomial, and the negative of a polynomial $p(x) = a_0 + a_1x + a_2x^2 + \dots$ is the polynomial $-p(x) = -a_0 - a_1x - a_2x^2 - \dots$ obtained by negating all the coefficients.

There is another vector space of polynomials that will be referred to later.

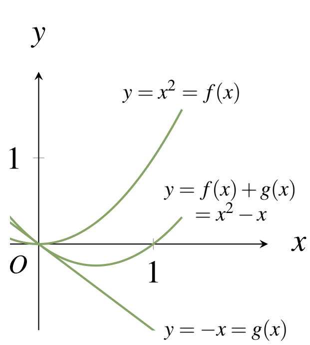
Example 6.1.6

Given $n \geq 1$, let $\mathbf{P}_n$ denote the set of all polynomials of degree at most n, together with the zero polynomial. That is

$$\mathbf{P}_n = \{a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \mid a_0,\ a_1,\ a_2,\ \dots,\ a_n \text{ in } \mathbb{R}\}.$$

Then $\mathbf{P}_n$ is a vector space. Indeed, sums and scalar multiples of polynomials in $\mathbf{P}_n$ are again in $\mathbf{P}_n$, and the other vector space axioms are inherited from $\mathbf{P}$. In particular, the zero vector and the negative of a polynomial in $\mathbf{P}_n$ are the same as those in $\mathbf{P}$.

If a and b are real numbers and $a < b$, the **interval** $[a,\ b]$ is defined to be the set of all real numbers x such that $a \leq x \leq b$. A (real-valued) **function** f on $[a,\ b]$ is a rule that associates to every number x in $[a,\ b]$ a real number denoted $f(x)$. The rule is frequently specified by giving a formula for $f(x)$ in terms of x. For example, $f(x) = 2^x$, $f(x) = \sin x$, and $f(x) = x^2 + 1$ are familiar functions. In fact, every polynomial $p(x)$ can be regarded as the formula for a function p.



The set of all functions on $[a,\ b]$ is denoted $\mathbf{F}[a,\ b]$. Two functions f and g in $\mathbf{F}[a,\ b]$ are **equal** if $f(x) = g(x)$ for every x in $[a,\ b]$, and we describe this by saying that f and g have the **same action**. Note that two polynomials are equal in $\mathbf{P}$ (defined prior to Example 6.1.5) if and only if they are equal as functions.

If f and g are two functions in $\mathbf{F}[a,\ b]$, and if r is a real number, define the sum $f + g$ and the scalar product rf by

$$(f + g)(x) = f(x) + g(x) \quad \text{for each } x \text{ in } [a,\ b]$$

$$(rf)(x) = rf(x) \qquad \text{for each } x \text{ in } [a,\ b]$$

In other words, the action of $f+g$ upon x is to associate x with the number $f(x)+g(x)$, and rf associates x with $rf(x)$. The sum of $f(x) = x^2$ and $g(x) = -x$ is shown in the diagram. These operations on $\mathbf{F}[a,\ b]$ are called **pointwise addition and scalar multiplication** of functions and they are the usual operations familiar from elementary algebra and calculus.

Example 6.1.7

The set $\mathbf{F}[a,\ b]$ of all functions on the interval $[a,\ b]$ is a vector space using pointwise addition and scalar multiplication. The zero function (in axiom A4), denoted 0, is the constant function defined by

$$0(x) = 0 \qquad \text{for each } x \text{ in } [a,\ b]$$

The negative of a function f is denoted $-f$ and has action defined by

$$(-f)(x) = -f(x) \qquad \text{for each } x \text{ in } [a,\ b]$$

Axioms A1 and S1 are clearly satisfied because, if f and g are functions on $[a,\ b]$, then $f+g$ and rf are again such functions. The verification of the remaining axioms is left as Exercise **??**.

Other examples of vector spaces will appear later, but these are sufficiently varied to indicate the scope of the concept and to illustrate the properties of vector spaces to be discussed. With such a variety of examples, it may come as a surprise that a well-developed *theory* of vector spaces exists. That is, many properties can be shown to hold for *all* vector spaces and hence hold in every example. Such properties are called *theorems* and can be deduced from the axioms. Here is an important example.

Theorem 6.1.1: Cancellation

Let $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ be vectors in a vector space V. If $\mathbf{v}+\mathbf{u} = \mathbf{v}+\mathbf{w}$, then $\mathbf{u} = \mathbf{w}$.

Proof. We are given $\mathbf{v}+\mathbf{u} = \mathbf{v}+\mathbf{w}$. If these were numbers instead of vectors, we would simply subtract $\mathbf{v}$ from both sides of the equation to obtain $\mathbf{u} = \mathbf{w}$. This can be accomplished with vectors by adding $-\mathbf{v}$ to both sides of the equation. The steps (using only the axioms) are as follows:

$$\begin{aligned}
\mathbf{v}+\mathbf{u} &= \mathbf{v}+\mathbf{w} \\
-\mathbf{v}+(\mathbf{v}+\mathbf{u}) &= -\mathbf{v}+(\mathbf{v}+\mathbf{w}) && \text{(axiom A5)} \\
(-\mathbf{v}+\mathbf{v})+\mathbf{u} &= (-\mathbf{v}+\mathbf{v})+\mathbf{w} && \text{(axiom A3)} \\
\mathbf{0}+\mathbf{u} &= \mathbf{0}+\mathbf{w} && \text{(axiom A5)} \\
\mathbf{u} &= \mathbf{w} && \text{(axiom A4)}
\end{aligned}$$

This is the desired conclusion.[3] □

As with many good mathematical theorems, the technique of the proof of Theorem 6.1.1 is at least as important as the theorem itself. The idea was to mimic the well-known process of numerical subtraction

[3]Observe that none of the scalar multiplication axioms are needed here.

in a vector space V as follows: To subtract a vector $\mathbf{v}$ from both sides of a vector equation, we added $-\mathbf{v}$ to both sides. With this in mind, we define **difference** $\mathbf{u}-\mathbf{v}$ of two vectors in V as

$$\mathbf{u}-\mathbf{v}=\mathbf{u}+(-\mathbf{v})$$

We shall say that this vector is the result of having **subtracted** $\mathbf{v}$ from $\mathbf{u}$ and, as in arithmetic, this operation has the property given in Theorem 6.1.2.

Theorem 6.1.2

If $\mathbf{u}$ and $\mathbf{v}$ are vectors in a vector space V, the equation

$$\mathbf{x}+\mathbf{v}=\mathbf{u}$$

has one and only one solution $\mathbf{x}$ in V given by

$$\mathbf{x}=\mathbf{u}-\mathbf{v}$$

Proof. The difference $\mathbf{x}=\mathbf{u}-\mathbf{v}$ is indeed a solution to the equation because (using several axioms)

$$\mathbf{x}+\mathbf{v}=(\mathbf{u}-\mathbf{v})+\mathbf{v}=[\mathbf{u}+(-\mathbf{v})]+\mathbf{v}=\mathbf{u}+(-\mathbf{v}+\mathbf{v})=\mathbf{u}+\mathbf{0}=\mathbf{u}$$

To see that this is the only solution, suppose $\mathbf{x}_1$ is another solution so that $\mathbf{x}_1+\mathbf{v}=\mathbf{u}$. Then $\mathbf{x}+\mathbf{v}=\mathbf{x}_1+\mathbf{v}$ (they both equal $\mathbf{u}$), so $\mathbf{x}=\mathbf{x}_1$ by cancellation. □

Similarly, cancellation shows that there is only one zero vector in any vector space and only one negative of each vector (Exercises **??** and **??**). Hence we speak of *the* zero vector and *the* negative of a vector.

The next theorem derives some basic properties of scalar multiplication that hold in every vector space, and will be used extensively.

Theorem 6.1.3

Let $\mathbf{v}$ denote a vector in a vector space V and let a denote a real number.

1. $0\mathbf{v}=\mathbf{0}$.
2. $a\mathbf{0}=\mathbf{0}$.
3. *If* $a\mathbf{v}=\mathbf{0}$, *then either* $a=0$ *or* $\mathbf{v}=\mathbf{0}$.
4. $(-1)\mathbf{v}=-\mathbf{v}$.
5. $(-a)\mathbf{v}=-(a\mathbf{v})=a(-\mathbf{v})$.

Proof.

1. Observe that $0\mathbf{v}+0\mathbf{v}=(0+0)\mathbf{v}=0\mathbf{v}=0\mathbf{v}+\mathbf{0}$ where the first equality is by axiom S3. It follows that $0\mathbf{v}=\mathbf{0}$ by cancellation.

2. The proof is similar to that of (1), and is left as Exercise **??**(a).

3. Assume that $a\mathbf{v} = \mathbf{0}$. If $a = 0$, there is nothing to prove; if $a \neq 0$, we must show that $\mathbf{v} = \mathbf{0}$. But $a \neq 0$ means we can scalar-multiply the equation $a\mathbf{v} = \mathbf{0}$ by the scalar $\frac{1}{a}$. The result (using (2) and Axioms S5 and S4) is
$$\mathbf{v} = 1\mathbf{v} = \left(\frac{1}{a}a\right)\mathbf{v} = \frac{1}{a}(a\mathbf{v}) = \frac{1}{a}\mathbf{0} = \mathbf{0}$$

4. We have $-\mathbf{v} + \mathbf{v} = \mathbf{0}$ by axiom A5. On the other hand,
$$(-1)\mathbf{v} + \mathbf{v} = (-1)\mathbf{v} + 1\mathbf{v} = (-1+1)\mathbf{v} = 0\mathbf{v} = \mathbf{0}$$
using (1) and axioms S5 and S3. Hence $(-1)\mathbf{v} + \mathbf{v} = -\mathbf{v} + \mathbf{v}$ (because both are equal to $\mathbf{0}$), so $(-1)\mathbf{v} = -\mathbf{v}$ by cancellation.

5. The proof is left as Exercise **??**.[4] □

The properties in Theorem 6.1.3 are familiar for matrices; the point here is that they hold in *every* vector space. It is hard to exaggerate the importance of this observation.

Axiom A3 ensures that the sum $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$ is the same however it is formed, and we write it simply as $\mathbf{u} + \mathbf{v} + \mathbf{w}$. Similarly, there are different ways to form any sum $\mathbf{v}_1 + \mathbf{v}_2 + \cdots + \mathbf{v}_n$, and Axiom A3 guarantees that they are all equal. Moreover, Axiom A2 shows that the order in which the vectors are written does not matter (for example: $\mathbf{u} + \mathbf{v} + \mathbf{w} + \mathbf{z} = \mathbf{z} + \mathbf{u} + \mathbf{w} + \mathbf{v}$).

Similarly, Axioms S2 and S3 extend. For example
$$a(\mathbf{u} + \mathbf{v} + \mathbf{w}) = a\left[\mathbf{u} + (\mathbf{v} + \mathbf{w})\right] = a\mathbf{u} + a(\mathbf{v} + \mathbf{w}) = a\mathbf{u} + a\mathbf{v} + a\mathbf{w}$$
for all a, $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$. Similarly $(a+b+c)\mathbf{v} = a\mathbf{v} + b\mathbf{v} + c\mathbf{v}$ hold for all values of a, b, c, and $\mathbf{v}$ (verify). More generally,
$$\begin{aligned} a(\mathbf{v}_1 + \mathbf{v}_2 + \cdots + \mathbf{v}_n) &= a\mathbf{v}_1 + a\mathbf{v}_2 + \cdots + a\mathbf{v}_n \\ (a_1 + a_2 + \cdots + a_n)\mathbf{v} &= a_1\mathbf{v} + a_2\mathbf{v} + \cdots + a_n\mathbf{v} \end{aligned}$$
hold for all $n \geq 1$, all numbers a, $a_1, \dots, a_n$, and all vectors, $\mathbf{v}$, $\mathbf{v}_1, \dots, \mathbf{v}_n$. The verifications are by induction and are left to the reader (Exercise **??**). These facts—together with the axioms, Theorem 6.1.3, and the definition of subtraction—enable us to simplify expressions involving sums of scalar multiples of vectors by collecting like terms, expanding, and taking out common factors. This has been discussed for the vector space of matrices in Section 2.1 (and for geometric vectors in Section 4.1); the manipulations in an arbitrary vector space are carried out in the same way. Here is an illustration.

Example 6.1.8

If $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ are vectors in a vector space V, simplify the expression
$$2(\mathbf{u} + 3\mathbf{w}) - 3(2\mathbf{w} - \mathbf{v}) - 3[2(2\mathbf{u} + \mathbf{v} - 4\mathbf{w}) - 4(\mathbf{u} - 2\mathbf{w})]$$

Solution. The reduction proceeds as though $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ were matrices or variables.
$$2(\mathbf{u} + 3\mathbf{w}) - 3(2\mathbf{w} - \mathbf{v}) - 3[2(2\mathbf{u} + \mathbf{v} - 4\mathbf{w}) - 4(\mathbf{u} - 2\mathbf{w})]$$

$$\begin{aligned}
&= 2\mathbf{u} + 6\mathbf{w} - 6\mathbf{w} + 3\mathbf{v} - 3[4\mathbf{u} + 2\mathbf{v} - 8\mathbf{w} - 4\mathbf{u} + 8\mathbf{w}] \\
&= 2\mathbf{u} + 3\mathbf{v} - 3[2\mathbf{v}] \\
&= 2\mathbf{u} + 3\mathbf{v} - 6\mathbf{v} \\
&= 2\mathbf{u} - 3\mathbf{v}
\end{aligned}$$

Condition (2) in Theorem 6.1.3 points to another example of a vector space.

Example 6.1.9

A set $\{\mathbf{0}\}$ with one element becomes a vector space if we define

$$\mathbf{0} + \mathbf{0} = \mathbf{0} \quad \text{and} \quad a\mathbf{0} = \mathbf{0} \text{ for all scalars } a.$$

The resulting space is called the **zero vector space** and is denoted $\{\mathbf{0}\}$.

The vector space axioms are easily verified for $\{\mathbf{0}\}$. In any vector space V, Theorem 6.1.3 shows that the zero subspace (consisting of the zero vector of V alone) is a copy of the zero vector space.

6.2 Subspaces and Spanning Sets

Chapter 5 is essentially about the subspaces of $\mathbb{R}^n$. We now extend this notion.

Definition 6.2 Subspaces of a Vector Space

*If V is a vector space, a nonempty subset $U \subseteq V$ is called a **subspace** of V if U is itself a vector space using the addition and scalar multiplication of V.*

Subspaces of $\mathbb{R}^n$ (as defined in Section 5.1) are subspaces in the present sense by Example 6.1.3. Moreover, the defining properties for a subspace of $\mathbb{R}^n$ actually *characterize* subspaces in general.

Theorem 6.2.1: Subspace Test

A subset U of a vector space is a subspace of V if and only if it satisfies the following three conditions:

1. *$\mathbf{0}$ lies in U where $\mathbf{0}$ is the zero vector of V.*
2. *If $\mathbf{u}_1$ and $\mathbf{u}_2$ are in U, then $\mathbf{u}_1 + \mathbf{u}_2$ is also in U.*
3. *If $\mathbf{u}$ is in U, then $a\mathbf{u}$ is also in U for each scalar a.*

Proof. If U is a subspace of V, then (2) and (3) hold by axioms A1 and S1 respectively, applied to the vector space U. Since U is nonempty (it is a vector space), choose $\mathbf{u}$ in U. Then (1) holds because $\mathbf{0} = 0\mathbf{u}$ is in U by (3) and Theorem 6.1.3.

Conversely, if (1), (2), and (3) hold, then axioms A1 and S1 hold because of (2) and (3), and axioms A2, A3, S2, S3, S4, and S5 hold in U because they hold in V. Axiom A4 holds because the zero vector $\mathbf{0}$ of V is actually in U by (1), and so serves as the zero of U. Finally, given $\mathbf{u}$ in U, then its negative $-\mathbf{u}$ in V is again in U by (3) because $-\mathbf{u} = (-1)\mathbf{u}$ (again using Theorem 6.1.3). Hence $-\mathbf{u}$ serves as the negative of $\mathbf{u}$ in U. □

Note that the proof of Theorem 6.2.1 shows that if U is a subspace of V, then U and V share the same zero vector, and that the negative of a vector in the space U is the same as its negative in V.

Example 6.2.1

If V is any vector space, show that $\{\mathbf{0}\}$ and V are subspaces of V.

Solution. $U = V$ clearly satisfies the conditions of the subspace test. As to $U = \{\mathbf{0}\}$, it satisfies the conditions because $\mathbf{0} + \mathbf{0} = \mathbf{0}$ and $a\mathbf{0} = \mathbf{0}$ for all a in $\mathbb{R}$.

The vector space $\{\mathbf{0}\}$ is called the **zero subspace** of V.

Example 6.2.2

Let $\mathbf{v}$ be a vector in a vector space V. Show that the set

$$\mathbb{R}\mathbf{v} = \{a\mathbf{v} \mid a \text{ in } \mathbb{R}\}$$

of all scalar multiples of $\mathbf{v}$ is a subspace of V.

Solution. Because $\mathbf{0} = 0\mathbf{v}$, it is clear that $\mathbf{0}$ lies in $\mathbb{R}\mathbf{v}$. Given two vectors $a\mathbf{v}$ and $a_1\mathbf{v}$ in $\mathbb{R}\mathbf{v}$, their sum $a\mathbf{v} + a_1\mathbf{v} = (a + a_1)\mathbf{v}$ is also a scalar multiple of $\mathbf{v}$ and so lies in $\mathbb{R}\mathbf{v}$. Hence $\mathbb{R}\mathbf{v}$ is closed under addition. Finally, given $a\mathbf{v}$, $r(a\mathbf{v}) = (ra)\mathbf{v}$ lies in $\mathbb{R}\mathbf{v}$ for all $r \in \mathbb{R}$, so $\mathbb{R}\mathbf{v}$ is closed under scalar multiplication. Hence the subspace test applies.

In particular, given $\mathbf{d} \neq \mathbf{0}$ in $\mathbb{R}^3$, $\mathbb{R}\mathbf{d}$ is the line through the origin with direction vector $\mathbf{d}$.

The space $\mathbb{R}\mathbf{v}$ in Example 6.2.2 is described by giving the *form* of each vector in $\mathbb{R}\mathbf{v}$. The next example describes a subset U of the space $\mathbf{M}_{nn}$ by giving a *condition* that each matrix of U must satisfy.

Example 6.2.3

Let A be a fixed matrix in $\mathbf{M}_{nn}$. Show that $U = \{X \text{ in } \mathbf{M}_{nn} \mid AX = XA\}$ is a subspace of $\mathbf{M}_{nn}$.

Solution. If 0 is the $n \times n$ zero matrix, then $A0 = 0A$, so 0 satisfies the condition for membership in U. Next suppose that X and X_1 lie in U so that $AX = XA$ and $AX_1 = X_1A$. Then

$$\begin{aligned} A(X + X_1) &= AX + AX_1 = XA + X_1A = (X + X_1)A \\ A(aX) &= a(AX) = a(XA) = (aX)A \end{aligned}$$

for all a in $\mathbb{R}$, so both $X + X_1$ and aX lie in U. Hence U is a subspace of $\mathbf{M}_{nn}$.

Suppose $p(x)$ is a polynomial and a is a number. Then the number $p(a)$ obtained by replacing x by a in the expression for $p(x)$ is called the **evaluation** of $p(x)$ at a. For example, if $p(x) = 5 - 6x + 2x^2$, then the evaluation of $p(x)$ at $a = 2$ is $p(2) = 5 - 12 + 8 = 1$. If $p(a) = 0$, the number a is called a **root** of $p(x)$.

Example 6.2.4

Consider the set U of all polynomials in $\mathbf{P}$ that have 3 as a root:

$$U = \{p(x) \in \mathbf{P} \mid p(3) = 0\}$$

Show that U is a subspace of $\mathbf{P}$.

Solution. Clearly, the zero polynomial lies in U. Now let $p(x)$ and $q(x)$ lie in U so $p(3) = 0$ and $q(3) = 0$. We have $(p + q)(x) = p(x) + q(x)$ for all x, so $(p + q)(3) = p(3) + q(3) = 0 + 0 = 0$, and U is closed under addition. The verification that U is closed under scalar multiplication is similar.

Recall that the space $\mathbf{P}_n$ consists of all polynomials of the form

$$a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$$

where $a_0, a_1, a_2, \dots, a_n$ are real numbers, and so is closed under the addition and scalar multiplication in $\mathbf{P}$. Moreover, the zero polynomial is included in $\mathbf{P}_n$. Thus the subspace test gives Example 6.2.5.

Example 6.2.5

$\mathbf{P}_n$ is a subspace of $\mathbf{P}$ for each $n \geq 0$.

The next example involves the notion of the derivative f' of a function f. (If the reader is not familiar with calculus, this example may be omitted.) A function f defined on the interval $[a, b]$ is called **differentiable** if the derivative $f'(r)$ exists at every r in $[a, b]$.

Example 6.2.6

Show that the subset $\mathbf{D}[a, b]$ of all **differentiable functions** on $[a, b]$ is a subspace of the vector space $\mathbf{F}[a, b]$ of all functions on $[a, b]$.

Solution. The derivative of any constant function is the constant function 0; in particular, 0 itself is differentiable and so lies in $\mathbf{D}[a, b]$. If f and g both lie in $\mathbf{D}[a, b]$ (so that f' and g' exist), then it is a theorem of calculus that $f+g$ and rf are both differentiable for any $r \in \mathbb{R}$. In fact, $(f+g)' = f' + g'$ and $(rf)' = rf'$, so both lie in $\mathbf{D}[a, b]$. This shows that $\mathbf{D}[a, b]$ is a subspace of $\mathbf{F}[a, b]$.

Linear Combinations and Spanning Sets

One of the crucial concept in linear algebra is that of a span of a set of vectors, obtained by considering all possible linear combinations of vectors in that set.

Definition 6.3 Linear Combinations and Spanning

*Let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ be a set of vectors in a vector space V. As in $\mathbb{R}^n$, a vector $\mathbf{v}$ is called a **linear combination** of the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ if it can be expressed in the form*

$$\mathbf{v} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n$$

*where $a_1, a_2, \dots, a_n$ are scalars, called the **coefficients** of $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$. The set of all linear combinations of these vectors is called their **span**, and is denoted by*

$$\text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\} = \{a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n \mid a_i \text{ in } \mathbb{R}\}$$

If it happens that $V = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$, these vectors are called a **spanning set** for V. For example, the span of two vectors $\mathbf{v}$ and $\mathbf{w}$ is the set

$$\text{span}\{\mathbf{v}, \mathbf{w}\} = \{s\mathbf{v} + t\mathbf{w} \mid s \text{ and } t \text{ in } \mathbb{R}\}$$

of all sums of scalar multiples of these vectors.

Example 6.2.7

Consider the vectors $p_1 = 1 + x + 4x^2$ and $p_2 = 1 + 5x + x^2$ in $\mathbf{P}_2$. Determine whether p_1 and p_2 lie in $\text{span}\{1 + 2x - x^2, 3 + 5x + 2x^2\}$.

Solution. For p_1, we want to determine if s and t exist such that

$$p_1 = s(1 + 2x - x^2) + t(3 + 5x + 2x^2)$$

Equating coefficients of powers of x (where $x^0 = 1$) gives

$$1 = s + 3t, \quad 1 = 2s + 5t, \quad \text{and} \quad 4 = -s + 2t$$

These equations have the solution $s = -2$ and $t = 1$, so p_1 is indeed in $\text{span}\{1 + 2x - x^2, 3 + 5x + 2x^2\}$.
Turning to $p_2 = 1 + 5x + x^2$, we are looking for s and t such that

$$p_2 = s(1 + 2x - x^2) + t(3 + 5x + 2x^2)$$

Again equating coefficients of powers of x gives equations $1 = s + 3t$, $5 = 2s + 5t$, and $1 = -s + 2t$. But in this case there is no solution, so p_2 is *not* in $\text{span}\{1 + 2x - x^2, 3 + 5x + 2x^2\}$.

We saw in Example 5.1.6 that $\mathbb{R}^m = \text{span}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m\}$ where the vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m$ are the columns of the $m \times m$ identity matrix. Of course $\mathbb{R}^m = \mathbf{M}_{m1}$ is the set of all $m \times 1$ matrices, and there is an analogous spanning set for each space $\mathbf{M}_{mn}$. For example, each 2×2 matrix has the form

$$\left[\begin{array}{cc} a & b \\ c & d \end{array}\right] = a\left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array}\right] + b\left[\begin{array}{cc} 0 & 1 \\ 0 & 0 \end{array}\right] + c\left[\begin{array}{cc} 0 & 0 \\ 1 & 0 \end{array}\right] + d\left[\begin{array}{cc} 0 & 0 \\ 0 & 1 \end{array}\right]$$

so

$$\mathbf{M}_{22} = \text{span}\left\{\left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array}\right], \left[\begin{array}{cc} 0 & 1 \\ 0 & 0 \end{array}\right], \left[\begin{array}{cc} 0 & 0 \\ 1 & 0 \end{array}\right], \left[\begin{array}{cc} 0 & 0 \\ 0 & 1 \end{array}\right]\right\}$$

Similarly, we obtain

Example 6.2.8

$\mathbf{M}_{mn}$ is the span of the set of all $m \times n$ matrices with exactly one entry equal to 1, and all other entries zero.

The fact that every polynomial in $\mathbf{P}_n$ has the form $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ where each a_i is in $\mathbb{R}$ shows that

Example 6.2.9

$\mathbf{P}_n = \text{span}\{1, x, x^2, \dots, x^n\}$.

In Example 6.2.2 we saw that $\text{span}\{\mathbf{v}\} = \{a\mathbf{v} \mid a \text{ in } \mathbb{R}\} = \mathbb{R}\mathbf{v}$ is a subspace for any vector $\mathbf{v}$ in a vector space V. More generally, the span of *any* set of vectors is a subspace. In fact, the proof of Theorem 5.1.1 goes through to prove:

Theorem 6.2.2

Let $U = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ *in a vector space* V*. Then:*

1. U *is a subspace of* V *containing each of* $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$*.*

2. U *is the "smallest" subspace containing these vectors in the sense that any subspace that contains each of* $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ *must contain* U*.*

Here is how condition 2 in Theorem 6.2.2 is used. Given vectors $\mathbf{v}_1, \dots, \mathbf{v}_k$ in a vector space V and a subspace $U \subseteq V$, then:

$$\text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_n\} \subseteq U \Leftrightarrow \text{ each } \mathbf{v}_i \in U$$

The following examples illustrate this.

Example 6.2.10

Show that $\mathbf{P}_3 = \text{span}\{x^2 + x^3, x, 2x^2 + 1, 3\}$.

Solution. Write $U = \text{span}\{x^2 + x^3, x, 2x^2 + 1, 3\}$. Then $U \subseteq \mathbf{P}_3$, and we use the fact that $\mathbf{P}_3 = \text{span}\{1, x, x^2, x^3\}$ to show that $\mathbf{P}_3 \subseteq U$. In fact, x and $1 = \frac{1}{3} \cdot 3$ clearly lie in U. But then successively,

$$x^2 = \tfrac{1}{2}[(2x^2 + 1) - 1] \quad \text{and} \quad x^3 = (x^2 + x^3) - x^2$$

also lie in U. Hence $\mathbf{P}_3 \subseteq U$ by Theorem 6.2.2.

Example 6.2.11

Let $\mathbf{u}$ and $\mathbf{v}$ be two vectors in a vector space V. Show that

$$\text{span}\{\mathbf{u}, \mathbf{v}\} = \text{span}\{\mathbf{u} + 2\mathbf{v}, \mathbf{u} - \mathbf{v}\}$$

Solution. We have $\text{span}\{\mathbf{u} + 2\mathbf{v}, \mathbf{u} - \mathbf{v}\} \subseteq \text{span}\{\mathbf{u}, \mathbf{v}\}$ by Theorem 6.2.2 because both $\mathbf{u} + 2\mathbf{v}$ and $\mathbf{u} - \mathbf{v}$ lie in $\text{span}\{\mathbf{u}, \mathbf{v}\}$. On the other hand,

$$\mathbf{u} = \tfrac{1}{3}(\mathbf{u} + 2\mathbf{v}) + \tfrac{2}{3}(\mathbf{u} - \mathbf{v}) \quad \text{and} \quad \mathbf{v} = \tfrac{1}{3}(\mathbf{u} + 2\mathbf{v}) - \tfrac{1}{3}(\mathbf{u} - \mathbf{v})$$

so $\text{span}\{\mathbf{u}, \mathbf{v}\} \subseteq \text{span}\{\mathbf{u} + 2\mathbf{v}, \mathbf{u} - \mathbf{v}\}$, again by Theorem 6.2.2.

6.3 Linear Independence and Dimension

In addition to the span of a set of vectors covered in the previous section, another central concept in linear algebra is that of an independent set of vectors, meaning roughly that the set does not contain any redundant vector.

Definition 6.4 Linear Independence and Dependence

*As in $\mathbb{R}^n$, a set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ in a vector space V is called **linearly independent** (or simply **independent**) if it satisfies the following condition:*

$$\text{If} \quad s_1\mathbf{v}_1 + s_2\mathbf{v}_2 + \cdots + s_n\mathbf{v}_n = \mathbf{0}, \quad \text{then} \quad s_1 = s_2 = \cdots = s_n = 0.$$

*A set of vectors that is not linearly independent is said to be **linearly dependent** (or simply **dependent**).*

The **trivial linear combination** of the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is the one with every coefficient zero:

$$0\mathbf{v}_1 + 0\mathbf{v}_2 + \cdots + 0\mathbf{v}_n$$

This is obviously one way of expressing $\mathbf{0}$ as a linear combination of the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$, and they are linearly independent when it is the *only* way.

Example 6.3.1

Show that $\{1+x,\ 3x+x^2,\ 2+x-x^2\}$ is independent in $\mathbf{P}_2$.

Solution. Suppose a linear combination of these polynomials vanishes.

$$s_1(1+x)+s_2(3x+x^2)+s_3(2+x-x^2)=0$$

Equating the coefficients of 1, x, and x^2 gives a set of linear equations.

$$\begin{aligned} s_1 + + 2s_3 &= 0 \\ s_1 + 3s_2 + s_3 &= 0 \\ s_2 - s_3 &= 0 \end{aligned}$$

The only solution is $s_1=s_2=s_3=0$.

Example 6.3.2

Show that $\{\sin x,\ \cos x\}$ is independent in the vector space $\mathbf{F}[0,\ 2\pi]$ of functions defined on the interval $[0,\ 2\pi]$.

Solution. Suppose that a linear combination of these functions vanishes.

$$s_1(\sin x)+s_2(\cos x)=0$$

This must hold for *all* values of x in $[0,\ 2\pi]$ (by the definition of equality in $\mathbf{F}[0,\ 2\pi]$). Taking $x=0$ yields $s_2=0$ (because $\sin 0=0$ and $\cos 0=1$). Similarly, $s_1=0$ follows from taking $x=\frac{\pi}{2}$ (because $\sin\frac{\pi}{2}=1$ and $\cos\frac{\pi}{2}=0$).

Example 6.3.3

Suppose that $\{\mathbf{u},\ \mathbf{v}\}$ is an independent set in a vector space V. Show that $\{\mathbf{u}+2\mathbf{v},\ \mathbf{u}-3\mathbf{v}\}$ is also independent.

Solution. Suppose a linear combination of $\mathbf{u}+2\mathbf{v}$ and $\mathbf{u}-3\mathbf{v}$ vanishes:

$$s(\mathbf{u}+2\mathbf{v})+t(\mathbf{u}-3\mathbf{v})=\mathbf{0}$$

We must deduce that $s=t=0$. Collecting terms involving $\mathbf{u}$ and $\mathbf{v}$ gives

$$(s+t)\mathbf{u}+(2s-3t)\mathbf{v}=\mathbf{0}$$

Because $\{\mathbf{u},\ \mathbf{v}\}$ is independent, this yields linear equations $s+t=0$ and $2s-3t=0$. The only solution is $s=t=0$.

Example 6.3.4

Show that any set of polynomials of distinct degrees is independent.

Solution. Let $p_1, p_2, \dots, p_m$ be polynomials where $\deg(p_i) = d_i$. By relabelling if necessary, we may assume that $d_1 > d_2 > \cdots > d_m$. Suppose that a linear combination vanishes:

$$t_1 p_1 + t_2 p_2 + \cdots + t_m p_m = 0$$

where each t_i is in $\mathbb{R}$. As $\deg(p_1) = d_1$, let ax^{d_1} be the term in p_1 of highest degree, where $a \neq 0$. Since $d_1 > d_2 > \cdots > d_m$, it follows that $t_1 ax^{d_1}$ is the only term of degree d_1 in the linear combination $t_1 p_1 + t_2 p_2 + \cdots + t_m p_m = 0$. This means that $t_1 ax^{d_1} = 0$, whence $t_1 a = 0$, hence $t_1 = 0$ (because $a \neq 0$). But then $t_2 p_2 + \cdots + t_m p_m = 0$ so we can repeat the argument to show that $t_2 = 0$. Continuing, we obtain $t_i = 0$ for each i, as desired.

Example 6.3.5

Suppose that A is an $n \times n$ matrix such that $A^k = 0$ but $A^{k-1} \neq 0$. Show that $B = \{I, A, A^2, \dots, A^{k-1}\}$ is independent in $\mathbf{M}_{nn}$.

Solution. Suppose $r_0 I + r_1 A + r_2 A^2 + \cdots + r_{k-1} A^{k-1} = 0$. Multiply by A^{k-1}:

$$r_0 A^{k-1} + r_1 A^k + r_2 A^{k+1} + \cdots + r_{k-1} A^{2k-2} = 0$$

Since $A^k = 0$, all the higher powers are zero, so this becomes $r_0 A^{k-1} = 0$. But $A^{k-1} \neq 0$, so $r_0 = 0$, and we have $r_1 A^1 + r_2 A^2 + \cdots + r_{k-1} A^{k-1} = 0$. Now multiply by A^{k-2} to conclude that $r_1 = 0$. Continuing, we obtain $r_i = 0$ for each i, so B is independent.

The next example collects several useful properties of independence for reference.

Example 6.3.6

Let V denote a vector space.

1. If $\mathbf{v} \neq \mathbf{0}$ in V, then $\{\mathbf{v}\}$ is an independent set.
2. No independent set of vectors in V can contain the zero vector.

Solution.

1. Let $t\mathbf{v} = \mathbf{0}$, t in $\mathbb{R}$. If $t \neq 0$, then $\mathbf{v} = 1\mathbf{v} = \frac{1}{t}(t\mathbf{v}) = \frac{1}{t}\mathbf{0} = \mathbf{0}$, contrary to assumption. So $t = 0$.
2. If $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ is independent and (say) $\mathbf{v}_2 = \mathbf{0}$, then $0\mathbf{v}_1 + 1\mathbf{v}_2 + \cdots + 0\mathbf{v}_k = \mathbf{0}$ is a nontrivial linear combination that vanishes, contrary to the independence of $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$.

A set of vectors is independent if $\mathbf{0}$ is a linear combination in a unique way. The following theorem shows that *every* linear combination of these vectors has uniquely determined coefficients, and so extends Theorem 5.2.1.

Theorem 6.3.1

Let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ be a linearly independent set of vectors in a vector space V. If a vector $\mathbf{v}$ has two (ostensibly different) representations

$$\begin{aligned}\mathbf{v} &= s_1\mathbf{v}_1 + s_2\mathbf{v}_2 + \cdots + s_n\mathbf{v}_n\\ \mathbf{v} &= t_1\mathbf{v}_1 + t_2\mathbf{v}_2 + \cdots + t_n\mathbf{v}_n\end{aligned}$$

as linear combinations of these vectors, then $s_1 = t_1, s_2 = t_2, \dots, s_n = t_n$. In other words, every vector in V can be written in a unique way as a linear combination of the $\mathbf{v}_i$.

Proof. Subtracting the equations given in the theorem gives

$$(s_1 - t_1)\mathbf{v}_1 + (s_2 - t_2)\mathbf{v}_2 + \cdots + (s_n - t_n)\mathbf{v}_n = \mathbf{0}$$

The independence of $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ gives $s_i - t_i = 0$ for each i, as required. □

The following theorem extends (and proves) Theorem 5.2.4, and is one of the most useful results in linear algebra.

Theorem 6.3.2: Fundamental Theorem

Suppose a vector space V can be spanned by n vectors. If a set of m vectors in V is linearly independent, then $m \leq n$.

Proof. Let $V = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$, and suppose that $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m\}$ is an independent set in V. Then $\mathbf{u}_1 = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n$ where each a_i is in $\mathbb{R}$. As $\mathbf{u}_1 \neq \mathbf{0}$ (Example 6.3.6), not all of the a_i are zero, say $a_1 \neq 0$ (after relabelling the $\mathbf{v}_i$). Then $V = \text{span}\{\mathbf{u}_1, \mathbf{v}_2, \mathbf{v}_3, \dots, \mathbf{v}_n\}$ as the reader can verify. Hence, write $\mathbf{u}_2 = b_1\mathbf{u}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 + \cdots + c_n\mathbf{v}_n$. Then some $c_i \neq 0$ because $\{\mathbf{u}_1, \mathbf{u}_2\}$ is independent; so, as before, $V = \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}_3, \dots, \mathbf{v}_n\}$, again after possible relabelling of the $\mathbf{v}_i$. If $m > n$, this procedure continues until all the vectors $\mathbf{v}_i$ are replaced by the vectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$. In particular, $V = \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$. But then $\mathbf{u}_{n+1}$ is a linear combination of $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ contrary to the independence of the $\mathbf{u}_i$. Hence, the assumption $m > n$ cannot be valid, so $m \leq n$ and the theorem is proved. □

If $V = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$, and if $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m\}$ is an independent set in V, the above proof shows not only that $m \leq n$ but also that m of the (spanning) vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ can be replaced by the (independent) vectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m$ and the resulting set will still span V. In this form the result is called the **Steinitz Exchange Lemma**.

Definition 6.5 Basis of a Vector Space

As in $\mathbb{R}^n$, a set $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ of vectors in a vector space V is called a **basis** *of V if it satisfies the following two conditions:*

1. $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ *is linearly independent*

2. $V = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$

Thus if a set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is a basis, then *every* vector in V can be written as a linear combination of these vectors in a *unique* way (Theorem 6.3.1). But even more is true: Any two (finite) bases of V contain the same number of vectors.

Theorem 6.3.3: Invariance Theorem

Let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ *and* $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m\}$ *be two bases of a vector space* V*. Then* $n = m$*.*

Proof. Because $V = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ and $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m\}$ is independent, it follows from Theorem 6.3.2 that $m \leq n$. Similarly $n \leq m$, so $n = m$, as asserted. □

Theorem 6.3.3 guarantees that no matter which basis of V is chosen it contains the same number of vectors as any other basis. Hence there is no ambiguity about the following definition.

Definition 6.6 Dimension of a Vector Space

If $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ *is a basis of the nonzero vector space* V*, the number* n *of vectors in the basis is called the* ***dimension*** *of* V*, and we write*

$$\dim V = n$$

The zero vector space $\{\mathbf{0}\}$ *is defined to have dimension* 0*:*

$$\dim\{\mathbf{0}\} = 0$$

In our discussion to this point we have always assumed that a basis is nonempty and hence that the dimension of the space is at least 1. However, the zero space $\{\mathbf{0}\}$ has *no* basis (by Example 6.3.6) so our insistence that $\dim\{\mathbf{0}\} = 0$ amounts to saying that the *empty* set of vectors is a basis of $\{\mathbf{0}\}$. Thus the statement that “the dimension of a vector space is the number of vectors in any basis” holds even for the zero space.

We saw in Example 5.2.10 that $\dim(\mathbb{R}^n) = n$ due to the standard basis $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ of $\mathbb{R}^n$. In Example 6.3.7 below, similar considerations apply to the space $\mathbf{M}_{mn}$ of all $m \times n$ matrices; the verifications are left to the reader.

Example 6.3.7

The space $\mathbf{M}_{mn}$ has dimension mn, and one basis consists of all $m \times n$ matrices with exactly one entry equal to 1 and all other entries equal to 0. We call this the **standard basis** of $\mathbf{M}_{mn}$.

Example 6.3.8

Show that $\dim \mathbf{P}_n = n+1$ and that $\{1,\ x,\ x^2,\ \dots,\ x^n\}$ is a basis, called the **standard basis** of $\mathbf{P}_n$.

Solution. Each polynomial $p(x) = a_0 + a_1x + \cdots + a_nx^n$ in $\mathbf{P}_n$ is clearly a linear combination of $1,\ x,\ \dots,\ x^n$, so $\mathbf{P}_n = \operatorname{span}\{1,\ x,\ \dots,\ x^n\}$. However, if a linear combination of these vectors vanishes, $a_0 1 + a_1 x + \cdots + a_n x^n = 0$, then $a_0 = a_1 = \cdots = a_n = 0$ because x is an indeterminate. So $\{1,\ x,\ \dots,\ x^n\}$ is linearly independent and hence is a basis containing $n+1$ vectors. Thus, $\dim(\mathbf{P}_n) = n+1$.

Example 6.3.9

If $\mathbf{v} \neq \mathbf{0}$ is any nonzero vector in a vector space V, show that $\operatorname{span}\{\mathbf{v}\} = \mathbb{R}\mathbf{v}$ has dimension 1.

Solution. $\{\mathbf{v}\}$ clearly spans $\mathbb{R}\mathbf{v}$, and it is linearly independent by Example 6.3.6. Hence $\{\mathbf{v}\}$ is a basis of $\mathbb{R}\mathbf{v}$, and so $\dim \mathbb{R}\mathbf{v} = 1$.

Example 6.3.10

Let $A = \left[\begin{array}{cc} 1 & 1 \\ 0 & 0 \end{array}\right]$ and consider the subspace

$$U = \{X \text{ in } \mathbf{M}_{22} \mid AX = XA\}$$

of $\mathbf{M}_{22}$. Show that $\dim U = 2$ and find a basis of U.

Solution. It was shown in Example 6.2.3 that U is a subspace for any choice of the matrix A. In the present case, if $X = \left[\begin{array}{cc} x & y \\ z & w \end{array}\right]$ is in U, the condition $AX = XA$ gives $z = 0$ and $x = y + w$. Hence each matrix X in U can be written

$$X = \left[\begin{array}{cc} y+w & y \\ 0 & w \end{array}\right] = y\left[\begin{array}{cc} 1 & 1 \\ 0 & 0 \end{array}\right] + w\left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right]$$

so $U = \operatorname{span} B$ where $B = \left\{\left[\begin{array}{cc} 1 & 1 \\ 0 & 0 \end{array}\right], \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right]\right\}$. Moreover, the set B is linearly independent (verify this), so it is a basis of U and $\dim U = 2$.

Example 6.3.11

Show that the set V of all symmetric 2×2 matrices is a vector space, and find the dimension of V.

Solution. A matrix A is symmetric if $A^T = A$. If A and B lie in V, then

$$(A+B)^T = A^T + B^T = A + B \quad \text{and} \quad (kA)^T = kA^T = kA$$

using Theorem 2.1.2. Hence $A+B$ and kA are also symmetric. As the 2×2 zero matrix is also in

V, this shows that V is a vector space (being a subspace of $\mathbf{M}_{22}$). Now a matrix A is symmetric when entries directly across the main diagonal are equal, so each 2×2 symmetric matrix has the form

$$\left[\begin{array}{cc} a & c \\ c & b \end{array}\right] = a\left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array}\right] + b\left[\begin{array}{cc} 0 & 0 \\ 0 & 1 \end{array}\right] + c\left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array}\right]$$

Hence the set $B = \left\{ \left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array}\right], \left[\begin{array}{cc} 0 & 0 \\ 0 & 1 \end{array}\right], \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array}\right] \right\}$ spans V, and the reader can verify that B is linearly independent. Thus B is a basis of V, so $\dim V = 3$.

It is frequently convenient to alter a basis by multiplying each basis vector by a nonzero scalar. The next example shows that this always produces another basis. The proof is left as Exercise **??**.

Example 6.3.12

Let $B = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ be nonzero vectors in a vector space V. Given nonzero scalars $a_1, a_2, \dots, a_n$, write $D = \{a_1\mathbf{v}_1, a_2\mathbf{v}_2, \dots, a_n\mathbf{v}_n\}$. If B is independent or spans V, the same is true of D. In particular, if B is a basis of V, so also is D.

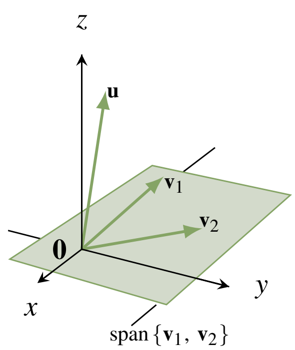
6.4 Finite Dimensional Spaces

Up to this point, we have had no guarantee that an arbitrary vector space *has* a basis—and hence no guarantee that one can speak *at all* of the dimension of V. However, Theorem 6.4.1 will show that any space that is spanned by a finite set of vectors has a (finite) basis: The proof requires the following basic lemma, of interest in itself, that gives a way to enlarge a given independent set of vectors.

Lemma 6.4.1: Independent Lemma

Let $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ be an independent set of vectors in a vector space V. If $\mathbf{u} \in V$ but[5] *$\mathbf{u} \notin \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$, then $\{\mathbf{u}, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ is also independent.*

Proof. Let $t\mathbf{u} + t_1\mathbf{v}_1 + t_2\mathbf{v}_2 + \cdots + t_k\mathbf{v}_k = \mathbf{0}$; we must show that all the coefficients are zero. First, $t = 0$ because, otherwise, $\mathbf{u} = -\frac{t_1}{t}\mathbf{v}_1 - \frac{t_2}{t}\mathbf{v}_2 - \cdots - \frac{t_k}{t}\mathbf{v}_k$ is in $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$, contrary to our assumption. Hence $t = 0$. But then $t_1\mathbf{v}_1 + t_2\mathbf{v}_2 + \cdots + t_k\mathbf{v}_k = \mathbf{0}$ so the rest of the t_i are zero by the independence of $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$. This is what we wanted. □



Note that the converse of Lemma 6.4.1 is also true: if $\{\mathbf{u}, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ is independent, then $\mathbf{u}$ is not in $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$.

As an illustration, suppose that $\{\mathbf{v}_1, \mathbf{v}_2\}$ is independent in $\mathbb{R}^3$. Then $\mathbf{v}_1$ and $\mathbf{v}_2$ are not parallel, so $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\}$ is a plane through the origin (shaded in the diagram). By Lemma 6.4.1, $\mathbf{u}$ is not in this plane if and only if $\{\mathbf{u}, \mathbf{v}_1, \mathbf{v}_2\}$ is independent.

Definition 6.7 Finite Dimensional and Infinite Dimensional Vector Spaces

*A vector space V is called **finite dimensional** if it is spanned by a finite set of vectors. Otherwise, V is called **infinite dimensional**.*

Thus the zero vector space $\{\mathbf{0}\}$ is finite dimensional because $\{\mathbf{0}\}$ is a spanning set.

Lemma 6.4.2

Let V be a finite dimensional vector space. If U is any subspace of V, then any independent subset of U can be enlarged to a finite basis of U.

Proof. Suppose that I is an independent subset of U. If $\operatorname{span} I = U$ then I is already a basis of U. If $\operatorname{span} I \neq U$, choose $\mathbf{u}_1 \in U$ such that $\mathbf{u}_1 \notin \operatorname{span} I$. Hence the set $I \cup \{\mathbf{u}_1\}$ is independent by Lemma 6.4.1. If $\operatorname{span}(I \cup \{\mathbf{u}_1\}) = U$ we are done; otherwise choose $\mathbf{u}_2 \in U$ such that $\mathbf{u}_2 \notin \operatorname{span}(I \cup \{\mathbf{u}_1\})$. Hence $I \cup \{\mathbf{u}_1, \mathbf{u}_2\}$ is independent, and the process continues. We claim that a basis of U will be reached eventually. Indeed, if no basis of U is ever reached, the process creates arbitrarily large independent sets in V. But this is impossible by the fundamental theorem because V is finite dimensional and so is spanned by a finite set of vectors. □

[5] If X is a set, we write $a \in X$ to indicate that a is an element of the set X. If a is not an element of X, we write $a \notin X$.

Theorem 6.4.1

Let V be a finite dimensional vector space spanned by m vectors.

1. *V has a finite basis, and* $\dim V \leq m$.
2. *Every independent set of vectors in V can be enlarged to a basis of V by adding vectors from any fixed basis of V.*
3. *If U is a subspace of V, then*
   a. *U is finite dimensional and* $\dim U \leq \dim V$.
   b. *If* $\dim U = \dim V$ *then* $U = V$.

Proof.

1. If $V = \{\mathbf{0}\}$, then V has an empty basis and $\dim V = 0 \leq m$. Otherwise, let $\mathbf{v} \neq \mathbf{0}$ be a vector in V. Then $\{\mathbf{v}\}$ is independent, so (1) follows from Lemma 6.4.2 with $U = V$.

2. We refine the proof of Lemma 6.4.2. Fix a basis B of V and let I be an independent subset of V. If $\operatorname{span} I = V$ then I is already a basis of V. If $\operatorname{span} I \neq V$, then B is not contained in I (because B spans V). Hence choose $\mathbf{b}_1 \in B$ such that $\mathbf{b}_1 \notin \operatorname{span} I$. Hence the set $I \cup \{\mathbf{b}_1\}$ is independent by Lemma 6.4.1. If $\operatorname{span}(I \cup \{\mathbf{b}_1\}) = V$ we are done; otherwise a similar argument shows that $(I \cup \{\mathbf{b}_1, \mathbf{b}_2\})$ is independent for some $\mathbf{b}_2 \in B$. Continue this process. As in the proof of Lemma 6.4.2, a basis of V will be reached eventually.

3. a. This is clear if $U = \{\mathbf{0}\}$. Otherwise, let $\mathbf{u} \neq \mathbf{0}$ in U. Then $\{\mathbf{u}\}$ can be enlarged to a finite basis B of U by Lemma 6.4.2, proving that U is finite dimensional. But B is independent in V, so $\dim U \leq \dim V$ by the fundamental theorem (Theorem 6.3.2).

   b. This is clear if $U = \{\mathbf{0}\}$ because V *has* a basis. Otherwise, assume $\dim V = n$. Then $\dim U = n$, so U has a basis B of n vectors. If $U \neq V$, then (by Lemma 6.4.2) B can be enlarged to a basis of V containing more than n vectors. This contradicts the invariance theorem (Theorem 6.3.3) because $\dim V = n$. So we conclude that $U = V$. □

Theorem 6.4.1 shows that a vector space V is finite dimensional if and only if it has a finite basis (possibly empty), and that every subspace of a finite dimensional space is again finite dimensional.

Example 6.4.1

Enlarge the independent set $D = \left\{ \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \right\}$ to a basis of $\mathbf{M}_{22}$.

Solution. The standard basis of $\mathbf{M}_{22}$ is $\left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}$, so including one of these in D will produce a basis by Theorem 6.4.1. In fact including *any* of these matrices in D produces an independent set (verify), and hence a basis by Theorem 6.4.4. Of course these vectors are not the only possibilities, for example, including $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ works as well.

Example 6.4.2

Find a basis of $\mathbf{P}_3$ containing the independent set $\{1+x,\ 1+x^2\}$.

Solution. The standard basis of $\mathbf{P}_3$ is $\{1,\ x,\ x^2,\ x^3\}$, so including two of these vectors will do. If we use 1 and x^3, the result is $\{1,\ 1+x,\ 1+x^2,\ x^3\}$. This is independent because the polynomials have distinct degrees (Example 6.3.4), and so is a basis by Theorem 6.4.1. Of course, including $\{1,\ x\}$ or $\{1,\ x^2\}$ would *not* work!

Example 6.4.3

Show that the space $\mathbf{P}$ of all polynomials is infinite dimensional.

Solution. For each $n \geq 1$, $\mathbf{P}$ has a subspace $\mathbf{P}_n$ of dimension $n+1$. Suppose $\mathbf{P}$ is finite dimensional, say $\dim \mathbf{P} = m$. Then $\dim \mathbf{P}_n \leq \dim \mathbf{P}$ by Theorem 6.4.1, that is $n+1 \leq m$. This is impossible since n is arbitrary, so $\mathbf{P}$ must be infinite dimensional.

The next example illustrates how (2) of Theorem 6.4.1 can be used.

Example 6.4.4

If $\mathbf{c}_1,\ \mathbf{c}_2,\ \dots,\ \mathbf{c}_k$ are independent columns in $\mathbb{R}^n$, show that they are the first k columns in some invertible $n \times n$ matrix.

Solution. By Theorem 6.4.1, expand $\{\mathbf{c}_1,\ \mathbf{c}_2,\ \dots,\ \mathbf{c}_k\}$ to a basis $\{\mathbf{c}_1,\ \mathbf{c}_2,\ \dots,\ \mathbf{c}_k,\ \mathbf{c}_{k+1},\ \dots,\ \mathbf{c}_n\}$ of $\mathbb{R}^n$. Then the matrix $A = \left[\begin{array}{cccccc} \mathbf{c}_1 & \mathbf{c}_2 & \dots & \mathbf{c}_k & \mathbf{c}_{k+1} & \dots & \mathbf{c}_n \end{array} \right]$ with this basis as its columns is an $n \times n$ matrix and it is invertible by Theorem 5.2.3.

Theorem 6.4.2

Let U and W be subspaces of the finite dimensional space V.

1. *If $U \subseteq W$, then $\dim U \leq \dim W$.*
2. *If $U \subseteq W$ and $\dim U = \dim W$, then $U = W$.*

Proof. Since W is finite dimensional, (1) follows by taking $V = W$ in part (3) of Theorem 6.4.1. Now assume $\dim U = \dim W = n$, and let B be a basis of U. Then B is an independent set in W. If $U \neq W$, then $\operatorname{span} B \neq W$, so B can be extended to an independent set of $n+1$ vectors in W by Lemma 6.4.1. This contradicts the fundamental theorem (Theorem 6.3.2) because W is spanned by $\dim W = n$ vectors. Hence $U = W$, proving (2). □

Theorem 6.4.2 is very useful. This was illustrated in Example 5.2.13 for $\mathbb{R}^2$ and $\mathbb{R}^3$; here is another example.

Example 6.4.5

If a is a number, let W denote the subspace of all polynomials in $\mathbf{P}_n$ that have a as a root:

$$W = \{p(x) \mid p(x) \in \mathbf{P}_n \text{ and } p(a) = 0\}$$

Show that $\{(x-a),\ (x-a)^2,\ \dots,\ (x-a)^n\}$ is a basis of W.

Solution. Observe first that $(x-a),\ (x-a)^2,\ \dots,\ (x-a)^n$ are members of W, and that they are independent because they have distinct degrees (Example 6.3.4). Write

$$U = \text{span}\{(x-a),\ (x-a)^2,\ \dots,\ (x-a)^n\}$$

Then we have $U \subseteq W \subseteq \mathbf{P}_n$, $\dim U = n$, and $\dim \mathbf{P}_n = n+1$. Hence $n \leq \dim W \leq n+1$ by Theorem 6.4.2. Since $\dim W$ is an integer, we must have $\dim W = n$ or $\dim W = n+1$. But then $W = U$ or $W = \mathbf{P}_n$, again by Theorem 6.4.2. Because $W \neq \mathbf{P}_n$, it follows that $W = U$, as required.

A set of vectors is called **dependent** if it is *not* independent, that is if some nontrivial linear combination vanishes. The next result is a convenient test for dependence.

Lemma 6.4.3: Dependent Lemma

A set $D = \{\mathbf{v}_1,\ \mathbf{v}_2,\ \dots,\ \mathbf{v}_k\}$ of vectors in a vector space V is dependent if and only if some vector in D is a linear combination of the others.

Proof. Let $\mathbf{v}_2$ (say) be a linear combination of the rest: $\mathbf{v}_2 = s_1\mathbf{v}_1 + s_3\mathbf{v}_3 + \cdots + s_k\mathbf{v}_k$. Then

$$s_1\mathbf{v}_1 + (-1)\mathbf{v}_2 + s_3\mathbf{v}_3 + \cdots + s_k\mathbf{v}_k = \mathbf{0}$$

is a nontrivial linear combination that vanishes, so D is dependent. Conversely, if D is dependent, let $t_1\mathbf{v}_1 + t_2\mathbf{v}_2 + \cdots + t_k\mathbf{v}_k = \mathbf{0}$ where some coefficient is nonzero. If (say) $t_2 \neq 0$, then $\mathbf{v}_2 = -\frac{t_1}{t_2}\mathbf{v}_1 - \frac{t_3}{t_2}\mathbf{v}_3 - \cdots - \frac{t_k}{t_2}\mathbf{v}_k$ is a linear combination of the others. □

Lemma 6.4.1 gives a way to enlarge independent sets to a basis; by contrast, Lemma 6.4.3 shows that spanning sets can be cut down to a basis.

Theorem 6.4.3

Let V be a finite dimensional vector space. Any spanning set for V can be cut down (by deleting vectors) to a basis of V.

Proof. Since V is finite dimensional, it has a finite spanning set S. Among all spanning sets contained in S, choose S_0 containing the smallest number of vectors. It suffices to show that S_0 is independent (then S_0 is a basis, proving the theorem). Suppose, on the contrary, that S_0 is not independent. Then, by Lemma 6.4.3, some vector $\mathbf{u} \in S_0$ is a linear combination of the set $S_1 = S_0 \setminus \{\mathbf{u}\}$ of vectors in S_0 other than $\mathbf{u}$. It follows that $\text{span}\, S_0 = \text{span}\, S_1$, that is, $V = \text{span}\, S_1$. But S_1 has fewer elements than S_0 so this contradicts the choice of S_0. Hence S_0 is independent after all. □

Note that, with Theorem 6.4.1, Theorem 6.4.3 completes the promised proof of Theorem 5.2.6 for the case $V = \mathbb{R}^n$.

Example 6.4.6

Find a basis of $\mathbf{P}_3$ in the spanning set $S = \{1,\ x+x^2,\ 2x-3x^2,\ 1+3x-2x^2,\ x^3\}$.

Solution. Since $\dim \mathbf{P}_3 = 4$, we must eliminate one polynomial from S. It cannot be x^3 because the span of the rest of S is contained in $\mathbf{P}_2$. But eliminating $1+3x-2x^2$ does leave a basis (verify). Note that $1+3x-2x^2$ is the sum of the first three polynomials in S.

Theorems 6.4.1 and 6.4.3 have other useful consequences.

Theorem 6.4.4

Let V be a vector space with $\dim V = n$, and suppose S is a set of exactly n vectors in V. Then S is independent if and only if S spans V.

Proof. Assume first that S is independent. By Theorem 6.4.1, S is contained in a basis B of V. Hence $|S| = n = |B|$ so, since $S \subseteq B$, it follows that $S = B$. In particular S spans V.

Conversely, assume that S spans V, so S contains a basis B by Theorem 6.4.3. Again $|S| = n = |B|$ so, since $S \supseteq B$, it follows that $S = B$. Hence S is independent. □

One of independence or spanning is often easier to establish than the other when showing that a set of vectors is a basis. For example if $V = \mathbb{R}^n$ it is easy to check whether a subset S of $\mathbb{R}^n$ is orthogonal (hence independent) but checking spanning can be tedious. Here are three more examples.

Example 6.4.7

Consider the set $S = \{p_0(x),\ p_1(x),\ \dots,\ p_n(x)\}$ of polynomials in $\mathbf{P}_n$. If $\deg p_k(x) = k$ for each k, show that S is a basis of $\mathbf{P}_n$.

Solution. The set S is independent—the degrees are distinct—see Example 6.3.4. Hence S is a basis of $\mathbf{P}_n$ by Theorem 6.4.4 because $\dim \mathbf{P}_n = n+1$.

Example 6.4.8

Let V denote the space of all symmetric 2×2 matrices. Find a basis of V consisting of invertible matrices.

Solution. We know that $\dim V = 3$ (Example 6.3.11), so what is needed is a set of three invertible, symmetric matrices that (using Theorem 6.4.4) is either independent or spans V. The set $\left\{ \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right], \left[\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array} \right], \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \right] \right\}$ is independent (verify) and so is a basis of the required type.

Example 6.4.9

Let A be any $n \times n$ matrix. Show that there exist n^2+1 scalars $a_0, a_1, a_2, \dots, a_{n^2}$ not all zero, such that

$$a_0 I + a_1 A + a_2 A^2 + \cdots + a_{n^2} A^{n^2} = 0$$

where I denotes the $n \times n$ identity matrix.

Solution. The space $\mathbf{M}_{nn}$ of all $n \times n$ matrices has dimension n^2 by Example 6.3.7. Hence the n^2+1 matrices $I, A, A^2, \dots, A^{n^2}$ cannot be independent by Theorem 6.4.4, so a nontrivial linear combination vanishes. This is the desired conclusion.

The result in Example 6.4.9 can be written as $f(A) = 0$ where $f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{n^2} x^{n^2}$. In other words, A satisfies a nonzero polynomial $f(x)$ of degree at most n^2. In fact we know that A satisfies a nonzero polynomial of degree n (this is the Cayley-Hamilton theorem—see Theorem 8.7.10), but the brevity of the solution in Example 6.4.6 is an indication of the power of these methods.

If U and W are subspaces of a vector space V, there are two related subspaces that are of interest, their **sum** $U + W$ and their **intersection** $U \cap W$, defined by

$$U + W = \{\mathbf{u} + \mathbf{w} \mid \mathbf{u} \in U \text{ and } \mathbf{w} \in W\}$$
$$U \cap W = \{\mathbf{v} \in V \mid \mathbf{v} \in U \text{ and } \mathbf{v} \in W\}$$

It is routine to verify that these are indeed subspaces of V, that $U \cap W$ is contained in both U and W, and that $U + W$ contains both U and W. We conclude this section with a useful fact about the dimensions of these spaces. The proof is a good illustration of how the theorems in this section are used.

Theorem 6.4.5

Suppose that U and W are finite dimensional subspaces of a vector space V. Then $U + W$ is finite dimensional and

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

Proof. Since $U \cap W \subseteq U$, it has a finite basis, say $\{\mathbf{x}_1, \dots, \mathbf{x}_d\}$. Extend it to a basis $\{\mathbf{x}_1, \dots, \mathbf{x}_d, \mathbf{u}_1, \dots, \mathbf{u}_m\}$ of U by Theorem 6.4.1. Similarly extend $\{\mathbf{x}_1, \dots, \mathbf{x}_d\}$ to a basis $\{\mathbf{x}_1, \dots, \mathbf{x}_d, \mathbf{w}_1, \dots, \mathbf{w}_p\}$ of W. Then

$$U + W = \operatorname{span}\{\mathbf{x}_1, \dots, \mathbf{x}_d, \mathbf{u}_1, \dots, \mathbf{u}_m, \mathbf{w}_1, \dots, \mathbf{w}_p\}$$

as the reader can verify, so $U + W$ is finite dimensional. For the rest, it suffices to show that $\{\mathbf{x}_1, \dots, \mathbf{x}_d, \mathbf{u}_1, \dots, \mathbf{u}_m, \mathbf{w}_1, \dots, \mathbf{w}_p\}$ is independent (verify). Suppose that

$$r_1\mathbf{x}_1 + \cdots + r_d\mathbf{x}_d + s_1\mathbf{u}_1 + \cdots + s_m\mathbf{u}_m + t_1\mathbf{w}_1 + \cdots + t_p\mathbf{w}_p = \mathbf{0} \tag{6.1}$$

where the r_i, s_j, and t_k are scalars. Then

$$r_1\mathbf{x}_1 + \cdots + r_d\mathbf{x}_d + s_1\mathbf{u}_1 + \cdots + s_m\mathbf{u}_m = -(t_1\mathbf{w}_1 + \cdots + t_p\mathbf{w}_p)$$

is in U (left side) and also in W (right side), and so is in $U \cap W$. Hence $(t_1\mathbf{w}_1 + \cdots + t_p\mathbf{w}_p)$ is a linear combination of $\{\mathbf{x}_1, \dots, \mathbf{x}_d\}$, so $t_1 = \cdots = t_p = 0$, because $\{\mathbf{x}_1, \dots, \mathbf{x}_d, \mathbf{w}_1, \dots, \mathbf{w}_p\}$ is independent.

Similarly, $s_1 = \cdots = s_m = 0$, so (6.1) becomes $r_1\mathbf{x}_1 + \cdots + r_d\mathbf{x}_d = \mathbf{0}$. It follows that $r_1 = \cdots = r_d = 0$, as required. □

Theorem 6.4.5 is particularly interesting if $U \cap W = \{\mathbf{0}\}$. Then there are *no* vectors $\mathbf{x}_i$ in the above proof, and the argument shows that if $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ and $\{\mathbf{w}_1, \dots, \mathbf{w}_p\}$ are bases of U and W respectively, then $\{\mathbf{u}_1, \dots, \mathbf{u}_m, \mathbf{w}_1, \dots, \mathbf{w}_p\}$ is a basis of $U + W$. In this case $U + W$ is said to be a **direct sum** (written $U \oplus W$); we return to this in Chapter 9.

6.5 An Application to Polynomials

The vector space of all polynomials of degree at most n is denoted $\mathbf{P}_n$, and it was established in Section 6.3 that $\mathbf{P}_n$ has dimension $n+1$; in fact, $\{1, x, x^2, \dots, x^n\}$ is a basis. More generally, *any* $n+1$ polynomials of distinct degrees form a basis, by Theorem 6.4.4 (they are independent by Example 6.3.4). This proves

Theorem 6.5.1

Let $p_0(x), p_1(x), p_2(x), \dots, p_n(x)$ be polynomials in $\mathbf{P}_n$ of degrees $0, 1, 2, \dots, n$, respectively. Then $\{p_0(x), \dots, p_n(x)\}$ is a basis of $\mathbf{P}_n$.

An immediate consequence is that $\{1, (x-a), (x-a)^2, \dots, (x-a)^n\}$ is a basis of $\mathbf{P}_n$ for any number a. Hence we have the following:

Corollary 6.5.1

If a is any number, every polynomial $f(x)$ of degree at most n has an expansion in powers of $(x-a)$:

$$f(x) = a_0 + a_1(x-a) + a_2(x-a)^2 + \cdots + a_n(x-a)^n \tag{6.2}$$

If $f(x)$ is evaluated at $x = a$, then equation (6.2) becomes

$$f(a) = a_0 + a_1(a-a) + \cdots + a_n(a-a)^n = a_0$$

Hence $a_0 = f(a)$, and equation (6.2) can be written $f(x) = f(a) + (x-a)g(x)$, where $g(x)$ is a polynomial of degree $n-1$ (this assumes that $n \geq 1$). If it happens that $f(a) = 0$, then it is clear that $f(x)$ has the form $f(x) = (x-a)g(x)$. Conversely, every such polynomial certainly satisfies $f(a) = 0$, and we obtain:

Corollary 6.5.2

Let $f(x)$ be a polynomial of degree $n \geq 1$ and let a be any number. Then:

Remainder Theorem

1. *$f(x) = f(a) + (x-a)g(x)$ for some polynomial $g(x)$ of degree $n-1$.*

Factor Theorem

2. *$f(a) = 0$ if and only if $f(x) = (x-a)g(x)$ for some polynomial $g(x)$.*

The polynomial $g(x)$ can be computed easily by using "long division" to divide $f(x)$ by $(x-a)$—see Appendix D.

All the coefficients in the expansion (6.2) of $f(x)$ in powers of $(x-a)$ can be determined in terms of the derivatives of $f(x)$.[6] These will be familiar to students of calculus. Let $f^{(n)}(x)$ denote the nth derivative of the polynomial $f(x)$, and write $f^{(0)}(x) = f(x)$. Then, if

$$f(x) = a_0 + a_1(x-a) + a_2(x-a)^2 + \cdots + a_n(x-a)^n$$

it is clear that $a_0 = f(a) = f^{(0)}(a)$. Differentiation gives

$$f^{(1)}(x) = a_1 + 2a_2(x-a) + 3a_3(x-a)^2 + \cdots + na_n(x-a)^{n-1}$$

and substituting $x = a$ yields $a_1 = f^{(1)}(a)$. This continues to give $a_2 = \frac{f^{(2)}(a)}{2!}$, $a_3 = \frac{f^{(3)}(a)}{3!}$, $\dots$, $a_k = \frac{f^{(k)}(a)}{k!}$, where $k!$ is defined as $k! = k(k-1)\cdots 2 \cdot 1$. Hence we obtain the following:

Corollary 6.5.3: Taylor's Theorem

If $f(x)$ is a polynomial of degree n, then

$$f(x) = f(a) + \frac{f^{(1)}(a)}{1!}(x-a) + \frac{f^{(2)}(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

[6] The discussion of Taylor's theorem can be omitted with no loss of continuity.

Example 6.5.1

Expand $f(x) = 5x^3 + 10x + 2$ as a polynomial in powers of $x - 1$.

Solution. The derivatives are $f^{(1)}(x) = 15x^2 + 10$, $f^{(2)}(x) = 30x$, and $f^{(3)}(x) = 30$. Hence the Taylor expansion is

$$\begin{aligned} f(x) &= f(1) + \tfrac{f^{(1)}(1)}{1!}(x-1) + \tfrac{f^{(2)}(1)}{2!}(x-1)^2 + \tfrac{f^{(3)}(1)}{3!}(x-1)^3 \\ &= 17 + 25(x-1) + 15(x-1)^2 + 5(x-1)^3 \end{aligned}$$

Taylor's theorem is useful in that it provides a formula for the coefficients in the expansion. It is dealt with in calculus texts and will not be pursued here.

Theorem 6.5.1 produces bases of $\mathbf{P}_n$ consisting of polynomials of distinct degrees. A different criterion is involved in the next theorem.

Theorem 6.5.2

Let $f_0(x),\ f_1(x),\ \dots,\ f_n(x)$ be nonzero polynomials in $\mathbf{P}_n$. Assume that numbers $a_0,\ a_1,\ \dots,\ a_n$ exist such that

$$\begin{aligned} f_i(a_i) &\neq 0 \quad \text{for each } i \\ f_i(a_j) &= 0 \quad \text{if } i \neq j \end{aligned}$$

Then

1. $\{f_0(x),\ \dots,\ f_n(x)\}$ *is a basis of* $\mathbf{P}_n$.
2. *If $f(x)$ is any polynomial in $\mathbf{P}_n$, its expansion as a linear combination of these basis vectors is*

$$f(x) = \tfrac{f(a_0)}{f_0(a_0)} f_0(x) + \tfrac{f(a_1)}{f_1(a_1)} f_1(x) + \cdots + \tfrac{f(a_n)}{f_n(a_n)} f_n(x)$$

Proof.

1. It suffices (by Theorem 6.4.4) to show that $\{f_0(x),\ \dots,\ f_n(x)\}$ is linearly independent (because $\dim \mathbf{P}_n = n+1$). Suppose that

$$r_0 f_0(x) + r_1 f_1(x) + \cdots + r_n f_n(x) = 0,\ r_i \in \mathbb{R}$$

Because $f_i(a_0) = 0$ for all $i > 0$, taking $x = a_0$ gives $r_0 f_0(a_0) = 0$. But then $r_0 = 0$ because $f_0(a_0) \neq 0$. The proof that $r_i = 0$ for $i > 0$ is analogous.

2. By (1), $f(x) = r_0 f_0(x) + \cdots + r_n f_n(x)$ for *some* numbers r_i. Once again, evaluating at a_0 gives $f(a_0) = r_0 f_0(a_0)$, so $r_0 = f(a_0)/f_0(a_0)$. Similarly, $r_i = f(a_i)/f_i(a_i)$ for each i. □

Example 6.5.2

Show that $\{x^2 - x,\ x^2 - 2x,\ x^2 - 3x + 2\}$ is a basis of $\mathbf{P}_2$.

Solution. Write $f_0(x) = x^2 - x = x(x-1)$, $f_1(x) = x^2 - 2x = x(x-2)$, and $f_2(x) = x^2 - 3x + 2 = (x-1)(x-2)$. Then the conditions of Theorem 6.5.2 are satisfied with $a_0 = 2$, $a_1 = 1$, and $a_2 = 0$.

We investigate one natural choice of the polynomials $f_i(x)$ in Theorem 6.5.2. To illustrate, let a_0, a_1, and a_2 be distinct numbers and write

$$f_0(x) = \frac{(x-a_1)(x-a_2)}{(a_0-a_1)(a_0-a_2)} \quad f_1(x) = \frac{(x-a_0)(x-a_2)}{(a_1-a_0)(a_1-a_2)} \quad f_2(x) = \frac{(x-a_0)(x-a_1)}{(a_2-a_0)(a_2-a_1)}$$

Then $f_0(a_0) = f_1(a_1) = f_2(a_2) = 1$, and $f_i(a_j) = 0$ for $i \neq j$. Hence Theorem 6.5.2 applies, and because $f_i(a_i) = 1$ for each i, the formula for expanding any polynomial is simplified.

In fact, this can be generalized with no extra effort. If $a_0, a_1, \dots, a_n$ are distinct numbers, define the **Lagrange polynomials** $\delta_0(x), \delta_1(x), \dots, \delta_n(x)$ relative to these numbers as follows:

$$\delta_k(x) = \frac{\prod_{i \neq k}(x-a_i)}{\prod_{i \neq k}(a_k-a_i)} \quad k = 0, 1, 2, \dots, n$$

Here the numerator is the product of all the terms $(x-a_0), (x-a_1), \dots, (x-a_n)$ with $(x-a_k)$ omitted, and a similar remark applies to the denominator. If $n = 2$, these are just the polynomials in the preceding paragraph. For another example, if $n = 3$, the polynomial $\delta_1(x)$ takes the form

$$\delta_1(x) = \frac{(x-a_0)(x-a_2)(x-a_3)}{(a_1-a_0)(a_1-a_2)(a_1-a_3)}$$

In the general case, it is clear that $\delta_i(a_i) = 1$ for each i and that $\delta_i(a_j) = 0$ if $i \neq j$. Hence Theorem 6.5.2 specializes as Theorem 6.5.3.

Theorem 6.5.3: Lagrange Interpolation Expansion

Let $a_0, a_1, \dots, a_n$ be distinct numbers. The corresponding set

$$\{\delta_0(x), \delta_1(x), \dots, \delta_n(x)\}$$

of Lagrange polynomials is a basis of $\mathbf{P}_n$, and any polynomial $f(x)$ in $\mathbf{P}_n$ has the following unique expansion as a linear combination of these polynomials.

$$f(x) = f(a_0)\delta_0(x) + f(a_1)\delta_1(x) + \cdots + f(a_n)\delta_n(x)$$

Example 6.5.3

Find the Lagrange interpolation expansion for $f(x) = x^2 - 2x + 1$ relative to $a_0 = -1$, $a_1 = 0$, and $a_2 = 1$.

Solution. The Lagrange polynomials are

$$\begin{aligned}
\delta_0 &= \frac{(x-0)(x-1)}{(-1-0)(-1-1)} = \frac{1}{2}(x^2 - x) \\
\delta_1 &= \frac{(x+1)(x-1)}{(0+1)(0-1)} = -(x^2 - 1) \\
\delta_2 &= \frac{(x+1)(x-0)}{(1+1)(1-0)} = \frac{1}{2}(x^2 + x)
\end{aligned}$$

Because $f(-1) = 4$, $f(0) = 1$, and $f(1) = 0$, the expansion is

$$f(x) = 2(x^2 - x) - (x^2 - 1)$$

The Lagrange interpolation expansion gives an easy proof of the following important fact.

Theorem 6.5.4

Let $f(x)$ be a polynomial in $\mathbf{P}_n$, and let $a_0, a_1, \dots, a_n$ denote distinct numbers. If $f(a_i) = 0$ for all i, then $f(x)$ is the zero polynomial (that is, all coefficients are zero).

Proof. All the coefficients in the Lagrange expansion of $f(x)$ are zero. □

6.6 An Application to Differential Equations

Call a function $f : \mathbb{R} \to \mathbb{R}$ **differentiable** if it can be differentiated as many times as we want. If f is a differentiable function, the nth derivative $f^{(n)}$ of f is the result of differentiating n times. Thus $f^{(0)} = f$, $f^{(1)} = f'$, $f^{(2)} = f^{(1)\prime}$, $\dots$ and, in general, $f^{(n+1)} = f^{(n)\prime}$ for each $n \geq 0$. For small values of n these are often written as $f, f', f'', f''', \dots$.

If a, b, and c are numbers, the differential equations

$$f'' + af' + bf = 0 \quad \text{or} \quad f''' + af'' + bf' + cf = 0$$

are said to be of **second-order** and **third-order**, respectively. In general, an equation

$$f^{(n)} + a_{n-1}f^{(n-1)} + a_{n-2}f^{(n-2)} + \cdots + a_2f^{(2)} + a_1f^{(1)} + a_0f^{(0)} = 0, \quad a_i \text{ in } \mathbb{R} \tag{6.3}$$

is called a **differential equation of order** n. In this section we investigate the set of solutions to (6.3) and, if n is 1 or 2, find explicit solutions. Of course an acquaintance with calculus is required.

Let f and g be solutions to (6.3). Then $f+g$ is also a solution because $(f+g)^{(k)} = f^{(k)} + g^{(k)}$ for all k, and af is a solution for any a in $\mathbb{R}$ because $(af)^{(k)} = af^{(k)}$. It follows that the set of solutions to (6.3) is a vector space, and we ask for the dimension of this space.

We have already dealt with the simplest case (see Theorem 3.7.1):

Theorem 6.6.1

The set of solutions of the first-order differential equation $f' + af = 0$ is a one-dimensional vector space and $\{e^{-ax}\}$ is a basis.

There is a far-reaching generalization of Theorem 6.6.1 that will be proved in Theorem 7.4.1.

Theorem 6.6.2

The set of solutions to the nth order equation (6.3) has dimension n.

Remark

Every differential equation of order n can be converted into a system of n linear first-order equations (see Exercises **??** and **??**). In the case that the matrix of this system is diagonalizable, this approach provides a proof of Theorem 6.6.2. But if the matrix is not diagonalizable, Theorem 7.4.1 is required.

Theorem 6.6.1 suggests that we look for solutions to (6.3) of the form $e^{\lambda x}$ for some number λ. This is a good idea. If we write $f(x) = e^{\lambda x}$, it is easy to verify that $f^{(k)}(x) = \lambda^k e^{\lambda x}$ for each $k \geq 0$, so substituting f in (6.3) gives

$$(\lambda^n + a_{n-1}\lambda^{n-1} + a_{n-2}\lambda^{n-2} + \cdots + a_2\lambda^2 + a_1\lambda^1 + a_0)e^{\lambda x} = 0$$

Since $e^{\lambda x} \neq 0$ for all x, this shows that $e^{\lambda x}$ is a solution of (6.3) if and only if λ is a root of the **characteristic polynomial** $c(x)$, defined to be

$$c(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_2x^2 + a_1x + a_0$$

This proves Theorem 6.6.3.

Theorem 6.6.3

If λ is real, the function $e^{\lambda x}$ is a solution of (6.3) if and only if λ is a root of the characteristic polynomial $c(x)$.

Example 6.6.1

Find a basis of the space U of solutions of $f''' - 2f'' - f' - 2f = 0$.

Solution. The characteristic polynomial is $x^3 - 2x^2 - x - 1 = (x-1)(x+1)(x-2)$, with roots $\lambda_1 = 1$, $\lambda_2 = -1$, and $\lambda_3 = 2$. Hence e^x, e^{-x}, and e^{2x} are all in U. Moreover they are independent (by Lemma 6.6.1 below) so, since $\dim(U) = 3$ by Theorem 6.6.2, $\{e^x, e^{-x}, e^{2x}\}$ is a basis of U.

Lemma 6.6.1

If $\lambda_1, \lambda_2, \dots, \lambda_k$ are distinct, then $\{e^{\lambda_1 x}, e^{\lambda_2 x}, \dots, e^{\lambda_k x}\}$ is linearly independent.

Proof. Suppose that $a_1 e^{\lambda_1 x} + a_2 e^{\lambda_2 x} + \cdots + a_k e^{\lambda_k x} = 0$ for all x. By repeatedly differentiating this equation $k-1$ times, we obtain the following k equations:

$$\begin{array}{ccccccc}
a_1 e^{\lambda_1 x} & + & a_2 e^{\lambda_2 x} & + \cdots + & a_k e^{\lambda_k x} & = 0 \\
a_1 \lambda_1 e^{\lambda_1 x} & + & a_2 \lambda_2 e^{\lambda_2 x} & + \cdots + & a_k \lambda_k e^{\lambda_k x} & = 0 \\
a_1 \lambda_1^2 e^{\lambda_1 x} & + & a_2 \lambda_2^2 e^{\lambda_2 x} & + \cdots + & a_k \lambda_k^2 e^{\lambda_k x} & = 0 \\
\cdots & & & & & \\
\multicolumn{6}{r}{a_1 \lambda_1^{(k-1)} e^{\lambda_1 x} + a_2 \lambda_2^{(k-1)} e^{\lambda_2 x} + \cdots + a_k \lambda_k^{(k-1)} e^{\lambda_k x} = 0}
\end{array}$$

These can be written as the following matrix equation.

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ \lambda_1 & \lambda_2 & \dots & \lambda_k \\ \lambda_1^2 & \lambda_2^2 & \dots & \lambda_k^2 \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_1^{(k-1)} & \lambda_2^{(k-1)} & \dots & \lambda_k^{(k-1)} \end{bmatrix} \begin{bmatrix} a_1 e^{\lambda_1 x} \\ a_2 e^{\lambda_2 x} \\ \vdots \\ a_k e^{\lambda_k x} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

The matrix on the left is a Vandermonde matrix, and by Theorem 3.2.7 its determinant is given by $\prod_{1 \leq j < i \leq k} (\lambda_i - \lambda_j)$ and hence nonzero as the λ_i's are assumed to all be distinct. Hence that matrix is invertible and it follows that $a_i e^{\lambda_i x} = 0$ for each i and any value of x. Using $x = 0$ we conclude that $a_i = 0$ for each i and the proof is complete. □

Theorem 6.6.4

Let U denote the space of solutions to the second-order equation

$$f'' + af' + bf = 0$$

where a and b are real constants. Assume that the characteristic polynomial x^2+ax+b has two real roots λ and μ. Then

1. *If $\lambda \neq \mu$, then $\{e^{\lambda x},\ e^{\mu x}\}$ is a basis of U.*

2. *If $\lambda = \mu$, then $\{e^{\lambda x},\ xe^{\lambda x}\}$ is a basis of U.*

Proof. Since $\dim(U) = 2$ by Theorem 6.6.2, (1) follows by Lemma 6.6.1, and (2) follows because the set $\{e^{\lambda x},\ xe^{\lambda x}\}$ is independent (Exercise **??**). $\square$

Example 6.6.2

Find the solution of $f'' + 4f' + 4f = 0$ that satisfies the **boundary conditions** $f(0) = 1$, $f(1) = -1$.

Solution. The characteristic polynomial is $x^2+4x+4 = (x+2)^2$, so -2 is a double root. Hence $\{e^{-2x},\ xe^{-2x}\}$ is a basis for the space of solutions, and the general solution takes the form $f(x) = ce^{-2x} + dxe^{-2x}$. Applying the boundary conditions gives $1 = f(0) = c$ and $-1 = f(1) = (c+d)e^{-2}$. Hence $c = 1$ and $d = -(1+e^2)$, so the required solution is

$$f(x) = e^{-2x} - (1+e^2)xe^{-2x}$$

One other question remains: What happens if the roots of the characteristic polynomial are not real? To answer this, we must first state precisely what $e^{\lambda x}$ means when λ is not real. If q is a real number, define

$$e^{iq} = \cos q + i \sin q$$

where $i^2 = -1$. Then the relationship $e^{iq}e^{iq_1} = e^{i(q+q_1)}$ holds for all real q and q_1, as is easily verified. If $\lambda = p + iq$, where p and q are real numbers, we define

$$e^{\lambda} = e^p e^{iq} = e^p(\cos q + i \sin q)$$

Then it is a routine exercise to show that

1. $e^{\lambda}e^{\mu} = e^{\lambda+\mu}$

2. $e^{\lambda} = 1$ if and only if $\lambda = 0$

3. $(e^{\lambda x})' = \lambda e^{\lambda x}$

These easily imply that $f(x) = e^{\lambda x}$ is a solution to $f'' + af' + bf = 0$ if λ is a (possibly complex) root of the characteristic polynomial x^2+ax+b. Now write $\lambda = p + iq$ so that

$$f(x) = e^{\lambda x} = e^{px}\cos(qx) + ie^{px}\sin(qx)$$

For convenience, denote the real and imaginary parts of $f(x)$ as $u(x) = e^{px}\cos(qx)$ and $v(x) = e^{px}\sin(qx)$. Then the fact that $f(x)$ satisfies the differential equation gives

$$0 = f'' + af' + bf = (u'' + au' + bu) + i(v'' + av' + bv)$$

Equating real and imaginary parts shows that $u(x)$ and $v(x)$ are both solutions to the differential equation. This proves part of Theorem 6.6.5.

Theorem 6.6.5

Let U denote the space of solutions of the second-order differential equation

$$f'' + af' + bf = 0$$

where a and b are real. Suppose λ is a nonreal root of the characteristic polynomial $x^2 + ax + b$. If $\lambda = p + iq$, where p and q are real, then

$$\{e^{px}\cos(qx),\ e^{px}\sin(qx)\}$$

is a basis of U.

Proof. The foregoing discussion shows that these functions lie in U. Because $\dim U = 2$ by Theorem 6.6.2, it suffices to show that they are linearly independent. But if

$$re^{px}\cos(qx) + se^{px}\sin(qx) = 0$$

for all x, then $r\cos(qx) + s\sin(qx) = 0$ for all x (because $e^{px} \neq 0$). Taking $x = 0$ gives $r = 0$, and taking $x = \frac{\pi}{2q}$ gives $s = 0$ ($q \neq 0$ because λ is not real). This is what we wanted. □

Example 6.6.3

Find the solution $f(x)$ to $f'' - 2f' + 2f = 0$ that satisfies $f(0) = 2$ and $f(\frac{\pi}{2}) = 0$.

Solution. The characteristic polynomial $x^2 - 2x + 2$ has roots $1 + i$ and $1 - i$. Taking $\lambda = 1 + i$ (quite arbitrarily) gives $p = q = 1$ in the notation of Theorem 6.6.5, so $\{e^x\cos x,\ e^x\sin x\}$ is a basis for the space of solutions. The general solution is thus $f(x) = e^x(r\cos x + s\sin x)$. The boundary conditions yield $2 = f(0) = r$ and $0 = f(\frac{\pi}{2}) = e^{\pi/2}s$. Thus $r = 2$ and $s = 0$, and the required solution is $f(x) = 2e^x\cos x$.

The following theorem is an important special case of Theorem 6.6.5.

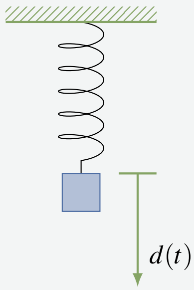
Theorem 6.6.6

If $q \neq 0$ is a real number, the space of solutions to the differential equation $f'' + q^2 f = 0$ has basis $\{\cos(qx),\ \sin(qx)\}$.

Proof. The characteristic polynomial $x^2 + q^2$ has roots qi and $-qi$, so Theorem 6.6.5 applies with $p = 0$. □

In many situations, the displacement $s(t)$ of some object at time t turns out to have an oscillating form $s(t) = c\sin(at) + d\cos(at)$. These are called **simple harmonic motions**. An example follows.

Example 6.6.4



A weight is attached to an extension spring (see diagram). If it is pulled from the equilibrium position and released, it is observed to oscillate up and down. Let $d(t)$ denote the distance of the weight below the equilibrium position t seconds later. It is known (**Hooke's law**) that the acceleration $d''(t)$ of the weight is proportional to the displacement $d(t)$ and in the opposite direction. That is,

$$d''(t) = -kd(t)$$

where $k > 0$ is called the **spring constant**. Find $d(t)$ if the maximum extension is 10 cm below the equilibrium position and find the **period** of the oscillation (time taken for the weight to make a full oscillation).

Solution. It follows from Theorem 6.6.6 (with $q^2 = k$) that

$$d(t) = r\sin(\sqrt{k}\,t) + s\cos(\sqrt{k}\,t)$$

where r and s are constants. The condition $d(0) = 0$ gives $s = 0$, so $d(t) = r\sin(\sqrt{k}\,t)$. Now the maximum value of the function $\sin x$ is 1 (when $x = \frac{\pi}{2}$), so $r = 10$ (when $t = \frac{\pi}{2\sqrt{k}}$). Hence

$$d(t) = 10\sin(\sqrt{k}\,t)$$

Finally, the weight goes through a full oscillation as $\sqrt{k}\,t$ increases from 0 to 2π. The time taken is $t = \frac{2\pi}{\sqrt{k}}$, the period of the oscillation.

Chapter 7

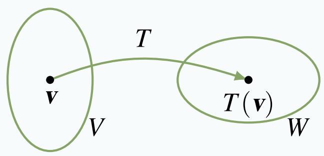
Linear Transformations

If V and W are vector spaces, a function $T : V \to W$ is a rule that assigns to each vector $\mathbf{v}$ in V a uniquely determined vector $T(\mathbf{v})$ in W. As mentioned in Section 2.2, two functions $S : V \to W$ and $T : V \to W$ are equal if $S(\mathbf{v}) = T(\mathbf{v})$ for every $\mathbf{v}$ in V. A function $T : V \to W$ is called a *linear transformation* if $T(\mathbf{v} + \mathbf{v}_1) = T(\mathbf{v}) + T(\mathbf{v}_1)$ for all $\mathbf{v}$, $\mathbf{v}_1$ in V and $T(r\mathbf{v}) = rT(\mathbf{v})$ for all $\mathbf{v}$ in V and all scalars r. $T(\mathbf{v})$ is called the *image* of $\mathbf{v}$ under T. We have already studied linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ and shown (in Section 2.6) that they are all given by multiplication by a uniquely determined $m \times n$ matrix A; that is $T(\mathbf{x}) = A\mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^n$. In the case of linear operators $\mathbb{R}^2 \to \mathbb{R}^2$, this yields an important way to describe geometric functions such as rotations about the origin and reflections in a line through the origin.

In the present chapter we will describe linear transformations in general, introduce the *kernel* and *image* of a linear transformation, and prove a useful result (called the *dimension theorem*) that relates the dimensions of the kernel and image, and unifies and extends several earlier results. Finally we study the notion of *isomorphic* vector spaces, that is, spaces that are identical except for notation, and relate this to composition of transformations that was introduced in Section 2.3.

7.1 Examples and Elementary Properties

Definition 7.1 Linear Transformations of Vector Spaces



If V and W are two vector spaces, a function $T : V \to W$ is called a ***linear transformation*** *if it satisfies the following axioms.*

T1.	$T(\mathbf{v}+\mathbf{v}_1) = T(\mathbf{v}) + T(\mathbf{v}_1)$	*for all* $\mathbf{v}$ *and* $\mathbf{v}_1$ *in* V.
T2.	$T(r\mathbf{v}) = rT(\mathbf{v})$	*for all* $\mathbf{v}$ *in* V *and* r *in* $\mathbb{R}$.

A linear transformation $T : V \to V$ is called a ***linear operator*** *on V. The situation can be visualized as in the diagram.*

Axiom T1 is just the requirement that T *preserves* vector addition. It asserts that the result $T(\mathbf{v}+\mathbf{v}_1)$ of adding $\mathbf{v}$ and $\mathbf{v}_1$ first and then applying T is the same as applying T first to get $T(\mathbf{v})$ and $T(\mathbf{v}_1)$ and then adding. Similarly, axiom T2 means that T *preserves* scalar multiplication. Note that, even though the additions in axiom T1 are both denoted by the same symbol $+$, the addition on the left forming $\mathbf{v}+\mathbf{v}_1$ is carried out in V, whereas the addition $T(\mathbf{v})+T(\mathbf{v}_1)$ is done in W. Similarly, the scalar multiplications $r\mathbf{v}$ and $rT(\mathbf{v})$ in axiom T2 refer to the spaces V and W, respectively.

