Jin Ho Kwak
Sungpyo Hong

Linear Algebra

Birkhäuser
Boston • Basel • Berlin

Jin Ho Kwak
Sungpyo Hong
Department of Mathematics
Pohang University of Science and Technology
Pohang, The Republic of Korea

Library of Congress Cataloging-in-Publication Data

Kwak, Jin Ho, 1948-
Linear Algebra / Jin Ho Kwak, Sungpyo Hong.
p. cm.
Includes index.
ISBN 0-8176-3999-3 (alk. paper). -- ISBN 3-7643-3999-3 (alk. paper)
1. Algebras, Linear. I. Hong, Sungpyo, 1948- . II. Title.
QA184.K94 1997
512'.5--dc21 97-9062
CIP

Printed on acid-free paper

Birkhäuser

ISBN 0-8176-3999-3
ISBN 3-7643-3999-3
Typesetting by the authors in LaTeX
Printed and bound by Hamilton Printing, Rensselaer, NY
Printed in the U.S.A.

9 8 7 6 5 4 3 2 1

Preface

Linear algebra is one of the most important subjects in the study of science and engineering because of its widespread applications in social or natural science, computer science, physics, or economics. As one of the most useful courses in undergraduate mathematics, it has provided essential tools for industrial scientists. The basic concepts of linear algebra are vector spaces, linear transformations, matrices and determinants, and they serve as an abstract language for stating ideas and solving problems.

This book is based on the lectures delivered several years in a sophomore-level linear algebra course designed for science and engineering students. The primary purpose of this book is to give a careful presentation of the basic concepts of linear algebra as a coherent part of mathematics, and to illustrate its power and usefulness through applications to other disciplines. We have tried to emphasize the computational skills along with the mathematical abstractions, which have also an integrity and beauty of their own. The book includes a variety of interesting applications with many examples not only to help students understand new concepts but also to practice wide applications of the subject to such areas as differential equations, statistics, geometry, and physics. Some of those applications may not be central to the mathematical development and may be omitted or selected in a syllabus at the discretion of the instructor. Most basic concepts and introductory motivations begin with examples in Euclidean space or solving a system of linear equations, and are gradually examined from different points of views to derive general principles.

For those students who have completed a year of calculus, linear algebra may be the first course in which the subject is developed in an abstract way, and we often find that many students struggle with the abstraction and miss the applications. Our experience is that, to understand the material, students should practice with many problems, which are sometimes omitted because of a lack of time. To encourage the students to do repeated practice, we placed in the middle of the text not only many examples but also some carefully selected problems, with answers or helpful hints. We have tried to make this book as easily accessible and clear as possible, but certainly there may be some awkward expressions in several ways. Any criticism or comment from the readers will be appreciated.

We are very grateful to many colleagues in Korea, especially to the faculty members in the mathematics department at Pohang University of Science and Technology (POSTECH), who helped us over the years with various aspects of this book. For their valuable suggestions and comments, we would like to thank the students at POSTECH, who have used photocopied versions of the text over the past several years. We would also like to acknowledge the invaluable assistance we have received from the teaching assistants who have checked and added some answers or hints for the problems and exercises in this book. Our thanks also go to Mrs. Kathleen Roush who made this book much more legible with her grammatical corrections in the final manuscript. Our thanks finally go to the editing staff of Birkhäuser for gladly accepting our book for publication.

Jin Ho Kwak
Sungpyo Hong
E-mail: jinkwak@postech.ac.kr
sungpyo@postech.ac.kr

April 1997, in Pohang, Korea

"Linear algebra is the mathematics of our modern technological world of complex multivariable systems and computers"

– Alan Tucker –

"We (Halmos and Kaplansky) share a love of linear algebra. I think it is our conviction that we'll never understand infinite-dimensional operators properly until we have a decent mastery of finite matrices. And we share a philosophy about linear algebra: we think basis-free, we write basis-free, but when the chips are down we close the office door and compute with matrices like fury"

– Irving Kaplansky –

Contents

Linear Algebra

Chapter 1

Linear Equations and Matrices

1.1 Introduction

One of the central motivations for linear algebra is solving systems of linear equations. We thus begin with the problem of finding the solutions of a system of m linear equations in n unknowns of the following form:

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

$= 0 \Rightarrow$ homogeneous

where $x_1,\ x_2,\ \ldots,\ x_n$ are the unknowns and a_{ij}'s and b_i's denote constant (real or complex) numbers.

A sequence of numbers $(s_1,\ s_2,\ \ldots,\ s_n)$ is called a **solution** of the system if $x_1 = s_1,\ x_2 = s_2,\ \ldots,\ x_n = s_n$ satisfy each equation in the system simultaneously. When $b_1 = b_2 = \cdots = b_m = 0$, we say that the system is **homogeneous**.

The central topic of this chapter is to examine whether or not a given system has a solution, and to find a solution if it has one. For instance, any homogeneous system always has at least one solution $x_1 = x_2 = \cdots = x_n = 0$, called the **trivial solution**. A natural question is whether such a homogeneous system has a nontrivial solution. If so, we would like to have a systematic method of finding all the solutions. A system of linear equations is said to be **consistent** if it has at least one solution, and **inconsistent** if it has no solution. The following example gives us an idea how to answer the above questions.

Example 1.1 When $m = n = 2$, the system reduces to two equations in two unknowns x and y:

$$\begin{cases} a_1x + b_1y = c_1 \\ a_2x + b_2y = c_2. \end{cases}$$

Geometrically, each equation in the system represents a straight line when we interpret x and y as coordinates in the xy-plane. Therefore, a point $P = (x, y)$ is a solution if and only if the point P lies on both lines. Hence there are three possible types of solution set:

(1) the empty set if the lines are parallel,

(2) only one point if they intersect,

(3) a straight line: *i.e.*, infinitely many solutions, if they coincide.

The following examples and diagrams illustrate the three types:

Case (1)

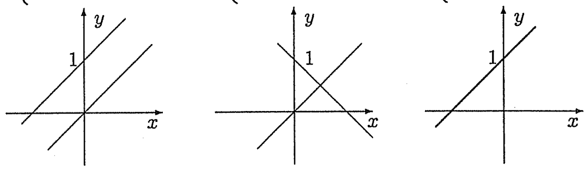
$$\begin{cases} x - y = -1 \\ x - y = 0 \end{cases}$$

Case (2)

$$\begin{cases} x + y = 1 \\ x - y = 0 \end{cases}$$

Case (3)

$$\begin{cases} x - y = -1 \\ 2x - 2y = -2 \end{cases}$$



To decide whether the given system has a solution and to find a general method of solving the system when it has a solution, we repeat here a well-known elementary method of *elimination* and *substitution*.

Suppose first that the system consists of only one equation $ax + by = c$. Then the system has either infinitely many solutions (*i.e.*, points on the straight line $x = -\frac{b}{a}y + \frac{c}{a}$ or $y = -\frac{a}{b}x + \frac{c}{b}$ depending on whether $a \neq 0$ or $b \neq 0$) or no solutions when $a = b = 0$ and $c \neq 0$.

We now assume that the system has two equations representing two lines in the plane. Then clearly the two lines are parallel with the same slopes if and only if $a_2 = \lambda a_1$ and $b_2 = \lambda b_1$ for some $\lambda \neq 0$, or $a_1b_2 - a_2b_1 = 0$. Furthermore, the two lines either coincide (infinitely many solutions) or are distinct and parallel (no solutions) according to whether $c_2 = \lambda c_1$ holds or not.

Suppose now that the lines are not parallel, or $a_1