We have already seen many examples of linear transformations $T : \mathbb{R}^n \to \mathbb{R}^m$. In fact, writing vectors in $\mathbb{R}^n$ as columns, Theorem 2.6.2 shows that, for each such T, there is an $m \times n$ matrix A such that $T(\mathbf{x}) = A\mathbf{x}$ for every $\mathbf{x}$ in $\mathbb{R}^n$. Moreover, the matrix A is given by $A = \left[\begin{array}{cccc} T(\mathbf{e}_1) & T(\mathbf{e}_2) & \cdots & T(\mathbf{e}_n) \end{array}\right]$ where $\{\mathbf{e}_1,\ \mathbf{e}_2,\ \dots,\ \mathbf{e}_n\}$ is the standard basis of $\mathbb{R}^n$. We denote this transformation by $T_A : \mathbb{R}^n \to \mathbb{R}^m$, defined by

$$T_A(\mathbf{x}) = A\mathbf{x} \quad \text{for all } \mathbf{x} \text{ in } \mathbb{R}^n$$

Example 7.1.1 lists three important linear transformations that will be referred to later. The verification of axioms T1 and T2 is left to the reader.

Example 7.1.1

If V and W are vector spaces, the following are linear transformations:

Identity operator $V \to V$	$1_V : V \to V$	where $1_V(\mathbf{v}) = \mathbf{v}$ for all $\mathbf{v}$ in V
Zero transformation $V \to W$	$0 : V \to W$	where $0(\mathbf{v}) = \mathbf{0}$ for all $\mathbf{v}$ in V
Scalar operator $V \to V$	$a : V \to V$	where $a(\mathbf{v}) = a\mathbf{v}$ for all $\mathbf{v}$ in V
		(Here a is any real number.)

The symbol 0 will be used to denote the zero transformation from V to W for *any* spaces V and W. It was also used earlier to denote the zero function $[a,\ b] \to \mathbb{R}$.

The next example gives two important transformations of matrices. Recall that the trace $\operatorname{tr} A$ of an $n \times n$ matrix A is the sum of the entries on the main diagonal.

Example 7.1.2

Show that the transposition and trace are linear transformations. More precisely,

$$\begin{aligned} R : \mathbf{M}_{mn} &\to \mathbf{M}_{nm} && \text{where } R(A) = A^T \text{ for all } A \text{ in } \mathbf{M}_{mn} \\ S : \mathbf{M}_{mn} &\to \mathbb{R} && \text{where } S(A) = \text{tr } A \text{ for all } A \text{ in } \mathbf{M}_{nn} \end{aligned}$$

are both linear transformations.

Solution. Axioms T1 and T2 for transposition are $(A+B)^T = A^T + B^T$ and $(rA)^T = r(A^T)$, respectively (using Theorem 2.1.2). The verifications for the trace are left to the reader.

Example 7.1.3

If a is a scalar, define $E_a : \mathbf{P}_n \to \mathbb{R}$ by $E_a(p) = p(a)$ for each polynomial p in $\mathbf{P}_n$. Show that E_a is a linear transformation (called **evaluation** at a).

Solution. If p and q are polynomials and r is in $\mathbb{R}$, we use the fact that the sum $p+q$ and scalar product rp are defined as for functions:

$$(p+q)(x) = p(x) + q(x) \quad \text{and} \quad (rp)(x) = rp(x)$$

for all x. Hence, for all p and q in $\mathbf{P}_n$ and all r in $\mathbb{R}$:

$$\begin{aligned} E_a(p+q) &= (p+q)(a) = p(a) + q(a) = E_a(p) + E_a(q), \quad \text{and} \\ E_a(rp) &= (rp)(a) = rp(a) = rE_a(p). \end{aligned}$$

Hence E_a is a linear transformation.

The next example involves some calculus.

Example 7.1.4

Show that the differentiation and integration operations on $\mathbf{P}_n$ are linear transformations. More precisely,

$$\begin{aligned} D : \mathbf{P}_n &\to \mathbf{P}_{n-1} && \text{where } D[p(x)] = p'(x) \text{ for all } p(x) \text{ in } \mathbf{P}_n \\ I : \mathbf{P}_n &\to \mathbf{P}_{n+1} && \text{where } I[p(x)] = \int_0^x p(t)dt \text{ for all } p(x) \text{ in } \mathbf{P}_n \end{aligned}$$

are linear transformations.

Solution. These restate the following fundamental properties of differentiation and integration.

$$[p(x) + q(x)]' = p'(x) + q'(x) \quad \text{and} \quad [rp(x)]' = (rp)'(x)$$

$$\textstyle\int_0^x [p(t) + q(t)]\,dt = \int_0^x p(t)dt + \int_0^x q(t)dt \quad \text{and} \quad \int_0^x rp(t)dt = r\int_0^x p(t)dt$$

The next theorem collects three useful properties of *all* linear transformations. They can be described by saying that, in addition to preserving addition and scalar multiplication (these are the axioms), linear transformations preserve the zero vector, negatives, and linear combinations.

Theorem 7.1.1

Let $T : V \to W$ be a linear transformation.

1. $T(\mathbf{0}) = \mathbf{0}$.
2. $T(-\mathbf{v}) = -T(\mathbf{v})$ *for all* $\mathbf{v}$ *in* V.
3. $T(r_1\mathbf{v}_1 + r_2\mathbf{v}_2 + \cdots + r_k\mathbf{v}_k) = r_1T(\mathbf{v}_1) + r_2T(\mathbf{v}_2) + \cdots + r_kT(\mathbf{v}_k)$ *for all* $\mathbf{v}_i$ *in* V *and all* r_i *in* $\mathbb{R}$.

Proof.

1. $T(\mathbf{0}) = T(0\mathbf{v}) = 0T(\mathbf{v}) = \mathbf{0}$ for any $\mathbf{v}$ in V.
2. $T(-\mathbf{v}) = T[(-1)\mathbf{v}] = (-1)T(\mathbf{v}) = -T(\mathbf{v})$ for any $\mathbf{v}$ in V.
3. The proof of Theorem 2.6.1 goes through. □

The ability to use the last part of Theorem 7.1.1 effectively is vital to obtaining the benefits of linear transformations. Example 7.1.5 and Theorem 7.1.2 provide illustrations.

Example 7.1.5

Let $T : V \to W$ be a linear transformation. If $T(\mathbf{v} - 3\mathbf{v}_1) = \mathbf{w}$ and $T(2\mathbf{v} - \mathbf{v}_1) = \mathbf{w}_1$, find $T(\mathbf{v})$ and $T(\mathbf{v}_1)$ in terms of $\mathbf{w}$ and $\mathbf{w}_1$.

Solution. The given relations imply that

$$\begin{aligned} T(\mathbf{v}) - 3T(\mathbf{v}_1) &= \mathbf{w} \\ 2T(\mathbf{v}) - T(\mathbf{v}_1) &= \mathbf{w}_1 \end{aligned}$$

by Theorem 7.1.1. Subtracting twice the first from the second gives $T(\mathbf{v}_1) = \frac{1}{5}(\mathbf{w}_1 - 2\mathbf{w})$. Then substitution gives $T(\mathbf{v}) = \frac{1}{5}(3\mathbf{w}_1 - \mathbf{w})$.

The full effect of property (3) in Theorem 7.1.1 is this: If $T : V \to W$ is a linear transformation and $T(\mathbf{v}_1), T(\mathbf{v}_2), \dots, T(\mathbf{v}_n)$ are known, then $T(\mathbf{v})$ can be computed for *every* vector $\mathbf{v}$ in span $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$. In particular, if $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ spans V, then $T(\mathbf{v})$ is determined for all $\mathbf{v}$ in V by the choice of $T(\mathbf{v}_1), T(\mathbf{v}_2), \dots, T(\mathbf{v}_n)$. The next theorem states this somewhat differently. As for functions in general, two linear transformations $T : V \to W$ and $S : V \to W$ are called **equal** (written $T = S$) if they have the same **action**; that is, if $T(\mathbf{v}) = S(\mathbf{v})$ for all $\mathbf{v}$ in V.

Theorem 7.1.2

Let $T : V \to W$ and $S : V \to W$ be two linear transformations. Suppose that $V = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$. If $T(\mathbf{v}_i) = S(\mathbf{v}_i)$ for each i, then $T = S$.

Proof. If $\mathbf{v}$ is any vector in $V = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$, write $\mathbf{v} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n$ where each a_i is in $\mathbb{R}$. Since $T(\mathbf{v}_i) = S(\mathbf{v}_i)$ for each i, Theorem 7.1.1 gives

$$\begin{aligned} T(\mathbf{v}) &= T(a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n) \\ &= a_1T(\mathbf{v}_1) + a_2T(\mathbf{v}_2) + \cdots + a_nT(\mathbf{v}_n) \\ &= a_1S(\mathbf{v}_1) + a_2S(\mathbf{v}_2) + \cdots + a_nS(\mathbf{v}_n) \\ &= S(a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n) \\ &= S(\mathbf{v}) \end{aligned}$$

Since $\mathbf{v}$ was arbitrary in V, this shows that $T = S$. □

Example 7.1.6

Let $V = \text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$. Let $T : V \to W$ be a linear transformation. If $T(\mathbf{v}_1) = \cdots = T(\mathbf{v}_n) = \mathbf{0}$, show that $T = 0$, the zero transformation from V to W.

Solution. The zero transformation $0 : V \to W$ is defined by $0(\mathbf{v}) = \mathbf{0}$ for all $\mathbf{v}$ in V (Example 7.1.1), so $T(\mathbf{v}_i) = 0(\mathbf{v}_i)$ holds for each i. Hence $T = 0$ by Theorem 7.1.2.

Theorem 7.1.2 can be expressed as follows: If we know what a linear transformation $T : V \to W$ does to each vector in a spanning set for V, then we know what T does to *every* vector in V. If the spanning set is a basis, we can say much more.

Theorem 7.1.3

Let V and W be vector spaces and let $\{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ be a basis of V. Given any vectors $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$ in W (they need not be distinct), there exists a unique linear transformation $T : V \to W$ satisfying $T(\mathbf{b}_i) = \mathbf{w}_i$ for each $i = 1, 2, \dots, n$. In fact, the action of T is as follows: Given $\mathbf{v} = v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + \cdots + v_n\mathbf{b}_n$ in V, v_i in $\mathbb{R}$, then

$$T(\mathbf{v}) = T(v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + \cdots + v_n\mathbf{b}_n) = v_1\mathbf{w}_1 + v_2\mathbf{w}_2 + \cdots + v_n\mathbf{w}_n.$$

Proof. If a transformation T *does* exist with $T(\mathbf{b}_i) = \mathbf{w}_i$ for each i, and if S is any other such transformation, then $T(\mathbf{b}_i) = \mathbf{w}_i = S(\mathbf{b}_i)$ holds for each i, so $S = T$ by Theorem 7.1.2. Hence T is unique if it exists, and it remains to show that there really is such a linear transformation. Given $\mathbf{v}$ in V, we must specify $T(\mathbf{v})$ in W. Because $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ is a basis of V, we have $\mathbf{v} = v_1\mathbf{b}_1 + \cdots + v_n\mathbf{b}_n$, where $v_1, \dots, v_n$ are *uniquely* determined by $\mathbf{v}$ (this is Theorem 6.3.1). Hence we may define $T : V \to W$ by

$$T(\mathbf{v}) = T(v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + \cdots + v_n\mathbf{b}_n) = v_1\mathbf{w}_1 + v_2\mathbf{w}_2 + \cdots + v_n\mathbf{w}_n$$

for all $\mathbf{v} = v_1\mathbf{b}_1 + \cdots + v_n\mathbf{b}_n$ in V. This satisfies $T(\mathbf{b}_i) = \mathbf{w}_i$ for each i; the verification that T is linear is left to the reader. □

This theorem shows that linear transformations can be defined almost at will: Simply specify where the basis vectors go, and the rest of the action is dictated by the linearity. Moreover, Theorem 7.1.2 shows that deciding whether two linear transformations are equal comes down to determining whether they have the same effect on the basis vectors. So, given a basis $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ of a vector space V, there is a different linear transformation $V \to W$ for every ordered selection $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$ of vectors in W (not necessarily distinct).

Example 7.1.7

Find a linear transformation $T : \mathbf{P}_2 \to \mathbf{M}_{22}$ such that

$$T(1+x) = \left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array} \right], \quad T(x+x^2) = \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \right], \quad \text{and} \quad T(1+x^2) = \left[\begin{array}{cc} 0 & 0 \\ 0 & 1 \end{array} \right].$$

Solution. The set $\{1+x,\ x+x^2,\ 1+x^2\}$ is a basis of $\mathbf{P}_2$, so every vector $p = a + bx + cx^2$ in $\mathbf{P}_2$ is a linear combination of these vectors. In fact

$$p(x) = \frac{1}{2}(a+b-c)(1+x) + \frac{1}{2}(-a+b+c)(x+x^2) + \frac{1}{2}(a-b+c)(1+x^2)$$

Hence Theorem 7.1.3 gives

$$\begin{aligned} T\left[p(x)\right] &= \frac{1}{2}(a+b-c) \left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array} \right] + \frac{1}{2}(-a+b+c) \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \right] + \frac{1}{2}(a-b+c) \left[\begin{array}{cc} 0 & 0 \\ 0 & 1 \end{array} \right] \\ &= \frac{1}{2} \left[\begin{array}{cc} a+b-c & -a+b+c \\ -a+b+c & a-b+c \end{array} \right] \end{aligned}$$

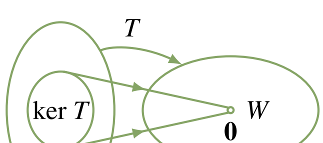
7.2 Kernel and Image of a Linear Transformation

This section is devoted to two important subspaces associated with a linear transformation $T : V \to W$.

Definition 7.2 Kernel and Image of a Linear Transformation

The ***kernel*** *of* T *(denoted* ker T*) and the* ***image*** *of* T *(denoted* im T *or* $T(V)$*) are defined by*

$$\begin{aligned} \ker T &= \{\mathbf{v} \text{ in } V \mid T(\mathbf{v}) = \mathbf{0}\} \\ \operatorname{im} T &= \{T(\mathbf{v}) \mid \mathbf{v} \text{ in } V\} = T(V) \end{aligned}$$



The kernel of T is often called the **nullspace** of T because it consists of all vectors $\mathbf{v}$ in V satisfying the *condition* that $T(\mathbf{v}) = \mathbf{0}$. The image of T is often called the **range** of T and consists of all vectors $\mathbf{w}$ in W of the *form* $\mathbf{w} = T(\mathbf{v})$ for some $\mathbf{v}$ in V. These subspaces are depicted in the diagrams.

Example 7.2.1

Let $T_A : \mathbb{R}^n \to \mathbb{R}^m$ be the linear transformation induced by the $m \times n$ matrix A, that is $T_A(\mathbf{x}) = A\mathbf{x}$ for all columns $\mathbf{x}$ in $\mathbb{R}^n$. Then

$$\ker T_A = \{\mathbf{x} \mid A\mathbf{x} = \mathbf{0}\} = \text{null } A \quad \text{ and}$$

$$\text{im}\, T_A = \{A\mathbf{x} \mid \mathbf{x} \text{ in } \mathbb{R}^n\} = \text{im}\, A$$

Hence the following theorem extends Example 5.1.2.

Theorem 7.2.1

Let $T : V \to W$ be a linear transformation.

1. $\ker T$ *is a subspace of* V.
2. $\operatorname{im} T$ *is a subspace of* W.

Proof. The fact that $T(\mathbf{0}) = \mathbf{0}$ shows that $\ker T$ and $\operatorname{im} T$ contain the zero vector of V and W respectively.

1. If $\mathbf{v}$ and $\mathbf{v}_1$ lie in $\ker T$, then $T(\mathbf{v}) = \mathbf{0} = T(\mathbf{v}_1)$, so

$$\begin{aligned} T(\mathbf{v}+\mathbf{v}_1) &= T(\mathbf{v}) + T(\mathbf{v}_1) = \mathbf{0}+\mathbf{0} = \mathbf{0} \\ T(r\mathbf{v}) &= rT(\mathbf{v}) = r\mathbf{0} = \mathbf{0} \quad \text{for all } r \text{ in } \mathbb{R} \end{aligned}$$

   Hence $\mathbf{v}+\mathbf{v}_1$ and $r\mathbf{v}$ lie in $\ker T$ (they satisfy the required condition), so $\ker T$ is a subspace of V by the subspace test (Theorem 6.2.1).

2. If $\mathbf{w}$ and $\mathbf{w}_1$ lie in $\operatorname{im} T$, write $\mathbf{w} = T(\mathbf{v})$ and $\mathbf{w}_1 = T(\mathbf{v}_1)$ where $\mathbf{v},\ \mathbf{v}_1 \in V$. Then

$$\begin{aligned} \mathbf{w}+\mathbf{w}_1 &= T(\mathbf{v}) + T(\mathbf{v}_1) = T(\mathbf{v}+\mathbf{v}_1) \\ r\mathbf{w} &= rT(\mathbf{v}) = T(r\mathbf{v}) \quad \text{for all } r \text{ in } \mathbb{R} \end{aligned}$$

   Hence $\mathbf{w}+\mathbf{w}_1$ and $r\mathbf{w}$ both lie in $\operatorname{im} T$ (they have the required form), so $\operatorname{im} T$ is a subspace of W.

□

Given a linear transformation $T : V \to W$:

$\dim(\ker T)$ is called the **nullity** of T and denoted as $\operatorname{nullity}(T)$
$\dim(\operatorname{im} T)$ is called the **rank** of T and denoted as $\operatorname{rank}(T)$

The rank of a matrix A was defined earlier to be the dimension of $\operatorname{col} A$, the column space of A. The two usages of the word *rank* are consistent in the following sense. Recall the definition of T_A in Example 7.2.1.

Example 7.2.2

Given an $m \times n$ matrix A, show that $\operatorname{im} T_A = \operatorname{col} A$, so $\operatorname{rank} T_A = \operatorname{rank} A$.

Solution. Write $A = \left[\begin{array}{ccc} \mathbf{c}_1 & \cdots & \mathbf{c}_n \end{array} \right]$ in terms of its columns. Then

$$\operatorname{im} T_A = \{A\mathbf{x} \mid \mathbf{x} \text{ in } \mathbb{R}^n\} = \{x_1\mathbf{c}_1 + \cdots + x_n\mathbf{c}_n \mid x_i \text{ in } \mathbb{R}\}$$

using Definition 2.5. Hence $\operatorname{im} T_A$ is the column space of A; the rest follows.

Often, a useful way to study a subspace of a vector space is to exhibit it as the kernel or image of a linear transformation. Here is an example.

Example 7.2.3

Define a transformation $P : \mathbf{M}_{nn} \to \mathbf{M}_{nn}$ by $P(A) = A - A^T$ for all A in $\mathbf{M}_{nn}$. Show that P is linear and that:

a. $\ker P$ consists of all symmetric matrices.

b. $\operatorname{im} P$ consists of all skew-symmetric matrices.

Solution. The verification that P is linear is left to the reader. To prove part (a), note that a matrix A lies in $\ker P$ just when $0 = P(A) = A - A^T$, and this occurs if and only if $A = A^T$—that is, A is symmetric. Turning to part (b), the space $\operatorname{im} P$ consists of all matrices $P(A)$, A in $\mathbf{M}_{nn}$. Every such matrix is skew-symmetric because

$$P(A)^T = (A - A^T)^T = A^T - A = -P(A)$$

On the other hand, if S is skew-symmetric (that is, $S^T = -S$), then S lies in $\operatorname{im} P$. In fact,

$$P\left[\tfrac{1}{2}S\right] = \tfrac{1}{2}S - \left[\tfrac{1}{2}S\right]^T = \tfrac{1}{2}(S - S^T) = \tfrac{1}{2}(S + S) = S$$

One-to-One and Onto Transformations

Definition 7.3 One-to-one and Onto Linear Transformations

Let $T : V \to W$ be a linear transformation.

1. *T is said to be* **onto** *if* $\operatorname{im} T = W$.
2. *T is said to be* **one-to-one** *if $T(\mathbf{v}) = T(\mathbf{v}_1)$ implies $\mathbf{v} = \mathbf{v}_1$.*

A vector $\mathbf{w}$ in W is said to be **hit** by T if $\mathbf{w} = T(\mathbf{v})$ for some $\mathbf{v}$ in V. Then T is onto if every vector in W is hit at least once, and T is one-to-one if no element of W gets hit twice. Clearly the onto transformations T are those for which $\operatorname{im} T = W$ is as large a subspace of W as possible. By contrast, Theorem 7.2.2 shows that the one-to-one transformations T are the ones with $\ker T$ as *small* a subspace of V as possible.

Theorem 7.2.2

If $T : V \to W$ is a linear transformation, then T is one-to-one if and only if $\ker T = \{\mathbf{0}\}$.

Proof. If T is one-to-one, let $\mathbf{v}$ be any vector in $\ker T$. Then $T(\mathbf{v}) = \mathbf{0}$, so $T(\mathbf{v}) = T(\mathbf{0})$. Hence $\mathbf{v} = \mathbf{0}$ because T is one-to-one. Hence $\ker T = \{\mathbf{0}\}$.

Conversely, assume that $\ker T = \{0\}$ and let $T(\mathbf{v}) = T(\mathbf{v}_1)$ with $\mathbf{v}$ and $\mathbf{v}_1$ in V. Then $T(\mathbf{v} - \mathbf{v}_1) = T(\mathbf{v}) - T(\mathbf{v}_1) = \mathbf{0}$, so $\mathbf{v} - \mathbf{v}_1$ lies in $\ker T = \{\mathbf{0}\}$. This means that $\mathbf{v} - \mathbf{v}_1 = \mathbf{0}$, so $\mathbf{v} = \mathbf{v}_1$, proving that T is one-to-one. □

Example 7.2.4

The identity transformation $1_V : V \to V$ is both one-to-one and onto for any vector space V.

Example 7.2.5

Consider the linear transformations

$$\begin{aligned} S : \mathbb{R}^3 \to \mathbb{R}^2 &\quad \text{given by } S(x,\ y,\ z) = (x+y,\ x-y) \\ T : \mathbb{R}^2 \to \mathbb{R}^3 &\quad \text{given by } T(x,\ y) = (x+y,\ x-y,\ x) \end{aligned}$$

Show that T is one-to-one but not onto, whereas S is onto but not one-to-one.

Solution. The verification that they are linear is omitted. T is one-to-one because

$$\ker T = \{(x,\ y) \mid x+y = x-y = x = 0\} = \{(0,\ 0)\}$$

However, it is not onto. For example $(0,\ 0,\ 1)$ does not lie in $\operatorname{im} T$ because if $(0,\ 0,\ 1) = (x+y,\ x-y,\ x)$ for some x and y, then $x+y = 0 = x-y$ and $x = 1$, an impossibility. Turning to S, it is not one-to-one by Theorem 7.2.2 because $(0,\ 0,\ 1)$ lies in $\ker S$. But every element $(s,\ t)$ in $\mathbb{R}^2$ lies in $\operatorname{im} S$ because $(s,\ t) = (x+y,\ x-y) = S(x,\ y,\ z)$ for some x, y, and z (in fact, $x = \frac{1}{2}(s+t)$, $y = \frac{1}{2}(s-t)$, and $z = 0$). Hence S is onto.

Example 7.2.6

Let U be an invertible $m \times m$ matrix and define

$$T : \mathbf{M}_{mn} \to \mathbf{M}_{mn} \quad \text{by} \quad T(X) = UX \text{ for all } X \text{ in } \mathbf{M}_{mn}$$

Show that T is a linear transformation that is both one-to-one and onto.

Solution. The verification that T is linear is left to the reader. To see that T is one-to-one, let $T(X) = 0$. Then $UX = 0$, so left-multiplication by U^{-1} gives $X = 0$. Hence $\ker T = \{\mathbf{0}\}$, so T is one-to-one. Finally, if Y is any member of $\mathbf{M}_{mn}$, then $U^{-1}Y$ lies in $\mathbf{M}_{mn}$ too, and $T(U^{-1}Y) = U(U^{-1}Y) = Y$. This shows that T is onto.

The linear transformations $\mathbb{R}^n \to \mathbb{R}^m$ all have the form T_A for some $m \times n$ matrix A (Theorem 2.6.2). The next theorem gives conditions under which they are onto or one-to-one. Note the connection with Theorem 5.4.3 and Theorem 5.4.4.

Theorem 7.2.3

Let A be an $m \times n$ matrix, and let $T_A : \mathbb{R}^n \to \mathbb{R}^m$ be the linear transformation induced by A, that is $T_A(\mathbf{x}) = A\mathbf{x}$ for all columns $\mathbf{x}$ in $\mathbb{R}^n$.

1. *T_A is onto if and only if* $\operatorname{rank} A = m$.
2. *T_A is one-to-one if and only if* $\operatorname{rank} A = n$.

Proof.

1. We have that $\operatorname{im} T_A$ is the column space of A (see Example 7.2.2), so T_A is onto if and only if the column space of A is $\mathbb{R}^m$. Because the rank of A is the dimension of the column space, this holds if and only if $\operatorname{rank} A = m$.
2. $\ker T_A = \{\mathbf{x} \text{ in } \mathbb{R}^n \mid A\mathbf{x} = \mathbf{0}\}$, so (using Theorem 7.2.2) T_A is one-to-one if and only if $A\mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$. This is equivalent to $\operatorname{rank} A = n$ by Theorem 5.4.3. □

The Dimension Theorem

Let A denote an $m \times n$ matrix of rank r and let $T_A : \mathbb{R}^n \to \mathbb{R}^m$ denote the corresponding matrix transformation given by $T_A(\mathbf{x}) = A\mathbf{x}$ for all columns $\mathbf{x}$ in $\mathbb{R}^n$. It follows from Example 7.2.1 and Example 7.2.2 that $\operatorname{im} T_A = \operatorname{col} A$, so $\dim(\operatorname{im} T_A) = \dim(\operatorname{col} A) = r$. On the other hand Theorem 5.4.2 shows that $\dim(\ker T_A) = \dim(\operatorname{null} A) = n - r$. Combining these we see that

$$\dim(\operatorname{im} T_A) + \dim(\ker T_A) = n \quad \text{for every } m \times n \text{ matrix } A$$

The main result of this section is a deep generalization of this observation.

Theorem 7.2.4: Dimension Theorem

Let $T : V \to W$ be any linear transformation and assume that $\ker T$ and $\operatorname{im} T$ are both finite dimensional. Then V is also finite dimensional and

$$\dim V = \dim(\ker T) + \dim(\operatorname{im} T)$$

In other words, $\dim V = \operatorname{nullity}(T) + \operatorname{rank}(T)$.

Proof. Every vector in $\operatorname{im} T = T(V)$ has the form $T(\mathbf{v})$ for some $\mathbf{v}$ in V. Hence let $\{T(\mathbf{e}_1), T(\mathbf{e}_2), \dots, T(\mathbf{e}_r)\}$ be a basis of $\operatorname{im} T$, where the $\mathbf{e}_i$ lie in V. Let $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_k\}$ be any basis of $\ker T$. Then $\dim(\operatorname{im} T) = r$ and $\dim(\ker T) = k$, so it suffices to show that $B = \{\mathbf{e}_1, \dots, \mathbf{e}_r, \mathbf{f}_1, \dots, \mathbf{f}_k\}$ is a basis of V.

1. *B spans V.* If $\mathbf{v}$ lies in V, then $T(\mathbf{v})$ lies in $\operatorname{im} V$, so

$$T(\mathbf{v}) = t_1 T(\mathbf{e}_1) + t_2 T(\mathbf{e}_2) + \cdots + t_r T(\mathbf{e}_r) \quad t_i \text{ in } \mathbb{R}$$

This implies that $\mathbf{v} - t_1\mathbf{e}_1 - t_2\mathbf{e}_2 - \cdots - t_r\mathbf{e}_r$ lies in $\ker T$ and so is a linear combination of $\mathbf{f}_1, \dots, \mathbf{f}_k$. Hence $\mathbf{v}$ is a linear combination of the vectors in B.

2. B is *linearly independent*. Suppose that t_i and s_j in $\mathbb{R}$ satisfy

$$t_1\mathbf{e}_1 + \cdots + t_r\mathbf{e}_r + s_1\mathbf{f}_1 + \cdots + s_k\mathbf{f}_k = \mathbf{0} \tag{7.1}$$

Applying T gives $t_1T(\mathbf{e}_1) + \cdots + t_rT(\mathbf{e}_r) = \mathbf{0}$ (because $T(\mathbf{f}_i) = \mathbf{0}$ for each i). Hence the independence of $\{T(\mathbf{e}_1),\ \dots,\ T(\mathbf{e}_r)\}$ yields $t_1 = \cdots = t_r = 0$. But then (7.1) becomes

$$s_1\mathbf{f}_1 + \cdots + s_k\mathbf{f}_k = \mathbf{0}$$

so $s_1 = \cdots = s_k = 0$ by the independence of $\{\mathbf{f}_1,\ \dots,\ \mathbf{f}_k\}$. This proves that B is linearly independent.

□

Note that the vector space V is not assumed to be finite dimensional in Theorem 7.2.4. In fact, verifying that $\ker T$ and $\operatorname{im} T$ are both finite dimensional is often an important way to *prove* that V is finite dimensional.

Note further that $r + k = n$ in the proof so, after relabelling, we end up with a basis

$$B = \{\mathbf{e}_1,\ \mathbf{e}_2,\ \dots,\ \mathbf{e}_r,\ \mathbf{e}_{r+1},\ \dots,\ \mathbf{e}_n\}$$

of V with the property that $\{\mathbf{e}_{r+1},\ \dots,\ \mathbf{e}_n\}$ is a basis of $\ker T$ and $\{T(\mathbf{e}_1),\ \dots,\ T(\mathbf{e}_r)\}$ is a basis of $\operatorname{im} T$. In fact, if V is known in advance to be finite dimensional, then *any* basis $\{\mathbf{e}_{r+1},\ \dots,\ \mathbf{e}_n\}$ of $\ker T$ can be extended to a basis $\{\mathbf{e}_1,\ \mathbf{e}_2,\ \dots,\ \mathbf{e}_r,\ \mathbf{e}_{r+1},\ \dots,\ \mathbf{e}_n\}$ of V by Theorem 6.4.1. Moreover, it turns out that, no matter how this is done, the vectors $\{T(\mathbf{e}_1),\ \dots,\ T(\mathbf{e}_r)\}$ will be a basis of $\operatorname{im} T$. This result is useful, and we record it for reference. The proof is much like that of Theorem 7.2.4 and is left as Exercise **??**.

Theorem 7.2.5

Let $T : V \to W$ be a linear transformation, and let $\{\mathbf{e}_1,\ \dots,\ \mathbf{e}_r,\ \mathbf{e}_{r+1},\ \dots,\ \mathbf{e}_n\}$ be a basis of V such that $\{\mathbf{e}_{r+1},\ \dots,\ \mathbf{e}_n\}$ is a basis of $\ker T$. Then $\{T(\mathbf{e}_1),\ \dots,\ T(\mathbf{e}_r)\}$ is a basis of $\operatorname{im} T$, and hence $r = \operatorname{rank} T$.

The dimension theorem is one of the most useful results in all of linear algebra. It shows that if either $\dim(\ker T)$ or $\dim(\operatorname{im} T)$ can be found, then the other is automatically known. In many cases it is easier to compute one than the other, so the theorem is a real asset. The rest of this section is devoted to illustrations of this fact. The next example uses the dimension theorem to give a different proof of the first part of Theorem 5.4.2.

Example 7.2.7

Let A be an $m \times n$ matrix of rank r. Show that the space $\operatorname{null} A$ of all solutions of the system $A\mathbf{x} = \mathbf{0}$ of m homogeneous equations in n variables has dimension $n - r$.

Solution. The space in question is just $\ker T_A$, where $T_A : \mathbb{R}^n \to \mathbb{R}^m$ is defined by $T_A(\mathbf{x}) = A\mathbf{x}$ for all columns $\mathbf{x}$ in $\mathbb{R}^n$. But $\dim(\operatorname{im} T_A) = \operatorname{rank} T_A = \operatorname{rank} A = r$ by Example 7.2.2, so $\dim(\ker T_A) = n - r$ by the dimension theorem.

Example 7.2.8

If $T : V \to W$ is a linear transformation where V is finite dimensional, then

$$\dim(\ker T) \leq \dim V \quad \text{and} \quad \dim(\operatorname{im} T) \leq \dim V$$

Indeed, $\dim V = \dim(\ker T) + \dim(\operatorname{im} T)$ by Theorem 7.2.4. Of course, the first inequality also follows because $\ker T$ is a subspace of V.

Example 7.2.9

Let $D : \mathbf{P}_n \to \mathbf{P}_{n-1}$ be the differentiation map defined by $D[p(x)] = p'(x)$. Compute $\ker D$ and hence conclude that D is onto.

Solution. Because $p'(x) = 0$ means $p(x)$ is constant, we have $\dim(\ker D) = 1$. Since $\dim \mathbf{P}_n = n+1$, the dimension theorem gives

$$\dim(\operatorname{im} D) = (n+1) - \dim(\ker D) = n = \dim(\mathbf{P}_{n-1})$$

This implies that $\operatorname{im} D = \mathbf{P}_{n-1}$, so D is onto.

Of course it is not difficult to verify directly that each polynomial $q(x)$ in $\mathbf{P}_{n-1}$ is the derivative of some polynomial in $\mathbf{P}_n$ (simply integrate $q(x)$!), so the dimension theorem is not needed in this case. However, in some situations it is difficult to see directly that a linear transformation is onto, and the method used in Example 7.2.9 may be by far the easiest way to prove it. Here is another illustration.

Example 7.2.10

Given a in $\mathbb{R}$, the evaluation map $E_a : \mathbf{P}_n \to \mathbb{R}$ is given by $E_a[p(x)] = p(a)$. Show that E_a is linear and onto, and hence conclude that $\{(x-a), (x-a)^2, \dots, (x-a)^n\}$ is a basis of $\ker E_a$, the subspace of all polynomials $p(x)$ for which $p(a) = 0$.

Solution. E_a is linear by Example 7.1.3; the verification that it is onto is left to the reader. Hence $\dim(\operatorname{im} E_a) = \dim(\mathbb{R}) = 1$, so $\dim(\ker E_a) = (n+1) - 1 = n$ by the dimension theorem. Now each of the n polynomials $(x-a), (x-a)^2, \dots, (x-a)^n$ clearly lies in $\ker E_a$, and they are linearly independent (they have distinct degrees). Hence they are a basis because $\dim(\ker E_a) = n$.

We conclude by applying the dimension theorem to the rank of a matrix.

Example 7.2.11

If A is any $m \times n$ matrix, show that $\operatorname{rank} A = \operatorname{rank} A^T A = \operatorname{rank} AA^T$.

Solution. It suffices to show that $\operatorname{rank} A = \operatorname{rank} A^T A$ (the rest follows by replacing A with A^T). Write $B = A^T A$, and consider the associated matrix transformations

$$T_A : \mathbb{R}^n \to \mathbb{R}^m \quad \text{and} \quad T_B : \mathbb{R}^n \to \mathbb{R}^n$$

The dimension theorem and Example 7.2.2 give

$$\begin{aligned} \text{rank } A &= \text{rank } T_A = \dim(\text{im } T_A) = n - \dim(\ker T_A) \\ \text{rank } B &= \text{rank } T_B = \dim(\text{im } T_B) = n - \dim(\ker T_B) \end{aligned}$$

so it suffices to show that $\ker T_A = \ker T_B$. Now $A\mathbf{x} = \mathbf{0}$ implies that $B\mathbf{x} = A^T A\mathbf{x} = \mathbf{0}$, so $\ker T_A$ is contained in $\ker T_B$. On the other hand, if $B\mathbf{x} = \mathbf{0}$, then $A^T A\mathbf{x} = \mathbf{0}$, so

$$\|A\mathbf{x}\|^2 = (A\mathbf{x})^T(A\mathbf{x}) = \mathbf{x}^T A^T A\mathbf{x} = \mathbf{x}^T\mathbf{0} = 0$$

This implies that $A\mathbf{x} = \mathbf{0}$, so $\ker T_B$ is contained in $\ker T_A$.

7.3 Isomorphisms and Composition

Often two vector spaces can consist of quite different types of vectors but, on closer examination, turn out to be the same underlying space displayed in different symbols. For example, consider the spaces

$$\mathbb{R}^2 = \{(a, b) \mid a, b \in \mathbb{R}\} \quad \text{and} \quad \mathbf{P}_1 = \{a + bx \mid a, b \in \mathbb{R}\}$$

Compare the addition and scalar multiplication in these spaces:

$$(a, b) + (a_1, b_1) = (a + a_1, b + b_1) \qquad (a + bx) + (a_1 + b_1x) = (a + a_1) + (b + b_1)x$$

$$r(a,\ b) = (ra,\ rb) \qquad\qquad r(a+bx) = (ra) + (rb)x$$

Clearly these are the *same* vector space expressed in different notation: if we change each $(a,\ b)$ in $\mathbb{R}^2$ to $a+bx$, then $\mathbb{R}^2$ *becomes* $\mathbf{P}_1$, complete with addition and scalar multiplication. This can be expressed by noting that the map $(a,\ b) \mapsto a+bx$ is a linear transformation $\mathbb{R}^2 \to \mathbf{P}_1$ that is both one-to-one and onto. In this form, we can describe the general situation.

Definition 7.4 Isomorphic Vector Spaces

A linear transformation $T : V \to W$ is called an ***isomorphism*** *if it is both onto and one-to-one. The vector spaces V and W are said to be* ***isomorphic*** *if there exists an isomorphism $T : V \to W$, and we write $V \cong W$ when this is the case.*

Example 7.3.1

The identity transformation $1_V : V \to V$ is an isomorphism for any vector space V.

Example 7.3.2

If $T : \mathbf{M}_{mn} \to \mathbf{M}_{nm}$ is defined by $T(A) = A^T$ for all A in $\mathbf{M}_{mn}$, then T is an isomorphism (verify). Hence $\mathbf{M}_{mn} \cong \mathbf{M}_{nm}$.

Example 7.3.3

Isomorphic spaces can "look" quite different. For example, $\mathbf{M}_{22} \cong \mathbf{P}_3$ because the map $T : \mathbf{M}_{22} \to \mathbf{P}_3$ given by $T\left[\begin{array}{cc} a & b \\ c & d \end{array}\right] = a+bx+cx^2+dx^3$ is an isomorphism (verify).

The word *isomorphism* comes from two Greek roots: *iso*, meaning "same," and *morphos*, meaning "form." An isomorphism $T : V \to W$ induces a pairing

$$\mathbf{v} \leftrightarrow T(\mathbf{v})$$

between vectors $\mathbf{v}$ in V and vectors $T(\mathbf{v})$ in W that preserves vector addition and scalar multiplication. Hence, *as far as their vector space properties are concerned*, the spaces V and W are identical except for notation. Because addition and scalar multiplication in either space are completely determined by the same operations in the other space, all *vector space* properties of either space are completely determined by those of the other.

One of the most important examples of isomorphic spaces was considered in Chapter 4. Let A denote the set of all "arrows" with tail at the origin in space, and make A into a vector space using the parallelogram law and the scalar multiple law (see Section 4.1). Then define a transformation $T : \mathbb{R}^3 \to A$ by taking

$$T\left[\begin{array}{c} x \\ y \\ z \end{array}\right] = \text{ the arrow } \mathbf{v} \text{ from the origin to the point } P(x,\ y,\ z).$$

In Section 4.1 matrix addition and scalar multiplication were shown to correspond to the parallelogram law and the scalar multiplication law for these arrows, so the map T is a linear transformation. Moreover T is an isomorphism: it is one-to-one by Theorem 4.1.2, and it is onto because, given an arrow $\mathbf{v}$ in A with tip $P(x, y, z)$, we have $T\left[\begin{array}{c} x \\ y \\ z \end{array}\right] = \mathbf{v}$. This justifies the identification $\mathbf{v} = \left[\begin{array}{c} x \\ y \\ z \end{array}\right]$ in Chapter 4 of the geometric arrows with the algebraic matrices. This identification is very useful. The arrows give a "picture" of the matrices and so bring geometric intuition into $\mathbb{R}^3$; the matrices are useful for detailed calculations and so bring analytic precision into geometry. This is one of the best examples of the power of an isomorphism to shed light on both spaces being considered.

The following theorem gives a very useful characterization of isomorphisms: They are the linear transformations that preserve bases.

Theorem 7.3.1

If V and W are finite dimensional spaces, the following conditions are equivalent for a linear transformation $T : V \to W$.

1. *T is an isomorphism.*
2. *If $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ is any basis of V, then $\{T(\mathbf{e}_1), T(\mathbf{e}_2), \dots, T(\mathbf{e}_n)\}$ is a basis of W.*
3. *There exists a basis $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ of V such that $\{T(\mathbf{e}_1), T(\mathbf{e}_2), \dots, T(\mathbf{e}_n)\}$ is a basis of W.*

Proof. (1) $\Rightarrow$ (2). Let $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ be a basis of V. If $t_1T(\mathbf{e}_1) + \cdots + t_nT(\mathbf{e}_n) = \mathbf{0}$ with t_i in $\mathbb{R}$, then $T(t_1\mathbf{e}_1 + \cdots + t_n\mathbf{e}_n) = \mathbf{0}$, so $t_1\mathbf{e}_1 + \cdots + t_n\mathbf{e}_n = \mathbf{0}$ (because $\ker T = \{\mathbf{0}\}$). But then each $t_i = 0$ by the independence of the $\mathbf{e}_i$, so $\{T(\mathbf{e}_1), \dots, T(\mathbf{e}_n)\}$ is independent. To show that it spans W, choose $\mathbf{w}$ in W. Because T is onto, $\mathbf{w} = T(\mathbf{v})$ for some $\mathbf{v}$ in V, so write $\mathbf{v} = t_1\mathbf{e}_1 + \cdots + t_n\mathbf{e}_n$. Hence we obtain $\mathbf{w} = T(\mathbf{v}) = t_1T(\mathbf{e}_1) + \cdots + t_nT(\mathbf{e}_n)$, proving that $\{T(\mathbf{e}_1), \dots, T(\mathbf{e}_n)\}$ spans W.

(2) $\Rightarrow$ (3). This is because V has a basis.

(3) $\Rightarrow$ (1). If $T(\mathbf{v}) = \mathbf{0}$, write $\mathbf{v} = v_1\mathbf{e}_1 + \cdots + v_n\mathbf{e}_n$ where each v_i is in $\mathbb{R}$. Then

$$\mathbf{0} = T(\mathbf{v}) = v_1T(\mathbf{e}_1) + \cdots + v_nT(\mathbf{e}_n)$$

so $v_1 = \cdots = v_n = 0$ by (3). Hence $\mathbf{v} = \mathbf{0}$, so $\ker T = \{\mathbf{0}\}$ and T is one-to-one. To show that T is onto, let $\mathbf{w}$ be any vector in W. By (3) there exist $w_1, \dots, w_n$ in $\mathbb{R}$ such that

$$\mathbf{w} = w_1T(\mathbf{e}_1) + \cdots + w_nT(\mathbf{e}_n) = T(w_1\mathbf{e}_1 + \cdots + w_n\mathbf{e}_n)$$

Thus T is onto. □

Theorem 7.3.1 dovetails nicely with Theorem 7.1.3 as follows. Let V and W be vector spaces of dimension n, and suppose that $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ and $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ are bases of V and W, respectively. Theorem 7.1.3 asserts that there exists a linear transformation $T : V \to W$ such that

$$T(\mathbf{e}_i) = \mathbf{f}_i \quad \text{for each } i = 1, 2, \dots, n$$

Then $\{T(\mathbf{e}_1),\ \dots,\ T(\mathbf{e}_n)\}$ is evidently a basis of W, so T is an isomorphism by Theorem 7.3.1. Furthermore, the action of T is prescribed by

$$T(r_1\mathbf{e}_1 + \cdots + r_n\mathbf{e}_n) = r_1\mathbf{f}_1 + \cdots + r_n\mathbf{f}_n$$

so isomorphisms between spaces of equal dimension can be easily defined as soon as bases are known. In particular, this shows that if two vector spaces V and W have the same dimension then they are isomorphic, that is $V \cong W$. This is half of the following theorem.

Theorem 7.3.2

If V and W are finite dimensional vector spaces, then $V \cong W$ if and only if $\dim V = \dim W$.

Proof. It remains to show that if $V \cong W$ then $\dim V = \dim W$. But if $V \cong W$, then there exists an isomorphism $T : V \to W$. Since V is finite dimensional, let $\{\mathbf{e}_1,\ \dots,\ \mathbf{e}_n\}$ be a basis of V. Then $\{T(\mathbf{e}_1),\ \dots,\ T(\mathbf{e}_n)\}$ is a basis of W by Theorem 7.3.1, so $\dim W = n = \dim V$. □

Corollary 7.3.1

Let U, V, and W denote vector spaces. Then:

1. *$V \cong V$ for every vector space V.*
2. *If $V \cong W$ then $W \cong V$.*
3. *If $U \cong V$ and $V \cong W$, then $U \cong W$.*

The proof is left to the reader. By virtue of these properties, the relation $\cong$ is called an *equivalence relation* on the class of finite dimensional vector spaces. Since $\dim(\mathbb{R}^n) = n$ it follows that

Corollary 7.3.2

If V is a vector space and $\dim V = n$, then V is isomorphic to $\mathbb{R}^n$.

If V is a vector space of dimension n, note that there are important explicit isomorphisms $V \to \mathbb{R}^n$. Fix a basis $B = \{\mathbf{b}_1,\ \mathbf{b}_2,\ \dots,\ \mathbf{b}_n\}$ of V and write $\{\mathbf{e}_1,\ \mathbf{e}_2,\ \dots,\ \mathbf{e}_n\}$ for the standard basis of $\mathbb{R}^n$. By Theorem 7.1.3 there is a unique linear transformation $C_B : V \to \mathbb{R}^n$ given by

$$C_B(v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + \cdots + v_n\mathbf{b}_n) = v_1\mathbf{e}_1 + v_2\mathbf{e}_2 + \cdots + v_n\mathbf{e}_n = \left[\begin{array}{c} v_1 \\ v_2 \\ \vdots \\ v_n \end{array}\right]$$

where each v_i is in $\mathbb{R}$. Moreover, $C_B(\mathbf{b}_i) = \mathbf{e}_i$ for each i so C_B is an isomorphism by Theorem 7.3.1, called the **coordinate isomorphism** corresponding to the basis B. These isomorphisms will play a central role in Chapter 9.

The conclusion in the above corollary can be phrased as follows: As far as vector space properties are concerned, every n-dimensional vector space V is essentially the same as $\mathbb{R}^n$; they are the "same" vector space except for a change of symbols. This appears to make the process of abstraction seem less important—just study $\mathbb{R}^n$ and be done with it! But consider the different "feel" of the spaces $\mathbf{P}_8$ and $\mathbf{M}_{33}$ even though they are both the "same" as $\mathbb{R}^9$: For example, vectors in $\mathbf{P}_8$ can have roots, while vectors in $\mathbf{M}_{33}$ can be multiplied. So the merit in the abstraction process lies in identifying *common* properties of the vector spaces in the various examples. This is important even for finite dimensional spaces. However, the payoff from abstraction is much greater in the infinite dimensional case, particularly for spaces of functions.

Example 7.3.4

Let V denote the space of all 2×2 symmetric matrices. Find an isomorphism $T : \mathbf{P}_2 \to V$ such that $T(1) = I$, where I is the 2×2 identity matrix.

Solution. $\{1, x, x^2\}$ is a basis of $\mathbf{P}_2$, and we want a basis of V containing I. The set $\left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}$ is independent in V, so it is a basis because $\dim V = 3$ (by Example 6.3.11). Hence define $T : \mathbf{P}_2 \to V$ by taking $T(1) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $T(x) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $T(x^2) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$, and extending linearly as in Theorem 7.1.3. Then T is an isomorphism by Theorem 7.3.1, and its action is given by

$$T(a + bx + cx^2) = aT(1) + bT(x) + cT(x^2) = \begin{bmatrix} a & b \\ b & a+c \end{bmatrix}$$

The dimension theorem (Theorem 7.2.4) gives the following useful fact about isomorphisms.

Theorem 7.3.3

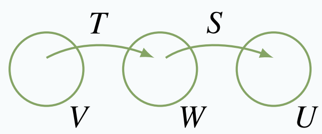
If V and W have the same dimension n, a linear transformation $T : V \to W$ is an isomorphism if it is either one-to-one or onto.

Proof. The dimension theorem asserts that $\dim(\ker T) + \dim(\operatorname{im} T) = n$, so $\dim(\ker T) = 0$ if and only if $\dim(\operatorname{im} T) = n$. Thus T is one-to-one if and only if T is onto, and the result follows. □

Composition

Suppose that $T : V \to W$ and $S : W \to U$ are linear transformations. They link together as in the diagram so, as in Section 2.3, it is possible to define a new function $V \to U$ by first applying T and then S.

Definition 7.5 Composition of Linear Transformations



Given linear transformations $V \xrightarrow{T} W \xrightarrow{S} U$, the ***composite*** *$ST : V \to U$ of T and S is defined by*

$$ST(\mathbf{v}) = S[T(\mathbf{v})] \quad \text{for all } \mathbf{v} \text{ in } V$$

The operation of forming the new function ST is called ***composition****.*[1]

The action of ST can be described compactly as follows: ST means first T then S.

Not all pairs of linear transformations can be composed. For example, if $T : V \to W$ and $S : W \to U$ are linear transformations then $ST : V \to U$ is defined, but TS *cannot* be formed unless $U = V$ because $S : W \to U$ and $T : V \to W$ do not "link" in that order.[2]

Moreover, even if ST and TS can both be formed, they may not be equal. In fact, if $S : \mathbb{R}^m \to \mathbb{R}^n$ and $T : \mathbb{R}^n \to \mathbb{R}^m$ are induced by matrices A and B respectively, then ST and TS can both be formed (they are induced by AB and BA respectively), but the matrix products AB and BA may not be equal (they may not even be the same *size*). Here is another example.

Example 7.3.5

Define: $S : \mathbf{M}_{22} \to \mathbf{M}_{22}$ and $T : \mathbf{M}_{22} \to \mathbf{M}_{22}$ by $S\left[\begin{array}{cc} a & b \\ c & d \end{array}\right] = \left[\begin{array}{cc} c & d \\ a & b \end{array}\right]$ and $T(A) = A^T$ for $A \in \mathbf{M}_{22}$. Describe the action of ST and TS, and show that $ST \neq TS$.

Solution. $ST\left[\begin{array}{cc} a & b \\ c & d \end{array}\right] = S\left[\begin{array}{cc} a & c \\ b & d \end{array}\right] = \left[\begin{array}{cc} b & d \\ a & c \end{array}\right]$, whereas

$TS\left[\begin{array}{cc} a & b \\ c & d \end{array}\right] = T\left[\begin{array}{cc} c & d \\ a & b \end{array}\right] = \left[\begin{array}{cc} c & a \\ d & b \end{array}\right]$.

It is clear that $TS\left[\begin{array}{cc} a & b \\ c & d \end{array}\right]$ need not equal $ST\left[\begin{array}{cc} a & b \\ c & d \end{array}\right]$, so $TS \neq ST$.

The next theorem collects some basic properties of the composition operation.

Theorem 7.3.4: [3]

Let $V \xrightarrow{T} W \xrightarrow{S} U \xrightarrow{R} Z$ be linear transformations.

1. *The composite ST is again a linear transformation.*
2. *$T1_V = T$ and $1_W T = T$.*
3. *$(RS)T = R(ST)$.*

[1] In Section 2.3 we denoted the composite as $S \circ T$. However, it is more convenient to use the simpler notation ST.
[2] Actually, all that is required is $U \subseteq V$.

Proof. The proofs of (1) and (2) are left as Exercise **??**. To prove (3), observe that, for all $\mathbf{v}$ in V:

$$\{(RS)T\}(\mathbf{v}) = (RS)[T(\mathbf{v})] = R\{S[T(\mathbf{v})]\} = R\{(ST)(\mathbf{v})\} = \{R(ST)\}(\mathbf{v})$$

□

Up to this point, composition seems to have no connection with isomorphisms. In fact, the two notions are closely related.

Theorem 7.3.5

Let V and W be finite dimensional vector spaces. The following conditions are equivalent for a linear transformation $T : V \to W$.

1. *T is an isomorphism.*
2. *There exists a linear transformation $S : W \to V$ such that $ST = 1_V$ and $TS = 1_W$.*

Moreover, in this case S is also an isomorphism and is uniquely determined by T:

$$\textit{If } \mathbf{w} \textit{ in } W \textit{ is written as } \mathbf{w} = T(\mathbf{v}), \textit{ then } S(\mathbf{w}) = \mathbf{v}.$$

Proof. (1) $\Rightarrow$ (2). If $B = \{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is a basis of V, then $D = \{T(\mathbf{e}_1), \dots, T(\mathbf{e}_n)\}$ is a basis of W by Theorem 7.3.1. Hence (using Theorem 7.1.3), define a linear transformation $S : W \to V$ by

$$S[T(\mathbf{e}_i)] = \mathbf{e}_i \quad \text{for each } i \tag{7.2}$$

Since $\mathbf{e}_i = 1_V(\mathbf{e}_i)$, this gives $ST = 1_V$ by Theorem 7.1.2. But applying T gives $T[S[T(\mathbf{e}_i)]] = T(\mathbf{e}_i)$ for each i, so $TS = 1_W$ (again by Theorem 7.1.2, using the basis D of W).

(2) $\Rightarrow$ (1). If $T(\mathbf{v}) = T(\mathbf{v}_1)$, then $S[T(\mathbf{v})] = S[T(\mathbf{v}_1)]$. Because $ST = 1_V$ by (2), this reads $\mathbf{v} = \mathbf{v}_1$; that is, T is one-to-one. Given $\mathbf{w}$ in W, the fact that $TS = 1_W$ means that $\mathbf{w} = T[S(\mathbf{w})]$, so T is onto.

Finally, S is uniquely determined by the condition $ST = 1_V$ because this condition implies (7.2). S is an isomorphism because it carries the basis D to B. As to the last assertion, given $\mathbf{w}$ in W, write $\mathbf{w} = r_1T(\mathbf{e}_1) + \cdots + r_nT(\mathbf{e}_n)$. Then $\mathbf{w} = T(\mathbf{v})$, where $\mathbf{v} = r_1\mathbf{e}_1 + \cdots + r_n\mathbf{e}_n$. Then $S(\mathbf{w}) = \mathbf{v}$ by (7.2). □

Given an isomorphism $T : V \to W$, the unique isomorphism $S : W \to V$ satisfying condition (2) of Theorem 7.3.5 is called the **inverse** of T and is denoted by T^{-1}. Hence $T : V \to W$ and $T^{-1} : W \to V$ are related by the **fundamental identities**:

$$T^{-1}[T(\mathbf{v})] = \mathbf{v} \text{ for all } \mathbf{v} \text{ in } V \quad \text{ and } \quad T\left[T^{-1}(\mathbf{w})\right] = \mathbf{w} \text{ for all } \mathbf{w} \text{ in } W$$

In other words, each of T and T^{-1} reverses the action of the other. In particular, equation (7.2) in the proof of Theorem 7.3.5 shows how to define T^{-1} using the image of a basis under the isomorphism T. Here is an example.

[3] Theorem 7.3.4 can be expressed by saying that vector spaces and linear transformations are an example of a category. In general a category consists of certain objects and, for any two objects X and Y, a set $\text{mor}(X, Y)$. The elements α of $\text{mor}(X, Y)$ are called morphisms from X to Y and are written $\alpha : X \to Y$. It is assumed that identity morphisms and composition are defined in such a way that Theorem 7.3.4 holds. Hence, in the category of vector spaces the objects are the vector spaces themselves and the morphisms are the linear transformations. Another example is the category of metric spaces, in which the objects are sets equipped with a distance function (called a metric), and the morphisms are continuous functions (with respect to the metric). The category of sets and functions is a very basic example.

Example 7.3.6

Define $T : \mathbf{P}_1 \to \mathbf{P}_1$ by $T(a+bx) = (a-b)+ax$. Show that T has an inverse, and find the action of T^{-1}.

Solution. The transformation T is linear (verify). Because $T(1) = 1+x$ and $T(x) = -1$, T carries the basis $B = \{1, x\}$ to the basis $D = \{1+x, -1\}$. Hence T is an isomorphism, and T^{-1} carries D back to B, that is,

$$T^{-1}(1+x) = 1 \quad \text{and} \quad T^{-1}(-1) = x$$

Because $a+bx = b(1+x)+(b-a)(-1)$, we obtain

$$T^{-1}(a+bx) = bT^{-1}(1+x)+(b-a)T^{-1}(-1) = b+(b-a)x$$

Sometimes the action of the inverse of a transformation is apparent.

Example 7.3.7

If $B = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ is a basis of a vector space V, the coordinate transformation $C_B : V \to \mathbb{R}^n$ is an isomorphism defined by

$$C_B(v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + \cdots + v_n\mathbf{b}_n) = (v_1, v_2, \dots, v_n)^T$$

The way to reverse the action of C_B is clear: $C_B^{-1} : \mathbb{R}^n \to V$ is given by

$$C_B^{-1}(v_1, v_2, \dots, v_n) = v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + \cdots + v_n\mathbf{b}_n \quad \text{for all } v_i \text{ in } V$$

Condition (2) in Theorem 7.3.5 characterizes the inverse of a linear transformation $T : V \to W$ as the (unique) transformation $S : W \to V$ that satisfies $ST = 1_V$ and $TS = 1_W$. This often determines the inverse.

Example 7.3.8

Define $T : \mathbb{R}^3 \to \mathbb{R}^3$ by $T(x, y, z) = (z, x, y)$. Show that $T^3 = 1_{\mathbb{R}^3}$, and hence find T^{-1}.

Solution. $T^2(x, y, z) = T[T(x, y, z)] = T(z, x, y) = (y, z, x)$. Hence

$$T^3(x, y, z) = T\left[T^2(x, y, z)\right] = T(y, z, x) = (x, y, z)$$

Since this holds for all (x, y, z), it shows that $T^3 = 1_{\mathbb{R}^3}$, so $T(T^2) = 1_{\mathbb{R}^3} = (T^2)T$. Thus $T^{-1} = T^2$ by (2) of Theorem 7.3.5.

Example 7.3.9

Define $T : \mathbf{P}_n \to \mathbb{R}^{n+1}$ by $T(p) = (p(0), p(1), \dots, p(n))$ for all p in $\mathbf{P}_n$. Show that T^{-1} exists.

Solution. The verification that T is linear is left to the reader. If $T(p) = 0$, then $p(k) = 0$ for $k = 0, 1, \dots, n$, so p has $n+1$ distinct roots. Because p has degree at most n, this implies that

$p = 0$ is the zero polynomial (Theorem 6.5.4) and hence that T is one-to-one. But $\dim \mathbf{P}_n = n+1 = \dim \mathbb{R}^{n+1}$, so this means that T is also onto and hence is an isomorphism. Thus T^{-1} exists by Theorem 7.3.5. Note that we have not given a description of the action of T^{-1}, we have merely shown that such a description exists. To give it explicitly requires some ingenuity; one method involves the Lagrange interpolation expansion (Theorem 6.5.3).

7.4 A Theorem about Differential Equations

Differential equations are instrumental in solving a variety of problems throughout science, social science, and engineering. In this brief section, we will see that the set of solutions of a linear differential equation (with constant coefficients) is a vector space and we will calculate its dimension. The proof is pure linear algebra, although the applications are primarily in analysis. However, a key result (Lemma 7.4.3 below) can be applied much more widely.

We denote the derivative of a function $f : \mathbb{R} \to \mathbb{R}$ by f', and f will be called **differentiable** if it can be differentiated any number of times. If f is a differentiable function, the nth derivative $f^{(n)}$ of f is the result of differentiating n times. Thus $f^{(0)} = f,\ f^{(1)} = f',\ f^{(2)} = f^{(1)\prime}, \dots$, and in general $f^{(n+1)} = f^{(n)\prime}$ for each $n \geq 0$. For small values of n these are often written as $f,\ f',\ f'',\ f''',\ \dots$.

If a, b, and c are numbers, the differential equations

$$f'' - af' - bf = 0 \quad \text{or} \quad f''' - af'' - bf' - cf = 0$$

are said to be of **second order** and **third-order**, respectively. In general, an equation

$$f^{(n)} - a_{n-1}f^{(n-1)} - a_{n-2}f^{(n-2)} - \cdots - a_2 f^{(2)} - a_1 f^{(1)} - a_0 f^{(0)} = 0, \; a_i \text{ in } \mathbb{R} \tag{7.3}$$

is called a **differential equation of order** n. We want to describe all solutions of this equation. Of course a knowledge of calculus is required.

The set $\mathbf{F}$ of all functions $\mathbb{R} \to \mathbb{R}$ is a vector space with operations as described in Example 6.1.7. If f and g are differentiable, we have $(f+g)' = f' + g'$ and $(af)' = af'$ for all a in $\mathbb{R}$. With this it is a routine matter to verify that the following set is a subspace of $\mathbf{F}$:

$$\mathbf{D}_n = \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ is differentiable and is a solution to (7.3)}\}$$

Our sole objective in this section is to prove

Theorem 7.4.1

The space $\mathbf{D}_n$ has dimension n.

As will be clear later, the proof of Theorem 7.4.1 requires that we enlarge $\mathbf{D}_n$ somewhat and allow our differentiable functions to take values in the set $\mathbb{C}$ of complex numbers. To do this, we must clarify what it means for a function $f : \mathbb{R} \to \mathbb{C}$ to be differentiable. For each real number x write $f(x)$ in terms of its real and imaginary parts $f_r(x)$ and $f_i(x)$:

$$f(x) = f_r(x) + i f_i(x)$$

This produces new functions $f_r : \mathbb{R} \to \mathbb{R}$ and $f_i : \mathbb{R} \to \mathbb{R}$, called the **real** and **imaginary parts** of f, respectively. We say that f is **differentiable** if both f_r and f_i are differentiable (as real functions), and we define the **derivative** f' of f by

$$f' = f_r' + i f_i' \tag{7.4}$$

We refer to this frequently in what follows.[4]

With this, write $\mathbf{D}_\infty$ for the set of all differentiable complex valued functions $f : \mathbb{R} \to \mathbb{C}$. This is a *complex* vector space using pointwise addition (see Example 6.1.7), and the following scalar multiplication: For any w in $\mathbb{C}$ and f in $\mathbf{D}_\infty$, we define $wf : \mathbb{R} \to \mathbb{C}$ by $(wf)(x) = wf(x)$ for all x in $\mathbb{R}$. We will be working in $\mathbf{D}_\infty$ for the rest of this section. In particular, consider the following complex subspace of $\mathbf{D}_\infty$:

$$\mathbf{D}_n^* = \{f : \mathbb{R} \to \mathbb{C} \mid f \text{ is a solution to (7.3)}\}$$

Clearly, $\mathbf{D}_n \subseteq \mathbf{D}_n^*$, and our interest in $\mathbf{D}_n^*$ comes from

Lemma 7.4.1

If $\dim_{\mathbb{C}}(\mathbf{D}_n^*) = n$, *then* $\dim_{\mathbb{R}}(\mathbf{D}_n) = n$.

[4] Write $|w|$ for the absolute value of any complex number w. As for functions $\mathbb{R} \to \mathbb{R}$, we say that $\lim_{t\to 0} f(t) = w$ if, for all $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(t) - w| <\in$ whenever $|t| < \delta$. (Note that t represents a real number here.) In particular, given a real number x, we define the *derivative* f' of a function $f : \mathbb{R} \to \mathbb{C}$ by $f'(x) = \lim_{t\to 0}\left\{\frac{1}{t}[f(x+t) - f(x)]\right\}$ and we say that f is *differentiable* if $f'(x)$ exists for all x in $\mathbb{R}$. Then we can *prove* that f is differentiable if and only if both f_r and f_i are differentiable, and that $f' = f_r' + i f_i'$ in this case.

Proof. Observe first that if $\dim_{\mathbb{C}}(\mathbf{D}_n^*) = n$, then $\dim_{\mathbb{R}}(\mathbf{D}_n^*) = 2n$. [In fact, if $\{g_1, \ldots, g_n\}$ is a $\mathbb{C}$-basis of $\mathbf{D}_n^*$ then $\{g_1, \ldots, g_n, ig_1, \ldots, ig_n\}$ is a $\mathbb{R}$-basis of $\mathbf{D}_n^*$]. Now observe that the set $\mathbf{D}_n \times \mathbf{D}_n$ of all ordered pairs (f, g) with f and g in $\mathbf{D}_n$ is a real vector space with componentwise operations. Define

$$\theta : \mathbf{D}_n^* \to \mathbf{D}_n \times \mathbf{D}_n \quad \text{given by} \quad \theta(f) = (f_r, f_i) \text{ for } f \text{ in } \mathbf{D}_n^*$$

One verifies that θ is onto and one-to-one, and it is $\mathbb{R}$-linear because $f \to f_r$ and $f \to f_i$ are both $\mathbb{R}$-linear. Hence $\mathbf{D}_n^* \cong \mathbf{D}_n \times \mathbf{D}_n$ as $\mathbb{R}$-spaces. Since $\dim_{\mathbb{R}}(\mathbf{D}_n^*)$ is finite, it follows that $\dim_{\mathbb{R}}(\mathbf{D}_n)$ is finite, and we have

$$2\dim_{\mathbb{R}}(\mathbf{D}_n) = \dim_{\mathbb{R}}(\mathbf{D}_n \times \mathbf{D}_n) = \dim_{\mathbb{R}}(\mathbf{D}_n^*) = 2n$$

Hence $\dim_{\mathbb{R}}(\mathbf{D}_n) = n$, as required. □

It follows that to prove Theorem 7.4.1 it suffices to show that $\dim_{\mathbb{C}}(\mathbf{D}_n^*) = n$.

There is one function that arises frequently in any discussion of differential equations. Given a complex number $w = a + ib$ (where a and b are real), we have $e^w = e^a(\cos b + i \sin b)$. The law of exponents, $e^w e^v = e^{w+v}$ for all w, v in $\mathbb{C}$ is easily verified using the formulas for $\sin(b + b_1)$ and $\cos(b + b_1)$. If x is a variable and $w = a + ib$ is a complex number, define the **exponential function** e^{wx} by

$$e^{wx} = e^{ax}(\cos bx + i \sin bx)$$

Hence e^{wx} is differentiable because its real and imaginary parts are differentiable for all x. Moreover, the following can be proved using (7.4):

$$(e^{wx})' = we^{wx}$$

In addition, (7.4) gives the **product rule** for differentiation:

$$\text{If } f \text{ and } g \text{ are in } \mathbf{D}_\infty, \ \text{ then } (fg)' = f'g + fg'$$

We omit the verifications.

To prove that $\dim_{\mathbb{C}}(\mathbf{D}_n^*) = n$, two preliminary results are required. Here is the first.

Lemma 7.4.2

Given f in $\mathbf{D}_\infty$ and w in $\mathbb{C}$, there exists g in $\mathbf{D}_\infty$ such that $g' - wg = f$.

Proof. Define $p(x) = f(x)e^{-wx}$. Then p is differentiable, whence p_r and p_i are both differentiable, hence continuous, and so both have antiderivatives, say $p_r = q_r'$ and $p_i = q_i'$. Then the function $q = q_r + iq_i$ is in $\mathbf{D}_\infty$, and $q' = p$ by (7.4). Finally define $g(x) = q(x)e^{wx}$. Then

$$g' = q'e^{wx} + qwe^{wx} = pe^{wx} + w(qe^{wx}) = f + wg$$

by the product rule, as required. □

The second preliminary result is important in its own right.

Lemma 7.4.3: Kernel Lemma

Let V be a vector space, and let S and T be linear operators $V \to V$. If S is onto and both $\ker(S)$ and $\ker(T)$ are finite dimensional, then $\ker(TS)$ is also finite dimensional and $\dim[\ker(TS)] = \dim[\ker(T)] + \dim[\ker(S)]$.

Proof. Let $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m\}$ be a basis of $\ker(T)$ and let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ be a basis of $\ker(S)$. Since S is onto, let $\mathbf{u}_i = S(\mathbf{w}_i)$ for some $\mathbf{w}_i$ in V. It suffices to show that

$$B = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m, \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$$

is a basis of $\ker(TS)$. Note $B \subseteq \ker(TS)$ because $TS(\mathbf{w}_i) = T(\mathbf{u}_i) = \mathbf{0}$ for each i and $TS(\mathbf{v}_j) = T(\mathbf{0}) = \mathbf{0}$ for each j.

Spanning. If $\mathbf{v}$ is in $\ker(TS)$, then $S(\mathbf{v})$ is in $\ker(T)$, say $S(\mathbf{v}) = \sum r_i\mathbf{u}_i = \sum r_i S(\mathbf{w}_i) = S(\sum r_i\mathbf{w}_i)$. It follows that $\mathbf{v} - \sum r_i\mathbf{w}_i$ is in $\ker(S) = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$, proving that $\mathbf{v}$ is in $\text{span}(B)$.

Independence. Let $\sum r_i\mathbf{w}_i + \sum t_j\mathbf{v}_j = \mathbf{0}$. Applying S, and noting that $S(\mathbf{v}_j) = \mathbf{0}$ for each j, yields $\mathbf{0} = \sum r_i S(\mathbf{w}_i) = \sum r_i\mathbf{u}_i$. Hence $r_i = 0$ for each i, and so $\sum t_j\mathbf{v}_j = \mathbf{0}$. This implies that each $t_j = 0$, and so proves the independence of B. □

Proof of Theorem 7.4.1. By Lemma 7.4.1, it suffices to prove that $\dim_{\mathbb{C}}(\mathbf{D}_n^*) = n$. This holds for $n = 1$ because the proof of Theorem 3.7.1 goes through to show that $\mathbf{D}_1^* = \mathbb{C}e^{a_0x}$. Hence we proceed by induction on n. With an eye on equation (7.3), consider the polynomial

$$p(t) = t^n - a_{n-1}t^{n-1} - a_{n-2}t^{n-2} - \cdots - a_2t^2 - a_1t - a_0$$

(called the *characteristic polynomial* of equation (7.3)). Now define a map $D : \mathbf{D}_\infty \to \mathbf{D}_\infty$ by $D(f) = f'$ for all f in $\mathbf{D}_\infty$. Then D is a linear operator, whence $p(D) : \mathbf{D}_\infty \to \mathbf{D}_\infty$ is also a linear operator. Moreover, since $D^k(f) = f^{(k)}$ for each $k \geq 0$, equation (7.3) takes the form $p(D)(f) = 0$. In other words,

$$\mathbf{D}_n^* = \ker[p(D)]$$

By the fundamental theorem of algebra,[5] let w be a complex root of $p(t)$, so that $p(t) = q(t)(t-w)$ for some complex polynomial $q(t)$ of degree $n-1$. It follows that $p(D) = q(D)(D - w1_{\mathbf{D}_\infty})$. Moreover $D - w1_{\mathbf{D}_\infty}$ is onto by Lemma 7.4.2, $\dim_{\mathbb{C}}[\ker(D - w1_{\mathbf{D}_\infty})] = 1$ by the case $n = 1$ above, and $\dim_{\mathbb{C}}(\ker[q(D)]) = n - 1$ by induction. Hence Lemma 7.4.3 shows that $\ker[P(D)]$ is also finite dimensional and

$$\dim_{\mathbb{C}}(\ker[p(D)]) = \dim_{\mathbb{C}}(\ker[q(D)]) + \dim_{\mathbb{C}}(\ker[D - w1_{\mathbf{D}_\infty}]) = (n-1) + 1 = n.$$

Since $\mathbf{D}_n^* = \ker[p(D)]$, this completes the induction, and so proves Theorem 7.4.1. □

[5] This is the reason for allowing our solutions to (7.3) to be *complex* valued.

7.5 More on Linear Recurrences[6]

In Section 3.4 we used diagonalization to study linear recurrences, and gave several examples. We now apply the theory of vector spaces and linear transformations to study the problem in more generality.

Consider the linear recurrence

$$x_{n+2} = 6x_n - x_{n+1} \quad \text{for } n \geq 0$$

If the initial values x_0 and x_1 are prescribed, this gives a sequence of numbers. For example, if $x_0 = 1$ and $x_1 = 1$ the sequence continues

$$x_2 = 5,\ x_3 = 1,\ x_4 = 29,\ x_5 = -23,\ x_6 = 197,\ \dots$$

as the reader can verify. Clearly, the entire sequence is uniquely determined by the recurrence and the two initial values. In this section we define a vector space structure on the set of *all* sequences, and study the subspace of those sequences that satisfy a particular recurrence.

Sequences will be considered entities in their own right, so it is useful to have a special notation for them. Let

$$[x_n) \quad \text{denote the sequence } x_0,\ x_1,\ x_2,\ \dots,\ x_n,\ \dots$$

[6]This section requires only Sections 7.1-7.3.

Example 7.5.1

$[n)$	is the sequence $0,\ 1,\ 2,\ 3,\ \dots$
$[n+1)$	is the sequence $1,\ 2,\ 3,\ 4,\ \dots$
$[2^n)$	is the sequence $1,\ 2,\ 2^2,\ 2^3,\ \dots$
$[(-1)^n)$	is the sequence $1,\ -1,\ 1,\ -1,\ \dots$
$[5)$	is the sequence $5,\ 5,\ 5,\ 5,\ \dots$

Sequences of the form $[c)$ for a fixed number c will be referred to as **constant sequences**, and those of the form $[\lambda^n)$, λ some number, are **power sequences**.

Two sequences are regarded as **equal** when they are identical:

$$[x_n) = [y_n) \quad \text{means} \quad x_n = y_n \quad \text{for all } n = 0,\ 1,\ 2,\ \dots$$

Addition and scalar multiplication of sequences are defined by

$$\begin{aligned} [x_n) + [y_n) &= [x_n + y_n) \\ r[x_n) &= [rx_n) \end{aligned}$$

These operations are analogous to the addition and scalar multiplication in $\mathbb{R}^n$, and it is easy to check that the vector-space axioms are satisfied. The zero vector is the constant sequence $[0)$, and the negative of a sequence $[x_n)$ is given by $-[x_n) = [-x_n)$.

Now suppose k real numbers $r_0,\ r_1,\ \dots,\ r_{k-1}$ are given, and consider the **linear recurrence relation** determined by these numbers.

$$x_{n+k} = r_0 x_n + r_1 x_{n+1} + \cdots + r_{k-1} x_{n+k-1} \tag{7.5}$$

When $r_0 \neq 0$, we say this recurrence has **length** k.[7] For example, the relation $x_{n+2} = 2x_n + x_{n+1}$ is of length 2.

A sequence $[x_n)$ is said to **satisfy** the relation (7.5) if (7.5) holds for all $n \geq 0$. Let V denote the set of all sequences that satisfy the relation. In symbols,

$$V = \{[x_n) \mid x_{n+k} = r_0 x_n + r_1 x_{n+1} + \cdots + r_{k-1} x_{n+k-1} \text{ hold for all } n \geq 0\}$$

It is easy to see that the constant sequence $[0)$ lies in V and that V is closed under addition and scalar multiplication of sequences. Hence V is vector space (being a subspace of the space of all sequences). The following important observation about V is needed (it was used implicitly earlier): If the first k terms of two sequences agree, then the sequences are identical. More formally,

Lemma 7.5.1

Let $[x_n)$ and $[y_n)$ denote two sequences in V. Then

$$[x_n) = [y_n) \quad \textit{if and only if} \quad x_0 = y_0,\ \ x_1 = y_1,\ \dots,\ x_{k-1} = y_{k-1}$$

[7] We shall usually assume that $r_0 \neq 0$; otherwise, we are essentially dealing with a recurrence of shorter length than k.

Proof. If $[x_n) = [y_n)$ then $x_n = y_n$ for *all* $n = 0, 1, 2, \dots$. Conversely, if $x_i = y_i$ for all $i = 0, 1, \dots, k-1$, use the recurrence (7.5) for $n = 0$.

$$x_k = r_0x_0 + r_1x_1 + \cdots + r_{k-1}x_{k-1} = r_0y_0 + r_1y_1 + \cdots + r_{k-1}y_{k-1} = y_k$$

Next the recurrence for $n = 1$ establishes $x_{k+1} = y_{k+1}$. The process continues to show that $x_{n+k} = y_{n+k}$ holds for *all* $n \geq 0$ by induction on n. Hence $[x_n) = [y_n)$. □

This shows that a sequence in V is completely determined by its first k terms. In particular, given a k-tuple $\mathbf{v} = (v_0, v_1, \dots, v_{k-1})$ in $\mathbb{R}^k$, define

$$T(\mathbf{v}) \text{ to be the sequence in } V \text{ whose first } k \text{ terms are } v_0, v_1, \dots, v_{k-1}$$

The rest of the sequence $T(\mathbf{v})$ is determined by the recurrence, so $T : \mathbb{R}^k \to V$ is a function. In fact, it is an isomorphism.

Theorem 7.5.1

Given real numbers $r_0, r_1, \dots, r_{k-1}$, let

$$V = \{[x_n) \mid x_{n+k} = r_0x_n + r_1x_{n+1} + \cdots + r_{k-1}x_{n+k-1}, \quad \text{for all } n \geq 0\}$$

denote the vector space of all sequences satisfying the linear recurrence relation (7.5) determined by $r_0, r_1, \dots, r_{k-1}$. Then the function

$$T : \mathbb{R}^k \to V$$

defined above is an isomorphism. In particular:

1. $\dim V = k$.
2. *If $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ is any basis of $\mathbb{R}^k$, then $\{T(\mathbf{v}_1), \dots, T(\mathbf{v}_k)\}$ is a basis of V.*

Proof. (1) and (2) will follow from Theorem 7.3.1 and Theorem 7.3.2 as soon as we show that T is an isomorphism. Given $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^k$, write $\mathbf{v} = (v_0, v_1, \dots, v_{k-1})$ and $\mathbf{w} = (w_0, w_1, \dots, w_{k-1})$. The first k terms of $T(\mathbf{v})$ and $T(\mathbf{w})$ are $v_0, v_1, \dots, v_{k-1}$ and $w_0, w_1, \dots, w_{k-1}$, respectively, so the first k terms of $T(\mathbf{v}) + T(\mathbf{w})$ are $v_0 + w_0, v_1 + w_1, \dots, v_{k-1} + w_{k-1}$. Because these terms agree with the first k terms of $T(\mathbf{v} + \mathbf{w})$, Lemma 7.5.1 implies that $T(\mathbf{v} + \mathbf{w}) = T(\mathbf{v}) + T(\mathbf{w})$. The proof that $T(r\mathbf{v}) + rT(\mathbf{v})$ is similar, so T is linear.

Now let $[x_n)$ be any sequence in V, and let $\mathbf{v} = (x_0, x_1, \dots, x_{k-1})$. Then the first k terms of $[x_n)$ and $T(\mathbf{v})$ agree, so $T(\mathbf{v}) = [x_n)$. Hence T is onto. Finally, if $T(\mathbf{v}) = [0)$ is the zero sequence, then the first k terms of $T(\mathbf{v})$ are all zero (*all* terms of $T(\mathbf{v})$ are zero!) so $\mathbf{v} = \mathbf{0}$. This means that $\ker T = \{\mathbf{0}\}$, so T is one-to-one. □

Example 7.5.2

Show that the sequences $[1)$, $[n)$, and $[(-1)^n)$ are a basis of the space V of all solutions of the recurrence

$$x_{n+3} = -x_n + x_{n+1} + x_{n+2}$$

Then find the solution satisfying $x_0 = 1$, $x_1 = 2$, $x_2 = 5$.

Solution. The verifications that these sequences satisfy the recurrence (and hence lie in V) are left to the reader. They are a basis because $[1) = T(1, 1, 1)$, $[n) = T(0, 1, 2)$, and $[(-1)^n) = T(1, -1, 1)$; and $\{(1, 1, 1), (0, 1, 2), (1, -1, 1)\}$ is a basis of $\mathbb{R}^3$. Hence the sequence $[x_n)$ in V satisfying $x_0 = 1$, $x_1 = 2$, $x_2 = 5$ is a linear combination of this basis:

$$[x_n) = t_1[1) + t_2[n) + t_3[(-1)^n)$$

The nth term is $x_n = t_1 + nt_2 + (-1)^n t_3$, so taking $n = 0, 1, 2$ gives

$$\begin{aligned} 1 = x_0 &= t_1 + 0 + t_3 \\ 2 = x_1 &= t_1 + t_2 - t_3 \\ 5 = x_2 &= t_1 + 2t_2 + t_3 \end{aligned}$$

This has the solution $t_1 = t_3 = \frac{1}{2}$, $t_2 = 2$, so $x_n = \frac{1}{2} + 2n + \frac{1}{2}(-1)^n$.

This technique clearly works for any linear recurrence of length k: Simply take your favourite basis $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ of $\mathbb{R}^k$—perhaps the standard basis—and compute $T(\mathbf{v}_1), \dots, T(\mathbf{v}_k)$. This is a basis of V all right, but the nth term of $T(\mathbf{v}_i)$ is not usually given as an explicit function of n. (The basis in Example 7.5.2 was carefully chosen so that the nth terms of the three sequences were 1, n, and $(-1)^n$, respectively, each a simple function of n.)

However, it turns out that an explicit basis of V can be given in the general situation. Given the recurrence (7.5) again:

$$x_{n+k} = r_0x_n + r_1x_{n+1} + \cdots + r_{k-1}x_{n+k-1}$$

the idea is to look for numbers λ such that the power sequence $[\lambda^n)$ satisfies (7.5). This happens if and only if

$$\lambda^{n+k} = r_0\lambda^n + r_1\lambda^{n+1} + \cdots + r_{k-1}\lambda^{n+k-1}$$

holds for all $n \geq 0$. This is true just when the case $n = 0$ holds; that is,

$$\lambda^k = r_0 + r_1\lambda + \cdots + r_{k-1}\lambda^{k-1}$$

The polynomial

$$p(x) = x^k - r_{k-1}x^{k-1} - \cdots - r_1x - r_0$$

is called the polynomial **associated** with the linear recurrence (7.5). Thus every root λ of $p(x)$ provides a sequence $[\lambda^n)$ satisfying (7.5). If there are k distinct roots, the power sequences provide a basis. Incidentally, if $\lambda = 0$, the sequence $[\lambda^n)$ is $1, 0, 0, \dots$; that is, we accept the convention that $0^0 = 1$.

Theorem 7.5.2

Let $r_0, r_1, \dots, r_{k-1}$ be real numbers; let

$$V = \{[x_n) \mid x_{n+k} = r_0x_n + r_1x_{n+1} + \cdots + r_{k-1}x_{n+k-1} \text{ for all } n \geq 0\}$$

denote the vector space of all sequences satisfying the linear recurrence relation determined by $r_0, r_1, \dots, r_{k-1}$; and let

$$p(x) = x^k - r_{k-1}x^{k-1} - \cdots - r_1x - r_0$$

denote the polynomial associated with the recurrence relation. Then

1. *$[\lambda^n)$ lies in V if and only if λ is a root of $p(x)$.*
2. *If $\lambda_1, \lambda_2, \dots, \lambda_k$ are distinct real roots of $p(x)$, then $\{[\lambda_1^n), [\lambda_2^n), \dots, [\lambda_k^n)\}$ is a basis of V.*

Proof. It remains to prove (2). But $[\lambda_i^n) = T(\mathbf{v}_i)$ where $\mathbf{v}_i = (1, \lambda_i, \lambda_i^2, \dots, \lambda_i^{k-1})$, so (2) follows by Theorem 7.5.1, provided that $(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$ is a basis of $\mathbb{R}^k$. This is true provided that the matrix with the $\mathbf{v}_i$ as its rows

$$\left[\begin{array}{ccccc} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{k-1} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{k-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_k & \lambda_k^2 & \cdots & \lambda_k^{k-1} \end{array}\right]$$

is invertible. But this is a Vandermonde matrix and so is invertible if the λ_i are distinct (Theorem 3.2.7). This proves (2). □

Example 7.5.3

Find the solution of $x_{n+2} = 2x_n + x_{n+1}$ that satisfies $x_0 = a$, $x_1 = b$.

Solution. The associated polynomial is $p(x) = x^2 - x - 2 = (x-2)(x+1)$. The roots are $\lambda_1 = 2$ and $\lambda_2 = -1$, so the sequences $[2^n)$ and $[(-1)^n)$ are a basis for the space of solutions by Theorem 7.5.2. Hence every solution $[x_n)$ is a linear combination

$$[x_n) = t_1[2^n) + t_2[(-1)^n)$$

This means that $x_n = t_1 2^n + t_2(-1)^n$ holds for $n = 0, 1, 2, \dots$, so (taking $n = 0, 1$) $x_0 = a$ and $x_1 = b$ give

$$\begin{aligned} t_1 + t_2 &= a \\ 2t_1 - t_2 &= b \end{aligned}$$

These are easily solved: $t_1 = \frac{1}{3}(a+b)$ and $t_2 = \frac{1}{3}(2a-b)$, so

$$t_n = \frac{1}{3}\left[(a+b)2^n + (2a-b)(-1)^n\right]$$

The Shift Operator

If $p(x)$ is the polynomial associated with a linear recurrence relation of length k, and if $p(x)$ has k distinct roots $\lambda_1, \lambda_2, \dots, \lambda_k$, then $p(x)$ factors completely:

$$p(x) = (x - \lambda_1)(x - \lambda_2) \cdots (x - \lambda_k)$$

Each root λ_i provides a sequence $[\lambda_i^n)$ satisfying the recurrence, and they are a basis of V by Theorem 7.5.2. In this case, each λ_i has multiplicity 1 as a root of $p(x)$. In general, a root λ has **multiplicity** m if $p(x) = (x - \lambda)^m q(x)$, where $q(\lambda) \neq 0$. In this case, there are fewer than k distinct roots and so fewer than k sequences $[\lambda^n)$ satisfying the recurrence. However, we can still obtain a basis because, if λ has multiplicity m (and $\lambda \neq 0$), it provides m linearly independent sequences that satisfy the recurrence. To prove this, it is convenient to give another way to describe the space V of all sequences satisfying a given linear recurrence relation.

Let $\mathbf{S}$ denote the vector space of *all* sequences and define a function

$$S : \mathbf{S} \to \mathbf{S} \quad \text{by} \quad S[x_n) = [x_{n+1}) = [x_1,\ x_2,\ x_3,\ \dots)$$

S is clearly a linear transformation and is called the **shift operator** on $\mathbf{S}$. Note that powers of S shift the sequence further: $S^2[x_n) = S[x_{n+1}) = [x_{n+2})$. In general,

$$S^k[x_n) = [x_{n+k}) = [x_k,\ x_{k+1},\ \dots) \quad \text{for all } k = 0,\ 1,\ 2,\ \dots$$

But then a linear recurrence relation

$$x_{n+k} = r_0 x_n + r_1 x_{n+1} + \cdots + r_{k-1} x_{n+k-1} \quad \text{for all } n = 0,\ 1,\ \dots$$

can be written

$$S^k[x_n) = r_0[x_n) + r_1 S[x_n) + \cdots + r_{k-1} S^{k-1}[x_n) \tag{7.6}$$

Now let $p(x) = x^k - r_{k-1}x^{k-1} - \cdots - r_1 x - r_0$ denote the polynomial associated with the recurrence relation. The set $\mathbf{L}[\mathbf{S},\ \mathbf{S}]$ of all linear transformations from $\mathbf{S}$ to itself is a vector space (verify[8]) that is closed under composition. In particular,

$$p(S) = S^k - r_{k-1}S^{k-1} - \cdots - r_1 S - r_0$$

is a linear transformation called the **evaluation** of p at S. The point is that condition (7.6) can be written as

$$p(S)\{[x_n)\} = 0$$

In other words, the space V of all sequences satisfying the recurrence relation is just $\ker[p(S)]$. This is the first assertion in the following theorem.

Theorem 7.5.3

Let $r_0,\ r_1,\ \dots,\ r_{k-1}$ be real numbers, and let

$$V = \{[x_n) \mid x_{n+k} = r_0 x_n + r_1 x_{n+1} + \cdots + r_{k-1} x_{n+k-1} \quad \textit{for all } n \geq 0\}$$

denote the space of all sequences satisfying the linear recurrence relation determined by

[8] See Exercises ?? and ??.

$r_0,\ r_1,\ \dots,\ r_{k-1}$. Let

$$p(x) = x^k - r_{k-1}x^{k-1} - \cdots - r_1 x - r_0$$

denote the corresponding polynomial. Then:

1. *$V = \ker[p(S)]$, where S is the shift operator.*
2. *If $p(x) = (x-\lambda)^m q(x)$, where $\lambda \neq 0$ and $m > 1$, then the sequences*

$$\{[\lambda^n),\ [n\lambda^n),\ [n^2\lambda^n),\ \dots,\ [n^{m-1}\lambda^n)\}$$

all lie in V and are linearly independent.

Proof (Sketch). It remains to prove (2). If $\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!}$ denotes the binomial coefficient, the idea is to use (1) to show that the sequence $s_k = \left[\binom{n}{k}\lambda^n\right)$ is a solution for each $k = 0,\ 1,\ \dots,\ m-1$. Then (2) of Theorem 7.5.1 can be applied to show that $\{s_0,\ s_1,\ \dots,\ s_{m-1}\}$ is linearly independent. Finally, the sequences $t_k = [n^k\lambda^n)$, $k = 0,\ 1,\ \dots,\ m-1$, in the present theorem can be given by $t_k = \sum_{j=0}^{m-1} a_{kj}s_j$, where $A = \left[a_{ij}\right]$ is an invertible matrix. Then (2) follows. We omit the details. □

This theorem combines with Theorem 7.5.2 to give a basis for V when $p(x)$ has k real roots (not necessarily distinct) none of which is zero. This last requirement means $r_0 \neq 0$, a condition that is unimportant in practice (see Remark 1 below).

Theorem 7.5.4

Let $r_0,\ r_1,\ \dots,\ r_{k-1}$ be real numbers with $r_0 \neq 0$; let

$$V = \{[x_n) \mid x_{n+k} = r_0x_n + r_1x_{n+1} + \cdots + r_{k-1}x_{n+k-1} \text{ for all } n \geq 0\}$$

denote the space of all sequences satisfying the linear recurrence relation of length k determined by $r_0,\ \dots,\ r_{k-1}$; and assume that the polynomial

$$p(x) = x^k - r_{k-1}x^{k-1} - \cdots - r_1 x - r_0$$

factors completely as

$$p(x) = (x-\lambda_1)^{m_1}(x-\lambda_2)^{m_2}\cdots(x-\lambda_p)^{m_p}$$

where $\lambda_1,\ \lambda_2,\ \dots,\ \lambda_p$ are distinct real numbers and each $m_i \geq 1$. Then $\lambda_i \neq 0$ for each i, and

$$\begin{array}{c}
\left[\lambda_1^n\right),\ \left[n\lambda_1^n\right),\ \dots,\ \left[n^{m_1-1}\lambda_1^n\right) \\
\left[\lambda_2^n\right),\ \left[n\lambda_2^n\right),\ \dots,\ \left[n^{m_2-1}\lambda_2^n\right) \\
\vdots \\
\left[\lambda_p^n\right),\ \left[n\lambda_p^n\right),\ \dots,\ \left[n^{m_p-1}\lambda_p^n\right)
\end{array}$$

is a basis of V.

Proof. There are $m_1 + m_2 + \cdots + m_p = k$ sequences in all so, because $\dim V = k$, it suffices to show that they are linearly independent. The assumption that $r_0 \neq 0$, implies that 0 is not a root of $p(x)$. Hence each

$\lambda_i \neq 0$, so $\{[\lambda_i^n),\ [n\lambda_i^n),\ \dots,\ [n^{m_i-1}\lambda_i^n)\}$ is linearly independent by Theorem 7.5.3. The proof that the whole set of sequences is linearly independent is omitted. □

Example 7.5.4

Find a basis for the space V of all sequences $[x_n)$ satisfying

$$x_{n+3} = -9x_n - 3x_{n+1} + 5x_{n+2}$$

Solution. The associated polynomial is

$$p(x) = x^3 - 5x^2 + 3x + 9 = (x-3)^2(x+1)$$

Hence 3 is a double root, so $[3_n)$ and $[n3^n)$ both lie in V by Theorem 7.5.3 (the reader should verify this). Similarly, $\lambda = -1$ is a root of multiplicity 1, so $[(-1)^n)$ lies in V. Hence $\{[3^n),\ [n3^n),\ [(-1)^n)\}$ is a basis by Theorem 7.5.4.

Remark 1

If $r_0 = 0$ [so $p(x)$ has 0 as a root], the recurrence reduces to one of shorter length. For example, consider

$$x_{n+4} = 0x_n + 0x_{n+1} + 3x_{n+2} + 2x_{n+3} \tag{7.7}$$

If we set $y_n = x_{n+2}$, this recurrence becomes $y_{n+2} = 3y_n + 2y_{n+1}$, which has solutions $[3^n)$ and $[(-1)^n)$. These give the following solution to (7.5):

$$[0,\ 0,\ 1,\ 3,\ 3^2,\ \dots)$$
$$[0,\ 0,\ 1,\ -1,\ (-1)^2,\ \dots)$$

In addition, it is easy to verify that

$$[1,\ 0,\ 0,\ 0,\ 0,\ \dots)$$
$$[0,\ 1,\ 0,\ 0,\ 0,\ \dots)$$

are also solutions to (7.7). The space of all solutions of (7.5) has dimension 4 (Theorem 7.5.1), so these sequences are a basis. This technique works whenever $r_0 = 0$.

Remark 2

Theorem 7.5.4 completely describes the space V of sequences that satisfy a linear recurrence relation for which the associated polynomial $p(x)$ has all real roots. However, in many cases of interest, $p(x)$ has complex roots that are not real. If $p(\mu) = 0$, μ complex, then $p(\overline{\mu}) = 0$ too ($\overline{\mu}$ the conjugate), and the main observation is that $[\mu^n + \overline{\mu}^n)$ and $[i(\mu^n + \overline{\mu}^n))$ are *real* solutions. Analogs of the preceding theorems can then be proved.

Chapter 8

Orthogonality

In Section 5.3 we introduced the dot product in $\mathbb{R}^n$ and extended the basic geometric notions of length and distance. A set $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$ of nonzero vectors in $\mathbb{R}^n$ was called an **orthogonal set** if $\mathbf{f}_i \cdot \mathbf{f}_j = 0$ for all $i \neq j$, and it was proved that every orthogonal set is independent. In particular, it was observed that the expansion of a vector as a linear combination of orthogonal basis vectors is easy to obtain because formulas exist for the coefficients. Hence the orthogonal bases are the "nice" bases, and much of this chapter is devoted to extending results about bases to orthogonal bases. This leads to some very powerful methods and theorems. Our first task is to show that every subspace of $\mathbb{R}^n$ *has* an orthogonal basis.

8.1 Orthogonal Complements and Projections

If $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ is linearly independent in a general vector space, and if $\mathbf{v}_{m+1}$ is not in $\operatorname{span}\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$, then $\{\mathbf{v}_1, \dots, \mathbf{v}_m, \mathbf{v}_{m+1}\}$ is independent (Lemma 6.4.1). Here is the analog for *orthogonal* sets in $\mathbb{R}^n$.

Lemma 8.1.1: Orthogonal Lemma

Let $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$ be an orthogonal set in $\mathbb{R}^n$. Given $\mathbf{x}$ in $\mathbb{R}^n$, write

$$\mathbf{f}_{m+1} = \mathbf{x} - \frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 - \cdots - \frac{\mathbf{x}\cdot\mathbf{f}_m}{\|\mathbf{f}_m\|^2}\mathbf{f}_m$$

Then:

1. $\mathbf{f}_{m+1} \cdot \mathbf{f}_k = 0$ *for* $k = 1, 2, \dots, m$.
2. *If* $\mathbf{x}$ *is not in* $\text{span}\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$, *then* $\mathbf{f}_{m+1} \neq \mathbf{0}$ *and* $\{\mathbf{f}_1, \dots, \mathbf{f}_m, \mathbf{f}_{m+1}\}$ *is an orthogonal set.*

Proof. For convenience, write $t_i = (\mathbf{x} \cdot \mathbf{f}_i)/\|\mathbf{f}_i\|^2$ for each i. Given $1 \leq k \leq m$:

$$\begin{aligned}
\mathbf{f}_{m+1} \cdot \mathbf{f}_k &= (\mathbf{x} - t_1\mathbf{f}_1 - \cdots - t_k\mathbf{f}_k - \cdots - t_m\mathbf{f}_m) \cdot \mathbf{f}_k \\
&= \mathbf{x} \cdot \mathbf{f}_k - t_1(\mathbf{f}_1 \cdot \mathbf{f}_k) - \cdots - t_k(\mathbf{f}_k \cdot \mathbf{f}_k) - \cdots - t_m(\mathbf{f}_m \cdot \mathbf{f}_k) \\
&= \mathbf{x} \cdot \mathbf{f}_k - t_k\|\mathbf{f}_k\|^2 \\
&= 0
\end{aligned}$$

This proves (1), and (2) follows because $\mathbf{f}_{m+1} \neq \mathbf{0}$ if $\mathbf{x}$ is not in $\text{span}\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$. □

The orthogonal lemma has three important consequences for $\mathbb{R}^n$. The first is an extension for orthogonal sets of the fundamental fact that any independent set is part of a basis (Theorem 6.4.1).

Theorem 8.1.1

Let U *be a subspace of* $\mathbb{R}^n$.

1. *Every orthogonal subset* $\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$ *in* U *is a subset of an orthogonal basis of* U.
2. U *has an orthogonal basis.*

Proof.

1. If $\text{span}\{\mathbf{f}_1, \dots, \mathbf{f}_m\} = U$, it is *already* a basis. Otherwise, there exists $\mathbf{x}$ in U outside $\text{span}\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$. If $\mathbf{f}_{m+1}$ is as given in the orthogonal lemma, then $\mathbf{f}_{m+1}$ is in U and $\{\mathbf{f}_1, \dots, \mathbf{f}_m, \mathbf{f}_{m+1}\}$ is orthogonal. If $\text{span}\{\mathbf{f}_1, \dots, \mathbf{f}_m, \mathbf{f}_{m+1}\} = U$, we are done. Otherwise, the process continues to create larger and larger orthogonal subsets of U. They are all independent by Theorem 5.3.5, so we have a basis when we reach a subset containing $\dim U$ vectors.
2. If $U = \{\mathbf{0}\}$, the empty basis is orthogonal. Otherwise, if $\mathbf{f} \neq \mathbf{0}$ is in U, then $\{\mathbf{f}\}$ is orthogonal, so (2) follows from (1). □

We can improve upon (2) of Theorem 8.1.1. In fact, the second consequence of the orthogonal lemma is a procedure by which *any* basis $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ of a subspace U of $\mathbb{R}^n$ can be systematically modified to yield an orthogonal basis $\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$ of U. The $\mathbf{f}_i$ are constructed one at a time from the $\mathbf{x}_i$.

To start the process, take $\mathbf{f}_1 = \mathbf{x}_1$. Then $\mathbf{x}_2$ is not in $\text{span}\{\mathbf{f}_1\}$ because $\{\mathbf{x}_1, \mathbf{x}_2\}$ is independent, so take

$$\mathbf{f}_2 = \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1$$

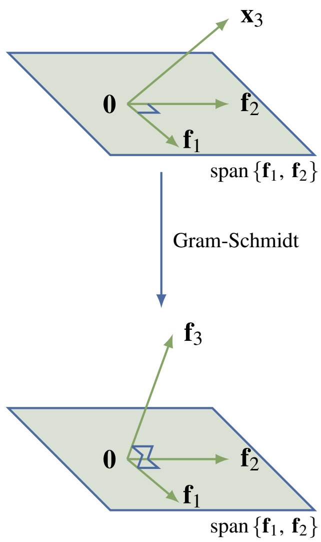
Thus $\{\mathbf{f}_1, \mathbf{f}_2\}$ is orthogonal by Lemma 8.1.1. Moreover, $\text{span}\{\mathbf{f}_1, \mathbf{f}_2\} = \text{span}\{\mathbf{x}_1, \mathbf{x}_2\}$ (verify), so $\mathbf{x}_3$ is not in $\text{span}\{\mathbf{f}_1, \mathbf{f}_2\}$. Hence $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ is orthogonal where

$$\mathbf{f}_3 = \mathbf{x}_3 - \frac{\mathbf{x}_3 \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \frac{\mathbf{x}_3 \cdot \mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2$$

Again, $\operatorname{span}\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\} = \operatorname{span}\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, so $\mathbf{x}_4$ is not in $\operatorname{span}\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ and the process continues. At the mth iteration we construct an orthogonal set $\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$ such that

$$\operatorname{span}\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\} = \operatorname{span}\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\} = U$$

Hence $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$ is the desired orthogonal basis of U. The procedure can be summarized as follows.



Theorem 8.1.2: Gram-Schmidt Orthogonalization Algorithm[1]

If $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ is any basis of a subspace U of $\mathbb{R}^n$, construct $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m$ in U successively as follows:

$$\begin{aligned}
\mathbf{f}_1 &= \mathbf{x}_1 \\
\mathbf{f}_2 &= \mathbf{x}_2 - \tfrac{\mathbf{x}_2 \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 \\
\mathbf{f}_3 &= \mathbf{x}_3 - \tfrac{\mathbf{x}_3 \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \tfrac{\mathbf{x}_3 \cdot \mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 \\
&\vdots \\
\mathbf{f}_k &= \mathbf{x}_k - \tfrac{\mathbf{x}_k \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \tfrac{\mathbf{x}_k \cdot \mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 - \cdots - \tfrac{\mathbf{x}_k \cdot \mathbf{f}_{k-1}}{\|\mathbf{f}_{k-1}\|^2}\mathbf{f}_{k-1}
\end{aligned}$$

for each $k = 2, 3, \dots, m$. Then

1. *$\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$ is an orthogonal basis of U.*
2. *$\operatorname{span}\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_k\} = \operatorname{span}\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ for each $k = 1, 2, \dots, m$.*

The process (for $k = 3$) is depicted in the diagrams. Of course, the algorithm converts any basis of $\mathbb{R}^n$ itself into an orthogonal basis.

Example 8.1.1

Find an orthogonal basis of the row space of $A = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 3 & 2 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$.

Solution. Let $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ denote the rows of A and observe that $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is linearly independent. Take $\mathbf{f}_1 = \mathbf{x}_1$. The algorithm gives

$$\begin{aligned}
\mathbf{f}_2 &= \mathbf{x}_2 - \tfrac{\mathbf{x}_2 \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 = (3, 2, 0, 1) - \tfrac{4}{4}(1, 1, -1, -1) = (2, 1, 1, 2) \\
\mathbf{f}_3 &= \mathbf{x}_3 - \tfrac{\mathbf{x}_3 \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \tfrac{\mathbf{x}_3 \cdot \mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 = \mathbf{x}_3 - \tfrac{0}{4}\mathbf{f}_1 - \tfrac{3}{10}\mathbf{f}_2 = \tfrac{1}{10}(4, -3, 7, -6)
\end{aligned}$$

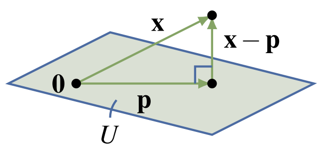
Hence $\{(1, 1, -1, -1), (2, 1, 1, 2), \tfrac{1}{10}(4, -3, 7, -6)\}$ is the orthogonal basis provided by the algorithm. In hand calculations it may be convenient to eliminate fractions (see the Remark below), so $\{(1, 1, -1, -1), (2, 1, 1, 2), (4, -3, 7, -6)\}$ is also an orthogonal basis for row A.

[1]Erhardt Schmidt (1876–1959) was a German mathematician who studied under the great David Hilbert and later developed the theory of Hilbert spaces. He first described the present algorithm in 1907. Jörgen Pederson Gram (1850–1916) was a Danish actuary.

Remark

Observe that the vector $\frac{\mathbf{x}\cdot\mathbf{f}_i}{\|\mathbf{f}_i\|^2}\mathbf{f}_i$ is unchanged if a nonzero scalar multiple of $\mathbf{f}_i$ is used in place of $\mathbf{f}_i$. Hence, if a newly constructed $\mathbf{f}_i$ is multiplied by a nonzero scalar at some stage of the Gram-Schmidt algorithm, the subsequent $\mathbf{f}$s will be unchanged. This is useful in actual calculations.

Projections



Suppose a point $\mathbf{x}$ and a plane U through the origin in $\mathbb{R}^3$ are given, and we want to find the point $\mathbf{p}$ in the plane that is closest to $\mathbf{x}$. Our geometric intuition assures us that such a point $\mathbf{p}$ exists. In fact (see the diagram), $\mathbf{p}$ must be chosen in such a way that $\mathbf{x}-\mathbf{p}$ is *perpendicular* to the plane.

Now we make two observations: first, the plane U is a *subspace of* $\mathbb{R}^3$ (because U contains the origin); and second, that the condition that $\mathbf{x}-\mathbf{p}$ is perpendicular to the plane U means that $\mathbf{x}-\mathbf{p}$ is *orthogonal* to every vector in U. In these terms the whole discussion makes sense in $\mathbb{R}^n$. Furthermore, the orthogonal lemma provides exactly what is needed to find $\mathbf{p}$ in this more general setting.

Definition 8.1 Orthogonal Complement of a Subspace of $\mathbb{R}^n$

If U is a subspace of $\mathbb{R}^n$, define the ***orthogonal complement*** *$U^\perp$ of U (pronounced "U-perp") by*

$$U^\perp = \{\mathbf{x} \text{ in } \mathbb{R}^n \mid \mathbf{x}\cdot\mathbf{y} = 0 \text{ for all } \mathbf{y} \text{ in } U\}$$

The following lemma collects some useful properties of the orthogonal complement; the proof of (1) and (2) is left as Exercise **??**.

Lemma 8.1.2

Let U be a subspace of $\mathbb{R}^n$.

1. *$U^\perp$ is a subspace of $\mathbb{R}^n$.*
2. *$\{\mathbf{0}\}^\perp = \mathbb{R}^n$ and $(\mathbb{R}^n)^\perp = \{\mathbf{0}\}$.*
3. *If $U = \text{span}\,\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$, then $U^\perp = \{\mathbf{x} \text{ in } \mathbb{R}^n \mid \mathbf{x}\cdot\mathbf{x}_i = 0 \text{ for } i = 1, 2, \dots, k\}$.*

Proof.

3. Let $U = \text{span}\,\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$; we must show that $U^\perp = \{\mathbf{x} \mid \mathbf{x}\cdot\mathbf{x}_i = 0 \text{ for each } i\}$. If $\mathbf{x}$ is in $U^\perp$ then $\mathbf{x}\cdot\mathbf{x}_i = 0$ for all i because each $\mathbf{x}_i$ is in U. Conversely, suppose that $\mathbf{x}\cdot\mathbf{x}_i = 0$ for all i; we must show that $\mathbf{x}$ is in $U^\perp$, that is, $\mathbf{x}\cdot\mathbf{y} = 0$ for each $\mathbf{y}$ in U. Write $\mathbf{y} = r_1\mathbf{x}_1 + r_2\mathbf{x}_2 + \cdots + r_k\mathbf{x}_k$, where each r_i is in $\mathbb{R}$. Then, using Theorem 5.3.1,

$$\mathbf{x}\cdot\mathbf{y} = r_1(\mathbf{x}\cdot\mathbf{x}_1) + r_2(\mathbf{x}\cdot\mathbf{x}_2) + \cdots + r_k(\mathbf{x}\cdot\mathbf{x}_k) = r_1 0 + r_2 0 + \cdots + r_k 0 = 0$$

as required. □

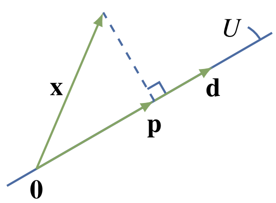
Example 8.1.2

Find $U^\perp$ if $U = \text{span}\{(1,\ -1,\ 2,\ 0),\ (1,\ 0,\ -2,\ 3)\}$ in $\mathbb{R}^4$.

Solution. By Lemma 8.1.2, $\mathbf{x} = (x,\ y,\ z,\ w)$ is in $U^\perp$ if and only if it is orthogonal to both $(1,\ -1,\ 2,\ 0)$ and $(1,\ 0,\ -2,\ 3)$; that is,

$$\begin{aligned} x - y + 2z \phantom{{}+3w} &= 0 \\ x \phantom{{}-y} - 2z + 3w &= 0 \end{aligned}$$

Gaussian elimination gives $U^\perp = \text{span}\{(2,\ 4,\ 1,\ 0),\ (3,\ 3,\ 0,\ -1)\}$.



Now consider vectors $\mathbf{x}$ and $\mathbf{d} \neq \mathbf{0}$ in $\mathbb{R}^3$. The projection $\mathbf{p} = \text{proj}_{\mathbf{d}}\,\mathbf{x}$ of $\mathbf{x}$ on $\mathbf{d}$ was defined in Section 4.2 as in the diagram.

The following formula for $\mathbf{p}$ was derived in Theorem 4.2.4

$$\mathbf{p} = \text{proj}_{\mathbf{d}}\,\mathbf{x} = \left(\frac{\mathbf{x}\cdot\mathbf{d}}{\|\mathbf{d}\|^2}\right)\mathbf{d}$$

where it is shown that $\mathbf{x} - \mathbf{p}$ is orthogonal to $\mathbf{d}$. Now observe that the line $U = \mathbb{R}\mathbf{d} = \{t\mathbf{d} \mid t \in \mathbb{R}\}$ is a subspace of $\mathbb{R}^3$, that $\{\mathbf{d}\}$ is an orthogonal basis of U, and that $\mathbf{p} \in U$ and $\mathbf{x} - \mathbf{p} \in U^\perp$ (by Theorem 4.2.4).

In this form, this makes sense for any vector $\mathbf{x}$ in $\mathbb{R}^n$ and any subspace U of $\mathbb{R}^n$, so we generalize it as follows. If $\{\mathbf{f}_1,\ \mathbf{f}_2,\ \dots,\ \mathbf{f}_m\}$ is an orthogonal basis of U, we define the projection $\mathbf{p}$ of $\mathbf{x}$ on U by the formula

$$\mathbf{p} = \left(\frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\right)\mathbf{f}_1 + \left(\frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\right)\mathbf{f}_2 + \cdots + \left(\frac{\mathbf{x}\cdot\mathbf{f}_m}{\|\mathbf{f}_m\|^2}\right)\mathbf{f}_m \tag{8.1}$$

Then $\mathbf{p} \in U$ and (by the orthogonal lemma) $\mathbf{x} - \mathbf{p} \in U^\perp$, so it looks like we have a generalization of Theorem 4.2.4.

However there is a potential problem: the formula (8.1) for $\mathbf{p}$ must be shown to be independent of the choice of the orthogonal basis $\{\mathbf{f}_1,\ \mathbf{f}_2,\ \dots,\ \mathbf{f}_m\}$. To verify this, suppose that $\{\mathbf{f}'_1,\ \mathbf{f}'_2,\ \dots,\ \mathbf{f}'_m\}$ is another orthogonal basis of U, and write

$$\mathbf{p}' = \left(\frac{\mathbf{x}\cdot\mathbf{f}'_1}{\|\mathbf{f}'_1\|^2}\right)\mathbf{f}'_1 + \left(\frac{\mathbf{x}\cdot\mathbf{f}'_2}{\|\mathbf{f}'_2\|^2}\right)\mathbf{f}'_2 + \cdots + \left(\frac{\mathbf{x}\cdot\mathbf{f}'_m}{\|\mathbf{f}'_m\|^2}\right)\mathbf{f}'_m$$

As before, $\mathbf{p}' \in U$ and $\mathbf{x} - \mathbf{p}' \in U^\perp$, and we must show that $\mathbf{p}' = \mathbf{p}$. To see this, write the vector $\mathbf{p} - \mathbf{p}'$ as follows:

$$\mathbf{p} - \mathbf{p}' = (\mathbf{x} - \mathbf{p}') - (\mathbf{x} - \mathbf{p})$$

This vector is in U (because $\mathbf{p}$ and $\mathbf{p}'$ are in U) and it is in $U^\perp$ (because $\mathbf{x} - \mathbf{p}'$ and $\mathbf{x} - \mathbf{p}$ are in $U^\perp$), and so it must be zero (it is orthogonal to itself!). This means $\mathbf{p}' = \mathbf{p}$ as desired.

Hence, the vector $\mathbf{p}$ in equation (8.1) depends only on $\mathbf{x}$ and the subspace U, and *not* on the choice of orthogonal basis $\{\mathbf{f}_1,\ \dots,\ \mathbf{f}_m\}$ of U used to compute it. Thus, we are entitled to make the following definition:

Definition 8.2 Projection onto a Subspace of $\mathbb{R}^n$

Let U be a subspace of $\mathbb{R}^n$ with orthogonal basis $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$. If $\mathbf{x}$ is in $\mathbb{R}^n$, the vector

$$\text{proj}_U \mathbf{x} = \frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 + \frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 + \cdots + \frac{\mathbf{x}\cdot\mathbf{f}_m}{\|\mathbf{f}_m\|^2}\mathbf{f}_m$$

is called the ***orthogonal projection*** *of $\mathbf{x}$ on U. For the zero subspace $U = \{\mathbf{0}\}$, we define*

$$\text{proj}_{\{\mathbf{0}\}} \mathbf{x} = \mathbf{0}$$

The preceding discussion proves (1) of the following theorem.

Theorem 8.1.3: Projection Theorem

If U is a subspace of $\mathbb{R}^n$ and $\mathbf{x}$ is in $\mathbb{R}^n$, write $\mathbf{p} = \text{proj}_U \mathbf{x}$. Then:

1. *$\mathbf{p}$ is in U and $\mathbf{x} - \mathbf{p}$ is in $U^\perp$.*
2. *$\mathbf{p}$ is the vector in U closest to $\mathbf{x}$ in the sense that*

$$\|\mathbf{x} - \mathbf{p}\| < \|\mathbf{x} - \mathbf{y}\| \quad \text{for all } \mathbf{y} \in U, \ \mathbf{y} \neq \mathbf{p}$$

Proof.

1. This is proved in the preceding discussion (it is clear if $U = \{\mathbf{0}\}$).
2. Write $\mathbf{x} - \mathbf{y} = (\mathbf{x} - \mathbf{p}) + (\mathbf{p} - \mathbf{y})$. Then $\mathbf{p} - \mathbf{y}$ is in U and so is orthogonal to $\mathbf{x} - \mathbf{p}$ by (1). Hence, the Pythagorean theorem gives

$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x} - \mathbf{p}\|^2 + \|\mathbf{p} - \mathbf{y}\|^2 > \|\mathbf{x} - \mathbf{p}\|^2$$

because $\mathbf{p} - \mathbf{y} \neq \mathbf{0}$. This gives (2). □

Example 8.1.3

Let $U = \text{span}\{\mathbf{x}_1, \mathbf{x}_2\}$ in $\mathbb{R}^4$ where $\mathbf{x}_1 = (1, 1, 0, 1)$ and $\mathbf{x}_2 = (0, 1, 1, 2)$. If $\mathbf{x} = (3, -1, 0, 2)$, find the vector in U closest to $\mathbf{x}$ and express $\mathbf{x}$ as the sum of a vector in U and a vector orthogonal to U.

Solution. $\{\mathbf{x}_1, \mathbf{x}_2\}$ is independent but not orthogonal. The Gram-Schmidt process gives an orthogonal basis $\{\mathbf{f}_1, \mathbf{f}_2\}$ of U where $\mathbf{f}_1 = \mathbf{x}_1 = (1, 1, 0, 1)$ and

$$\mathbf{f}_2 = \mathbf{x}_2 - \frac{\mathbf{x}_2\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 = \mathbf{x}_2 - \frac{3}{3}\mathbf{f}_1 = (-1, 0, 1, 1)$$

Hence, we can compute the projection using $\{\mathbf{f}_1, \mathbf{f}_2\}$:

$$\mathbf{p} = \text{proj}_U \mathbf{x} = \frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 + \frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 = \frac{4}{3}\mathbf{f}_1 + \frac{-1}{3}\mathbf{f}_2 = \frac{1}{3}\begin{bmatrix} 5 & 4 & -1 & 3 \end{bmatrix}$$

Thus, $\mathbf{p}$ is the vector in U closest to $\mathbf{x}$, and $\mathbf{x}-\mathbf{p} = \frac{1}{3}(4,\ -7,\ 1,\ 3)$ is orthogonal to every vector in U. (This can be verified by checking that it is orthogonal to the generators $\mathbf{x}_1$ and $\mathbf{x}_2$ of U.) The required decomposition of $\mathbf{x}$ is thus

$$\mathbf{x} = \mathbf{p} + (\mathbf{x}-\mathbf{p}) = \tfrac{1}{3}(5,\ 4,\ -1,\ 3) + \tfrac{1}{3}(4,\ -7,\ 1,\ 3)$$

Example 8.1.4

Find the point in the plane with equation $2x+y-z=0$ that is closest to the point $(2,\ -1,\ -3)$.

Solution. We write $\mathbb{R}^3$ as rows. The plane is the subspace U whose points $(x,\ y,\ z)$ satisfy $z = 2x+y$. Hence

$$U = \{(s,\ t,\ 2s+t) \mid s,\ t \text{ in } \mathbb{R}\} = \text{span}\,\{(0,\ 1,\ 1),\ (1,\ 0,\ 2)\}$$

The Gram-Schmidt process produces an orthogonal basis $\{\mathbf{f}_1,\ \mathbf{f}_2\}$ of U where $\mathbf{f}_1 = (0,\ 1,\ 1)$ and $\mathbf{f}_2 = (1,\ -1,\ 1)$. Hence, the vector in U closest to $\mathbf{x} = (2,\ -1,\ -3)$ is

$$\text{proj}_U\ \mathbf{x} = \tfrac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 + \tfrac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 = -2\mathbf{f}_1 + 0\mathbf{f}_2 = (0,\ -2,\ -2)$$

Thus, the point in U closest to $(2,\ -1,\ -3)$ is $(0,\ -2,\ -2)$.

The next theorem shows that projection on a subspace of $\mathbb{R}^n$ is actually a linear operator $\mathbb{R}^n \to \mathbb{R}^n$.

Theorem 8.1.4

Let U be a fixed subspace of $\mathbb{R}^n$. If we define $T : \mathbb{R}^n \to \mathbb{R}^n$ by

$$T(\mathbf{x}) = \text{proj}_U\ \mathbf{x} \quad \text{for all } \mathbf{x} \text{ in } \mathbb{R}^n$$

1. *T is a linear operator.*
2. *$\text{im}\ T = U$ and $\ker T = U^\perp$.*
3. *$\dim U + \dim U^\perp = n$.*

Proof. If $U = \{\mathbf{0}\}$, then $U^\perp = \mathbb{R}^n$, and so $T(\mathbf{x}) = \text{proj}_{\{\mathbf{0}\}}\ \mathbf{x} = \mathbf{0}$ for all $\mathbf{x}$. Thus $T = 0$ is the zero (linear) operator, so (1), (2), and (3) hold. Hence assume that $U \neq \{\mathbf{0}\}$.

1. If $\{\mathbf{f}_1,\ \mathbf{f}_2,\ \dots,\ \mathbf{f}_m\}$ is an orthonormal basis of U, then

$$T(\mathbf{x}) = (\mathbf{x}\cdot\mathbf{f}_1)\mathbf{f}_1 + (\mathbf{x}\cdot\mathbf{f}_2)\mathbf{f}_2 + \cdots + (\mathbf{x}\cdot\mathbf{f}_m)\mathbf{f}_m \quad \text{for all } \mathbf{x} \text{ in } \mathbb{R}^n \tag{8.2}$$

by the definition of the projection. Thus T is linear because

$$(\mathbf{x}+\mathbf{y})\cdot\mathbf{f}_i = \mathbf{x}\cdot\mathbf{f}_i + \mathbf{y}\cdot\mathbf{f}_i \quad \text{and} \quad (r\mathbf{x})\cdot\mathbf{f}_i = r(\mathbf{x}\cdot\mathbf{f}_i) \quad \text{for each } i$$

2. We have $\operatorname{im} T \subseteq U$ by (8.2) because each $\mathbf{f}_i$ is in U. But if $\mathbf{x}$ is in U, then $\mathbf{x} = T(\mathbf{x})$ by (8.2) and the expansion theorem applied to the space U. This shows that $U \subseteq \operatorname{im} T$, so $\operatorname{im} T = U$.

   Now suppose that $\mathbf{x}$ is in $U^{\perp}$. Then $\mathbf{x} \cdot \mathbf{f}_i = 0$ for each i (again because each $\mathbf{f}_i$ is in U) so $\mathbf{x}$ is in $\ker T$ by (8.2). Hence $U^{\perp} \subseteq \ker T$. On the other hand, Theorem 8.1.3 shows that $\mathbf{x} - T(\mathbf{x})$ is in $U^{\perp}$ for all $\mathbf{x}$ in $\mathbb{R}^n$, and it follows that $\ker T \subseteq U^{\perp}$. Hence $\ker T = U^{\perp}$, proving (2).

3. This follows from (1), (2), and the dimension theorem (Theorem 7.2.4). □

8.2 Orthogonal Diagonalization

Recall (Theorem 5.5.3) that an $n \times n$ matrix A is diagonalizable if and only if it has n linearly independent eigenvectors. Moreover, the matrix P with these eigenvectors as columns is a diagonalizing matrix for A, that is

$$P^{-1}AP \text{ is diagonal.}$$

As we have seen, the really nice bases of $\mathbb{R}^n$ are the orthogonal ones, so a natural question is: which $n \times n$ matrices have an *orthogonal* basis of eigenvectors? These turn out to be precisely the symmetric matrices, and this is the main result of this section.

Before proceeding, recall that an orthogonal set of vectors is called *orthonormal* if $\|\mathbf{v}\| = 1$ for each vector $\mathbf{v}$ in the set, and that any orthogonal set $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ can be "*normalized*", that is converted into an orthonormal set $\{\frac{1}{\|\mathbf{v}_1\|}\mathbf{v}_1, \frac{1}{\|\mathbf{v}_2\|}\mathbf{v}_2, \dots, \frac{1}{\|\mathbf{v}_k\|}\mathbf{v}_k\}$. In particular, if a matrix A has n orthogonal eigenvectors, they can (by normalizing) be taken to be orthonormal. The corresponding diagonalizing matrix P has orthonormal columns, and such matrices are very easy to invert.

Theorem 8.2.1

The following conditions are equivalent for an $n \times n$ matrix P.

1. *P is invertible and $P^{-1} = P^T$.*
2. *The rows of P are orthonormal.*
3. *The columns of P are orthonormal.*

Proof. First recall that condition (1) is equivalent to $PP^T = I$ by Corollary 2.4.2 of Theorem 2.4.5. Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ denote the rows of P. Then $\mathbf{x}_j^T$ is the jth column of P^T, so the (i, j)-entry of PP^T is $\mathbf{x}_i \cdot \mathbf{x}_j$. Thus $PP^T = I$ means that $\mathbf{x}_i \cdot \mathbf{x}_j = 0$ if $i \neq j$ and $\mathbf{x}_i \cdot \mathbf{x}_j = 1$ if $i = j$. Hence condition (1) is equivalent to (2). The proof of the equivalence of (1) and (3) is similar. □

Definition 8.3 Orthogonal Matrices

An $n \times n$ matrix P is called an ***orthogonal matrix***[2] *if it satisfies one (and hence all) of the conditions in Theorem 8.2.1.*

Example 8.2.1

The rotation matrix $\left[\begin{array}{rr} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array} \right]$ is orthogonal for any angle θ.

These orthogonal matrices have the virtue that they are easy to invert—simply take the transpose. But they have many other important properties as well. If $T : \mathbb{R}^n \to \mathbb{R}^n$ is a linear operator, we will prove (Theorem 10.4.3) that T is distance preserving if and only if its matrix is orthogonal. In particular, the matrices of rotations and reflections about the origin in $\mathbb{R}^2$ and $\mathbb{R}^3$ are all orthogonal (see Example 8.2.1).

It is not enough that the rows of a matrix A are merely orthogonal for A to be an orthogonal matrix. Here is an example.

Example 8.2.2

The matrix $\left[\begin{array}{rrr} 2 & 1 & 1 \\ -1 & 1 & 1 \\ 0 & -1 & 1 \end{array} \right]$ has orthogonal rows but the columns are not orthogonal. However, if the rows are normalized, the resulting matrix $\left[\begin{array}{ccc} \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{-1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{array} \right]$ is orthogonal (so the columns are now orthonormal as the reader can verify).

[2]In view of (2) and (3) of Theorem 8.2.1, *orthonormal matrix* might be a better name. But *orthogonal matrix* is standard.

Example 8.2.3

If P and Q are orthogonal matrices, then PQ is also orthogonal, as is $P^{-1} = P^T$.

Solution. P and Q are invertible, so PQ is also invertible and

$$(PQ)^{-1} = Q^{-1}P^{-1} = Q^T P^T = (PQ)^T$$

Hence PQ is orthogonal. Similarly,

$$(P^{-1})^{-1} = P = (P^T)^T = (P^{-1})^T$$

shows that P^{-1} is orthogonal.

Definition 8.4 Orthogonally Diagonalizable Matrices

*An $n \times n$ matrix A is said to be **orthogonally diagonalizable** when an orthogonal matrix P can be found such that $P^{-1}AP = P^T AP$ is diagonal.*

This condition turns out to characterize the symmetric matrices.

Theorem 8.2.2: Principal Axes Theorem

The following conditions are equivalent for an $n \times n$ matrix A.

1. *A has an orthonormal set of n eigenvectors.*
2. *A is orthogonally diagonalizable.*
3. *A is symmetric.*

Proof. (1) $\Leftrightarrow$ (2). Given (1), let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ be orthonormal eigenvectors of A. Then $P = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{bmatrix}$ is orthogonal, and $P^{-1}AP$ is diagonal by Theorem 3.4.1. This proves (2). Conversely, given (2) let $P^{-1}AP$ be diagonal where P is orthogonal. If $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ are the columns of P then $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ is an orthonormal basis of $\mathbb{R}^n$ that consists of eigenvectors of A by Theorem 3.4.1. This proves (1).

(2) $\Rightarrow$ (3). If $P^T AP = D$ is diagonal, where $P^{-1} = P^T$, then $A = PDP^T$. But $D^T = D$, so this gives $A^T = P^{TT} D^T P^T = PDP^T = A$.

(3) $\Rightarrow$ (2). If A is an $n \times n$ symmetric matrix, we proceed by induction on n. If $n = 1$, A is already diagonal. If $n > 1$, assume that (3) $\Rightarrow$ (2) for $(n-1) \times (n-1)$ symmetric matrices. By Theorem 5.5.7 let λ_1 be a (real) eigenvalue of A, and let $A\mathbf{x}_1 = \lambda_1 \mathbf{x}_1$, where $\|\mathbf{x}_1\| = 1$. Use the Gram-Schmidt algorithm to find an orthonormal basis $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ for $\mathbb{R}^n$. Let $P_1 = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{bmatrix}$, so P_1 is an orthogonal matrix and $P_1^T AP_1 = \begin{bmatrix} \lambda_1 & B \\ 0 & A_1 \end{bmatrix}$ in block form by Lemma 5.5.2. But $P_1^T AP_1$ is symmetric (A is), so it follows that $B = 0$ and A_1 is symmetric. Then, by induction, there exists an $(n-1) \times (n-1)$ orthogonal matrix Q such that $Q^T A_1 Q = D_1$ is diagonal. Observe that $P_2 = \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix}$ is orthogonal, and compute:

$$(P_1P_2)^T A(P_1P_2) = P_2^T (P_1^T AP_1)P_2$$

$$
\begin{aligned}
&= \begin{bmatrix} 1 & 0 \\ 0 & Q^T \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1 & 0 \\ 0 & D_1 \end{bmatrix}
\end{aligned}
$$

is diagonal. Because $P_1 P_2$ is orthogonal, this proves (2). □

A set of orthonormal eigenvectors of a symmetric matrix A is called a set of **principal axes** for A. The name comes from geometry, and this is discussed in Section 8.9. Because the eigenvalues of a (real) symmetric matrix are real, Theorem 8.2.2 is also called the **real spectral theorem**, and the set of distinct eigenvalues is called the **spectrum** of the matrix. In full generality, the spectral theorem is a similar result for matrices with complex entries (Theorem 8.7.8).

Example 8.2.4

Find an orthogonal matrix P such that $P^{-1}AP$ is diagonal, where $A = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ -1 & 2 & 5 \end{bmatrix}$.

Solution. The characteristic polynomial of A is (adding twice row 1 to row 2):

$$c_A(x) = \det \begin{bmatrix} x-1 & 0 & 1 \\ 0 & x-1 & -2 \\ 1 & -2 & x-5 \end{bmatrix} = x(x-1)(x-6)$$

Thus the eigenvalues are $\lambda = 0$, 1, and 6, and corresponding eigenvectors are

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \quad \mathbf{x}_3 = \begin{bmatrix} -1 \\ 2 \\ 5 \end{bmatrix}$$

respectively. Moreover, by what appears to be remarkably good luck, these eigenvectors are *orthogonal*. We have $\|\mathbf{x}_1\|^2 = 6$, $\|\mathbf{x}_2\|^2 = 5$, and $\|\mathbf{x}_3\|^2 = 30$, so

$$P = \begin{bmatrix} \frac{1}{\sqrt{6}}\mathbf{x}_1 & \frac{1}{\sqrt{5}}\mathbf{x}_2 & \frac{1}{\sqrt{30}}\mathbf{x}_3 \end{bmatrix} = \frac{1}{\sqrt{30}} \begin{bmatrix} \sqrt{5} & 2\sqrt{6} & -1 \\ -2\sqrt{5} & \sqrt{6} & 2 \\ \sqrt{5} & 0 & 5 \end{bmatrix}$$

is an orthogonal matrix. Thus $P^{-1} = P^T$ and

$$P^T AP = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 6 \end{bmatrix}$$

by the diagonalization algorithm.

Actually, the fact that the eigenvectors in Example 8.2.4 are orthogonal is no coincidence. Theorem 5.5.4 guarantees they are linearly independent (they correspond to distinct eigenvalues); the fact that the matrix is *symmetric* implies that they are orthogonal. To prove this we need the following useful fact about symmetric matrices.

Theorem 8.2.3

If A is an $n \times n$ symmetric matrix, then

$$(A\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (A\mathbf{y})$$

for all columns $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$.[3]

Proof. Recall that $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T\mathbf{y}$ for all columns $\mathbf{x}$ and $\mathbf{y}$. Because $A^T = A$, we get

$$(A\mathbf{x}) \cdot \mathbf{y} = (A\mathbf{x})^T\mathbf{y} = \mathbf{x}^T A^T \mathbf{y} = \mathbf{x}^T A\mathbf{y} = \mathbf{x} \cdot (A\mathbf{y})$$

□

Theorem 8.2.4

If A is a symmetric matrix, then eigenvectors of A corresponding to distinct eigenvalues are orthogonal.

Proof. Let $A\mathbf{x} = \lambda\mathbf{x}$ and $A\mathbf{y} = \mu\mathbf{y}$, where $\lambda \neq \mu$. Using Theorem 8.2.3, we compute

$$\lambda(\mathbf{x} \cdot \mathbf{y}) = (\lambda\mathbf{x}) \cdot \mathbf{y} = (A\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (A\mathbf{y}) = \mathbf{x} \cdot (\mu\mathbf{y}) = \mu(\mathbf{x} \cdot \mathbf{y})$$

Hence $(\lambda - \mu)(\mathbf{x} \cdot \mathbf{y}) = 0$, and so $\mathbf{x} \cdot \mathbf{y} = 0$ because $\lambda \neq \mu$. □

Now the procedure for diagonalizing a symmetric $n \times n$ matrix is clear. Find the distinct eigenvalues (all real by Theorem 5.5.7) and find orthonormal bases for each eigenspace (the Gram-Schmidt algorithm may be needed). Then the set of all these basis vectors is orthonormal (by Theorem 8.2.4) and contains n vectors. Here is an example.

Example 8.2.5

Orthogonally diagonalize the symmetric matrix $A = \begin{bmatrix} 8 & -2 & 2 \\ -2 & 5 & 4 \\ 2 & 4 & 5 \end{bmatrix}$.

Solution. The characteristic polynomial is

$$c_A(x) = \det \begin{bmatrix} x-8 & 2 & -2 \\ 2 & x-5 & -4 \\ -2 & -4 & x-5 \end{bmatrix} = x(x-9)^2$$

Hence the distinct eigenvalues are 0 and 9 of multiplicities 1 and 2, respectively, so $\dim(E_0) = 1$ and $\dim(E_9) = 2$ by Theorem 5.5.6 (A is diagonalizable, being symmetric). Gaussian elimination

[3] The converse also holds (Exercise **??**).

gives

$$E_0(A) = \text{span}\,\{\mathbf{x}_1\}, \quad \mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \\ -2 \end{bmatrix}, \qquad \text{and} \qquad E_9(A) = \text{span}\left\{ \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} \right\}$$

The eigenvectors in E_9 are both orthogonal to $\mathbf{x}_1$ as Theorem 8.2.4 guarantees, but not to each other. However, the Gram-Schmidt process yields an orthogonal basis

$$\{\mathbf{x}_2,\, \mathbf{x}_3\} \text{ of } E_9(A) \quad \text{where} \quad \mathbf{x}_2 = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} \text{ and } \mathbf{x}_3 = \begin{bmatrix} 2 \\ 4 \\ 5 \end{bmatrix}$$

Normalizing gives orthonormal vectors $\{\frac{1}{3}\mathbf{x}_1,\, \frac{1}{\sqrt{5}}\mathbf{x}_2,\, \frac{1}{3\sqrt{5}}\mathbf{x}_3\}$, so

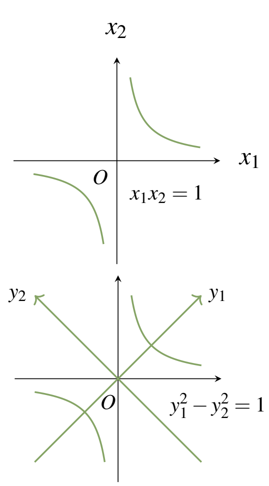
$$P = \left[\begin{array}{ccc} \frac{1}{3}\mathbf{x}_1 & \frac{1}{\sqrt{5}}\mathbf{x}_2 & \frac{1}{3\sqrt{5}}\mathbf{x}_3 \end{array}\right] = \frac{1}{3\sqrt{5}} \begin{bmatrix} \sqrt{5} & -6 & 2 \\ 2\sqrt{5} & 3 & 4 \\ -2\sqrt{5} & 0 & 5 \end{bmatrix}$$

is an orthogonal matrix such that $P^{-1}AP$ is diagonal.

It is worth noting that other, more convenient, diagonalizing matrices P exist. For example, $\mathbf{y}_2 = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}$ and $\mathbf{y}_3 = \begin{bmatrix} -2 \\ 2 \\ 1 \end{bmatrix}$ lie in $E_9(A)$ and they are orthogonal. Moreover, they both have norm 3 (as does $\mathbf{x}_1$), so

$$Q = \left[\begin{array}{ccc} \frac{1}{3}\mathbf{x}_1 & \frac{1}{3}\mathbf{y}_2 & \frac{1}{3}\mathbf{y}_3 \end{array}\right] = \frac{1}{3} \begin{bmatrix} 1 & 2 & -2 \\ 2 & 1 & 2 \\ -2 & 2 & 1 \end{bmatrix}$$

is a nicer orthogonal matrix with the property that $Q^{-1}AQ$ is diagonal.



If A is symmetric and a set of orthogonal eigenvectors of A is given, the eigenvectors are called principal axes of A. The name comes from geometry. An expression $q = ax_1^2 + bx_1x_2 + cx_2^2$ is called a **quadratic form** in the variables x_1 and x_2, and the graph of the equation $q = 1$ is called a **conic** in these variables. For example, if $q = x_1x_2$, the graph of $q = 1$ is given in the first diagram.

But if we introduce new variables y_1 and y_2 by setting $x_1 = y_1 + y_2$ and $x_2 = y_1 - y_2$, then q becomes $q = y_1^2 - y_2^2$, a diagonal form with no cross term y_1y_2 (see the second diagram). Because of this, the y_1 and y_2 axes are called the principal axes for the conic (hence the name). Orthogonal diagonalization provides a systematic method for finding principal axes. Here is an illustration.

Example 8.2.6

Find principal axes for the quadratic form $q = x_1^2 - 4x_1x_2 + x_2^2$.

Solution. In order to utilize diagonalization, we first express q in matrix form. Observe that

$$q = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & -4 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

The matrix here is not symmetric, but we can remedy that by writing

$$q = x_1^2 - 2x_1x_2 - 2x_2x_1 + x_2^2$$

Then we have

$$q = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{x}^T A \mathbf{x}$$

where $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $A = \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix}$ is symmetric. The eigenvalues of A are $\lambda_1 = 3$ and $\lambda_2 = -1$, with corresponding (orthogonal) eigenvectors $\mathbf{x}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and $\mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Since $\|\mathbf{x}_1\| = \|\mathbf{x}_2\| = \sqrt{2}$, so

$$P = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \text{ is orthogonal and } P^T AP = D = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}$$

Now define new variables $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{y}$ by $\mathbf{y} = P^T \mathbf{x}$, equivalently $\mathbf{x} = P\mathbf{y}$ (since $P^{-1} = P^T$). Hence

$$y_1 = \frac{1}{\sqrt{2}}(x_1 - x_2) \quad \text{and} \quad y_2 = \frac{1}{\sqrt{2}}(x_1 + x_2)$$

In terms of y_1 and y_2, q takes the form

$$q = \mathbf{x}^T A\mathbf{x} = (P\mathbf{y})^T A(P\mathbf{y}) = \mathbf{y}^T (P^T AP)\mathbf{y} = \mathbf{y}^T D\mathbf{y} = 3y_1^2 - y_2^2$$

Note that $\mathbf{y} = P^T \mathbf{x}$ is obtained from $\mathbf{x}$ by a counterclockwise rotation of $\frac{\pi}{4}$ (see Theorem 2.4.6).

Observe that the quadratic form q in Example 8.2.6 can be diagonalized in other ways. For example

$$q = x_1^2 - 4x_1x_2 + x_2^2 = z_1^2 - \tfrac{1}{3}z_2^2$$

where $z_1 = x_1 - 2x_2$ and $z_2 = 3x_2$. We examine this more carefully in Section 8.9.

If we are willing to replace "diagonal" by "upper triangular" in the principal axes theorem, we can weaken the requirement that A is symmetric to insisting only that A has real eigenvalues.

Theorem 8.2.5: Triangulation Theorem

If A is an $n \times n$ matrix with n real eigenvalues, an orthogonal matrix P exists such that $P^T AP$ is upper triangular.[4]

Proof. We modify the proof of Theorem 8.2.2. If $A\mathbf{x}_1 = \lambda_1 \mathbf{x}_1$ where $\|\mathbf{x}_1\| = 1$, let $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ be an orthonormal basis of $\mathbb{R}^n$, and let $P_1 = \left[\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{array} \right]$. Then P_1 is orthogonal and $P_1^T A P_1 = \left[\begin{array}{cc} \lambda_1 & B \\ 0 & A_1 \end{array} \right]$ in block form. By induction, let $Q^T A_1 Q = T_1$ be upper triangular where Q is of size $(n-1) \times (n-1)$ and orthogonal. Then $P_2 = \left[\begin{array}{cc} 1 & 0 \\ 0 & Q \end{array} \right]$ is orthogonal, so $P = P_1 P_2$ is also orthogonal and $P^T AP = \left[\begin{array}{cc} \lambda_1 & BQ \\ 0 & T_1 \end{array} \right]$ is upper triangular. □

The proof of Theorem 8.2.5 gives no way to construct the matrix P. However, an algorithm will be given in Section 11.1 where an improved version of Theorem 8.2.5 is presented. In a different direction, a version of Theorem 8.2.5 holds for an arbitrary matrix with complex entries (Schur's theorem in Section 8.7).

As for a diagonal matrix, the eigenvalues of an upper triangular matrix are displayed along the main diagonal. Because A and $P^T AP$ have the same determinant and trace whenever P is orthogonal, Theorem 8.2.5 gives:

Corollary 8.2.1

If A is an $n \times n$ matrix with real eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ (possibly not all distinct), then $\det A = \lambda_1 \lambda_2 \dots \lambda_n$ *and* $\operatorname{tr} A = \lambda_1 + \lambda_2 + \cdots + \lambda_n$.

This corollary remains true even if the eigenvalues are not real (using Schur's theorem).

[4]There is also a lower triangular version.

8.3 Positive Definite Matrices

All the eigenvalues of any symmetric matrix are real; this section is about the case in which the eigenvalues are positive. These matrices, which arise whenever optimization (maximum and minimum) problems are encountered, have countless applications throughout science and engineering. They also arise in statistics (for example, in factor analysis used in the social sciences) and in geometry (see Section 8.9). We will encounter them again in Chapter 10 when describing all inner products in $\mathbb{R}^n$.

Definition 8.5 Positive Definite Matrices

A square matrix is called ***positive definite*** *if it is symmetric and all its eigenvalues* λ *are positive, that is* $\lambda > 0$.

Because these matrices are symmetric, the principal axes theorem plays a central role in the theory.

Theorem 8.3.1

If A is positive definite, then it is invertible and $\det A > 0$.

Proof. If A is $n \times n$ and the eigenvalues are $\lambda_1, \lambda_2, \dots, \lambda_n$, then $\det A = \lambda_1 \lambda_2 \cdots \lambda_n > 0$ by the principal axes theorem (or the corollary to Theorem 8.2.5). $\square$

If $\mathbf{x}$ is a column in $\mathbb{R}^n$ and A is any real $n \times n$ matrix, we view the 1×1 matrix $\mathbf{x}^T A\mathbf{x}$ as a real number. With this convention, we have the following characterization of positive definite matrices.

Theorem 8.3.2

A symmetric matrix A is positive definite if and only if $\mathbf{x}^T A\mathbf{x} > 0$ for every column $\mathbf{x} \neq \mathbf{0}$ in $\mathbb{R}^n$.

Proof. A is symmetric so, by the principal axes theorem, let $P^T AP = D = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ where $P^{-1} = P^T$ and the λ_i are the eigenvalues of A. Given a column $\mathbf{x}$ in $\mathbb{R}^n$, write $\mathbf{y} = P^T\mathbf{x} = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T$. Then

$$\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T (PDP^T)\mathbf{x} = \mathbf{y}^T D\mathbf{y} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2 \tag{8.3}$$

If A is positive definite and $\mathbf{x} \neq \mathbf{0}$, then $\mathbf{x}^T A\mathbf{x} > 0$ by (8.3) because some $y_j \neq 0$ and every $\lambda_i > 0$. Conversely, if $\mathbf{x}^T A\mathbf{x} > 0$ whenever $\mathbf{x} \neq \mathbf{0}$, let $\mathbf{x} = P\mathbf{e}_j \neq \mathbf{0}$ where $\mathbf{e}_j$ is column j of I_n. Then $\mathbf{y} = \mathbf{e}_j$, so (8.3) reads $\lambda_j = \mathbf{x}^T A\mathbf{x} > 0$. □

Note that Theorem 8.3.2 shows that the positive definite matrices are exactly the symmetric matrices A for which the quadratic form $q = \mathbf{x}^T A\mathbf{x}$ takes only positive values.

Example 8.3.1

If U is any invertible $n \times n$ matrix, show that $A = U^T U$ is positive definite.

Solution. If $\mathbf{x}$ is in $\mathbb{R}^n$ and $\mathbf{x} \neq \mathbf{0}$, then

$$\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T (U^T U)\mathbf{x} = (U\mathbf{x})^T (U\mathbf{x}) = \|U\mathbf{x}\|^2 > 0$$

because $U\mathbf{x} \neq \mathbf{0}$ (U is invertible). Hence Theorem 8.3.2 applies.

It is remarkable that the converse to Example 8.3.1 is also true. In fact every positive definite matrix A can be factored as $A = U^T U$ where U is an upper triangular matrix with positive elements on the main diagonal. However, before verifying this, we introduce another concept that is central to any discussion of positive definite matrices.

If A is any $n \times n$ matrix, let ${}^{(r)}A$ denote the $r \times r$ submatrix in the upper left corner of A; that is, ${}^{(r)}A$ is the matrix obtained from A by deleting the last $n - r$ rows and columns. The matrices ${}^{(1)}A$, ${}^{(2)}A$, ${}^{(3)}A$, $\dots$, ${}^{(n)}A = A$ are called the **principal submatrices** of A.

Example 8.3.2

If $A = \begin{bmatrix} 10 & 5 & 2 \\ 5 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$ then ${}^{(1)}A = [10]$, ${}^{(2)}A = \begin{bmatrix} 10 & 5 \\ 5 & 3 \end{bmatrix}$ and ${}^{(3)}A = A$.

Lemma 8.3.1

If A is positive definite, so is each principal submatrix ${}^{(r)}A$ for $r = 1, 2, \dots, n$.

Proof. Write $A = \left[\begin{array}{cc} ^{(r)}A & P \\ Q & R \end{array} \right]$ in block form. If $\mathbf{y} \neq \mathbf{0}$ in $\mathbb{R}^r$, write $\mathbf{x} = \left[\begin{array}{c} \mathbf{y} \\ \mathbf{0} \end{array} \right]$ in $\mathbb{R}^n$.

Then $\mathbf{x} \neq \mathbf{0}$, so the fact that A is positive definite gives

$$0 < \mathbf{x}^T A \mathbf{x} = \left[\begin{array}{cc} \mathbf{y}^T & \mathbf{0} \end{array} \right] \left[\begin{array}{cc} ^{(r)}A & P \\ Q & R \end{array} \right] \left[\begin{array}{c} \mathbf{y} \\ \mathbf{0} \end{array} \right] = \mathbf{y}^T (^{(r)}A) \mathbf{y}$$

This shows that $^{(r)}A$ is positive definite by Theorem 8.3.2.[5] □

If A is positive definite, Lemma 8.3.1 and Theorem 8.3.1 show that $\det(^{(r)}A) > 0$ for every r. This proves part of the following theorem which contains the converse to Example 8.3.1, and characterizes the positive definite matrices among the symmetric ones.

Theorem 8.3.3

The following conditions are equivalent for a symmetric $n \times n$ matrix A:

1. *A is positive definite.*
2. *$\det(^{(r)}A) > 0$ for each $r = 1, 2, \dots, n$.*
3. *$A = U^T U$ where U is an upper triangular matrix with positive entries on the main diagonal.*

Furthermore, the factorization in (3) is unique (called the ***Cholesky factorization***[6] *of A).*

Proof. First, (3) $\Rightarrow$ (1) by Example 8.3.1, and (1) $\Rightarrow$ (2) by Lemma 8.3.1 and Theorem 8.3.1.

(2) $\Rightarrow$ (3). Assume (2) and proceed by induction on n. If $n = 1$, then $A = [a]$ where $a > 0$ by (2), so take $U = [\sqrt{a}]$. If $n > 1$, write $B = ^{(n-1)}A$. Then B is symmetric and satisfies (2) so, by induction, we have $B = U^T U$ as in (3) where U is of size $(n-1) \times (n-1)$. Then, as A is symmetric, it has block form $A = \left[\begin{array}{cc} B & \mathbf{p} \\ \mathbf{p}^T & b \end{array} \right]$ where $\mathbf{p}$ is a column in $\mathbb{R}^{n-1}$ and b is in $\mathbb{R}$. If we write $\mathbf{x} = (U^T)^{-1}\mathbf{p}$ and $c = b - \mathbf{x}^T \mathbf{x}$, block multiplication gives

$$A = \left[\begin{array}{cc} U^T U & \mathbf{p} \\ \mathbf{p}^T & b \end{array} \right] = \left[\begin{array}{cc} U^T & 0 \\ \mathbf{x}^T & 1 \end{array} \right] \left[\begin{array}{cc} U & \mathbf{x} \\ 0 & c \end{array} \right]$$

as the reader can verify. Taking determinants and applying Theorem 3.1.5 gives $\det A = \det(U^T) \det U \cdot c = c(\det U)^2$. Hence $c > 0$ because $\det A > 0$ by (2), so the above factorization can be written

$$A = \left[\begin{array}{cc} U^T & 0 \\ \mathbf{x}^T & \sqrt{c} \end{array} \right] \left[\begin{array}{cc} U & \mathbf{x} \\ 0 & \sqrt{c} \end{array} \right]$$

Since U has positive diagonal entries, this proves (3).

As to the uniqueness, suppose that $A = U^T U = U_1^T U_1$ are two Cholesky factorizations. Now write $D = UU_1^{-1} = (U^T)^{-1}U_1^T$. Then D is upper triangular, because $D = UU_1^{-1}$, and lower triangular, because $D = (U^T)^{-1}U_1^T$, and so it is a diagonal matrix. Thus $U = DU_1$ and $U_1 = DU$, so it suffices to show that

[5]A similar argument shows that, if B is any matrix obtained from a positive definite matrix A by deleting certain rows and deleting the *same* columns, then B is also positive definite.

[6]Andre-Louis Cholesky (1875–1918), was a French mathematician who died in World War I. His factorization was published in 1924 by a fellow officer.

$D = I$. But eliminating U_1 gives $U = D^2U$, so $D^2 = I$ because U is invertible. Since the diagonal entries of D are positive (this is true of U and U_1), it follows that $D = I$. □

The remarkable thing is that the matrix U in the Cholesky factorization is easy to obtain from A using row operations. The key is that Step 1 of the following algorithm is *possible* for any positive definite matrix A. A proof of the algorithm is given following Example 8.3.3.

Theorem: Algorithm for the Cholesky Factorization

If A is a positive definite matrix, the Cholesky factorization $A = U^TU$ can be obtained as follows:

Step 1. Carry A to an upper triangular matrix U_1 with positive diagonal entries using row operations each of which adds a multiple of a row to a lower row.

Step 2. Obtain U from U_1 by dividing each row of U_1 by the square root of the diagonal entry in that row.

Example 8.3.3

Find the Cholesky factorization of $A = \begin{bmatrix} 10 & 5 & 2 \\ 5 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$.

Solution. The matrix A is positive definite by Theorem 8.3.3 because $\det{}^{(1)}A = 10 > 0$, $\det{}^{(2)}A = 5 > 0$, and $\det{}^{(3)}A = \det A = 3 > 0$. Hence Step 1 of the algorithm is carried out as follows:

$$A = \begin{bmatrix} 10 & 5 & 2 \\ 5 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix} \to \begin{bmatrix} 10 & 5 & 2 \\ 0 & \frac{1}{2} & 1 \\ 0 & 1 & \frac{13}{5} \end{bmatrix} \to \begin{bmatrix} 10 & 5 & 2 \\ 0 & \frac{1}{2} & 1 \\ 0 & 0 & \frac{3}{5} \end{bmatrix} = U_1$$

Now carry out Step 2 on U_1 to obtain $U = \begin{bmatrix} \sqrt{10} & \frac{5}{\sqrt{10}} & \frac{2}{\sqrt{10}} \\ 0 & \frac{1}{\sqrt{2}} & \sqrt{2} \\ 0 & 0 & \frac{\sqrt{3}}{\sqrt{5}} \end{bmatrix}$.

The reader can verify that $U^TU = A$.

Proof of the Cholesky Algorithm. If A is positive definite, let $A = U^TU$ be the Cholesky factorization, and let $D = \text{diag}\,(d_1, \dots, d_n)$ be the common diagonal of U and U^T. Then U^TD^{-1} is lower triangular with ones on the diagonal (call such matrices LT-1). Hence $L = (U^TD^{-1})^{-1}$ is also LT-1, and so $I_n \to L$ by a sequence of row operations each of which adds a multiple of a row to a lower row (verify; modify columns right to left). But then $A \to LA$ by the same sequence of row operations (see the discussion preceding Theorem 2.5.1). Since $LA = [D(U^T)^{-1}][U^TU] = DU$ is upper triangular with positive entries on the diagonal, this shows that Step 1 of the algorithm is possible.

Turning to Step 2, let $A \to U_1$ as in Step 1 so that $U_1 = L_1A$ where L_1 is LT-1. Since A is symmetric, we get

$$L_1 U_1^T = L_1 (L_1 A)^T = L_1 A^T L_1^T = L_1 A L_1^T = U_1 L_1^T \tag{8.4}$$

Let $D_1 = \text{diag}(e_1, \dots, e_n)$ denote the diagonal of U_1. Then (8.4) gives $L_1 (U_1^T D_1^{-1}) = U_1 L_1^T D_1^{-1}$. This is both upper triangular (right side) and LT-1 (left side), and so must equal I_n. In particular, $U_1^T D_1^{-1} = L_1^{-1}$. Now let $D_2 = \text{diag}(\sqrt{e_1}, \dots, \sqrt{e_n})$, so that $D_2^2 = D_1$. If we write $U = D_2^{-1} U_1$ we have

$$U^T U = (U_1^T D_2^{-1})(D_2^{-1} U_1) = U_1^T (D_2^2)^{-1} U_1 = (U_1^T D_1^{-1}) U_1 = (L_1^{-1}) U_1 = A$$

This proves Step 2 because $U = D_2^{-1} U_1$ is formed by dividing each row of U_1 by the square root of its diagonal entry (verify). □

8.4 QR-Factorization[7]

One of the main virtues of orthogonal matrices is that they can be easily inverted—the transpose is the inverse. This fact, combined with the factorization theorem in this section, provides a useful way to simplify many matrix calculations (for example, in least squares approximation).

Definition 8.6 QR-factorization

*Let A be an $m \times n$ matrix with independent columns. A **QR-factorization** of A expresses it as $A = QR$ where Q is $m \times n$ with orthonormal columns and R is an invertible and upper triangular*

[7] This section is not used elsewhere in the book

matrix with positive diagonal entries.

The importance of the factorization lies in the fact that there are computer algorithms that accomplish it with good control over round-off error, making it particularly useful in matrix calculations. The factorization is a matrix version of the Gram-Schmidt process.

Suppose $A = \left[\begin{array}{cccc} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_n \end{array}\right]$ is an $m \times n$ matrix with linearly independent columns $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n$. The Gram-Schmidt algorithm can be applied to these columns to provide orthogonal columns $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n$ where $\mathbf{f}_1 = \mathbf{c}_1$ and

$$\mathbf{f}_k = \mathbf{c}_k - \frac{\mathbf{c}_k \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 + \frac{\mathbf{c}_k \cdot \mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 - \cdots - \frac{\mathbf{c}_k \cdot \mathbf{f}_{k-1}}{\|\mathbf{f}_{k-1}\|^2}\mathbf{f}_{k-1}$$

for each $k = 2, 3, \dots, n$. Now write $\mathbf{q}_k = \frac{1}{\|\mathbf{f}_k\|}\mathbf{f}_k$ for each k. Then $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n$ are orthonormal columns, and the above equation becomes

$$\|\mathbf{f}_k\|\mathbf{q}_k = \mathbf{c}_k - (\mathbf{c}_k \cdot \mathbf{q}_1)\mathbf{q}_1 - (\mathbf{c}_k \cdot \mathbf{q}_2)\mathbf{q}_2 - \cdots - (\mathbf{c}_k \cdot \mathbf{q}_{k-1})\mathbf{q}_{k-1}$$

Using these equations, express each $\mathbf{c}_k$ as a linear combination of the $\mathbf{q}_i$:

$$\begin{array}{lcl}
\mathbf{c}_1 & = & \|\mathbf{f}_1\|\mathbf{q}_1 \\
\mathbf{c}_2 & = & (\mathbf{c}_2 \cdot \mathbf{q}_1)\mathbf{q}_1 + \|\mathbf{f}_2\|\mathbf{q}_2 \\
\mathbf{c}_3 & = & (\mathbf{c}_3 \cdot \mathbf{q}_1)\mathbf{q}_1 + (\mathbf{c}_3 \cdot \mathbf{q}_2)\mathbf{q}_2 + \|\mathbf{f}_3\|\mathbf{q}_3 \\
\vdots & & \vdots \\
\mathbf{c}_n & = & (\mathbf{c}_n \cdot \mathbf{q}_1)\mathbf{q}_1 + (\mathbf{c}_n \cdot \mathbf{q}_2)\mathbf{q}_2 + (\mathbf{c}_n \cdot \mathbf{q}_3)\mathbf{q}_3 + \cdots + \|\mathbf{f}_n\|\mathbf{q}_n
\end{array}$$

These equations have a matrix form that gives the required factorization:

$$\begin{aligned}
A &= \left[\begin{array}{ccccc} \mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_3 & \cdots & \mathbf{c}_n \end{array}\right] \\
&= \left[\begin{array}{ccccc} \mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 & \cdots & \mathbf{q}_n \end{array}\right] \left[\begin{array}{ccccc}
\|\mathbf{f}_1\| & \mathbf{c}_2 \cdot \mathbf{q}_1 & \mathbf{c}_3 \cdot \mathbf{q}_1 & \cdots & \mathbf{c}_n \cdot \mathbf{q}_1 \\
0 & \|\mathbf{f}_2\| & \mathbf{c}_3 \cdot \mathbf{q}_2 & \cdots & \mathbf{c}_n \cdot \mathbf{q}_2 \\
0 & 0 & \|\mathbf{f}_3\| & \cdots & \mathbf{c}_n \cdot \mathbf{q}_3 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & \|\mathbf{f}_n\|
\end{array}\right]
\end{aligned} \tag{8.5}$$

Here the first factor $Q = \left[\begin{array}{ccccc} \mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 & \cdots & \mathbf{q}_n \end{array}\right]$ has orthonormal columns, and the second factor is an $n \times n$ upper triangular matrix R with positive diagonal entries (and so is invertible). We record this in the following theorem.

Theorem 8.4.1: QR-Factorization

Every $m \times n$ matrix A with linearly independent columns has a QR-factorization $A = QR$ where Q has orthonormal columns and R is upper triangular with positive diagonal entries.

The matrices Q and R in Theorem 8.4.1 are uniquely determined by A; we return to this below.

Example 8.4.1

Find the QR-factorization of $A = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$.

Solution. Denote the columns of A as $\mathbf{c}_1$, $\mathbf{c}_2$, and $\mathbf{c}_3$, and observe that $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$ is independent. If we apply the Gram-Schmidt algorithm to these columns, the result is:

$$\mathbf{f}_1 = \mathbf{c}_1 = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{f}_2 = \mathbf{c}_2 - \tfrac{1}{2}\mathbf{f}_1 = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 1 \\ 0 \end{bmatrix}, \text{ and } \quad \mathbf{f}_3 = \mathbf{c}_3 + \tfrac{1}{2}\mathbf{f}_1 - \mathbf{f}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Write $\mathbf{q}_j = \frac{1}{\|\mathbf{f}_j\|^2}\mathbf{f}_j$ for each j, so $\{\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3\}$ is orthonormal. Then equation (8.5) preceding Theorem 8.4.1 gives $A = QR$ where

$$Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & 0 \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & 0 \\ 0 & \frac{2}{\sqrt{6}} & 0 \\ 0 & 0 & 1 \end{bmatrix} = \frac{1}{\sqrt{6}} \begin{bmatrix} \sqrt{3} & 1 & 0 \\ -\sqrt{3} & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & \sqrt{6} \end{bmatrix}$$

$$R = \begin{bmatrix} \|\mathbf{f}_1\| & \mathbf{c}_2 \cdot \mathbf{q}_1 & \mathbf{c}_3 \cdot \mathbf{q}_1 \\ 0 & \|\mathbf{f}_2\| & \mathbf{c}_3 \cdot \mathbf{q}_2 \\ 0 & 0 & \|\mathbf{f}_3\| \end{bmatrix} = \begin{bmatrix} \sqrt{2} & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ 0 & \frac{\sqrt{3}}{\sqrt{2}} & \frac{\sqrt{3}}{\sqrt{2}} \\ 0 & 0 & 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 2 & 1 & -1 \\ 0 & \sqrt{3} & \sqrt{3} \\ 0 & 0 & \sqrt{2} \end{bmatrix}$$

The reader can verify that indeed $A = QR$.

If a matrix A has independent rows and we apply QR-factorization to A^T, the result is:

Corollary 8.4.1

If A has independent rows, then A factors uniquely as $A = LP$ where P has orthonormal rows and L is an invertible lower triangular matrix with positive main diagonal entries.

Since a square matrix with orthonormal columns is orthogonal, we have

Theorem 8.4.2

Every square, invertible matrix A has factorizations $A = QR$ and $A = LP$ where Q and P are orthogonal, R is upper triangular with positive diagonal entries, and L is lower triangular with positive diagonal entries.

Remark

In Section 5.6 we found how to find a best approximation $\mathbf{z}$ to a solution of a (possibly inconsistent) system $A\mathbf{x} = \mathbf{b}$ of linear equations: take $\mathbf{z}$ to be any solution of the "normal" equations $(A^T A)\mathbf{z} = A^T\mathbf{b}$. If A has independent columns this $\mathbf{z}$ is unique ($A^T A$ is invertible by Theorem 5.4.3), so it is often desirable to compute $(A^T A)^{-1}$. This is particularly useful in least squares approximation (Section 5.6). This is simplified if we have a QR-factorization of A (and is one of the main reasons for the importance of Theorem 8.4.1). For if $A = QR$ is such a factorization, then $Q^T Q = I_n$ because Q has orthonormal columns (verify), so we obtain

$$A^T A = R^T Q^T QR = R^T R$$

Hence computing $(A^T A)^{-1}$ amounts to finding R^{-1}, and this is a routine matter because R is upper triangular. Thus the difficulty in computing $(A^T A)^{-1}$ lies in obtaining the QR-factorization of A.

We conclude by proving the uniqueness of the QR-factorization.

Theorem 8.4.3

Let A be an $m \times n$ matrix with independent columns. If $A = QR$ and $A = Q_1 R_1$ are QR-factorizations of A, then $Q_1 = Q$ and $R_1 = R$.

Proof. Write $Q = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_n \end{bmatrix}$ and $Q_1 = \begin{bmatrix} \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_n \end{bmatrix}$ in terms of their columns, and observe first that $Q^T Q = I_n = Q_1^T Q_1$ because Q and Q_1 have orthonormal columns. Hence it suffices to show that $Q_1 = Q$ (then $R_1 = Q_1^T A = Q^T A = R$). Since $Q_1^T Q_1 = I_n$, the equation $QR = Q_1 R_1$ gives $Q_1^T Q = R_1 R^{-1}$; for convenience we write this matrix as

$$Q_1^T Q = R_1 R^{-1} = \begin{bmatrix} t_{ij} \end{bmatrix}$$

This matrix is upper triangular with positive diagonal elements (since this is true for R and R_1), so $t_{ii} > 0$ for each i and $t_{ij} = 0$ if $i > j$. On the other hand, the (i, j)-entry of $Q_1^T Q$ is $\mathbf{d}_i^T \mathbf{c}_j = \mathbf{d}_i \cdot \mathbf{c}_j$, so we have $\mathbf{d}_i \cdot \mathbf{c}_j = t_{ij}$ for all i and j. But each $\mathbf{c}_j$ is in $\text{span}\{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n\}$ because $Q = Q_1(R_1 R^{-1})$. Hence the expansion theorem gives

$$\mathbf{c}_j = (\mathbf{d}_1 \cdot \mathbf{c}_j)\mathbf{d}_1 + (\mathbf{d}_2 \cdot \mathbf{c}_j)\mathbf{d}_2 + \cdots + (\mathbf{d}_n \cdot \mathbf{c}_j)\mathbf{d}_n = t_{1j}\mathbf{d}_1 + t_{2j}\mathbf{d}_2 + \cdots + t_{jj}\mathbf{d}_i$$

because $\mathbf{d}_i \cdot \mathbf{c}_j = t_{ij} = 0$ if $i > j$. The first few equations here are

$$\begin{array}{lcl} \mathbf{c}_1 & = & t_{11}\mathbf{d}_1 \\ \mathbf{c}_2 & = & t_{12}\mathbf{d}_1 + t_{22}\mathbf{d}_2 \\ \mathbf{c}_3 & = & t_{13}\mathbf{d}_1 + t_{23}\mathbf{d}_2 + t_{33}\mathbf{d}_3 \\ \mathbf{c}_4 & = & t_{14}\mathbf{d}_1 + t_{24}\mathbf{d}_2 + t_{34}\mathbf{d}_3 + t_{44}\mathbf{d}_4 \\ \vdots & & \vdots \end{array}$$

The first of these equations gives $1 = \|\mathbf{c}_1\| = \|t_{11}\mathbf{d}_1\| = |t_{11}|\|\mathbf{d}_1\| = t_{11}$, whence $\mathbf{c}_1 = \mathbf{d}_1$. But then we have $t_{12} = \mathbf{d}_1 \cdot \mathbf{c}_2 = \mathbf{c}_1 \cdot \mathbf{c}_2 = 0$, so the second equation becomes $\mathbf{c}_2 = t_{22}\mathbf{d}_2$. Now a similar argument gives $\mathbf{c}_2 = \mathbf{d}_2$, and then $t_{13} = 0$ and $t_{23} = 0$ follows in the same way. Hence $\mathbf{c}_3 = t_{33}\mathbf{d}_3$ and $\mathbf{c}_3 = \mathbf{d}_3$. Continue in this way to get $\mathbf{c}_i = \mathbf{d}_i$ for all i. This means that $Q_1 = Q$, which is what we wanted. $\square$

8.5 Computing Eigenvalues

In practice, the problem of finding eigenvalues of a matrix is virtually never solved by finding the roots of the characteristic polynomial. This is difficult for large matrices and iterative methods are much better. Two such methods are described briefly in this section.

The Power Method

In Chapter 3 our initial rationale for diagonalizing matrices was to be able to compute the powers of a square matrix, and the eigenvalues were needed to do this. In this section, we are interested in efficiently computing eigenvalues, and it may come as no surprise that the first method we discuss uses the powers of a matrix.

Recall that an eigenvalue λ of an $n \times n$ matrix A is called a **dominant eigenvalue** if λ has multiplicity 1, and

$$|\lambda| > |\mu| \quad \text{for all eigenvalues } \mu \neq \lambda$$

Any corresponding eigenvector is called a **dominant eigenvector** of A. When such an eigenvalue exists, one technique for finding it is as follows: Let $\mathbf{x}_0$ in $\mathbb{R}^n$ be a first approximation to a dominant eigenvector λ, and compute successive approximations $\mathbf{x}_1, \mathbf{x}_2, \dots$ as follows:

$$\mathbf{x}_1 = A\mathbf{x}_0 \quad \mathbf{x}_2 = A\mathbf{x}_1 \quad \mathbf{x}_3 = A\mathbf{x}_2 \quad \cdots$$

In general, we define

$$\mathbf{x}_{k+1} = A\mathbf{x}_k \quad \text{for each } k \geq 0$$

If the first estimate $\mathbf{x}_0$ is good enough, these vectors $\mathbf{x}_n$ will approximate the dominant eigenvector λ (see below). This technique is called the **power method** (because $\mathbf{x}_k = A^k\mathbf{x}_0$ for each $k \geq 1$). Observe that if $\mathbf{z}$ is any eigenvector corresponding to λ, then

$$\frac{\mathbf{z}\cdot(A\mathbf{z})}{\|\mathbf{z}\|^2} = \frac{\mathbf{z}\cdot(\lambda\mathbf{z})}{\|\mathbf{z}\|^2} = \lambda$$

Because the vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \dots$ approximate dominant eigenvectors, this suggests that we define the **Rayleigh quotients** as follows:

$$r_k = \frac{\mathbf{x}_k\cdot\mathbf{x}_{k+1}}{\|\mathbf{x}_k\|^2} \quad \text{for } k \geq 1$$

Then the numbers r_k approximate the dominant eigenvalue λ.

Example 8.5.1

Use the power method to approximate a dominant eigenvector and eigenvalue of $A = \left[\begin{array}{rr} 1 & 1 \\ 2 & 0 \end{array}\right]$.

Solution. The eigenvalues of A are 2 and -1, with eigenvectors $\left[\begin{array}{r} 1 \\ 1 \end{array}\right]$ and $\left[\begin{array}{r} 1 \\ -2 \end{array}\right]$. Take $\mathbf{x}_0 = \left[\begin{array}{r} 1 \\ 0 \end{array}\right]$ as the first approximation and compute $\mathbf{x}_1, \mathbf{x}_2, \dots,$ successively, from $\mathbf{x}_1 = A\mathbf{x}_0, \mathbf{x}_2 = A\mathbf{x}_1, \dots$. The result is

$$\mathbf{x}_1 = \left[\begin{array}{r} 1 \\ 2 \end{array}\right], \ \mathbf{x}_2 = \left[\begin{array}{r} 3 \\ 2 \end{array}\right], \ \mathbf{x}_3 = \left[\begin{array}{r} 5 \\ 6 \end{array}\right], \ \mathbf{x}_4 = \left[\begin{array}{r} 11 \\ 10 \end{array}\right], \ \mathbf{x}_3 = \left[\begin{array}{r} 21 \\ 22 \end{array}\right], \ \dots$$

These vectors are approaching scalar multiples of the dominant eigenvector $\left[\begin{array}{r} 1 \\ 1 \end{array}\right]$. Moreover, the Rayleigh quotients are

$$r_1 = \frac{7}{5}, \ r_2 = \frac{27}{13}, \ r_3 = \frac{115}{61}, \ r_4 = \frac{451}{221}, \ \dots$$

and these are approaching the dominant eigenvalue 2.

To see why the power method works, let $\lambda_1, \lambda_2, \dots, \lambda_m$ be eigenvalues of A with λ_1 dominant and let $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$ be corresponding eigenvectors. What is required is that the first approximation $\mathbf{x}_0$ be a linear combination of these eigenvectors:

$$\mathbf{x}_0 = a_1\mathbf{y}_1 + a_2\mathbf{y}_2 + \cdots + a_m\mathbf{y}_m \quad \text{with } a_1 \neq 0$$

If $k \geq 1$, the fact that $\mathbf{x}_k = A^k\mathbf{x}_0$ and $A^k\mathbf{y}_i = \lambda_i^k\mathbf{y}_i$ for each i gives

$$\mathbf{x}_k = a_1\lambda_1^k\mathbf{y}_1 + a_2\lambda_2^k\mathbf{y}_2 + \cdots + a_m\lambda_m^k\mathbf{y}_m \quad \text{for } k \geq 1$$

Hence

$$\frac{1}{\lambda_1^k}\mathbf{x}_k = a_1\mathbf{y}_1 + a_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\mathbf{y}_2 + \cdots + a_m\left(\frac{\lambda_m}{\lambda_1}\right)^k\mathbf{y}_m$$

The right side approaches $a_1\mathbf{y}_1$ as k increases because λ_1 is dominant $\left(\left|\frac{\lambda_i}{\lambda_1}\right| < 1 \text{ for each } i > 1\right)$. Because $a_1 \neq 0$, this means that $\mathbf{x}_k$ approximates the dominant eigenvector $a_1\lambda_1^k\mathbf{y}_1$.

The power method requires that the first approximation $\mathbf{x}_0$ be a linear combination of eigenvectors. (In Example 8.5.1 the eigenvectors form a basis of $\mathbb{R}^2$.) But even in this case the method fails if $a_1 = 0$, where a_1 is the coefficient of the dominant eigenvector (try $\mathbf{x}_0 = \left[\begin{array}{r} -1 \\ 2 \end{array}\right]$ in Example 8.5.1). In general, the rate of convergence is quite slow if any of the ratios $\left|\frac{\lambda_i}{\lambda_1}\right|$ is near 1. Also, because the method requires repeated multiplications by A, it is not recommended unless these multiplications are easy to carry out (for example, if most of the entries of A are zero).

QR-Algorithm

A much better method for approximating the eigenvalues of an invertible matrix A depends on the factorization (using the Gram-Schmidt algorithm) of A in the form

$$A = QR$$

where Q is orthogonal and R is invertible and upper triangular (see Theorem 8.4.2). The **QR-algorithm** uses this repeatedly to create a sequence of matrices $A_1 = A, A_2, A_3, \dots,$ as follows:

1. Define $A_1 = A$ and factor it as $A_1 = Q_1R_1$.
2. Define $A_2 = R_1Q_1$ and factor it as $A_2 = Q_2R_2$.
3. Define $A_3 = R_2Q_2$ and factor it as $A_3 = Q_3R_3$.

$$\vdots$$

In general, A_k is factored as $A_k = Q_kR_k$ and we define $A_{k+1} = R_kQ_k$. Then A_{k+1} is similar to A_k [in fact, $A_{k+1} = R_kQ_k = (Q_k^{-1}A_k)Q_k$], and hence each A_k has the same eigenvalues as A. If the eigenvalues of A are real and have distinct absolute values, the remarkable thing is that the sequence of matrices $A_1, A_2, A_3, \dots$ converges to an upper triangular matrix with these eigenvalues on the main diagonal. [See below for the case of complex eigenvalues.]

Example 8.5.2

If $A = \left[\begin{array}{rr} 1 & 1 \\ 2 & 0 \end{array}\right]$ as in Example 8.5.1, use the QR-algorithm to approximate the eigenvalues.

Solution. The matrices A_1, A_2, and A_3 are as follows:

$$A_1 = \left[\begin{array}{rr} 1 & 1 \\ 2 & 0 \end{array}\right] = Q_1R_1 \quad \text{where } Q_1 = \frac{1}{\sqrt{5}}\left[\begin{array}{rr} 1 & 2 \\ 2 & -1 \end{array}\right] \text{ and } R_1 = \frac{1}{\sqrt{5}}\left[\begin{array}{rr} 5 & 1 \\ 0 & 2 \end{array}\right]$$

$$A_2 = \frac{1}{5}\left[\begin{array}{rr} 7 & 9 \\ 4 & -2 \end{array}\right] = \left[\begin{array}{rr} 1.4 & -1.8 \\ -0.8 & -0.4 \end{array}\right] = Q_2R_2$$

$$\text{where } Q_2 = \frac{1}{\sqrt{65}}\left[\begin{array}{rr} 7 & 4 \\ 4 & -7 \end{array}\right] \text{ and } R_2 = \frac{1}{\sqrt{65}}\left[\begin{array}{rr} 13 & 11 \\ 0 & 10 \end{array}\right]$$

$$A_3 = \frac{1}{13}\left[\begin{array}{rr} 27 & -5 \\ 8 & -14 \end{array}\right] = \left[\begin{array}{rr} 2.08 & -0.38 \\ 0.62 & -1.08 \end{array}\right]$$

This is converging to $\begin{bmatrix} 2 & * \\ 0 & -1 \end{bmatrix}$ and so is approximating the eigenvalues 2 and -1 on the main diagonal.

It is beyond the scope of this book to pursue a detailed discussion of these methods. The reader is referred to J. M. Wilkinson, *The Algebraic Eigenvalue Problem* (Oxford, England: Oxford University Press, 1965) or G. W. Stewart, *Introduction to Matrix Computations* (New York: Academic Press, 1973). We conclude with some remarks on the QR-algorithm.

Shifting. Convergence is accelerated if, at stage k of the algorithm, a number s_k is chosen and $A_k - s_k I$ is factored in the form $Q_k R_k$ rather than A_k itself. Then

$$Q_k^{-1} A_k Q_k = Q_k^{-1}(Q_k R_k + s_k I) Q_k = R_k Q_k + s_k I$$

so we take $A_{k+1} = R_k Q_k + s_k I$. If the shifts s_k are carefully chosen, convergence can be greatly improved.

Preliminary Preparation. A matrix such as

$$\begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix}$$

is said to be in **upper Hessenberg** form, and the QR-factorizations of such matrices are greatly simplified. Given an $n \times n$ matrix A, a series of orthogonal matrices $H_1, H_2, \dots, H_m$ (called **Householder matrices**) can be easily constructed such that

$$B = H_m^T \cdots H_1^T A H_1 \cdots H_m$$

is in upper Hessenberg form. Then the QR-algorithm can be efficiently applied to B and, because B is similar to A, it produces the eigenvalues of A.

Complex Eigenvalues. If some of the eigenvalues of a real matrix A are not real, the QR-algorithm converges to a block upper triangular matrix where the diagonal blocks are either 1×1 (the real eigenvalues) or 2×2 (each providing a pair of conjugate complex eigenvalues of A).

8.6 The Singular Value Decomposition

When working with a square matrix A it is clearly useful to be able to "diagonalize" A, that is to find a factorization $A = Q^{-1}DQ$ where Q is invertible and D is diagonal. Unfortunately such a factorization may not exist for A. However, even if A is not square gaussian elimination provides a factorization of the form $A = PDQ$ where P and Q are invertible and D is diagonal—the Smith Normal form (Theorem 2.5.3). However, if A is real we can choose P and Q to be *orthogonal* real matrices and D to be real. Such a factorization is called a **singular value decomposition (SVD)** for A, one of the most useful tools in applied linear algebra. In this Section we show how to explicitly compute an SVD for any real matrix A, and illustrate some of its many applications.

We need a fact about two subspaces associated with an $m \times n$ matrix A:

$$\operatorname{im} A = \{A\mathbf{x} \mid \mathbf{x} \text{ in } \mathbb{R}^n\} \qquad \text{and} \qquad \operatorname{col} A = \operatorname{span}\{\mathbf{a} \mid \mathbf{a} \text{ is a column of } A\}$$

Then $\operatorname{im} A$ is called the **image** of A (so named because of the linear transformation $\mathbb{R}^n \to \mathbb{R}^m$ with $\mathbf{x} \mapsto A\mathbf{x}$); and $\operatorname{col} A$ is called the **column space** of A (Definition 5.10). Surprisingly, these spaces are equal:

Lemma 8.6.1

For any $m \times n$ matrix A, $\operatorname{im} A = \operatorname{col} A$.

Proof. Let $A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix}$ in terms of its columns. Let $\mathbf{x} \in \operatorname{im} A$, say $\mathbf{x} = A\mathbf{y}$, $\mathbf{y}$ in $\mathbb{R}^n$. If

$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^T$, then $A\mathbf{y} = y_1\mathbf{a}_1 + y_2\mathbf{a}_2 + \cdots + y_n\mathbf{a}_n \in \text{col}\, A$ by Definition 2.5. This shows that $\text{im}\, A \subseteq \text{col}\, A$. For the other inclusion, each $\mathbf{a}_k = A\mathbf{e}_k$ where $\mathbf{e}_k$ is column k of I_n. □

8.6.1 Singular Value Decompositions

We know a lot about any real symmetric matrix: Its eigenvalues are real (Theorem 5.5.7), and it is orthogonally diagonalizable by the Principal Axes Theorem (Theorem 8.2.2). So for any real matrix A (square or not), the fact that both A^TA and AA^T are real and symmetric suggests that we can learn a lot about A by studying them. This section shows just how true this is.

The following Lemma reveals some similarities between A^TA and AA^T which simplify the statement and the proof of the SVD we are constructing.

Lemma 8.6.2

Let A be a real $m \times n$ matrix. Then:

1. *The eigenvalues of A^TA and AA^T are real and non-negative.*
2. *A^TA and AA^T have the same set of positive eigenvalues.*

Proof.

1. Since both matrices A^TA and AA^T are real and symmetric, then their eigenvalues are also real by Theorem 5.5.7. Not let λ be an eigenvalue of A^TA, with eigenvector $\mathbf{0} \neq \mathbf{q} \in \mathbb{R}^n$. Then:
$$\|A\mathbf{q}\|^2 = (A\mathbf{q})^T(A\mathbf{q}) = \mathbf{q}^T(A^TA\mathbf{q}) = \mathbf{q}^T(\lambda\mathbf{q}) = \lambda(\mathbf{q}^T\mathbf{q}) = \lambda\|\mathbf{q}\|^2$$
Then (1.) follows for A^TA, and the case AA^T follows by replacing A by A^T.

2. Write $N(B)$ for the set of positive eigenvalues of a matrix B. We must show that $N(A^TA) = N(AA^T)$. If $\lambda \in N(A^TA)$ with eigenvector $\mathbf{0} \neq \mathbf{q} \in \mathbb{R}^n$, then $A\mathbf{q} \in \mathbb{R}^m$ and
$$AA^T(A\mathbf{q}) = A[(A^TA)\mathbf{q}] = A(\lambda\mathbf{q}) = \lambda(A\mathbf{q})$$
Moreover, $A\mathbf{q} \neq \mathbf{0}$ since $A^TA\mathbf{q} = \lambda\mathbf{q} \neq \mathbf{0}$ as both $\lambda \neq 0$ and $\mathbf{q} \neq \mathbf{0}$. Hence λ is also a positive eigenvalue of AA^T, proving $N(A^TA) \subseteq N(AA^T)$. For the other inclusion replace A by A^T.

□

To analyze an $m \times n$ matrix A we have two symmetric matrices to work with: A^TA and AA^T. In view of Lemma 8.6.2, we choose A^TA (sometimes called the **Gram** matrix of A), and derive a series of facts which we will need. This narrative is a bit long, but trust that it will be worth the effort. We parse it out in several steps:

1. The $n \times n$ matrix A^TA is real and symmetric so, by the Principal Axes Theorem 8.2.2, let $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\} \subseteq \mathbb{R}^n$ be an orthonormal basis of eigenvectors of A^TA, with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. By Lemma 8.6.2(1), λ_i is real for each i and $\lambda_i \geq 0$. By re-ordering the $\mathbf{q}_i$ we may (and do) assume that
$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0 \quad \text{and}^{8} \quad \lambda_i = 0 \text{ if } i > r \tag{i}$$

[8] Of course they could *all* be positive ($r = n$) or *all* zero (so $A^TA = 0$, and hence $A = 0$ by Exercise **??**).

By Theorems 8.2.1 and 3.4.1, the matrix

$$Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_n \end{bmatrix} \text{ is orthogonal} \quad \text{and} \quad \text{orthogonally diagonalizes } A^TA. \tag{ii}$$

2. Even though the λ_i are the eigenvalues of A^TA, the number r in (**i**) turns out to be rank A. To understand why, consider the vectors $A\mathbf{q}_i \in \text{im } A$. For all i, j:

$$A\mathbf{q}_i \cdot A\mathbf{q}_j = (A\mathbf{q}_i)^T A\mathbf{q}_j = \mathbf{q}_i^T(A^TA)\mathbf{q}_j = \mathbf{q}_i^T(\lambda_j\mathbf{q}_j) = \lambda_j(\mathbf{q}_i^T\mathbf{q}_j) = \lambda_j(\mathbf{q}_i \cdot \mathbf{q}_j)$$

Because $\{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n\}$ is an orthonormal set, this gives

$$A\mathbf{q}_i \cdot A\mathbf{q}_j = 0 \text{ if } i \neq j \quad \text{and} \quad \|A\mathbf{q}_i\|^2 = \lambda_i\|\mathbf{q}_i\|^2 = \lambda_i \text{ for each } i \tag{iii}$$

We can extract two conclusions from (**iii**) and (**i**):

$$\{A\mathbf{q}_1, A\mathbf{q}_2, \dots, A\mathbf{q}_r\} \subseteq \text{im } A \text{ is an orthogonal set} \quad \text{and} \quad A\mathbf{q}_i = \mathbf{0} \text{ if } i > r \tag{iv}$$

With this write $U = \text{span}\{A\mathbf{q}_1, A\mathbf{q}_2, \dots, A\mathbf{q}_r\} \subseteq \text{im } A$; we claim that $U = \text{im } A$, that is $\text{im } A \subseteq U$. For this we must show that $A\mathbf{x} \in U$ for each $\mathbf{x} \in \mathbb{R}^n$. Since $\{\mathbf{q}_1, \dots, \mathbf{q}_r, \dots, \mathbf{q}_n\}$ is a basis of $\mathbb{R}^n$ (it is orthonormal), we can write $\mathbf{x} = t_1\mathbf{q}_1 + \cdots + t_r\mathbf{q}_r + \cdots + t_n\mathbf{q}_n$ where each $t_j \in \mathbb{R}$. Then, using (**iv**) we obtain

$$A\mathbf{x} = t_1A\mathbf{q}_1 + \cdots + t_rA\mathbf{q}_r + \cdots + t_nA\mathbf{q}_n = t_1A\mathbf{q}_1 + \cdots + t_rA\mathbf{q}_r \in U$$

This shows that $U = \text{im } A$, and so

$$\{A\mathbf{q}_1, A\mathbf{q}_2, \dots, A\mathbf{q}_r\} \text{ is an } \textit{orthogonal} \text{ basis of } \text{im}(A) \tag{v}$$

But $\text{col } A = \text{im } A$ by Lemma 8.6.1, and $\text{rank } A = \dim(\text{col } A)$ by Theorem 5.4.1, so

$$\text{rank } A = \dim(\text{col } A) = \dim(\text{im } A) \overset{\text{(v)}}{=} r \tag{vi}$$

3. Before proceeding, some definitions are in order:

Definition 8.7

The real numbers $\sigma_i = \sqrt{\lambda_i} \overset{\text{(iii)}}{=} \|A\mathbf{q}_i\|$ for $i = 1, 2, \dots, n$, are called the ***singular values*** *of the matrix A.*

Clearly $\sigma_1, \sigma_2, \dots, \sigma_r$ are the *positive* singular values of A. By (**i**) we have

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 \quad \text{and} \quad \sigma_i = 0 \text{ if } i > r \tag{vii}$$

With (**vi**) this makes the following definitions depend only upon A.

Definition 8.8

Let A be a real, $m \times n$ matrix of rank r, with positive singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ and $\sigma_i = 0$ if $i > r$. Define:

$$D_A = \operatorname{diag}(\sigma_1, \dots, \sigma_r) \qquad \text{and} \qquad \Sigma_A = \begin{bmatrix} D_A & 0 \\ 0 & 0 \end{bmatrix}_{m \times n}$$

Here Σ_A is in block form and is called the ***singular matrix*** *of A.*

The singular values σ_i and the matrices D_A and Σ_A will be referred to frequently below.

4. Returning to our narrative, normalize the vectors $A\mathbf{q}_1, A\mathbf{q}_2, \dots, A\mathbf{q}_r$, by defining

$$\mathbf{p}_i = \frac{1}{\|A\mathbf{q}_i\|} A\mathbf{q}_i \in \mathbb{R}^m \quad \text{for each } i = 1, 2, \dots, r \tag{viii}$$

By (**v**) and Lemma 8.6.1, we conclude that

$$\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_r\} \text{ is an } \textit{orthonormal} \text{ basis of } \operatorname{col} A \subseteq \mathbb{R}^m \tag{ix}$$

Employing the Gram-Schmidt algorithm (or otherwise), construct $\mathbf{p}_{r+1}, \dots, \mathbf{p}_m$ so that

$$\{\mathbf{p}_1, \dots, \mathbf{p}_r, \dots, \mathbf{p}_m\} \text{ is an orthonormal basis of } \mathbb{R}^m \tag{x}$$

5. By (**x**) and (**ii**) we have *two* orthogonal matrices

$$P = \begin{bmatrix} \mathbf{p}_1 & \cdots & \mathbf{p}_r & \cdots & \mathbf{p}_m \end{bmatrix} \text{ of size } m \times m \quad \text{and} \quad Q = \begin{bmatrix} \mathbf{q}_1 & \cdots & \mathbf{q}_r & \cdots & \mathbf{q}_n \end{bmatrix} \text{ of size } n \times n$$

These matrices are related. In fact we have:

$$\sigma_i \mathbf{p}_i = \sqrt{\lambda_i}\mathbf{p}_i \overset{\text{(iii)}}{=} \|A\mathbf{q}_i\|\mathbf{p}_i \overset{\text{(viii)}}{=} A\mathbf{q}_i \quad \text{for each } i = 1, 2, \dots, r \tag{xi}$$

This yields the following expression for AQ in terms of its columns:

$$AQ = \begin{bmatrix} A\mathbf{q}_1 & \cdots & A\mathbf{q}_r & A\mathbf{q}_{r+1} & \cdots & A\mathbf{q}_n \end{bmatrix} \overset{\text{(iv)}}{=} \begin{bmatrix} \sigma_1\mathbf{p}_1 & \cdots & \sigma_r\mathbf{p}_r & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \tag{xii}$$

Then we compute:

$$\begin{aligned} P\Sigma_A &= \begin{bmatrix} \mathbf{p}_1 & \cdots & \mathbf{p}_r & \mathbf{p}_{r+1} & \cdots & \mathbf{p}_m \end{bmatrix} \begin{bmatrix} \sigma_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1\mathbf{p}_1 & \cdots & \sigma_r\mathbf{p}_r & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \\ &\overset{\text{(xii)}}{=} AQ \end{aligned}$$

Finally, as $Q^{-1} = Q^T$ it follows that $A = P\Sigma_A Q^T$.

With this we can state the main theorem of this Section.

Theorem 8.6.1

Let A be a real $m \times n$ matrix, and let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ be the positive singular values of A. Then r is the rank of A and we have the factorization

$$A = P\Sigma_A Q^T \qquad \text{where } P \text{ and } Q \text{ are orthogonal matrices}$$

The factorization $A = P\Sigma_A Q^T$ in Theorem 8.6.1, where P and Q are orthogonal matrices, is called a *Singular Value Decomposition* (*SVD*) of A. This decomposition is not unique. For example if $r < m$ then the vectors $\mathbf{p}_{r+1}, \dots, \mathbf{p}_m$ can be *any* extension of $\{\mathbf{p}_1, \dots, \mathbf{p}_r\}$ to an orthonormal basis of $\mathbb{R}^m$, and each will lead to a different matrix P in the decomposition. For a more dramatic example, if $A = I_n$ then $\Sigma_A = I_n$, and $A = P\Sigma_A P^T$ is a SVD of A for *any* orthogonal $n \times n$ matrix P.

Example 8.6.1

Find a singular value decomposition for $A = \left[\begin{array}{rrr} 1 & 0 & 1 \\ -1 & 1 & 0 \end{array} \right]$.

Solution. We have $A^TA = \left[\begin{array}{rrr} 2 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{array} \right]$, so the characteristic polynomial is

$$c_{A^TA}(x) = \det \left[\begin{array}{ccc} x-2 & 1 & -1 \\ 1 & x-1 & 0 \\ -1 & 0 & x-1 \end{array} \right] = (x-3)(x-1)x$$

Hence the eigenvalues of A^TA (in descending order) are $\lambda_1 = 3$, $\lambda_2 = 1$ and $\lambda_3 = 0$ with, respectively, unit eigenvectors

$$\mathbf{q}_1 = \tfrac{1}{\sqrt{6}} \left[\begin{array}{r} 2 \\ -1 \\ 1 \end{array} \right], \quad \mathbf{q}_2 = \tfrac{1}{\sqrt{2}} \left[\begin{array}{r} 0 \\ 1 \\ 1 \end{array} \right], \quad \text{and} \quad \mathbf{q}_3 = \tfrac{1}{\sqrt{3}} \left[\begin{array}{r} -1 \\ -1 \\ 1 \end{array} \right]$$

It follows that the orthogonal matrix Q in Theorem 8.6.1 is

$$Q = \left[\begin{array}{ccc} \mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 \end{array} \right] = \tfrac{1}{\sqrt{6}} \left[\begin{array}{rrr} 2 & 0 & -\sqrt{2} \\ -1 & \sqrt{3} & -\sqrt{2} \\ 1 & \sqrt{3} & \sqrt{2} \end{array} \right]$$

The singular values here are $\sigma_1 = \sqrt{3}$, $\sigma_2 = 1$ and $\sigma_3 = 0$, so $\operatorname{rank}(A) = 2$—clear in this case—and the singular matrix is

$$\Sigma_A = \left[\begin{array}{ccc} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \end{array} \right] = \left[\begin{array}{ccc} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{array} \right]$$

So it remains to find the 2×2 orthogonal matrix P in Theorem 8.6.1. This involves the vectors

$$A\mathbf{q}_1 = \tfrac{\sqrt{6}}{2} \left[\begin{array}{r} 1 \\ -1 \end{array} \right], \quad A\mathbf{q}_2 = \tfrac{\sqrt{2}}{2} \left[\begin{array}{r} 1 \\ 1 \end{array} \right], \quad \text{and} \quad A\mathbf{q}_3 = \left[\begin{array}{r} 0 \\ 0 \end{array} \right]$$

Normalize $A\mathbf{q}_1$ and $A\mathbf{q}_2$ to get

$$\mathbf{p}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad \text{and} \quad \mathbf{p}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

In this case, $\{\mathbf{p}_1, \mathbf{p}_2\}$ is *already* a basis of $\mathbb{R}^2$ (so the Gram-Schmidt algorithm is not needed), and we have the 2×2 orthogonal matrix

$$P = \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

Finally (by Theorem 8.6.1) the singular value decomposition for A is

$$A = P\Sigma_A Q^T = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}\frac{1}{\sqrt{6}}\begin{bmatrix} 2 & -1 & 1 \\ 0 & \sqrt{3} & \sqrt{3} \\ -\sqrt{2} & -\sqrt{2} & \sqrt{2} \end{bmatrix}$$

Of course this can be confirmed by direct matrix multiplication.

Thus, computing an SVD for a real matrix A is a routine matter, and we now describe a systematic procedure for doing so.

Theorem: SVD Algorithm

Given a real $m \times n$ matrix A, find an SVD $A = P\Sigma_A Q^T$ as follows:

1. *Use the Diagonalization Algorithm (see page 156) to find the (real and non-negative) eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ of $A^T A$ with corresponding (orthonormal) eigenvectors $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n$. Reorder the $\mathbf{q}_i$ (if necessary) to ensure that the nonzero eigenvalues are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ and $\lambda_i = 0$ if $i > r$.*

2. *The integer r is the rank of the matrix A.*

3. *The $n \times n$ orthogonal matrix Q in the SVD is $Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_n \end{bmatrix}$.*

4. *Define $\mathbf{p}_i = \frac{1}{\|A\mathbf{q}_i\|}A\mathbf{q}_i$ for $i = 1, 2, \dots, r$ (where r is as in step 1). Then $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_r\}$ is orthonormal in $\mathbb{R}^m$ so (using Gram-Schmidt or otherwise) extend it to an orthonormal basis $\{\mathbf{p}_1, \dots, \mathbf{p}_r, \dots, \mathbf{p}_m\}$ in $\mathbb{R}^m$.*

5. *The $m \times m$ orthogonal matrix P in the SVD is $P = \begin{bmatrix} \mathbf{p}_1 & \cdots & \mathbf{p}_r & \cdots & \mathbf{p}_m \end{bmatrix}$.*

6. *The singular values for A are $\sigma_1, \sigma_2, \dots, \sigma_n$ where $\sigma_i = \sqrt{\lambda_i}$ for each i. Hence the nonzero singular values are $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, and so the singular matrix of A in the SVD is $\Sigma_A = \begin{bmatrix} \operatorname{diag}(\sigma_1, \dots, \sigma_r) & 0 \\ 0 & 0 \end{bmatrix}_{m \times n}$.*

7. *Thus $A = P\Sigma Q^T$ is a SVD for A.*

In practice the singular values σ_i, the matrices P and Q, and even the rank of an $m \times n$ matrix are not calculated this way. There are sophisticated numerical algorithms for calculating them maybe not exactly but to a high degree of accuracy. The reader is referred to books on numerical linear algebra.

So the main virtue of Theorem 8.6.1 is that it provides a way of *constructing* an SVD for every real matrix A. In particular it shows that every real matrix A *has* a singular value decomposition[9] in the following, more general, sense:

Definition 8.9

A ***Singular Value Decomposition (SVD)*** *of an $m \times n$ matrix A is a factorization $A = P\Sigma Q^T$ where P and Q are orthogonal and $\Sigma = \left[\begin{array}{cc} D & 0 \\ 0 & 0 \end{array} \right]_{m \times n}$ in block form where $D = \mathrm{diag}\,(d_1,\, d_2,\, \dots,\, d_r)$ where each $d_i > 0$, and $r \leq m$ and $r \leq n$.*

Note that for *any* SVD $A = P\Sigma Q^T$ we immediately obtain some information about A:

Lemma 8.6.3

If $A = P\Sigma Q^T$ is any SVD for A as in Definition 8.9, then:

1. $r = \mathrm{rank}\, A$.
2. *The numbers $d_1,\, d_2,\, \dots,\, d_r$ are the singular values of A in some order.*

Proof. Use the notation of Definition 8.9. We have

$$A^T A = (Q\Sigma^T P^T)(P\Sigma Q^T) = Q(\Sigma^T \Sigma)Q^T$$

so $\Sigma^T\Sigma$ and $A^T A$ are similar $n \times n$ matrices (Definition 5.12). Hence $r = \mathrm{rank}\, A$ by Corollary 5.4.3, proving (1.). Furthermore, $\Sigma^T\Sigma$ and $A^T A$ have the same eigenvalues by Theorem 5.5.1; that is (using (1.)):

$$\{d_1^2,\, d_2^2,\, \dots,\, d_r^2\} = \{\lambda_1,\, \lambda_2,\, \dots,\, \lambda_r\} \quad \text{are equal as sets}$$

where $\lambda_1,\, \lambda_2,\, \dots,\, \lambda_r$ are the positive eigenvalues of $A^T A$. Hence there is a permutation τ of $\{1,\, 2,\, \cdots,\, r\}$ such that $d_i^2 = \lambda_{i\tau}$ for each $i = 1,\, 2,\, \dots,\, r$. Hence $d_i = \sqrt{\lambda_{i\tau}} = \sigma_{i\tau}$ for each i by Definition 8.7. This proves (2.). □

We note in passing that more is true. Let A be $m \times n$ of rank r, and let $A = P\Sigma Q^T$ be any SVD for A. Using the proof of Lemma 8.6.3 we have $d_i = \sigma_{i\tau}$ for some permutation τ of $\{1,\, 2,\, \dots,\, r\}$. In fact, it can be shown that there exist orthogonal matrices P_1 and Q_1 obtained from P and Q by τ-permuting columns and rows respectively, such that $A = P_1\Sigma_A Q_1^T$ is an SVD of A.

---

[9] In fact every complex matrix has an SVD [J.T. Scheick, Linear Algebra with Applications, McGraw-Hill, 1997]

8.6.2 Fundamental Subspaces

It turns out that any singular value decomposition contains a great deal of information about an $m \times n$ matrix A and the subspaces associated with A. For example, in addition to Lemma 8.6.3, the set $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_r\}$ of vectors constructed in the proof of Theorem 8.6.1 is an orthonormal basis of $\text{col}\, A$ (by (**v**) and (**viii**) in the proof). There are more such examples, which is the thrust of this subsection. In particular, there are four subspaces associated to a real $m \times n$ matrix A that have come to be called fundamental:

Definition 8.10

*The **fundamental subspaces** of an $m \times n$ matrix A are:*

$$\text{row}\, A = \text{span}\, \{\mathbf{x} \mid \mathbf{x} \textit{ is a row of } A\}$$

$$\text{col}\, A = \text{span}\, \{\mathbf{x} \mid \mathbf{x} \textit{ is a column of } A\}$$

$$\text{null}\, A = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{0}\}$$

$$\text{null}\, A^T = \{\mathbf{x} \in \mathbb{R}^n \mid A^T\mathbf{x} = \mathbf{0}\}$$

If $A = P\Sigma Q^T$ is *any* SVD for the real $m \times n$ matrix A, then orthonormal bases for each of these fundamental subspaces can be obtained from the columns of P and Q. We are going to show how exactly, but first we need three properties related to the *orthogonal complement* $U^\perp$ of a subspace U of $\mathbb{R}^n$, where (Definition 8.1):

$$U^\perp = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{u} \cdot \mathbf{x} = 0 \text{ for all } \mathbf{u} \in U\}$$

The orthogonal complement plays an important role in the Projection Theorem (Theorem 8.1.3), and we return to it in Section 10.2. For now we need:

Lemma 8.6.4

If A is any matrix then:

1. $(\text{row}\, A)^\perp = \text{null}\, A$ *and* $(\text{col}\, A)^\perp = \text{null}\, A^T$.
2. *If U is any subspace of $\mathbb{R}^n$ then $U^{\perp\perp} = U$.*
3. *Let $\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$ be an orthonormal basis of $\mathbb{R}^m$. If $U = \text{span}\, \{\mathbf{f}_1, \dots, \mathbf{f}_k\}$, then*

$$U^\perp = \text{span}\, \{\mathbf{f}_{k+1}, \dots, \mathbf{f}_m\}$$

Proof.

1. Assume A is $m \times n$, and let $\mathbf{b}_1, \dots, \mathbf{b}_m$ be the rows of A. If $\mathbf{x}$ is a column in $\mathbb{R}^n$, then entry i of $A\mathbf{x}$ is $\mathbf{b}_i \cdot \mathbf{x}$, so $A\mathbf{x} = \mathbf{0}$ if and only if $\mathbf{b}_i \cdot \mathbf{x} = 0$ for each i. Thus:

$$\mathbf{x} \in \text{null}\, A \quad \Leftrightarrow \quad \mathbf{b}_i \cdot \mathbf{x} = 0 \text{ for each } i \quad \Leftrightarrow \quad \mathbf{x} \in (\text{span}\, \{\mathbf{b}_1, \dots, \mathbf{b}_m\})^\perp = (\text{row}\, A)^\perp$$

Hence $\text{null}\, A = (\text{row}\, A)^\perp$. Now replace A by A^T to get $\text{null}\, A^T = (\text{row}\, A^T)^\perp = (\text{col}\, A)^\perp$, which is the other identity in (1).

2. If $\mathbf{x} \in U$ then $\mathbf{y} \cdot \mathbf{x} = 0$ for all $\mathbf{y} \in U^{\perp}$, that is $\mathbf{x} \in U^{\perp\perp}$. This proves that $U \subseteq U^{\perp\perp}$, so it is enough to show that $\dim U = \dim U^{\perp\perp}$. By Theorem 8.1.4 we see that $\dim V^{\perp} = n - \dim V$ for any subspace $V \subseteq \mathbb{R}^n$. Hence

$$\dim U^{\perp\perp} = n - \dim U^{\perp} = n - (n - \dim U) = \dim U, \quad \text{as required}$$

3. We have $\operatorname{span}\{\mathbf{f}_{k+1}, \dots, \mathbf{f}_m\} \subseteq U^{\perp}$ because $\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$ is orthogonal. For the other inclusion, let $\mathbf{x} \in U^{\perp}$ so $\mathbf{f}_i \cdot \mathbf{x} = 0$ for $i = 1, 2, \dots, k$. By the Expansion Theorem 5.3.6:

$$\begin{array}{rcccccccccccc} \mathbf{x} & = & (\mathbf{f}_1 \cdot \mathbf{x})\mathbf{f}_1 & + & \cdots & + & (\mathbf{f}_k \cdot \mathbf{x})\mathbf{f}_k & + & (\mathbf{f}_{k+1} \cdot \mathbf{x})\mathbf{f}_{k+1} & + & \cdots & + & (\mathbf{f}_m \cdot \mathbf{x})\mathbf{f}_m \\ & = & \mathbf{0} & + & \cdots & + & \mathbf{0} & + & (\mathbf{f}_{k+1} \cdot \mathbf{x})\mathbf{f}_{k+1} & + & \cdots & + & (\mathbf{f}_m \cdot \mathbf{x})\mathbf{f}_m \end{array}$$

Hence $U^{\perp} \subseteq \operatorname{span}\{\mathbf{f}_{k+1}, \dots, \mathbf{f}_m\}$.

□

With this we can see how *any* SVD for a matrix A provides orthonormal bases for each of the four fundamental subspaces of A.

Theorem 8.6.2

Let A be an $m \times n$ real matrix, let $A = P\Sigma Q^T$ be any SVD for A where P and Q are orthogonal of size $m \times m$ and $n \times n$ respectively, and let

$$\Sigma = \left[\begin{array}{cc} D & 0 \\ 0 & 0 \end{array}\right]_{m \times n} \qquad \text{where} \qquad D = \operatorname{diag}(d_1, d_2, \dots, d_r), \ \text{ with each } d_i > 0$$

Write $P = \left[\begin{array}{ccccc} \mathbf{p}_1 & \cdots & \mathbf{p}_r & \cdots & \mathbf{p}_m \end{array}\right]$ and $Q = \left[\begin{array}{ccccc} \mathbf{q}_1 & \cdots & \mathbf{q}_r & \cdots & \mathbf{q}_n \end{array}\right]$, so $\{\mathbf{p}_1, \dots, \mathbf{p}_r, \dots, \mathbf{p}_m\}$ and $\{\mathbf{q}_1, \dots, \mathbf{q}_r, \dots, \mathbf{q}_n\}$ are orthonormal bases of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. Then

1. *$r = \operatorname{rank} A$, and the singular values of A are $d_1, d_2, \dots, d_r$.*
2. *The fundamental spaces are described as follows:*
   a. *$\{\mathbf{p}_1, \dots, \mathbf{p}_r\}$ is an orthonormal basis of $\operatorname{col} A$.*
   b. *$\{\mathbf{p}_{r+1}, \dots, \mathbf{p}_m\}$ is an orthonormal basis of $\operatorname{null} A^T$.*
   c. *$\{\mathbf{q}_{r+1}, \dots, \mathbf{q}_n\}$ is an orthonormal basis of $\operatorname{null} A$.*
   d. *$\{\mathbf{q}_1, \dots, \mathbf{q}_r\}$ is an orthonormal basis of $\operatorname{row} A$.*

Proof.

1. This is Lemma 8.6.3.
2. a. As $\operatorname{col} A = \operatorname{col}(AQ)$ by Lemma 5.4.3 and $AQ = P\Sigma$, (a.) follows from

$$P\Sigma = \left[\begin{array}{ccccc} \mathbf{p}_1 & \cdots & \mathbf{p}_r & \cdots & \mathbf{p}_m \end{array}\right] \left[\begin{array}{cc} \operatorname{diag}(d_1, d_2, \dots, d_r) & 0 \\ 0 & 0 \end{array}\right] = \left[\begin{array}{cccccc} d_1\mathbf{p}_1 & \cdots & d_r\mathbf{p}_r & \mathbf{0} & \cdots & \mathbf{0} \end{array}\right]$$

b. We have $(\text{col } A)^{\perp} \overset{\text{(a.)}}{=} (\text{span}\{\mathbf{p}_1, \dots, \mathbf{p}_r\})^{\perp} = \text{span}\{\mathbf{p}_{r+1}, \dots, \mathbf{p}_m\}$ by Lemma 8.6.4(3). This proves (b.) because $(\text{col } A)^{\perp} = \text{null } A^T$ by Lemma 8.6.4(1).

c. We have $\dim(\text{null } A) + \dim(\text{im } A) = n$ by the Dimension Theorem 7.2.4, applied to $T : \mathbb{R}^n \to \mathbb{R}^m$ where $T(\mathbf{x}) = A\mathbf{x}$. Since also $\text{im } A = \text{col } A$ by Lemma 8.6.1, we obtain

$$\dim(\text{null } A) = n - \dim(\text{col } A) = n - r = \dim(\text{span}\{\mathbf{q}_{r+1}, \dots, \mathbf{q}_n\})$$

So to prove (c.) it is enough to show that $\mathbf{q}_j \in \text{null } A$ whenever $j > r$. To this end write

$$d_{r+1} = \cdots = d_n = 0, \quad \text{so} \quad \Sigma^T\Sigma = \text{diag}(d_1^2, \dots, d_r^2, d_{r+1}^2, \dots, d_n^2)$$

Observe that each d_j is an eigenvalue of $\Sigma^T\Sigma$ with eigenvector $\mathbf{e}_j$ = column j of I_n. Thus $\mathbf{q}_j = Q\mathbf{e}_j$ for each j. As $A^TA = Q\Sigma^T\Sigma Q^T$ (proof of Lemma 8.6.3), we obtain

$$(A^TA)\mathbf{v}_j = (Q\Sigma^T\Sigma Q^T)(Q\mathbf{e}_j) = Q(\Sigma^T\Sigma\mathbf{e}_j) = Q\left(d_j^2\mathbf{e}_j\right) = d_j^2 Q\mathbf{e}_j = d_j^2\mathbf{q}_j$$

for $1 \leq j \leq n$. Thus each $\mathbf{q}_j$ is an eigenvector of A^TA corresponding to d_j^2. But then

$$\|A\mathbf{q}_j\|^2 = (A\mathbf{q}_j)^T A\mathbf{q}_j = \mathbf{q}_j^T(A^TA\mathbf{q}_j) = \mathbf{q}_j^T(d_j^2\mathbf{q}_j) = d_j^2\|\mathbf{q}_j\|^2 = d_j^2 \quad \text{for } i = 1, \dots, n$$

In particular, $A\mathbf{q}_j = \mathbf{0}$ whenever $j > r$, so $\mathbf{q}_j \in \text{null } A$ if $j > r$, as desired. This proves (c).

d. Observe that $\text{span}\{\mathbf{q}_{r+1}, \dots, \mathbf{q}_n\} \overset{\text{(c.)}}{=} \text{null } A = (\text{row } A)^{\perp}$ by Lemma 8.6.4(1). But then parts (2) and (3) of Lemma 8.6.4 show

$$\text{row } A = \left((\text{row } A)^{\perp}\right)^{\perp} = (\text{span}\{\mathbf{q}_{r+1}, \dots, \mathbf{q}_n\})^{\perp} = \text{span}\{\mathbf{q}_1, \dots, \mathbf{q}_r\}$$

This proves (d.), and hence Theorem 8.6.2.

□

Example 8.6.2

Consider the homogeneous linear system

$$A\mathbf{x} = \mathbf{0} \text{ of } m \text{ equations in } n \text{ variables}$$

Then the set of all solutions is $\text{null } A$. Hence if $A = P\Sigma Q^T$ is any SVD for A then (in the notation of Theorem 8.6.2) $\{\mathbf{q}_{r+1}, \dots, \mathbf{q}_n\}$ is an orthonormal basis of the set of solutions for the system. As such they are a set of **basic solutions** for the system, the most basic notion in Chapter 1.

8.6.3 The Polar Decomposition of a Real Square Matrix

If A is real and $n \times n$ the factorization in the title is related to the polar decomposition A. Unlike the SVD, in this case the decomposition is *uniquely* determined by A.

Recall (Section 8.3) that a symmetric matrix A is called positive definite if and only if $\mathbf{x}^T A\mathbf{x} > 0$ for every column $\mathbf{x} \neq \mathbf{0} \in \mathbb{R}^n$. Before proceeding, we must explore the following weaker notion:

Definition 8.11

A real $n \times n$ matrix G is called **positive**[10] *if it is symmetric and*

$$\mathbf{x}^T G\mathbf{x} \geq 0 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n$$

Clearly every positive definite matrix is positive, but the converse fails. Indeed, $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ is positive because, if $\mathbf{x} = \begin{bmatrix} a & b \end{bmatrix}^T$ in $\mathbb{R}^2$, then $\mathbf{x}^T A\mathbf{x} = (a+b)^2 \geq 0$. But $\mathbf{y}^T A\mathbf{y} = 0$ if $\mathbf{y} = \begin{bmatrix} 1 & -1 \end{bmatrix}^T$, so A is not positive definite.

Lemma 8.6.5

Let G denote an $n \times n$ positive matrix.

1. *If A is any $\times m$ matrix and G is positive, then $A^T GA$ is positive (and $m \times m$).*
2. *If $G = \operatorname{diag}(d_1, d_2, \cdots, d_n)$ and each $d_i \geq 0$ then G is positive.*

Proof.

1. $\mathbf{x}^T (A^T GA)\mathbf{x} = (A\mathbf{x})^T G(A\mathbf{x}) \geq 0$ because G is positive.
2. If $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T$, then

$$\mathbf{x}^T G\mathbf{x} = d_1 x_1^2 + d_2 x_2^2 + \cdots + d_n x_n^2 \geq 0$$

because $d_i \geq 0$ for each i.

□

Definition 8.12

If A is a real $n \times n$ matrix, a factorization

$$A = GQ \text{ where } G \text{ is positive and } Q \text{ is orthogonal}$$

is called a **polar decomposition** *for A.*

Any SVD for a real square matrix A yields a polar form for A.

[10] Also called **positive semi-definite.**

Theorem 8.6.3

Every square real matrix has a polar form.

Proof. Let $A = U\Sigma V^T$ be a SVD for A with Σ as in Definition 8.9 and $m = n$. Since $U^TU = I_n$ here we have

$$A = U\Sigma V^T = (U\Sigma)(U^TU)V^T = (U\Sigma U^T)(UV^T)$$

So if we write $G = U\Sigma U^T$ and $Q = UV^T$, then Q is orthogonal, and it remains to show that G is positive. But this follows from Lemma 8.6.5. □

The SVD for a square matrix A is not unique ($I_n = PI_nP^T$ for any orthogonal matrix P). But given the proof of Theorem 8.6.3 it is surprising that the polar decomposition *is* unique.[11] We omit the proof.

The name "polar form" is reminiscent of the same form for complex numbers (see Appendix A). This is no coincidence. To see why, we represent the complex numbers as real 2×2 matrices. Write $\mathbf{M}_2(\mathbb{R})$ for the set of all real 2×2 matrices, and define

$$\sigma : \mathbb{C} \to \mathbf{M}_2(\mathbb{R}) \quad \text{by} \quad \sigma(a+bi) = \left[\begin{array}{rr} a & -b \\ b & a \end{array}\right] \text{ for all } a+bi \text{ in } \mathbb{C}$$

One verifies that σ preserves addition and multiplication in the sense that

$$\sigma(zw) = \sigma(z)\sigma(w) \qquad \text{and} \qquad \sigma(z+w) = \sigma(z) + \sigma(w)$$

for all complex numbers z and w. Since θ is one-to-one we may *identify* each complex number $a+bi$ with the matrix $\theta(a+bi)$, that is we write

$$a+bi = \left[\begin{array}{rr} a & -b \\ b & a \end{array}\right] \quad \text{for all } a+bi \text{ in } \mathbb{C}$$

Thus $0 = \left[\begin{array}{rr} 0 & 0 \\ 0 & 0 \end{array}\right]$, $1 = \left[\begin{array}{rr} 1 & 0 \\ 0 & 1 \end{array}\right] = I_2$, $i = \left[\begin{array}{rr} 0 & -1 \\ 1 & 0 \end{array}\right]$, and $r = \left[\begin{array}{rr} r & 0 \\ 0 & r \end{array}\right]$ if r is real.

If $z = a+bi$ is nonzero then the *absolute value* $r = |z| = \sqrt{a^2+b^2} \neq 0$. If θ is the *angle* of z in standard position, then $\cos\theta = a/r$ and $\sin\theta = b/r$. Observe:

$$\left[\begin{array}{rr} a & -b \\ b & a \end{array}\right] = \left[\begin{array}{rr} r & 0 \\ 0 & r \end{array}\right]\left[\begin{array}{rr} a/r & -b/r \\ b/r & a/r \end{array}\right] = \left[\begin{array}{rr} r & 0 \\ 0 & r \end{array}\right]\left[\begin{array}{rr} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right] = GQ \tag{xiii}$$

where $G = \left[\begin{array}{rr} r & 0 \\ 0 & r \end{array}\right]$ is positive and $Q = \left[\begin{array}{rr} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right]$ is orthogonal. But in $\mathbb{C}$ we have $G = r$ and $Q = \cos\theta + i\sin\theta$ so (**xiii**) reads $z = r(\cos\theta + i\sin\theta) = re^{i\theta}$ which is the *classical polar form* for the complex number $a+bi$. This is why (**xiii**) is called the polar form of the matrix $\left[\begin{array}{rr} a & -b \\ b & a \end{array}\right]$; Definition 8.12 simply adopts the terminology for $n \times n$ matrices.

[11]See J.T. Scheick, Linear Algebra with Applications, McGraw-Hill, 1997, page 379.

8.6.4 The Pseudoinverse of a Matrix

It is impossible for a non-square matrix A to have an inverse (see the footnote to Definition 2.11). Nonetheless, one candidate for an "inverse" of A is an $m \times n$ matrix B such that

$$ABA = A \qquad \text{and} \qquad BAB = B$$

Such a matrix B is called a *middle inverse* for A. If A is invertible then A^{-1} is the unique middle inverse for A, but a middle inverse is not unique in general, even for square matrices. For example, if $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ then $B = \begin{bmatrix} 1 & 0 & 0 \\ b & 0 & 0 \end{bmatrix}$ is a middle inverse for A for any b.

If $ABA = A$ and $BAB = B$ it is easy to see that AB and BA are both idempotent matrices. In 1955 Roger Penrose observed that the middle inverse is unique if both AB and BA are symmetric. We omit the proof.

Theorem 8.6.4: Penrose' Theorem[12]

Given any real $m \times n$ matrix A, there is exactly one $n \times m$ matrix B such that A and B satisfy the following conditions:

P1 $ABA = A$ *and* $BAB = B$.

P2 *Both AB and BA are symmetric.*

Definition 8.13

Let A be a real $m \times n$ matrix. The **pseudoinverse** *of A is the unique $n \times m$ matrix A^+ such that A and A^+ satisfy* **P1** *and* **P2**, *that is:*

$$AA^+A = A, \qquad A^+AA^+ = A^+, \qquad \text{and both } AA^+ \text{ and } A^+A \text{ are symmetric}^{13}$$

If A is invertible then $A^+ = A^{-1}$ as expected. In general, the symmetry in conditions P1 and P2 shows that A is the pseudoinverse of A^+, that is $A^{++} = A$.

[12]R. Penrose, *A generalized inverse for matrices*, Proceedings of the Cambridge Philosophical Society **5l** (1955), 406-413. In fact Penrose proved this for any complex matrix, where AB and BA are both required to be hermitian (see Definition 8.18 in the following section).

[13]Penrose called the matrix A^+ the generalized inverse of A, but the term pseudoinverse is now commonly used. The matrix A^+ is also called the **Moore-Penrose** inverse after E.H. Moore who had the idea in 1935 as part of a larger work on "General Analysis". Penrose independently re-discovered it 20 years later.

Theorem 8.6.5

Let A be an $m \times n$ matrix.

1. *If* rank $A = m$ *then* AA^T *is invertible and* $A^+ = A^T(AA^T)^{-1}$.
2. *If* rank $A = n$ *then* A^TA *is invertible and* $A^+ = (A^TA)^{-1}A^T$.

Proof. Here AA^T (respectively A^TA) is invertible by Theorem 5.4.4 (respectively Theorem 5.4.3). The rest is a routine verification. □

In general, given an $m \times n$ matrix A, the pseudoinverse A^+ can be computed from any SVD for A. To see how, we need some notation. Let $A = P\Sigma Q^T$ be an SVD for A (as in Definition 8.9) where P and Q are orthogonal and $\Sigma = \left[\begin{array}{cc} D & 0 \\ 0 & 0 \end{array}\right]_{m\times n}$ in block form where $D = \operatorname{diag}(d_1, d_2, \dots, d_r)$ where each $d_i > 0$. Hence D is invertible, so we make:

Definition 8.14

$\Sigma' = \left[\begin{array}{cc} D^{-1} & 0 \\ 0 & 0 \end{array}\right]_{n\times m}$.

A routine calculation gives:

Lemma 8.6.6

- $\Sigma\Sigma'\Sigma = \Sigma$
- $\Sigma\Sigma' = \left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right]_{m\times m}$
- $\Sigma'\Sigma\Sigma' = \Sigma'$
- $\Sigma'\Sigma = \left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right]_{n\times n}$

That is, Σ' is the pseudoinverse of Σ.

Now given $A = P\Sigma Q^T$, define $B = Q\Sigma' P^T$. Then

$$ABA = (P\Sigma Q^T)(Q\Sigma' P^T)(P\Sigma Q^T) = P(\Sigma\Sigma'\Sigma)Q^T = P\Sigma V^T = A$$

by Lemma 8.6.6. Similarly $BAB = B$. Moreover $AB = P(\Sigma\Sigma')P^T$ and $BA = Q(\Sigma'\Sigma)Q^T$ are both symmetric again by Lemma 8.6.6. This proves

Theorem 8.6.6

Let A be real and $m \times n$, and let $A = P\Sigma Q^T$ is any SVD for A as in Definition 8.9. Then $A^+ = Q\Sigma' P^T$.

Of course we can always use the SVD constructed in Theorem 8.6.1 to find the pseudoinverse. If $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, we observed above that $B = \begin{bmatrix} 1 & 0 & 0 \\ b & 0 & 0 \end{bmatrix}$ is a middle inverse for A for any b. Furthermore AB is symmetric, and BA is symmetric exactly when $b = 0$. In this case, B is the pseudoinverse of A found in Example 8.6.3.

Example 8.6.3

Find A^+ if $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$.

Solution. $A^T A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ with eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 0$ and corresponding eigenvectors $\mathbf{q}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mathbf{q}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Hence $Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 \end{bmatrix} = I_2$. Also A has rank 1 with singular values $\sigma_1 = 1$ and $\sigma_2 = 0$, so $\Sigma_A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = A$ and $\Sigma_A' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = A^T$ in this case.

Since $A\mathbf{q}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ and $A\mathbf{q}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$, we have $\mathbf{p}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ which extends to an orthonormal basis $\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$ of $\mathbb{R}^3$ where (say) $\mathbf{p}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ and $\mathbf{p}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$. Hence $P = \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_3 \end{bmatrix} = I$, so the SVD for A is $A = P\Sigma_A Q^T$. Finally, the pseudoinverse of A is $A^+ = Q\Sigma_A' P^T = \Sigma_A' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$. Note that $A^+ = A^T$ in this case.

The following Lemma collects some properties of the pseudoinverse that mimic those of the inverse. Its verification is left as an exercise.

Lemma 8.6.7

Let A be an $m \times n$ matrix.

1. $A^{++} = A$.
2. *If A is invertible then* $A^+ = A^{-1}$.
3. $(A^T)^+ = (A^+)^T$.
4. $(kA)^+ = k^{-1}A^+$ *for any real* $k \neq 0$.
5. $(PAQ)^+ = P^T(A^+)Q^T$ *whenever P and Q are orthogonal.*

8.7 Complex Matrices

If A is an $n \times n$ matrix, the characteristic polynomial $c_A(x)$ is a polynomial of degree n and the eigenvalues of A are just the roots of $c_A(x)$. In most of our examples these roots have been *real* numbers (in fact, the examples have been carefully chosen so this will be the case!); but it need not happen, even when the characteristic polynomial has real coefficients. For example, if $A = \left[\begin{array}{rr} 0 & 1 \\ -1 & 0 \end{array} \right]$ then $c_A(x) = x^2 + 1$ has roots i and $-i$, where i is a complex number satisfying $i^2 = -1$. Therefore, we have to deal with the possibility that the eigenvalues of a (real) square matrix might be complex numbers.

In fact, nearly everything in this book would remain true if the phrase *real number* were replaced by *complex number* wherever it occurs. Then we would deal with matrices with complex entries, systems of linear equations with complex coefficients (and complex solutions), determinants of complex matrices, and vector spaces with scalar multiplication by any complex number allowed. Moreover, the proofs of most theorems about (the real version of) these concepts extend easily to the complex case. It is not our intention here to give a full treatment of complex linear algebra. However, we will carry the theory far enough to give another proof that the eigenvalues of a real symmetric matrix A are real (Theorem 5.5.7) and to prove the spectral theorem, an extension of the principal axes theorem (Theorem 8.2.2).

The set of complex numbers is denoted $\mathbb{C}$. We will use only the most basic properties of these numbers (mainly conjugation and absolute values), and the reader can find this material in Appendix A.

If $n \geq 1$, we denote the set of all n-tuples of complex numbers by $\mathbb{C}^n$. As with $\mathbb{R}^n$, these n-tuples will be written either as row or column matrices and will be referred to as **vectors**. We define vector operations on $\mathbb{C}^n$ as follows:

$$\begin{aligned}(v_1,\ v_2,\ \dots,\ v_n) + (w_1,\ w_2,\ \dots,\ w_n) &= (v_1 + w_1,\ v_2 + w_2,\ \dots,\ v_n + w_n) \\ u(v_1,\ v_2,\ \dots,\ v_n) &= (uv_1,\ uv_2,\ \dots,\ uv_n) \quad \text{for } u \text{ in } \mathbb{C}\end{aligned}$$

With these definitions, $\mathbb{C}^n$ satisfies the axioms for a vector space (with complex scalars) given in Chapter 6. Thus we can speak of spanning sets for $\mathbb{C}^n$, of linearly independent subsets, and of bases. In all cases, the definitions are identical to the real case, except that the scalars are allowed to be complex numbers. In particular, the standard basis of $\mathbb{R}^n$ remains a basis of $\mathbb{C}^n$, called the **standard basis** of $\mathbb{C}^n$.

A matrix $A = \left[a_{ij}\right]$ is called a **complex matrix** if every entry a_{ij} is a complex number. The notion of conjugation for complex numbers extends to matrices as follows: Define the **conjugate** of $A = \left[a_{ij}\right]$ to be the matrix

$$\overline{A} = \left[\ \overline{a}_{ij}\ \right]$$

obtained from A by conjugating every entry. Then (using Appendix A)

$$\overline{A+B} = \overline{A} + \overline{B} \quad \text{and} \quad \overline{AB} = \overline{A}\,\overline{B}$$

holds for all (complex) matrices of appropriate size.

The Standard Inner Product

There is a natural generalization to $\mathbb{C}^n$ of the dot product in $\mathbb{R}^n$.

Definition 8.15 Standard Inner Product in $\mathbb{R}^n$

*Given $\mathbf{z} = (z_1,\ z_2,\ \dots,\ z_n)$ and $\mathbf{w} = (w_1,\ w_2,\ \dots,\ w_n)$ in $\mathbb{C}^n$, define their **standard inner product** $\langle \mathbf{z},\ \mathbf{w} \rangle$ by*

$$\langle \mathbf{z},\ \mathbf{w} \rangle = z_1\overline{w}_1 + z_2\overline{w}_2 + \cdots + z_n\overline{w}_n = \mathbf{z} \cdot \overline{\mathbf{w}}$$

where $\overline{w}$ is the conjugate of the complex number w.

Clearly, if $\mathbf{z}$ and $\mathbf{w}$ actually lie in $\mathbb{R}^n$, then $\langle \mathbf{z},\ \mathbf{w} \rangle = \mathbf{z} \cdot \mathbf{w}$ is the usual dot product.

Example 8.7.1

If $\mathbf{z} = (2,\ 1-i,\ 2i,\ 3-i)$ and $\mathbf{w} = (1-i,\ -1,\ -i,\ 3+2i)$, then

$$\begin{aligned}\langle \mathbf{z},\ \mathbf{w} \rangle &= 2(1+i) + (1-i)(-1) + (2i)(i) + (3-i)(3-2i) = 6-6i \\ \langle \mathbf{z},\ \mathbf{z} \rangle &= 2 \cdot 2 + (1-i)(1+i) + (2i)(-2i) + (3-i)(3+i) = 20\end{aligned}$$

Note that $\langle \mathbf{z},\ \mathbf{w} \rangle$ is a complex number in general. However, if $\mathbf{w} = \mathbf{z} = (z_1,\ z_2,\ \dots,\ z_n)$, the definition gives $\langle \mathbf{z},\ \mathbf{z} \rangle = |z_1|^2 + \cdots + |z_n|^2$ which is a nonnegative real number, equal to 0 if and only if $\mathbf{z} = \mathbf{0}$. This explains the conjugation in the definition of $\langle \mathbf{z},\ \mathbf{w} \rangle$, and it gives (4) of the following theorem.

Theorem 8.7.1

Let $\mathbf{z}$, $\mathbf{z}_1$, $\mathbf{w}$, and $\mathbf{w}_1$ denote vectors in $\mathbb{C}^n$, and let λ denote a complex number.

1. $\langle \mathbf{z}+\mathbf{z}_1,\ \mathbf{w}\rangle = \langle \mathbf{z},\ \mathbf{w}\rangle + \langle \mathbf{z}_1,\ \mathbf{w}\rangle$ *and* $\langle \mathbf{z},\ \mathbf{w}+\mathbf{w}_1\rangle = \langle \mathbf{z},\ \mathbf{w}\rangle + \langle \mathbf{z},\ \mathbf{w}_1\rangle$.
2. $\langle \lambda\mathbf{z},\ \mathbf{w}\rangle = \lambda\langle \mathbf{z},\ \mathbf{w}\rangle$ *and* $\langle \mathbf{z},\ \lambda\mathbf{w}\rangle = \overline{\lambda}\langle \mathbf{z},\ \mathbf{w}\rangle$.
3. $\langle \mathbf{z},\ \mathbf{w}\rangle = \overline{\langle \mathbf{w},\ \mathbf{z}\rangle}$.
4. $\langle \mathbf{z},\ \mathbf{z}\rangle \geq 0$, *and* $\langle \mathbf{z},\ \mathbf{z}\rangle = 0$ *if and only if* $\mathbf{z} = \mathbf{0}$.

Proof. We leave (1) and (2) to the reader (Exercise **??**), and (4) has already been proved. To prove (3), write $\mathbf{z} = (z_1,\ z_2,\ \dots,\ z_n)$ and $\mathbf{w} = (w_1,\ w_2,\ \dots,\ w_n)$. Then

$$\begin{aligned}\overline{\langle \mathbf{w},\ \mathbf{z}\rangle} = (\overline{w_1\overline{z}_1 + \cdots + w_n\overline{z}_n}) &= \overline{w}_1\overline{\overline{z}}_1 + \cdots + \overline{w}_n\overline{\overline{z}}_n \\ &= z_1\overline{w}_1 + \cdots + z_n\overline{w}_n = \langle \mathbf{z},\ \mathbf{w}\rangle\end{aligned}$$

□

Definition 8.16 Norm and Length in $\mathbb{C}^n$

*As for the dot product on $\mathbb{R}^n$, property (4) enables us to define the **norm** or **length** $\|\mathbf{z}\|$ of a vector $\mathbf{z} = (z_1,\ z_2,\ \dots,\ z_n)$ in $\mathbb{C}^n$:*

$$\|\mathbf{z}\| = \sqrt{\langle \mathbf{z},\ \mathbf{z}\rangle} = \sqrt{|z_1|^2 + |z_2|^2 + \cdots + |z_n|^2}$$

The only properties of the norm function we will need are the following (the proofs are left to the reader):

Theorem 8.7.2

If $\mathbf{z}$ is any vector in $\mathbb{C}^n$, then

1. $\|\mathbf{z}\| \geq 0$ *and* $\|\mathbf{z}\| = 0$ *if and only if* $\mathbf{z} = \mathbf{0}$.
2. $\|\lambda\mathbf{z}\| = |\lambda|\|\mathbf{z}\|$ *for all complex numbers* λ.

A vector $\mathbf{u}$ in $\mathbb{C}^n$ is called a **unit vector** if $\|\mathbf{u}\| = 1$. Property (2) in Theorem 8.7.2 then shows that if $\mathbf{z} \neq \mathbf{0}$ is any nonzero vector in $\mathbb{C}^n$, then $\mathbf{u} = \frac{1}{\|\mathbf{z}\|}\mathbf{z}$ is a unit vector.

Example 8.7.2

In $\mathbb{C}^4$, find a unit vector $\mathbf{u}$ that is a positive real multiple of $\mathbf{z} = (1-i,\ i,\ 2,\ 3+4i)$.

Solution. $\|\mathbf{z}\| = \sqrt{2+1+4+25} = \sqrt{32} = 4\sqrt{2}$, so take $\mathbf{u} = \frac{1}{4\sqrt{2}}\mathbf{z}$.

Transposition of complex matrices is defined just as in the real case, and the following notion is fundamental.

Definition 8.17 Conjugate Transpose in $\mathbb{C}^n$

The ***conjugate transpose*** A^H *of a complex matrix* A *is defined by*

$$A^H = (\overline{A})^T = \overline{(A^T)}$$

Observe that $A^H = A^T$ when A is real.[14]

Example 8.7.3

$$\left[\begin{array}{ccc} 3 & 1-i & 2+i \\ 2i & 5+2i & -i \end{array}\right]^H = \left[\begin{array}{cc} 3 & -2i \\ 1+i & 5-2i \\ 2-i & i \end{array}\right]$$

The following properties of A^H follow easily from the rules for transposition of real matrices and extend these rules to complex matrices. Note the conjugate in property (3).

Theorem 8.7.3

Let A *and* B *denote complex matrices, and let* λ *be a complex number.*

1. $(A^H)^H = A$.
2. $(A+B)^H = A^H + B^H$.
3. $(\lambda A)^H = \overline{\lambda} A^H$.
4. $(AB)^H = B^H A^H$.

Hermitian and Unitary Matrices

If A is a real symmetric matrix, it is clear that $A^H = A$. The complex matrices that satisfy this condition turn out to be the most natural generalization of the real symmetric matrices:

Definition 8.18 Hermitian Matrices

A square complex matrix A *is called* ***hermitian***[15] *if* $A^H = A$*, equivalently if* $\overline{A} = A^T$.

Hermitian matrices are easy to recognize because the entries on the main diagonal must be real, and the "reflection" of each nondiagonal entry in the main diagonal must be the conjugate of that entry.

[14] Other notations for A^H are A^* and $A^\dagger$.

[15] The name hermitian honours Charles Hermite (1822–1901), a French mathematician who worked primarily in analysis and is remembered as the first to show that the number e from calculus is transcendental—that is, e is not a root of any polynomial with integer coefficients.

Example 8.7.4

$\begin{bmatrix} 3 & i & 2+i \\ -i & -2 & -7 \\ 2-i & -7 & 1 \end{bmatrix}$ is hermitian, whereas $\begin{bmatrix} 1 & i \\ i & -2 \end{bmatrix}$ and $\begin{bmatrix} 1 & i \\ -i & i \end{bmatrix}$ are not.

The following Theorem extends Theorem 8.2.3, and gives a very useful characterization of hermitian matrices in terms of the standard inner product in $\mathbb{C}^n$.

Theorem 8.7.4

An $n \times n$ complex matrix A is hermitian if and only if

$$\langle A\mathbf{z},\ \mathbf{w}\rangle = \langle \mathbf{z},\ A\mathbf{w}\rangle$$

for all n-tuples $\mathbf{z}$ and $\mathbf{w}$ in $\mathbb{C}^n$.

Proof. If A is hermitian, we have $A^T = \overline{A}$. If $\mathbf{z}$ and $\mathbf{w}$ are columns in $\mathbb{C}^n$, then $\langle \mathbf{z},\ \mathbf{w}\rangle = \mathbf{z}^T\overline{\mathbf{w}}$, so

$$\langle A\mathbf{z},\ \mathbf{w}\rangle = (A\mathbf{z})^T\overline{\mathbf{w}} = \mathbf{z}^T A^T \overline{\mathbf{w}} = \mathbf{z}^T\overline{A}\overline{\mathbf{w}} = \mathbf{z}^T(\overline{A\mathbf{w}}) = \langle \mathbf{z},\ A\mathbf{w}\rangle$$

To prove the converse, let $\mathbf{e}_j$ denote column j of the identity matrix. If $A = \left[a_{ij}\right]$, the condition gives

$$\overline{a}_{ij} = \langle \mathbf{e}_i,\ A\mathbf{e}_j\rangle = \langle A\mathbf{e}_i,\ \mathbf{e}_j\rangle = a_{ij}$$

Hence $\overline{A} = A^T$, so A is hermitian. □

Let A be an $n \times n$ complex matrix. As in the real case, a complex number λ is called an **eigenvalue** of A if $A\mathbf{x} = \lambda\mathbf{x}$ holds for some column $\mathbf{x} \neq \mathbf{0}$ in $\mathbb{C}^n$. In this case $\mathbf{x}$ is called an **eigenvector** of A corresponding to λ. The **characteristic polynomial** $c_A(x)$ is defined by

$$c_A(x) = \det(xI - A)$$

This polynomial has complex coefficients (possibly nonreal). However, the proof of Theorem 3.3.2 goes through to show that the eigenvalues of A are the roots (possibly complex) of $c_A(x)$.

It is at this point that the advantage of working with complex numbers becomes apparent. The real numbers are incomplete in the sense that the characteristic polynomial of a real matrix may fail to have all its roots real. However, this difficulty does not occur for the complex numbers. The so-called fundamental theorem of algebra ensures that *every* polynomial of positive degree with complex coefficients has a complex root. Hence every square complex matrix A has a (complex) eigenvalue. Indeed (Appendix A), $c_A(x)$ factors completely as follows:

$$c_A(x) = (x - \lambda_1)(x - \lambda_2)\cdots(x - \lambda_n)$$

where $\lambda_1,\ \lambda_2,\ \dots,\ \lambda_n$ are the eigenvalues of A (with possible repetitions due to multiple roots).

The next result shows that, for hermitian matrices, the eigenvalues are actually real. Because symmetric real matrices are hermitian, this re-proves Theorem 5.5.7. It also extends Theorem 8.2.4, which asserts that eigenvectors of a symmetric real matrix corresponding to distinct eigenvalues are actually orthogonal. In the complex context, two n-tuples $\mathbf{z}$ and $\mathbf{w}$ in $\mathbb{C}^n$ are said to be **orthogonal** if $\langle \mathbf{z},\ \mathbf{w}\rangle = 0$.

Theorem 8.7.5

Let A denote a hermitian matrix.

1. *The eigenvalues of A are real.*
2. *Eigenvectors of A corresponding to distinct eigenvalues are orthogonal.*

Proof. Let λ and μ be eigenvalues of A with (nonzero) eigenvectors $\mathbf{z}$ and $\mathbf{w}$. Then $A\mathbf{z} = \lambda\mathbf{z}$ and $A\mathbf{w} = \mu\mathbf{w}$, so Theorem 8.7.4 gives

$$\lambda\langle\mathbf{z},\ \mathbf{w}\rangle = \langle\lambda\mathbf{z},\ \mathbf{w}\rangle = \langle A\mathbf{z},\ \mathbf{w}\rangle = \langle\mathbf{z},\ A\mathbf{w}\rangle = \langle\mathbf{z},\ \mu\mathbf{w}\rangle = \overline{\mu}\langle\mathbf{z},\ \mathbf{w}\rangle \tag{8.6}$$

If $\mu = \lambda$ and $\mathbf{w} = \mathbf{z}$, this becomes $\lambda\langle\mathbf{z},\ \mathbf{z}\rangle = \overline{\lambda}\langle\mathbf{z},\ \mathbf{z}\rangle$. Because $\langle\mathbf{z},\ \mathbf{z}\rangle = \|\mathbf{z}\|^2 \neq 0$, this implies $\lambda = \overline{\lambda}$. Thus λ is real, proving (1). Similarly, μ is real, so equation (8.6) gives $\lambda\langle\mathbf{z},\ \mathbf{w}\rangle = \mu\langle\mathbf{z},\ \mathbf{w}\rangle$. If $\lambda \neq \mu$, this implies $\langle\mathbf{z},\ \mathbf{w}\rangle = 0$, proving (2). □

The principal axes theorem (Theorem 8.2.2) asserts that every real symmetric matrix A is orthogonally diagonalizable—that is P^TAP is diagonal where P is an orthogonal matrix $(P^{-1} = P^T)$. The next theorem identifies the complex analogs of these orthogonal real matrices.

Definition 8.19 Orthogonal and Orthonormal Vectors in $\mathbb{C}^n$

*As in the real case, a set of nonzero vectors $\{\mathbf{z}_1,\ \mathbf{z}_2,\ \dots,\ \mathbf{z}_m\}$ in $\mathbb{C}^n$ is called **orthogonal** if $\langle\mathbf{z}_i,\ \mathbf{z}_j\rangle = 0$ whenever $i \neq j$, and it is **orthonormal** if, in addition, $\|\mathbf{z}_i\| = 1$ for each i.*

Theorem 8.7.6

The following are equivalent for an $n \times n$ complex matrix A.

1. *A is invertible and $A^{-1} = A^H$.*
2. *The rows of A are an orthonormal set in $\mathbb{C}^n$.*
3. *The columns of A are an orthonormal set in $\mathbb{C}^n$.*

Proof. If $A = \left[\begin{array}{cccc} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_n \end{array}\right]$ is a complex matrix with jth column $\mathbf{c}_j$, then $A^T\overline{A} = \left[\langle\mathbf{c}_i,\ \mathbf{c}_j\rangle\right]$, as in Theorem 8.2.1. Now (1) ⇔ (2) follows, and (1) ⇔ (3) is proved in the same way. □

Definition 8.20 Unitary Matrices

*A square complex matrix U is called **unitary** if $U^{-1} = U^H$.*

Thus a real matrix is unitary if and only if it is orthogonal.

Example 8.7.5

The matrix $A = \begin{bmatrix} 1+i & 1 \\ 1-i & i \end{bmatrix}$ has orthogonal columns, but the rows are not orthogonal.

Normalizing the columns gives the unitary matrix $\frac{1}{2}\begin{bmatrix} 1+i & \sqrt{2} \\ 1-i & \sqrt{2}i \end{bmatrix}$.

Given a real symmetric matrix A, the diagonalization algorithm in Section 3.3 leads to a procedure for finding an orthogonal matrix P such that P^TAP is diagonal (see Example 8.2.4). The following example illustrates Theorem 8.7.5 and shows that the technique works for complex matrices.

Example 8.7.6

Consider the hermitian matrix $A = \begin{bmatrix} 3 & 2+i \\ 2-i & 7 \end{bmatrix}$. Find the eigenvalues of A, find two orthonormal eigenvectors, and so find a unitary matrix U such that U^HAU is diagonal.

Solution. The characteristic polynomial of A is

$$c_A(x) = \det(xI - A) = \det \begin{bmatrix} x-3 & -2-i \\ -2+i & x-7 \end{bmatrix} = (x-2)(x-8)$$

Hence the eigenvalues are 2 and 8 (both real as expected), and corresponding eigenvectors are $\begin{bmatrix} 2+i \\ -1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 2-i \end{bmatrix}$ (orthogonal as expected). Each has length $\sqrt{6}$ so, as in the (real) diagonalization algorithm, let $U = \frac{1}{\sqrt{6}}\begin{bmatrix} 2+i & 1 \\ -1 & 2-i \end{bmatrix}$ be the unitary matrix with the normalized eigenvectors as columns.

Then $U^HAU = \begin{bmatrix} 2 & 0 \\ 0 & 8 \end{bmatrix}$ is diagonal.

Unitary Diagonalization

An $n \times n$ complex matrix A is called **unitarily diagonalizable** if U^HAU is diagonal for some unitary matrix U. As Example 8.7.6 suggests, we are going to prove that every hermitian matrix is unitarily diagonalizable. However, with only a little extra effort, we can get a very important theorem that has this result as an easy consequence.

A complex matrix is called **upper triangular** if every entry below the main diagonal is zero. We owe the following theorem to Issai Schur.[16]

[16] Issai Schur (1875–1941) was a German mathematician who did fundamental work in the theory of representations of groups as matrices.

Theorem 8.7.7: Schur's Theorem

If A is any $n \times n$ complex matrix, there exists a unitary matrix U such that

$$U^H AU = T$$

is upper triangular. Moreover, the entries on the main diagonal of T are the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ of A (including multiplicities).

Proof. We use induction on n. If $n = 1$, A is already upper triangular. If $n > 1$, assume the theorem is valid for $(n-1) \times (n-1)$ complex matrices. Let λ_1 be an eigenvalue of A, and let $\mathbf{y}_1$ be an eigenvector with $\|\mathbf{y}_1\| = 1$. Then $\mathbf{y}_1$ is part of a basis of $\mathbb{C}^n$ (by the analog of Theorem 6.4.1), so the (complex analog of the) Gram-Schmidt process provides $\mathbf{y}_2, \dots, \mathbf{y}_n$ such that $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}$ is an orthonormal basis of $\mathbb{C}^n$. If $U_1 = \left[\begin{array}{cccc} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_n \end{array}\right]$ is the matrix with these vectors as its columns, then (see Lemma 5.4.3)

$$U_1^H AU_1 = \left[\begin{array}{cc} \lambda_1 & X_1 \\ 0 & A_1 \end{array}\right]$$

in block form. Now apply induction to find a unitary $(n-1) \times (n-1)$ matrix W_1 such that $W_1^H A_1 W_1 = T_1$ is upper triangular. Then $U_2 = \left[\begin{array}{cc} 1 & 0 \\ 0 & W_1 \end{array}\right]$ is a unitary $n \times n$ matrix. Hence $U = U_1 U_2$ is unitary (using Theorem 8.7.6), and

$$\begin{aligned} U^H AU &= U_2^H (U_1^H AU_1) U_2 \\ &= \left[\begin{array}{cc} 1 & 0 \\ 0 & W_1^H \end{array}\right] \left[\begin{array}{cc} \lambda_1 & X_1 \\ 0 & A_1 \end{array}\right] \left[\begin{array}{cc} 1 & 0 \\ 0 & W_1 \end{array}\right] = \left[\begin{array}{cc} \lambda_1 & X_1 W_1 \\ 0 & T_1 \end{array}\right] \end{aligned}$$

is upper triangular. Finally, A and $U^H AU = T$ have the same eigenvalues by (the complex version of) Theorem 5.5.1, and they are the diagonal entries of T because T is upper triangular. □

The fact that similar matrices have the same traces and determinants gives the following consequence of Schur's theorem.

Corollary 8.7.1

Let A be an $n \times n$ complex matrix, and let $\lambda_1, \lambda_2, \dots, \lambda_n$ denote the eigenvalues of A, including multiplicities. Then

$$\det A = \lambda_1 \lambda_2 \cdots \lambda_n \quad \textit{and} \quad \operatorname{tr} A = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

Schur's theorem asserts that every complex matrix can be "unitarily triangularized." However, we cannot substitute "unitarily diagonalized" here. In fact, if $A = \left[\begin{array}{cc} 1 & 1 \\ 0 & 1 \end{array}\right]$, there is no invertible complex matrix U at all such that $U^{-1}AU$ is diagonal. However, the situation is much better for hermitian matrices.

Theorem 8.7.8: Spectral Theorem

If A is hermitian, there is a unitary matrix U such that $U^H AU$ is diagonal.

Proof. By Schur's theorem, let $U^HAU = T$ be upper triangular where U is unitary. Since A is hermitian, this gives

$$T^H = (U^HAU)^H = U^HA^HU^{HH} = U^HAU = T$$

This means that T is both upper and lower triangular. Hence T is actually diagonal. □

The principal axes theorem asserts that a real matrix A is symmetric if and only if it is orthogonally diagonalizable (that is, P^TAP is diagonal for some real orthogonal matrix P). Theorem 8.7.8 is the complex analog of half of this result. However, the converse is false for complex matrices: There exist unitarily diagonalizable matrices that are not hermitian.

Example 8.7.7

Show that the non-hermitian matrix $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ is unitarily diagonalizable.

Solution. The characteristic polynomial is $c_A(x) = x^2 + 1$. Hence the eigenvalues are i and $-i$, and it is easy to verify that $\begin{bmatrix} i \\ -1 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ i \end{bmatrix}$ are corresponding eigenvectors. Moreover, these eigenvectors are orthogonal and both have length $\sqrt{2}$, so $U = \frac{1}{\sqrt{2}}\begin{bmatrix} i & -1 \\ -1 & i \end{bmatrix}$ is a unitary matrix such that $U^HAU = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}$ is diagonal.

There is a very simple way to characterize those complex matrices that are unitarily diagonalizable. To this end, an $n \times n$ complex matrix N is called **normal** if $NN^H = N^HN$. It is clear that every hermitian or unitary matrix is normal, as is the matrix $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ in Example 8.7.7. In fact we have the following result.

Theorem 8.7.9

An $n \times n$ complex matrix A is unitarily diagonalizable if and only if A is normal.

Proof. Assume first that $U^HAU = D$, where U is unitary and D is diagonal. Then $DD^H = D^HD$ as is easily verified. Because $DD^H = U^H(AA^H)U$ and $D^HD = U^H(A^HA)U$, it follows by cancellation that $AA^H = A^HA$.

Conversely, assume A is normal—that is, $AA^H = A^HA$. By Schur's theorem, let $U^HAU = T$, where T is upper triangular and U is unitary. Then T is normal too:

$$TT^H = U^H(AA^H)U = U^H(A^HA)U = T^HT$$

Hence it suffices to show that a normal $n \times n$ upper triangular matrix T must be diagonal. We induct on n; it is clear if $n = 1$. If $n > 1$ and $T = \left[t_{ij}\right]$, then equating $(1,\ 1)$-entries in TT^H and T^HT gives

$$|t_{11}|^2 + |t_{12}|^2 + \cdots + |t_{1n}|^2 = |t_{11}|^2$$

This implies $t_{12} = t_{13} = \cdots = t_{1n} = 0$, so $T = \begin{bmatrix} t_{11} & 0 \\ 0 & T_1 \end{bmatrix}$ in block form. Hence $T = \begin{bmatrix} \bar{t}_{11} & 0 \\ 0 & T_1^H \end{bmatrix}$ so $TT^H = T^HT$ implies $T_1T_1^H = T_1T_1^H$. Thus T_1 is diagonal by induction, and the proof is complete. □

We conclude this section by using Schur's theorem (Theorem 8.7.7) to prove a famous theorem about matrices. Recall that the characteristic polynomial of a square matrix A is defined by $c_A(x) = \det(xI - A)$, and that the eigenvalues of A are just the roots of $c_A(x)$.

Theorem 8.7.10: Cayley-Hamilton Theorem[17]

If A is an $n \times n$ complex matrix, then $c_A(A) = 0$; that is, A is a root of its characteristic polynomial.

Proof. If $p(x)$ is any polynomial with complex coefficients, then $p(P^{-1}AP) = P^{-1}p(A)P$ for any invertible complex matrix P. Hence, by Schur's theorem, we may assume that A is upper triangular. Then the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ of A appear along the main diagonal, so

$$c_A(x) = (x - \lambda_1)(x - \lambda_2)(x - \lambda_3) \cdots (x - \lambda_n)$$

Thus

$$c_A(A) = (A - \lambda_1 I)(A - \lambda_2 I)(A - \lambda_3 I) \cdots (A - \lambda_n I)$$

Note that each matrix $A - \lambda_i I$ is upper triangular. Now observe:

1. $A - \lambda_1 I$ has zero first column because column 1 of A is $(\lambda_1, 0, 0, \dots, 0)^T$.
2. Then $(A - \lambda_1 I)(A - \lambda_2 I)$ has the first two columns zero because the second column of $(A - \lambda_2 I)$ is $(b, 0, 0, \dots, 0)^T$ for some constant b.
3. Next $(A - \lambda_1 I)(A - \lambda_2 I)(A - \lambda_3 I)$ has the first three columns zero because column 3 of $(A - \lambda_3 I)$ is $(c, d, 0, \dots, 0)^T$ for some constants c and d.

Continuing in this way we see that $(A - \lambda_1 I)(A - \lambda_2 I)(A - \lambda_3 I) \cdots (A - \lambda_n I)$ has all n columns zero; that is, $c_A(A) = 0$. □

[17] Named after the English mathematician Arthur Cayley (1821–1895) and William Rowan Hamilton (1805–1865), an Irish mathematician famous for his work on physical dynamics.

8.8 An Application to Linear Codes over Finite Fields

For centuries mankind has been using codes to transmit messages. In many cases, for example transmitting financial, medical, or military information, the message is disguised in such a way that it cannot be understood by an intruder who intercepts it, but can be easily "decoded" by the intended receiver. This subject is called *cryptography* and, while intriguing, is not our focus here. Instead, we investigate methods for detecting and correcting errors in the transmission of the message.

The stunning photos of the planet Saturn sent by the space probe are a very good example of how successful these methods can be. These messages are subject to "noise" such as solar interference which causes errors in the message. The signal is received on Earth with errors that must be detected and corrected before the high-quality pictures can be printed. This is done using error-correcting codes. To see how, we first discuss a system of adding and multiplying integers while ignoring multiples of a fixed integer.

Modular Arithmetic

We work in the set $\mathbb{Z} = \{0,\ \pm 1,\ \pm 2,\ \pm 3,\ \dots\}$ of **integers**, that is the set of whole numbers. Everyone is familiar with the process of "long division" from arithmetic. For example, we can divide an integer a by 5 and leave a remainder "modulo 5" in the set $\{0,\ 1,\ 2,\ 3,\ 4\}$. As an illustration

$$19 = 3 \cdot 5 + 4$$

so the remainder of 19 modulo 5 is 4. Similarly, the remainder of 137 modulo 5 is 2 because we have $137 = 27 \cdot 5 + 2$. This works even for negative integers: For example,

$$-17 = (-4) \cdot 5 + 3$$

so the remainder of -17 modulo 5 is 3.

This process is called the **division algorithm**. More formally, let $n \geq 2$ denote an integer. Then every integer a can be written uniquely in the form

$$a = qn + r \quad \text{ where } q \text{ and } r \text{ are integers and } 0 \leq r \leq n-1$$

Here q is called the **quotient** of a **modulo** n, and r is called the **remainder** of a **modulo** n. We refer to n as the **modulus**. Thus, if $n = 6$, the fact that $134 = 22 \cdot 6 + 2$ means that 134 has quotient 22 and remainder 2 modulo 6.

Our interest here is in the set of *all* possible remainders modulo n. This set is denoted

$$\mathbb{Z}_n = \{0,\ 1,\ 2,\ 3,\ \dots,\ n-1\}$$

and is called the set of **integers modulo** n. Thus every integer is uniquely represented in $\mathbb{Z}_n$ by its remainder modulo n.

We are going to show how to do arithmetic in $\mathbb{Z}_n$ by adding and multiplying modulo n. That is, we add or multiply two numbers in $\mathbb{Z}_n$ by calculating the usual sum or product in $\mathbb{Z}$ and taking the remainder modulo n. It is proved in books on abstract algebra that the usual laws of arithmetic hold in $\mathbb{Z}_n$ for any modulus $n \geq 2$. This seems remarkable until we remember that these laws are true for ordinary addition and multiplication and all we are doing is reducing modulo n.

To illustrate, consider the case $n = 6$, so that $\mathbb{Z}_6 = \{0,\ 1,\ 2,\ 3,\ 4,\ 5\}$. Then $2 + 5 = 1$ in $\mathbb{Z}_6$ because 7 leaves a remainder of 1 when divided by 6. Similarly, $2 \cdot 5 = 4$ in $\mathbb{Z}_6$, while $3 + 5 = 2$, and $3 + 3 = 0$. In this way we can fill in the addition and multiplication tables for $\mathbb{Z}_6$; the result is:

Tables for $\mathbb{Z}_6$

$+$	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

$\times$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

Calculations in $\mathbb{Z}_6$ are carried out much as in $\mathbb{Z}$. As an illustration, consider the familiar "distributive law" $a(b+c) = ab+ac$ from ordinary arithmetic. This holds for all a, b, and c in $\mathbb{Z}_6$; we verify a particular case:

$$3(5+4) = 3\cdot 5+3\cdot 4 \quad \text{in } \mathbb{Z}_6$$

In fact, the left side is $3(5+4) = 3\cdot 3 = 3$, and the right side is $(3\cdot 5)+(3\cdot 4) = 3+0 = 3$ too. Hence doing arithmetic in $\mathbb{Z}_6$ is familiar. However, there are differences. For example, $3\cdot 4 = 0$ in $\mathbb{Z}_6$, in contrast to the fact that $a\cdot b = 0$ in $\mathbb{Z}$ can only happen when either $a = 0$ or $b = 0$. Similarly, $3^2 = 3$ in $\mathbb{Z}_6$, unlike $\mathbb{Z}$.

Note that we will make statements like $-30 = 19$ in $\mathbb{Z}_7$; it means that -30 and 19 leave the same remainder 5 when divided by 7, and so are equal in $\mathbb{Z}_7$ because they both equal 5. In general, if $n \geq 2$ is any modulus, the operative fact is that

$$a = b \text{ in } \mathbb{Z}_n \quad \text{if and only if} \quad a-b \text{ is a multiple of } n$$

In this case we say that a and b are **equal modulo** n, and write $a = b(\bmod n)$.

Arithmetic in $\mathbb{Z}_n$ is, in a sense, simpler than that for the integers. For example, consider negatives. Given the element 8 in $\mathbb{Z}_{17}$, what is -8? The answer lies in the observation that $8+9 = 0$ in $\mathbb{Z}_{17}$, so $-8 = 9$ (and $-9 = 8$). In the same way, finding negatives is not difficult in $\mathbb{Z}_n$ for any modulus n.

Finite Fields

In our study of linear algebra so far the scalars have been real (possibly complex) numbers. The set $\mathbb{R}$ of real numbers has the property that it is closed under addition and multiplication, that the usual laws of arithmetic hold, and that every nonzero real number has an inverse in $\mathbb{R}$. Such a system is called a **field**. Hence the real numbers $\mathbb{R}$ form a field, as does the set $\mathbb{C}$ of complex numbers. Another example is the set $\mathbb{Q}$ of all rational numbers (fractions); however the set $\mathbb{Z}$ of integers is *not* a field—for example, 2 has no inverse *in* the set $\mathbb{Z}$ because $2\cdot x = 1$ has no solution x in $\mathbb{Z}$.

Our motivation for isolating the concept of a field is that nearly everything we have done remains valid if the scalars are restricted to some field: The gaussian algorithm can be used to solve systems of linear equations with coefficients in the field; a square matrix with entries from the field is invertible if and only if its determinant is nonzero; the matrix inversion algorithm works in the same way; and so on. The reason is that the field has all the properties used in the proofs of these results for the field $\mathbb{R}$, so all the theorems remain valid.

It turns out that there are *finite* fields—that is, finite sets that satisfy the usual laws of arithmetic and in which every nonzero element a has an **inverse**, that is an element b in the field such that $ab = 1$. If $n \geq 2$ is an integer, the modular system $\mathbb{Z}_n$ certainly satisfies the basic laws of arithmetic, but it need not be a field. For example we have $2\cdot 3 = 0$ in $\mathbb{Z}_6$ so 3 has no inverse in $\mathbb{Z}_6$ (if $3a = 1$ then $2 = 2\cdot 1 = 2(3a) = 0a = 0$ in $\mathbb{Z}_6$, a contradiction). The problem is that $6 = 2\cdot 3$ can be properly factored in $\mathbb{Z}$.

An integer $p \geq 2$ is called a **prime** if p *cannot* be factored as $p = ab$ where a and b are positive integers and neither a nor b equals 1. Thus the first few primes are 2, 3, 5, 7, 11, 13, 17, ... If $n \geq 2$ is not a prime and $n = ab$ where $2 \leq a$, $b \leq n-1$, then $ab = 0$ in $\mathbb{Z}_n$ and it follows (as above in the case $n = 6$) that b cannot have an inverse in $\mathbb{Z}_n$, and hence that $\mathbb{Z}_n$ is not a field. In other words, if $\mathbb{Z}_n$ is a field, then n must be a prime. Surprisingly, the converse is true:

Theorem 8.8.1

If p is a prime, then $\mathbb{Z}_p$ is a field using addition and multiplication modulo p.

The proof can be found in books on abstract algebra.[18] If p is a prime, the field $\mathbb{Z}_p$ is called the **field of integers modulo** p.

For example, consider the case $n = 5$. Then $\mathbb{Z}_5 = \{0, 1, 2, 3, 4\}$ and the addition and multiplication tables are:

$$\begin{array}{c|ccccc} + & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 0 & 1 & 2 & 3 & 4 \\ 1 & 1 & 2 & 3 & 4 & 0 \\ 2 & 2 & 3 & 4 & 0 & 1 \\ 3 & 3 & 4 & 0 & 1 & 2 \\ 4 & 4 & 0 & 1 & 2 & 3 \end{array} \qquad \begin{array}{c|ccccc} \times & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 2 & 3 & 4 \\ 2 & 0 & 2 & 4 & 1 & 3 \\ 3 & 0 & 3 & 1 & 4 & 2 \\ 4 & 0 & 4 & 3 & 2 & 1 \end{array}$$

Hence 1 and 4 are self-inverse in $\mathbb{Z}_5$, and 2 and 3 are inverses of each other, so $\mathbb{Z}_5$ is indeed a field. Here is another important example.

Example 8.8.1

If $p = 2$, then $\mathbb{Z}_2 = \{0, 1\}$ is a field with addition and multiplication modulo 2 given by the tables

$$\begin{array}{c|cc} + & 0 & 1 \\ \hline 0 & 0 & 1 \\ 1 & 1 & 0 \end{array} \quad \text{and} \quad \begin{array}{c|cc} \times & 0 & 1 \\ \hline 0 & 0 & 0 \\ 1 & 0 & 1 \end{array}$$

This is binary arithmetic, the basic algebra of computers.

While it is routine to find negatives of elements of $\mathbb{Z}_p$, it is a bit more difficult to find inverses in $\mathbb{Z}_p$. For example, how does one find 14^{-1} in $\mathbb{Z}_{17}$? Since we want $14^{-1} \cdot 14 = 1$ in $\mathbb{Z}_{17}$, we are looking for an integer a with the property that $a \cdot 14 = 1$ modulo 17. Of course we can try all possibilities in $\mathbb{Z}_{17}$ (there are only 17 of them!), and the result is $a = 11$ (verify). However this method is of little use for large primes p, and it is a comfort to know that there is a systematic procedure (called the **euclidean algorithm**) for finding inverses in $\mathbb{Z}_p$ for any prime p. Furthermore, this algorithm is easy to program for a computer. To illustrate the method, let us once again find the inverse of 14 in $\mathbb{Z}_{17}$.

Example 8.8.2

Find the inverse of 14 in $\mathbb{Z}_{17}$.

Solution. The idea is to first divide $p = 17$ by 14:

$$17 = 1 \cdot 14 + 3$$

Now divide (the previous divisor) 14 by the new remainder 3 to get

$$14 = 4 \cdot 3 + 2$$

[18] See, for example, W. Keith Nicholson, *Introduction to Abstract Algebra*, 4th ed., (New York: Wiley, 2012).

and then divide (the previous divisor) 3 by the new remainder 2 to get

$$3 = 1 \cdot 2 + 1$$

It is a theorem of number theory that, because 17 is a prime, this procedure will *always* lead to a remainder of 1. At this point we eliminate remainders in these equations from the bottom up:

$$\begin{aligned} 1 &= 3 - 1 \cdot 2 && \text{since } 3 = 1 \cdot 2 + 1 \\ &= 3 - 1 \cdot (14 - 4 \cdot 3) = 5 \cdot 3 - 1 \cdot 14 && \text{since } 2 = 14 - 4 \cdot 3 \\ &= 5 \cdot (17 - 1 \cdot 14) - 1 \cdot 14 = 5 \cdot 17 - 6 \cdot 14 && \text{since } 3 = 17 - 1 \cdot 14 \end{aligned}$$

Hence $(-6) \cdot 14 = 1$ in $\mathbb{Z}_{17}$, that is, $11 \cdot 14 = 1$. So $14^{-1} = 11$ in $\mathbb{Z}_{17}$.

As mentioned above, nearly everything we have done with matrices over the field of real numbers can be done in the same way for matrices with entries from $\mathbb{Z}_p$. We illustrate this with one example. Again the reader is referred to books on abstract algebra.

Example 8.8.3

Determine if the matrix $A = \begin{bmatrix} 1 & 4 \\ 6 & 5 \end{bmatrix}$ from $\mathbb{Z}_7$ is invertible and, if so, find its inverse.

Solution. Working in $\mathbb{Z}_7$ we have $\det A = 1 \cdot 5 - 6 \cdot 4 = 5 - 3 = 2 \neq 0$ in $\mathbb{Z}_7$, so A is invertible. Hence Example 2.4.4 gives $A^{-1} = 2^{-1} \begin{bmatrix} 5 & -4 \\ -6 & 1 \end{bmatrix}$. Note that $2^{-1} = 4$ in $\mathbb{Z}_7$ (because $2 \cdot 4 = 1$ in $\mathbb{Z}_7$). Note also that $-4 = 3$ and $-6 = 1$ in $\mathbb{Z}_7$, so finally $A^{-1} = 4 \begin{bmatrix} 5 & 3 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 5 \\ 4 & 4 \end{bmatrix}$. The reader can verify that indeed $\begin{bmatrix} 1 & 4 \\ 6 & 5 \end{bmatrix} \begin{bmatrix} 6 & 5 \\ 4 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ in $\mathbb{Z}_7$.

While we shall not use them, there are finite fields other than $\mathbb{Z}_p$ for the various primes p. Surprisingly, for every prime p and every integer $n \geq 1$, there *exists* a field with exactly p^n elements, and this field is *unique*.[19] It is called the **Galois field** of order p^n, and is denoted $GF(p^n)$.

[19]See, for example, W. K. Nicholson, *Introduction to Abstract Algebra*, 4th ed., (New York: Wiley, 2012).

Error Correcting Codes

Coding theory is concerned with the transmission of information over a *channel* that is affected by *noise*. The noise causes errors, so the aim of the theory is to find ways to detect such errors and correct at least some of them. General coding theory originated with the work of Claude Shannon (1916–2001) who showed that information can be transmitted at near optimal rates with arbitrarily small chance of error.

Let F denote a finite field and, if $n \geq 1$, let

$$F^n \text{ denote the } F\text{-vector space of } 1 \times n \text{ row matrices over } F$$

with the usual componentwise addition and scalar multiplication. In this context, the rows in F^n are called **words** (or n-**words**) and, as the name implies, will be written as $[a\ b\ c\ d] = abcd$. The individual components of a word are called its **digits**. A nonempty subset C of F^n is called a **code** (or an n-**code**), and the elements in C are called **code words**. If $F = \mathbb{Z}_2$, these are called **binary** codes.

If a code word $\mathbf{w}$ is transmitted and an error occurs, the resulting word $\mathbf{v}$ is decoded as the code word "closest" to $\mathbf{v}$ in F^n. To make sense of what "closest" means, we need a distance function on F^n analogous to that in $\mathbb{R}^n$ (see Theorem 5.3.3). The usual definition in $\mathbb{R}^n$ does not work in this situation. For example, if $\mathbf{w} = 1111$ in $(\mathbb{Z}_2)^4$ then the square of the distance of $\mathbf{w}$ from $\mathbf{0}$ is

$$(1-0)^2 + (1-0)^2 + (1-0)^2 + (1-0)^2 = 0$$

even though $\mathbf{w} \neq \mathbf{0}$.

However there is a satisfactory notion of distance in F^n due to Richard Hamming (1915–1998). Given a word $\mathbf{w} = a_1 a_2 \cdots a_n$ in F^n, we first define the **Hamming weight** $wt(\mathbf{w})$ to be the number of nonzero digits in $\mathbf{w}$:

$$wt(\mathbf{w}) = wt(a_1 a_2 \cdots a_n) = |\{i \mid a_i \neq 0\}|$$

Clearly, $0 \leq wt(\mathbf{w}) \leq n$ for every word $\mathbf{w}$ in F^n. Given another word $\mathbf{v} = b_1 b_2 \cdots b_n$ in F^n, the **Hamming distance** $d(\mathbf{v},\ \mathbf{w})$ between $\mathbf{v}$ and $\mathbf{w}$ is defined by

$$d(\mathbf{v},\ \mathbf{w}) = wt(\mathbf{v} - \mathbf{w}) = |\{i \mid b_i \neq a_i\}|$$

In other words, $d(\mathbf{v},\ \mathbf{w})$ is the number of places at which the digits of $\mathbf{v}$ and $\mathbf{w}$ differ. The next result justifies using the term *distance* for this function d.

Theorem 8.8.2

Let $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ denote words in F^n. Then:

1. $d(\mathbf{v},\ \mathbf{w}) \geq 0$.
2. $d(\mathbf{v},\ \mathbf{w}) = 0$ *if and only if* $\mathbf{v} = \mathbf{w}$.
3. $d(\mathbf{v},\ \mathbf{w}) = d(\mathbf{w},\ \mathbf{v})$.
4. $d(\mathbf{v},\ \mathbf{w}) \leq d(\mathbf{v},\ \mathbf{u}) + d(\mathbf{u},\ \mathbf{w})$

Proof. (1) and (3) are clear, and (2) follows because $wt(\mathbf{v}) = 0$ if and only if $\mathbf{v} = \mathbf{0}$. To prove (4), write $\mathbf{x} = \mathbf{v} - \mathbf{u}$ and $\mathbf{y} = \mathbf{u} - \mathbf{w}$. Then (4) reads $wt(\mathbf{x} + \mathbf{y}) \leq wt(\mathbf{x}) + wt(\mathbf{y})$. If $\mathbf{x} = a_1 a_2 \cdots a_n$ and $\mathbf{y} = b_1 b_2 \cdots b_n$, this follows because $a_i + b_i \neq 0$ implies that either $a_i \neq 0$ or $b_i \neq 0$. □

Given a word $\mathbf{w}$ in F^n and a real number $r > 0$, define the **ball** $B_r(\mathbf{w})$ of radius r (or simply the r-**ball**) about $\mathbf{w}$ as follows:

$$B_r(\mathbf{w}) = \{\mathbf{x} \in F^n \mid d(\mathbf{w},\ \mathbf{x}) \leq r\}$$

Using this we can describe one of the most useful decoding methods.

Theorem: Nearest Neighbour Decoding

Let C be an n-code, and suppose a word $\mathbf{v}$ is transmitted and $\mathbf{w}$ is received. Then $\mathbf{w}$ is decoded as the code word in C closest to it. (If there is a tie, choose arbitrarily.)

Using this method, we can describe how to construct a code C that can detect (or correct) t errors. Suppose a code word $\mathbf{c}$ is transmitted and a word $\mathbf{w}$ is received with s errors where $1 \leq s \leq t$. Then s is the number of places at which the $\mathbf{c}$- and $\mathbf{w}$-digits differ, that is, $s = d(\mathbf{c},\ \mathbf{w})$. Hence $B_t(\mathbf{c})$ consists of all possible received words where at most t errors have occurred.

Assume first that C has the property that no code word lies in the t-ball of another code word. Because $\mathbf{w}$ is in $B_t(\mathbf{c})$ and $\mathbf{w} \neq \mathbf{c}$, this means that $\mathbf{w}$ is not a code word and the error has been detected. If we strengthen the assumption on C to require that the t-balls about code words are pairwise disjoint, then $\mathbf{w}$ belongs to a unique ball (the one about $\mathbf{c}$), and so $\mathbf{w}$ will be correctly decoded as $\mathbf{c}$.

To describe when this happens, let C be an n-code. The **minimum distance** d of C is defined to be the smallest distance between two distinct code words in C; that is,

$$d = \min\{d(\mathbf{v},\ \mathbf{w}) \mid \mathbf{v} \text{ and } \mathbf{w} \text{ in } C; \mathbf{v} \neq \mathbf{w}\}$$

Theorem 8.8.3

Let C be an n-code with minimum distance d. Assume that nearest neighbour decoding is used. Then:

1. *If $t < d$, then C can detect t errors.*[20]
2. *If $2t < d$, then C can correct t errors.*

Proof.

1. Let $\mathbf{c}$ be a code word in C. If $\mathbf{w} \in B_t(\mathbf{c})$, then $d(\mathbf{w},\ \mathbf{c}) \leq t < d$ by hypothesis. Thus the t-ball $B_t(\mathbf{c})$ contains no other code word, so C can detect t errors by the preceding discussion.

2. If $2t < d$, it suffices (again by the preceding discussion) to show that the t-balls about distinct code words are pairwise disjoint. But if $\mathbf{c} \neq \mathbf{c}'$ are code words in C and $\mathbf{w}$ is in $B_t(\mathbf{c}') \cap B_t(\mathbf{c})$, then Theorem 8.8.2 gives
$$d(\mathbf{c},\ \mathbf{c}') \leq d(\mathbf{c},\ \mathbf{w}) + d(\mathbf{w},\ \mathbf{c}') \leq t + t = 2t < d$$
by hypothesis, contradicting the minimality of d. □

[20] We say that C detects (corrects) t errors if C can detect (or correct) t or *fewer* errors.

Example 8.8.4

If $F = \mathbb{Z}_3 = \{0, 1, 2\}$, the 6-code $\{111111, 111222, 222111\}$ has minimum distance 3 and so can detect 2 errors and correct 1 error.

Let $\mathbf{c}$ be any word in F^n. A word $\mathbf{w}$ satisfies $d(\mathbf{w}, \mathbf{c}) = r$ if and only if $\mathbf{w}$ and $\mathbf{c}$ differ in exactly r digits. If $|F| = q$, there are exactly $\binom{n}{r}(q-1)^r$ such words where $\binom{n}{r}$ is the binomial coefficient. Indeed, choose the r places where they differ in $\binom{n}{r}$ ways, and then fill those places in $\mathbf{w}$ in $(q-1)^r$ ways. It follows that the number of words in the t-ball about $\mathbf{c}$ is

$$|B_t(\mathbf{c})| = \binom{n}{0} + \binom{n}{1}(q-1) + \binom{n}{2}(q-1)^2 + \cdots + \binom{n}{t}(q-1)^t = \sum_{i=0}^{t} \binom{n}{i}(q-1)^i$$

This leads to a useful bound on the size of error-correcting codes.

Theorem 8.8.4: Hamming Bound

Let C be an n-code over a field F that can correct t errors using nearest neighbour decoding. If $|F| = q$, then

$$|C| \leq \frac{q^n}{\sum_{i=0}^{t} \binom{n}{i}(q-1)^i}$$

Proof. Write $k = \sum_{i=0}^{t} \binom{n}{i}(q-1)^i$. The t-balls centred at distinct code words each contain k words, and there are $|C|$ of them. Moreover they are pairwise disjoint because the code corrects t errors (see the discussion preceding Theorem 8.8.3). Hence they contain $k \cdot |C|$ distinct words, and so $k \cdot |C| \leq |F^n| = q^n$, proving the theorem. □

A code is called **perfect** if there is equality in the Hamming bound; equivalently, if every word in F^n lies in exactly one t-ball about a code word. For example, if $F = \mathbb{Z}_2$, $n = 3$, and $t = 1$, then $q = 2$ and $\binom{3}{0} + \binom{3}{1} = 4$, so the Hamming bound is $\frac{2^3}{4} = 2$. The 3-code $C = \{000, 111\}$ has minimum distance 3 and so can correct 1 error by Theorem 8.8.3. Hence C is perfect.

Linear Codes

Up to this point we have been regarding *any* nonempty subset of the F-vector space F^n as a code. However many important codes are actually subspaces. A subspace $C \subseteq F^n$ of dimension $k \geq 1$ over F is called an (n, k)**-linear code**, or simply an (n, k)**-code**. We do not regard the zero subspace (that is, $k = 0$) as a code.

Example 8.8.5

If $F = \mathbb{Z}_2$ and $n \geq 2$, the n-**parity-check code** is constructed as follows: An extra digit is added to each word in F^{n-1} to make the number of 1s in the resulting word even (we say such words have **even parity**). The resulting $(n, n-1)$-code is linear because the sum of two words of even parity again has even parity.

Many of the properties of general codes take a simpler form for linear codes. The following result gives a much easier way to find the minimal distance of a linear code, and sharpens the results in Theorem 8.8.3.

Theorem 8.8.5

Let C be an $(n,\ k)$-code with minimum distance d over a finite field F, and use nearest neighbour decoding.

1. $d = \min\{wt(\mathbf{w}) \mid \mathbf{0} \neq \mathbf{w} \in C\}$.
2. *C can detect $t \geq 1$ errors if and only if $t < d$.*
3. *C can correct $t \geq 1$ errors if and only if $2t < d$.*
4. *If C can correct $t \geq 1$ errors and $|F| = q$, then*

$$\binom{n}{0} + \binom{n}{1}(q-1) + \binom{n}{2}(q-1)^2 + \cdots + \binom{n}{t}(q-1)^t \leq q^{n-k}$$

Proof.

1. Write $d' = \min\{wt(\mathbf{w}) \mid \mathbf{0} \neq \mathbf{w} \text{ in } C\}$. If $\mathbf{v} \neq \mathbf{w}$ are words in C, then $d(\mathbf{v},\ \mathbf{w}) = wt(\mathbf{v} - \mathbf{w}) \geq d'$ because $\mathbf{v} - \mathbf{w}$ is in the subspace C. Hence $d \geq d'$. Conversely, given $\mathbf{w} \neq \mathbf{0}$ in C then, since $\mathbf{0}$ is in C, we have $wt(\mathbf{w}) = d(\mathbf{w},\ \mathbf{0}) \geq d$ by the definition of d. Hence $d' \geq d$ and (1) is proved.

2. Assume that C can detect t errors. Given $\mathbf{w} \neq \mathbf{0}$ in C, the t-ball $B_t(\mathbf{w})$ about $\mathbf{w}$ contains no other code word (see the discussion preceding Theorem 8.8.3). In particular, it does not contain the code word $\mathbf{0}$, so $t < d(\mathbf{w},\ \mathbf{0}) = wt(\mathbf{w})$. Hence $t < d$ by (1). The converse is part of Theorem 8.8.3.

3. We require a result of interest in itself.

   Claim. Suppose $\mathbf{c}$ in C has $wt(\mathbf{c}) \leq 2t$. Then $B_t(\mathbf{0}) \cap B_t(\mathbf{c})$ is nonempty.

   Proof. If $wt(\mathbf{c}) \leq t$, then $\mathbf{c}$ itself is in $B_t(\mathbf{0}) \cap B_t(\mathbf{c})$. So assume $t < wt(\mathbf{c}) \leq 2t$. Then $\mathbf{c}$ has more than t nonzero digits, so we can form a new word $\mathbf{w}$ by changing exactly t of these nonzero digits to zero. Then $d(\mathbf{w},\ \mathbf{c}) = t$, so $\mathbf{w}$ is in $B_t(\mathbf{c})$. But $wt(\mathbf{w}) = wt(\mathbf{c}) - t \leq t$, so $\mathbf{w}$ is also in $B_t(\mathbf{0})$. Hence $\mathbf{w}$ is in $B_t(\mathbf{0}) \cap B_t(\mathbf{c})$, proving the Claim.

   If C corrects t errors, the t-balls about code words are pairwise disjoint (see the discussion preceding Theorem 8.8.3). Hence the claim shows that $wt(\mathbf{c}) > 2t$ for all $\mathbf{c} \neq \mathbf{0}$ in C, from which $d > 2t$ by (1). The other inequality comes from Theorem 8.8.3.

4. We have $|C| = q^k$ because $\dim_F C = k$, so this assertion restates Theorem 8.8.4. □

Example 8.8.6

If $F = \mathbb{Z}_2$, then

$$C = \{0000000,\ 0101010,\ 1010101,\ 1110000,\ 1011010,\ 0100101,\ 0001111,\ 1111111\}$$

is a $(7,\ 3)$-code; in fact $C = \text{span}\{0101010,\ 1010101,\ 1110000\}$. The minimum distance for C is 3, the minimum weight of a nonzero word in C.

Matrix Generators

Given a linear n-code C over a finite field F, the way encoding works in practice is as follows. A message stream is blocked off into segments of length $k \leq n$ called **messages**. Each message $\mathbf{u}$ in F^k is encoded as a code word, the code word is transmitted, the receiver decodes the received word as the nearest code word, and then re-creates the original message. A fast and convenient method is needed to encode the incoming messages, to decode the received word after transmission (with or without error), and finally to retrieve messages from code words. All this can be achieved for any linear code using matrix multiplication.

Let G denote a $k \times n$ matrix over a finite field F, and encode each message $\mathbf{u}$ in F^k as the word $\mathbf{u}G$ in F^n using matrix multiplication (thinking of words as rows). This amounts to saying that the set of code words is the subspace $C = \{\mathbf{u}G \mid \mathbf{u} \text{ in } F^k\}$ of F^n. This subspace need not have dimension k for every $k \times n$ matrix G. But, if $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k\}$ is the standard basis of F^k, then $\mathbf{e}_iG$ is row i of G for each I and $\{\mathbf{e}_1G, \mathbf{e}_2G, \dots, \mathbf{e}_kG\}$ spans C. Hence $\dim C = k$ if and only if the rows of G are independent in F^n, and these matrices turn out to be exactly the ones we need. For reference, we state their main properties in Lemma 8.8.1 below (see Theorem 5.4.4).

Lemma 8.8.1

The following are equivalent for a $k \times n$ matrix G over a finite field F:

1. $\text{rank } G = k$.
2. *The columns of G span F^k.*
3. *The rows of G are independent in F^n.*
4. *The system $GX = B$ is consistent for every column B in $\mathbb{R}^k$.*
5. *$GK = I_k$ for some $n \times k$ matrix K.*

Proof. (1) $\Rightarrow$ (2). This is because $\dim(\text{col } G) = k$ by (1).
(2) $\Rightarrow$ (4). $G\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^T = x_1\mathbf{c}_1 + \cdots + x_n\mathbf{c}_n$ where $\mathbf{c}_j$ is column j of G.
(4) $\Rightarrow$ (5). $G\begin{bmatrix} \mathbf{k}_1 & \cdots & \mathbf{k}_k \end{bmatrix} = \begin{bmatrix} G\mathbf{k}_1 & \cdots & G\mathbf{k}_k \end{bmatrix}$ for columns $\mathbf{k}_j$.
(5) $\Rightarrow$ (3). If $a_1R_1 + \cdots + a_kR_k = 0$ where R_i is row i of G, then $\begin{bmatrix} a_1 & \cdots & a_k \end{bmatrix} G = 0$, so by (5), $\begin{bmatrix} a_1 & \cdots & a_k \end{bmatrix} = 0$. Hence each $a_i = 0$, proving (3).
(3) $\Rightarrow$ (1). $\text{rank } G = \dim(\text{row } G) = k$ by (3). □

Note that Theorem 5.4.4 asserts that, over the real field $\mathbb{R}$, the properties in Lemma 8.8.1 hold if and only if GG^T is invertible. But this need not be true in general. For example, if $F = \mathbb{Z}_2$ and $G = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$, then $GG^T = 0$. The reason is that the dot product $\mathbf{w} \cdot \mathbf{w}$ can be zero for $\mathbf{w}$ in F^n even if $\mathbf{w} \neq \mathbf{0}$. However, even though GG^T is not invertible, we do have $GK = I_2$ for some 4×2 matrix K over F as Lemma 8.8.1 asserts (in fact, $K = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}^T$ is one such matrix).

Let $C \subseteq F^n$ be an (n, k)-code over a finite field F. If $\{\mathbf{w}_1, \dots, \mathbf{w}_k\}$ is a basis of C, let $G = \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_k \end{bmatrix}$

be the $k \times n$ matrix with the $\mathbf{w}_i$ as its rows. Let $\{\mathbf{e}_1, \dots, \mathbf{e}_k\}$ is the standard basis of F^k regarded as rows. Then $\mathbf{w}_i = \mathbf{e}_i G$ for each i, so $C = \text{span}\{\mathbf{w}_1, \dots, \mathbf{w}_k\} = \text{span}\{\mathbf{e}_1 G, \dots, \mathbf{e}_k G\}$. It follows (verify) that

$$C = \{\mathbf{u}G \mid \mathbf{u} \text{ in } F^k\}$$

Because of this, the $k \times n$ matrix G is called a **generator** of the code C, and G has rank k by Lemma 8.8.1 because its rows $\mathbf{w}_i$ are independent.

In fact, every linear code C in F^n has a generator of a simple, convenient form. If G is a generator matrix for C, let R be the reduced row-echelon form of G. We claim that C is also generated by R. Since $G \to R$ by row operations, Theorem 2.5.1 shows that these same row operations $\left[\begin{array}{cc} G & I_k \end{array}\right] \to \left[\begin{array}{cc} R & W \end{array}\right]$, performed on $\left[\begin{array}{cc} G & I_k \end{array}\right]$, produce an invertible $k \times k$ matrix W such that $R = WG$. Then $C = \{\mathbf{u}R \mid \mathbf{u} \text{ in } F^k\}$. [In fact, if $\mathbf{u}$ is in F^k, then $\mathbf{u}G = \mathbf{u}_1 R$ where $\mathbf{u}_1 = \mathbf{u}W^{-1}$ is in F^k, and $\mathbf{u}R = \mathbf{u}_2 G$ where $\mathbf{u}_2 = \mathbf{u}W$ is in F^k]. Thus R is a generator of C, so we may assume that G is in reduced row-echelon form.

In that case, G has no row of zeros (since $\text{rank } G = k$) and so contains all the columns of I_k. Hence a series of *column* interchanges will carry G to the block form $G'' = \left[\begin{array}{cc} I_k & A \end{array}\right]$ for some $k \times (n-k)$ matrix A. Hence the code $C'' = \{\mathbf{u}G'' \mid \mathbf{u} \text{ in } F^k\}$ is essentially the same as C; the code words in C'' are obtained from those in C by a series of column interchanges. Hence if C is a linear (n, k)-code, we may (and shall) assume that the generator matrix G has the form

$$G = \left[\begin{array}{cc} I_k & A \end{array}\right] \quad \text{for some} \quad k \times (n-k) \text{ matrix } A$$

Such a matrix is called a **standard generator**, or a **systematic generator**, for the code C. In this case, if $\mathbf{u}$ is a message word in F^k, the first k digits of the encoded word $\mathbf{u}G$ are just the first k digits of $\mathbf{u}$, so retrieval of $\mathbf{u}$ from $\mathbf{u}G$ is very simple indeed. The last $n-k$ digits of $\mathbf{u}G$ are called **parity digits**.

Parity-Check Matrices

We begin with an important theorem about matrices over a finite field.

Theorem 8.8.6

Let F be a finite field, let G be a $k \times n$ matrix of rank k, let H be an $(n-k) \times n$ matrix of rank $n-k$, and let $C = \{\mathbf{u}G \mid \mathbf{u} \text{ in } F^k\}$ and $D = \{\mathbf{v}H \mid \mathbf{V} \text{ in } F^{n-k}\}$ be the codes they generate. Then the following conditions are equivalent:

1. $GH^T = 0$.
2. $HG^T = 0$.
3. $C = \{\mathbf{w} \text{ in } F^n \mid \mathbf{w}H^T = \mathbf{0}\}$.
4. $D = \{\mathbf{w} \text{ in } F^n \mid \mathbf{w}G^T = \mathbf{0}\}$.

Proof. First, (1) $\Leftrightarrow$ (2) holds because HG^T and GH^T are transposes of each other.

(1) $\Rightarrow$ (3) Consider the linear transformation $T : F^n \to F^{n-k}$ defined by $T(\mathbf{w}) = \mathbf{w}H^T$ for all $\mathbf{w}$ in F^n. To prove (3) we must show that $C = \ker T$. We have $C \subseteq \ker T$ by (1) because $T(\mathbf{u}G) = \mathbf{u}GH^T = \mathbf{0}$ for all $\mathbf{u}$ in F^k. Since $\dim C = \text{rank } G = k$, it is enough (by Theorem 6.4.2) to show $\dim(\ker T) = k$. However

the dimension theorem (Theorem 7.2.4) shows that $\dim(\ker T) = n - \dim(\operatorname{im} T)$, so it is enough to show that $\dim(\operatorname{im} T) = n - k$. But if $R_1, \dots, R_n$ are the rows of H^T, then block multiplication gives

$$\operatorname{im} T = \{\mathbf{w}H^T \mid \mathbf{w} \text{ in } \mathbb{R}^n\} = \operatorname{span}\{R_1, \dots, R_n\} = \operatorname{row}(H^T)$$

Hence $\dim(\operatorname{im} T) = \operatorname{rank}(H^T) = \operatorname{rank} H = n - k$, as required. This proves (3).

(3) $\Rightarrow$ (1) If $\mathbf{u}$ is in F^k, then $\mathbf{u}G$ is in C so, by (3), $\mathbf{u}(GH^T) = (\mathbf{u}G)H^T = \mathbf{0}$. Since $\mathbf{u}$ is arbitrary in F^k, it follows that $GH^T = 0$.

(2) $\Leftrightarrow$ (4) The proof is analogous to (1) $\Leftrightarrow$ (3). □

The relationship between the codes C and D in Theorem 8.8.6 will be characterized in another way in the next subsection.

If C is an (n, k)-code, an $(n-k) \times n$ matrix H is called a **parity-check matrix** for C if $C = \{\mathbf{w} \mid \mathbf{w}H^T = \mathbf{0}\}$ as in Theorem 8.8.6. Such matrices are easy to find for a given code C. If $G = \left[\begin{array}{cc} I_k & A \end{array}\right]$ is a standard generator for C where A is $k \times (n-k)$, the $(n-k) \times n$ matrix

$$H = \left[\begin{array}{cc} -A^T & I_{n-k} \end{array}\right]$$

is a parity-check matrix for C. Indeed, $\operatorname{rank} H = n - k$ because the rows of H are independent (due to the presence of I_{n-k}), and

$$GH^T = \left[\begin{array}{cc} I_k & A \end{array}\right] \left[\begin{array}{c} -A \\ I_{n-k} \end{array}\right] = -A + A = 0$$

by block multiplication. Hence H is a parity-check matrix for C and we have $C = \{\mathbf{w} \text{ in } F^n \mid \mathbf{w}H^T = \mathbf{0}\}$. Since $\mathbf{w}H^T$ and $H\mathbf{w}^T$ are transposes of each other, this shows that C can be characterized as follows:

$$C = \{\mathbf{w} \text{ in } F^n \mid H\mathbf{w}^T = \mathbf{0}\}$$

by Theorem 8.8.6.

This is useful in decoding. The reason is that decoding is done as follows: If a code word $\mathbf{c}$ is transmitted and $\mathbf{v}$ is received, then $\mathbf{z} = \mathbf{v} - \mathbf{c}$ is called the **error**. Since $H\mathbf{c}^T = \mathbf{0}$, we have $H\mathbf{z}^T = H\mathbf{v}^T$ and this word

$$\mathbf{s} = H\mathbf{z}^T = H\mathbf{v}^T$$

is called the **syndrome**. The receiver knows $\mathbf{v}$ and $\mathbf{s} = H\mathbf{v}^T$, and wants to recover $\mathbf{c}$. Since $\mathbf{c} = \mathbf{v} - \mathbf{z}$, it is enough to find $\mathbf{z}$. But the possibilities for $\mathbf{z}$ are the solutions of the linear system

$$H\mathbf{z}^T = \mathbf{s}$$

where $\mathbf{s}$ is known. Now recall that Theorem 2.2.3 shows that these solutions have the form $\mathbf{z} = \mathbf{x} + \mathbf{s}$ where $\mathbf{x}$ is any solution of the homogeneous system $H\mathbf{x}^T = \mathbf{0}$, that is, $\mathbf{x}$ is any word in C (by Lemma 8.8.1). In other words, the errors $\mathbf{z}$ are the elements of the set

$$C + \mathbf{s} = \{\mathbf{c} + \mathbf{s} \mid \mathbf{c} \text{ in } C\}$$

The set $C + \mathbf{s}$ is called a **coset** of C. Let $|F| = q$. Since $|C + \mathbf{s}| = |C| = q^{n-k}$ the search for $\mathbf{z}$ is reduced from q^n possibilities in F^n to q^{n-k} possibilities in $C + \mathbf{s}$. This is called **syndrome decoding**, and various methods for improving efficiency and accuracy have been devised. The reader is referred to books on coding for more details.[21]

[21]For an elementary introduction, see V. Pless, *Introduction to the Theory of Error-Correcting Codes*, 3rd ed., (New York: Wiley, 1998).

Orthogonal Codes

Let F be a finite field. Given two words $\mathbf{v} = a_1a_2\cdots a_n$ and $\mathbf{w} = b_1b_2\cdots b_n$ in F^n, the dot product $\mathbf{v}\cdot\mathbf{w}$ is defined (as in $\mathbb{R}^n$) by

$$\mathbf{v}\cdot\mathbf{w} = a_1b_1 + a_2b_2 + \cdots + a_nb_n$$

Note that $\mathbf{v}\cdot\mathbf{w}$ is an element of F, and it can be computed as a matrix product: $\mathbf{v}\cdot\mathbf{w} = \mathbf{v}\mathbf{w}^T$.

If $C \subseteq F^n$ is an $(n,\ k)$-code, the **orthogonal complement** $C^\perp$ is defined as in $\mathbb{R}^n$:

$$C^\perp = \{\mathbf{v} \text{ in } F^n \mid \mathbf{v}\cdot\mathbf{c} = 0 \text{ for all } \mathbf{c} \text{ in } C\}$$

This is easily seen to be a subspace of F^n, and it turns out to be an $(n,\ n-k)$-code. This follows when $F = \mathbb{R}$ because we showed (in the projection theorem) that $n = \dim U^\perp + \dim U$ for any subspace U of $\mathbb{R}^n$. However the proofs break down for a finite field F because the dot product in F^n has the property that $\mathbf{w}\cdot\mathbf{w} = 0$ can happen even if $\mathbf{w} \neq \mathbf{0}$. Nonetheless, the result remains valid.

Theorem 8.8.7

Let C be an $(n,\ k)$-code over a finite field F, let $G = \begin{bmatrix} I_k & A \end{bmatrix}$ be a standard generator for C where A is $k\times(n-k)$, and write $H = \begin{bmatrix} -A^T & I_{n-k} \end{bmatrix}$ for the parity-check matrix. Then:

1. *H is a generator of $C^\perp$.*
2. $\dim(C^\perp) = n-k = \operatorname{rank} H$.
3. *$C^{\perp\perp} = C$ and $\dim(C^\perp) + \dim C = n$.*

Proof. As in Theorem 8.8.6, let $D = \{\mathbf{v}H \mid \mathbf{v} \text{ in } F^{n-k}\}$ denote the code generated by H. Observe first that, for all $\mathbf{w}$ in F^n and all $\mathbf{u}$ in F^k, we have

$$\mathbf{w}\cdot(\mathbf{u}G) = \mathbf{w}(\mathbf{u}G)^T = \mathbf{w}(G^T\mathbf{u}^T) = (\mathbf{w}G^T)\cdot\mathbf{u}$$

Since $C = \{\mathbf{u}G \mid \mathbf{u} \text{ in } F^k\}$, this shows that $\mathbf{w}$ is in $C^\perp$ if and only if $(\mathbf{w}G^T)\cdot\mathbf{u} = 0$ for all $\mathbf{u}$ in F^k; if and only if[22] $\mathbf{w}G^T = \mathbf{0}$; if and only if $\mathbf{w}$ is in D (by Theorem 8.8.6). Thus $C^\perp = D$ and a similar argument shows that $D^\perp = C$.

1. H generates $C^\perp$ because $C^\perp = D = \{\mathbf{v}H \mid \mathbf{v} \text{ in } F^{n-k}\}$.
2. This follows from (1) because, as we observed above, $\operatorname{rank} H = n-k$.
3. Since $C^\perp = D$ and $D^\perp = C$, we have $C^{\perp\perp} = (C^\perp)^\perp = D^\perp = C$. Finally the second equation in (3) restates (2) because $\dim C = k$.

□

We note in passing that, if C is a subspace of $\mathbb{R}^k$, we have $C + C^\perp = \mathbb{R}^k$ by the projection theorem (Theorem 8.1.3), and $C \cap C^\perp = \{\mathbf{0}\}$ because any vector $\mathbf{x}$ in $C\cap C^\perp$ satisfies $\|\mathbf{x}\|^2 = \mathbf{x}\cdot\mathbf{x} = 0$. However, this fails in general. For example, if $F = \mathbb{Z}_2$ and $C = \operatorname{span}\{1010,\ 0101\}$ in F^4 then $C^\perp = C$, so $C + C^\perp = C = C\cap C^\perp$.

[22] If $\mathbf{v}\cdot\mathbf{u} = 0$ for every $\mathbf{u}$ in F^k, then $\mathbf{v} = \mathbf{0}$—let $\mathbf{u}$ range over the standard basis of F^k.

We conclude with one more example. If $F = \mathbb{Z}_2$, consider the standard matrix G below, and the corresponding parity-check matrix H:

$$G = \left[\begin{array}{ccccccc} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{array}\right] \quad \text{and} \quad H = \left[\begin{array}{ccccccc} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right]$$

The code $C = \{\mathbf{u}G \mid \mathbf{u} \text{ in } F^4\}$ generated by G has dimension $k = 4$, and is called the **Hamming (7, 4)-code**. The vectors in C are listed in the first table below. The dual code generated by H has dimension $n - k = 3$ and is listed in the second table.

C:

$\mathbf{u}$	$\mathbf{u}G$
0000	0000000
0001	0001011
0010	0010101
0011	0011110
0100	0100110
0101	0101101
0110	0110011
0111	0111000
1000	1000111
1001	1001100
1010	1010010
1011	1011001
1100	1100001
1101	1101010
1110	1110100
1111	1111111

$C^{\perp}$:

$\mathbf{v}$	$\mathbf{v}H$
000	0000000
001	1011001
010	1101010
011	0110011
100	1110100
101	0101101
110	0011110
111	1000111

Clearly each nonzero code word in C has weight at least 3, so C has minimum distance $d = 3$. Hence C can detect two errors and correct one error by Theorem 8.8.5. The dual code has minimum distance 4 and so can detect 3 errors and correct 1 error.

8.9 An Application to Quadratic Forms

An expression like $x_1^2 + x_2^2 + x_3^2 - 2x_1x_3 + x_2x_3$ is called a quadratic form in the variables x_1, x_2, and x_3. In this section we show that new variables y_1, y_2, and y_3 can always be found so that the quadratic form, when expressed in terms of the new variables, has no cross terms y_1y_2, y_1y_3, or y_2y_3. Moreover, we do this for forms involving any finite number of variables using orthogonal diagonalization. This has far-reaching applications; quadratic forms arise in such diverse areas as statistics, physics, the theory of functions of several variables, number theory, and geometry.

Definition 8.21 Quadratic Form

*A **quadratic form** q in the n variables $x_1, x_2, \dots, x_n$ is a linear combination of terms $x_1^2, x_2^2, \dots, x_n^2$, and cross terms $x_1x_2, x_1x_3, x_2x_3, \dots$.*

If $n = 3$, q has the form

$$q = a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 + a_{12}x_1x_2 + a_{21}x_2x_1 + a_{13}x_1x_3 + a_{31}x_3x_1 + a_{23}x_2x_3 + a_{32}x_3x_2$$

In general

$$q = a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{nn}x_n^2 + a_{12}x_1x_2 + a_{13}x_1x_3 + \cdots$$

This sum can be written compactly as a matrix product

$$q = q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$$

where $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is thought of as a column, and $A = \left[a_{ij}\right]$ is a real $n \times n$ matrix. Note that if $i \neq j$, two separate terms $a_{ij}x_ix_j$ and $a_{ji}x_jx_i$ are listed, each of which involves x_ix_j, and they can (rather cleverly) be replaced by

$$\tfrac{1}{2}(a_{ij} + a_{ji})x_ix_j \quad \text{and} \quad \tfrac{1}{2}(a_{ij} + a_{ji})x_jx_i$$

respectively, *without altering the quadratic form*. Hence there is no loss of generality in assuming that x_ix_j and x_jx_i have the same coefficient in the sum for q. In other words, **we may assume that A is symmetric**.

Example 8.9.1

Write $q = x_1^2 + 3x_3^2 + 2x_1x_2 - x_1x_3$ in the form $q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$, where A is a symmetric 3×3 matrix.

Solution. The cross terms are $2x_1x_2 = x_1x_2 + x_2x_1$ and $-x_1x_3 = -\frac{1}{2}x_1x_3 - \frac{1}{2}x_3x_1$.
Of course, x_2x_3 and x_3x_2 both have coefficient zero, as does x_2^2. Hence

$$q(\mathbf{x}) = \left[\begin{array}{ccc} x_1 & x_2 & x_3 \end{array}\right] \left[\begin{array}{rrr} 1 & 1 & -\frac{1}{2} \\ 1 & 0 & 0 \\ -\frac{1}{2} & 0 & 3 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right]$$

is the required form (verify).

We shall assume from now on that all quadratic forms are given by

$$q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$$

where A is symmetric. Given such a form, the problem is to find new variables $y_1, y_2, \dots, y_n$, related to $x_1, x_2, \dots, x_n$, with the property that when q is expressed in terms of $y_1, y_2, \dots, y_n$, there are no cross terms. If we write

$$\mathbf{y} = (y_1, y_2, \dots, y_n)^T$$

this amounts to asking that $q = \mathbf{y}^T D\mathbf{y}$ where D is diagonal. It turns out that this can always be accomplished and, not surprisingly, that D is the matrix obtained when the symmetric matrix A is orthogonally diagonalized. In fact, as Theorem 8.2.2 shows, a matrix P can be found that is orthogonal (that is, $P^{-1} = P^T$) and diagonalizes A:

$$P^T AP = D = \left[\begin{array}{cccc} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{array}\right]$$

The diagonal entries $\lambda_1, \lambda_2, \dots, \lambda_n$ are the (not necessarily distinct) eigenvalues of A, repeated according to their multiplicities in $c_A(x)$, and the columns of P are corresponding (orthonormal) eigenvectors of A. As A is symmetric, the λ_i are real by Theorem 5.5.7.

Now define new variables $\mathbf{y}$ by the equations

$$\mathbf{x} = P\mathbf{y} \quad \text{equivalently} \quad \mathbf{y} = P^T\mathbf{x}$$

Then substitution in $q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ gives

$$q = (P\mathbf{y})^T A(P\mathbf{y}) = \mathbf{y}^T(P^T AP)\mathbf{y} = \mathbf{y}^T D\mathbf{y} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2$$

Hence this change of variables produces the desired simplification in q.

Theorem 8.9.1: Diagonalization Theorem

Let $q = \mathbf{x}^T A\mathbf{x}$ be a quadratic form in the variables $x_1, x_2, \dots, x_n$, where $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ and A is a symmetric $n \times n$ matrix. Let P be an orthogonal matrix such that $P^T AP$ is diagonal, and define new variables $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ by

$$\mathbf{x} = P\mathbf{y} \quad \text{equivalently} \quad \mathbf{y} = P^T\mathbf{x}$$

If q is expressed in terms of these new variables $y_1, y_2, \dots, y_n$, the result is

$$q = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2$$

where $\lambda_1, \lambda_2, \dots, \lambda_n$ are the eigenvalues of A repeated according to their multiplicities.

Let $q = \mathbf{x}^T A\mathbf{x}$ be a quadratic form where A is a symmetric matrix and let $\lambda_1, \dots, \lambda_n$ be the (real) eigenvalues of A repeated according to their multiplicities. A corresponding set $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$ of orthonormal eigenvectors for A is called a set of **principal axes** for the quadratic form q. (The reason for the name will become clear later.) The orthogonal matrix P in Theorem 8.9.1 is given as $P = \begin{bmatrix} \mathbf{f}_1 & \cdots & \mathbf{f}_n \end{bmatrix}$, so the variables X and Y are related by

$$\mathbf{x} = P\mathbf{y} = \begin{bmatrix} \mathbf{f}_1 & \mathbf{f}_2 & \cdots & \mathbf{f}_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = y_1\mathbf{f}_1 + y_2\mathbf{f}_2 + \cdots + y_n\mathbf{f}_n$$

Thus the new variables y_i are the coefficients when $\mathbf{x}$ is expanded in terms of the orthonormal basis $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$ of $\mathbb{R}^n$. In particular, the coefficients y_i are given by $y_i = \mathbf{x} \cdot \mathbf{f}_i$ by the expansion theorem (Theorem 5.3.6). Hence q itself is easily computed from the eigenvalues λ_i and the principal axes $\mathbf{f}_i$:

$$q = q(\mathbf{x}) = \lambda_1(\mathbf{x} \cdot \mathbf{f}_1)^2 + \cdots + \lambda_n(\mathbf{x} \cdot \mathbf{f}_n)^2$$

Example 8.9.2

Find new variables y_1, y_2, y_3, and y_4 such that

$$q = 3(x_1^2 + x_2^2 + x_3^2 + x_4^2) + 2x_1x_2 - 10x_1x_3 + 10x_1x_4 + 10x_2x_3 - 10x_2x_4 + 2x_3x_4$$

has diagonal form, and find the corresponding principal axes.

Solution. The form can be written as $q = \mathbf{x}^T A\mathbf{x}$, where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} 3 & 1 & -5 & 5 \\ 1 & 3 & 5 & -5 \\ -5 & 5 & 3 & 1 \\ 5 & -5 & 1 & 3 \end{bmatrix}$$

A routine calculation yields

$$c_A(x) = \det(xI - A) = (x - 12)(x + 8)(x - 4)^2$$

so the eigenvalues are $\lambda_1 = 12$, $\lambda_2 = -8$, and $\lambda_3 = \lambda_4 = 4$. Corresponding orthonormal eigenvectors are the principal axes:

$$\mathbf{f}_1 = \frac{1}{2}\left[\begin{array}{r} 1 \\ -1 \\ -1 \\ 1 \end{array}\right] \quad \mathbf{f}_2 = \frac{1}{2}\left[\begin{array}{r} 1 \\ -1 \\ 1 \\ -1 \end{array}\right] \quad \mathbf{f}_3 = \frac{1}{2}\left[\begin{array}{r} 1 \\ 1 \\ 1 \\ 1 \end{array}\right] \quad \mathbf{f}_4 = \frac{1}{2}\left[\begin{array}{r} 1 \\ 1 \\ -1 \\ -1 \end{array}\right]$$

The matrix

$$P = \left[\begin{array}{cccc} \mathbf{f}_1 & \mathbf{f}_2 & \mathbf{f}_3 & \mathbf{f}_4 \end{array}\right] = \frac{1}{2}\left[\begin{array}{rrrr} 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 \end{array}\right]$$

is thus orthogonal, and $P^{-1}AP = P^TAP$ is diagonal. Hence the new variables $\mathbf{y}$ and the old variables $\mathbf{x}$ are related by $\mathbf{y} = P^T\mathbf{x}$ and $\mathbf{x} = P\mathbf{y}$. Explicitly,

$$\begin{aligned} y_1 &= \tfrac{1}{2}(x_1 - x_2 - x_3 + x_4) & \qquad x_1 &= \tfrac{1}{2}(y_1 + y_2 + y_3 + y_4) \\ y_2 &= \tfrac{1}{2}(x_1 - x_2 + x_3 - x_4) & x_2 &= \tfrac{1}{2}(-y_1 - y_2 + y_3 + y_4) \\ y_3 &= \tfrac{1}{2}(x_1 + x_2 + x_3 + x_4) & x_3 &= \tfrac{1}{2}(-y_1 + y_2 + y_3 - y_4) \\ y_4 &= \tfrac{1}{2}(x_1 + x_2 - x_3 - x_4) & x_4 &= \tfrac{1}{2}(y_1 - y_2 + y_3 - y_4) \end{aligned}$$

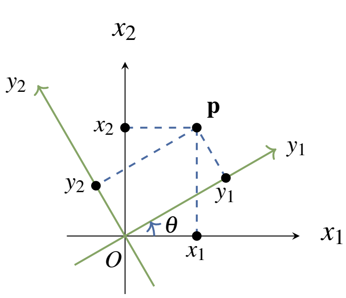
If these x_i are substituted in the original expression for q, the result is

$$q = 12y_1^2 - 8y_2^2 + 4y_3^2 + 4y_4^2$$

This is the required diagonal form.

It is instructive to look at the case of quadratic forms in two variables x_1 and x_2. Then the principal axes can always be found by rotating the x_1 and x_2 axes counterclockwise about the origin through an angle θ. This rotation is a linear transformation $R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$, and it is shown in Theorem 2.6.4 that R_θ has matrix $P = \left[\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right]$. If $\{\mathbf{e}_1,\ \mathbf{e}_2\}$ denotes the standard basis of $\mathbb{R}^2$, the rotation produces a new basis $\{\mathbf{f}_1,\ \mathbf{f}_2\}$ given by

$$\mathbf{f}_1 = R_\theta(\mathbf{e}_1) = \left[\begin{array}{c} \cos\theta \\ \sin\theta \end{array}\right] \quad \text{and} \quad \mathbf{f}_2 = R_\theta(\mathbf{e}_2) = \left[\begin{array}{c} -\sin\theta \\ \cos\theta \end{array}\right] \tag{8.7}$$



Given a point $\mathbf{p} = \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] = x_1\mathbf{e}_1 + x_2\mathbf{e}_2$ in the original system, let y_1 and y_2 be the coordinates of $\mathbf{p}$ in the new system (see the diagram). That is,

$$\left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] = \mathbf{p} = y_1\mathbf{f}_1 + y_2\mathbf{f}_2 = \left[\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right]\left[\begin{array}{c} y_1 \\ y_2 \end{array}\right] \tag{8.8}$$

Writing $\mathbf{x} = \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right]$ and $\mathbf{y} = \left[\begin{array}{c} y_1 \\ y_2 \end{array}\right]$, this reads $\mathbf{x} = P\mathbf{y}$ so, since P is orthogonal, this is the change of variables formula for the rotation as in Theorem 8.9.1.

If $r \neq 0 \neq s$, the graph of the equation $rx_1^2 + sx_2^2 = 1$ is called an **ellipse** if $rs > 0$ and a **hyperbola** if $rs < 0$. More generally, given a quadratic form

$$q = ax_1^2 + bx_1x_2 + cx_2^2 \quad \text{where not all of } a,\ b, \text{ and } c \text{ are zero}$$

the graph of the equation $q = 1$ is called a **conic**. We can now completely describe this graph. There are two special cases which we leave to the reader.

1. *If exactly one of a and c is zero, then the graph of $q = 1$ is a* **parabola**.

So we assume that $a \neq 0$ and $c \neq 0$. In this case, the description depends on the quantity $b^2 - 4ac$, called the **discriminant** of the quadratic form q.

2. *If $b^2 - 4ac = 0$, then either both $a \geq 0$ and $c \geq 0$, or both $a \leq 0$ and $c \leq 0$.*
   Hence $q = (\sqrt{a}x_1 + \sqrt{c}x_2)^2$ or $q = (\sqrt{-a}x_1 + \sqrt{-c}x_2)^2$, so the graph of $q = 1$ is a **pair of straight lines** *in either case.*

So we also assume that $b^2 - 4ac \neq 0$. But then the next theorem asserts that there exists a rotation of the plane about the origin which transforms the equation $ax_1^2 + bx_1x_2 + cx_2^2 = 1$ into either an ellipse or a hyperbola, and the theorem also provides a simple way to decide which conic it is.

Theorem 8.9.2

Consider the quadratic form $q = ax_1^2 + bx_1x_2 + cx_2^2$ where a, c, and $b^2 - 4ac$ are all nonzero.

1. *There is a counterclockwise rotation of the coordinate axes about the origin such that, in the new coordinate system, q has no cross term.*
2. *The graph of the equation*
$$ax_1^2 + bx_1x_2 + cx_2^2 = 1$$
   is an ellipse if $b^2 - 4ac < 0$ and an hyperbola if $b^2 - 4ac > 0$.

Proof. If $b = 0$, q *already* has no cross term and (1) and (2) are clear. So assume $b \neq 0$. The matrix $A = \begin{bmatrix} a & \frac{1}{2}b \\ \frac{1}{2}b & c \end{bmatrix}$ of q has characteristic polynomial $c_A(x) = x^2 - (a+c)x - \frac{1}{4}(b^2 - 4ac)$. If we write $d = \sqrt{b^2 + (a-c)^2}$ for convenience; then the quadratic formula gives the eigenvalues

$$\lambda_1 = \tfrac{1}{2}[a + c - d] \quad \text{and} \quad \lambda_2 = \tfrac{1}{2}[a + c + d]$$

with corresponding principal axes

$$\mathbf{f}_1 = \frac{1}{\sqrt{b^2 + (a-c-d)^2}} \begin{bmatrix} a - c - d \\ b \end{bmatrix} \quad \text{and}$$
$$\mathbf{f}_2 = \frac{1}{\sqrt{b^2 + (a-c-d)^2}} \begin{bmatrix} -b \\ a - c - d \end{bmatrix}$$

as the reader can verify. These agree with equation (8.7) above if θ is an angle such that

$$\cos\theta = \frac{a-c-d}{\sqrt{b^2+(a-c-d)^2}} \qquad \text{and} \qquad \sin\theta = \frac{b}{\sqrt{b^2+(a-c-d)^2}}$$

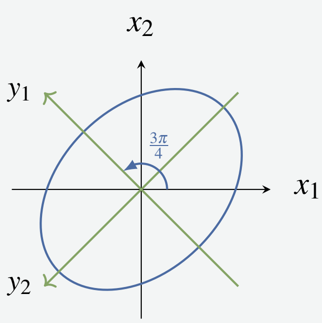
Then $P = \left[\begin{array}{cc} \mathbf{f}_1 & \mathbf{f}_2 \end{array}\right] = \left[\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right]$ diagonalizes A and equation (8.8) becomes the formula $\mathbf{x} = P\mathbf{y}$ in Theorem 8.9.1. This proves (1).

Finally, A is similar to $\left[\begin{array}{cc} \lambda_1 & 0 \\ 0 & \lambda_2 \end{array}\right]$ so $\lambda_1\lambda_2 = \det A = \frac{1}{4}(4ac - b^2)$. Hence the graph of $\lambda_1 y_1^2 + \lambda_2 y_2^2 = 1$ is an ellipse if $b^2 < 4ac$ and an hyperbola if $b^2 > 4ac$. This proves (2). □

Example 8.9.3

Consider the equation $x^2 + xy + y^2 = 1$. Find a rotation so that the equation has no cross term.

Solution.



Here $a = b = c = 1$ in the notation of Theorem 8.9.2, so $\cos\theta = \frac{-1}{\sqrt{2}}$ and $\sin\theta = \frac{1}{\sqrt{2}}$. Hence $\theta = \frac{3\pi}{4}$ will do it. The new variables are $y_1 = \frac{1}{\sqrt{2}}(x_2 - x_1)$ and $y_2 = \frac{-1}{\sqrt{2}}(x_2 + x_1)$ by (8.8), and the equation becomes $y_1^2 + 3y_2^2 = 2$. The angle θ has been chosen such that the new y_1 and y_2 axes are the axes of symmetry of the ellipse (see the diagram). The eigenvectors $\mathbf{f}_1 = \frac{1}{\sqrt{2}}\left[\begin{array}{c} -1 \\ 1 \end{array}\right]$ and $\mathbf{f}_2 = \frac{1}{\sqrt{2}}\left[\begin{array}{c} -1 \\ -1 \end{array}\right]$ point along these axes of symmetry, and this is the reason for the name *principal axes*.

The determinant of any orthogonal matrix P is either 1 or -1 (because $PP^T = I$). The orthogonal matrices $\left[\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right]$ arising from rotations all have determinant 1. More generally, given any quadratic form $q = \mathbf{x}^T A\mathbf{x}$, the orthogonal matrix P such that $P^T AP$ is diagonal can always be chosen so that $\det P = 1$ by interchanging two eigenvalues (and hence the corresponding columns of P). It is shown in Theorem 10.4.4 that orthogonal 2×2 matrices with determinant 1 correspond to rotations. Similarly, it can be shown that orthogonal 3×3 matrices with determinant 1 correspond to rotations about a line through the origin. This extends Theorem 8.9.2: Every quadratic form in two or three variables can be diagonalized by a rotation of the coordinate system.

Congruence

We return to the study of quadratic forms in general.

Theorem 8.9.3

If $q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ is a quadratic form given by a symmetric matrix A, then A is uniquely determined by q.

Proof. Let $q(\mathbf{x}) = \mathbf{x}^T B\mathbf{x}$ for all $\mathbf{x}$ where $B^T = B$. If $C = A - B$, then $C^T = C$ and $\mathbf{x}^T C\mathbf{x} = 0$ for all $\mathbf{x}$. We must show that $C = 0$. Given $\mathbf{y}$ in $\mathbb{R}^n$,

$$\begin{aligned} 0 = (\mathbf{x}+\mathbf{y})^T C(\mathbf{x}+\mathbf{y}) &= \mathbf{x}^T C\mathbf{x} + \mathbf{x}^T C\mathbf{y} + \mathbf{y}^T C\mathbf{x} + \mathbf{y}^T C\mathbf{y} \\ &= \mathbf{x}^T C\mathbf{y} + \mathbf{y}^T C\mathbf{x} \end{aligned}$$

But $\mathbf{y}^T C\mathbf{x} = (\mathbf{x}^T C\mathbf{y})^T = \mathbf{x}^T C\mathbf{y}$ (it is 1×1). Hence $\mathbf{x}^T C\mathbf{y} = 0$ for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$. If $\mathbf{e}_j$ is column j of I_n, then the $(i,\ j)$-entry of C is $\mathbf{e}_i^T C\mathbf{e}_j = 0$. Thus $C = 0$. □

Hence we can speak of *the* symmetric matrix of a quadratic form.

On the other hand, a quadratic form q in variables x_i can be written in several ways as a linear combination of squares of new variables, even if the new variables are required to be linear combinations of the x_i. For example, if $q = 2x_1^2 - 4x_1x_2 + x_2^2$ then

$$q = 2(x_1 - x_2)^2 - x_2^2 \quad \text{and} \quad q = -2x_1^2 + (2x_1 - x_2)^2$$

The question arises: How are these changes of variables related, and what properties do they share? To investigate this, we need a new concept.

Let a quadratic form $q = q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ be given in terms of variables $\mathbf{x} = (x_1,\ x_2,\ \dots,\ x_n)^T$. If the new variables $\mathbf{y} = (y_1,\ y_2,\ \dots,\ y_n)^T$ are to be linear combinations of the x_i, then $\mathbf{y} = A\mathbf{x}$ for some $n \times n$ matrix A. Moreover, since we want to be able to solve for the x_i in terms of the y_i, we ask that the matrix A be invertible. Hence suppose U is an invertible matrix and that the new variables $\mathbf{y}$ are given by

$$\mathbf{y} = U^{-1}\mathbf{x}, \qquad \text{equivalently } \mathbf{x} = U\mathbf{y}$$

In terms of these new variables, q takes the form

$$q = q(\mathbf{x}) = (U\mathbf{y})^T A(U\mathbf{y}) = \mathbf{y}^T(U^T AU)\mathbf{y}$$

That is, q has matrix $U^T AU$ with respect to the new variables $\mathbf{y}$. Hence, to study changes of variables in quadratic forms, we study the following relationship on matrices: Two $n \times n$ matrices A and B are called **congruent**, written $A \overset{c}{\sim} B$, if $B = U^T AU$ for some invertible matrix U. Here are some properties of congruence:

1. *$A \overset{c}{\sim} A$ for all A.*
2. *If $A \overset{c}{\sim} B$, then $B \overset{c}{\sim} A$.*
3. *If $A \overset{c}{\sim} B$ and $B \overset{c}{\sim} C$, then $A \overset{c}{\sim} C$.*

4. *If* $A \overset{c}{\sim} B$, *then* A *is symmetric if and only if* B *is symmetric.*
5. *If* $A \overset{c}{\sim} B$, *then* $\text{rank } A = \text{rank } B$.

The converse to (5) can fail even for symmetric matrices.

Example 8.9.4

The symmetric matrices $A = \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right]$ and $B = \left[\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array} \right]$ have the same rank but are not congruent. Indeed, if $A \overset{c}{\sim} B$, an invertible matrix U exists such that $B = U^T AU = U^T U$. But then $-1 = \det B = (\det U)^2$, a contradiction.

The key distinction between A and B in Example 8.9.4 is that A has two positive eigenvalues (counting multiplicities) whereas B has only one.

Theorem 8.9.4: Sylvester's Law of Inertia

If $A \overset{c}{\sim} B$, *then* A *and* B *have the same number of positive eigenvalues, counting multiplicities.*

The proof is given at the end of this section.

The **index** of a symmetric matrix A is the number of positive eigenvalues of A. If $q = q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ is a quadratic form, the **index** and **rank** of q are defined to be, respectively, the index and rank of the matrix A. As we saw before, if the variables expressing a quadratic form q are changed, the new matrix is congruent to the old one. Hence the index and rank depend only on q and not on the way it is expressed.

Now let $q = q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ be any quadratic form in n variables, of index k and rank r, where A is symmetric. We claim that new variables $\mathbf{z}$ can be found so that q is **completely diagonalized**—that is,

$$q(\mathbf{z}) = z_1^2 + \cdots + z_k^2 - z_{k+1}^2 - \cdots - z_r^2$$

If $k \leq r \leq n$, let $D_n(k, r)$ denote the $n \times n$ diagonal matrix whose main diagonal consists of k ones, followed by $r-k$ minus ones, followed by $n-r$ zeros. Then we seek new variables $\mathbf{z}$ such that

$$q(\mathbf{z}) = \mathbf{z}^T D_n(k, r)\mathbf{z}$$

To determine $\mathbf{z}$, first diagonalize A as follows: Find an orthogonal matrix P_0 such that

$$P_0^T AP_0 = D = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r, 0, \dots, 0)$$

is diagonal with the nonzero eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_r$ of A on the main diagonal (followed by $n-r$ zeros). By reordering the columns of P_0, if necessary, we may assume that $\lambda_1, \dots, \lambda_k$ are positive and $\lambda_{k+1}, \dots, \lambda_r$ are negative. This being the case, let D_0 be the $n \times n$ diagonal matrix

$$D_0 = \text{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \dots, \frac{1}{\sqrt{\lambda_k}}, \frac{1}{\sqrt{-\lambda_{k+1}}}, \dots, \frac{1}{\sqrt{-\lambda_r}}, 1, \dots, 1\right)$$

Then $D_0^T DD_0 = D_n(k, r)$, so if new variables $\mathbf{z}$ are given by $\mathbf{x} = (P_0 D_0)\mathbf{z}$, we obtain

$$q(\mathbf{z}) = \mathbf{z}^T D_n(k, r)\mathbf{z} = z_1^2 + \cdots + z_k^2 - z_{k+1}^2 - \cdots - z_r^2$$

as required. Note that the change-of-variables matrix $P_0 D_0$ from $\mathbf{z}$ to $\mathbf{x}$ has orthogonal columns (in fact, scalar multiples of the columns of P_0).

Example 8.9.5

Completely diagonalize the quadratic form q in Example 8.9.2 and find the index and rank.

Solution. In the notation of Example 8.9.2, the eigenvalues of the matrix A of q are 12, -8, 4, 4; so the index is 3 and the rank is 4. Moreover, the corresponding orthogonal eigenvectors are $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3$ (see Example 8.9.2), and $\mathbf{f}_4$. Hence $P_0 = \left[\begin{array}{cccc} \mathbf{f}_1 & \mathbf{f}_3 & \mathbf{f}_4 & \mathbf{f}_2 \end{array} \right]$ is orthogonal and

$$P_0^T AP_0 = \text{diag}\,(12,\ 4,\ 4,\ -8)$$

As before, take $D_0 = \text{diag}\,(\frac{1}{\sqrt{12}},\ \frac{1}{2},\ \frac{1}{2},\ \frac{1}{\sqrt{8}})$ and define the new variables $\mathbf{z}$ by $\mathbf{x} = (P_0D_0)\mathbf{z}$. Hence the new variables are given by $\mathbf{z} = D_0^{-1}P_0^T\mathbf{x}$. The result is

$$\begin{aligned}
z_1 &= \sqrt{3}(x_1 - x_2 - x_3 + x_4) \\
z_2 &= x_1 + x_2 + x_3 + x_4 \\
z_3 &= x_1 + x_2 - x_3 - x_4 \\
z_4 &= \sqrt{2}(x_1 - x_2 + x_3 - x_4)
\end{aligned}$$

This discussion gives the following information about symmetric matrices.

Theorem 8.9.5

Let A and B be symmetric $n \times n$ matrices, and let $0 \leq k \leq r \leq n$.

1. *A has index k and rank r if and only if $A \overset{c}{\sim} D_n(k,\ r)$.*

2. *$A \overset{c}{\sim} B$ if and only if they have the same rank and index.*

Proof.

1. If A has index k and rank r, take $U = P_0D_0$ where P_0 and D_0 are as described prior to Example 8.9.5. Then $U^TAU = D_n(k,\ r)$. The converse is true because $D_n(k,\ r)$ has index k and rank r (using Theorem 8.9.4).

2. If A and B both have index k and rank r, then $A \overset{c}{\sim} D_n(k,\ r) \overset{c}{\sim} B$ by (1). The converse was given earlier.

□

Proof of Theorem 8.9.4.

By Theorem 8.9.1, $A \overset{c}{\sim} D_1$ and $B \overset{c}{\sim} D_2$ where D_1 and D_2 are diagonal and have the same eigenvalues as A and B, respectively. We have $D_1 \overset{c}{\sim} D_2$, (because $A \overset{c}{\sim} B$), so we may assume that A and B are both diagonal. Consider the quadratic form $q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$. If A has k positive eigenvalues, q has the form

$$q(\mathbf{x}) = a_1x_1^2 + \cdots + a_kx_k^2 - a_{k+1}x_{k+1}^2 - \cdots - a_rx_r^2, \quad a_i > 0$$

where $r = \text{rank } A = \text{rank } B$. The subspace $W_1 = \{\mathbf{x} \mid x_{k+1} = \cdots = x_r = 0\}$ of $\mathbb{R}^n$ has dimension $n - r + k$ and satisfies $q(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \mathbf{0}$ in W_1.

On the other hand, if $B = U^T AU$, define new variables $\mathbf{y}$ by $\mathbf{x} = U\mathbf{y}$. If B has k' positive eigenvalues, q has the form

$$q(\mathbf{x}) = b_1y_1^2 + \cdots + b_{k'}y_{k'}^2 - b_{k'+1}y_{k'+1}^2 - \cdots - b_ry_r^2, \quad b_i > 0$$

Let $\mathbf{f}_1, \ldots, \mathbf{f}_n$ denote the columns of U. They are a basis of $\mathbb{R}^n$ and

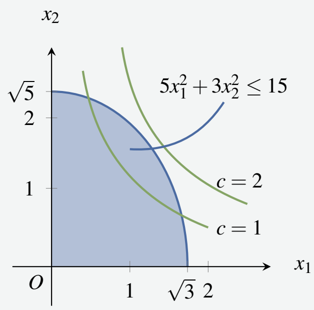
$$\mathbf{x} = U\mathbf{y} = \left[\begin{array}{ccc} \mathbf{f}_1 & \cdots & \mathbf{f}_n \end{array} \right] \left[\begin{array}{c} y_1 \\ \vdots \\ y_n \end{array} \right] = y_1\mathbf{f}_1 + \cdots + y_n\mathbf{f}_n$$

Hence the subspace $W_2 = \text{span}\{\mathbf{f}_{k'+1}, \ldots, \mathbf{f}_r\}$ satisfies $q(\mathbf{x}) < 0$ for all $\mathbf{x} \neq \mathbf{0}$ in W_2. Note $\dim W_2 = r - k'$. It follows that W_1 and W_2 have only the zero vector in common. Hence, if B_1 and B_2 are bases of W_1 and W_2, respectively, then (Exercise **??**) $B_1 \cup B_2$ is an independent set of $(n - r + k) + (r - k') = n + k - k'$ vectors in $\mathbb{R}^n$. This implies that $k \leq k'$, and a similar argument shows $k' \leq k$. □

8.10 An Application to Constrained Optimization

It is a frequent occurrence in applications that a function $q = q(x_1, x_2, \dots, x_n)$ of n variables, called an **objective function**, is to be made as large or as small as possible among all vectors $\mathbf{x} = (x_1, x_2, \dots, x_n)$ lying in a certain region of $\mathbb{R}^n$ called the **feasible region**. A wide variety of objective functions q arise in practice; our primary concern here is to examine one important situation where q is a quadratic form. The next example gives some indication of how such problems arise.

Example 8.10.1



A politician proposes to spend x_1 dollars annually on health care and x_2 dollars annually on education. She is constrained in her spending by various budget pressures, and one model of this is that the expenditures x_1 and x_2 should satisfy a constraint like

$$5x_1^2 + 3x_2^2 \leq 15$$

Since $x_i \geq 0$ for each i, the feasible region is the shaded area shown in the diagram. Any choice of feasible point (x_1, x_2) in this region will satisfy the budget constraints. However, these choices have different effects on voters, and the politician wants to choose $\mathbf{x} = (x_1, x_2)$ to maximize some measure $q = q(x_1, x_2)$ of voter satisfaction. Thus the assumption is that, for any value of c, all points on the graph of $q(x_1, x_2) = c$ have the same appeal to voters. Hence the goal is to find the largest value of c for which the graph of $q(x_1, x_2) = c$ contains a feasible point.

The choice of the function q depends upon many factors; we will show how to solve the problem for any quadratic form q (even with more than two variables). In the diagram the function q is given by

$$q(x_1, x_2) = x_1x_2$$

and the graphs of $q(x_1, x_2) = c$ are shown for $c = 1$ and $c = 2$. As c increases the graph of $q(x_1, x_2) = c$ moves up and to the right. From this it is clear that there will be a solution for some value of c between 1 and 2 (in fact the largest value is $c = \frac{1}{2}\sqrt{15} = 1.94$ to two decimal places).

The constraint $5x_1^2 + 3x_2^2 \leq 15$ in Example 8.10.1 can be put in a standard form. If we divide through by 15, it becomes $\left(\frac{x_1}{\sqrt{3}}\right)^2 + \left(\frac{x_2}{\sqrt{5}}\right)^2 \leq 1$. This suggests that we introduce new variables $\mathbf{y} = (y_1, y_2)$ where $y_1 = \frac{x_1}{\sqrt{3}}$ and $y_2 = \frac{x_2}{\sqrt{5}}$. Then the constraint becomes $\|\mathbf{y}\|^2 \leq 1$, equivalently $\|\mathbf{y}\| \leq 1$. In terms of these new variables, the objective function is $q = \sqrt{15}y_1y_2$, and we want to maximize this subject to $\|\mathbf{y}\| \leq 1$. When this is done, the maximizing values of x_1 and x_2 are obtained from $x_1 = \sqrt{3}y_1$ and $x_2 = \sqrt{5}y_2$.

Hence, for constraints like that in Example 8.10.1, there is no real loss in generality in assuming that the constraint takes the form $\|\mathbf{x}\| \leq 1$. In this case the principal axes theorem solves the problem. Recall that a vector in $\mathbb{R}^n$ of length 1 is called a *unit vector*.

Theorem 8.10.1

Consider the quadratic form $q = q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ where A is an $n \times n$ symmetric matrix, and let λ_1 and λ_n denote the largest and smallest eigenvalues of A, respectively. Then:

1. $\max\{q(\mathbf{x}) \mid \|\mathbf{x}\| \leq 1\} = \lambda_1$, *and* $q(\mathbf{f}_1) = \lambda_1$ *where* $\mathbf{f}_1$ *is any unit* λ_1*-eigenvector.*
2. $\min\{q(\mathbf{x}) \mid \|\mathbf{x}\| \leq 1\} = \lambda_n$, *and* $q(\mathbf{f}_n) = \lambda_n$ *where* $\mathbf{f}_n$ *is any unit* λ_n*-eigenvector.*

Proof. Since A is symmetric, let the (real) eigenvalues λ_i of A be ordered as to size as follows:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

By the principal axes theorem, let P be an orthogonal matrix such that $P^T AP = D = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$. Define $\mathbf{y} = P^T \mathbf{x}$, equivalently $\mathbf{x} = P\mathbf{y}$, and note $\|\mathbf{y}\| = \|\mathbf{x}\|$ because $\|\mathbf{y}\|^2 = \mathbf{y}^T \mathbf{y} = \mathbf{x}^T (PP^T)\mathbf{x} = \mathbf{x}^T \mathbf{x} = \|\mathbf{x}\|^2$. If we write $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$, then

$$\begin{aligned} q(\mathbf{x}) &= q(P\mathbf{y}) = (P\mathbf{y})^T A(P\mathbf{y}) \\ &= \mathbf{y}^T (P^T AP)\mathbf{y} = \mathbf{y}^T D\mathbf{y} \\ &= \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2 \end{aligned} \tag{8.9}$$

Now assume that $\|\mathbf{x}\| \leq 1$. Since $\lambda_i \leq \lambda_1$ for each i, (8.9) gives

$$q(\mathbf{x}) = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2 \leq \lambda_1 y_1^2 + \lambda_1 y_2^2 + \cdots + \lambda_1 y_n^2 = \lambda_1 \|\mathbf{y}\|^2 \leq \lambda_1$$

because $\|\mathbf{y}\| = \|\mathbf{x}\| \leq 1$. This shows that $q(\mathbf{x})$ cannot exceed λ_1 when $\|\mathbf{x}\| \leq 1$. To see that this maximum is actually achieved, let $\mathbf{f}_1$ be a unit eigenvector corresponding to λ_1. Then

$$q(\mathbf{f}_1) = \mathbf{f}_1^T A\mathbf{f}_1 = \mathbf{f}_1^T (\lambda_1 \mathbf{f}_1) = \lambda_1 (\mathbf{f}_1^T \mathbf{f}_1) = \lambda_1 \|\mathbf{f}_1\|^2 = \lambda_1$$

Hence λ_1 is the maximum value of $q(\mathbf{x})$ when $\|\mathbf{x}\| \leq 1$, proving (1). The proof of (2) is analogous. □

The set of all vectors $\mathbf{x}$ in $\mathbb{R}^n$ such that $\|\mathbf{x}\| \leq 1$ is called the **unit ball**. If $n = 2$, it is often called the unit disk and consists of the unit circle and its interior; if $n = 3$, it is the unit sphere and its interior. It is worth noting that the maximum value of a quadratic form $q(\mathbf{x})$ as $\mathbf{x}$ ranges *throughout* the unit ball is (by Theorem 8.10.1) actually attained for a unit vector $\mathbf{x}$ on the *boundary* of the unit ball.

Theorem 8.10.1 is important for applications involving vibrations in areas as diverse as aerodynamics and particle physics, and the maximum and minimum values in the theorem are often found using advanced calculus to minimize the quadratic form on the unit ball. The algebraic approach using the principal axes theorem gives a geometrical interpretation of the optimal values because they are eigenvalues.

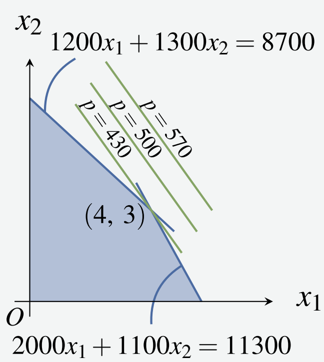
Example 8.10.2

Maximize and minimize the form $q(\mathbf{x}) = 3x_1^2 + 14x_1x_2 + 3x_2^2$ subject to $\|\mathbf{x}\| \leq 1$.

Solution. The matrix of q is $A = \begin{bmatrix} 3 & 7 \\ 7 & 3 \end{bmatrix}$, with eigenvalues $\lambda_1 = 10$ and $\lambda_2 = -4$, and corresponding unit eigenvectors $\mathbf{f}_1 = \frac{1}{\sqrt{2}}(1, 1)$ and $\mathbf{f}_2 = \frac{1}{\sqrt{2}}(1, -1)$. Hence, among all unit vectors $\mathbf{x}$ in $\mathbb{R}^2$, $q(\mathbf{x})$ takes its maximal value 10 at $\mathbf{x} = \mathbf{f}_1$, and the minimum value of $q(\mathbf{x})$ is -4 when $\mathbf{x} = \mathbf{f}_2$.

As noted above, the objective function in a constrained optimization problem need not be a quadratic form. We conclude with an example where the objective function is linear, and the feasible region is determined by linear constraints.

Example 8.10.3



A manufacturer makes x_1 units of product 1, and x_2 units of product 2, at a profit of \$70 and \$50 per unit respectively, and wants to choose x_1 and x_2 to maximize the total profit $p(x_1, x_2) = 70x_1 + 50x_2$. However x_1 and x_2 are not arbitrary; for example, $x_1 \geq 0$ and $x_2 \geq 0$. Other conditions also come into play. Each unit of product 1 costs \$1200 to produce and requires 2000 square feet of warehouse space; each unit of product 2 costs \$1300 to produce and requires 1100 square feet of space. If the total warehouse space is 11 300 square feet, and if the total production budget is \$8700, x_1 and x_2 must also satisfy the conditions

$$\begin{aligned} 2000x_1 + 1100x_2 &\leq 11300 \\ 1200x_1 + 1300x_2 &\leq 8700 \end{aligned}$$

The feasible region in the plane satisfying these constraints (and $x_1 \geq 0$, $x_2 \geq 0$) is shaded in the diagram. If the profit equation $70x_1 + 50x_2 = p$ is plotted for various values of p, the resulting lines are parallel, with p increasing with distance from the origin. Hence the best choice occurs for the line $70x_1 + 50x_2 = 430$ that touches the shaded region at the point $(4, 3)$. So the profit p has a maximum of $p = 430$ for $x_1 = 4$ units and $x_2 = 3$ units.

Example 8.10.3 is a simple case of the general **linear programming** problem[23] which arises in economic, management, network, and scheduling applications. Here the objective function is a linear combination $q = a_1x_1 + a_2x_2 + \cdots + a_nx_n$ of the variables, and the feasible region consists of the vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ in $\mathbb{R}^n$ which satisfy a set of linear inequalities of the form $b_1x_1 + b_2x_2 + \cdots + b_nx_n \leq b$. There is a good method (an extension of the gaussian algorithm) called the **simplex algorithm** for finding the maximum and minimum values of q when $\mathbf{x}$ ranges over such a feasible set. As Example 8.10.3 suggests, the optimal values turn out to be vertices of the feasible set. In particular, they are on the boundary of the feasible region, as is the case in Theorem 8.10.1.

[23]More information is available in "Linear Programming and Extensions" by N. Wu and R. Coppins, McGraw-Hill, 1981.

8.11 An Application to Statistical Principal Component Analysis

Linear algebra is important in multivariate analysis in statistics, and we conclude with a very short look at one application of diagonalization in this area. A main feature of probability and statistics is the idea of a *random variable* X, that is a real-valued function which takes its values according to a probability law (called its *distribution*). Random variables occur in a wide variety of contexts; examples include the number of meteors falling per square kilometre in a given region, the price of a share of a stock, or the duration of a long distance telephone call from a certain city.

The values of a random variable X are distributed about a central number μ, called the *mean* of X. The mean can be calculated from the distribution as the *expectation* $E(X) = \mu$ of the random variable X. Functions of a random variable are again random variables. In particular, $(X-\mu)^2$ is a random variable, and the *variance* of the random variable X, denoted $\text{var}(X)$, is defined to be the number

$$\text{var}(X) = E\{(X-\mu)^2\} \quad \text{where } \mu = E(X)$$

It is not difficult to see that $\text{var}(X) \geq 0$ for every random variable X. The number $\sigma = \sqrt{\text{var}(X)}$ is called the *standard deviation* of X, and is a measure of how much the values of X are spread about the mean μ of X. A main goal of statistical inference is finding reliable methods for estimating the mean and the standard deviation of a random variable X by sampling the values of X.

If two random variables X and Y are given, and their joint distribution is known, then functions of X and Y are also random variables. In particular, $X+Y$ and aX are random variables for any real number a, and we have

$$E(X+Y) = E(X) + E(Y) \quad \text{and} \quad E(aX) = aE(X).$$ [24]

An important question is how much the random variables X and Y depend on each other. One measure of this is the *covariance* of X and Y, denoted $\text{cov}\,(X,\, Y)$, defined by

$$\text{cov}\,(X,\, Y) = E\{(X-\mu)(Y-\upsilon)\} \quad \text{where } \mu = E(X) \text{ and } \upsilon = E(Y)$$

Clearly, $\text{cov}\,(X,\, X) = \text{var}\,(X)$. If $\text{cov}\,(X,\, Y) = 0$ then X and Y have little relationship to each other and are said to be *uncorrelated*.[25]

Multivariate statistical analysis deals with a family $X_1,\, X_2,\, \dots,\, X_n$ of random variables with means $\mu_i = E(X_i)$ and variances $\sigma_i^2 = \text{var}\,(X_i)$ for each i. Let $\sigma_{ij} = \text{cov}\,(X_i,\, X_j)$ denote the covariance of X_i and X_j. Then the *covariance matrix* of the random variables $X_1,\, X_2,\, \dots,\, X_n$ is defined to be the $n \times n$ matrix

$$\Sigma = [\sigma_{ij}]$$

whose $(i,\, j)$-entry is σ_{ij}. The matrix Σ is clearly symmetric; in fact it can be shown that Σ is **positive semidefinite** in the sense that $\lambda \geq 0$ for every eigenvalue λ of Σ. (In reality, Σ is positive definite in most cases of interest.) So suppose that the eigenvalues of Σ are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. The principal axes theorem (Theorem 8.2.2) shows that an orthogonal matrix P exists such that

$$P^T \Sigma P = \text{diag}\,(\lambda_1,\, \lambda_2,\, \dots,\, \lambda_n)$$

If we write $\overline{X} = (X_1,\, X_2,\, \dots,\, X_n)$, the procedure for diagonalizing a quadratic form gives new variables $\overline{Y} = (Y_1,\, Y_2,\, \dots,\, Y_n)$ defined by

$$\overline{Y} = P^T \overline{X}$$

These new random variables $Y_1,\, Y_2,\, \dots,\, Y_n$ are called the **principal components** of the original random variables X_i, and are linear combinations of the X_i. Furthermore, it can be shown that

$$\text{cov}\,(Y_i,\, Y_j) = 0 \text{ if } i \neq j \quad \text{and} \quad \text{var}\,(Y_i) = \lambda_i \quad \text{for each } i$$

Of course the principal components Y_i point along the principal axes of the quadratic form $q = \overline{X}^T \Sigma \overline{X}$.

The sum of the variances of a set of random variables is called the **total variance** of the variables, and determining the source of this total variance is one of the benefits of principal component analysis. The fact that the matrices Σ and $\text{diag}\,(\lambda_1,\, \lambda_2,\, \dots,\, \lambda_n)$ are similar means that they have the same trace, that is,

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{nn} = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

This means that the principal components Y_i have the same total variance as the original random variables X_i. Moreover, the fact that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ means that most of this variance resides in the first few Y_i. In practice, statisticians find that studying these first few Y_i (and ignoring the rest) gives an accurate analysis of the total system variability. This results in substantial data reduction since often only a few Y_i suffice for all practical purposes. Furthermore, these Y_i are easily obtained as linear combinations of the X_i. Finally, the analysis of the principal components often reveals relationships among the X_i that were not previously suspected, and so results in interpretations that would not otherwise have been made.

[24] Hence $E(\ )$ is a linear transformation from the vector space of all random variables to the space of real numbers.

[25] If X and Y are independent in the sense of probability theory, then they are uncorrelated; however, the converse is not true in general.

Chapter 9

Change of Basis

If A is an $m \times n$ matrix, the corresponding **matrix transformation** $T_A : \mathbb{R}^n \to \mathbb{R}^m$ is defined by

$$T_A(\mathbf{x}) = A\mathbf{x} \quad \text{for all columns } \mathbf{x} \text{ in } \mathbb{R}^n$$

It was shown in Theorem 2.6.2 that every linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ is a matrix transformation; that is, $T = T_A$ for some $m \times n$ matrix A. Furthermore, the matrix A is uniquely determined by T. In fact, A is given in terms of its columns by

$$A = \left[\begin{array}{cccc} T(\mathbf{e}_1) & T(\mathbf{e}_2) & \cdots & T(\mathbf{e}_n) \end{array} \right]$$

where $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ is the standard basis of $\mathbb{R}^n$.

In this chapter we show how to associate a matrix with *any* linear transformation $T : V \to W$ where V and W are finite-dimensional vector spaces, and we describe how the matrix can be used to compute $T(\mathbf{v})$ for any $\mathbf{v}$ in V. The matrix depends on the choice of a basis B in V and a basis D in W, and is denoted $M_{DB}(T)$. The case when $W = V$ is particularly important. If B and D are two bases of V, we show that the matrices $M_{BB}(T)$ and $M_{DD}(T)$ are similar, that is $M_{DD}(T) = P^{-1}M_{BB}(T)P$ for some invertible matrix P. Moreover, we give an explicit method for constructing P depending only on the bases B and D. This leads to some of the most important theorems in linear algebra, as we shall see in Chapter 11.

9.1 The Matrix of a Linear Transformation

Let $T : V \to W$ be a linear transformation where $\dim V = n$ and $\dim W = m$. The aim in this section is to describe the action of T as multiplication by an $m \times n$ matrix A. The idea is to convert a vector $\mathbf{v}$ in V into a column in $\mathbb{R}^n$, multiply that column by A to get a column in $\mathbb{R}^m$, and convert this column back to get $T(\mathbf{v})$ in W.

Converting vectors to columns is a simple matter, but one small change is needed. Up to now the *order* of the vectors in a basis has been of no importance. However, in this section, we shall speak of an **ordered basis** $\{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$, which is just a basis where the order in which the vectors are listed is taken into account. Hence $\{\mathbf{b}_2, \mathbf{b}_1, \mathbf{b}_3\}$ is a different *ordered* basis from $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$.

If $B = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ is an ordered basis in a vector space V, and if

$$\mathbf{v} = v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + \cdots + v_n\mathbf{b}_n, \qquad v_i \in \mathbb{R}$$

is a vector in V, then the (uniquely determined) numbers $v_1, v_2, \dots, v_n$ are called the **coordinates** of $\mathbf{v}$ with respect to the basis B.

Definition 9.1 Coordinate Vector $C_B(\mathbf{v})$ of $\mathbf{v}$ for a basis B

*The **coordinate vector** of $\mathbf{v}$ with respect to B is defined to be*

$$C_B(\mathbf{v}) = C_B(v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + \cdots + v_n\mathbf{b}_n) = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

The reason for writing $C_B(\mathbf{v})$ as a column instead of a row will become clear later. Note that $C_B(\mathbf{b}_i) = \mathbf{e}_i$ is column i of I_n.

Example 9.1.1

The coordinate vector for $\mathbf{v} = (2, 1, 3)$ with respect to the ordered basis $B = \{(1, 1, 0), (1, 0, 1), (0, 1, 1)\}$ of $\mathbb{R}^3$ is $C_B(\mathbf{v}) = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$ because

$$\mathbf{v} = (2, 1, 3) = 0(1, 1, 0) + 2(1, 0, 1) + 1(0, 1, 1)$$

Theorem 9.1.1

If V has dimension n and $B = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ is any ordered basis of V, the coordinate *transformation $C_B : V \to \mathbb{R}^n$ is an isomorphism. In fact, $C_B^{-1} : \mathbb{R}^n \to V$ is given by*

$$C_B^{-1}\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + \cdots + v_n\mathbf{b}_n \quad \textit{for all} \quad \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \textit{ in } \mathbb{R}^n.$$

Proof. The verification that C_B is linear is Exercise **??**. If $T : \mathbb{R}^n \to V$ is the map denoted C_B^{-1} in the theorem, one verifies (Exercise **??**) that $TC_B = 1_V$ and $C_BT = 1_{\mathbb{R}^n}$. Note that $C_B(\mathbf{b}_j)$ is column j of the identity matrix, so C_B carries the basis B to the standard basis of $\mathbb{R}^n$, proving again that it is an isomorphism (Theorem 7.3.1) □

$$\begin{tikzcd} V \arrow[r, "T"] \arrow[d, "C_B"] & W \arrow[d, "C_D"] \\ \mathbb{R}^n \arrow[r, "T_A"] & \mathbb{R}^m \end{tikzcd}$$

Now let $T : V \to W$ be any linear transformation where $\dim V = n$ and $\dim W = m$, and let $B = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ and D be ordered bases of V and W, respectively. Then $C_B : V \to \mathbb{R}^n$ and $C_D : W \to \mathbb{R}^m$ are isomorphisms and we have the situation shown in the diagram where A is an $m \times n$ matrix (to be determined). In fact, the composite

$$C_DTC_B^{-1} : \mathbb{R}^n \to \mathbb{R}^m \text{ is a linear transformation}$$

so Theorem 2.6.2 shows that a unique $m \times n$ matrix A exists such that

$$C_DTC_B^{-1} = T_A, \qquad \text{equivalently } C_DT = T_AC_B$$

T_A acts by left multiplication by A, so this latter condition is

$$C_D[T(\mathbf{v})] = AC_B(\mathbf{v}) \text{ for all } \mathbf{v} \text{ in } V$$

This requirement completely determines A. Indeed, the fact that $C_B(\mathbf{b}_j)$ is column j of the identity matrix gives

$$\text{column } j \text{ of } A = AC_B(\mathbf{b}_j) = C_D[T(\mathbf{b}_j)]$$

for all j. Hence, in terms of its columns,

$$A = \begin{bmatrix} C_D[T(\mathbf{b}_1)] & C_D[T(\mathbf{b}_2)] & \cdots & C_D[T(\mathbf{b}_n)] \end{bmatrix}$$

Definition 9.2 Matrix $M_{DB}(T)$ of $T : V \to W$ for bases D and B

This is called the ***matrix of T corresponding to the ordered bases B and D****, and we use the following notation:*

$$M_{DB}(T) = \begin{bmatrix} C_D[T(\mathbf{b}_1)] & C_D[T(\mathbf{b}_2)] & \cdots & C_D[T(\mathbf{b}_n)] \end{bmatrix}$$

This discussion is summarized in the following important theorem.

Theorem 9.1.2

Let $T : V \to W$ be a linear transformation where $\dim V = n$ and $\dim W = m$, and let $B = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ and D be ordered bases of V and W, respectively. Then the matrix $M_{DB}(T)$ just given is the unique $m \times n$ matrix A that satisfies

$$C_D T = T_A C_B$$

Hence the defining property of $M_{DB}(T)$ is

$$C_D[T(\mathbf{v})] = M_{DB}(T)C_B(\mathbf{v}) \text{ for all } \mathbf{v} \text{ in } V$$

The matrix $M_{DB}(T)$ is given in terms of its columns by

$$M_{DB}(T) = \left[\begin{array}{cccc} C_D[T(\mathbf{b}_1)] & C_D[T(\mathbf{b}_2)] & \cdots & C_D[T(\mathbf{b}_n)] \end{array}\right]$$

The fact that $T = C_D^{-1} T_A C_B$ means that the action of T on a vector $\mathbf{v}$ in V can be performed by first taking coordinates (that is, applying C_B to $\mathbf{v}$), then multiplying by A (applying T_A), and finally converting the resulting m-tuple back to a vector in W (applying C_D^{-1}).

Example 9.1.2

Define $T : \mathbf{P}_2 \to \mathbb{R}^2$ by $T(a + bx + cx^2) = (a + c, b - a - c)$ for all polynomials $a + bx + cx^2$. If $B = \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ and $D = \{\mathbf{d}_1, \mathbf{d}_2\}$ where

$$\mathbf{b}_1 = 1,\ \mathbf{b}_2 = x,\ \mathbf{b}_3 = x^2 \quad \text{and} \quad \mathbf{d}_1 = (1, 0),\ \mathbf{d}_2 = (0, 1)$$

compute $M_{DB}(T)$ and verify Theorem 9.1.2.

Solution. We have $T(\mathbf{b}_1) = \mathbf{d}_1 - \mathbf{d}_2$, $T(\mathbf{b}_2) = \mathbf{d}_2$, and $T(\mathbf{b}_3) = \mathbf{d}_1 - \mathbf{d}_2$. Hence

$$M_{DB}(T) = \left[\begin{array}{ccc} C_D[T(\mathbf{b}_1)] & C_D[T(\mathbf{b}_2)] & C_D[T(\mathbf{b}_n)] \end{array}\right] = \left[\begin{array}{rrr} 1 & 0 & 1 \\ -1 & 1 & -1 \end{array}\right]$$

If $\mathbf{v} = a + bx + cx^2 = a\mathbf{b}_1 + b\mathbf{b}_2 + c\mathbf{b}_3$, then $T(\mathbf{v}) = (a + c)\mathbf{d}_1 + (b - a - c)\mathbf{d}_2$, so

$$C_D[T(\mathbf{v})] = \left[\begin{array}{c} a + c \\ b - a - c \end{array}\right] = \left[\begin{array}{rrr} 1 & 0 & 1 \\ -1 & 1 & -1 \end{array}\right] \left[\begin{array}{c} a \\ b \\ c \end{array}\right] = M_{DB}(T)C_B(\mathbf{v})$$

as Theorem 9.1.2 asserts.

The next example shows how to determine the action of a transformation from its matrix.

Example 9.1.3

Suppose $T : \mathbf{M}_{22}(\mathbb{R}) \to \mathbb{R}^3$ is linear with matrix $M_{DB}(T) = \left[\begin{array}{rrrr} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{array}\right]$ where

$$B = \left\{ \left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array}\right], \left[\begin{array}{cc} 0 & 1 \\ 0 & 0 \end{array}\right], \left[\begin{array}{cc} 0 & 0 \\ 1 & 0 \end{array}\right], \left[\begin{array}{cc} 0 & 0 \\ 0 & 1 \end{array}\right] \right\} \text{ and } D = \{(1,\ 0,\ 0),\ (0,\ 1,\ 0),\ (0,\ 0,\ 1)\}$$

Compute $T(\mathbf{v})$ where $\mathbf{v} = \left[\begin{array}{cc} a & b \\ c & d \end{array}\right]$.

Solution. The idea is to compute $C_D[T(\mathbf{v})]$ first, and then obtain $T(\mathbf{v})$. We have

$$C_D[T(\mathbf{v})] = M_{DB}(T)C_B(\mathbf{v}) = \left[\begin{array}{rrrr} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{array}\right] \left[\begin{array}{c} a \\ b \\ c \\ d \end{array}\right] = \left[\begin{array}{c} a-b \\ b-c \\ c-d \end{array}\right]$$

$$\begin{aligned} \text{Hence } T(\mathbf{v}) &= (a-b)(1,\ 0,\ 0) + (b-c)(0,\ 1,\ 0) + (c-d)(0,\ 0,\ 1) \\ &= (a-b,\ b-c,\ c-d) \end{aligned}$$

The next two examples will be referred to later.

Example 9.1.4

Let A be an $m \times n$ matrix, and let $T_A : \mathbb{R}^n \to \mathbb{R}^m$ be the matrix transformation induced by $A : T_A(\mathbf{x}) = A\mathbf{x}$ for all columns $\mathbf{x}$ in $\mathbb{R}^n$. If B and D are the standard bases of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively (ordered as usual), then

$$M_{DB}(T_A) = A$$

In other words, the matrix of T_A corresponding to the standard bases is A itself.

Solution. Write $B = \{\mathbf{e}_1,\ \dots,\ \mathbf{e}_n\}$. Because D is the standard basis of $\mathbb{R}^m$, it is easy to verify that $C_D(\mathbf{y}) = \mathbf{y}$ for all columns $\mathbf{y}$ in $\mathbb{R}^m$. Hence

$$M_{DB}(T_A) = \left[\begin{array}{cccc} T_A(\mathbf{e}_1) & T_A(\mathbf{e}_2) & \cdots & T_A(\mathbf{e}_n) \end{array}\right] = \left[\begin{array}{cccc} A\mathbf{e}_1 & A\mathbf{e}_2 & \cdots & A\mathbf{e}_n \end{array}\right] = A$$

because $A\mathbf{e}_j$ is the jth column of A.

Example 9.1.5

Let V and W have ordered bases B and D, respectively. Let $\dim V = n$.

1. The identity transformation $1_V : V \to V$ has matrix $M_{BB}(1_V) = I_n$.
2. The zero transformation $0 : V \to W$ has matrix $M_{DB}(0) = 0$.

The first result in Example 9.1.5 is false if the two bases of V are not equal. In fact, if B is the standard basis of $\mathbb{R}^n$, then the basis D of $\mathbb{R}^n$ can be chosen so that $M_{DB}(1_{\mathbb{R}^n})$ turns out to be any invertible matrix we wish (Exercise **??**).

The next two theorems show that composition of linear transformations is compatible with multiplication of the corresponding matrices.

Theorem 9.1.3

$$\begin{tikzcd} V \arrow[r, "T"] \arrow[rr, bend right, "ST"'] & W \arrow[r, "S"] & U \end{tikzcd}$$

Let $V \xrightarrow{T} W \xrightarrow{S} U$ be linear transformations and let B, D, and E be finite ordered bases of V, W, and U, respectively. Then

$$M_{EB}(ST) = M_{ED}(S) \cdot M_{DB}(T)$$

Proof. We use the property in Theorem 9.1.2 three times. If $\mathbf{v}$ is in V,

$$M_{ED}(S)M_{DB}(T)C_B(\mathbf{v}) = M_{ED}(S)C_D[T(\mathbf{v})] = C_E[ST(\mathbf{v})] = M_{EB}(ST)C_B(\mathbf{v})$$

If $B = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, then $C_B(\mathbf{e}_j)$ is column j of I_n. Hence taking $\mathbf{v} = \mathbf{e}_j$ shows that $M_{ED}(S)M_{DB}(T)$ and $M_{EB}(ST)$ have equal jth columns. The theorem follows. □

Theorem 9.1.4

Let $T : V \to W$ be a linear transformation, where $\dim V = \dim W = n$. The following are equivalent.

1. *T is an isomorphism.*
2. *$M_{DB}(T)$ is invertible for all ordered bases B and D of V and W.*
3. *$M_{DB}(T)$ is invertible for some pair of ordered bases B and D of V and W.*

When this is the case, $[M_{DB}(T)]^{-1} = M_{BD}(T^{-1})$.

Proof. (1) $\Rightarrow$ (2). We have $V \xrightarrow{T} W \xrightarrow{T^{-1}} V$, so Theorem 9.1.3 and Example 9.1.5 give

$$M_{BD}(T^{-1})M_{DB}(T) = M_{BB}(T^{-1}T) = M_{BB}(1v) = I_n$$

Similarly, $M_{DB}(T)M_{BD}(T^{-1}) = I_n$, proving (2) (and the last statement in the theorem).

(2) $\Rightarrow$ (3). This is clear.

$$\begin{tikzcd} \mathbb{R}^n \arrow[r, "T_{A^{-1}}"] \arrow[rr, bend right, "T_A T_{A^{-1}}"'] & \mathbb{R}^n \arrow[r, "T_A"] & \mathbb{R}^n \end{tikzcd}$$

(3) $\Rightarrow$ (1). Suppose that $T_{DB}(T)$ is invertible for some bases B and D and, for convenience, write $A = M_{DB}(T)$. Then we have $C_D T = T_A C_B$ by Theorem 9.1.2, so

$$T = (C_D)^{-1} T_A C_B$$

by Theorem 9.1.1 where $(C_D)^{-1}$ and C_B are isomorphisms. Hence (1) follows if we can demonstrate that $T_A : \mathbb{R}^n \to \mathbb{R}^n$ is also an isomorphism. But A is invertible by (3) and one verifies that $T_A T_{A^{-1}} = 1_{\mathbb{R}^n} = T_{A^{-1}} T_A$. So T_A is indeed invertible (and $(T_A)^{-1} = T_{A^{-1}}$). □

In Section 7.2 we defined the rank of a linear transformation $T : V \to W$ by $\operatorname{rank} T = \dim(\operatorname{im} T)$. Moreover, if A is any $m \times n$ matrix and $T_A : \mathbb{R}^n \to \mathbb{R}^m$ is the matrix transformation, we showed that $\operatorname{rank}(T_A) = \operatorname{rank} A$. So it may not be surprising that $\operatorname{rank} T$ equals the rank of any matrix of T.

Theorem 9.1.5

Let $T : V \to W$ be a linear transformation where $\dim V = n$ and $\dim W = m$. If B and D are any ordered bases of V and W, then $\operatorname{rank} T = \operatorname{rank}[M_{DB}(T)]$.

Proof. Write $A = M_{DB}(T)$ for convenience. The column space of A is $U = \{A\mathbf{x} \mid \mathbf{x} \text{ in } \mathbb{R}^n\}$. This means $\operatorname{rank} A = \dim U$ and so, because $\operatorname{rank} T = \dim(\operatorname{im} T)$, it suffices to find an isomorphism $S : \operatorname{im} T \to U$. Now every vector in $\operatorname{im} T$ has the form $T(\mathbf{v})$, $\mathbf{v}$ in V. By Theorem 9.1.2, $C_D[T(\mathbf{v})] = AC_B(\mathbf{v})$ lies in U. So define $S : \operatorname{im} T \to U$ by

$$S[T(\mathbf{v})] = C_D[T(\mathbf{v})] \text{ for all vectors } T(\mathbf{v}) \in \operatorname{im} T$$

The fact that C_D is linear and one-to-one implies immediately that S is linear and one-to-one. To see that S is onto, let $A\mathbf{x}$ be any member of U, $\mathbf{x}$ in $\mathbb{R}^n$. Then $\mathbf{x} = C_B(\mathbf{v})$ for some $\mathbf{v}$ in V because C_B is onto. Hence $A\mathbf{x} = AC_B(\mathbf{v}) = C_D[T(\mathbf{v})] = S[T(\mathbf{v})]$, so S is onto. This means that S is an isomorphism. □

Example 9.1.6

Define $T : \mathbf{P}_2 \to \mathbb{R}^3$ by $T(a+bx+cx^2) = (a-2b,\ 3c-2a,\ 3c-4b)$ for $a, b, c \in \mathbb{R}$. Compute $\operatorname{rank} T$.

Solution. Since $\operatorname{rank} T = \operatorname{rank}[M_{DB}(T)]$ for *any* bases $B \subseteq \mathbf{P}_2$ and $D \subseteq \mathbb{R}^3$, we choose the most convenient ones: $B = \{1,\ x,\ x^2\}$ and $D = \{(1,\ 0,\ 0),\ (0,\ 1,\ 0),\ (0,\ 0,\ 1)\}$. Then $M_{DB}(T) = \left[\begin{array}{ccc} C_D[T(1)] & C_D[T(x)] & C_D[T(x^2)] \end{array}\right] = A$ where

$$A = \left[\begin{array}{rrr} 1 & -2 & 0 \\ -2 & 0 & 3 \\ 0 & -4 & 3 \end{array}\right]. \quad \text{Since } A \to \left[\begin{array}{rrr} 1 & -2 & 0 \\ 0 & -4 & 3 \\ 0 & -4 & 3 \end{array}\right] \to \left[\begin{array}{rrr} 1 & -2 & 0 \\ 0 & 1 & -\frac{3}{4} \\ 0 & 0 & 0 \end{array}\right]$$

we have $\operatorname{rank} A = 2$. Hence $\operatorname{rank} T = 2$ as well.

We conclude with an example showing that the matrix of a linear transformation can be made very simple by a careful choice of the two bases.

Example 9.1.7

Let $T : V \to W$ be a linear transformation where $\dim V = n$ and $\dim W = m$. Choose an ordered basis $B = \{\mathbf{b}_1,\ \dots,\ \mathbf{b}_r,\ \mathbf{b}_{r+1},\ \dots,\ \mathbf{b}_n\}$ of V in which $\{\mathbf{b}_{r+1},\ \dots,\ \mathbf{b}_n\}$ is a basis of $\ker T$, possibly empty. Then $\{T(\mathbf{b}_1),\ \dots,\ T(\mathbf{b}_r)\}$ is a basis of $\operatorname{im} T$ by Theorem 7.2.5, so extend it to an ordered basis $D = \{T(\mathbf{b}_1),\ \dots,\ T(\mathbf{b}_r),\ \mathbf{f}_{r+1},\ \dots,\ \mathbf{f}_m\}$ of W. Because $T(\mathbf{b}_{r+1}) = \cdots = T(\mathbf{b}_n) = \mathbf{0}$, we have

$$M_{DB}(T) = \left[\begin{array}{ccccccc} C_D[T(\mathbf{b}_1)] & \cdots & C_D[T(\mathbf{b}_r)] & C_D[T(\mathbf{b}_{r+1})] & \cdots & C_D[T(\mathbf{b}_n)] \end{array}\right] = \left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right]$$

Incidentally, this shows that $\operatorname{rank} T = r$ by Theorem 9.1.5.

9.2 Operators and Similarity

While the study of linear transformations from one vector space to another is important, the central problem of linear algebra is to understand the structure of a linear transformation $T : V \to V$ from a space V to itself. Such transformations are called **linear operators**. If $T : V \to V$ is a linear operator where $\dim(V) = n$, it is possible to choose bases B and D of V such that the matrix $M_{DB}(T)$ has a very simple form: $M_{DB}(T) = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}$ where $r = \text{rank } T$ (see Example 9.1.7). Consequently, only the rank of T is revealed by determining the simplest matrices $M_{DB}(T)$ of T where the bases B and D can be chosen arbitrarily. But if we insist that $B = D$ and look for bases B such that $M_{BB}(T)$ is as simple as possible, we learn a great deal about the operator T. We begin this task in this section.

The B-matrix of an Operator

Definition 9.3 Matrix $M_{DB}(T)$ of $T : V \to W$ for basis B

*If $T : V \to V$ is an operator on a vector space V, and if B is an ordered basis of V, define $M_B(T) = M_{BB}(T)$ and call this the **B-matrix** of T.*

Recall that if $T : \mathbb{R}^n \to \mathbb{R}^n$ is a linear operator and $E = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ is the standard basis of $\mathbb{R}^n$, then $C_E(\mathbf{x}) = \mathbf{x}$ for every $\mathbf{x} \in \mathbb{R}^n$, so $M_E(T) = [T(\mathbf{e}_1), T(\mathbf{e}_2), \dots, T(\mathbf{e}_n)]$ is the matrix obtained in Theorem 2.6.2. Hence $M_E(T)$ will be called the **standard matrix** of the operator T.

For reference the following theorem collects some results from Theorem 9.1.2, Theorem 9.1.3, and Theorem 9.1.4, specialized for operators. As before, $C_B(\mathbf{v})$ denoted the coordinate vector of $\mathbf{v}$ with respect to the basis B.

Theorem 9.2.1

Let $T : V \to V$ be an operator where $\dim V = n$, and let B be an ordered basis of V.

1. $C_B(T(\mathbf{v})) = M_B(T)C_B(\mathbf{v})$ *for all* $\mathbf{v}$ *in* V.
2. *If* $S : V \to V$ *is another operator on* V, *then* $M_B(ST) = M_B(S)M_B(T)$.
3. T *is an isomorphism if and only if* $M_B(T)$ *is invertible. In this case* $M_D(T)$ *is invertible for every ordered basis* D *of* V.
4. *If* T *is an isomorphism, then* $M_B(T^{-1}) = [M_B(T)]^{-1}$.
5. *If* $B = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$, *then* $M_B(T) = \left[\begin{array}{cccc} C_B[T(\mathbf{b}_1)] & C_B[T(\mathbf{b}_2)] & \cdots & C_B[T(\mathbf{b}_n)] \end{array}\right]$.

For a fixed operator T on a vector space V, we are going to study how the matrix $M_B(T)$ changes when the basis B changes. This turns out to be closely related to how the coordinates $C_B(\mathbf{v})$ change for a vector $\mathbf{v}$ in V. If B and D are two ordered bases of V, and if we take $T = 1_V$ in Theorem 9.1.2, we obtain

$$C_D(\mathbf{v}) = M_{DB}(1_V)C_B(\mathbf{v}) \qquad \text{for all } \mathbf{v} \text{ in } V$$

Definition 9.4 Change Matrix $P_{D \leftarrow B}$ for bases B and D

With this in mind, define the ***change matrix*** $P_{D \leftarrow B}$ *by*

$$P_{D \leftarrow B} = M_{DB}(1_V) \qquad \textit{for any ordered bases } B \textit{ and } D \textit{ of } V$$

This proves equation 9.2 in the following theorem:

Theorem 9.2.2

Let $B = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ and D denote ordered bases of a vector space V. Then the change matrix $P_{D \leftarrow B}$ is given in terms of its columns by

$$P_{D \leftarrow B} = \left[\begin{array}{cccc} C_D(\mathbf{b}_1) & C_D(\mathbf{b}_2) & \cdots & C_D(\mathbf{b}_n) \end{array} \right] \tag{9.1}$$

and has the property that

$$C_D(\mathbf{v}) = P_{D \leftarrow B} C_B(\mathbf{v}) \text{ for all } \mathbf{v} \text{ in } V \tag{9.2}$$

Moreover, if E is another ordered basis of V, we have

1. $P_{B \leftarrow B} = I_n$
2. $P_{D \leftarrow B}$ *is invertible and* $(P_{D \leftarrow B})^{-1} = P_{B \leftarrow D}$
3. $P_{E \leftarrow D} P_{D \leftarrow B} = P_{E \leftarrow B}$

Proof. The formula 9.2 is derived above, and 9.1 is immediate from the definition of $P_{D \leftarrow B}$ and the formula for $M_{DB}(T)$ in Theorem 9.1.2.

1. $P_{B \leftarrow B} = M_{BB}(1_V) = I_n$ as is easily verified.
2. This follows from (1) and (3).
3. Let $V \xrightarrow{T} W \xrightarrow{S} U$ be operators, and let B, D, and E be ordered bases of V, W, and U respectively. We have $M_{EB}(ST) = M_{ED}(S) M_{DB}(T)$ by Theorem 9.1.3. Now (3) is the result of specializing $V = W = U$ and $T = S = 1_V$. □

Property (3) in Theorem 9.2.2 explains the notation $\mathbf{P}_{D \leftarrow B}$.

Example 9.2.1

In $\mathbf{P}_2$ find $P_{D \leftarrow B}$ if $B = \{1, x, x^2\}$ and $D = \{1, (1-x), (1-x)^2\}$. Then use this to express $p = p(x) = a + bx + cx^2$ as a polynomial in powers of $(1-x)$.

Solution. To compute the change matrix $P_{D \leftarrow B}$, express $1, x, x^2$ in the basis D:

$$\begin{aligned} 1 &= 1 + 0(1-x) + 0(1-x)^2 \\ x &= 1 - 1(1-x) + 0(1-x)^2 \\ x^2 &= 1 - 2(1-x) + 1(1-x)^2 \end{aligned}$$

Hence $P_{D\leftarrow B} = \left[C_D(1),\, C_D(x),\, C_D(x)^2\right] = \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 1 \end{bmatrix}$. We have $C_B(p) = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$, so

$$C_D(p) = P_{D\leftarrow B}C_B(p) = \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} a+b+c \\ -b-2c \\ c \end{bmatrix}$$

Hence $p(x) = (a+b+c) - (b+2c)(1-x) + c(1-x)^2$ by Definition 9.1.[1]

Now let $B = \{\mathbf{b}_1,\, \mathbf{b}_2,\, \dots,\, \mathbf{b}_n\}$ and B_0 be two ordered bases of a vector space V. An operator $T : V \to V$ has different matrices $M_B[T]$ and $M_{B_0}[T]$ with respect to B and B_0. We can now determine how these matrices are related. Theorem 9.2.2 asserts that

$$C_{B_0}(\mathbf{v}) = P_{B_0\leftarrow B}C_B(\mathbf{v}) \text{ for all } \mathbf{v} \text{ in } V$$

On the other hand, Theorem 9.2.1 gives

$$C_B[T(\mathbf{v})] = M_B(T)C_B(\mathbf{v}) \text{ for all } \mathbf{v} \text{ in } V$$

Combining these (and writing $P = P_{B_0\leftarrow B}$ for convenience) gives

$$\begin{aligned} PM_B(T)C_B(\mathbf{v}) &= PC_B[T(\mathbf{v})] \\ &= C_{B_0}[T(\mathbf{v})] \\ &= M_{B_0}(T)C_{B_0}(\mathbf{v}) \\ &= M_{B_0}(T)PC_B(\mathbf{v}) \end{aligned}$$

This holds for all $\mathbf{v}$ in V. Because $C_B(\mathbf{b}_j)$ is the jth column of the identity matrix, it follows that

$$PM_B(T) = M_{B_0}(T)P$$

Moreover P is invertible (in fact, $P^{-1} = P_{B\leftarrow B_0}$ by Theorem 9.2.2), so this gives

$$M_B(T) = P^{-1}M_{B_0}(T)P$$

This asserts that $M_{B_0}(T)$ and $M_B(T)$ are similar matrices, and proves Theorem 9.2.3.

Theorem 9.2.3: Similarity Theorem

Let B_0 and B be two ordered bases of a finite dimensional vector space V. If $T : V \to V$ is any linear operator, the matrices $M_B(T)$ and $M_{B_0}(T)$ of T with respect to these bases are similar. More precisely,

$$M_B(T) = P^{-1}M_{B_0}(T)P$$

where $P = P_{B_0\leftarrow B}$ is the change matrix from B to B_0.

[1] This also follows from Taylor's theorem (Corollary 6.5.3 of Theorem 6.5.1 with $a = 1$).

Example 9.2.2

Let $T : \mathbb{R}^3 \to \mathbb{R}^3$ be defined by $T(a,\, b,\, c) = (2a-b,\, b+c,\, c-3a)$. If B_0 denotes the standard basis of $\mathbb{R}^3$ and $B = \{(1,\, 1,\, 0),\, (1,\, 0,\, 1),\, (0,\, 1,\, 0)\}$, find an invertible matrix P such that $P^{-1}M_{B_0}(T)P = M_B(T)$.

Solution. We have

$$M_{B_0}(T) = \left[\begin{array}{ccc} C_{B_0}(2,\, 0,\, -3) & C_{B_0}(-1,\, 1,\, 0) & C_{B_0}(0,\, 1,\, 1) \end{array}\right] = \left[\begin{array}{rrr} 2 & -1 & 0 \\ 0 & 1 & 1 \\ -3 & 0 & 1 \end{array}\right]$$

$$M_B(T) = \left[\begin{array}{ccc} C_B(1,\, 1,\, -3) & C_B(2,\, 1,\, -2) & C_B(-1,\, 1,\, 0) \end{array}\right] = \left[\begin{array}{rrr} 4 & 4 & -1 \\ -3 & -2 & 0 \\ -3 & -3 & 2 \end{array}\right]$$

$$P = P_{B_0 \leftarrow B} = \left[\begin{array}{ccc} C_{B_0}(1,\, 1,\, 0) & C_{B_0}(1,\, 0,\, 1) & C_{B_0}(0,\, 1,\, 0) \end{array}\right] = \left[\begin{array}{rrr} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{array}\right]$$

The reader can verify that $P^{-1}M_{B_0}(T)P = M_B(T)$; equivalently that $M_{B_0}(T)P = PM_B(T)$.

A square matrix is diagonalizable if and only if it is similar to a diagonal matrix. Theorem 9.2.3 comes into this as follows: Suppose an $n \times n$ matrix $A = M_{B_0}(T)$ is the matrix of some operator $T : V \to V$ with respect to an ordered basis B_0. If another ordered basis B of V can be found such that $M_B(T) = D$ is diagonal, then Theorem 9.2.3 shows how to find an invertible P such that $P^{-1}AP = D$. In other words, the "algebraic" problem of finding P such that $P^{-1}AP$ is diagonal comes down to the "geometric" problem of finding a basis B such that $M_B(T)$ is diagonal. This shift of emphasis is one of the most important techniques in linear algebra.

Each $n \times n$ matrix A can be easily realized as the matrix of an operator. In fact, (Example 9.1.4),

$$M_E(T_A) = A$$

where $T_A : \mathbb{R}^n \to \mathbb{R}^n$ is the matrix operator given by $T_A(\mathbf{x}) = A\mathbf{x}$, and E is the standard basis of $\mathbb{R}^n$. The first part of the next theorem gives the converse of Theorem 9.2.3: Any pair of similar matrices can be realized as the matrices of the same linear operator with respect to different bases. This is part 1 of the following theorem.

Theorem 9.2.4

Let A be an $n \times n$ matrix and let E be the standard basis of $\mathbb{R}^n$.

1. *Let A' be similar to A, say $A' = P^{-1}AP$, and let B be the ordered basis of $\mathbb{R}^n$ consisting of the columns of P in order. Then $T_A : \mathbb{R}^n \to \mathbb{R}^n$ is linear and*

$$M_E(T_A) = A \quad \text{and} \quad M_B(T_A) = A'$$

2. *If B is any ordered basis of $\mathbb{R}^n$, let P be the (invertible) matrix whose columns are the vectors in B in order. Then*

$$M_B(T_A) = P^{-1}AP$$

Proof.

1. We have $M_E(T_A) = A$ by Example 9.1.4. Write $P = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_n \end{bmatrix}$ in terms of its columns so $B = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ is a basis of $\mathbb{R}^n$. Since E is the standard basis,
$$P_{E \leftarrow B} = \begin{bmatrix} C_E(\mathbf{b}_1) & \cdots & C_E(\mathbf{b}_n) \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_n \end{bmatrix} = P$$
Hence Theorem 9.2.3 (with $B_0 = E$) gives $M_B(T_A) = P^{-1}M_E(T_A)P = P^{-1}AP = A'$.

2. Here P and B are as above, so again $P_{E \leftarrow B} = P$ and $M_B(T_A) = P^{-1}AP$. □

Example 9.2.3

Given $A = \begin{bmatrix} 10 & 6 \\ -18 & -11 \end{bmatrix}$, $P = \begin{bmatrix} 2 & -1 \\ -3 & 2 \end{bmatrix}$, and $D = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}$, verify that $P^{-1}AP = D$ and use this fact to find a basis B of $\mathbb{R}^2$ such that $M_B(T_A) = D$.

Solution. $P^{-1}AP = D$ holds if $AP = PD$; this verification is left to the reader. Let B consist of the columns of P in order, that is $B = \left\{ \begin{bmatrix} 2 \\ -3 \end{bmatrix}, \begin{bmatrix} -1 \\ 2 \end{bmatrix} \right\}$. Then Theorem 9.2.4 gives $M_B(T_A) = P^{-1}AP = D$. More explicitly,
$$M_B(T_A) = \left[C_B\left(T_A \begin{bmatrix} 2 \\ -3 \end{bmatrix}\right) \quad C_B\left(T_A \begin{bmatrix} -1 \\ 2 \end{bmatrix}\right) \right] = \left[C_B \begin{bmatrix} 2 \\ -3 \end{bmatrix} \quad C_B \begin{bmatrix} 2 \\ -4 \end{bmatrix} \right] = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix} = D$$

Let A be an $n \times n$ matrix. As in Example 9.2.3, Theorem 9.2.4 provides a new way to find an invertible matrix P such that $P^{-1}AP$ is diagonal. The idea is to find a basis $B = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ of $\mathbb{R}^n$ such that $M_B(T_A) = D$ is diagonal and take $P = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_n \end{bmatrix}$ to be the matrix with the $\mathbf{b}_j$ as columns. Then, by Theorem 9.2.4,
$$P^{-1}AP = M_B(T_A) = D$$
As mentioned above, this converts the algebraic problem of diagonalizing A into the geometric problem of finding the basis B. This new point of view is very powerful and will be explored in the next two sections.

Theorem 9.2.4 enables facts about matrices to be deduced from the corresponding properties of operators. Here is an example.

Example 9.2.4

1. If $T : V \to V$ is an operator where V is finite dimensional, show that $TST = T$ for some invertible operator $S : V \to V$.

2. If A is an $n \times n$ matrix, show that $AUA = A$ for some invertible matrix U.

Solution.

1. Let $B = \{\mathbf{b}_1, \dots, \mathbf{b}_r, \mathbf{b}_{r+1}, \dots, \mathbf{b}_n\}$ be a basis of V chosen so that $\ker T = \operatorname{span}\{\mathbf{b}_{r+1}, \dots, \mathbf{b}_n\}$. Then $\{T(\mathbf{b}_1), \dots, T(\mathbf{b}_r)\}$ is independent (Theorem 7.2.5), so complete it to a basis $\{T(\mathbf{b}_1), \dots, T(\mathbf{b}_r), \mathbf{f}_{r+1}, \dots, \mathbf{f}_n\}$ of V.

By Theorem 7.1.3, define $S : V \to V$ by

$$S[T(\mathbf{b}_i)] = \mathbf{b}_i \quad \text{for } 1 \leq i \leq r$$
$$S(\mathbf{f}_j) = \mathbf{b}_j \quad \text{for } r < j \leq n$$

Then S is an isomorphism by Theorem 7.3.1, and $TST = T$ because these operators agree on the basis B. In fact,

$$(TST)(\mathbf{b}_i) = T[ST(\mathbf{b}_i)] = T(\mathbf{b}_i) \text{ if } 1 \leq i \leq r, \text{ and}$$
$$(TST)(\mathbf{b}_j) = TS[T(\mathbf{b}_j)] = TS(\mathbf{0}) = \mathbf{0} = T(\mathbf{b}_j) \text{ for } r < j \leq n$$

2. Given A, let $T = T_A : \mathbb{R}^n \to \mathbb{R}^n$. By (1) let $TST = T$ where $S : \mathbb{R}^n \to \mathbb{R}^n$ is an isomorphism. If E is the standard basis of $\mathbb{R}^n$, then $A = M_E(T)$ by Theorem 9.2.4. If $U = M_E(S)$ then, by Theorem 9.2.1, U is invertible and

$$AUA = M_E(T)M_E(S)M_E(T) = M_E(TST) = M_E(T) = A$$

as required.

The reader will appreciate the power of these methods if he/she tries to find U directly in part 2 of Example 9.2.4, even if A is 2×2.

A property of $n \times n$ matrices is called a **similarity invariant** if, whenever a given $n \times n$ matrix A has the property, every matrix similar to A also has the property. Theorem 5.5.1 shows that rank, determinant, trace, and characteristic polynomial are all similarity invariants.

To illustrate how such similarity invariants are related to linear operators, consider the case of rank. If $T : V \to V$ is a linear operator, the matrices of T with respect to various bases of V all have the same rank (being similar), so it is natural to regard the common rank of all these matrices as a property of T itself and not of the particular matrix used to describe T. Hence the rank of T could be *defined* to be the rank of A, where A is *any* matrix of T. This would be unambiguous because rank is a similarity invariant. Of course, this is unnecessary in the case of rank because rank T was defined earlier to be the dimension of im T, and this was *proved* to equal the rank of every matrix representing T (Theorem 9.1.5). This definition of rank T is said to be *intrinsic* because it makes no reference to the matrices representing T. However, the technique serves to identify an intrinsic property of T with *every* similarity invariant, and some of these properties are not so easily defined directly.

In particular, if $T : V \to V$ is a linear operator on a finite dimensional space V, define the **determinant** of T (denoted $\det T$) by

$$\det T = \det M_B(T), \quad B \text{ any basis of } V$$

This is independent of the choice of basis B because, if D is any other basis of V, the matrices $M_B(T)$ and $M_D(T)$ are similar and so have the same determinant. In the same way, the **trace** of T (denoted $\operatorname{tr} T$) can be defined by

$$\operatorname{tr} T = \operatorname{tr} M_B(T), \quad B \text{ any basis of } V$$

This is unambiguous for the same reason.

Theorems about matrices can often be translated to theorems about linear operators. Here is an example.

Example 9.2.5

Let S and T denote linear operators on the finite dimensional space V. Show that

$$\det(ST) = \det S \det T$$

Solution. Choose a basis B of V and use Theorem 9.2.1.

$$\begin{aligned}\det(ST) = \det M_B(ST) &= \det[M_B(S)M_B(T)] \\ &= \det[M_B(S)]\det[M_B(T)] = \det S \det T\end{aligned}$$

Recall next that the characteristic polynomial of a matrix is another similarity invariant: If A and A' are similar matrices, then $c_A(x) = c_{A'}(x)$ (Theorem 5.5.1). As discussed above, the discovery of a similarity invariant means the discovery of a property of linear operators. In this case, if $T : V \to V$ is a linear operator on the finite dimensional space V, define the **characteristic polynomial** of T by

$$c_T(x) = c_A(x) \text{ where } A = M_B(T),\ B \text{ any basis of } V$$

In other words, the characteristic polynomial of an operator T is the characteristic polynomial of *any* matrix representing T. This is unambiguous because any two such matrices are similar by Theorem 9.2.3.

Example 9.2.6

Compute the characteristic polynomial $c_T(x)$ of the operator $T : \mathbf{P}_2 \to \mathbf{P}_2$ given by $T(a+bx+cx^2) = (b+c)+(a+c)x+(a+b)x^2$.

Solution. If $B = \{1, x, x^2\}$, the corresponding matrix of T is

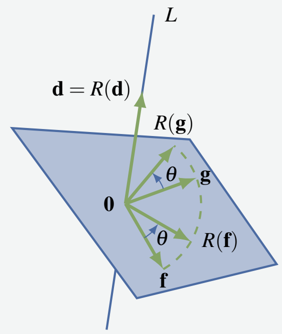
$$M_B(T) = \left[\begin{array}{ccc} C_B[T(1)] & C_B[T(x)] & C_B[T(x^2)] \end{array}\right] = \left[\begin{array}{ccc} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{array}\right]$$

Hence $c_T(x) = \det[xI - M_B(T)] = x^3 - 3x - 2 = (x+1)^2(x-2)$.

In Section 4.4 we computed the matrix of various projections, reflections, and rotations in $\mathbb{R}^3$. However, the methods available then were not adequate to find the matrix of a rotation about a line through the origin. We conclude this section with an example of how Theorem 9.2.3 can be used to compute such a matrix.

Example 9.2.7

Let L be the line in $\mathbb{R}^3$ through the origin with (unit) direction vector $\mathbf{d} = \frac{1}{3}\left[\begin{array}{ccc} 2 & 1 & 2 \end{array}\right]^T$. Compute the matrix of the rotation about L through an angle θ measured counterclockwise when viewed in the direction of $\mathbf{d}$.



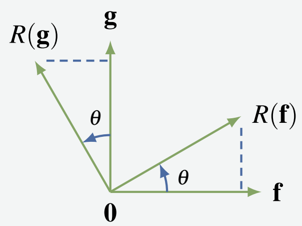
Solution. Let $R : \mathbb{R}^3 \to \mathbb{R}^3$ be the rotation. The idea is to first find a basis B_0 for which the matrix of $M_{B_0}(R)$ of R is easy to compute, and then use Theorem 9.2.3 to compute the "standard" matrix $M_E(R)$ with respect to the standard basis $E = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ of $\mathbb{R}^3$.

To construct the basis B_0, let K denote the plane through the origin with $\mathbf{d}$ as normal, shaded in the diagram. Then the vectors $\mathbf{f} = \frac{1}{3}\left[\begin{array}{ccc} -2 & 2 & 1 \end{array}\right]^T$ and $\mathbf{g} = \frac{1}{3}\left[\begin{array}{ccc} 1 & 2 & -2 \end{array}\right]^T$ are both in K (they are orthogonal to $\mathbf{d}$) and are independent (they are orthogonal to each other).

Hence $B_0 = \{\mathbf{d}, \mathbf{f}, \mathbf{g}\}$ is an orthonormal basis of $\mathbb{R}^3$, and the effect of R on B_0 is easy to determine. In fact $R(\mathbf{d}) = \mathbf{d}$ and (as in Theorem 2.6.4) the second diagram gives

$$R(\mathbf{f}) = \cos\theta\mathbf{f} + \sin\theta\mathbf{g} \quad \text{and} \quad R(\mathbf{g}) = -\sin\theta\mathbf{f} + \cos\theta\mathbf{g}$$

because $\|\mathbf{f}\| = 1 = \|\mathbf{g}\|$. Hence



$$M_{B_0}(R) = \left[\begin{array}{ccc} C_{B_0}(\mathbf{d}) & C_{B_0}(\mathbf{f}) & C_{B_0}(\mathbf{g}) \end{array}\right] = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{array}\right]$$

Now Theorem 9.2.3 (with $B = E$) asserts that $M_E(R) = P^{-1}M_{B_0}(R)P$ where

$$P = P_{B_0 \leftarrow E} = \left[\begin{array}{ccc} C_{B_0}(\mathbf{e}_1) & C_{B_0}(\mathbf{e}_2) & C_{B_0}(\mathbf{e}_3) \end{array}\right] = \frac{1}{3}\left[\begin{array}{rrr} 2 & 1 & 2 \\ -2 & 2 & 1 \\ 1 & 2 & -2 \end{array}\right]$$

using the expansion theorem (Theorem 5.3.6). Since $P^{-1} = P^T$ (P is orthogonal), the matrix of R with respect to E is

$$\begin{aligned} M_E(R) &= P^T M_{B_0}(R)P \\ &= \frac{1}{9}\left[\begin{array}{ccc} 5\cos\theta + 4 & 6\sin\theta - 2\cos\theta + 2 & 4 - 3\sin\theta - 4\cos\theta \\ 2 - 6\sin\theta - 2\cos\theta & 8\cos\theta + 1 & 6\sin\theta - 2\cos\theta + 2 \\ 3\sin\theta - 4\cos\theta + 4 & 2 - 6\sin\theta - 2\cos\theta & 5\cos\theta + 4 \end{array}\right] \end{aligned}$$

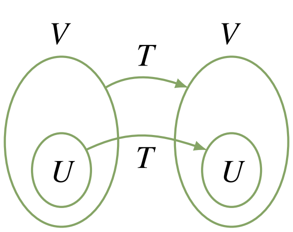
As a check one verifies that this is the identity matrix when $\theta = 0$, as it should.

Note that in Example 9.2.7 not much motivation was given to the choices of the (orthonormal) vectors $\mathbf{f}$ and $\mathbf{g}$ in the basis B_0, which is the key to the solution. However, if we begin with *any* basis containing $\mathbf{d}$ the Gram-Schmidt algorithm will produce an orthogonal basis containing $\mathbf{d}$, and the other two vectors will automatically be in $L^\perp = K$.

9.3 Invariant Subspaces and Direct Sums

A fundamental question in linear algebra is the following: If $T : V \to V$ is a linear operator, how can a basis B of V be chosen so the matrix $M_B(T)$ is as simple as possible? A basic technique for answering such questions will be explained in this section. If U is a subspace of V, write its image under T as

$$T(U) = \{T(\mathbf{u}) \mid \mathbf{u} \text{ in } U\}$$



Definition 9.5 T-invariant Subspace

*Let $T : V \to V$ be an operator. A subspace $U \subseteq V$ is called **T-invariant** if $T(U) \subseteq U$, that is, $T(\mathbf{u}) \in U$ for every vector $\mathbf{u} \in U$. Hence T is a linear operator on the vector space U.*

This is illustrated in the diagram, and the fact that $T : U \to U$ is an operator on U is the primary reason for our interest in T-invariant subspaces.

Example 9.3.1

Let $T : V \to V$ be any linear operator. Then:

1. $\{\mathbf{0}\}$ and V are T-invariant subspaces.

2. Both $\ker T$ and $\operatorname{im} T = T(V)$ are T-invariant subspaces.

3. If U and W are T-invariant subspaces, so are $T(U)$, $U \cap W$, and $U + W$.

Solution. Item 1 is clear, and the rest is left as Exercises **??** and **??**.

Example 9.3.2

Define $T : \mathbb{R}^3 \to \mathbb{R}^3$ by $T(a,\ b,\ c) = (3a + 2b,\ b - c,\ 4a + 2b - c)$. Then $U = \{(a,\ b,\ a) \mid a,\ b \text{ in } \mathbb{R}\}$ is T-invariant because

$$T(a,\ b,\ a) = (3a + 2b,\ b - a,\ 3a + 2b)$$

is in U for all a and b (the first and last entries are equal).

If a spanning set for a subspace U is known, it is easy to check whether U is T-invariant.

Example 9.3.3

Let $T : V \to V$ be a linear operator, and suppose that $U = \operatorname{span}\{\mathbf{u}_1,\ \mathbf{u}_2,\ \dots,\ \mathbf{u}_k\}$ is a subspace of V. Show that U is T-invariant if and only if $T(\mathbf{u}_i)$ lies in U for each $i = 1,\ 2,\ \dots,\ k$.

Solution. Given $\mathbf{u}$ in U, write it as $\mathbf{u} = r_1\mathbf{u}_1 + \cdots + r_k\mathbf{u}_k$, r_i in $\mathbb{R}$. Then

$$T(\mathbf{u}) = r_1 T(\mathbf{u}_1) + \cdots + r_k T(\mathbf{u}_k)$$

and this lies in U if each $T(\mathbf{u}_i)$ lies in U. This shows that U is T-invariant if each $T(\mathbf{u}_i)$ lies in U; the converse is clear.

Example 9.3.4

Define $T : \mathbb{R}^2 \to \mathbb{R}^2$ by $T(a,\ b) = (b,\ -a)$. Show that $\mathbb{R}^2$ contains no T-invariant subspace except 0 and $\mathbb{R}^2$.

Solution. Suppose, if possible, that U is T-invariant, but $U \neq 0$, $U \neq \mathbb{R}^2$. Then U has dimension 1 so $U = \mathbb{R}\mathbf{x}$ where $\mathbf{x} \neq \mathbf{0}$. Now $T(\mathbf{x})$ lies in U—say $T(\mathbf{x}) = r\mathbf{x}$, r in $\mathbb{R}$. If we write $\mathbf{x} = (a,\ b)$, this is $(b,\ -a) = r(a,\ b)$, which gives $b = ra$ and $-a = rb$. Eliminating b gives $r^2 a = rb = -a$, so $(r^2 + 1)a = 0$. Hence $a = 0$. Then $b = ra = 0$ too, contrary to the assumption that $\mathbf{x} \neq \mathbf{0}$. Hence no one-dimensional T-invariant subspace exists.

Definition 9.6 Restriction of an Operator

Let $T : V \to V$ be a linear operator. If U is any T-invariant subspace of V, then

$$T : U \to U$$

is a linear operator on the subspace U, called the **restriction** *of T to U.*

This is the reason for the importance of T-invariant subspaces and is the first step toward finding a basis that simplifies the matrix of T.

Theorem 9.3.1

Let $T : V \to V$ be a linear operator where V has dimension n and suppose that U is any T-invariant subspace of V. Let $B_1 = \{\mathbf{b}_1, \dots, \mathbf{b}_k\}$ be any basis of U and extend it to a basis $B = \{\mathbf{b}_1, \dots, \mathbf{b}_k, \mathbf{b}_{k+1}, \dots, \mathbf{b}_n\}$ of V in any way. Then $M_B(T)$ has the block triangular form

$$M_B(T) = \left[\begin{array}{cc} M_{B_1}(T) & Y \\ 0 & Z \end{array} \right]$$

where Z is $(n-k) \times (n-k)$ and $M_{B_1}(T)$ is the matrix of the restriction of T to U.

Proof. The matrix of (the restriction) $T : U \to U$ with respect to the basis B_1 is the $k \times k$ matrix

$$M_{B_1}(T) = \left[\begin{array}{cccc} C_{B_1}[T(\mathbf{b}_1)] & C_{B_1}[T(\mathbf{b}_2)] & \cdots & C_{B_1}[T(\mathbf{b}_k)] \end{array} \right]$$

Now compare the first column $C_{B_1}[T(\mathbf{b}_1)]$ here with the first column $C_B[T(\mathbf{b}_1)]$ of $M_B(T)$. The fact that $T(\mathbf{b}_1)$ lies in U (because U is T-invariant) means that $T(\mathbf{b}_1)$ has the form

$$T(\mathbf{b}_1) = t_1\mathbf{b}_1 + t_2\mathbf{b}_2 + \cdots + t_k\mathbf{b}_k + 0\mathbf{b}_{k+1} + \cdots + 0\mathbf{b}_n$$

Consequently,

$$C_{B_1}[T(\mathbf{b}_1)] = \left[\begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_k \end{array} \right] \text{ in } \mathbb{R}^k \quad \text{whereas} \quad C_B[T(\mathbf{b}_1)] = \left[\begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_k \\ 0 \\ \vdots \\ 0 \end{array} \right] \text{ in } \mathbb{R}^n$$

This shows that the matrices $M_B(T)$ and $\left[\begin{array}{cc} M_{B_1}(T) & Y \\ 0 & Z \end{array} \right]$ have identical first columns.

Similar statements apply to columns $2, 3, \dots, k$, and this proves the theorem. □

The block upper triangular form for the matrix $M_B(T)$ in Theorem 9.3.1 is very useful because the determinant of such a matrix equals the product of the determinants of each of the diagonal blocks. This is recorded in Theorem 9.3.2 for reference, together with an important application to characteristic polynomials.

Theorem 9.3.2

Let A be a block upper triangular matrix, say

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} \\ 0 & A_{22} & A_{23} & \cdots & A_{2n} \\ 0 & 0 & A_{33} & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & A_{nn} \end{bmatrix}$$

where the diagonal blocks are square. Then:

1. $\det A = (\det A_{11})(\det A_{22})(\det A_{33})\cdots(\det A_{nn})$.
2. $c_A(x) = c_{A_{11}}(x)c_{A_{22}}(x)c_{A_{33}}(x)\cdots c_{A_{nn}}(x)$.

Proof. If $n = 2$, (1) is Theorem 3.1.5; the general case (by induction on n) is left to the reader. Then (2) follows from (1) because

$$xI - A = \begin{bmatrix} xI - A_{11} & -A_{12} & -A_{13} & \cdots & -A_{1n} \\ 0 & xI - A_{22} & -A_{23} & \cdots & -A_{2n} \\ 0 & 0 & xI - A_{33} & \cdots & -A_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & xI - A_{nn} \end{bmatrix}$$

where, in each diagonal block, the symbol I stands for the identity matrix of the appropriate size. □

Example 9.3.5

Consider the linear operator $T : \mathbf{P}_2 \to \mathbf{P}_2$ given by

$$T(a + bx + cx^2) = (-2a - b + 2c) + (a + b)x + (-6a - 2b + 5c)x^2$$

Show that $U = \text{span}\,\{x,\ 1 + 2x^2\}$ is T-invariant, use it to find a block upper triangular matrix for T, and use that to compute $c_T(x)$.

Solution. U is T-invariant by Example 9.3.3 because $U = \text{span}\,\{x,\ 1 + 2x^2\}$ and both $T(x)$ and $T(1 + 2x^2)$ lie in U:

$$\begin{aligned} T(x) &= -1 + x - 2x^2 = x - (1 + 2x^2) \\ T(1 + 2x^2) &= 2 + x + 4x^2 = x + 2(1 + 2x^2) \end{aligned}$$

Extend the basis $B_1 = \{x,\ 1 + 2x^2\}$ of U to a basis B of $\mathbf{P}_2$ in any way at all—say, $B = \{x,\ 1 + 2x^2,\ x^2\}$. Then

$$\begin{aligned} M_B(T) &= \left[\ C_B[T(x)] \quad C_B[T(1 + 2x^2)] \quad C_B[T(x^2)]\ \right] \\ &= \left[\ C_B(-1 + x - 2x^2) \quad C_B(2 + x + 4x^2) \quad C_B(2 + 5x^2)\ \right] \end{aligned}$$

$$= \left[\begin{array}{cc|c} 1 & 1 & 0 \\ -1 & 2 & 2 \\ \hline 0 & 0 & 1 \end{array}\right]$$

is in block upper triangular form as expected. Finally,

$$c_T(x) = \det \left[\begin{array}{cc|c} x-1 & -1 & 0 \\ 1 & x-2 & -2 \\ \hline 0 & 0 & x-1 \end{array}\right] = (x^2 - 3x + 3)(x-1)$$

Eigenvalues

Let $T : V \to V$ be a linear operator. A one-dimensional subspace $\mathbb{R}\mathbf{v}$, $\mathbf{v} \neq \mathbf{0}$, is T-invariant if and only if $T(r\mathbf{v}) = rT(\mathbf{v})$ lies in $\mathbb{R}\mathbf{v}$ for all r in $\mathbb{R}$. This holds if and only if $T(\mathbf{v})$ lies in $\mathbb{R}\mathbf{v}$; that is, $T(\mathbf{v}) = \lambda\mathbf{v}$ for some λ in $\mathbb{R}$. A real number λ is called an **eigenvalue** of an operator $T : V \to V$ if

$$T(\mathbf{v}) = \lambda\mathbf{v}$$

holds for some nonzero vector $\mathbf{v}$ in V. In this case, $\mathbf{v}$ is called an **eigenvector** of T corresponding to λ. The subspace

$$E_\lambda(T) = \{\mathbf{v} \text{ in } V \mid T(\mathbf{v}) = \lambda\mathbf{v}\}$$

is called the **eigenspace** of T corresponding to λ. These terms are consistent with those used in Section 5.5 for matrices. If A is an $n \times n$ matrix, a real number λ is an eigenvalue of the matrix operator $T_A : \mathbb{R}^n \to \mathbb{R}^n$ if and only if λ is an eigenvalue of the matrix A. Moreover, the eigenspaces agree:

$$E_\lambda(T_A) = \{\mathbf{x} \text{ in } \mathbb{R}^n \mid A\mathbf{x} = \lambda\mathbf{x}\} = E_\lambda(A)$$

The following theorem reveals the connection between the eigenspaces of an operator T and those of the matrices representing T.

Theorem 9.3.3

Let $T : V \to V$ be a linear operator where $\dim V = n$, let B denote any ordered basis of V, and let $C_B : V \to \mathbb{R}^n$ denote the coordinate isomorphism. Then:

1. *The eigenvalues λ of T are precisely the eigenvalues of the matrix $M_B(T)$ and thus are the roots of the characteristic polynomial $c_T(x)$.*
2. *In this case the eigenspaces $E_\lambda(T)$ and $E_\lambda[M_B(T)]$ are isomorphic via the restriction $C_B : E_\lambda(T) \to E_\lambda[M_B(T)]$.*

Proof. Write $A = M_B(T)$ for convenience. If $T(\mathbf{v}) = \lambda\mathbf{v}$, then $\lambda C_B(\mathbf{v}) = C_B[T(\mathbf{v})] = AC_B(\mathbf{v})$ because C_B is linear. Hence $C_B(\mathbf{v})$ lies in $E_\lambda(A)$, so we do have a function $C_B : E_\lambda(T) \to E_\lambda(A)$. It is clearly linear and one-to-one; we claim it is onto. If $\mathbf{x}$ is in $E_\lambda(A)$, write $\mathbf{x} = C_B(\mathbf{v})$ for some $\mathbf{v}$ in V (C_B is onto). This $\mathbf{v}$ actually lies in $E_\lambda(T)$. To see why, observe that

$$C_B[T(\mathbf{v})] = AC_B(\mathbf{v}) = A\mathbf{x} = \lambda\mathbf{x} = \lambda C_B(\mathbf{v}) = C_B(\lambda\mathbf{v})$$

Hence $T(\mathbf{v}) = \lambda\mathbf{v}$ because C_B is one-to-one, and this proves (2). As to (1), we have already shown that eigenvalues of T are eigenvalues of A. The converse follows, as in the foregoing proof that C_B is onto. □

Theorem 9.3.3 shows how to pass back and forth between the eigenvectors of an operator T and the eigenvectors of any matrix $M_B(T)$ of T:

$$\mathbf{v} \text{ lies in } E_\lambda(T) \quad \text{if and only if} \quad C_B(\mathbf{v}) \text{ lies in } E_\lambda[M_B(T)]$$

Example 9.3.6

Find the eigenvalues and eigenspaces for $T : \mathbf{P}_2 \to \mathbf{P}_2$ given by

$$T(a+bx+cx^2) = (2a+b+c)+(2a+b-2c)x-(a+2c)x^2$$

Solution. If $B = \{1,\ x,\ x^2\}$, then

$$M_B(T) = \left[\begin{array}{ccc} C_B[T(1)] & C_B[T(x)] & C_B[T(x^2)] \end{array}\right] = \left[\begin{array}{rrr} 2 & 1 & 1 \\ 2 & 1 & -2 \\ -1 & 0 & -2 \end{array}\right]$$

Hence $c_T(x) = \det[xI - M_B(T)] = (x+1)^2(x-3)$ as the reader can verify.

Moreover, $E_{-1}[M_B(T)] = \mathbb{R}\left[\begin{array}{r} -1 \\ 2 \\ 1 \end{array}\right]$ and $E_3[M_B(T)] = \mathbb{R}\left[\begin{array}{r} 5 \\ 6 \\ -1 \end{array}\right]$, so Theorem 9.3.3 gives $E_{-1}(T) = \mathbb{R}(-1+2x+x^2)$ and $E_3(T) = \mathbb{R}(5+6x-x^2)$.

Theorem 9.3.4

Each eigenspace of a linear operator $T : V \to V$ is a T-invariant subspace of V.

Proof. If $\mathbf{v}$ lies in the eigenspace $E_\lambda(T)$, then $T(\mathbf{v}) = \lambda\mathbf{v}$, so $T[T(\mathbf{v})] = T(\lambda\mathbf{v}) = \lambda T(\mathbf{v})$. This shows that $T(\mathbf{v})$ lies in $E_\lambda(T)$ too. □

Direct Sums

Sometimes vectors in a space V can be written naturally as a sum of vectors in two subspaces. For example, in the space $\mathbf{M}_{nn}$ of all $n \times n$ matrices, we have subspaces

$$U = \{P \text{ in } \mathbf{M}_{nn} \mid P \text{ is symmetric }\} \quad \text{and} \quad W = \{Q \text{ in } \mathbf{M}_{nn} \mid Q \text{ is skew symmetric}\}$$

where a matrix Q is called **skew-symmetric** if $Q^T = -Q$. Then every matrix A in $\mathbf{M}_{nn}$ can be written as the sum of a matrix in U and a matrix in W; indeed,

$$A = \tfrac{1}{2}(A+A^T) + \tfrac{1}{2}(A-A^T)$$

where $\frac{1}{2}(A+A^T)$ is symmetric and $\frac{1}{2}(A-A^T)$ is skew symmetric. Remarkably, this representation is unique: If $A=P+Q$ where $P^T=P$ and $Q^T=-Q$, then $A^T=P^T+Q^T=P-Q$; adding this to $A=P+Q$ gives $P=\frac{1}{2}(A+A^T)$, and subtracting gives $Q=\frac{1}{2}(A-A^T)$. In addition, this uniqueness turns out to be closely related to the fact that the only matrix in both U and W is 0. This is a useful way to view matrices, and the idea generalizes to the important notion of a direct sum of subspaces.

If U and W are subspaces of V, their *sum* $U+W$ and their *intersection* $U\cap W$ were defined in Section 6.4 as follows:

$$U+W=\{\mathbf{u}+\mathbf{w}\mid \mathbf{u}\text{ in }U\text{ and }\mathbf{w}\text{ in }W\}$$
$$U\cap W=\{\mathbf{v}\mid \mathbf{v}\text{ lies in both }U\text{ and }W\}$$

These are subspaces of V, the sum containing both U and W and the intersection contained in both U and W. It turns out that the most interesting pairs U and W are those for which $U\cap W$ is as small as possible and $U+W$ is as large as possible.

Definition 9.7 Direct Sum of Subspaces

A vector space V is said to be the ***direct sum*** *of subspaces U and W if*

$$U\cap W=\{\mathbf{0}\}\quad\text{and}\quad U+W=V$$

In this case we write $V=U\oplus W$. Given a subspace U, any subspace W such that $V=U\oplus W$ is called a ***complement*** *of U in V.*

Example 9.3.7

In the space $\mathbb{R}^5$, consider the subspaces $U=\{(a,\ b,\ c,\ 0,\ 0)\mid a,\ b,\ \text{and } c\text{ in }\mathbb{R}\}$ and $W=\{(0,\ 0,\ 0,\ d,\ e)\mid d\text{ and }e\text{ in }\mathbb{R}\}$. Show that $\mathbb{R}^5=U\oplus W$.

Solution. If $\mathbf{x}=(a,\ b,\ c,\ d,\ e)$ is any vector in $\mathbb{R}^5$, then $\mathbf{x}=(a,\ b,\ c,\ 0,\ 0)+(0,\ 0,\ 0,\ d,\ e)$, so $\mathbf{x}$ lies in $U+W$. Hence $\mathbb{R}^5=U+W$. To show that $U\cap W=\{\mathbf{0}\}$, let $\mathbf{x}=(a,\ b,\ c,\ d,\ e)$ lie in $U\cap W$. Then $d=e=0$ because $\mathbf{x}$ lies in U, and $a=b=c=0$ because $\mathbf{x}$ lies in W. Thus $\mathbf{x}=(0,\ 0,\ 0,\ 0,\ 0)=\mathbf{0}$, so $\mathbf{0}$ is the only vector in $U\cap W$. Hence $U\cap W=\{\mathbf{0}\}$.

Example 9.3.8

If U is a subspace of $\mathbb{R}^n$, show that $\mathbb{R}^n=U\oplus U^{\perp}$.

Solution. The equation $\mathbb{R}^n=U+U^{\perp}$ holds because, given $\mathbf{x}$ in $\mathbb{R}^n$, the vector $\text{proj}_U\ \mathbf{x}$ lies in U and $\mathbf{x}-\text{proj}_U\ \mathbf{x}$ lies in $U^{\perp}$. To see that $U\cap U^{\perp}=\{\mathbf{0}\}$, observe that any vector in $U\cap U^{\perp}$ is orthogonal to itself and hence must be zero.

Example 9.3.9

Let $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ be a basis of a vector space V, and partition it into two parts: $\{\mathbf{e}_1, \dots, \mathbf{e}_k\}$ and $\{\mathbf{e}_{k+1}, \dots, \mathbf{e}_n\}$. If $U = \operatorname{span}\{\mathbf{e}_1, \dots, \mathbf{e}_k\}$ and $W = \operatorname{span}\{\mathbf{e}_{k+1}, \dots, \mathbf{e}_n\}$, show that $V = U \oplus W$.

Solution. If $\mathbf{v}$ lies in $U \cap W$, then $\mathbf{v} = a_1\mathbf{e}_1 + \cdots + a_k\mathbf{e}_k$ and $\mathbf{v} = b_{k+1}\mathbf{e}_{k+1} + \cdots + b_n\mathbf{e}_n$ hold for some a_i and b_j in $\mathbb{R}$. The fact that the $\mathbf{e}_i$ are linearly independent forces all $a_i = b_j = 0$, so $\mathbf{v} = \mathbf{0}$. Hence $U \cap W = \{\mathbf{0}\}$. Now, given $\mathbf{v}$ in V, write $\mathbf{v} = v_1\mathbf{e}_1 + \cdots + v_n\mathbf{e}_n$ where the v_i are in $\mathbb{R}$. Then $\mathbf{v} = \mathbf{u} + \mathbf{w}$, where $\mathbf{u} = v_1\mathbf{e}_1 + \cdots + v_k\mathbf{e}_k$ lies in U and $\mathbf{w} = v_{k+1}\mathbf{e}_{k+1} + \cdots + v_n\mathbf{e}_n$ lies in W. This proves that $V = U + W$.

Example 9.3.9 is typical of all direct sum decompositions.

Theorem 9.3.5

Let U and W be subspaces of a finite dimensional vector space V. The following three conditions are equivalent:

1. $V = U \oplus W$.
2. *Each vector $\mathbf{v}$ in V can be written uniquely in the form*

$$\mathbf{v} = \mathbf{u} + \mathbf{w} \qquad \mathbf{u} \text{ in } U, \ \mathbf{w} \text{ in } W$$

3. *If $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ and $\{\mathbf{w}_1, \dots, \mathbf{w}_m\}$ are bases of U and W, respectively, then $B = \{\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{w}_1, \dots, \mathbf{w}_m\}$ is a basis of V.*

(The uniqueness in (2) means that if $\mathbf{v} = \mathbf{u}_1 + \mathbf{w}_1$ is another such representation, then $\mathbf{u}_1 = \mathbf{u}$ and $\mathbf{w}_1 = \mathbf{w}$.)

Proof. Example 9.3.9 shows that (3) $\Rightarrow$ (1).

(1) $\Rightarrow$ (2). Given $\mathbf{v}$ in V, we have $\mathbf{v} = \mathbf{u} + \mathbf{w}$, $\mathbf{u}$ in U, $\mathbf{w}$ in W, because $V = U + W$. If also $\mathbf{v} = \mathbf{u}_1 + \mathbf{w}_1$, then $\mathbf{u} - \mathbf{u}_1 = \mathbf{w}_1 - \mathbf{w}$ lies in $U \cap W = \{\mathbf{0}\}$, so $\mathbf{u} = \mathbf{u}_1$ and $\mathbf{w} = \mathbf{w}_1$.

(2) $\Rightarrow$ (3). Given $\mathbf{v}$ in V, we have $\mathbf{v} = \mathbf{u} + \mathbf{w}$, $\mathbf{u}$ in U, $\mathbf{w}$ in W. Hence $\mathbf{v}$ lies in span B; that is, $V = \operatorname{span} B$. To see that B is independent, let $a_1\mathbf{u}_1 + \cdots + a_k\mathbf{u}_k + b_1\mathbf{w}_1 + \cdots + b_m\mathbf{w}_m = \mathbf{0}$. Write $\mathbf{u} = a_1\mathbf{u}_1 + \cdots + a_k\mathbf{u}_k$ and $\mathbf{w} = b_1\mathbf{w}_1 + \cdots + b_m\mathbf{w}_m$. Then $\mathbf{u} + \mathbf{w} = \mathbf{0}$, and so $\mathbf{u} = \mathbf{0}$ and $\mathbf{w} = \mathbf{0}$ by the uniqueness in (2). Hence $a_i = 0$ for all i and $b_j = 0$ for all j. $\square$

Condition (3) in Theorem 9.3.5 gives the following useful result.

Theorem 9.3.6

If a finite dimensional vector space V is the direct sum $V = U \oplus W$ of subspaces U and W, then

$$\dim V = \dim U + \dim W$$

These direct sum decompositions of V play an important role in any discussion of invariant subspaces.

If $T: V \to V$ is a linear operator and if U_1 is a T-invariant subspace, the block upper triangular matrix

$$M_B(T) = \begin{bmatrix} M_{B_1}(T) & Y \\ 0 & Z \end{bmatrix} \tag{9.3}$$

in Theorem 9.3.1 is achieved by choosing any basis $B_1 = \{\mathbf{b}_1, \dots, \mathbf{b}_k\}$ of U_1 and completing it to a basis $B = \{\mathbf{b}_1, \dots, \mathbf{b}_k, \mathbf{b}_{k+1}, \dots, \mathbf{b}_n\}$ of V in any way at all. The fact that U_1 is T-invariant ensures that the first k columns of $M_B(T)$ have the form in (9.3) (that is, the last $n-k$ entries are zero), and the question arises whether the additional basis vectors $\mathbf{b}_{k+1}, \dots, \mathbf{b}_n$ can be chosen such that

$$U_2 = \text{span}\,\{\mathbf{b}_{k+1}, \dots, \mathbf{b}_n\}$$

is *also* T-invariant. In other words, does each T-invariant subspace of V have a T-invariant complement? Unfortunately the answer in general is no (see Example 9.3.11 below); but when it is possible, the matrix $M_B(T)$ simplifies further. The assumption that the complement $U_2 = \text{span}\,\{\mathbf{b}_{k+1}, \dots, \mathbf{b}_n\}$ is T-invariant too means that $Y = 0$ in equation 9.3 above, and that $Z = M_{B_2}(T)$ is the matrix of the restriction of T to U_2 (where $B_2 = \{\mathbf{b}_{k+1}, \dots, \mathbf{b}_n\}$). The verification is the same as in the proof of Theorem 9.3.1.

Theorem 9.3.7

Let $T: V \to V$ be a linear operator where V has dimension n. Suppose $V = U_1 \oplus U_2$ where both U_1 and U_2 are T-invariant. If $B_1 = \{\mathbf{b}_1, \dots, \mathbf{b}_k\}$ and $B_2 = \{\mathbf{b}_{k+1}, \dots, \mathbf{b}_n\}$ are bases of U_1 and U_2 respectively, then

$$B = \{\mathbf{b}_1, \dots, \mathbf{b}_k, \mathbf{b}_{k+1}, \dots, \mathbf{b}_n\}$$

is a basis of V, and $M_B(T)$ has the block diagonal form

$$M_B(T) = \begin{bmatrix} M_{B_1}(T) & 0 \\ 0 & M_{B_2}(T) \end{bmatrix}$$

where $M_{B_1}(T)$ and $M_{B_2}(T)$ are the matrices of the restrictions of T to U_1 and to U_2 respectively.

Definition 9.8 Reducible Linear Operator

The linear operator $T: V \to V$ is said to be ***reducible*** *if nonzero T-invariant subspaces U_1 and U_2 can be found such that $V = U_1 \oplus U_2$.*

Then T has a matrix in block diagonal form as in Theorem 9.3.7, and the study of T is reduced to studying its restrictions to the lower-dimensional spaces U_1 and U_2. If these can be determined, so can T. Here is an example in which the action of T on the invariant subspaces U_1 and U_2 is very simple indeed. The result for operators is used to derive the corresponding similarity theorem for matrices.

Example 9.3.10

Let $T: V \to V$ be a linear operator satisfying $T^2 = 1_V$ (such operators are called **involutions**). Define

$$U_1 = \{\mathbf{v} \mid T(\mathbf{v}) = \mathbf{v}\} \quad \text{and} \quad U_2 = \{\mathbf{v} \mid T(\mathbf{v}) = -\mathbf{v}\}$$

a. Show that $V = U_1 \oplus U_2$.

b. If $\dim V = n$, find a basis B of V such that $M_B(T) = \left[\begin{array}{cc} I_k & 0 \\ 0 & -I_{n-k} \end{array} \right]$ for some k.

c. Conclude that, if A is an $n \times n$ matrix such that $A^2 = I$, then A is similar to $\left[\begin{array}{cc} I_k & 0 \\ 0 & -I_{n-k} \end{array} \right]$ for some k.

Solution.

a. The verification that U_1 and U_2 are subspaces of V is left to the reader. If $\mathbf{v}$ lies in $U_1 \cap U_2$, then $\mathbf{v} = T(\mathbf{v}) = -\mathbf{v}$, and it follows that $\mathbf{v} = \mathbf{0}$. Hence $U_1 \cap U_2 = \{\mathbf{0}\}$. Given $\mathbf{v}$ in V, write

$$\mathbf{v} = \tfrac{1}{2}\{[\mathbf{v} + T(\mathbf{v})] + [\mathbf{v} - T(\mathbf{v})]\}$$

Then $\mathbf{v} + T(\mathbf{v})$ lies in U_1, because $T[\mathbf{v} + T(\mathbf{v})] = T(\mathbf{v}) + T^2(\mathbf{v}) = \mathbf{v} + T(\mathbf{v})$. Similarly, $\mathbf{v} - T(\mathbf{v})$ lies in U_2, and it follows that $V = U_1 + U_2$. This proves part (a).

b. U_1 and U_2 are easily shown to be T-invariant, so the result follows from Theorem 9.3.7 if bases $B_1 = \{\mathbf{b}_1, \dots, \mathbf{b}_k\}$ and $B_2 = \{\mathbf{b}_{k+1}, \dots, \mathbf{b}_n\}$ of U_1 and U_2 can be found such that $M_{B_1}(T) = I_k$ and $M_{B_2}(T) = -I_{n-k}$. But this is true for *any* choice of B_1 and B_2:

$$\begin{aligned} M_{B_1}(T) &= \left[\begin{array}{cccc} C_{B_1}[T(\mathbf{b}_1)] & C_{B_1}[T(\mathbf{b}_2)] & \cdots & C_{B_1}[T(\mathbf{b}_k)] \end{array} \right] \\ &= \left[\begin{array}{cccc} C_{B_1}(\mathbf{b}_1) & C_{B_1}(\mathbf{b}_2) & \cdots & C_{B_1}(\mathbf{b}_k) \end{array} \right] \\ &= I_k \end{aligned}$$

A similar argument shows that $M_{B_2}(T) = -I_{n-k}$, so part (b) follows with $B = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$.

c. Given A such that $A^2 = I$, consider $T_A : \mathbb{R}^n \to \mathbb{R}^n$. Then $(T_A)^2(\mathbf{x}) = A^2\mathbf{x} = \mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^n$, so $(T_A)^2 = 1_V$. Hence, by part (b), there exists a basis B of $\mathbb{R}^n$ such that

$$M_B(T_A) = \left[\begin{array}{cc} I_r & 0 \\ 0 & -I_{n-r} \end{array} \right]$$

But Theorem 9.2.4 shows that $M_B(T_A) = P^{-1}AP$ for some invertible matrix P, and this proves part (c).

Note that the passage from the result for operators to the analogous result for matrices is routine and can be carried out in any situation, as in the verification of part (c) of Example 9.3.10. The key is the analysis of the operators. In this case, the involutions are just the operators satisfying $T^2 = 1_V$, and the simplicity of this condition means that the invariant subspaces U_1 and U_2 are easy to find.

Unfortunately, not every linear operator $T : V \to V$ is reducible. In fact, the linear operator in Example 9.3.4 has *no* invariant subspaces except 0 and V. On the other hand, one might expect that this is the only type of nonreducible operator; that is, if the operator *has* an invariant subspace that is not 0 or V, then *some* invariant complement must exist. The next example shows that even this is not valid.

Example 9.3.11

Consider the operator $T : \mathbb{R}^2 \to \mathbb{R}^2$ given by $T \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a+b \\ b \end{bmatrix}$. Show that $U_1 = \mathbb{R} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is T-invariant but that U_1 has not T-invariant complement in $\mathbb{R}^2$.

Solution. Because $U_1 = \text{span} \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$ and $T \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, it follows (by Example 9.3.3) that U_1 is T-invariant. Now assume, if possible, that U_1 has a T-invariant complement U_2 in $\mathbb{R}^2$. Then $U_1 \oplus U_2 = \mathbb{R}^2$ and $T(U_2) \subseteq U_2$. Theorem 9.3.6 gives

$$2 = \dim \mathbb{R}^2 = \dim U_1 + \dim U_2 = 1 + \dim U_2$$

so $\dim U_2 = 1$. Let $U_2 = \mathbb{R}\mathbf{u}_2$, and write $\mathbf{u}_2 = \begin{bmatrix} p \\ q \end{bmatrix}$. We claim that $\mathbf{u}_2$ is not in U_1. For if $\mathbf{u}_2 \in U_1$, then $\mathbf{u}_2 \in U_1 \cap U_2 = \{\mathbf{0}\}$, so $\mathbf{u}_2 = \mathbf{0}$. But then $U_2 = \mathbb{R}\mathbf{u}_2 = \{\mathbf{0}\}$, a contradiction, as $\dim U_2 = 1$. So $\mathbf{u}_2 \notin U_1$, from which $q \neq 0$. On the other hand, $T(\mathbf{u}_2) \in U_2 = \mathbb{R}\mathbf{u}_2$ (because U_2 is T-invariant), say $T(\mathbf{u}_2) = \lambda \mathbf{u}_2 = \lambda \begin{bmatrix} p \\ q \end{bmatrix}$.

Thus

$$\begin{bmatrix} p+q \\ q \end{bmatrix} = T \begin{bmatrix} p \\ q \end{bmatrix} = \lambda \begin{bmatrix} p \\ q \end{bmatrix} \text{ where } \lambda \in \mathbb{R}$$

Hence $p+q = \lambda p$ and $q = \lambda q$. Because $q \neq 0$, the second of these equations implies that $\lambda = 1$, so the first equation implies $q = 0$, a contradiction. So a T-invariant complement of U_1 does not exist.

This is as far as we take the theory here, but in Chapter 11 the techniques introduced in this section will be refined to show that every matrix is similar to a very nice matrix indeed—its Jordan canonical form.

Chapter 10

Inner Product Spaces

10.1 Inner Products and Norms

The dot product was introduced in $\mathbb{R}^n$ to provide a natural generalization of the geometrical notions of length and orthogonality that were so important in Chapter 4. The plan in this chapter is to define an *inner product* on an arbitrary real vector space V (of which the dot product is an example in $\mathbb{R}^n$) and use it to introduce these concepts in V. While this causes some repetition of arguments in Chapter 8, it is well worth the effort because of the much wider scope of the results when stated in full generality.

Definition 10.1 Inner Product Spaces

*An **inner product** on a real vector space V is a function that assigns a real number $\langle \mathbf{v}, \mathbf{w} \rangle$ to every pair $\mathbf{v}$, $\mathbf{w}$ of vectors in V in such a way that the following axioms are satisfied.*

P1. $\langle \mathbf{v}, \mathbf{w} \rangle$ *is a real number for all* $\mathbf{v}$ *and* $\mathbf{w}$ *in* V.

P2. $\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle$ *for all* $\mathbf{v}$ *and* $\mathbf{w}$ *in* V.

P3. $\langle \mathbf{v}+\mathbf{w}, \mathbf{u} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle + \langle \mathbf{w}, \mathbf{u} \rangle$ *for all* $\mathbf{u}$, $\mathbf{v}$, *and* $\mathbf{w}$ *in* V.

P4. $\langle r\mathbf{v}, \mathbf{w} \rangle = r\langle \mathbf{v}, \mathbf{w} \rangle$ *for all* $\mathbf{v}$ *and* $\mathbf{w}$ *in* V *and all* r *in* $\mathbb{R}$.

P5. $\langle \mathbf{v}, \mathbf{v} \rangle > 0$ *for all* $\mathbf{v} \neq \mathbf{0}$ *in* V.

A real vector space V with an inner product $\langle \, , \, \rangle$ will be called an **inner product space**. Note that every subspace of an inner product space is again an inner product space using the same inner product.[1]

Example 10.1.1

$\mathbb{R}^n$ is an inner product space with the dot product as inner product:

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v} \cdot \mathbf{w} \quad \text{for all } \mathbf{v}, \mathbf{w} \in \mathbb{R}^n$$

See Theorem 5.3.1. This is also called the **euclidean** inner product, and $\mathbb{R}^n$, equipped with the dot product, is called **euclidean n-space**.

Example 10.1.2

If A and B are $m \times n$ matrices, define $\langle A, B \rangle = \text{tr}\,(AB^T)$ where $\text{tr}\,(X)$ is the trace of the square matrix X. Show that $\langle \, , \, \rangle$ is an inner product in $\mathbf{M}_{mn}$.

[1] If we regard $\mathbb{C}^n$ as a vector space over the field $\mathbb{C}$ of complex numbers, then the "standard inner product" on $\mathbb{C}^n$ defined in Section 8.7 does not satisfy Axiom P4 (see Theorem 8.7.1(3)).

Solution. P1 is clear. Since $\text{tr}(P) = \text{tr}(P^T)$ for every square matrix P, we have P2:

$$\langle A, B\rangle = \text{tr}(AB^T) = \text{tr}[(AB^T)^T] = \text{tr}(BA^T) = \langle B, A\rangle$$

Next, P3 and P4 follow because trace is a linear transformation $\mathbf{M}_{mn} \to \mathbb{R}$ (Exercise **??**). Turning to P5, let $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_m$ denote the rows of the matrix A. Then the (i, j)-entry of AA^T is $\mathbf{r}_i \cdot \mathbf{r}_j$, so

$$\langle A, A\rangle = \text{tr}(AA^T) = \mathbf{r}_1 \cdot \mathbf{r}_1 + \mathbf{r}_2 \cdot \mathbf{r}_2 + \cdots + \mathbf{r}_m \cdot \mathbf{r}_m$$

But $\mathbf{r}_j \cdot \mathbf{r}_j$ is the sum of the squares of the entries of $\mathbf{r}_j$, so this shows that $\langle A, A\rangle$ is the sum of the squares of all nm entries of A. Axiom P5 follows.

The importance of the next example in analysis is difficult to overstate.

Example 10.1.3: [2]

Let $\mathbf{C}[a, b]$ denote the vector space of **continuous functions** from $[a, b]$ to $\mathbb{R}$, a subspace of $\mathbf{F}[a, b]$. Show that

$$\langle f, g\rangle = \int_a^b f(x)g(x)dx$$

defines an inner product on $\mathbf{C}[a, b]$.

Solution. Axioms P1 and P2 are clear. As to axiom P4,

$$\langle rf, g\rangle = \int_a^b rf(x)g(x)dx = r\int_a^b f(x)g(x)dx = r\langle f, g\rangle$$

Axiom P3 is similar. Finally, theorems of calculus show that $\langle f, f\rangle = \int_a^b f(x)^2dx \geq 0$ and, if f is continuous, that this is zero if and only if f is the zero function. This gives axiom P5.

If $\mathbf{v}$ is any vector, then, using axiom P3, we get

$$\langle \mathbf{0}, \mathbf{v}\rangle = \langle \mathbf{0}+\mathbf{0}, \mathbf{v}\rangle = \langle \mathbf{0}, \mathbf{v}\rangle + \langle \mathbf{0}, \mathbf{v}\rangle$$

and it follows that the number $\langle \mathbf{0}, \mathbf{v}\rangle$ must be zero. This observation is recorded for reference in the following theorem, along with several other properties of inner products. The other proofs are left as Exercise **??**.

Theorem 10.1.1

Let $\langle\ ,\ \rangle$ be an inner product on a space V; let $\mathbf{v}$, $\mathbf{u}$, and $\mathbf{w}$ denote vectors in V; and let r denote a real number.

1. $\langle \mathbf{u}, \mathbf{v}+\mathbf{w}\rangle = \langle \mathbf{u}, \mathbf{v}\rangle + \langle \mathbf{u}, \mathbf{w}\rangle$

[2] This example (and others later that refer to it) can be omitted with no loss of continuity by students with no calculus background.

2. $\langle \mathbf{v},\ r\mathbf{w} \rangle = r\langle \mathbf{v},\ \mathbf{w} \rangle = \langle r\mathbf{v},\ \mathbf{w} \rangle$

3. $\langle \mathbf{v},\ \mathbf{0} \rangle = 0 = \langle \mathbf{0},\ \mathbf{v} \rangle$

4. $\langle \mathbf{v},\ \mathbf{v} \rangle = 0$ *if and only if* $\mathbf{v} = \mathbf{0}$

If $\langle\ ,\ \rangle$ is an inner product on a space V, then, given $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ in V,

$$\langle r\mathbf{u} + s\mathbf{v},\ \mathbf{w} \rangle = \langle r\mathbf{u},\ \mathbf{w} \rangle + \langle s\mathbf{v},\ \mathbf{w} \rangle = r\langle \mathbf{u},\ \mathbf{w} \rangle + s\langle \mathbf{v},\ \mathbf{w} \rangle$$

for all r and s in $\mathbb{R}$ by axioms P3 and P4. Moreover, there is nothing special about the fact that there are two terms in the linear combination or that it is in the first component:

$$\begin{aligned} \langle r_1\mathbf{v}_1 + r_2\mathbf{v}_2 + \cdots + r_n\mathbf{v}_n,\ \mathbf{w} \rangle &= r_1\langle \mathbf{v}_1,\ \mathbf{w} \rangle + r_2\langle \mathbf{v}_2,\ \mathbf{w} \rangle + \cdots + r_n\langle \mathbf{v}_n,\ \mathbf{w} \rangle, \text{ and} \\ \langle \mathbf{v},\ s_1\mathbf{w}_1 + s_2\mathbf{w}_2 + \cdots + s_m\mathbf{w}_m \rangle &= s_1\langle \mathbf{v},\ \mathbf{w}_1 \rangle + s_2\langle \mathbf{v},\ \mathbf{w}_2 \rangle + \cdots + s_m\langle \mathbf{v},\ \mathbf{w}_m \rangle \end{aligned}$$

hold for all r_i and s_i in $\mathbb{R}$ and all $\mathbf{v}$, $\mathbf{w}$, $\mathbf{v}_i$, and $\mathbf{w}_j$ in V. These results are described by saying that inner products "preserve" linear combinations. For example,

$$\begin{aligned} \langle 2\mathbf{u} - \mathbf{v},\ 3\mathbf{u} + 2\mathbf{v} \rangle &= \langle 2\mathbf{u},\ 3\mathbf{u} \rangle + \langle 2\mathbf{u},\ 2\mathbf{v} \rangle + \langle -\mathbf{v},\ 3\mathbf{u} \rangle + \langle -\mathbf{v},\ 2\mathbf{v} \rangle \\ &= 6\langle \mathbf{u},\ \mathbf{u} \rangle + 4\langle \mathbf{u},\ \mathbf{v} \rangle - 3\langle \mathbf{v},\ \mathbf{u} \rangle - 2\langle \mathbf{v},\ \mathbf{v} \rangle \\ &= 6\langle \mathbf{u},\ \mathbf{u} \rangle + \langle \mathbf{u},\ \mathbf{v} \rangle - 2\langle \mathbf{v},\ \mathbf{v} \rangle \end{aligned}$$

If A is a symmetric $n \times n$ matrix and $\mathbf{x}$ and $\mathbf{y}$ are columns in $\mathbb{R}^n$, we regard the 1×1 matrix $\mathbf{x}^T A\mathbf{y}$ as a number. If we write

$$\langle \mathbf{x},\ \mathbf{y} \rangle = \mathbf{x}^T A\mathbf{y} \quad \text{for all columns } \mathbf{x},\ \mathbf{y} \text{ in } \mathbb{R}^n$$

then axioms P1–P4 follow from matrix arithmetic (only P2 requires that A is symmetric). Axiom P5 reads

$$\mathbf{x}^T A\mathbf{x} > 0 \quad \text{for all columns } \mathbf{x} \neq \mathbf{0} \text{ in } \mathbb{R}^n$$

and this condition characterizes the positive definite matrices (Theorem 8.3.2). This proves the first assertion in the next theorem.

Theorem 10.1.2

If A is any $n \times n$ positive definite matrix, then

$$\langle \mathbf{x},\ \mathbf{y} \rangle = \mathbf{x}^T A\mathbf{y} \text{ for all columns } \mathbf{x},\ \mathbf{y} \text{ in } \mathbb{R}^n$$

defines an inner product on $\mathbb{R}^n$, and every inner product on $\mathbb{R}^n$ arises in this way.

Proof. Given an inner product $\langle\ ,\ \rangle$ on $\mathbb{R}^n$, let $\{\mathbf{e}_1,\ \mathbf{e}_2,\ \dots,\ \mathbf{e}_n\}$ be the standard basis of $\mathbb{R}^n$. If $\mathbf{x} = \sum_{i=1}^{n} x_i\mathbf{e}_i$ and $\mathbf{y} = \sum_{j=1}^{n} y_j\mathbf{e}_j$ are two vectors in $\mathbb{R}^n$, compute $\langle \mathbf{x},\ \mathbf{y} \rangle$ by adding the inner product of each term $x_i\mathbf{e}_i$ to each term $y_j\mathbf{e}_j$. The result is a double sum.

$$\langle \mathbf{x},\ \mathbf{y} \rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} \langle x_i\mathbf{e}_i,\ y_j\mathbf{e}_j \rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i\langle \mathbf{e}_i,\ \mathbf{e}_j \rangle y_j$$

As the reader can verify, this is a matrix product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \left[\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array} \right] \left[\begin{array}{cccc} \langle \mathbf{e}_1, \mathbf{e}_1 \rangle & \langle \mathbf{e}_1, \mathbf{e}_2 \rangle & \cdots & \langle \mathbf{e}_1, \mathbf{e}_n \rangle \\ \langle \mathbf{e}_2, \mathbf{e}_1 \rangle & \langle \mathbf{e}_2, \mathbf{e}_2 \rangle & \cdots & \langle \mathbf{e}_2, \mathbf{e}_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{e}_n, \mathbf{e}_1 \rangle & \langle \mathbf{e}_n, \mathbf{e}_2 \rangle & \cdots & \langle \mathbf{e}_n, \mathbf{e}_n \rangle \end{array} \right] \left[\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array} \right]$$

Hence $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T A \mathbf{y}$, where A is the $n \times n$ matrix whose (i, j)-entry is $\langle \mathbf{e}_i, \mathbf{e}_j \rangle$. The fact that

$$\langle \mathbf{e}_i, \mathbf{e}_j \rangle = \langle \mathbf{e}_j, \mathbf{e}_i \rangle$$

shows that A is symmetric. Finally, A is positive definite by Theorem 8.3.2. □

Thus, just as every linear operator $\mathbb{R}^n \to \mathbb{R}^n$ corresponds to an $n \times n$ matrix, every inner product on $\mathbb{R}^n$ corresponds to a positive definite $n \times n$ matrix. In particular, the dot product corresponds to the identity matrix I_n.

Remark

If we refer to the inner product space $\mathbb{R}^n$ without specifying the inner product, we mean that the dot product is to be used.

Example 10.1.4

Let the inner product $\langle\ ,\ \rangle$ be defined on $\mathbb{R}^2$ by

$$\left\langle \left[\begin{array}{c} v_1 \\ v_2 \end{array} \right], \left[\begin{array}{c} w_1 \\ w_2 \end{array} \right] \right\rangle = 2v_1w_1 - v_1w_2 - v_2w_1 + v_2w_2$$

Find a symmetric 2×2 matrix A such that $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T A \mathbf{y}$ for all $\mathbf{x}$, $\mathbf{y}$ in $\mathbb{R}^2$.

Solution. The (i, j)-entry of the matrix A is the coefficient of v_iw_j in the expression, so $A = \left[\begin{array}{rr} 2 & -1 \\ -1 & 1 \end{array} \right]$. Incidentally, if $\mathbf{x} = \left[\begin{array}{c} x \\ y \end{array} \right]$, then

$$\langle \mathbf{x}, \mathbf{x} \rangle = 2x^2 - 2xy + y^2 = x^2 + (x-y)^2 \geq 0$$

for all $\mathbf{x}$, so $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ implies $\mathbf{x} = \mathbf{0}$. Hence $\langle\ ,\ \rangle$ is indeed an inner product, so A is positive definite.

Let $\langle\ ,\ \rangle$ be an inner product on $\mathbb{R}^n$ given as in Theorem 10.1.2 by a positive definite matrix A. If $\mathbf{x} = \left[\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array} \right]^T$, then $\langle \mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^T A \mathbf{x}$ is an expression in the variables $x_1, x_2, \dots, x_n$ called a **quadratic form**. These are studied in detail in Section 8.9.

Norm and Distance

Definition 10.2 Norm and Distance

*As in $\mathbb{R}^n$, if $\langle\ ,\ \rangle$ is an inner product on a space V, the **norm**[3] $\|\mathbf{v}\|$ of a vector $\mathbf{v}$ in V is defined by*

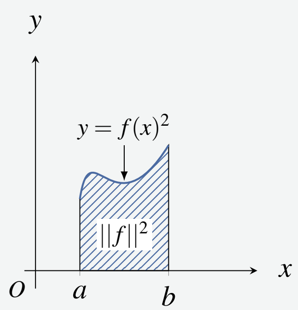
$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v},\ \mathbf{v}\rangle}$$

*We define the **distance** between vectors $\mathbf{v}$ and $\mathbf{w}$ in an inner product space V to be*

$$\mathrm{d}\,(\mathbf{v},\ \mathbf{w}) = \|\mathbf{v}-\mathbf{w}\|$$

Note that axiom P5 guarantees that $\langle \mathbf{v},\ \mathbf{v}\rangle \geq 0$, so $\|\mathbf{v}\|$ is a real number.

Example 10.1.5



The norm of a continuous function $f = f(x)$ in $\mathbf{C}[a,\ b]$ (with the inner product from Example 10.1.3) is given by

$$\|f\| = \sqrt{\int_a^b f(x)^2 dx}$$

Hence $\|f\|^2$ is the area beneath the graph of $y = f(x)^2$ between $x = a$ and $x = b$ (shaded in the diagram).

Example 10.1.6

Show that $\langle \mathbf{u}+\mathbf{v},\ \mathbf{u}-\mathbf{v}\rangle = \|\mathbf{u}\|^2 - \|\mathbf{v}\|^2$ in any inner product space.

Solution.

$$\begin{aligned}\langle \mathbf{u}+\mathbf{v},\ \mathbf{u}-\mathbf{v}\rangle &= \langle \mathbf{u},\ \mathbf{u}\rangle - \langle \mathbf{u},\ \mathbf{v}\rangle + \langle \mathbf{v},\ \mathbf{u}\rangle - \langle \mathbf{v},\ \mathbf{v}\rangle \\ &= \|\mathbf{u}\|^2 - \langle \mathbf{u},\ \mathbf{v}\rangle + \langle \mathbf{u},\ \mathbf{v}\rangle - \|\mathbf{v}\|^2 \\ &= \|\mathbf{u}\|^2 - \|\mathbf{v}\|^2\end{aligned}$$

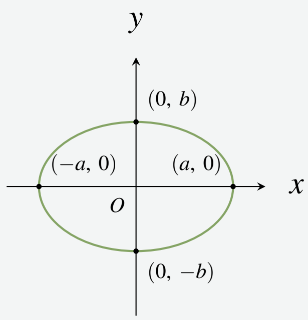
A vector $\mathbf{v}$ in an inner product space V is called a **unit vector** if $\|\mathbf{v}\| = 1$. The set of all unit vectors in V is called the **unit ball** in V. For example, if $V = \mathbb{R}^2$ (with the dot product) and $\mathbf{v} = (x,\ y)$, then

$$\|\mathbf{v}\|^2 = 1 \quad \text{if and only if} \quad x^2 + y^2 = 1$$

Hence the unit ball in $\mathbb{R}^2$ is the **unit circle** $x^2 + y^2 = 1$ with centre at the origin and radius 1. However, the shape of the unit ball varies with the choice of inner product.

[3]If the dot product is used in $\mathbb{R}^n$, the norm $\|\mathbf{x}\|$ of a vector $\mathbf{x}$ is usually called the **length** of $\mathbf{x}$.

Example 10.1.7



Let $a > 0$ and $b > 0$. If $\mathbf{v} = (x, y)$ and $\mathbf{w} = (x_1, y_1)$, define an inner product on $\mathbb{R}^2$ by

$$\langle \mathbf{v}, \mathbf{w} \rangle = \frac{xx_1}{a^2} + \frac{yy_1}{b^2}$$

The reader can verify (Exercise **??**) that this is indeed an inner product. In this case

$$\|\mathbf{v}\|^2 = 1 \quad \text{if and only if} \quad \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

so the unit ball is the ellipse shown in the diagram.

Example 10.1.7 graphically illustrates the fact that norms and distances in an inner product space V vary with the choice of inner product in V.

Theorem 10.1.3

If $\mathbf{v} \neq \mathbf{0}$ is any vector in an inner product space V, then $\frac{1}{\|\mathbf{v}\|}\mathbf{v}$ is the unique unit vector that is a positive multiple of $\mathbf{v}$.

The next theorem reveals an important and useful fact about the relationship between norms and inner products, extending the Cauchy inequality for $\mathbb{R}^n$ (Theorem 5.3.2).

Theorem 10.1.4: Cauchy-Schwarz Inequality[4]

If $\mathbf{v}$ and $\mathbf{w}$ are two vectors in an inner product space V, then

$$\langle \mathbf{v}, \mathbf{w} \rangle^2 \leq \|\mathbf{v}\|^2 \|\mathbf{w}\|^2$$

Moreover, equality occurs if and only if one of $\mathbf{v}$ and $\mathbf{w}$ is a scalar multiple of the other.

Proof. Write $\|\mathbf{v}\| = a$ and $\|\mathbf{w}\| = b$. Using Theorem 10.1.1 we compute:

$$\begin{aligned} \|b\mathbf{v} - a\mathbf{w}\|^2 &= b^2\|\mathbf{v}\|^2 - 2ab\langle \mathbf{v}, \mathbf{w} \rangle + a^2\|\mathbf{w}\|^2 = 2ab(ab - \langle \mathbf{v}, \mathbf{w} \rangle) \\ \|b\mathbf{v} + a\mathbf{w}\|^2 &= b^2\|\mathbf{v}\|^2 + 2ab\langle \mathbf{v}, \mathbf{w} \rangle + a^2\|\mathbf{w}\|^2 = 2ab(ab + \langle \mathbf{v}, \mathbf{w} \rangle) \end{aligned} \tag{10.1}$$

It follows that $ab - \langle \mathbf{v}, \mathbf{w} \rangle \geq 0$ and $ab + \langle \mathbf{v}, \mathbf{w} \rangle \geq 0$, and hence that $-ab \leq \langle \mathbf{v}, \mathbf{w} \rangle \leq ab$. But then $|\langle \mathbf{v}, \mathbf{w} \rangle| \leq ab = \|\mathbf{v}\|\|\mathbf{w}\|$, as desired.

Conversely, if $|\langle \mathbf{v}, \mathbf{w} \rangle| = \|\mathbf{v}\|\|\mathbf{w}\| = ab$ then $\langle \mathbf{v}, \mathbf{w} \rangle = \pm ab$. Hence (10.1) shows that $b\mathbf{v} - a\mathbf{w} = \mathbf{0}$ or $b\mathbf{v} + a\mathbf{w} = \mathbf{0}$. It follows that one of $\mathbf{v}$ and $\mathbf{w}$ is a scalar multiple of the other, even if $a = 0$ or $b = 0$. □

[4]Hermann Amandus Schwarz (1843–1921) was a German mathematician at the University of Berlin. He had strong geometric intuition, which he applied with great ingenuity to particular problems. A version of the inequality appeared in 1885.

Example 10.1.8

If f and g are continuous functions on the interval $[a, b]$, then (see Example 10.1.3)

$$\left(\int_a^b f(x)g(x)dx\right)^2 \leq \int_a^b f(x)^2 dx \int_a^b g(x)^2 dx$$

Another famous inequality, the so-called *triangle inequality*, also comes from the Cauchy-Schwarz inequality. It is included in the following list of basic properties of the norm of a vector.

Theorem 10.1.5

If V is an inner product space, the norm $\|\cdot\|$ has the following properties.

1. $\|\mathbf{v}\| \geq 0$ *for every vector* $\mathbf{v}$ *in* V.
2. $\|\mathbf{v}\| = 0$ *if and only if* $\mathbf{v} = \mathbf{0}$.
3. $\|r\mathbf{v}\| = |r|\|\mathbf{v}\|$ *for every* $\mathbf{v}$ *in* V *and every* r *in* $\mathbb{R}$.
4. $\|\mathbf{v}+\mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|$ *for all* $\mathbf{v}$ *and* $\mathbf{w}$ *in* V *(**triangle inequality**).*

Proof. Because $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$, properties (1) and (2) follow immediately from (3) and (4) of Theorem 10.1.1. As to (3), compute

$$\|r\mathbf{v}\|^2 = \langle r\mathbf{v}, r\mathbf{v} \rangle = r^2 \langle \mathbf{v}, \mathbf{v} \rangle = r^2 \|\mathbf{v}\|^2$$

Hence (3) follows by taking positive square roots. Finally, the fact that $\langle \mathbf{v}, \mathbf{w} \rangle \leq \|\mathbf{v}\|\|\mathbf{w}\|$ by the Cauchy-Schwarz inequality gives

$$\begin{aligned} \|\mathbf{v}+\mathbf{w}\|^2 = \langle \mathbf{v}+\mathbf{w}, \mathbf{v}+\mathbf{w} \rangle &= \|\mathbf{v}\|^2 + 2\langle \mathbf{v}, \mathbf{w} \rangle + \|\mathbf{w}\|^2 \\ &\leq \|\mathbf{v}\|^2 + 2\|\mathbf{v}\|\|\mathbf{w}\| + \|\mathbf{w}\|^2 \\ &= (\|\mathbf{v}\| + \|\mathbf{w}\|)^2 \end{aligned}$$

Hence (4) follows by taking positive square roots. □

It is worth noting that the usual triangle inequality for absolute values,

$$|r+s| \leq |r| + |s| \text{ for all real numbers } r \text{ and } s$$

is a special case of (4) where $V = \mathbb{R} = \mathbb{R}^1$ and the dot product $\langle r, s \rangle = rs$ is used.

In many calculations in an inner product space, it is required to show that some vector $\mathbf{v}$ is zero. This is often accomplished most easily by showing that its norm $\|\mathbf{v}\|$ is zero. Here is an example.

Example 10.1.9

Let $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ be a spanning set for an inner product space V. If $\mathbf{v}$ in V satisfies $\langle \mathbf{v}, \mathbf{v}_i \rangle = 0$ for each $i = 1, 2, \dots, n$, show that $\mathbf{v} = \mathbf{0}$.

Solution. Write $\mathbf{v} = r_1\mathbf{v}_1 + \cdots + r_n\mathbf{v}_n$, r_i in $\mathbb{R}$. To show that $\mathbf{v} = \mathbf{0}$, we show that $\|\mathbf{v}\|^2 = \langle \mathbf{v}, \mathbf{v} \rangle = 0$. Compute:

$$\langle \mathbf{v}, \mathbf{v} \rangle = \langle \mathbf{v}, r_1\mathbf{v}_1 + \cdots + r_n\mathbf{v}_n \rangle = r_1\langle \mathbf{v}, \mathbf{v}_1 \rangle + \cdots + r_n\langle \mathbf{v}, \mathbf{v}_n \rangle = 0$$

by hypothesis, and the result follows.

The norm properties in Theorem 10.1.5 translate to the following properties of distance familiar from geometry. The proof is Exercise **??**.

Theorem 10.1.6

Let V be an inner product space.

1. $\mathrm{d}(\mathbf{v}, \mathbf{w}) \geq 0$ *for all* $\mathbf{v}$, $\mathbf{w}$ *in* V.
2. $\mathrm{d}(\mathbf{v}, \mathbf{w}) = 0$ *if and only if* $\mathbf{v} = \mathbf{w}$.
3. $\mathrm{d}(\mathbf{v}, \mathbf{w}) = \mathrm{d}(\mathbf{w}, \mathbf{v})$ *for all* $\mathbf{v}$ *and* $\mathbf{w}$ *in* V.
4. $\mathrm{d}(\mathbf{v}, \mathbf{w}) \leq \mathrm{d}(\mathbf{v}, \mathbf{u}) + \mathrm{d}(\mathbf{u}, \mathbf{w})$ *for all* $\mathbf{v}$, $\mathbf{u}$, *and* $\mathbf{w}$ *in* V.

10.2 Orthogonal Sets of Vectors

The idea that two lines can be perpendicular is fundamental in geometry, and this section is devoted to introducing this notion into a general inner product space V. To motivate the definition, recall that two nonzero geometric vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$ are perpendicular (or orthogonal) if and only if $\mathbf{x} \cdot \mathbf{y} = 0$. In general, two vectors $\mathbf{v}$ and $\mathbf{w}$ in an inner product space V are said to be **orthogonal** if

$$\langle \mathbf{v}, \mathbf{w} \rangle = 0$$

A set $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ of vectors is called an **orthogonal set of vectors** if

1. *Each* $\mathbf{f}_i \neq \mathbf{0}$.
2. $\langle \mathbf{f}_i, \mathbf{f}_j \rangle = 0$ *for all* $i \neq j$.

If, in addition, $\|\mathbf{f}_i\| = 1$ for each i, the set $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ is called an **orthonormal** set.

Example 10.2.1

$\{\sin x, \cos x\}$ is orthogonal in $\mathbf{C}[-\pi, \pi]$ because

$$\int_{-\pi}^{\pi} \sin x \cos x \, dx = \left[-\tfrac{1}{4} \cos 2x\right]_{-\pi}^{\pi} = 0$$

The first result about orthogonal sets extends Pythagoras' theorem in $\mathbb{R}^n$ (Theorem 5.3.4) and the same proof works.

Theorem 10.2.1: Pythagoras' Theorem

If $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ *is an orthogonal set of vectors, then*

$$\|\mathbf{f}_1 + \mathbf{f}_2 + \cdots + \mathbf{f}_n\|^2 = \|\mathbf{f}_1\|^2 + \|\mathbf{f}_2\|^2 + \cdots + \|\mathbf{f}_n\|^2$$

The proof of the next result is left to the reader.

Theorem 10.2.2

Let $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ *be an orthogonal set of vectors.*

1. $\{r_1\mathbf{f}_1, r_2\mathbf{f}_2, \dots, r_n\mathbf{f}_n\}$ *is also orthogonal for any* $r_i \neq 0$ *in* $\mathbb{R}$.
2. $\left\{\frac{1}{\|\mathbf{f}_1\|}\mathbf{f}_1, \frac{1}{\|\mathbf{f}_2\|}\mathbf{f}_2, \dots, \frac{1}{\|\mathbf{f}_n\|}\mathbf{f}_n\right\}$ *is an orthonormal set.*

As before, the process of passing from an orthogonal set to an orthonormal one is called **normalizing** the orthogonal set. The proof of Theorem 5.3.5 goes through to give

Theorem 10.2.3

Every orthogonal set of vectors is linearly independent.

Example 10.2.2

Show that $\left\{ \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ 2 \end{bmatrix} \right\}$ is an orthogonal basis of $\mathbb{R}^3$ with inner product

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T A \mathbf{w}, \quad \text{where } A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Solution. We have

$$\left\langle \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \right\rangle = \begin{bmatrix} 2 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = 0$$

and the reader can verify that the other pairs are orthogonal too. Hence the set is orthogonal, so it is linearly independent by Theorem 10.2.3. Because $\dim \mathbb{R}^3 = 3$, it is a basis.

The proof of Theorem 5.3.6 generalizes to give the following:

Theorem 10.2.4: Expansion Theorem

Let $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ be an orthogonal basis of an inner product space V. If $\mathbf{v}$ is any vector in V, then

$$\mathbf{v} = \frac{\langle \mathbf{v}, \mathbf{f}_1 \rangle}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 + \frac{\langle \mathbf{v}, \mathbf{f}_2 \rangle}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 + \cdots + \frac{\langle \mathbf{v}, \mathbf{f}_n \rangle}{\|\mathbf{f}_n\|^2}\mathbf{f}_n$$

is the expansion of $\mathbf{v}$ as a linear combination of the basis vectors.

The coefficients $\frac{\langle \mathbf{v}, \mathbf{f}_1 \rangle}{\|\mathbf{f}_1\|^2}, \frac{\langle \mathbf{v}, \mathbf{f}_2 \rangle}{\|\mathbf{f}_2\|^2}, \dots, \frac{\langle \mathbf{v}, \mathbf{f}_n \rangle}{\|\mathbf{f}_n\|^2}$ in the expansion theorem are sometimes called the **Fourier coefficients** of $\mathbf{v}$ with respect to the orthogonal basis $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$. This is in honour of the French mathematician J.B.J. Fourier (1768–1830). His original work was with a particular orthogonal set in the space $\mathbf{C}[a, b]$, about which there will be more to say in Section 10.5.

Example 10.2.3

If $a_0, a_1, \dots, a_n$ are distinct numbers and $p(x)$ and $q(x)$ are in $\mathbf{P}_n$, define

$$\langle p(x), q(x) \rangle = p(a_0)q(a_0) + p(a_1)q(a_1) + \cdots + p(a_n)q(a_n)$$

This is an inner product on $\mathbf{P}_n$. (Axioms P1–P4 are routinely verified, and P5 holds because 0 is the only polynomial of degree n with $n+1$ distinct roots. See Theorem 6.5.4 or Appendix D.) Recall that the **Lagrange polynomials** $\delta_0(x), \delta_1(x), \dots, \delta_n(x)$ relative to the numbers

$a_0, a_1, \dots, a_n$ are defined as follows (see Section 6.5):

$$\delta_k(x) = \frac{\prod_{i \neq k}(x - a_i)}{\prod_{i \neq k}(a_k - a_i)} \quad k = 0, 1, 2, \dots, n$$

where $\prod_{i \neq k}(x - a_i)$ means the product of all the terms

$$(x - a_0), (x - a_1), (x - a_2), \dots, (x - a_n)$$

except that the kth term is omitted. Then $\{\delta_0(x), \delta_1(x), \dots, \delta_n(x)\}$ is orthonormal with respect to $\langle \, , \, \rangle$ because $\delta_k(a_i) = 0$ if $i \neq k$ and $\delta_k(a_k) = 1$. These facts also show that $\langle p(x), \delta_k(x) \rangle = p(a_k)$ so the expansion theorem gives

$$p(x) = p(a_0)\delta_0(x) + p(a_1)\delta_1(x) + \cdots + p(a_n)\delta_n(x)$$

for each $p(x)$ in $\mathbf{P}_n$. This is the **Lagrange interpolation expansion** of $p(x)$, Theorem 6.5.3, which is important in numerical integration.

Lemma 10.2.1: Orthogonal Lemma

Let $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$ be an orthogonal set of vectors in an inner product space V, and let $\mathbf{v}$ be any vector not in $\text{span}\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$. *Define*

$$\mathbf{f}_{m+1} = \mathbf{v} - \frac{\langle \mathbf{v}, \mathbf{f}_1 \rangle}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \frac{\langle \mathbf{v}, \mathbf{f}_2 \rangle}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 - \cdots - \frac{\langle \mathbf{v}, \mathbf{f}_m \rangle}{\|\mathbf{f}_m\|^2}\mathbf{f}_m$$

Then $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m, \mathbf{f}_{m+1}\}$ is an orthogonal set of vectors.

The proof of this result (and the next) is the same as for the dot product in $\mathbb{R}^n$ (Lemma 8.1.1 and Theorem 8.1.2).

Theorem 10.2.5: Gram-Schmidt Orthogonalization Algorithm

Let V be an inner product space and let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ be any basis of V. Define vectors $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n$ in V successively as follows:

$$\begin{aligned}
\mathbf{f}_1 &= \mathbf{v}_1 \\
\mathbf{f}_2 &= \mathbf{v}_2 - \frac{\langle \mathbf{v}_2, \mathbf{f}_1 \rangle}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 \\
\mathbf{f}_3 &= \mathbf{v}_3 - \frac{\langle \mathbf{v}_3, \mathbf{f}_1 \rangle}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \frac{\langle \mathbf{v}_3, \mathbf{f}_2 \rangle}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 \\
\vdots & \qquad \vdots \\
\mathbf{f}_k &= \mathbf{v}_k - \frac{\langle \mathbf{v}_k, \mathbf{f}_1 \rangle}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \frac{\langle \mathbf{v}_k, \mathbf{f}_2 \rangle}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 - \cdots - \frac{\langle \mathbf{v}_k, \mathbf{f}_{k-1} \rangle}{\|\mathbf{f}_{k-1}\|^2}\mathbf{f}_{k-1}
\end{aligned}$$

for each $k = 2, 3, \dots, n$. Then

1. *$\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ is an orthogonal basis of V.*

2. $\text{span}\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_k\} = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ *holds for each $k = 1, 2, \dots, n$.*

The purpose of the Gram-Schmidt algorithm is to convert a basis of an inner product space into an *orthogonal* basis. In particular, it shows that every finite dimensional inner product space *has* an orthogonal basis.

Example 10.2.4

Consider $V = \mathbf{P}_3$ with the inner product $\langle p, q\rangle = \int_{-1}^{1} p(x)q(x)dx$. If the Gram-Schmidt algorithm is applied to the basis $\{1, x, x^2, x^3\}$, show that the result is the orthogonal basis

$$\{1, x, \tfrac{1}{3}(3x^2-1), \tfrac{1}{5}(5x^3-3x)\}$$

Solution. Take $\mathbf{f}_1 = 1$. Then the algorithm gives

$$\begin{aligned}
\mathbf{f}_2 &= x - \tfrac{\langle x, \mathbf{f}_1\rangle}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 = x - \tfrac{0}{2}\mathbf{f}_1 = x \\
\mathbf{f}_3 &= x^2 - \tfrac{\langle x^2, \mathbf{f}_1\rangle}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \tfrac{\langle x^2, \mathbf{f}_2\rangle}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 \\
&= x^2 - \tfrac{\frac{2}{3}}{2}1 - \tfrac{0}{\frac{2}{3}}x \\
&= \tfrac{1}{3}(3x^2-1)
\end{aligned}$$

The verification that $\mathbf{f}_4 = \frac{1}{5}(5x^3-3x)$ is omitted.

The polynomials in Example 10.2.4 are such that the leading coefficient is 1 in each case. In other contexts (the study of differential equations, for example) it is customary to take multiples $p(x)$ of these polynomials such that $p(1) = 1$. The resulting orthogonal basis of $\mathbf{P}_3$ is

$$\{1, x, \tfrac{1}{3}(3x^2-1), \tfrac{1}{5}(5x^3-3x)\}$$

and these are the first four **Legendre polynomials**, so called to honour the French mathematician A. M. Legendre (1752–1833). They are important in the study of differential equations.

If V is an inner product space of dimension n, let $E = \{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ be an orthonormal basis of V (by Theorem 10.2.5). If $\mathbf{v} = v_1\mathbf{f}_1 + v_2\mathbf{f}_2 + \cdots + v_n\mathbf{f}_n$ and $\mathbf{w} = w_1\mathbf{f}_1 + w_2\mathbf{f}_2 + \cdots + w_n\mathbf{f}_n$ are two vectors in V, we have $C_E(\mathbf{v}) = \left[\begin{array}{cccc} v_1 & v_2 & \cdots & v_n \end{array}\right]^T$ and $C_E(\mathbf{w}) = \left[\begin{array}{cccc} w_1 & w_2 & \cdots & w_n \end{array}\right]^T$. Hence

$$\langle \mathbf{v}, \mathbf{w}\rangle = \langle \sum_i v_i\mathbf{f}_i, \sum_j w_j\mathbf{f}_j\rangle = \sum_{i, j} v_i w_j \langle \mathbf{f}_i, \mathbf{f}_j\rangle = \sum_i v_i w_i = C_E(\mathbf{v}) \cdot C_E(\mathbf{w})$$

This shows that the coordinate isomorphism $C_E : V \to \mathbb{R}^n$ preserves inner products, and so proves

Corollary 10.2.1

If V is any n-dimensional inner product space, then V is isomorphic to $\mathbb{R}^n$ as inner product spaces. More precisely, if E is any orthonormal basis of V, the coordinate isomorphism

$$C_E : V \to \mathbb{R}^n \text{ satisfies } \langle \mathbf{v}, \mathbf{w}\rangle = C_E(\mathbf{v}) \cdot C_E(\mathbf{w})$$

for all $\mathbf{v}$ and $\mathbf{w}$ in V.

The orthogonal complement of a subspace U of $\mathbb{R}^n$ was defined (in Chapter 8) to be the set of all vectors in $\mathbb{R}^n$ that are orthogonal to every vector in U. This notion has a natural extension in an arbitrary inner product space. Let U be a subspace of an inner product space V. As in $\mathbb{R}^n$, the **orthogonal complement** $U^\perp$ of U in V is defined by

$$U^\perp = \{\mathbf{v} \mid \mathbf{v} \in V,\ \langle \mathbf{v},\ \mathbf{u}\rangle = 0 \text{ for all } \mathbf{u} \in U\}$$

Theorem 10.2.6

Let U be a finite dimensional subspace of an inner product space V.

1. *$U^\perp$ is a subspace of V and $V = U \oplus U^\perp$.*
2. *If $\dim V = n$, then $\dim U + \dim U^\perp = n$.*
3. *If $\dim V = n$, then $U^{\perp\perp} = U$.*

Proof.

1. $U^\perp$ is a subspace by Theorem 10.1.1. If $\mathbf{v}$ is in $U \cap U^\perp$, then $\langle \mathbf{v},\ \mathbf{v}\rangle = 0$, so $\mathbf{v} = \mathbf{0}$ again by Theorem 10.1.1. Hence $U \cap U^\perp = \{\mathbf{0}\}$, and it remains to show that $U + U^\perp = V$. Given $\mathbf{v}$ in V, we must show that $\mathbf{v}$ is in $U + U^\perp$, and this is clear if $\mathbf{v}$ is in U. If $\mathbf{v}$ is not in U, let $\{\mathbf{f}_1,\ \mathbf{f}_2,\ \dots,\ \mathbf{f}_m\}$ be an orthogonal basis of U. Then the orthogonal lemma shows that $\mathbf{v} - \left(\frac{\langle \mathbf{v},\ \mathbf{f}_1\rangle}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 + \frac{\langle \mathbf{v},\ \mathbf{f}_2\rangle}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 + \cdots + \frac{\langle \mathbf{v},\ \mathbf{f}_m\rangle}{\|\mathbf{f}_m\|^2}\mathbf{f}_m\right)$ is in $U^\perp$, so $\mathbf{v}$ is in $U + U^\perp$ as required.
2. This follows from Theorem 9.3.6.
3. We have $\dim U^{\perp\perp} = n - \dim U^\perp = n - (n - \dim U) = \dim U$, using (2) twice. As $U \subseteq U^{\perp\perp}$ always holds (verify), (3) follows by Theorem 6.4.2. □

We digress briefly and consider a subspace U of an arbitrary vector space V. As in Section 9.3, if W is any complement of U in V, that is, $V = U \oplus W$, then each vector $\mathbf{v}$ in V has a *unique* representation as a sum $\mathbf{v} = \mathbf{u} + \mathbf{w}$ where $\mathbf{u}$ is in U and $\mathbf{w}$ is in W. Hence we may define a function $T : V \to V$ as follows:

$$T(\mathbf{v}) = \mathbf{u} \quad \text{where } \mathbf{v} = \mathbf{u} + \mathbf{w},\ \ \mathbf{u} \text{ in } U,\ \ \mathbf{w} \text{ in } W$$

Thus, to compute $T(\mathbf{v})$, express $\mathbf{v}$ in any way at all as the sum of a vector $\mathbf{u}$ in U and a vector in W; then $T(\mathbf{v}) = \mathbf{u}$.

This function T is a linear operator on V. Indeed, if $\mathbf{v}_1 = \mathbf{u}_1 + \mathbf{w}_1$ where $\mathbf{u}_1$ is in U and $\mathbf{w}_1$ is in W, then $\mathbf{v} + \mathbf{v}_1 = (\mathbf{u} + \mathbf{u}_1) + (\mathbf{w} + \mathbf{w}_1)$ where $\mathbf{u} + \mathbf{u}_1$ is in U and $\mathbf{w} + \mathbf{w}_1$ is in W, so

$$T(\mathbf{v} + \mathbf{v}_1) = \mathbf{u} + \mathbf{u}_1 = T(\mathbf{v}) + T(\mathbf{v}_1)$$

Similarly, $T(a\mathbf{v}) = aT(\mathbf{v})$ for all a in $\mathbb{R}$, so T is a linear operator. Furthermore, $\operatorname{im} T = U$ and $\ker T = W$ as the reader can verify, and T is called the **projection on U with kernel W**.

If U is a subspace of V, there are many projections on U, one for each complementary subspace W with $V = U \oplus W$. If V is an *inner product space*, we single out one for special attention. Let U be a finite dimensional subspace of an inner product space V.

Definition 10.3 Orthogonal Projection on a Subspace

The projection on U with kernel $U^\perp$ is called the ***orthogonal projection*** *on U (or simply the* ***projection*** *on U) and is denoted $\operatorname{proj}_U : V \to V$.*

Theorem 10.2.7: Projection Theorem

Let U be a finite dimensional subspace of an inner product space V and let $\mathbf{v}$ be a vector in V.

1. $\operatorname{proj}_U : V \to V$ *is a linear operator with image U and kernel $U^\perp$.*
2. $\operatorname{proj}_U \mathbf{v}$ *is in U and $\mathbf{v} - \operatorname{proj}_U \mathbf{v}$ is in $U^\perp$.*
3. *If $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$ is any orthogonal basis of U, then*

$$\operatorname{proj}_U \mathbf{v} = \frac{\langle \mathbf{v}, \mathbf{f}_1 \rangle}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 + \frac{\langle \mathbf{v}, \mathbf{f}_2 \rangle}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 + \cdots + \frac{\langle \mathbf{v}, \mathbf{f}_m \rangle}{\|\mathbf{f}_m\|^2}\mathbf{f}_m$$

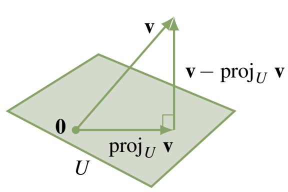
Proof. Only (3) remains to be proved. But since $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ is an orthogonal basis of U and since $\operatorname{proj}_U \mathbf{v}$ is in U, the result follows from the expansion theorem (Theorem 10.2.4) applied to the finite dimensional space U. □

Note that there is no requirement in Theorem 10.2.7 that V is finite dimensional.

Example 10.2.5

Let U be a subspace of the finite dimensional inner product space V. Show that $\operatorname{proj}_{U^\perp} \mathbf{v} = \mathbf{v} - \operatorname{proj}_U \mathbf{v}$ for all $\mathbf{v} \in V$.

Solution. We have $V = U^\perp \oplus U^{\perp\perp}$ by Theorem 10.2.6. If we write $\mathbf{p} = \operatorname{proj}_U \mathbf{v}$, then $\mathbf{v} = (\mathbf{v} - \mathbf{p}) + \mathbf{p}$ where $\mathbf{v} - \mathbf{p}$ is in $U^\perp$ and $\mathbf{p}$ is in $U = U^{\perp\perp}$ by Theorem 10.2.7. Hence $\operatorname{proj}_{U^\perp} \mathbf{v} = \mathbf{v} - \mathbf{p}$. See Exercise **??**.



The vectors $\mathbf{v}$, $\operatorname{proj}_U \mathbf{v}$, and $\mathbf{v} - \operatorname{proj}_U \mathbf{v}$ in Theorem 10.2.7 can be visualized geometrically as in the diagram (where U is shaded and $\dim U = 2$). This suggests that $\operatorname{proj}_U \mathbf{v}$ is the vector in U closest to $\mathbf{v}$. This is, in fact, the case.

Theorem 10.2.8: Approximation Theorem

Let U be a finite dimensional subspace of an inner product space V. If $\mathbf{v}$ is any vector in V, then $\operatorname{proj}_U \mathbf{v}$ is the vector in U that is closest to $\mathbf{v}$. Here ***closest*** *means that*

$$\|\mathbf{v} - \operatorname{proj}_U \mathbf{v}\| < \|\mathbf{v} - \mathbf{u}\|$$

for all $\mathbf{u}$ in U, $\mathbf{u} \neq \operatorname{proj}_U \mathbf{v}$.

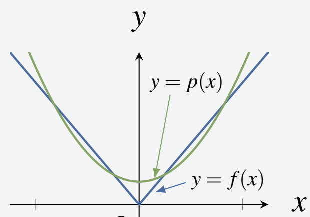
Proof. Write $\mathbf{p} = \text{proj}_U \mathbf{v}$, and consider $\mathbf{v} - \mathbf{u} = (\mathbf{v} - \mathbf{p}) + (\mathbf{p} - \mathbf{u})$. Because $\mathbf{v} - \mathbf{p}$ is in $U^\perp$ and $\mathbf{p} - \mathbf{u}$ is in U, Pythagoras' theorem gives

$$\|\mathbf{v} - \mathbf{u}\|^2 = \|\mathbf{v} - \mathbf{p}\|^2 + \|\mathbf{p} - \mathbf{u}\|^2 > \|\mathbf{v} - \mathbf{p}\|^2$$

because $\mathbf{p} - \mathbf{u} \neq 0$. The result follows. □

Example 10.2.6

Consider the space $\mathbf{C}[-1, 1]$ of real-valued continuous functions on the interval $[-1, 1]$ with inner product $\langle f, g \rangle = \int_{-1}^{1} f(x)g(x)dx$. Find the polynomial $p = p(x)$ of degree at most 2 that best approximates the absolute-value function f given by $f(x) = |x|$.



Solution. Here we want the vector p in the subspace $U = \mathbf{P}_2$ of $\mathbf{C}[-1, 1]$ that is closest to f. In Example 10.2.4 the Gram-Schmidt algorithm was applied to give an orthogonal basis $\{\mathbf{f}_1 = 1, \mathbf{f}_2 = x, \mathbf{f}_3 = 3x^2 - 1\}$ of $\mathbf{P}_2$ (where, for convenience, we have changed $\mathbf{f}_3$ by a numerical factor). Hence the required polynomial is

$$\begin{aligned} p &= \text{proj}_{\mathbf{P}_2} f \\ &= \tfrac{\langle f, \mathbf{f}_1 \rangle}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 + \tfrac{\langle f, \mathbf{f}_2 \rangle}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 + \tfrac{\langle f, \mathbf{f}_3 \rangle}{\|\mathbf{f}_3\|^2}\mathbf{f}_3 \\ &= \tfrac{1}{2}\mathbf{f}_1 + 0\mathbf{f}_2 + \tfrac{1/2}{8/5}\mathbf{f}_3 \\ &= \tfrac{3}{16}(5x^2 + 1) \end{aligned}$$

The graphs of $p(x)$ and $f(x)$ are given in the diagram.

If polynomials of degree at most n are allowed in Example 10.2.6, the polynomial in $\mathbf{P}_n$ is $\text{proj}_{\mathbf{P}_n} f$, and it is calculated in the same way. Because the subspaces $\mathbf{P}_n$ get larger as n increases, it turns out that the approximating polynomials $\text{proj}_{\mathbf{P}_n} f$ get closer and closer to f. In fact, solving many practical problems comes down to approximating some interesting vector $\mathbf{v}$ (often a function) in an infinite dimensional inner product space V by vectors in finite dimensional subspaces (which can be computed). If $U_1 \subseteq U_2$ are finite dimensional subspaces of V, then

$$\|\mathbf{v} - \text{proj}_{U_2} \mathbf{v}\| \leq \|\mathbf{v} - \text{proj}_{U_1} \mathbf{v}\|$$

by Theorem 10.2.8 (because $\text{proj}_{U_1} \mathbf{v}$ lies in U_1 and hence in U_2). Thus $\text{proj}_{U_2} \mathbf{v}$ is a better approximation to $\mathbf{v}$ than $\text{proj}_{U_1} \mathbf{v}$. Hence a general method in approximation theory might be described as follows: Given $\mathbf{v}$, use it to construct a sequence of finite dimensional subspaces

$$U_1 \subseteq U_2 \subseteq U_3 \subseteq \cdots$$

of V in such a way that $\|\mathbf{v} - \text{proj}_{U_k} \mathbf{v}\|$ approaches zero as k increases. Then $\text{proj}_{U_k} \mathbf{v}$ is a suitable approximation to $\mathbf{v}$ if k is large enough. For more information, the interested reader may wish to consult *Interpolation and Approximation* by Philip J. Davis (New York: Blaisdell, 1963).

10.3 Orthogonal Diagonalization

There is a natural way to define a symmetric linear operator T on a finite dimensional inner product space V. If T is such an operator, it is shown in this section that V has an orthogonal basis consisting of eigenvectors of T. This yields another proof of the principal axes theorem in the context of inner product spaces.

Theorem 10.3.1

Let $T : V \to V$ be a linear operator on a finite dimensional space V. Then the following conditions are equivalent.

1. *V has a basis consisting of eigenvectors of T.*
2. *There exists a basis B of V such that $M_B(T)$ is diagonal.*

Proof. We have $M_B(T) = \left[\begin{array}{cccc} C_B[T(\mathbf{b}_1)] & C_B[T(\mathbf{b}_2)] & \cdots & C_B[T(\mathbf{b}_n)] \end{array} \right]$ where $B = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ is any basis of V. By comparing columns:

$$M_B(T) = \left[\begin{array}{cccc} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{array} \right] \text{ if and only if } T(\mathbf{b}_i) = \lambda_i \mathbf{b}_i \text{ for each } i$$

Theorem 10.3.1 follows. □

Definition 10.4 Diagonalizable Linear Operators

A linear operator T on a finite dimensional space V is called ***diagonalizable*** *if V has a basis consisting of eigenvectors of T.*

Example 10.3.1

Let $T : \mathbf{P}_2 \to \mathbf{P}_2$ be given by

$$T(a+bx+cx^2) = (a+4c) - 2bx + (3a+2c)x^2$$

Find the eigenspaces of T and hence find a basis of eigenvectors.

Solution. If $B_0 = \{1,\ x,\ x^2\}$, then

$$M_{B_0}(T) = \begin{bmatrix} 1 & 0 & 4 \\ 0 & -2 & 0 \\ 3 & 0 & 2 \end{bmatrix}$$

so $c_T(x) = (x+2)^2(x-5)$, and the eigenvalues of T are $\lambda = -2$ and $\lambda = 5$. One sees that $\left\{ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 \\ 0 \\ -3 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \right\}$ is a basis of eigenvectors of $M_{B_0}(T)$, so $B = \{x,\ 4-3x^2,\ 1+x^2\}$ is a basis of $\mathbf{P}_2$ consisting of eigenvectors of T.

If V is an inner product space, the expansion theorem gives a simple formula for the matrix of a linear operator with respect to an orthogonal basis.

Theorem 10.3.2

Let $T : V \to V$ be a linear operator on an inner product space V. If $B = \{\mathbf{b}_1,\ \mathbf{b}_2,\ \dots,\ \mathbf{b}_n\}$ is an orthogonal basis of V, then

$$M_B(T) = \left[\frac{\langle \mathbf{b}_i,\ T(\mathbf{b}_j) \rangle}{\|\mathbf{b}_i\|^2} \right]$$

Proof. Write $M_B(T) = \left[a_{ij} \right]$. The jth column of $M_B(T)$ is $C_B[T(\mathbf{e}_j)]$, so

$$T(\mathbf{b}_j) = a_{1j}\mathbf{b}_1 + \cdots + a_{ij}\mathbf{b}_i + \cdots + a_{nj}\mathbf{b}_n$$

On the other hand, the expansion theorem (Theorem 10.2.4) gives

$$\mathbf{v} = \frac{\langle \mathbf{b}_1,\ \mathbf{v} \rangle}{\|\mathbf{b}_1\|^2}\mathbf{b}_1 + \cdots + \frac{\langle \mathbf{b}_i,\ \mathbf{v} \rangle}{\|\mathbf{b}_i\|^2}\mathbf{b}_i + \cdots + \frac{\langle \mathbf{b}_n,\ \mathbf{v} \rangle}{\|\mathbf{b}_n\|^2}\mathbf{b}_n$$

for any $\mathbf{v}$ in V. The result follows by taking $\mathbf{v} = T(\mathbf{b}_j)$. □

Example 10.3.2

Let $T : \mathbb{R}^3 \to \mathbb{R}^3$ be given by

$$T(a,\ b,\ c) = (a + 2b - c,\ 2a + 3c,\ -a + 3b + 2c)$$

If the dot product in $\mathbb{R}^3$ is used, find the matrix of T with respect to the standard basis $B = \{\mathbf{e}_1,\ \mathbf{e}_2,\ \mathbf{e}_3\}$ where $\mathbf{e}_1 = (1,\ 0,\ 0)$, $\mathbf{e}_2 = (0,\ 1,\ 0)$, $\mathbf{e}_3 = (0,\ 0,\ 1)$.

Solution. The basis B is orthonormal, so Theorem 10.3.2 gives

$$M_B(T) = \left[\begin{array}{ccc} \mathbf{e}_1 \cdot T(\mathbf{e}_1) & \mathbf{e}_1 \cdot T(\mathbf{e}_2) & \mathbf{e}_1 \cdot T(\mathbf{e}_3) \\ \mathbf{e}_2 \cdot T(\mathbf{e}_1) & \mathbf{e}_2 \cdot T(\mathbf{e}_2) & \mathbf{e}_2 \cdot T(\mathbf{e}_3) \\ \mathbf{e}_3 \cdot T(\mathbf{e}_1) & \mathbf{e}_3 \cdot T(\mathbf{e}_2) & \mathbf{e}_3 \cdot T(\mathbf{e}_3) \end{array} \right] = \left[\begin{array}{rrr} 1 & 2 & -1 \\ 2 & 0 & 3 \\ -1 & 3 & 2 \end{array} \right]$$

Of course, this can also be found in the usual way.

It is not difficult to verify that an $n \times n$ matrix A is symmetric if and only if $\mathbf{x} \cdot (A\mathbf{y}) = (A\mathbf{x}) \cdot \mathbf{y}$ holds for all columns $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$. The analog for operators is as follows:

Theorem 10.3.3

Let V be a finite dimensional inner product space. The following conditions are equivalent for a linear operator $T : V \to V$.

1. $\langle \mathbf{v},\ T(\mathbf{w}) \rangle = \langle T(\mathbf{v}),\ \mathbf{w} \rangle$ *for all* $\mathbf{v}$ *and* $\mathbf{w}$ *in* V.
2. *The matrix of T is symmetric with respect to every orthonormal basis of V.*
3. *The matrix of T is symmetric with respect to some orthonormal basis of V.*
4. *There is an orthonormal basis $B = \{\mathbf{f}_1,\ \mathbf{f}_2,\ \dots,\ \mathbf{f}_n\}$ of V such that $\langle \mathbf{f}_i,\ T(\mathbf{f}_j) \rangle = \langle T(\mathbf{f}_i),\ \mathbf{f}_j \rangle$ holds for all i and j.*

Proof. (1) $\Rightarrow$ (2). Let $B = \{\mathbf{f}_1,\ \dots,\ \mathbf{f}_n\}$ be an orthonormal basis of V, and write $M_B(T) = \left[a_{ij} \right]$. Then $a_{ij} = \langle \mathbf{f}_i,\ T(\mathbf{f}_j) \rangle$ by Theorem 10.3.2. Hence (1) and axiom P2 give

$$a_{ij} = \langle \mathbf{f}_i,\ T(\mathbf{f}_j) \rangle = \langle T(\mathbf{f}_i),\ \mathbf{f}_j \rangle = \langle \mathbf{f}_j,\ T(\mathbf{f}_i) \rangle = a_{ji}$$

for all i and j. This shows that $M_B(T)$ is symmetric.

(2) $\Rightarrow$ (3). This is clear.

(3) $\Rightarrow$ (4). Let $B = \{\mathbf{f}_1,\ \dots,\ \mathbf{f}_n\}$ be an orthonormal basis of V such that $M_B(T)$ is symmetric. By (3) and Theorem 10.3.2, $\langle \mathbf{f}_i,\ T(\mathbf{f}_j) \rangle = \langle \mathbf{f}_j,\ T(\mathbf{f}_i) \rangle$ for all i and j, so (4) follows from axiom P2.

(4) $\Rightarrow$ (1). Let $\mathbf{v}$ and $\mathbf{w}$ be vectors in V and write them as $\mathbf{v} = \sum_{i=1}^{n} v_i \mathbf{f}_i$ and $\mathbf{w} = \sum_{j=1}^{n} w_j \mathbf{f}_j$. Then

$$\langle \mathbf{v},\ T(\mathbf{w}) \rangle = \left\langle \sum_i v_i \mathbf{f}_i,\ \sum_j w_j T \mathbf{f}_j \right\rangle = \sum_i \sum_j v_i w_j \langle \mathbf{f}_i,\ T(\mathbf{f}_j) \rangle$$

$$= \sum_i \sum_j v_i w_j \langle T(\mathbf{f}_i),\, \mathbf{f}_j \rangle$$
$$= \left\langle \sum_i v_i T(\mathbf{f}_i),\, \sum_j w_j \mathbf{f}_j \right\rangle$$
$$= \langle T(\mathbf{v}),\, \mathbf{w} \rangle$$

where we used (4) at the third stage. This proves (1). □

A linear operator T on an inner product space V is called **symmetric** if $\langle \mathbf{v},\, T(\mathbf{w}) \rangle = \langle T(\mathbf{v}),\, \mathbf{w} \rangle$ holds for all $\mathbf{v}$ and $\mathbf{w}$ in V.

Example 10.3.3

If A is an $n \times n$ matrix, let $T_A : \mathbb{R}^n \to \mathbb{R}^n$ be the matrix operator given by $T_A(\mathbf{v}) = A\mathbf{v}$ for all columns $\mathbf{v}$. If the dot product is used in $\mathbb{R}^n$, then T_A is a symmetric operator if and only if A is a symmetric matrix.

Solution. If E is the standard basis of $\mathbb{R}^n$, then E is orthonormal when the dot product is used. We have $M_E(T_A) = A$ (by Example 9.1.4), so the result follows immediately from part (3) of Theorem 10.3.3.

It is important to note that whether an operator is symmetric depends on which inner product is being used (see Exercise **??**).

If V is a finite dimensional inner product space, the eigenvalues of an operator $T : V \to V$ are the same as those of $M_B(T)$ for any orthonormal basis B (see Theorem 9.3.3). If T is symmetric, $M_B(T)$ is a symmetric matrix and so has real eigenvalues by Theorem 5.5.7. Hence we have the following:

Theorem 10.3.4

A symmetric linear operator on a finite dimensional inner product space has real eigenvalues.

If U is a subspace of an inner product space V, recall that its orthogonal complement is the subspace $U^\perp$ of V defined by

$$U^\perp = \{\mathbf{v} \text{ in } V \mid \langle \mathbf{v},\, \mathbf{u} \rangle = 0 \text{ for all } \mathbf{u} \text{ in } U\}$$

Theorem 10.3.5

Let $T : V \to V$ be a symmetric linear operator on an inner product space V, and let U be a T-invariant subspace of V. Then:

1. *The restriction of T to U is a symmetric linear operator on U.*
2. *$U^\perp$ is also T-invariant.*

Proof.

1. U is itself an inner product space using the same inner product, and condition 1 in Theorem 10.3.3 that T is symmetric is clearly preserved.

2. If $\mathbf{v}$ is in $U^{\perp}$, our task is to show that $T(\mathbf{v})$ is also in $U^{\perp}$; that is, $\langle T(\mathbf{v}),\, \mathbf{u}\rangle = 0$ for all $\mathbf{u}$ in U. But if $\mathbf{u}$ is in U, then $T(\mathbf{u})$ also lies in U because U is T-invariant, so

$$\langle T(\mathbf{v}),\, \mathbf{u}\rangle = \langle \mathbf{v},\, T(\mathbf{u})\rangle$$

using the symmetry of T and the definition of $U^{\perp}$. □

The principal axes theorem (Theorem 8.2.2) asserts that an $n \times n$ matrix A is symmetric if and only if $\mathbb{R}^n$ has an orthogonal basis of eigenvectors of A. The following result not only extends this theorem to an arbitrary n-dimensional inner product space, but the proof is much more intuitive.

Theorem 10.3.6: Principal Axes Theorem

The following conditions are equivalent for a linear operator T on a finite dimensional inner product space V.

1. *T is symmetric.*
2. *V has an orthogonal basis consisting of eigenvectors of T.*

Proof. (1) $\Rightarrow$ (2). Assume that T is symmetric and proceed by induction on $n = \dim V$. If $n = 1$, *every* nonzero vector in V is an eigenvector of T, so there is nothing to prove. If $n \geq 2$, assume inductively that the theorem holds for spaces of dimension less than n. Let λ_1 be a real eigenvalue of T (by Theorem 10.3.4) and choose an eigenvector $\mathbf{f}_1$ corresponding to λ_1. Then $U = \mathbb{R}\mathbf{f}_1$ is T-invariant, so $U^{\perp}$ is also T-invariant by Theorem 10.3.5 (T is symmetric). Because $\dim U^{\perp} = n-1$ (Theorem 10.2.6), and because the restriction of T to $U^{\perp}$ is a symmetric operator (Theorem 10.3.5), it follows by induction that $U^{\perp}$ has an orthogonal basis $\{\mathbf{f}_2,\, \dots,\, \mathbf{f}_n\}$ of eigenvectors of T. Hence $B = \{\mathbf{f}_1,\, \mathbf{f}_2,\, \dots,\, \mathbf{f}_n\}$ is an orthogonal basis of V, which proves (2).

(2) $\Rightarrow$ (1). If $B = \{\mathbf{f}_1,\, \dots,\, \mathbf{f}_n\}$ is a basis as in (2), then $M_B(T)$ is symmetric (indeed diagonal), so T is symmetric by Theorem 10.3.3. □

The matrix version of the principal axes theorem is an immediate consequence of Theorem 10.3.6. If A is an $n \times n$ symmetric matrix, then $T_A : \mathbb{R}^n \to \mathbb{R}^n$ is a symmetric operator, so let B be an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of T_A (and hence of A). Then P^TAP is diagonal where P is the orthogonal matrix whose columns are the vectors in B (see Theorem 9.2.4).

Similarly, let $T : V \to V$ be a symmetric linear operator on the n-dimensional inner product space V and let B_0 be any convenient orthonormal basis of V. Then an orthonormal basis of eigenvectors of T can be computed from $M_{B_0}(T)$. In fact, if $P^T M_{B_0}(T)P$ is diagonal where P is orthogonal, let $B = \{\mathbf{f}_1,\, \dots,\, \mathbf{f}_n\}$ be the vectors in V such that $C_{B_0}(\mathbf{f}_j)$ is column j of P for each j. Then B consists of eigenvectors of T by Theorem 9.3.3, and they are orthonormal because B_0 is orthonormal. Indeed

$$\langle \mathbf{f}_i,\, \mathbf{f}_j\rangle = C_{B_0}(\mathbf{f}_i) \cdot C_{B_0}(\mathbf{f}_j)$$

holds for all i and j, as the reader can verify. Here is an example.

Example 10.3.4

Let $T : \mathbf{P}_2 \to \mathbf{P}_2$ be given by

$$T(a+bx+cx^2) = (8a-2b+2c)+(-2a+5b+4c)x+(2a+4b+5c)x^2$$

Using the inner product $\langle a+bx+cx^2,\ a'+b'x+c'x^2\rangle = aa'+bb'+cc'$, show that T is symmetric and find an orthonormal basis of $\mathbf{P}_2$ consisting of eigenvectors.

Solution. If $B_0 = \{1,\ x,\ x^2\}$, then $M_{B_0}(T) = \left[\begin{array}{rrr} 8 & -2 & 2 \\ -2 & 5 & 4 \\ 2 & 4 & 5 \end{array}\right]$ is symmetric, so T is symmetric.

This matrix was analyzed in Example 8.2.5, where it was found that an *orthonormal* basis of eigenvectors is $\left\{\frac{1}{3}\left[\begin{array}{ccc} 1 & 2 & -2 \end{array}\right]^T,\ \frac{1}{3}\left[\begin{array}{ccc} 2 & 1 & 2 \end{array}\right]^T,\ \frac{1}{3}\left[\begin{array}{ccc} -2 & 2 & 1 \end{array}\right]^T\right\}$. Because B_0 is orthonormal, the corresponding orthonormal basis of $\mathbf{P}_2$ is

$$B = \left\{\tfrac{1}{3}(1+2x-2x^2),\ \tfrac{1}{3}(2+x+2x^2),\ \tfrac{1}{3}(-2+2x+x^2)\right\}$$

10.4 Isometries

We saw in Section 2.6 that rotations about the origin and reflections in a line through the origin are linear operators on $\mathbb{R}^2$. Similar geometric arguments (in Section 4.4) establish that, in $\mathbb{R}^3$, rotations about a line through the origin and reflections in a plane through the origin are linear. We are going to give an algebraic proof of these results that is valid in any inner product space. The key observation is that reflections and rotations are distance preserving in the following sense. If V is an inner product space, a transformation $S : V \to V$ (not necessarily linear) is said to be **distance preserving** if the distance between $S(\mathbf{v})$ and $S(\mathbf{w})$ is the same as the distance between $\mathbf{v}$ and $\mathbf{w}$ for all vectors $\mathbf{v}$ and $\mathbf{w}$; more formally, if

$$\|S(\mathbf{v}) - S(\mathbf{w})\| = \|\mathbf{v} - \mathbf{w}\| \quad \text{for all } \mathbf{v} \text{ and } \mathbf{w} \text{ in } V \tag{10.2}$$

Distance-preserving maps need not be linear. For example, if $\mathbf{u}$ is any vector in V, the transformation $S_{\mathbf{u}} : V \to V$ defined by $S_{\mathbf{u}}(\mathbf{v}) = \mathbf{v} + \mathbf{u}$ for all $\mathbf{v}$ in V is called **translation** by $\mathbf{u}$, and it is routine to verify that $S_{\mathbf{u}}$ is distance preserving for any $\mathbf{u}$. However, $S_{\mathbf{u}}$ is linear only if $\mathbf{u} = \mathbf{0}$ (since then $S_{\mathbf{u}}(\mathbf{0}) = \mathbf{0}$). Remarkably, distance-preserving operators that do fix the origin are necessarily linear.

Lemma 10.4.1

Let V be an inner product space of dimension n, and consider a distance-preserving transformation $S : V \to V$. If $S(\mathbf{0}) = \mathbf{0}$, then S is linear.

Proof. We have $\|S(\mathbf{v}) - S(\mathbf{w})\|^2 = \|\mathbf{v} - \mathbf{w}\|^2$ for all $\mathbf{v}$ and $\mathbf{w}$ in V by (10.2), which gives

$$\langle S(\mathbf{v}),\, S(\mathbf{w})\rangle = \langle \mathbf{v},\, \mathbf{w}\rangle \qquad \text{for all } \mathbf{v} \text{ and } \mathbf{w} \text{ in } V \tag{10.3}$$

Now let $\{\mathbf{f}_1,\, \mathbf{f}_2,\, \dots,\, \mathbf{f}_n\}$ be an orthonormal basis of V. Then $\{S(\mathbf{f}_1),\, S(\mathbf{f}_2),\, \dots,\, S(\mathbf{f}_n)\}$ is orthonormal by (10.3) and so is a basis because $\dim V = n$. Now compute:

$$\begin{aligned}
\langle S(\mathbf{v}+\mathbf{w}) - S(\mathbf{v}) - S(\mathbf{w}),\, S(\mathbf{f}_i)\rangle &= \langle S(\mathbf{v}+\mathbf{w}),\, S(\mathbf{f}_i)\rangle - \langle S(\mathbf{v}),\, S(\mathbf{f}_i)\rangle - \langle S(\mathbf{w}),\, S(\mathbf{f}_i)\rangle \\
&= \langle \mathbf{v}+\mathbf{w},\, \mathbf{f}_i\rangle - \langle \mathbf{v},\, \mathbf{f}_i\rangle - \langle \mathbf{w},\, \mathbf{f}_i\rangle \\
&= 0
\end{aligned}$$

for each i. It follows from the expansion theorem (Theorem 10.2.4) that $S(\mathbf{v}+\mathbf{w}) - S(\mathbf{v}) - S(\mathbf{w}) = 0$; that is, $S(\mathbf{v}+\mathbf{w}) = S(\mathbf{v}) + S(\mathbf{w})$. A similar argument shows that $S(a\mathbf{v}) = aS(\mathbf{v})$ holds for all a in $\mathbb{R}$ and $\mathbf{v}$ in V, so S is linear after all. □

Definition 10.5 Isometries

Distance-preserving linear operators are called ***isometries****.*

It is routine to verify that the composite of two distance-preserving transformations is again distance preserving. In particular the composite of a translation and an isometry is distance preserving. Surprisingly, the converse is true.

Theorem 10.4.1

If V is a finite dimensional inner product space, then every distance-preserving transformation $S : V \to V$ is the composite of a translation and an isometry.

Proof. If $S : V \to V$ is distance preserving, write $S(\mathbf{0}) = \mathbf{u}$ and define $T : V \to V$ by $T(\mathbf{v}) = S(\mathbf{v}) - \mathbf{u}$ for all $\mathbf{v}$ in V. Then $\|T(\mathbf{v}) - T(\mathbf{w})\| = \|\mathbf{v} - \mathbf{w}\|$ for all vectors $\mathbf{v}$ and $\mathbf{w}$ in V as the reader can verify; that is, T is distance preserving. Clearly, $T(\mathbf{0}) = \mathbf{0}$, so it is an isometry by Lemma 10.4.1. Since

$$S(\mathbf{v}) = \mathbf{u} + T(\mathbf{v}) = (S_{\mathbf{u}} \circ T)(\mathbf{v}) \quad \text{for all } \mathbf{v} \text{ in } V$$

we have $S = S_{\mathbf{u}} \circ T$, and the theorem is proved. □

In Theorem 10.4.1, $S = S_{\mathbf{u}} \circ T$ factors as the composite of an isometry T followed by a translation $S_{\mathbf{u}}$. More is true: this factorization is unique in that $\mathbf{u}$ and T are uniquely determined by S; and $\mathbf{w} \in V$ exists such that $S = T \circ S_{\mathbf{w}}$ is uniquely the composite of translation by $\mathbf{w}$ followed by the same isometry T (Exercise **??**).

Theorem 10.4.1 focuses our attention on the isometries, and the next theorem shows that, while they preserve distance, they are characterized as those operators that preserve other properties.

Theorem 10.4.2

Let $T : V \to V$ be a linear operator on a finite dimensional inner product space V.
The following conditions are equivalent:

1. *T is an isometry.* *(T preserves distance)*
2. *$\|T(\mathbf{v})\| = \|\mathbf{v}\|$ for all $\mathbf{v}$ in V.* *(T preserves norms)*
3. *$\langle T(\mathbf{v}), T(\mathbf{w})\rangle = \langle \mathbf{v}, \mathbf{w}\rangle$ for all $\mathbf{v}$ and $\mathbf{w}$ in V.* *(T preserves inner products)*
4. *If $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ is an orthonormal basis of V, then $\{T(\mathbf{f}_1), T(\mathbf{f}_2), \dots, T(\mathbf{f}_n)\}$ is also an orthonormal basis.* *(T preserves orthonormal bases)*
5. *T carries some orthonormal basis to an orthonormal basis.*

Proof. (1) $\Rightarrow$ (2). Take $\mathbf{w} = \mathbf{0}$ in (10.2).

(2) $\Rightarrow$ (3). Since T is linear, (2) gives $\|T(\mathbf{v}) - T(\mathbf{w})\|^2 = \|T(\mathbf{v} - \mathbf{w})\|^2 = \|\mathbf{v} - \mathbf{w}\|^2$. Now (3) follows.

(3) $\Rightarrow$ (4). By (3), $\{T(\mathbf{f}_1), T(\mathbf{f}_2), \dots, T(\mathbf{f}_n)\}$ is orthogonal and $\|T(\mathbf{f}_i)\|^2 = \|\mathbf{f}_i\|^2 = 1$. Hence it is a basis because $\dim V = n$.

(4) $\Rightarrow$ (5). This needs no proof.

(5) $\Rightarrow$ (1). By (5), let $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$ be an orthonormal basis of V such that $\{T(\mathbf{f}_1), \dots, T(\mathbf{f}_n)\}$ is also orthonormal. Given $\mathbf{v} = v_1\mathbf{f}_1 + \cdots + v_n\mathbf{f}_n$ in V, we have $T(\mathbf{v}) = v_1T(\mathbf{f}_1) + \cdots + v_nT(\mathbf{f}_n)$ so Pythagoras' theorem gives

$$\|T(\mathbf{v})\|^2 = v_1^2 + \cdots + v_n^2 = \|\mathbf{v}\|^2$$

Hence $\|T(\mathbf{v})\| = \|\mathbf{v}\|$ for all $\mathbf{v}$, and (1) follows by replacing $\mathbf{v}$ by $\mathbf{v} - \mathbf{w}$. □

Before giving examples, we note some consequences of Theorem 10.4.2.

Corollary 10.4.1

Let V be a finite dimensional inner product space.

1. *Every isometry of V is an isomorphism.*[5]
2. a. *$1_V : V \to V$ is an isometry.*
   b. *The composite of two isometries of V is an isometry.*
   c. *The inverse of an isometry of V is an isometry.*

Proof. (1) is by (4) of Theorem 10.4.2 and Theorem 7.3.1. (2a) is clear, and (2b) is left to the reader. If $T : V \to V$ is an isometry and $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$ is an orthonormal basis of V, then (2c) follows because T^{-1} carries the orthonormal basis $\{T(\mathbf{f}_1), \dots, T(\mathbf{f}_n)\}$ back to $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$. □

The conditions in part (2) of the corollary assert that the set of isometries of a finite dimensional inner product space forms an algebraic system called a **group**. The theory of groups is well developed, and groups of operators are important in geometry. In fact, geometry itself can be fruitfully viewed as the study of those properties of a vector space that are preserved by a group of invertible linear operators.

Example 10.4.1

Rotations of $\mathbb{R}^2$ about the origin are isometries, as are reflections in lines through the origin: They clearly preserve distance and so are linear by Lemma 10.4.1. Similarly, rotations about lines through the origin and reflections in planes through the origin are isometries of $\mathbb{R}^3$.

Example 10.4.2

Let $T : \mathbf{M}_{nn} \to \mathbf{M}_{nn}$ be the transposition operator: $T(A) = A^T$. Then T is an isometry if the inner product is $\langle A, B\rangle = \operatorname{tr}(AB^T) = \sum\limits_{i,\, j} a_{ij}b_{ij}$. In fact, T permutes the basis consisting of all matrices with one entry 1 and the other entries 0.

The proof of the next result requires the fact (see Theorem 10.4.2) that, if B is an orthonormal basis, then $\langle \mathbf{v}, \mathbf{w}\rangle = C_B(\mathbf{v}) \cdot C_B(\mathbf{w})$ for all vectors $\mathbf{v}$ and $\mathbf{w}$.

Theorem 10.4.3

Let $T : V \to V$ be an operator where V is a finite dimensional inner product space. The following conditions are equivalent.

1. *T is an isometry.*
2. *$M_B(T)$ is an orthogonal matrix for every orthonormal basis B.*
3. *$M_B(T)$ is an orthogonal matrix for some orthonormal basis B.*

[5] V must be finite dimensional—see Exercise **??**.

Proof. (1) $\Rightarrow$ (2). Let $B = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ be an orthonormal basis. Then the jth column of $M_B(T)$ is $C_B[T(\mathbf{e}_j)]$, and we have

$$C_B[T(\mathbf{e}_j)] \cdot C_B[T(\mathbf{e}_k)] = \langle T(\mathbf{e}_j), T(\mathbf{e}_k) \rangle = \langle \mathbf{e}_j, \mathbf{e}_k \rangle$$

using (1). Hence the columns of $M_B(T)$ are orthonormal in $\mathbb{R}^n$, which proves (2).

(2) $\Rightarrow$ (3). This is clear.

(3) $\Rightarrow$ (1). Let $B = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ be as in (3). Then, as before,

$$\langle T(\mathbf{e}_j), T(\mathbf{e}_k) \rangle = C_B[T(\mathbf{e}_j)] \cdot C_B[T(\mathbf{e}_k)]$$

so $\{T(\mathbf{e}_1), \ldots, T(\mathbf{e}_n)\}$ is orthonormal by (3). Hence Theorem 10.4.2 gives (1). □

It is important that B is *orthonormal* in Theorem 10.4.3. For example, $T : V \to V$ given by $T(\mathbf{v}) = 2\mathbf{v}$ preserves *orthogonal* sets but is not an isometry, as is easily checked.

If P is an orthogonal square matrix, then $P^{-1} = P^T$. Taking determinants yields $(\det P)^2 = 1$, so $\det P = \pm 1$. Hence:

Corollary 10.4.2

If $T : V \to V$ is an isometry where V is a finite dimensional inner product space, then $\det T = \pm 1$.

Example 10.4.3

If A is any $n \times n$ matrix, the matrix operator $T_A : \mathbb{R}^n \to \mathbb{R}^n$ is an isometry if and only if A is orthogonal using the dot product in $\mathbb{R}^n$. Indeed, if E is the standard basis of $\mathbb{R}^n$, then $M_E(T_A) = A$ by Theorem 9.2.4.

Rotations and reflections that fix the origin are isometries in $\mathbb{R}^2$ and $\mathbb{R}^3$ (Example 10.4.1); we are going to show that these isometries (and compositions of them in $\mathbb{R}^3$) are the only possibilities. In fact, this will follow from a general structure theorem for isometries. Surprisingly enough, much of the work involves the two–dimensional case.

Theorem 10.4.4

Let $T : V \to V$ be an isometry on the two-dimensional inner product space V. Then there are two possibilities.

Either (1) *There is an orthonormal basis B of V such that*

$$M_B(T) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \quad 0 \le \theta < 2\pi$$

or (2) *There is an orthonormal basis B of V such that*

$$M_B(T) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

Furthermore, type (1) *occurs if and only if* $\det T = 1$, *and type* (2) *occurs if and only if* $\det T = -1$.

Proof. The final statement follows from the rest because $\det T = \det[M_B(T)]$ for any basis B. Let $B_0 = \{\mathbf{e}_1,\ \mathbf{e}_2\}$ be any ordered orthonormal basis of V and write

$$A = M_{B_0}(T) = \left[\begin{array}{cc} a & b \\ c & d \end{array}\right];\ \text{that is,}\ \begin{array}{l} T(\mathbf{e}_1) = a\mathbf{e}_1 + c\mathbf{e}_2 \\ T(\mathbf{e}_2) = b\mathbf{e}_1 + d\mathbf{e}_2 \end{array}$$

Then A is orthogonal by Theorem 10.4.3, so its columns (and rows) are orthonormal. Hence

$$a^2 + c^2 = 1 = b^2 + d^2$$

so $(a,\ c)$ and $(d,\ b)$ lie on the unit circle. Thus angles θ and φ exist such that

$$\begin{array}{lll} a = \cos\theta, & c = \sin\theta & 0 \leq \theta < 2\pi \\ d = \cos\varphi, & b = \sin\varphi & 0 \leq \varphi < 2\pi \end{array}$$

Then $\sin(\theta + \varphi) = cd + ab = 0$ because the columns of A are orthogonal, so $\theta + \varphi = k\pi$ for some integer k. This gives $d = \cos(k\pi - \theta) = (-1)^k \cos\theta$ and $b = \sin(k\pi - \theta) = (-1)^{k+1} \sin\theta$. Finally

$$A = \left[\begin{array}{cc} \cos\theta & (-1)^{k+1}\sin\theta \\ \sin\theta & (-1)^k\cos\theta \end{array}\right]$$

If k is even we are in type (1) with $B = B_0$, so assume k is odd. Then $A = \left[\begin{array}{cc} a & c \\ c & -a \end{array}\right]$. If $a = -1$ and $c = 0$, we are in type (1) with $B = \{\mathbf{e}_2,\ \mathbf{e}_2\}$. Otherwise A has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = -1$ with corresponding eigenvectors $\mathbf{x}_1 = \left[\begin{array}{c} 1+a \\ c \end{array}\right]$ and $\mathbf{x}_2 = \left[\begin{array}{c} -c \\ 1+a \end{array}\right]$ as the reader can verify. Write

$$\mathbf{f}_1 = (1+a)\mathbf{e}_1 + c\mathbf{e}_2 \qquad \text{and} \qquad \mathbf{f}_2 = -c\mathbf{e}_2 + (1+a)\mathbf{e}_2$$

Then $\mathbf{f}_1$ and $\mathbf{f}_2$ are orthogonal (verify) and $C_{B_0}(\mathbf{f}_i) = C_{B_0}(\lambda_i\mathbf{f}_i) = \mathbf{x}_i$ for each i. Moreover

$$C_{B_0}[T(\mathbf{f}_i)] = AC_{B_0}(\mathbf{f}_i) = A\mathbf{x}_i = \lambda_i\mathbf{x}_i = \lambda_i C_{B_0}(\mathbf{f}_i) = C_{B_0}(\lambda_i\mathbf{f}_i)$$

so $T(\mathbf{f}_i) = \lambda_i\mathbf{f}_i$ for each i. Hence $M_B(T) = \left[\begin{array}{cc} \lambda_1 & 0 \\ 0 & \lambda_2 \end{array}\right] = \left[\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right]$ and we are in type (2) with $B = \left\{\frac{1}{\|\mathbf{f}_1\|}\mathbf{f}_1,\ \frac{1}{\|\mathbf{f}_2\|}\mathbf{f}_2\right\}$. □

Corollary 10.4.3

An operator $T : \mathbb{R}^2 \to \mathbb{R}^2$ is an isometry if and only if T is a rotation or a reflection.

In fact, if E is the standard basis of $\mathbb{R}^2$, then the clockwise rotation R_θ about the origin through an angle θ has matrix

$$M_E(R_\theta) = \left[\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right]$$

(see Theorem 2.6.4). On the other hand, if $S : \mathbb{R}^2 \to \mathbb{R}^2$ is the reflection in a line through the origin (called the **fixed line** of the reflection), let $\mathbf{f}_1$ be a unit vector pointing along the fixed line and let $\mathbf{f}_2$ be a unit vector perpendicular to the fixed line. Then $B = \{\mathbf{f}_1,\ \mathbf{f}_2\}$ is an orthonormal basis, $S(\mathbf{f}_1) = \mathbf{f}_1$ and $S(\mathbf{f}_2) = -\mathbf{f}_2$, so

$$M_B(S) = \left[\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right]$$

Thus S is of type 2. Note that, in this case, 1 is an eigenvalue of S, and any eigenvector corresponding to 1 is a direction vector for the fixed line.

Example 10.4.4

In each case, determine whether $T_A : \mathbb{R}^2 \to \mathbb{R}^2$ is a rotation or a reflection, and then find the angle or fixed line:

$$\text{(a) } A = \frac{1}{2}\left[\begin{array}{rr} 1 & \sqrt{3} \\ -\sqrt{3} & 1 \end{array}\right] \qquad \text{(b) } A = \frac{1}{5}\left[\begin{array}{rr} -3 & 4 \\ 4 & 3 \end{array}\right]$$

Solution. Both matrices are orthogonal, so (because $M_E(T_A) = A$, where E is the standard basis) T_A is an isometry in both cases. In the first case, $\det A = 1$, so T_A is a counterclockwise rotation through θ, where $\cos\theta = \frac{1}{2}$ and $\sin\theta = -\frac{\sqrt{3}}{2}$. Thus $\theta = -\frac{\pi}{3}$. In (b), $\det A = -1$, so T_A is a reflection in this case. We verify that $\mathbf{d} = \left[\begin{array}{r} 1 \\ 2 \end{array}\right]$ is an eigenvector corresponding to the eigenvalue 1. Hence the fixed line $\mathbb{R}\mathbf{d}$ has equation $y = 2x$.

We now give a structure theorem for isometries. The proof requires three preliminary results, each of interest in its own right.

Lemma 10.4.2

Let $T : V \to V$ be an isometry of a finite dimensional inner product space V. If U is a T-invariant subspace of V, then $U^\perp$ is also T-invariant.

Proof. Let $\mathbf{w}$ lie in $U^\perp$. We are to prove that $T(\mathbf{w})$ is also in $U^\perp$; that is, $\langle T(\mathbf{w}), \mathbf{u}\rangle = 0$ for all $\mathbf{u}$ in U. At this point, observe that the restriction of T to U is an isometry $U \to U$ and so is an isomorphism by the corollary to Theorem 10.4.2. In particular, each $\mathbf{u}$ in U can be written in the form $\mathbf{u} = T(\mathbf{u}_1)$ for some $\mathbf{u}_1$ in U, so

$$\langle T(\mathbf{w}), \mathbf{u}\rangle = \langle T(\mathbf{w}), T(\mathbf{u}_1)\rangle = \langle \mathbf{w}, \mathbf{u}_1\rangle = 0$$

because $\mathbf{w}$ is in $U^\perp$. This is what we wanted. □

To employ Lemma 10.4.2 above to analyze an isometry $T : V \to V$ when $\dim V = n$, it is necessary to show that a T-invariant subspace U exists such that $U \neq 0$ and $U \neq V$. We will show, in fact, that such a subspace U can always be found of dimension 1 or 2. If T has a real eigenvalue λ then $\mathbb{R}\mathbf{u}$ is T-invariant where $\mathbf{u}$ is any λ-eigenvector. But, in case (1) of Theorem 10.4.4, the eigenvalues of T are $e^{i\theta}$ and $e^{-i\theta}$ (the reader should check this), and these are nonreal if $\theta \neq 0$ and $\theta \neq \pi$. It turns out that every complex eigenvalue λ of T has absolute value 1 (Lemma 10.4.3 below); and that U has a T-invariant subspace of dimension 2 if λ is not real (Lemma 10.4.4).

Lemma 10.4.3

Let $T : V \to V$ be an isometry of the finite dimensional inner product space V. If λ is a complex eigenvalue of T, then $|\lambda| = 1$.

Proof. Choose an orthonormal basis B of V, and let $A = M_B(T)$. Then A is a real orthogonal matrix so, using the standard inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \overline{\mathbf{y}}$ in $\mathbb{C}$, we get

$$||A\mathbf{x}||^2 = (A\mathbf{x})^T(\overline{A\mathbf{x}}) = \mathbf{x}^T A^T \overline{A\mathbf{x}} = \mathbf{x}^T I \mathbf{x} = ||\mathbf{x}||^2$$

for all $\mathbf{x}$ in $\mathbb{C}^n$. But $A\mathbf{x} = \lambda\mathbf{x}$ for some $\mathbf{x} \neq \mathbf{0}$, whence $||\mathbf{x}||^2 = ||\lambda\mathbf{x}||^2 = |\lambda|^2||\mathbf{x}||^2$. This gives $|\lambda| = 1$, as required. □

Lemma 10.4.4

Let $T : V \to V$ be an isometry of the n-dimensional inner product space V. If T has a nonreal eigenvalue, then V has a two-dimensional T-invariant subspace.

Proof. Let B be an orthonormal basis of V, let $A = M_B(T)$, and (using Lemma 10.4.3) let $\lambda = e^{i\alpha}$ be a nonreal eigenvalue of A, say $A\mathbf{x} = \lambda\mathbf{x}$ where $\mathbf{x} \neq \mathbf{0}$ in $\mathbb{C}^n$. Because A is real, complex conjugation gives $A\overline{\mathbf{x}} = \overline{\lambda}\overline{\mathbf{x}}$, so $\overline{\lambda}$ is also an eigenvalue. Moreover $\lambda \neq \overline{\lambda}$ (λ is nonreal), so $\{\mathbf{x}, \overline{\mathbf{x}}\}$ is linearly independent in $\mathbb{C}^n$ (the argument in the proof of Theorem 5.5.4 works). Now define

$$\mathbf{z}_1 = \mathbf{x} + \overline{\mathbf{x}} \quad \text{and} \quad \mathbf{z}_2 = i(\mathbf{x} - \overline{\mathbf{x}})$$

Then $\mathbf{z}_1$ and $\mathbf{z}_2$ lie in $\mathbb{R}^n$, and $\{\mathbf{z}_1, \mathbf{z}_2\}$ is linearly independent over $\mathbb{R}$ because $\{\mathbf{x}, \overline{\mathbf{x}}\}$ is linearly independent over $\mathbb{C}$. Moreover

$$\mathbf{x} = \tfrac{1}{2}(\mathbf{z}_1 - i\mathbf{z}_2) \quad \text{and} \quad \overline{\mathbf{x}} = \tfrac{1}{2}(\mathbf{z}_1 + i\mathbf{z}_2)$$

Now $\lambda + \overline{\lambda} = 2\cos\alpha$ and $\lambda - \overline{\lambda} = 2i\sin\alpha$, and a routine computation gives

$$\begin{aligned} A\mathbf{z}_1 &= \mathbf{z}_1\cos\alpha + \mathbf{z}_2\sin\alpha \\ A\mathbf{z}_2 &= -\mathbf{z}_1\sin\alpha + \mathbf{z}_2\cos\alpha \end{aligned}$$

Finally, let $\mathbf{e}_1$ and $\mathbf{e}_2$ in V be such that $\mathbf{z}_1 = C_B(\mathbf{e}_1)$ and $\mathbf{z}_2 = C_B(\mathbf{e}_2)$. Then

$$C_B[T(\mathbf{e}_1)] = AC_B(\mathbf{e}_1) = A\mathbf{z}_1 = C_B(\mathbf{e}_1\cos\alpha + \mathbf{e}_2\sin\alpha)$$

using Theorem 9.1.2. Because C_B is one-to-one, this gives the first of the following equations (the other is similar):

$$\begin{aligned} T(\mathbf{e}_1) &= \mathbf{e}_1\cos\alpha + \mathbf{e}_2\sin\alpha \\ T(\mathbf{e}_2) &= -\mathbf{e}_1\sin\alpha + \mathbf{e}_2\cos\alpha \end{aligned}$$

Thus $U = \text{span}\{\mathbf{e}_1, \mathbf{e}_2\}$ is T-invariant and two-dimensional. □

We can now prove the structure theorem for isometries.

Theorem 10.4.5

Let $T : V \to V$ be an isometry of the n-dimensional inner product space V. Given an angle θ, write $R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$. Then there exists an orthonormal basis B of V such that $M_B(T)$ has *one of the following block diagonal forms, classified for convenience by whether n is odd or even:*

$$n = 2k+1 \quad \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & R(\theta_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R(\theta_k) \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} -1 & 0 & \cdots & 0 \\ 0 & R(\theta_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R(\theta_k) \end{bmatrix}$$

$$n = 2k \quad \begin{bmatrix} R(\theta_1) & 0 & \cdots & 0 \\ 0 & R(\theta_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R(\theta_k) \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & R(\theta_1) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & R(\theta_{k-1}) \end{bmatrix}$$

Proof. We show first, by induction on n, that an orthonormal basis B of V can be found such that $M_B(T)$ is a block diagonal matrix of the following form:

$$M_B(T) = \begin{bmatrix} I_r & 0 & 0 & \cdots & 0 \\ 0 & -I_s & 0 & \cdots & 0 \\ 0 & 0 & R(\theta_1) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & R(\theta_t) \end{bmatrix}$$

where the identity matrix I_r, the matrix $-I_s$, or the matrices $R(\theta_i)$ may be missing. If $n = 1$ and $V = \mathbb{R}\mathbf{v}$, this holds because $T(\mathbf{v}) = \lambda\mathbf{v}$ and $\lambda = \pm 1$ by Lemma 10.4.3. If $n = 2$, this follows from Theorem 10.4.4. If $n \geq 3$, either T has a real eigenvalue and therefore has a one-dimensional T-invariant subspace $U = \mathbb{R}\mathbf{u}$ for any eigenvector $\mathbf{u}$, or T has no real eigenvalue and therefore has a two-dimensional T-invariant subspace U by Lemma 10.4.4. In either case $U^\perp$ is T-invariant (Lemma 10.4.2) and $\dim U^\perp = n - \dim U < n$. Hence, by induction, let B_1 and B_2 be orthonormal bases of U and $U^\perp$ such that $M_{B_1}(T)$ and $M_{B_2}(T)$ have the form given. Then $B = B_1 \cup B_2$ is an orthonormal basis of V, and $M_B(T)$ has the desired form with a suitable ordering of the vectors in B.

Now observe that $R(0) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $R(\pi) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$. It follows that an even number of 1s or -1s can be written as $R(\theta_1)$-blocks. Hence, with a suitable reordering of the basis B, the theorem follows. □

As in the dimension 2 situation, these possibilities can be given a geometric interpretation when $V = \mathbb{R}^3$ is taken as euclidean space. As before, this entails looking carefully at reflections and rotations in $\mathbb{R}^3$. If $Q : \mathbb{R}^3 \to \mathbb{R}^3$ is any reflection in a plane through the origin (called the **fixed plane** of the reflection), take $\{\mathbf{f}_2, \mathbf{f}_3\}$ to be any orthonormal basis of the fixed plane and take $\mathbf{f}_1$ to be a unit vector perpendicular to the fixed plane. Then $Q(\mathbf{f}_1) = -\mathbf{f}_1$, whereas $Q(\mathbf{f}_2) = \mathbf{f}_2$ and $Q(\mathbf{f}_3) = \mathbf{f}_3$. Hence $B = \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ is an orthonormal basis such that

$$M_B(Q) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Similarly, suppose that $R : \mathbb{R}^3 \to \mathbb{R}^3$ is any rotation about a line through the origin (called the **axis** of the rotation), and let $\mathbf{f}_1$ be a unit vector pointing along the axis, so $R(\mathbf{f}_1) = \mathbf{f}_1$. Now the plane through the origin perpendicular to the axis is an R-invariant subspace of $\mathbb{R}^2$ of dimension 2, and the restriction of R to this plane is a rotation. Hence, by Theorem 10.4.4, there is an orthonormal basis $B_1 = \{\mathbf{f}_2,\ \mathbf{f}_3\}$ of this plane such that $M_{B_1}(R) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$. But then $B = \{\mathbf{f}_1,\ \mathbf{f}_2,\ \mathbf{f}_3\}$ is an orthonormal basis of $\mathbb{R}^3$ such that the matrix of R is

$$M_B(R) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$$

However, Theorem 10.4.5 shows that there are isometries T in $\mathbb{R}^3$ of a third type: those with a matrix of the form

$$M_B(T) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$$

If $B = \{\mathbf{f}_1,\ \mathbf{f}_2,\ \mathbf{f}_3\}$, let Q be the reflection in the plane spanned by $\mathbf{f}_2$ and $\mathbf{f}_3$, and let R be the rotation corresponding to θ about the line spanned by $\mathbf{f}_1$. Then $M_B(Q)$ and $M_B(R)$ are as above, and $M_B(Q)M_B(R) = M_B(T)$ as the reader can verify. This means that $M_B(QR) = M_B(T)$ by Theorem 9.2.1, and this in turn implies that $QR = T$ because M_B is one-to-one (see Exercise **??**). A similar argument shows that $RQ = T$, and we have Theorem 10.4.6.

Theorem 10.4.6

If $T : \mathbb{R}^3 \to \mathbb{R}^3$ is an isometry, there are three possibilities.

a. *T is a rotation, and* $M_B(T) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$ *for some orthonormal basis B.*

b. *T is a reflection, and* $M_B(T) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ *for some orthonormal basis B.*

c. *$T = QR = RQ$ where Q is a reflection, R is a rotation about an axis perpendicular to the fixed plane of Q and* $M_B(T) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$ *for some orthonormal basis B.*

Hence T is a rotation if and only if $\det T = 1$.

Proof. It remains only to verify the final observation that T is a rotation if and only if $\det T = 1$. But clearly $\det T = -1$ in parts (b) and (c). □

A useful way of analyzing a given isometry $T : \mathbb{R}^3 \to \mathbb{R}^3$ comes from computing the eigenvalues of T. Because the characteristic polynomial of T has degree 3, it must have a real root. Hence, there must be at least one real eigenvalue, and the only possible real eigenvalues are ± 1 by Lemma 10.4.3. Thus Table 10.1 includes all possibilities.

Table 10.1

Eigenvalues of T	**Action of T**
(1) 1, no other real eigenvalues	Rotation about the line $\mathbb{R}\mathbf{f}$ where $\mathbf{f}$ is an eigenvector corresponding to 1. [Case (a) of Theorem 10.4.6.]
(2) -1, no other real eigenvalues	Rotation about the line $\mathbb{R}\mathbf{f}$ followed by reflection in the plane $(\mathbb{R}\mathbf{f})^{\perp}$ where $\mathbf{f}$ is an eigenvector corresponding to -1. [Case (c) of Theorem 10.4.6.]
(3) $-1,\ 1,\ 1$	Reflection in the plane $(\mathbb{R}\mathbf{f})^{\perp}$ where $\mathbf{f}$ is an eigenvector corresponding to -1. [Case (b) of Theorem 10.4.6.]
(4) $1,\ -1,\ -1$	This is as in (1) with a rotation of π.
(5) $-1,\ -1,\ -1$	Here $T(\mathbf{x}) = -\mathbf{x}$ for all x. This is (2) with a rotation of π.
(6) $1,\ 1,\ 1$	Here T is the identity isometry.

Example 10.4.5

Analyze the isometry $T : \mathbb{R}^3 \to \mathbb{R}^3$ given by $T\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} y \\ z \\ -x \end{bmatrix}$.

Solution. If B_0 is the standard basis of $\mathbb{R}^3$, then $M_{B_0}(T) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 0 \end{bmatrix}$, so $c_T(x) = x^3 + 1 = (x+1)(x^2 - x + 1)$. This is (2) in Table 10.1. Write:

$$\mathbf{f}_1 = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \quad \mathbf{f}_2 = \frac{1}{\sqrt{6}}\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \quad \mathbf{f}_3 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

Here $\mathbf{f}_1$ is a unit eigenvector corresponding to $\lambda_1 = -1$, so T is a rotation (through an angle θ) about the line $L = \mathbb{R}\mathbf{f}_1$, followed by reflection in the plane U through the origin perpendicular to $\mathbf{f}_1$ (with equation $x - y + z = 0$). Then, $\{\mathbf{f}_1,\ \mathbf{f}_2\}$ is chosen as an orthonormal basis of U, so $B = \{\mathbf{f}_1,\ \mathbf{f}_2,\ \mathbf{f}_3\}$ is an orthonormal basis of $\mathbb{R}^3$ and

$$M_B(T) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ 0 & \frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix}$$

Hence θ is given by $\cos\theta = \frac{1}{2}$, $\sin\theta = \frac{\sqrt{3}}{2}$, so $\theta = \frac{\pi}{3}$.

Let V be an n-dimensional inner product space. A subspace of V of dimension $n-1$ is called a **hyperplane** in V. Thus the hyperplanes in $\mathbb{R}^3$ and $\mathbb{R}^2$ are, respectively, the planes and lines through the origin. Let $Q : V \to V$ be an isometry with matrix

$$M_B(Q) = \left[\begin{array}{cc} -1 & 0 \\ 0 & I_{n-1} \end{array} \right]$$

for some orthonormal basis $B = \{\mathbf{f}_1,\ \mathbf{f}_2,\ \dots,\ \mathbf{f}_n\}$. Then $Q(\mathbf{f}_1) = -\mathbf{f}_1$ whereas $Q(\mathbf{u}) = \mathbf{u}$ for each $\mathbf{u}$ in $U = \text{span}\,\{\mathbf{f}_2,\ \dots,\ \mathbf{f}_n\}$. Hence U is called the **fixed hyperplane** of Q, and Q is called **reflection** in U. Note that each hyperplane in V is the fixed hyperplane of a (unique) reflection of V. Clearly, reflections in $\mathbb{R}^2$ and $\mathbb{R}^3$ are reflections in this more general sense.

Continuing the analogy with $\mathbb{R}^2$ and $\mathbb{R}^3$, an isometry $T : V \to V$ is called a **rotation** if there exists an orthonormal basis $\{\mathbf{f}_1,\ \dots,\ \mathbf{f}_n\}$ such that

$$M_B(T) = \left[\begin{array}{ccc} I_r & 0 & 0 \\ 0 & R(\theta) & 0 \\ 0 & 0 & I_s \end{array} \right]$$

in block form, where $R(\theta) = \left[\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array} \right]$, and where either I_r or I_s (or both) may be missing. If $R(\theta)$ occupies columns i and $i+1$ of $M_B(T)$, and if $W = \text{span}\,\{\mathbf{f}_i,\ \mathbf{f}_{i+1}\}$, then W is T-invariant and the matrix of $T : W \to W$ with respect to $\{\mathbf{f}_i,\ \mathbf{f}_{i+1}\}$ is $R(\theta)$. Clearly, if W is viewed as a copy of $\mathbb{R}^2$, then T is a rotation in W. Moreover, $T(\mathbf{u}) = \mathbf{u}$ holds for all vectors $\mathbf{u}$ in the $(n-2)$-dimensional subspace $U = \text{span}\,\{\mathbf{f}_1,\ \dots,\ \mathbf{f}_{i-1},\ \mathbf{f}_{i+1},\ \dots,\ \mathbf{f}_n\}$, and U is called the **fixed axis** of the rotation T. In $\mathbb{R}^3$, the axis of any rotation is a line (one-dimensional), whereas in $\mathbb{R}^2$ the axis is $U = \{\mathbf{0}\}$.

With these definitions, the following theorem is an immediate consequence of Theorem 10.4.5 (the details are left to the reader).

Theorem 10.4.7

Let $T : V \to V$ be an isometry of a finite dimensional inner product space V. Then there exist isometries $T_1,\ \dots,\ T$ such that

$$T = T_k T_{k-1} \cdots T_2 T_1$$

where each T_i is either a rotation or a reflection, at most one is a reflection, and $T_i T_j = T_j T_i$ holds for all i and j. Furthermore, T is a composite of rotations if and only if $\det T = 1$.

10.5 An Application to Fourier Approximation[6]

If U is an orthogonal basis of a vector space V, the expansion theorem (Theorem 10.2.4) presents a vector $v \in V$ as a linear combination of the vectors in U. Of course this requires that the set U is finite since otherwise the linear combination is an infinite sum and makes no sense in V.

However, given an infinite orthogonal set $U = \{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n, \dots\}$, we can use the expansion theorem for $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ for each n to get a series of "approximations" $\mathbf{v}_n$ for a given vector $\mathbf{v}$. A natural question is whether these $\mathbf{v}_n$ are getting closer and closer to $\mathbf{v}$ as n increases. This turns out to be a very fruitful idea.

In this section we shall investigate an important orthogonal set in the space $\mathbf{C}[-\pi, \pi]$ of continuous functions on the interval $[-\pi, \pi]$, using the inner product.

$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(x)g(x)dx$$

Of course, calculus will be needed. The orthogonal set in question is

$$\{1, \sin x, \cos x, \sin(2x), \cos(2x), \sin(3x), \cos(3x), \dots\}$$

Standard techniques of integration give

$$\|1\|^2 = \int_{-\pi}^{\pi} 1^2 dx = 2\pi$$

[6]The name honours the French mathematician J.B.J. Fourier (1768-1830) who used these techniques in 1822 to investigate heat conduction in solids.

$$\|\sin\ kx\|^2 = \int_{-\pi}^{\pi} \sin^2(kx)dx = \pi \quad \text{for any } k = 1,\ 2,\ 3,\ \ldots$$
$$\|\cos\ kx\|^2 = \int_{-\pi}^{\pi} \cos^2(kx)dx = \pi \quad \text{for any } k = 1,\ 2,\ 3,\ \ldots$$

We leave the verifications to the reader, together with the task of showing that these functions are orthogonal:

$$\langle \sin(kx),\ \sin(mx)\rangle = 0 = \langle \cos(kx),\ \cos(mx)\rangle \quad \text{if } k \neq m$$

and

$$\langle \sin(kx),\ \cos(mx)\rangle = 0 \quad \text{for all } k \geq 0 \text{ and } m \geq 0$$

(Note that $1 = \cos(0x)$, so the constant function 1 is included.)

Now define the following subspace of $\mathbf{C}[-\pi,\ \pi]$:

$$F_n = \text{span}\,\{1,\ \sin x,\ \cos x,\ \sin(2x),\ \cos(2x),\ \ldots,\ \sin(nx),\ \cos(nx)\}$$

The aim is to use the approximation theorem (Theorem 10.2.8); so, given a function f in $\mathbf{C}[-\pi,\ \pi]$, define the **Fourier coefficients** of f by

$$a_0 = \frac{\langle f(x),\ 1\rangle}{\|1\|^2} = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)dx$$
$$a_k = \frac{\langle f(x),\ \cos(kx)\rangle}{\|\cos(kx)\|^2} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(kx)dx \quad k = 1,\ 2,\ \ldots$$
$$b_k = \frac{\langle f(x),\ \sin(kx)\rangle}{\|\sin(kx)\|^2} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(kx)dx \quad k = 1,\ 2,\ \ldots$$

Then the approximation theorem (Theorem 10.2.8) gives Theorem 10.5.1.

Theorem 10.5.1

Let f be any continuous real-valued function defined on the interval $[-\pi,\ \pi]$. If $a_0,\ a_1,\ \ldots,$ and $b_0,\ b_1,\ \ldots$ are the Fourier coefficients of f, then given $n \geq 0$,

$$f_n(x) = a_0 + a_1\cos x + b_1\sin x + a_2\cos(2x) + b_2\sin(2x) + \cdots + a_n\cos(nx) + b_n\sin(nx)$$

is a function in F_n that is closest to f in the sense that

$$\|f - f_n\| \leq \|f - g\|$$

holds for all functions g in F_n.

The function f_n is called the nth **Fourier approximation** to the function f.

Example 10.5.1

Find the fifth Fourier approximation to the function $f(x)$ defined on $[-\pi,\ \pi]$ as follows:

$$f(x) = \begin{cases} \pi + x & \text{if } -\pi \leq x < 0 \\ \pi - x & \text{if } 0 \leq x \leq \pi \end{cases}$$

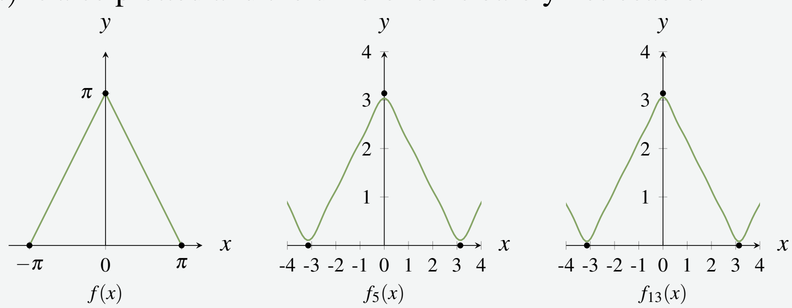
Solution. The graph of $y = f(x)$ appears on the left below. The Fourier coefficients are computed as follows, although the details of the integrations (usually by parts) are omitted.

$$
\begin{aligned}
a_0 &= \tfrac{1}{2\pi}\int_{-\pi}^{\pi} f(x)dx = \tfrac{\pi}{2} \\
a_k &= \tfrac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(kx)dx = \tfrac{2}{\pi k^2}[1-\cos(k\pi)] = \begin{cases} 0 & \text{if } k \text{ is even} \\ \frac{4}{\pi k^2} & \text{if } k \text{ is odd} \end{cases} \\
b_k &= \tfrac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(kx)dx = 0 \qquad \text{for all } k = 1,\ 2,\ \dots
\end{aligned}
$$

Hence the fifth Fourier approximation is

$$f_5(x) = \tfrac{\pi}{2} + \tfrac{4}{\pi}\left\{\cos x + \tfrac{1}{3^2}\cos(3x) + \tfrac{1}{5^2}\cos(5x)\right\}$$

This is plotted in the middle diagram and is already a reasonable approximation to $f(x)$. By comparison, $f_{13}(x)$ is also plotted and the difference is barely noticeable.



We say that a function f is an **even function** if $f(x) = f(-x)$ holds for all x; f is called an **odd function** if $f(-x) = -f(x)$ holds for all x. Examples of even functions include the function in Example 10.5.1, all constant functions, the even powers x^2, x^4, $\dots$, and $\cos(kx)$; these functions are characterized by the fact that the graph of $y = f(x)$ is symmetric about the y axis. Examples of odd functions are the odd powers x, x^3, $\dots$, and $\sin(kx)$ where $k > 0$, and the graph of $y = f(x)$ is symmetric about the origin if f is odd. The usefulness of these functions stems from the fact that

$$
\begin{array}{ll}
\int_{-\pi}^{\pi} f(x)dx = 0 & \text{if } f \text{ is odd} \\
\int_{-\pi}^{\pi} f(x)dx = 2\int_0^{\pi} f(x)dx & \text{if } f \text{ is even}
\end{array}
$$

These facts often simplify the computations of the Fourier coefficients. For example:

1. The Fourier sine coefficients b_k all vanish if f is even.

2. The Fourier cosine coefficients a_k all vanish if f is odd.

This is because $f(x)\sin(kx)$ is odd in the first case and $f(x)\cos(kx)$ is odd in the second case.

The functions 1, $\cos(kx)$, and $\sin(kx)$ that occur in the Fourier approximation for $f(x)$ are all easy to generate as an electrical voltage (when x is time). By summing these signals (with the amplitudes given by the Fourier coefficients), it is possible to produce an electrical signal with (the approximation to) $f(x)$ as the voltage. Hence these Fourier approximations play a fundamental role in electronics.

Finally, the Fourier approximations $f_1, f_2, \ldots$ of a function f in some cases get better and better as n increases. This is in particular the case when the function $f(x)$ is piecewise smooth, that is the function can be broken into distinct pieces and on each piece both the function and its derivative, $f'(x)$, are continuous. A piecewise smooth function may not be continuous everywhere however the only discontinuities that are allowed are a finite number of jump discontinuities. The reason is that the subspaces F_n increase:

$$F_1 \subseteq F_2 \subseteq F_3 \subseteq \cdots \subseteq F_n \subseteq \cdots$$

So, because $f_n = \text{proj}_{F_n} f$, we get (see the discussion following Example 10.2.6)

$$\|f - f_1\| \geq \|f - f_2\| \geq \cdots \geq \|f - f_n\| \geq \cdots$$

Under some conditions these numbers $\|f - f_n\|$ approach zero; in fact, we have the following fundamental theorem.

Theorem 10.5.2

Let f in $C[-\pi, \pi]$ be piecewise smooth. Then

$$f_n(x) \text{ approaches } f(x) \text{ for all } x \text{ such that } -\pi < x < \pi.$$[7]

It shows that f has a representation as an infinite series, called the **Fourier series** of f:

$$f(x) = a_0 + a_1 \cos x + b_1 \sin x + a_2 \cos(2x) + b_2 \sin(2x) + \cdots$$

whenever $-\pi < x < \pi$. A full discussion of Theorem 10.5.2 is beyond the scope of this book. This subject had great historical impact on the development of mathematics, and has become one of the standard tools in science and engineering.

Thus the Fourier series for the function f in Example 10.5.1 is

$$f(x) = \frac{\pi}{2} + \frac{4}{\pi}\left\{\cos x + \frac{1}{3^2}\cos(3x) + \frac{1}{5^2}\cos(5x) + \frac{1}{7^2}\cos(7x) + \cdots\right\}$$

Since $f(0) = \pi$ and $\cos(0) = 1$, taking $x = 0$ leads to the series

$$\frac{\pi^2}{8} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots$$

Example 10.5.2

Expand $f(x) = x$ on the interval $[-\pi, \pi]$ in a Fourier series, and so obtain a series expansion of $\frac{\pi}{4}$.

Solution. Here f is an odd function so all the Fourier cosine coefficients a_k are zero. As to the sine coefficients:

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} x\sin(kx)dx = \frac{2}{k}(-1)^{k+1} \quad \text{for } k \geq 1$$

where we omit the details of the integration by parts. Hence the Fourier series for x is

$$x = 2[\sin x - \frac{1}{2}\sin(2x) + \frac{1}{3}\sin(3x) - \frac{1}{4}\sin(4x) + \ldots]$$

[7] We have to be careful at the end points $x = \pi$ or $x = -\pi$ because $\sin(k\pi) = \sin(-k\pi)$ and $\cos(k\pi) = \cos(-k\pi)$.

for $-\pi < x < \pi$. In particular, taking $x = \frac{\pi}{2}$ gives an infinite series for $\frac{\pi}{4}$.

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots$$

Many other such formulas can be proved using Theorem 10.5.2.

Chapter 11

Canonical Forms

Given a matrix A, the effect of a sequence of row-operations on A is to produce UA where U is invertible. Under this "row-equivalence" operation the best that can be achieved is the reduced row-echelon form for A. If column operations are also allowed, the result is UAV where both U and V are invertible, and the best outcome under this "equivalence" operation is called the Smith canonical form of A (Theorem 2.5.3). There are other kinds of operations on a matrix and, in many cases, there is a "canonical" best possible result.

If A is square, the most important operation of this sort is arguably "similarity" wherein A is carried to $U^{-1}AU$ where U is invertible. In this case we say that matrices A and B are *similar*, and write $A \sim B$, when $B = U^{-1}AU$ for some invertible matrix U. Under similarity the canonical matrices, called *Jordan canonical matrices*, are block triangular with upper triangular "Jordan" blocks on the main diagonal. In this short chapter we are going to define these Jordan blocks and prove that every matrix is similar to a Jordan canonical matrix.

Here is the key to the method. Let $T : V \to V$ be an operator on an n-dimensional vector space V, and suppose that we can find an ordered basis B of V so that the matrix $M_B(T)$ is as simple as possible. Then, if B_0 is *any* ordered basis of V, the matrices $M_B(T)$ and $M_{B_0}(T)$ are similar; that is,

$$M_B(T) = P^{-1}M_{B_0}(T)P \quad \text{for some invertible matrix } P$$

Moreover, $P = P_{B_0 \leftarrow B}$ is easily computed from the bases B and B_0 (Theorem 9.2.3). This, combined with the invariant subspaces and direct sums studied in Section 9.3, enables us to calculate the Jordan canonical form of any square matrix A. Along the way we derive an explicit construction of an invertible matrix P such that $P^{-1}AP$ is block triangular.

This technique is important in many ways. For example, if we want to diagonalize an $n \times n$ matrix A, let $T_A : \mathbb{R}^n \to \mathbb{R}^n$ be the operator given by $T_A(\mathbf{x}) = A\mathbf{x}$ or all $\mathbf{x}$ in $\mathbb{R}^n$, and look for a basis B of $\mathbb{R}^n$ such that $M_B(T_A)$ is diagonal. If $B_0 = E$ is the standard basis of $\mathbb{R}^n$, then $M_E(T_A) = A$, so

$$P^{-1}AP = P^{-1}M_E(T_A)P = M_B(T_A)$$

and we have diagonalized A. Thus the "algebraic" problem of finding an invertible matrix P such that $P^{-1}AP$ is diagonal is converted into the "geometric" problem of finding a basis B such that $M_B(T_A)$ is diagonal. This change of perspective is one of the most important techniques in linear algebra.

11.1 Block Triangular Form

We have shown (Theorem 8.2.5) that any $n \times n$ matrix A with every eigenvalue real is orthogonally similar to an upper triangular matrix U. The following theorem shows that U can be chosen in a special way.

Theorem 11.1.1: Block Triangulation Theorem

Let A be an $n \times n$ matrix with every eigenvalue real and let

$$c_A(x) = (x-\lambda_1)^{m_1}(x-\lambda_2)^{m_2}\cdots(x-\lambda_k)^{m_k}$$

where $\lambda_1, \lambda_2, \dots, \lambda_k$ are the distinct eigenvalues of A. Then an invertible matrix P exists such that

$$P^{-1}AP = \begin{bmatrix} U_1 & 0 & 0 & \cdots & 0 \\ 0 & U_2 & 0 & \cdots & 0 \\ 0 & 0 & U_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & U_k \end{bmatrix}$$

where, for each i, U_i is an $m_i \times m_i$ upper triangular matrix with every entry on the main diagonal equal to λ_i.

The proof is given at the end of this section. For now, we focus on a method for *finding* the matrix P. The key concept is as follows.

Definition 11.1 Generalized Eigenspaces

If A is as in Theorem 11.1.1, the **generalized eigenspace** $G_{\lambda_i}(A)$ *is defined by*

$$G_{\lambda_i}(A) = \text{null}\,[(\lambda_i I - A)^{m_i}]$$

where m_i is the multiplicity of λ_i.

Observe that the eigenspace $E_{\lambda_i}(A) = \text{null}\,(\lambda_i I - A)$ is a subspace of $G_{\lambda_i}(A)$. We need three technical results.

Lemma 11.1.1

Using the notation of Theorem 11.1.1, we have $\dim\,[G_{\lambda_i}(A)] = m_i$.

Proof. Write $A_i = (\lambda_i I - A)^{m_i}$ for convenience and let P be as in Theorem 11.1.1. The spaces $G_{\lambda_i}(A) = \text{null}\,(A_i)$ and $\text{null}\,(P^{-1}A_iP)$ are isomorphic via $\mathbf{x} \leftrightarrow P^{-1}\mathbf{x}$, so we show $\dim\,[\text{null}\,(P^{-1}A_iP)] = m_i$. Now $P^{-1}A_iP = (\lambda_i I - P^{-1}AP)^{m_i}$. If we use the block form in Theorem 11.1.1, this becomes

$$\begin{aligned}
P^{-1}A_iP &= \begin{bmatrix} \lambda_i I - U_1 & 0 & \cdots & 0 \\ 0 & \lambda_i I - U_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_i I - U_k \end{bmatrix}^{m_i} \\
&= \begin{bmatrix} (\lambda_i I - U_1)^{m_i} & 0 & \cdots & 0 \\ 0 & (\lambda_i I - U_2)^{m_i} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & (\lambda_i I - U_k)^{m_i} \end{bmatrix}
\end{aligned}$$

The matrix $(\lambda_i I - U_j)^{m_i}$ is invertible if $j \neq i$ and zero if $j = i$ (because then U_i is an $m_i \times m_i$ upper triangular matrix with each entry on the main diagonal equal to λ_i). It follows that $m_i = \dim\,[\text{null}\,(P^{-1}A_iP)]$, as required. □

Lemma 11.1.2

If P is as in Theorem 11.1.1, denote the columns of P as follows:

$$\mathbf{p}_{11},\ \mathbf{p}_{12},\ \dots,\ \mathbf{p}_{1m_1};\quad \mathbf{p}_{21},\ \mathbf{p}_{22},\ \dots,\ \mathbf{p}_{2m_2};\quad \dots;\quad \mathbf{p}_{k1},\ \mathbf{p}_{k2},\ \dots,\ \mathbf{p}_{km_k}$$

Then $\{\mathbf{p}_{i1},\ \mathbf{p}_{i2},\ \dots,\ \mathbf{p}_{im_i}\}$ is a basis of $G_{\lambda_i}(A)$.

Proof. It suffices by Lemma 11.1.1 to show that each $\mathbf{p}_{ij}$ is *in* $G_{\lambda_i}(A)$. Write the matrix in Theorem 11.1.1 as $P^{-1}AP = \text{diag}\,(U_1,\ U_2,\ \dots,\ U_k)$. Then

$$AP = P\,\text{diag}\,(U_1,\ U_2,\ \dots,\ U_k)$$

Comparing columns gives, successively:

$$\begin{array}{ll}
A\mathbf{p}_{11} = \lambda_1 \mathbf{p}_{11}, & \text{so } (\lambda_1 I - A)\mathbf{p}_{11} = \mathbf{0} \\
A\mathbf{p}_{12} = u\mathbf{p}_{11} + \lambda_1 \mathbf{p}_{12}, & \text{so } (\lambda_1 I - A)^2\mathbf{p}_{12} = \mathbf{0} \\
A\mathbf{p}_{13} = w\mathbf{p}_{11} + v\mathbf{p}_{12} + \lambda_1 \mathbf{p}_{13} & \text{so } (\lambda_1 I - A)^3\mathbf{p}_{13} = \mathbf{0} \\
\vdots & \vdots
\end{array}$$

where u, v, w are in $\mathbb{R}$. In general, $(\lambda_1 I - A)^j \mathbf{p}_{1j} = \mathbf{0}$ for $j = 1, 2, \dots, m_1$, so $\mathbf{p}_{1j}$ is in $G_{\lambda_i}(A)$. Similarly, $\mathbf{p}_{ij}$ is in $G_{\lambda_i}(A)$ for each i and j. □

Lemma 11.1.3

If B_i is any basis of $G_{\lambda_i}(A)$, then $B = B_1 \cup B_2 \cup \cdots \cup B_k$ is a basis of $\mathbb{R}^n$.

Proof. It suffices by Lemma 11.1.1 to show that B is independent. If a linear combination from B vanishes, let $\mathbf{x}_i$ be the sum of the terms from B_i. Then $\mathbf{x}_1 + \cdots + \mathbf{x}_k = \mathbf{0}$. But $\mathbf{x}_i = \sum_j r_{ij}\mathbf{p}_{ij}$ by Lemma 11.1.2, so $\sum_{i,\, j} r_{ij}\mathbf{p}_{ij} = \mathbf{0}$. Hence each $\mathbf{x}_i = \mathbf{0}$, so each coefficient in $\mathbf{x}_i$ is zero. □

Lemma 11.1.2 suggests an algorithm for finding the matrix P in Theorem 11.1.1. Observe that there is an ascending chain of subspaces leading from $E_{\lambda_i}(A)$ to $G_{\lambda_i}(A)$:

$$E_{\lambda_i}(A) = \operatorname{null}[(\lambda_i I - A)] \subseteq \operatorname{null}[(\lambda_i I - A)^2] \subseteq \cdots \subseteq \operatorname{null}[(\lambda_i I - A)^{m_i}] = G_{\lambda_i}(A)$$

We construct a basis for $G_{\lambda_i}(A)$ by climbing up this chain.

Theorem: Triangulation Algorithm

Suppose A has characteristic polynomial

$$c_A(x) = (x - \lambda_1)^{m_1}(x - \lambda_2)^{m_2} \cdots (x - \lambda_k)^{m_k}$$

1. *Choose a basis of* $\operatorname{null}[(\lambda_1 I - A)]$*; enlarge it by adding vectors (possibly none) to a basis of* $\operatorname{null}[(\lambda_1 I - A)^2]$*; enlarge that to a basis of* $\operatorname{null}[(\lambda_1 I - A)^3]$*, and so on. Continue to obtain an ordered basis* $\{\mathbf{p}_{11}, \mathbf{p}_{12}, \dots, \mathbf{p}_{1m_1}\}$ *of* $G_{\lambda_1}(A)$.
2. *As in* (1) *choose a basis* $\{\mathbf{p}_{i1}, \mathbf{p}_{i2}, \dots, \mathbf{p}_{im_i}\}$ *of* $G_{\lambda_i}(A)$ *for each* i.
3. *Let* $P = \left[\ \mathbf{p}_{11}\mathbf{p}_{12}\cdots\mathbf{p}_{1m_1};\ \ \mathbf{p}_{21}\mathbf{p}_{22}\cdots\mathbf{p}_{2m_2};\ \ \cdots;\ \ \mathbf{p}_{k1}\mathbf{p}_{k2}\cdots\mathbf{p}_{km_k}\ \right]$ *be the matrix with these basis vectors (in order) as columns.*

Then $P^{-1}AP = \operatorname{diag}(U_1, U_2, \dots, U_k)$ *as in Theorem 11.1.1.*

Proof. Lemma 11.1.3 guarantees that $B = \{\mathbf{p}_{11}, \dots, \mathbf{p}_{km_1}\}$ is a basis of $\mathbb{R}^n$, and Theorem 9.2.4 shows that $P^{-1}AP = M_B(T_A)$. Now $G_{\lambda_i}(A)$ is T_A-invariant for each i because

$$(\lambda_i I - A)^{m_i}\mathbf{x} = \mathbf{0} \quad \text{implies} \quad (\lambda_i I - A)^{m_i}(A\mathbf{x}) = A(\lambda_i I - A)^{m_i}\mathbf{x} = \mathbf{0}$$

By Theorem 9.3.7 (and induction), we have

$$P^{-1}AP = M_B(T_A) = \operatorname{diag}(U_1,\, U_2,\, \dots,\, U_k)$$

where U_i is the matrix of the restriction of T_A to $G_{\lambda_i}(A)$, and it remains to show that U_i has the desired upper triangular form. Given s, let $\mathbf{p}_{ij}$ be a basis vector in $\operatorname{null}[(\lambda_i I - A)^{s+1}]$. Then $(\lambda_i I - A)\mathbf{p}_{ij}$ is in $\operatorname{null}[(\lambda_i I - A)^s]$, and therefore is a linear combination of the basis vectors $\mathbf{p}_{it}$ coming *before* $\mathbf{p}_{ij}$. Hence

$$T_A(\mathbf{p}_{ij}) = A\mathbf{p}_{ij} = \lambda_i \mathbf{p}_{ij} - (\lambda_i I - A)\mathbf{p}_{ij}$$

shows that the column of U_i corresponding to $\mathbf{p}_{ij}$ has λ_i on the main diagonal and zeros below the main diagonal. This is what we wanted. □

Example 11.1.1

If $A = \begin{bmatrix} 2 & 0 & 0 & 1 \\ 0 & 2 & 0 & -1 \\ -1 & 1 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$, find P such that $P^{-1}AP$ is block triangular.

Solution. $c_A(x) = \det[xI - A] = (x-2)^4$, so $\lambda_1 = 2$ is the only eigenvalue and we are in the case $k = 1$ of Theorem 11.1.1. Compute:

$$(2I - A) = \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad (2I - A)^2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad (2I - A)^3 = 0$$

By gaussian elimination find a basis $\{\mathbf{p}_{11},\, \mathbf{p}_{12}\}$ of $\operatorname{null}(2I - A)$; then extend in any way to a basis $\{\mathbf{p}_{11},\, \mathbf{p}_{12},\, \mathbf{p}_{13}\}$ of $\operatorname{null}[(2I - A)^2]$; and finally get a basis $\{\mathbf{p}_{11},\, \mathbf{p}_{12},\, \mathbf{p}_{13},\, \mathbf{p}_{14}\}$ of $\operatorname{null}[(2I - A)^3] = \mathbb{R}^4$. One choice is

$$\mathbf{p}_{11} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{p}_{12} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \mathbf{p}_{13} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{p}_{14} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Hence $P = \begin{bmatrix} \mathbf{p}_{11} & \mathbf{p}_{12} & \mathbf{p}_{13} & \mathbf{p}_{14} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ gives $P^{-1}AP = \begin{bmatrix} 2 & 0 & 0 & 1 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & 0 & 2 \end{bmatrix}$

Example 11.1.2

If $A = \begin{bmatrix} 2 & 0 & 1 & 1 \\ 3 & 5 & 4 & 1 \\ -4 & -3 & -3 & -1 \\ 1 & 0 & 1 & 2 \end{bmatrix}$, find P such that $P^{-1}AP$ is block triangular.

Solution. The eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = 2$ because

$$\begin{aligned}
c_A(x) &= \begin{vmatrix} x-2 & 0 & -1 & -1 \\ -3 & x-5 & -4 & -1 \\ 4 & 3 & x+3 & 1 \\ -1 & 0 & -1 & x-2 \end{vmatrix} = \begin{vmatrix} x-1 & 0 & 0 & -x+1 \\ -3 & x-5 & -4 & -1 \\ 4 & 3 & x+3 & 1 \\ -1 & 0 & -1 & x-2 \end{vmatrix} \\
&= \begin{vmatrix} x-1 & 0 & 0 & 0 \\ -3 & x-5 & -4 & -4 \\ 4 & 3 & x+3 & 5 \\ -1 & 0 & -1 & x-3 \end{vmatrix} = (x-1)\begin{vmatrix} x-5 & -4 & -4 \\ 3 & x+3 & 5 \\ 0 & -1 & x-3 \end{vmatrix} \\
&= (x-1)\begin{vmatrix} x-5 & -4 & 0 \\ 3 & x+3 & -x+2 \\ 0 & -1 & x-2 \end{vmatrix} = (x-1)\begin{vmatrix} x-5 & -4 & 0 \\ 3 & x+2 & 0 \\ 0 & -1 & x-2 \end{vmatrix} \\
&= (x-1)(x-2)\begin{vmatrix} x-5 & -4 \\ 3 & x+2 \end{vmatrix} = (x-1)^2(x-2)^2
\end{aligned}$$

By solving equations, we find $\text{null}\,(I-A) = \text{span}\,\{\mathbf{p}_{11}\}$ and $\text{null}\,(I-A)^2 = \text{span}\,\{\mathbf{p}_{11}, \mathbf{p}_{12}\}$ where

$$\mathbf{p}_{11} = \begin{bmatrix} 1 \\ 1 \\ -2 \\ 1 \end{bmatrix} \quad \mathbf{p}_{12} = \begin{bmatrix} 0 \\ 3 \\ -4 \\ 1 \end{bmatrix}$$

Since $\lambda_1 = 1$ has multiplicity 2 as a root of $c_A(x)$, $\dim G_{\lambda_1}(A) = 2$ by Lemma 11.1.1. Since $\mathbf{p}_{11}$ and $\mathbf{p}_{12}$ both lie in $G_{\lambda_1}(A)$, we have $G_{\lambda_1}(A) = \text{span}\,\{\mathbf{p}_{11}, \mathbf{p}_{12}\}$. Turning to $\lambda_2 = 2$, we find that $\text{null}\,(2I-A) = \text{span}\,\{\mathbf{p}_{21}\}$ and $\text{null}\,[(2I-A)^2] = \text{span}\,\{\mathbf{p}_{21}, \mathbf{p}_{22}\}$ where

$$\mathbf{p}_{21} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{p}_{22} = \begin{bmatrix} 0 \\ -4 \\ 3 \\ 0 \end{bmatrix}$$

Again, $\dim G_{\lambda_2}(A) = 2$ as λ_2 has multiplicity 2, so $G_{\lambda_2}(A) = \text{span}\,\{\mathbf{p}_{21}, \mathbf{p}_{22}\}$. Hence

$$P = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 3 & 0 & -4 \\ -2 & -4 & -1 & 3 \\ 1 & 1 & 1 & 0 \end{bmatrix} \text{ gives } P^{-1}AP = \begin{bmatrix} 1 & -3 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 3 \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

If $p(x)$ is a polynomial and A is an $n \times n$ matrix, then $p(A)$ is also an $n \times n$ matrix if we interpret $A^0 = I_n$.

For example, if $p(x) = x^2 - 2x + 3$, then $p(A) = A^2 - 2A + 3I$. Theorem 11.1.1 provides another proof of the Cayley-Hamilton theorem (see also Theorem 8.7.10). As before, let $c_A(x)$ denote the characteristic polynomial of A.

Theorem 11.1.2: Cayley-Hamilton Theorem

If A is a square matrix with every eigenvalue real, then $c_A(A) = 0$.

Proof. As in Theorem 11.1.1, write $c_A(x) = (x-\lambda_1)^{m_1} \cdots (x-\lambda_k)^{m_k} = \Pi_{i=1}^{k}(x-\lambda_i)^{m_i}$, and write

$$P^{-1}AP = D = \text{diag}\,(U_1, \dots, U_k)$$

Hence

$$c_A(U_i) = \Pi_{i=1}^{k}(U_i - \lambda_i I_{m_i})^{m_i} = 0 \text{ for each } i$$

because the factor $(U_i - \lambda_i I_{m_i})^{m_i} = 0$. In fact $U_i - \lambda_i I_{m_i}$ is $m_i \times m_i$ and has zeros on the main diagonal. But then

$$\begin{aligned} P^{-1}c_A(A)P = c_A(D) &= c_A[\text{diag}\,(U_1, \dots, U_k)] \\ &= \text{diag}\,[c_A(U_1), \dots, c_A(U_k)] \\ &= 0 \end{aligned}$$

It follows that $c_A(A) = 0$. □

Example 11.1.3

If $A = \begin{bmatrix} 1 & 3 \\ -1 & 2 \end{bmatrix}$, then $c_A(x) = \det \begin{bmatrix} x-1 & -3 \\ 1 & x-2 \end{bmatrix} = x^2 - 3x + 5$. Then $c_A(A) = A^2 - 3A + 5I_2 = \begin{bmatrix} -2 & 9 \\ -3 & 1 \end{bmatrix} - \begin{bmatrix} 3 & 9 \\ -3 & 6 \end{bmatrix} + \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$.

Theorem 11.1.1 will be refined even further in the next section.

Proof of Theorem 11.1.1

The proof of Theorem 11.1.1 requires the following simple fact about bases, the proof of which we leave to the reader.

Lemma 11.1.4

If $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is a basis of a vector space V, so also is $\{\mathbf{v}_1 + s\mathbf{v}_2, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ for any scalar s.

Proof of Theorem 11.1.1. Let A be as in Theorem 11.1.1, and let $T = T_A : \mathbb{R}^n \to \mathbb{R}^n$ be the matrix transformation induced by A. For convenience, call a matrix a λ-m-ut matrix if it is an $m \times m$ upper triangular matrix and every diagonal entry equals λ. Then we must find a basis B of $\mathbb{R}^n$ such that $M_B(T) = \text{diag}\,(U_1, U_2, \dots, U_k)$ where U_i is a λ_i-m_i-ut matrix for each i. We proceed by induction on n. If $n = 1$, take $B = \{\mathbf{v}\}$ where $\mathbf{v}$ is any eigenvector of T.

If $n > 1$, let $\mathbf{v}_1$ be a λ_1-eigenvector of T, and let $B_0 = \{\mathbf{v}_1, \mathbf{w}_1, \dots, \mathbf{w}_{n-1}\}$ be any basis of $\mathbb{R}^n$ containing $\mathbf{v}_1$. Then (see Lemma 5.5.2)

$$M_{B_0}(T) = \begin{bmatrix} \lambda_1 & X \\ 0 & A_1 \end{bmatrix}$$

in block form where A_1 is $(n-1) \times (n-1)$. Moreover, A and $M_{B_0}(T)$ are similar, so

$$c_A(x) = c_{M_{B_0}(T)}(x) = (x - \lambda_1)c_{A_1}(x)$$

Hence $c_{A_1}(x) = (x-\lambda_1)^{m_1-1}(x-\lambda_2)^{m_2}\cdots(x-\lambda_k)^{m_k}$ so (by induction) let

$$Q^{-1}A_1Q = \operatorname{diag}(Z_1, U_2, \dots, U_k)$$

where Z_1 is a λ_1-(m_1-1)-ut matrix and U_i is a λ_i-m_i-ut matrix for each $i > 1$.

If $P = \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix}$, then $P^{-1}MB_0(T) = \begin{bmatrix} \lambda_1 & XQ \\ 0 & Q^{-1}A_1Q \end{bmatrix} = A'$, say. Hence $A' \sim M_{B_0}(T) \sim A$ so by Theorem 9.2.4(2) there is a basis B of $\mathbb{R}^n$ such that $M_{B_1}(T_A) = A'$, that is $M_{B_1}(T) = A'$. Hence $M_{B_1}(T)$ takes the block form

$$M_{B_1}(T) = \begin{bmatrix} \lambda_1 & XQ \\ 0 & \operatorname{diag}(Z_1, U_2, \dots, U_k) \end{bmatrix} = \left[\begin{array}{cc|ccc} \lambda_1 & X_1 & & Y & \\ 0 & Z_1 & 0 & 0 & 0 \\ \hline & & U_2 & \cdots & 0 \\ & 0 & \vdots & & \vdots \\ & & 0 & \cdots & U_k \end{array}\right] \tag{11.1}$$

If we write $U_1 = \begin{bmatrix} \lambda_1 & X_1 \\ 0 & Z_1 \end{bmatrix}$, the basis B_1 fulfills our needs except that the row matrix Y may not be zero.

We remedy this defect as follows. Observe that the first vector in the basis B_1 is a λ_1 eigenvector of T, which we continue to denote as $\mathbf{v}_1$. The idea is to add suitable scalar multiples of $\mathbf{v}_1$ to the other vectors in B_1. This results in a new basis by Lemma 11.1.4, and the multiples can be chosen so that the new matrix of T is the same as (11.1) except that $Y = 0$. Let $\{\mathbf{w}_1, \dots, \mathbf{w}_{m_2}\}$ be the vectors in B_1 corresponding to λ_2 (giving rise to U_2 in (11.1)). Write

$$U_2 = \begin{bmatrix} \lambda_2 & u_{12} & u_{13} & \cdots & u_{1m_2} \\ 0 & \lambda_2 & u_{23} & \cdots & u_{2m_2} \\ 0 & 0 & \lambda_2 & \cdots & u_{3m_2} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_2 \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} y_1 & y_2 & \cdots & y_{m_2} \end{bmatrix}$$

We first replace $\mathbf{w}_1$ by $\mathbf{w}_1' = \mathbf{w}_1 + s\mathbf{v}_1$ where s is to be determined. Then (11.1) gives

$$\begin{aligned} T(\mathbf{w}_1') &= T(\mathbf{w}_1) + sT(\mathbf{v}_1) \\ &= (y_1\mathbf{v}_1 + \lambda_2\mathbf{w}_1) + s\lambda_1\mathbf{v}_1 \\ &= y_1\mathbf{v}_1 + \lambda_2(\mathbf{w}_1' - s\mathbf{v}_1) + s\lambda_1\mathbf{v}_1 \\ &= \lambda_2\mathbf{w}_1' + [(y_1 - s(\lambda_2 - \lambda_1)]\mathbf{v}_1 \end{aligned}$$

Because $\lambda_2 \neq \lambda_1$ we can choose s such that $T(\mathbf{w}'_1) = \lambda_2 \mathbf{w}'_1$. Similarly, let $\mathbf{w}'_2 = \mathbf{w}_2 + t\mathbf{v}_1$ where t is to be chosen. Then, as before,

$$\begin{aligned} T(\mathbf{w}'_2) &= T(\mathbf{w}_2) + tT(\mathbf{v}_1) \\ &= (y_2\mathbf{v}_1 + u_{12}\mathbf{w}_1 + \lambda_2\mathbf{w}_2) + t\lambda_1\mathbf{v}_1 \\ &= u_{12}\mathbf{w}'_1 + \lambda_2\mathbf{w}'_2 + [(y_2 - u_{12}s) - t(\lambda_2 - \lambda_1)]\mathbf{v}_1 \end{aligned}$$

Again, t can be chosen so that $T(\mathbf{w}'_2) = u_{12}\mathbf{w}'_1 + \lambda_2\mathbf{w}'_2$. Continue in this way to eliminate $y_1, \dots, y_{m_2}$. This procedure also works for $\lambda_3, \lambda_4, \dots$ and so produces a new basis B such that $M_B(T)$ is as in (11.1) but with $Y = 0$. □

11.2 The Jordan Canonical Form

Two $m \times n$ matrices A and B are called row-equivalent if A can be carried to B using row operations and, equivalently, if $B = UA$ for some invertible matrix U. We know (Theorem 2.6.4) that each $m \times n$ matrix is row-equivalent to a unique matrix in reduced row-echelon form, and we say that these reduced row-echelon matrices are *canonical forms* for $m \times n$ matrices using row operations. If we allow column operations as well, then $A \to UAV = \left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right]$ for invertible U and V, and the canonical forms are the matrices $\left[\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right]$ where r is the rank (this is the Smith normal form and is discussed in Theorem 2.6.3). In this section, we discover the canonical forms for square matrices under similarity: $A \to P^{-1}AP$.

If A is an $n \times n$ matrix with distinct real eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_k$, we saw in Theorem 11.1.1 that A is similar to a block triangular matrix; more precisely, an invertible matrix P exists such that

$$P^{-1}AP = \left[\begin{array}{cccc} U_1 & 0 & \cdots & 0 \\ 0 & U_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & U_k \end{array} \right] = \mathrm{diag}(U_1, U_2, \dots, U_k) \tag{11.2}$$

where, for each i, U_i is upper triangular with λ_i repeated on the main diagonal. The Jordan canonical form is a refinement of this theorem. The proof we gave of (11.2) is matrix theoretic because we wanted to give an algorithm for actually finding the matrix P. However, we are going to employ abstract methods here. Consequently, we reformulate Theorem 11.1.1 as follows:

Theorem 11.2.1

Let $T : V \to V$ be a linear operator where $\dim V = n$. Assume that $\lambda_1, \lambda_2, \dots, \lambda_k$ are the distinct eigenvalues of T, and that the λ_i are all real. Then there exists a basis F of V such that $M_F(T) = \mathrm{diag}(U_1, U_2, \dots, U_k)$ where, for each i, U_i is square, upper triangular, with λ_i repeated on the main diagonal.

Proof. Choose any basis $B = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ of V and write $A = M_B(T)$. Since A has the same eigenvalues as T, Theorem 11.1.1 shows that an invertible matrix P exists such that $P^{-1}AP = \mathrm{diag}(U_1, U_2, \dots, U_k)$ where the U_i are as in the statement of the Theorem. If $\mathbf{p}_j$ denotes column j of P and $C_B : V \to \mathbb{R}^n$ is the coordinate isomorphism, let $\mathbf{f}_j = C_B^{-1}(\mathbf{p}_j)$ for each j. Then $F = \{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ is a basis of V and $C_B(\mathbf{f}_j) = \mathbf{p}_j$ for each j. This means that $P_{B \leftarrow F} = \left[C_B(\mathbf{f}_j) \right] = \left[\mathbf{p}_j \right] = P$, and hence (by Theorem 9.2.2) that $P_{F \leftarrow B} = P^{-1}$. With this, column j of $M_F(T)$ is

$$C_F(T(\mathbf{f}_j)) = P_{F \leftarrow B} C_B(T(\mathbf{f}_j)) = P^{-1} M_B(T) C_B(\mathbf{f}_j) = P^{-1}A\mathbf{p}_j$$

for all j. Hence

$$M_F(T) = \left[C_F(T(\mathbf{f}_j)) \right] = \left[P^{-1}A\mathbf{p}_j \right] = P^{-1}A \left[\mathbf{p}_j \right] = P^{-1}AP = \mathrm{diag}(U_1, U_2, \dots, U_k)$$

as required. $\square$

Definition 11.2 Jordan Blocks

*If $n \geq 1$, define the **Jordan block** $J_n(\lambda)$ to be the $n \times n$ matrix with λs on the main diagonal, 1s on the diagonal above, and 0s elsewhere. We take $J_1(\lambda) = [\lambda]$.*

Hence

$$J_1(\lambda) = [\lambda], \quad J_2(\lambda) = \left[\begin{array}{cc} \lambda & 1 \\ 0 & \lambda \end{array} \right], \quad J_3(\lambda) = \left[\begin{array}{ccc} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{array} \right], \quad J_4(\lambda) = \left[\begin{array}{cccc} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{array} \right], \quad \dots$$

We are going to show that Theorem 11.2.1 holds with each block U_i replaced by Jordan blocks corresponding to eigenvalues. It turns out that the whole thing hinges on the case $\lambda = 0$. An operator T is called **nilpotent** if $T^m = 0$ for some $m \geq 1$, and in this case $\lambda = 0$ for every eigenvalue λ of T. Moreover, the converse holds by Theorem 11.1.1. Hence the following lemma is crucial.

Lemma 11.2.1

Let $T : V \to V$ be a linear operator where $\dim V = n$, and assume that T is nilpotent; that is, $T^m = 0$ for some $m \geq 1$. Then V has a basis B such that

$$M_B(T) = \operatorname{diag}(J_1,\, J_2,\, \dots,\, J_k)$$

where each J_i is a Jordan block corresponding to $\lambda = 0$.[1]

A proof is given at the end of this section.

Theorem 11.2.2: Real Jordan Canonical Form

Let $T : V \to V$ be a linear operator where $\dim V = n$, and assume that $\lambda_1,\, \lambda_2,\, \dots,\, \lambda_m$ are the distinct eigenvalues of T and that the λ_i are all real. Then there exists a basis E of V such that

$$M_E(T) = \operatorname{diag}(U_1,\, U_2,\, \dots,\, U_k)$$

in block form. Moreover, each U_j is itself block diagonal:

$$U_j = \operatorname{diag}(J_1,\, J_2,\, \dots,\, J_k)$$

where each J_i is a Jordan block corresponding to some λ_i.

Proof. Let $E = \{\mathbf{e}_1,\, \mathbf{e}_2,\, \dots,\, \mathbf{e}_n\}$ be a basis of V as in Theorem 11.2.1, and assume that U_i is an $n_i \times n_i$ matrix for each i. Let

$$E_1 = \{\mathbf{e}_1,\, \dots,\, \mathbf{e}_{n_1}\}, \quad E_2 = \{\mathbf{e}_{n_1+1},\, \dots,\, \mathbf{e}_{n_1+n_2}\}, \quad \dots, \quad E_k = \{\mathbf{e}_{n_1+\cdots+n_{k-1}+1},\, \dots,\, \mathbf{e}_{n_1+\cdots+n_k}\}$$

where $n_k = n$, and define $V_i = \operatorname{span}\{E_i\}$ for each i. Because the matrix $M_E(T) = \operatorname{diag}(U_1,\, U_2,\, \dots,\, U_m)$ is block diagonal, it follows that each V_i is T-invariant and $M_{E_i}(T) = U_i$ for each i. Let U_i have λ_i repeated along the main diagonal, and consider the restriction $T : V_i \to V_i$. Then $M_{E_i}(T - \lambda_i I_{n_i})$ is a nilpotent matrix, and hence $(T - \lambda_i I_{n_i})$ is a nilpotent operator on V_i. But then Lemma 11.2.1 shows that V_i has a basis B_i such that $M_{B_i}(T - \lambda_i I_{n_i}) = \operatorname{diag}(K_1,\, K_2,\, \dots,\, K_{t_i})$ where each K_i is a Jordan block corresponding to $\lambda = 0$. Hence

$$\begin{aligned} M_{B_i}(T) &= M_{B_i}(\lambda_i I_{n_i}) + M_{B_i}(T - \lambda_i I_{n_i}) \\ &= \lambda_i I_{n_i} + \operatorname{diag}(K_1,\, K_2,\, \dots,\, K_{t_i}) = \operatorname{diag}(J_1,\, J_2,\, \dots,\, J_k) \end{aligned}$$

where $J_i = \lambda_i I_{f_i} + K_i$ is a Jordan block corresponding to λ_i (where K_i is $f_i \times f_i$). Finally,

$$B = B_1 \cup B_2 \cup \cdots \cup B_k$$

is a basis of V with respect to which T has the desired matrix. □

[1]The converse is true too: If $M_B(T)$ has this form for some basis B of V, then T is nilpotent.

Corollary 11.2.1

If A is an $n \times n$ matrix with real eigenvalues, an invertible matrix P exists such that $P^{-1}AP = \text{diag}(J_1, J_2, \dots, J_k)$ where each J_i is a Jordan block corresponding to an eigenvalue λ_i.

Proof. Apply Theorem 11.2.2 to the matrix transformation $T_A : \mathbb{R}^n \to \mathbb{R}^n$ to find a basis B of $\mathbb{R}^n$ such that $M_B(T_A)$ has the desired form. If P is the (invertible) $n \times n$ matrix with the vectors of B as its columns, then $P^{-1}AP = M_B(T_A)$ by Theorem 9.2.4. □

Of course if we work over the field $\mathbb{C}$ of complex numbers rather than $\mathbb{R}$, the characteristic polynomial of a (complex) matrix A splits completely as a product of linear factors. The proof of Theorem 11.2.2 goes through to give

Theorem 11.2.3: Jordan Canonical Form[2]

Let $T : V \to V$ be a linear operator where $\dim V = n$, and assume that $\lambda_1, \lambda_2, \dots, \lambda_m$ are the distinct eigenvalues of T. Then there exists a basis F of V such that

$$M_F(T) = \text{diag}(U_1, U_2, \dots, U_k)$$

in block form. Moreover, each U_j is itself block diagonal:

$$U_j = \text{diag}(J_1, J_2, \dots, J_{t_j})$$

where each J_i is a Jordan block corresponding to some λ_i.

Except for the order of the Jordan blocks J_i, the Jordan canonical form is uniquely determined by the operator T. That is, for each eigenvalue λ the number and size of the Jordan blocks corresponding to λ is uniquely determined. Thus, for example, two matrices (or two operators) are similar if and only if they have the same Jordan canonical form. We omit the proof of uniqueness; it is best presented using modules in a course on abstract algebra.

Proof of Lemma 1

Lemma 11.2.2

Let $T : V \to V$ be a linear operator where $\dim V = n$, and assume that T is nilpotent; that is, $T^m = 0$ for some $m \geq 1$. Then V has a basis B such that

$$M_B(T) = \text{diag}(J_1, J_2, \dots, J_k)$$

where each $J_i = J_{n_i}(0)$ is a Jordan block corresponding to $\lambda = 0$.

Proof. The proof proceeds by induction on n. If $n = 1$, then T is a scalar operator, and so $T = 0$ and the lemma holds. If $n \geq 1$, we may assume that $T \neq 0$, so $m \geq 1$ and we may assume that m is chosen such

[2] This was first proved in 1870 by the French mathematician Camille Jordan (1838–1922) in his monumental *Traité des substitutions et des équations algébriques*.

that $T^m = 0$, but $T^{m-1} \neq 0$. Suppose $T^{m-1}\mathbf{u} \neq \mathbf{0}$ for some $\mathbf{u}$ in V.[3]

Claim. $\{\mathbf{u},\ T\mathbf{u},\ T^2\mathbf{u},\ \dots,\ T^{m-1}\mathbf{u}\}$ is independent.

Proof. Suppose $a_0\mathbf{u}+a_1T\mathbf{u}+a_2T^2\mathbf{u}+\cdots+a_{m-1}T^{m-1}\mathbf{u}=\mathbf{0}$ where each a_i is in $\mathbb{R}$. Since $T^m=0$, applying T^{m-1} gives $\mathbf{0}=T^{m-1}\mathbf{0}=a_0T^{m-1}\mathbf{u}$, whence $a_0=0$. Hence $a_1T\mathbf{u}+a_2T^2\mathbf{u}+\cdots+a_{m-1}T^{m-1}\mathbf{u}=\mathbf{0}$ and applying T^{m-2} gives $a_1=0$ in the same way. Continue in this fashion to obtain $a_i=0$ for each i. This proves the Claim.

Now define $P=\text{span}\{\mathbf{u},\ T\mathbf{u},\ T^2\mathbf{u},\ \dots,\ T^{m-1}\mathbf{u}\}$. Then P is a T-invariant subspace (because $T^m=0$), and $T:P\to P$ is nilpotent with matrix $M_B(T)=J_m(0)$ where $B=\{\mathbf{u},\ T\mathbf{u},\ T^2\mathbf{u},\ \dots,\ T^{m-1}\mathbf{u}\}$. Hence we are done, by induction, if $V=P\oplus Q$ where Q is T-invariant (then $\dim Q=n-\dim P<n$ because $P\neq 0$, and $T:Q\to Q$ is nilpotent). With this in mind, choose a T-invariant subspace Q of maximal dimension such that $P\cap Q=\{\mathbf{0}\}$.[4] We assume that $V\neq P\oplus Q$ and look for a contradiction.

Choose $\mathbf{x}\in V$ such that $\mathbf{x}\notin P\oplus Q$. Then $T^m\mathbf{x}=\mathbf{0}\in P\oplus Q$ while $T^0\mathbf{x}=\mathbf{x}\notin P\oplus Q$. Hence there exists k, $1\leq k\leq m$, such that $T^k\mathbf{x}\in P\oplus Q$ but $T^{k-1}\mathbf{x}\notin P\oplus Q$. Write $\mathbf{v}=T^{k-1}\mathbf{x}$, so that

$$\mathbf{v}\notin P\oplus Q \quad \text{and} \quad T\mathbf{v}\in P\oplus Q$$

Let $T\mathbf{v}=\mathbf{p}+\mathbf{q}$ with $\mathbf{p}$ in P and $\mathbf{q}$ in Q. Then $\mathbf{0}=T^{m-1}(T\mathbf{v})=T^{m-1}\mathbf{p}+T^{m-1}\mathbf{q}$ so, since P and Q are T-invariant, $T^{m-1}\mathbf{p}=-T^{m-1}\mathbf{q}\in P\cap Q=\{\mathbf{0}\}$. Hence

$$T^{m-1}\mathbf{p}=\mathbf{0}$$

Since $\mathbf{p}\in P$ we have $\mathbf{p}=a_0\mathbf{u}+a_1T\mathbf{u}+a_2T^2\mathbf{u}+\cdots+a_{m-1}T^{m-1}\mathbf{u}$ for $a_i\in\mathbb{R}$. Since $T^m=0$, applying T^{m-1} gives $\mathbf{0}=T^{m-1}\mathbf{p}=a_0T^{m-1}\mathbf{u}$, whence $a_0=0$. Thus $\mathbf{p}=T(\mathbf{p}_1)$ where

$$\mathbf{p}_1=a_1\mathbf{u}+a_2T\mathbf{u}+\cdots+a_{m-1}T^{m-2}\mathbf{u}\in P$$

If we write $\mathbf{v}_1=\mathbf{v}-\mathbf{p}_1$ we have

$$T(\mathbf{v}_1)=T(\mathbf{v}-\mathbf{p}_1)=T\mathbf{v}-\mathbf{p}=\mathbf{q}\in Q$$

Since $T(Q)\subseteq Q$, it follows that $T(Q+\mathbb{R}\mathbf{v}_1)\subseteq Q\subseteq Q+\mathbb{R}\mathbf{v}_1$. Moreover $\mathbf{v}_1\notin Q$ (otherwise $\mathbf{v}=\mathbf{v}_1+\mathbf{p}_1\in P\oplus Q$, a contradiction). Hence $Q\subset Q+\mathbb{R}\mathbf{v}_1$ so, by the maximality of Q, we have $(Q+\mathbb{R}\mathbf{v}_1)\cap P\neq\{\mathbf{0}\}$, say

$$\mathbf{0}\neq\mathbf{p}_2=\mathbf{q}_1+a\mathbf{v}_1 \quad \text{where} \quad \mathbf{p}_2\in P, \quad \mathbf{q}_1\in Q, \quad \text{and} \quad a\in\mathbb{R}$$

Thus $a\mathbf{v}_1=\mathbf{p}_2-\mathbf{q}_1\in P\oplus Q$. But since $\mathbf{v}_1=\mathbf{v}-\mathbf{p}_1$ we have

$$a\mathbf{v}=a\mathbf{v}_1+a\mathbf{p}_1\in(P\oplus Q)+P=P\oplus Q$$

Since $\mathbf{v}\notin P\oplus Q$, this implies that $a=0$. But then $\mathbf{p}_2=\mathbf{q}_1\in P\cap Q=\{\mathbf{0}\}$, a contradiction. This completes the proof. □

[3]If $S:V\to V$ is an operator, we abbreviate $S(\mathbf{u})$ by $S\mathbf{u}$ for simplicity.

[4]Observe that there *is* at least one such subspace: $Q=\{\mathbf{0}\}$.

Appendix A

Complex Numbers

The fact that the square of every real number is nonnegative shows that the equation $x^2 + 1 = 0$ has no real root; in other words, there is no real number u such that $u^2 = -1$. So the set of real numbers is inadequate for finding all roots of all polynomials. This kind of problem arises with other number systems as well. The set of integers contains no solution of the equation $3x + 2 = 0$, and the rational numbers had to be invented to solve such equations. But the set of rational numbers is also incomplete because, for example, it contains no root of the polynomial $x^2 - 2$. Hence the real numbers were invented. In the same way, the set of complex numbers was invented, which contains all real numbers together with a root of the equation $x^2 + 1 = 0$. However, the process ends here: the complex numbers have the property that *every* polynomial with complex coefficients has a (complex) root. This fact is known as the fundamental theorem of algebra.

One pleasant aspect of the complex numbers is that, whereas describing the real numbers in terms of the rationals is a rather complicated business, the complex numbers are quite easy to describe in terms of real numbers. Every **complex number** has the form

$$a + bi$$

where a and b are real numbers, and i is a root of the polynomial $x^2 + 1$. Here a and b are called the **real part** and the **imaginary part** of the complex number, respectively. The real numbers are now regarded as special complex numbers of the form $a + 0i = a$, with zero imaginary part. The complex numbers of the form $0 + bi = bi$ with zero real part are called **pure imaginary** numbers. The complex number i itself is called the **imaginary unit** and is distinguished by the fact that

$$i^2 = -1$$

As the terms *complex* and *imaginary* suggest, these numbers met with some resistance when they were first used. This has changed; now they are essential in science and engineering as well as mathematics, and they are used extensively. The names persist, however, and continue to be a bit misleading: These numbers are no more "*complex*" than the real numbers, and the number i is no more "*imaginary*" than -1.

Much as for polynomials, two complex numbers are declared to be **equal** if and only if they have the same real parts and the same imaginary parts. In symbols,

$$a + bi = a' + b'i \quad \text{if and only if } a = a' \text{ and } b = b'$$

The addition and subtraction of complex numbers is accomplished by adding and subtracting real and imaginary parts:

$$\begin{aligned}(a + bi) + (a' + b'i) &= (a + a') + (b + b')i \\ (a + bi) - (a' + b'i) &= (a - a') + (b - b')i\end{aligned}$$

This is analogous to these operations for linear polynomials $a + bx$ and $a' + b'x$, and the multiplication of complex numbers is also analogous with one difference: $i^2 = -1$. The definition is

$$(a + bi)(a' + b'i) = (aa' - bb') + (ab' + ba')i$$

With these definitions of equality, addition, and multiplication, the complex numbers *satisfy all the basic arithmetical axioms adhered to by the real numbers* (the verifications are omitted). One consequence of this is that they can be manipulated in the obvious fashion, except that i^2 is replaced by -1 wherever it occurs, and the rule for equality must be observed.

Example A.1

If $z = 2-3i$ and $w = -1+i$, write each of the following in the form $a+bi$: $z+w$, $z-w$, zw, $\frac{1}{3}z$, and z^2.

Solution.

$$\begin{aligned}
z+w &= (2-3i)+(-1+i) = (2-1)+(-3+1)i = 1-2i \\
z-w &= (2-3i)-(-1+i) = (2+1)+(-3-1)i = 3-4i \\
zw &= (2-3i)(-1+i) = (-2-3i^2)+(2+3)i = 1+5i \\
\tfrac{1}{3}z &= \tfrac{1}{3}(2-3i) = \tfrac{2}{3}-i \\
z^2 &= (2-3i)(2-3i) = (4+9i^2)+(-6-6)i = -5-12i
\end{aligned}$$

Example A.2

Find all complex numbers z such as that $z^2 = i$.

Solution. Write $z = a+bi$; we must determine a and b. Now $z^2 = (a^2-b^2)+(2ab)i$, so the condition $z^2 = i$ becomes

$$(a^2-b^2)+(2ab)i = 0+i$$

Equating real and imaginary parts, we find that $a^2 = b^2$ and $2ab = 1$. The solution is $a = b = \pm\frac{1}{\sqrt{2}}$, so the complex numbers required are $z = \frac{1}{\sqrt{2}}+\frac{1}{\sqrt{2}}i$ and $z = -\frac{1}{\sqrt{2}}-\frac{1}{\sqrt{2}}i$.

As for real numbers, it is possible to divide by every nonzero complex number z. That is, there exists a complex number w such that $wz = 1$. As in the real case, this number w is called the **inverse** of z and is denoted by z^{-1} or $\frac{1}{z}$. Moreover, if $z = a+bi$, the fact that $z \neq 0$ means that $a \neq 0$ or $b \neq 0$. Hence $a^2+b^2 \neq 0$, and an explicit formula for the inverse is

$$\frac{1}{z} = \frac{a}{a^2+b^2} - \frac{b}{a^2+b^2}i$$

In actual calculations, the work is facilitated by two useful notions: the conjugate and the absolute value of a complex number. The next example illustrates the technique.

Example A.3

Write $\frac{3+2i}{2+5i}$ in the form $a+bi$.

Solution. Multiply top and bottom by the complex number $2-5i$ (obtained from the denominator by negating the imaginary part). The result is

$$\frac{3+2i}{2+5i} = \frac{(2-5i)(3+2i)}{(2-5i)(2+5i)} = \frac{(6+10)+(4-15)i}{2^2-(5i)^2} = \frac{16}{29} - \frac{11}{29}i$$

Hence the simplified form is $\frac{16}{29} - \frac{11}{29}i$, as required.

The key to this technique is that the product $(2-5i)(2+5i) = 29$ in the denominator turned out to be a *real* number. The situation in general leads to the following notation: If $z = a+bi$ is a complex number, the **conjugate** of z is the complex number, denoted $\overline{z}$, given by

$$\overline{z} = a - bi \quad \text{where } z = a+bi$$

Hence $\overline{z}$ is obtained from z by negating the imaginary part. Thus $\overline{(2+3i)} = 2-3i$ and $\overline{(1-i)} = 1+i$. If we multiply $z = a+bi$ by $\overline{z}$, we obtain

$$z\overline{z} = a^2 + b^2 \quad \text{where } z = a+bi$$

The real number a^2+b^2 is always nonnegative, so we can state the following definition: The **absolute value** or **modulus** of a complex number $z = a+bi$, denoted by $|z|$, is the positive square root $\sqrt{a^2+b^2}$; that is,

$$|z| = \sqrt{a^2+b^2} \quad \text{where } z = a+bi$$

For example, $|2-3i| = \sqrt{2^2+(-3)^2} = \sqrt{13}$ and $|1+i| = \sqrt{1^2+1^2} = \sqrt{2}$.

Note that if a real number a is viewed as the complex number $a+0i$, its absolute value (as a complex number) is $|a| = \sqrt{a^2}$, which agrees with its absolute value as a *real* number.

With these notions in hand, we can describe the technique applied in Example A.3 as follows: When converting a quotient $\frac{z}{w}$ of complex numbers to the form $a+bi$, multiply top and bottom by the conjugate $\overline{w}$ of the denominator.

The following list contains the most important properties of conjugates and absolute values. Throughout, z and w denote complex numbers.

$C1.$	$\overline{z \pm w} = \overline{z} \pm \overline{w}$	$C7.$	$\frac{1}{z} = \frac{1}{|z|^2}\overline{z}$
$C2.$	$\overline{zw} = \overline{z}\,\overline{w}$	$C8.$	$|z| \geq 0$ for all complex numbers z
$C3.$	$\overline{\left(\frac{z}{w}\right)} = \frac{\overline{z}}{\overline{w}}$	$C9.$	$|z| = 0$ if and only if $z = 0$
$C4.$	$\overline{(\overline{z})} = z$	$C10.$	$|zw| = |z||w|$
$C5.$	z is real if and only if $\overline{z} = z$	$C11.$	$|\frac{z}{w}| = \frac{|z|}{|w|}$
$C6.$	$z\overline{z} = |z|^2$	$C12.$	$|z+w| \leq |z| + |w|$ (**triangle inequality**)

All these properties (except property C12) can (and should) be verified by the reader for arbitrary complex numbers $z = a+bi$ and $w = c+di$. They are not independent; for example, property C10 follows from properties C2 and C6.

The triangle inequality, as its name suggests, comes from a geometric representation of the complex numbers analogous to identification of the real numbers with the points of a line. The representation is achieved as follows:

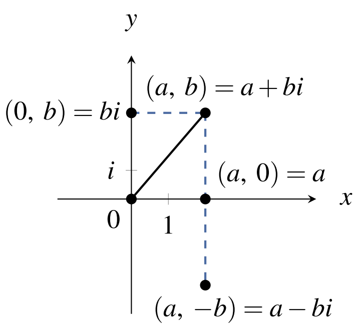


Figure A.1

Introduce a rectangular coordinate system in the plane (Figure A.1), and identify the complex number $a+bi$ with the point $(a,\ b)$. When this is done, the plane is called the **complex plane**. Note that the point $(a,\ 0)$ on the x axis now represents the *real* number $a = a + 0i$, and for this reason, the x axis is called the **real axis**. Similarly, the y axis is called the **imaginary axis**. The identification $(a,\ b) = a + bi$ of the geometric point $(a,\ b)$ and the complex number $a + bi$ will be used in what follows without comment. For example, the origin will be referred to as 0.

This representation of the complex numbers in the complex plane gives a useful way of describing the absolute value and conjugate of a complex number $z = a + bi$. The absolute value $|z| = \sqrt{a^2 + b^2}$ is just the distance from z to the origin. This makes properties C8 and C9 quite obvious. The conjugate $\overline{z} = a - bi$ of z is just the reflection of z in the real axis (x axis), a fact that makes properties C4 and C5 clear.

Given two complex numbers $z_1 = a_1 + b_1 i = (a_1,\ b_1)$ and $z_2 = a_2 + b_2 i = (a_2,\ b_2)$, the absolute value of their difference

$$|z_1 - z_2| = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2}$$

is just the distance between them. This gives the **complex distance formula**:

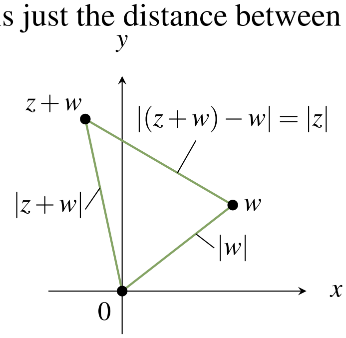


Figure A.2

$$|z_1 - z_2| \text{ is the distance between } z_1 \text{ and } z_2$$

This useful fact yields a simple verification of the triangle inequality, property C12. Suppose z and w are given complex numbers. Consider the triangle in Figure A.2 whose vertices are 0, w, and $z + w$. The three sides have lengths $|z|$, $|w|$, and $|z + w|$ by the complex distance formula, so the inequality

$$|z + w| \leq |z| + |w|$$

expresses the obvious geometric fact that the sum of the lengths of two sides of a triangle is at least as great as the length of the third side.

The representation of complex numbers as points in the complex plane has another very useful property: It enables us to give a geometric description of the sum and product of two complex numbers. To obtain the description for the sum, let

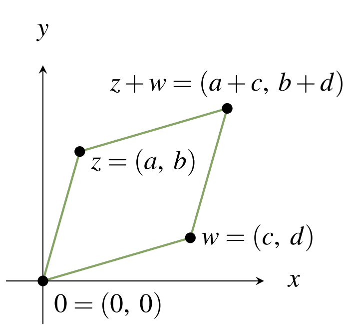


Figure A.3

$$z = a + bi = (a,\ b)$$
$$w = c + di = (c,\ d)$$

denote two complex numbers. We claim that the four points 0, z, w, and $z + w$ form the vertices of a parallelogram. In fact, in Figure A.3 the lines from 0 to z and from w to $z + w$ have slopes

$$\frac{b-0}{a-0} = \frac{b}{a} \quad \text{and} \quad \frac{(b+d)-d}{(a+c)-c} = \frac{b}{a}$$

respectively, so these lines are parallel. (If it happens that $a = 0$, then both these lines are vertical.) Similarly, the lines from z to $z+w$ and from 0 to w are also parallel, so the figure with vertices 0, z, w, and $z+w$ is indeed a parallelogram. Hence, the complex number $z+w$ can be obtained geometrically from z and w by *completing* the parallelogram. This is sometimes called the **parallelogram law** of complex addition. Readers who have studied mechanics will recall that velocities and accelerations add in the same way; in fact, these are all special cases of *vector* addition.

Polar Form

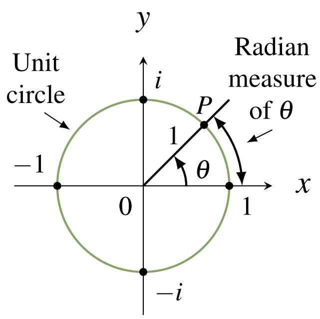


Figure A.4

The geometric description of what happens when two complex numbers are multiplied is at least as elegant as the parallelogram law of addition, but it requires that the complex numbers be represented in polar form. Before discussing this, we pause to recall the general definition of the trigonometric functions sine and cosine. An angle θ in the complex plane is in **standard position** if it is measured counterclockwise from the positive real axis as indicated in Figure A.4. Rather than using degrees to measure angles, it is more natural to use radian measure. This is defined as follows: The circle with its centre at the origin and radius 1 (called the **unit circle**) is drawn in Figure A.4. It has circumference 2π, and the **radian measure** of θ is the length of the arc on the unit circle counterclockwise from 1 to the point P on the unit circle determined by θ. Hence $90^\circ = \frac{\pi}{2}$, $45^\circ = \frac{\pi}{4}$, $180^\circ = \pi$, and a full circle has the angle $360^\circ = 2\pi$. Angles measured clockwise from 1 are negative; for example, $-i$ corresponds to $-\frac{\pi}{2}$ (or to $\frac{3\pi}{2}$).

Consider an angle θ in the range $0 \leq \theta \leq \frac{\pi}{2}$. If θ is plotted in standard position as in Figure A.4, it determines a unique point P on the unit circle, and P has coordinates $(\cos\theta,\ \sin\theta)$ by elementary trigonometry. However, *any* angle θ (acute or not) determines a unique point on the unit circle, so we *define* the **cosine** and **sine** of θ (written $\cos\theta$ and $\sin\theta$) to be the x and y coordinates of this point. For example, the points

$$1 = (1,\ 0) \quad i = (0,\ 1) \quad -1 = (-1,\ 0) \quad -i = (0,\ -1)$$

plotted in Figure A.4 are determined by the angles 0, $\frac{\pi}{2}$, π, $\frac{3\pi}{2}$, respectively. Hence

$$\begin{array}{llll} \cos 0 = 1 & \cos\frac{\pi}{2} = 0 & \cos\pi = -1 & \cos\frac{3\pi}{2} = 0 \\ \sin 0 = 0 & \sin\frac{\pi}{2} = 1 & \sin\pi = 0 & \sin\frac{3\pi}{2} = -1 \end{array}$$

Now we can describe the polar form of a complex number. Let $z = a+bi$ be a complex number, and write the absolute value of z as

$$r = |z| = \sqrt{a^2+b^2}$$

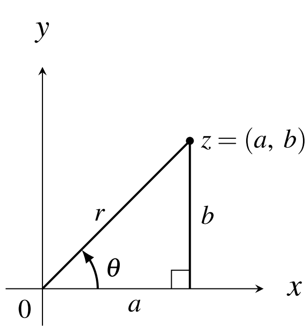


Figure A.5

If $z \neq 0$, the angle θ shown in Figure A.5 is called an **argument** of z and is denoted

$$\theta = \arg z$$

This angle is not unique ($\theta + 2\pi k$ would do as well for any $k = 0,\ \pm 1,\ \pm 2,\ \dots$). However, there is only one argument θ in the range $-\pi < \theta \leq \pi$, and this is sometimes called the **principal argument** of z.

Returning to Figure A.5, we find that the real and imaginary parts a and b of z are related to r and θ by

$$a = r\cos\theta$$
$$b = r\sin\theta$$

Hence the complex number $z = a + bi$ has the form

$$z = r(\cos\theta + i\sin\theta) \quad r = |z|,\ \theta = \arg(z)$$

The combination $\cos\theta + i\sin\theta$ is so important that a special notation is used:

$$e^{i\theta} = \cos\theta + i\sin\theta$$

is called **Euler's formula** after the great Swiss mathematician Leonhard Euler (1707–1783). With this notation, z is written

$$z = re^{i\theta} \quad r = |z|,\ \theta = \arg(z)$$

This is a **polar form** of the complex number z. Of course it is not unique, because the argument can be changed by adding a multiple of 2π.

Example A.4

Write $z_1 = -2 + 2i$ and $z_2 = -i$ in polar form.

Solution.

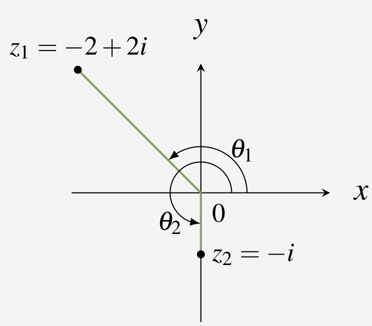


Figure A.6

The two numbers are plotted in the complex plane in Figure A.6. The absolute values are

$$r_1 = |-2 + 2i| = \sqrt{(-2)^2 + 2^2} = 2\sqrt{2}$$
$$r_2 = |-i| = \sqrt{0^2 + (-1)^2} = 1$$

By inspection of Figure A.6, arguments of z_1 and z_2 are

$$\theta_1 = \arg(-2 + 2i) = \tfrac{3\pi}{4}$$
$$\theta_2 = \arg(-i) = \tfrac{3\pi}{2}$$

The corresponding polar forms are $z_1 = -2 + 2i = 2\sqrt{2}e^{3\pi i/4}$ and $z_2 = -i = e^{3\pi i/2}$. Of course, we could have taken the argument $-\frac{\pi}{2}$ for z_2 and obtained the polar form $z_2 = e^{-\pi i/2}$.

In Euler's formula $e^{i\theta} = \cos\theta + i\sin\theta$, the number e is the familiar constant $e = 2.71828\dots$ from calculus. The reason for using e will not be given here; the reason why $\cos\theta + i\sin\theta$ is written as an *exponential* function of θ is that the **law of exponents** holds:

$$e^{i\theta} \cdot e^{i\phi} = e^{i(\theta+\phi)}$$

where θ and ϕ are any two angles. In fact, this is an immediate consequence of the addition identities for $\sin(\theta + \phi)$ and $\cos(\theta + \phi)$:

$$
\begin{aligned}
e^{i\theta}e^{i\phi} &= (\cos\theta + i\sin\theta)(\cos\phi + i\sin\phi) \\
&= (\cos\theta\cos\phi - \sin\theta\sin\phi) + i(\cos\theta\sin\phi + \sin\theta\cos\phi) \\
&= \cos(\theta+\phi) + i\sin(\theta+\phi) \\
&= e^{i(\theta+\phi)}
\end{aligned}
$$

This is analogous to the rule $e^a e^b = e^{a+b}$, which holds for real numbers a and b, so it is not unnatural to use the exponential notation $e^{i\theta}$ for the expression $\cos\theta + i\sin\theta$. In fact, a whole theory exists wherein functions such as e^z, $\sin z$, and $\cos z$ are studied, where z is a *complex* variable. Many deep and beautiful theorems can be proved in this theory, one of which is the so-called fundamental theorem of algebra mentioned later (Theorem A.4). We shall not pursue this here.

The geometric description of the multiplication of two complex numbers follows from the law of exponents.

Theorem A.1: Multiplication Rule

If $z_1 = r_1 e^{i\theta_1}$ and $z_2 = r_2 e^{i\theta_2}$ are complex numbers in polar form, then

$$z_1 z_2 = r_1 r_2 e^{i(\theta_1+\theta_2)}$$

In other words, to multiply two complex numbers, simply multiply the absolute values and add the arguments. This simplifies calculations considerably, particularly when we observe that it is valid for *any* arguments θ_1 and θ_2.

Example A.5

Multiply $(1-i)(1+\sqrt{3}i)$ in two ways.

Solution.

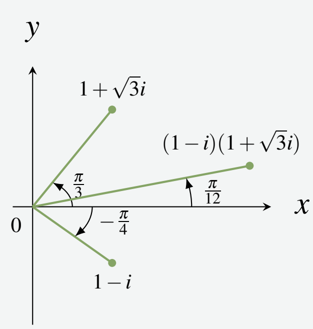


Figure A.7

We have $|1-i| = \sqrt{2}$ and $|1+\sqrt{3}i| = 2$ so, from Figure A.7,

$$
\begin{aligned}
1 - i &= \sqrt{2}e^{-i\pi/4} \\
1 + \sqrt{3}i &= 2e^{i\pi/3}
\end{aligned}
$$

Hence, by the multiplication rule,

$$
\begin{aligned}
(1-i)(1+\sqrt{3}i) &= (\sqrt{2}e^{-i\pi/4})(2e^{i\pi/3}) \\
&= 2\sqrt{2}e^{i(-\pi/4+\pi/3)} \\
&= 2\sqrt{2}e^{i\pi/12}
\end{aligned}
$$

This gives the required product in polar form. Of course, direct multiplication gives $(1-i)(1+\sqrt{3}i) = (\sqrt{3}+1) + (\sqrt{3}-1)i$. Hence, equating real and imaginary parts gives the formulas $\cos(\frac{\pi}{12}) = \frac{\sqrt{3}+1}{2\sqrt{2}}$ and $\sin(\frac{\pi}{12}) = \frac{\sqrt{3}-1}{2\sqrt{2}}$.

Roots of Unity

If a complex number $z = re^{i\theta}$ is given in polar form, the powers assume a particularly simple form. In fact, $z^2 = (re^{i\theta})(re^{i\theta}) = r^2e^{2i\theta}$, $z^3 = z^2 \cdot z = (r^2e^{2i\theta})(re^{i\theta}) = r^3e^{3i\theta}$, and so on. Continuing in this way, it follows by induction that the following theorem holds for any positive integer n. The name honours Abraham De Moivre (1667–1754).

Theorem A.2: De Moivre's Theorem

If θ is any angle, then $(e^{i\theta})^n = e^{in\theta}$ holds for all integers n.

Proof. The case $n > 0$ has been discussed, and the reader can verify the result for $n = 0$. To derive it for $n < 0$, first observe that

$$\text{if} \quad z = re^{i\theta} \neq 0 \quad \text{then} \quad z^{-1} = \frac{1}{r}\, e^{-i\theta}$$

In fact, $(re^{i\theta})(\frac{1}{r}e^{-i\theta}) = 1e^{i0} = 1$ by the multiplication rule. Now assume that n is negative and write it as $n = -m$, $m > 0$. Then

$$(re^{i\theta})^n = [(re^{i\theta})^{-1}]^m = (\frac{1}{r}\, e^{-i\theta})^m = r^{-m}e^{i(-m\theta)} = r^n e^{in\theta}$$

If $r = 1$, this is De Moivre's theorem for negative n. □

Example A.6

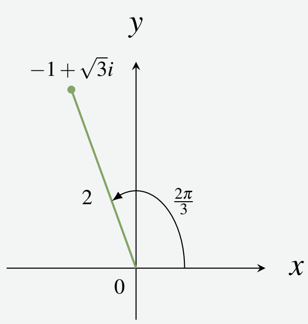


Figure A.8

Verify that $(-1+\sqrt{3}i)^3 = 8$.

Solution. We have $|-1+\sqrt{3}i| = 2$, so $-1+\sqrt{3}i = 2e^{2\pi i/3}$ (see Figure A.8). Hence De Moivre's theorem gives

$$(-1+\sqrt{3}i)^3 = (2e^{2\pi i/3})^3 = 8e^{3(2\pi i/3)} = 8e^{2\pi i} = 8$$

De Moivre's theorem can be used to find nth roots of complex numbers where n is positive. The next example illustrates this technique.

Example A.7

Find the cube roots of unity; that is, find all complex numbers z such that $z^3 = 1$.

Solution. First write $z = re^{i\theta}$ and $1 = 1e^{i0}$ in polar form. We must use the condition $z^3 = 1$ to determine r and θ. Because $z^3 = r^3e^{3i\theta}$ by De Moivre's theorem, this requirement becomes

$$r^3e^{3i\theta} = 1e^{0i}$$

These two complex numbers are equal, so their absolute values must be equal and the arguments must either be equal or differ by an integral multiple of 2π:

$$\begin{aligned} r^3 &= 1 \\ 3\theta &= 0 + 2k\pi, \quad k \text{ some integer} \end{aligned}$$

Because r is real and positive, the condition $r^3 = 1$ implies that $r = 1$. However,

$$\theta = \tfrac{2k\pi}{3}, \quad k \text{ some integer}$$

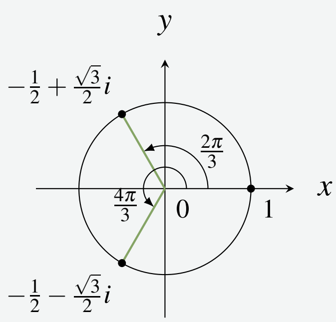


Figure A.9

seems at first glance to yield infinitely many different angles for z. However, choosing $k = 0,\ 1,\ 2$ gives three possible arguments θ (where $0 \leq \theta < 2\pi$), and the corresponding roots are

$$\begin{aligned} 1e^{0i} &= 1 \\ 1e^{2\pi i/3} &= -\tfrac{1}{2} + \tfrac{\sqrt{3}}{2}i \\ 1e^{4\pi i/3} &= -\tfrac{1}{2} - \tfrac{\sqrt{3}}{2}i \end{aligned}$$

These are displayed in Figure A.9. All other values of k yield values of θ that differ from one of these by a multiple of 2π—and so do not give new roots. Hence we have found all the roots.

The same type of calculation gives all complex n**th roots of unity**; that is, all complex numbers z such that $z^n = 1$. As before, write $1 = 1e^{0i}$ and

$$z = re^{i\theta}$$

in polar form. Then $z^n = 1$ takes the form

$$r^n e^{ni\theta} = 1e^{0i}$$

using De Moivre's theorem. Comparing absolute values and arguments yields

$$\begin{aligned} r^n &= 1 \\ n\theta &= 0 + 2k\pi, \quad k \text{ some integer} \end{aligned}$$

Hence $r = 1$, and the n values

$$\theta = \tfrac{2k\pi}{n}, \quad k = 0,\ 1,\ 2,\ \dots,\ n-1$$

of θ all lie in the range $0 \leq \theta < 2\pi$. As in Example A.7, *every* choice of k yields a value of θ that differs from one of these by a multiple of 2π, so these give the arguments of *all* the possible roots.

Theorem A.3: nth Roots of Unity

If $n \geq 1$ is an integer, the nth roots of unity (that is, the solutions to $z^n = 1$) are given by

$$z = e^{2\pi ki/n}, \quad k = 0,\ 1,\ 2,\ \dots,\ n-1$$

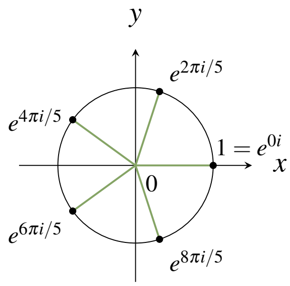


Figure A.10

The nth roots of unity can be found geometrically as the points on the unit circle that cut the circle into n equal sectors, starting at 1. The case $n = 5$ is shown in Figure A.10, where the five fifth roots of unity are plotted.

The method just used to find the nth roots of unity works equally well to find the nth roots of any complex number in polar form. We give one example.

Example A.8

Find the fourth roots of $\sqrt{2} + \sqrt{2}i$.

Solution. First write $\sqrt{2} + \sqrt{2}i = 2e^{\pi i/4}$ in polar form. If $z = re^{i\theta}$ satisfies $z^4 = \sqrt{2} + \sqrt{2}i$, then De Moivre's theorem gives

$$r^4 e^{i(4\theta)} = 2e^{\pi i/4}$$

Hence $r^4 = 2$ and $4\theta = \frac{\pi}{4} + 2k\pi$, k an integer. We obtain four distinct roots (and hence all) by

$$r = \sqrt[4]{2}, \qquad \theta = \tfrac{\pi}{16} = \tfrac{2k\pi}{16},\ k = 0,\ 1,\ 2,\ 3$$

Thus the four roots are

$$\sqrt[4]{2}e^{\pi i/16} \qquad \sqrt[4]{2}e^{9\pi i/16} \qquad \sqrt[4]{2}e^{17\pi i/16} \qquad \sqrt[4]{2}e^{25\pi i/16}$$

Of course, reducing these roots to the form $a + bi$ would require the computation of $\sqrt[4]{2}$ and the sine and cosine of the various angles.

An expression of the form $ax^2 + bx + c$, where the coefficients $a \neq 0$, b, and c are real numbers, is called a **real quadratic**. A complex number u is called a **root** of the quadratic if $au^2 + bu + c = 0$. The roots are given by the famous **quadratic formula**:

$$u = \tfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

The quantity $d = b^2 - 4ac$ is called the **discriminant** of the quadratic $ax^2 + bx + c$, and there is no real root if and only if $d < 0$. In this case the quadratic is said to be **irreducible**. Moreover, the fact that $d < 0$ means that $\sqrt{d} = i\sqrt{|d|}$, so the two (complex) roots are conjugates of each other:

$$u = \tfrac{1}{2a}(-b + i\sqrt{|d|}) \quad \text{and} \quad \overline{u} = \tfrac{1}{2a}(-b - i\sqrt{|d|})$$

The converse of this is true too: Given any nonreal complex number u, then u and $\overline{u}$ are the roots of some real irreducible quadratic. Indeed, the quadratic

$$x^2 - (u + \overline{u})x + u\overline{u} = (x - u)(x - \overline{u})$$

has real coefficients ($u\overline{u} = |u|^2$ and $u + \overline{u}$ is twice the real part of u) and so is irreducible because its roots u and $\overline{u}$ are not real.

Example A.9

Find a real irreducible quadratic with $u = 3 - 4i$ as a root.

Solution. We have $u + \overline{u} = 6$ and $|u|^2 = 25$, so $x^2 - 6x + 25$ is irreducible with u and $\overline{u} = 3 + 4i$ as roots.

Fundamental Theorem of Algebra

As we mentioned earlier, the complex numbers are the culmination of a long search by mathematicians to find a set of numbers large enough to contain a root of every polynomial. The fact that the complex numbers have this property was first proved by Gauss in 1797 when he was 20 years old. The proof is omitted.

Theorem A.4: Fundamental Theorem of Algebra

Every polynomial of positive degree with complex coefficients has a complex root.

If $f(x)$ is a polynomial with complex coefficients, and if u_1 is a root, then the factor theorem (Section 6.5) asserts that

$$f(x) = (x - u_1)g(x)$$

where $g(x)$ is a polynomial with complex coefficients and with degree one less than the degree of $f(x)$. Suppose that u_2 is a root of $g(x)$, again by the fundamental theorem. Then $g(x) = (x - u_2)h(x)$, so

$$f(x) = (x - u_1)(x - u_2)h(x)$$

This process continues until the last polynomial to appear is linear. Thus $f(x)$ has been expressed as a product of linear factors. The last of these factors can be written in the form $u(x - u_n)$, where u and u_n are complex (verify this), so the fundamental theorem takes the following form.

Theorem A.5

Every complex polynomial $f(x)$ of degree $n \geq 1$ has the form

$$f(x) = u(x - u_1)(x - u_2)\cdots(x - u_n)$$

where $u,\ u_1,\ \dots,\ u_n$ are complex numbers and $u \neq 0$. The numbers $u_1,\ u_2,\ \dots,\ u_n$ are the roots of $f(x)$ (and need not all be distinct), and u is the coefficient of x^n.

This form of the fundamental theorem, when applied to a polynomial $f(x)$ with *real* coefficients, can be used to deduce the following result.

Theorem A.6

Every polynomial $f(x)$ of positive degree with real coefficients can be factored as a product of linear and irreducible quadratic factors.

In fact, suppose $f(x)$ has the form

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

where the coefficients a_i are real. If u is a complex root of $f(x)$, then we claim first that $\overline{u}$ is also a root. In fact, we have $f(u) = 0$, so

$$\begin{aligned} 0 = \overline{0} = \overline{f(u)} &= \overline{a_n u^n + a_{n-1}u^{n-1} + \cdots + a_1 u + a_0} \\ &= \overline{a_n u^n} + \overline{a_{n-1}u^{n-1}} + \cdots + \overline{a_1 u} + \overline{a_0} \\ &= \overline{a}_n \overline{u}^n + \overline{a}_{n-1}\overline{u}^{n-1} + \cdots + \overline{a}_1 \overline{u} + \overline{a}_0 \\ &= a_n \overline{u}^n + a_{n-1}\overline{u}^{n-1} + \cdots + a_1 \overline{u} + a_0 \\ &= f(\overline{u}) \end{aligned}$$

where $\overline{a}_i = a_i$ for each i because the coefficients a_i are real. Thus if u is a root of $f(x)$, so is its conjugate $\overline{u}$. Of course some of the roots of $f(x)$ may be real (and so equal their conjugates), but the nonreal roots come in pairs, u and $\overline{u}$. By Theorem A.6, we can thus write $f(x)$ as a product:

$$f(x) = a_n(x - r_1)\cdots(x - r_k)(x - u_1)(x - \overline{u}_1)\cdots(x - u_m)(x - \overline{u}_m) \tag{A.1}$$

where a_n is the coefficient of x^n in $f(x)$; $r_1, r_2, \ldots, r_k$ are the real roots; and $u_1, \overline{u}_1, u_2, \overline{u}_2, \ldots, u_m, \overline{u}_m$ are the nonreal roots. But the product

$$(x - u_j)(x - \overline{u}_j) = x^2 - (u_j + \overline{u}_j)x + (u_j\overline{u}_j)$$

is a real irreducible quadratic for each j (see the discussion preceding Example A.9). Hence (A.1) shows that $f(x)$ is a product of linear and irreducible quadratic factors, each with real coefficients. This is the conclusion in Theorem A.6.

Appendix B

Proofs

Logic plays a basic role in human affairs. Scientists use logic to draw conclusions from experiments, judges use it to deduce consequences of the law, and mathematicians use it to prove theorems. Logic arises in ordinary speech with assertions such as "If John studies hard, he will pass the course," or "If an integer n is divisible by 6, then n is divisible by 3."[1] In each case, the aim is to assert that if a certain statement is true, then another statement must also be true. In fact, if p and q denote statements, most theorems take the form of an **implication**: "If p is true, then q is true." We write this in symbols as

$$p \Rightarrow q$$

and read it as "p implies q." Here p is the **hypothesis** and q the **conclusion** of the implication. The verification that $p \Rightarrow q$ is valid is called the **proof** of the implication. In this section we examine the most common methods of proof[2] and illustrate each technique with some examples.

Method of Direct Proof

To prove that $p \Rightarrow q$, demonstrate directly that q is true whenever p is true.

Example B.1

If n is an odd integer, show that n^2 is odd.

Solution. If n is odd, it has the form $n = 2k + 1$ for some integer k. Then $n^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$ also is odd because $2k^2 + 2k$ is an integer.

Note that the computation $n^2 = 4k^2 + 4k + 1$ in Example B.1 involves some simple properties of arithmetic that we did not prove. These properties, in turn, can be proved from certain more basic properties of numbers (called axioms)—more about that later. Actually, a whole body of mathematical information lies behind nearly every proof of any complexity, although this fact usually is not stated explicitly. Here is a geometrical example.

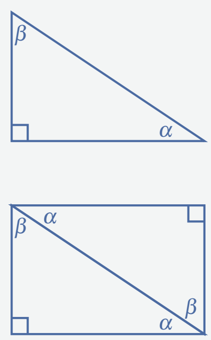
[1] By an *integer* we mean a "whole number"; that is, a number in the set $0, \pm 1, \pm 2, \pm 3, \dots$

[2] For a more detailed look at proof techniques see D. Solow, How to Read and Do Proofs, 2nd ed. (New York: Wiley, 1990); or J. F. Lucas. *Introduction to Abstract Mathematics*, Chapter 2 (Belmont, CA: Wadsworth, 1986).

Example B.2

In a right triangle, show that the sum of the two acute angles is 90 degrees.

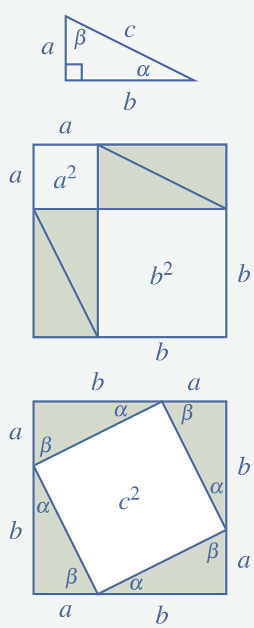
Solution.



The right triangle is shown in the diagram. Construct a rectangle with sides of the same length as the short sides of the original triangle, and draw a diagonal as shown. The original triangle appears on the bottom of the rectangle, and the top triangle is identical to the original (but rotated). Now it is clear that $\alpha + \beta$ is a right angle.

Geometry was one of the first subjects in which formal proofs were used—Euclid's *Elements* was published about 300 B.C. The *Elements* is the most successful textbook ever written, and contains many of the basic geometrical theorems that are taught in school today. In particular, Euclid included a proof of an earlier theorem (about 500 B.C.) due to Pythagoras. Recall that, in a right triangle, the side opposite the right angle is called the *hypotenuse* of the triangle.

Example B.3: Pythagoras' Theorem



In a right-angled triangle, show that the square of the length of the hypotenuse equals the sum of the squares of the lengths of the other two sides.

Solution. Let the sides of the right triangle have lengths a, b, and c as shown. Consider two squares with sides of length $a+b$, and place four copies of the triangle in these squares as in the diagram. The central rectangle in the second square shown is itself a square because the angles α and β add to 90 degrees (using Example B.2), so its area is c^2 as shown. Comparing areas shows that both $a^2 + b^2$ and c^2 each equal the area of the large square minus four times the area of the original triangle, and hence are equal.

Sometimes it is convenient (or even necessary) to break a proof into parts, and deal with each case separately. We formulate the general method as follows:

Method of Reduction to Cases

To prove that $p \Rightarrow q$, show that p implies at least one of a list $p_1, p_2, \dots, p_n$ of statements (the cases) and then show that $p_i \Rightarrow q$ for each i.

Example B.4

Show that $n^2 \geq 0$ for every integer n.

Solution. This statement can be expressed as an implication: If n is an integer, then $n^2 \geq 0$. To prove it, consider the following three cases:

$$(1)\ n > 0; \quad (2)\ n = 0; \quad (3)\ n < 0.$$

Then $n^2 > 0$ in Cases (1) and (3) because the product of two positive (or two negative) integers is positive. In Case (2) $n^2 = 0^2 = 0$, so $n^2 \geq 0$ in every case.

Example B.5

If n is an integer, show that $n^2 - n$ is even.

Solution. We consider two cases:

$$(1)\ n \text{ is even}; \quad (2)\ n \text{ is odd}.$$

We have $n^2 - n = n(n-1)$, so this is even in Case (1) because any multiple of an even number is again even. Similarly, $n - 1$ is even in Case (2) so $n(n-1)$ is again even for the same reason. Hence $n^2 - n$ is even in any case.

The statements used in mathematics are required to be either true or false. This leads to a proof technique which causes consternation in many beginning students. The method is a formal version of a debating strategy whereby the debater assumes the truth of an opponent's position and shows that it leads to an absurd conclusion.

Method of Proof by Contradiction

To prove that $p \Rightarrow q$, show that the assumption that both p is true and q is false leads to a contradiction. In other words, if p is true, then q must be true; that is, $p \Rightarrow q$.

Example B.6

If r is a rational number (fraction), show that $r^2 \neq 2$.

Solution. To argue by contradiction, we assume that r is a rational number and that $r^2 = 2$, and show that this assumption leads to a contradiction. Let m and n be integers such that $r = \frac{m}{n}$ is in lowest terms (so, in particular, m and n are not both even). Then $r^2 = 2$ gives $m^2 = 2n^2$, so m^2 is even. This means m is even (Example B.1), say $m = 2k$. But then $2n^2 = m^2 = 4k^2$, so $n^2 = 2k^2$ is even, and hence n is even. This shows that n and m are both even, contrary to the choice of these numbers.

Example B.7: Pigeonhole Principle

If $n+1$ pigeons are placed in n holes, then some hole contains at least 2 pigeons.

Solution. Assume the conclusion is false. Then each hole contains at most one pigeon and so, since there are n holes, there must be at most n pigeons, contrary to assumption.

The next example involves the notion of a *prime* number, that is an integer that is greater than 1 which cannot be factored as the product of two smaller positive integers both greater than 1. The first few primes are $2, 3, 5, 7, 11, \dots$.

Example B.8

If $2^n - 1$ is a prime number, show that n is a prime number.

Solution. We must show that $p \Rightarrow q$ where p is the statement "$2^n - 1$ is a prime", and q is the statement "n is a prime." Suppose that p is true but q is false so that n is not a prime, say $n = ab$ where $a \geq 2$ and $b \geq 2$ are integers. If we write $2^a = x$, then $2^n = 2^{ab} = (2^a)^b = x^b$. Hence $2^n - 1$ factors:

$$2^n - 1 = x^b - 1 = (x-1)(x^{b-1} + x^{b-2} + \cdots + x^2 + x + 1)$$

As $x \geq 4$, this expression is a factorization of $2^n - 1$ into smaller positive integers, contradicting the assumption that $2^n - 1$ is prime.

The next example exhibits one way to show that an implication is *not* valid.

Example B.9

Show that the implication "n is a prime $\Rightarrow 2^n - 1$ is a prime" is false.

Solution. The first four primes are 2, 3, 5, and 7, and the corresponding values for $2^n - 1$ are 3, 7, 31, 127 (when $n = 2, 3, 5, 7$). These are all prime as the reader can verify. This result seems to be evidence that the implication is true. However, the next prime is 11 and $2^{11} - 1 = 2047 = 23 \cdot 89$, which is clearly *not* a prime.

We say that $n = 11$ is a **counterexample** to the (proposed) implication in Example B.9. Note that, if you can find even one example for which an implication is not valid, the implication is false. Thus disproving implications is in a sense easier than proving them.

The implications in Example B.8 and Example B.9 are closely related: They have the form $p \Rightarrow q$ and $q \Rightarrow p$, where p and q are statements. Each is called the **converse** of the other and, as these examples show, an implication can be valid even though its converse is not valid. If *both* $p \Rightarrow q$ and $q \Rightarrow p$ are valid, the statements p and q are called **logically equivalent**. This is written in symbols as

$$p \Leftrightarrow q$$

and is read "p if and only if q". Many of the most satisfying theorems make the assertion that two statements, ostensibly quite different, are in fact logically equivalent.

Example B.10

If n is an integer, show that "n is odd $\Leftrightarrow n^2$ is odd."

Solution. In Example B.1 we proved the implication "n is odd $\Rightarrow n^2$ is odd." Here we prove the converse by contradiction. If n^2 is odd, we assume that n is not odd. Then n is even, say $n = 2k$, so $n^2 = 4k^2$, which is also even, a contradiction.

Many more examples of proofs can be found in this book and, although they are often more complex, most are based on one of these methods. In fact, linear algebra is one of the best topics on which the reader can sharpen his or her skill at constructing proofs. Part of the reason for this is that much of linear algebra is developed using the **axiomatic method**. That is, in the course of studying various examples it is observed that they all have certain properties in common. Then a general, abstract system is studied in which these basic properties are *assumed* to hold (and are called **axioms**). In this system, statements (called **theorems**) are deduced from the axioms using the methods presented in this appendix. These theorems will then be true in *all* the concrete examples, because the axioms hold in each case. But this procedure is more than just an efficient method for finding theorems in the examples. By reducing the proof to its essentials, we gain a better understanding of why the theorem is true and how it relates to analogous theorems in other abstract systems.

The axiomatic method is not new. Euclid first used it in about 300 B.C. to derive all the propositions of (euclidean) geometry from a list of 10 axioms. The method lends itself well to linear algebra. The axioms are simple and easy to understand, and there are only a few of them. For example, the theory of vector spaces contains a large number of theorems derived from only ten simple axioms.

Appendix C

Mathematical Induction

Suppose one is presented with the following sequence of equations:

$$\begin{aligned} 1 &= 1 \\ 1+3 &= 4 \\ 1+3+5 &= 9 \\ 1+3+5+7 &= 16 \\ 1+3+5+7+9 &= 25 \end{aligned}$$

It is clear that there is a pattern. The numbers on the right side of the equations are the squares 1^2, 2^2, 3^2, 4^2, and 5^2 and, in the equation with n^2 on the right side, the left side is the sum of the first n odd numbers. The odd numbers are

$$\begin{aligned} 1 &= 2 \cdot 1 - 1 \\ 3 &= 2 \cdot 2 - 1 \\ 5 &= 2 \cdot 3 - 1 \\ 7 &= 2 \cdot 4 - 1 \\ 9 &= 2 \cdot 5 - 1 \end{aligned}$$

and from this it is clear that the nth odd number is $2n-1$. Hence, at least for $n = 1, 2, 3, 4$, or 5, the following is true:

$$1 + 3 + \cdots + (2n-1) = n^2 \qquad (S_n)$$

The question arises whether the statement S_n is true for *every* n. There is no hope of separately verifying all these statements because there are infinitely many of them. A more subtle approach is required.

The idea is as follows: Suppose it is verified that the statement S_{n+1} will be true whenever S_n is true. That is, suppose we prove that, *if* S_n is true, then it necessarily follows that S_{n+1} is also true. Then, if we can show that S_1 is true, it follows that S_2 is true, and from this that S_3 is true, hence that S_4 is true, and so on and on. This is the principle of induction. To express it more compactly, it is useful to have a short way to express the assertion "If S_n is true, then S_{n+1} is true." As in Appendix B, we write this assertion as

$$S_n \Rightarrow S_{n+1}$$

and read it as " S_n implies S_{n+1}." We can now state the principle of mathematical induction.

Theorem: The Principle of Mathematical Induction

Suppose S_n is a statement about the natural number n for each $n = 1, 2, 3, \ldots$.
Suppose further that:

1. *S_1 is true.*
2. *$S_n \Rightarrow S_{n+1}$ for every $n \geq 1$.*

Then S_n is true for every $n \geq 1$.

This is one of the most useful techniques in all of mathematics. It applies in a wide variety of situations, as the following examples illustrate.

Example C.1

Show that $1 + 2 + \cdots + n = \frac{1}{2}n(n+1)$ for $n \geq 1$.

Solution. Let S_n be the statement: $1 + 2 + \cdots + n = \frac{1}{2}n(n+1)$ for $n \geq 1$. We apply induction.

1. S_1 is true. The statement S_1 is $1 = \frac{1}{2}1(1+1)$, which is true.
2. $S_n \Rightarrow S_{n+1}$. We *assume* that S_n is true for some $n \geq 1$—that is, that

$$1 + 2 + \cdots + n = \tfrac{1}{2}n(n+1)$$

We must prove that the statement

$$S_{n+1} : 1 + 2 + \cdots + (n+1) = \tfrac{1}{2}(n+1)(n+2)$$

is also true, and we are entitled to use S_n to do so. Now the left side of S_{n+1} is the sum of the first $n+1$ positive integers. Hence the second-to-last term is n, so we can write

$$\begin{aligned} 1 + 2 + \cdots + (n+1) &= (1 + 2 + \cdots + n) + (n+1) \\ &= \tfrac{1}{2}n(n+1) + (n+1) \quad \text{using } S_n \\ &= \tfrac{1}{2}(n+1)(n+2) \end{aligned}$$

This shows that S_{n+1} is true and so completes the induction.

In the verification that $S_n \Rightarrow S_{n+1}$, we *assume* that S_n is true and use it to deduce that S_{n+1} is true. The assumption that S_n is true is sometimes called the **induction hypothesis**.

Example C.2

If x is any number such that $x \neq 1$, show that $1 + x + x^2 + \cdots + x^n = \frac{x^{n+1}-1}{x-1}$ for $n \geq 1$.

Solution. Let S_n be the statement: $1 + x + x^2 + \cdots + x^n = \frac{x^{n+1}-1}{x-1}$.

1. S_1 is true. S_1 reads $1+x=\frac{x^2-1}{x-1}$, which is true because $x^2-1=(x-1)(x+1)$.

2. $S_n \Rightarrow S_{n+1}$. Assume the truth of $S_n : 1+x+x^2+\cdots+x^n=\frac{x^{n+1}-1}{x-1}$.

We must *deduce* from this the truth of S_{n+1}: $1+x+x^2+\cdot+x^{n+1}=\frac{x^{n+2}-1}{x-1}$. Starting with the left side of S_{n+1} and using the induction hypothesis, we find

$$\begin{aligned}1+x+x^2+\cdots+x^{n+1} &= (1+x+x^2+\cdots+x^n)+x^{n+1}\\ &= \tfrac{x^{n+1}-1}{x-1}+x^{n+1}\\ &= \tfrac{x^{n+1}-1+x^{n+1}(x-1)}{x-1}\\ &= \tfrac{x^{n+2}-1}{x-1}\end{aligned}$$

This shows that S_{n+1} is true and so completes the induction.

Both of these examples involve formulas for a certain sum, and it is often convenient to use summation notation. For example, $\sum_{k=1}^{n}(2k-1)$ means that in the expression $(2k-1)$, k is to be given the values $k=1,\ k=2,\ k=3,\ \dots,\ k=n$, and then the resulting n numbers are to be added. The same thing applies to other expressions involving k. For example,

$$\sum_{k=1}^{n} k^3 = 1^3+2^3+\cdots+n^3$$

$$\sum_{k=1}^{5}(3k-1)=(3\cdot 1-1)+(3\cdot 2-1)+(3\cdot 3-1)+(3\cdot 4-1)+(3\cdot 5-1)$$

The next example involves this notation.

Example C.3

Show that $\sum_{k=1}^{n}(3k^2-k)=n^2(n+1)$ for each $n\geq 1$.

Solution. Let S_n be the statement: $\sum_{k=1}^{n}(3k^2-k)=n^2(n+1)$.

1. S_1 is true. S_1 reads $(3\cdot 1^2-1)=1^2(1+1)$, which is true.

2. $S_n \Rightarrow S_{n+1}$. Assume that S_n is true. We must prove S_{n+1}:

$$\begin{aligned}\sum_{k=1}^{n+1}(3k^2-k) &= \sum_{k=1}^{n}(3k^2-k)+[3(n+1)^2-(n+1)]\\ &= n^2(n+1)+(n+1)[3(n+1)-1] && \text{(using } S_n\text{)}\\ &= (n+1)[n^2+3n+2]\\ &= (n+1)[(n+1)(n+2)]\\ &= (n+1)^2(n+2)\end{aligned}$$

This proves that S_{n+1} is true.

We now turn to examples wherein induction is used to prove propositions that do not involve sums.

Example C.4

Show that $7^n + 2$ is a multiple of 3 for all $n \geq 1$.

Solution.

1. S_1 is true: $7^1 + 2 = 9$ is a multiple of 3.

2. $S_n \Rightarrow S_{n+1}$. Assume that $7^n + 2$ is a multiple of 3 for some $n \geq 1$; say, $7^n + 2 = 3m$ for some integer m. Then

$$7^{n+1} + 2 = 7(7^n) + 2 = 7(3m-2) + 2 = 21m - 12 = 3(7m-4)$$

so $7^{n+1} + 2$ is also a multiple of 3. This proves that S_{n+1} is true.

In all the foregoing examples, we have used the principle of induction starting at 1; that is, we have verified that S_1 is true and that $S_n \Rightarrow S_{n+1}$ for each $n \geq 1$, and then we have concluded that S_n is true for every $n \geq 1$. But there is nothing special about 1 here. If m is some fixed integer and we verify that

1. S_m is true.

2. $S_n \Rightarrow S_{n+1}$ for every $n \geq m$.

then it follows that S_n is true for every $n \geq m$. This "extended" induction principle is just as plausible as the induction principle and can, in fact, be proved by induction. The next example will illustrate it. Recall that if n is a positive integer, the number $n!$ (which is read "n-factorial") is the product

$$n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$$

of all the numbers from n to 1. Thus $2! = 2$, $3! = 6$, and so on.

Example C.5

Show that $2^n < n!$ for all $n \geq 4$.

Solution. Observe that $2^n < n!$ is actually false if $n = 1,\ 2,\ 3$.

1. S_4 is true. $2^4 = 16 < 24 = 4!$.

2. $S_n \Rightarrow S_{n+1}$ if $n \geq 4$. Assume that S_n is true; that is, $2^n < n!$. Then

$$\begin{array}{rcll} 2^{n+1} & = & 2 \cdot 2^n & \\ & < & 2 \cdot n! & \text{because } 2^n < n! \\ & < & (n+1)n! & \text{because } 2 < n+1 \\ & = & (n+1)! & \end{array}$$

Hence S_{n+1} is true.

Appendix D

Polynomials

Expressions like $3-5x$ and $1+3x-2x^2$ are examples of polynomials. In general, a **polynomial** is an expression of the form

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$$

where the a_i are numbers, called the **coefficients** of the polynomial, and x is a variable called an **indeterminate**. The number a_0 is called the **constant** coefficient of the polynomial. The polynomial with every coefficient zero is called the **zero polynomial**, and is denoted simply as 0.

If $f(x) \neq 0$, the coefficient of the highest power of x appearing in $f(x)$ is called the **leading** coefficient of $f(x)$, and the highest power itself is called the **degree** of the polynomial and is denoted $\deg(f(x))$. Hence

$$\begin{array}{ll} -1+5x+3x^2 & \text{has constant coefficient } -1, \text{ leading coefficient } 3, \text{ and degree } 2, \\ 7 & \text{has constant coefficient } 7, \text{ leading coefficient } 7, \text{ and degree } 0, \\ 6x-3x^3+x^4-x^5 & \text{has constant coefficient } 0, \text{ leading coefficient } -1, \text{ and degree } 5. \end{array}$$

We do not define the degree of the zero polynomial.

Two polynomials $f(x)$ and $g(x)$ are called **equal** if every coefficient of $f(x)$ is the same as the corresponding coefficient of $g(x)$. More precisely, if

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots \quad \text{and} \quad g(x) = b_0 + b_1x + b_2x^2 + \cdots$$

are polynomials, then

$$f(x) = g(x) \quad \text{if and only if} \quad a_0 = b_0,\ a_1 = b_1,\ a_2 = b_2,\ \dots$$

In particular, this means that

$$f(x) = 0 \text{ is the zero polynomial if and only if } a_0 = 0,\ a_1 = 0,\ a_2 = 0,\ \dots$$

This is the reason for calling x an indeterminate.

Let $f(x)$ and $g(x)$ denote nonzero polynomials of degrees n and m respectively, say

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \quad \text{and} \quad g(x) = b_0 + b_1x + b_2x^2 + \cdots + b_mx^m$$

where $a_n \neq 0$ and $b_m \neq 0$. If these expressions are multiplied, the result is

$$f(x)g(x) = a_0b_0 + (a_0b_1 + a_1b_0)x + (a_0b_2 + a_1b_1 + a_2b_0)x^2 + \cdots + a_nb_mx^{n+m}$$

Since a_n and b_m are nonzero numbers, their product $a_nb_m \neq 0$ and we have

Theorem D.1

If $f(x)$ and $g(x)$ are nonzero polynomials of degrees n and m respectively, their product $f(x)g(x)$ is also nonzero and

$$\deg[f(x)g(x)] = n + m$$

Example D.1

$(2-x+3x^2)(3+x^2-5x^3) = 6-3x+11x^2-11x^3+8x^4-15x^5.$

If $f(x)$ is any polynomial, the next theorem shows that $f(x)-f(a)$ is a multiple of the polynomial $x-a$. In fact we have

Theorem D.2: Remainder Theorem

If $f(x)$ is a polynomial of degree $n \geq 1$ and a is any number, then there exists a polynomial $q(x)$ such that

$$f(x) = (x-a)q(x) + f(a)$$

where $\deg(q(x)) = n-1$.

Proof. Write $f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ where the a_i are numbers, so that

$$f(a) = a_0 + a_1a + a_2a^2 + \cdots + a_na^n$$

If these expressions are subtracted, the constant terms cancel and we obtain

$$f(x) - f(a) = a_1(x-a) + a_2(x^2-a^2) + \cdots + a_n(x^n-a^n).$$

Hence it suffices to show that, for each $k \geq 1$, $x^k - a^k = (x-a)p(x)$ for some polynomial $p(x)$ of degree $k-1$. This is clear if $k=1$. If it holds for some value k, the fact that

$$x^{k+1} - a^{k+1} = (x-a)x^k + a(x^k - a^k)$$

shows that it holds for $k+1$. Hence the proof is complete by induction. □

There is a systematic procedure for finding the polynomial $q(x)$ in the remainder theorem. It is illustrated below for $f(x) = x^3 - 3x^2 + x - 1$ and $a = 2$. The polynomial $q(x)$ is generated on the top line one term at a time as follows: First x^2 is chosen because $x^2(x-2)$ has the same x^3-term as $f(x)$, and this is subtracted from $f(x)$ to leave a "remainder" of $-x^2+x-1$. Next, the second term on top is $-x$ because $-x(x-2)$ has the same x^2-term, and this is subtracted to leave $-x-1$. Finally, the third term on top is -1, and the process ends with a "remainder" of -3.

$$\begin{array}{r|rrrr}
 & & x^2 & -\;x & -1 \\ \cline{2-5}
x-2 & x^3 & -\;3x^2 & +\;x & -1 \\
 & x^3 & -\;2x^2 & & \\ \cline{2-5}
 & & -x^2 & +\;x & -1 \\
 & & -x^2 & +\;2x & \\ \cline{3-5}
 & & & -x & -1 \\
 & & & -x & +2 \\ \cline{4-5}
 & & & & -3
\end{array}$$

Hence $x^3 - 3x^2 + x - 1 = (x-2)(x^2 - x - 1) + (-3)$. The final remainder is $-3 = f(2)$ as is easily verified. This procedure is called the **division algorithm**.[1]

A real number a is called a **root** of the polynomial $f(x)$ if

$$f(a) = 0$$

Hence for example, 1 is a root of $f(x) = 2 - x + 3x^2 - 4x^3$, but -1 is not a root because $f(-1) = 10 \neq 0$. If $f(x)$ is a multiple of $x - a$, we say that $x - a$ is a **factor** of $f(x)$. Hence the remainder theorem shows immediately that if a is root of $f(x)$, then $x - a$ is factor of $f(x)$. But the converse is also true: If $x - a$ is a factor of $f(x)$, say $f(x) = (x-a)q(x)$, then $f(a) = (a-a)q(a) = 0$. This proves the

Theorem D.3: Factor Theorem

If $f(x)$ is a polynomial and a is a number, then $x - a$ is a factor of $f(x)$ if and only if a is a root of $f(x)$.

Example D.2

If $f(x) = x^3 - 2x^2 - 6x + 4$, then $f(-2) = 0$, so $x - (-2) = x + 2$ is a factor of $f(x)$. In fact, the division algorithm gives $f(x) = (x+2)(x^2 - 4x + 2)$.

Consider the polynomial $f(x) = x^3 - 3x + 2$. Then 1 is clearly a root of $f(x)$, and the division algorithm gives $f(x) = (x-1)(x^2 + x - 2)$. But 1 is also a root of $x^2 + x - 2$; in fact, $x^2 + x - 2 = (x-1)(x+2)$. Hence

$$f(x) = (x-1)^2(x+2)$$

and we say that the root 1 has **multiplicity** 2.

Note that non-zero constant polynomials $f(x) = b \neq 0$ have *no* roots. However, there do exist non-constant polynomials with no roots. For example, if $g(x) = x^2 + 1$, then $g(a) = a^2 + 1 \geq 1$ for every real number a, so a is not a root. However the *complex* number i is a root of $g(x)$; we return to this below.

Now suppose that $f(x)$ is any nonzero polynomial. We claim that it can be factored in the following form:

$$f(x) = (x - a_1)(x - a_2)\cdots(x - a_m)g(x)$$

where $a_1, a_2, \dots, a_m$ are the roots of $f(x)$ and $g(x)$ has no root (where the a_i may have repetitions, and may not appear at all if $f(x)$ has no real root).

By the above calculation $f(x) = x^3 - 3x + 2 = (x-1)^2(x+2)$ has roots 1 and -2, with 1 of multiplicity two (and $g(x) = 1$). Counting the root -2 once, we say that $f(x)$ has three roots counting multiplicities. The next theorem shows that no polynomial can have more roots than its degree even if multiplicities are counted.

Theorem D.4

If $f(x)$ is a nonzero polynomial of degree n, then $f(x)$ has at most n roots counting multiplicities.

[1] This procedure can be used to divide $f(x)$ by any nonzero polynomial $d(x)$ in place of $x - a$; the remainder then is a polynomial that is either zero or of degree less than the degree of $d(x)$.

Proof. If $n = 0$, then $f(x)$ is a constant and has no roots. So the theorem is true if $n = 0$. (It also holds for $n = 1$ because, if $f(x) = a + bx$ where $b \neq 0$, then the only root is $-\frac{a}{b}$.) In general, suppose inductively that the theorem holds for some value of $n \geq 0$, and let $f(x)$ have degree $n+1$. We must show that $f(x)$ has at most $n+1$ roots counting multiplicities. This is certainly true if $f(x)$ has no root. On the other hand, if a is a root of $f(x)$, the factor theorem shows that $f(x) = (x-a)q(x)$ for some polynomial $q(x)$, and $q(x)$ has degree n by Theorem D.1. By induction, $q(x)$ has at most n roots. But if b is any root of $f(x)$, then

$$(b-a)q(b) = f(b) = 0$$

so either $b = a$ or b is a root of $q(x)$. It follows that $f(x)$ has at most n roots. This completes the induction and so proves Theorem D.4. $\square$

As we have seen, a polynomial may have *no* root, for example $f(x) = x^2 + 1$. Of course $f(x)$ has complex roots i and $-i$, where i is the complex number such that $i^2 = -1$. But Theorem D.4 even holds for complex roots: the number of complex roots (counting multiplicities) cannot exceed the degree of the polynomial. Moreover, the fundamental theorem of algebra asserts that the only nonzero polynomials with no complex root are the non-zero constant polynomials. This is discussed more in Appendix A, Theorems A.4 and A.5.

Selected Exercise Answers

Index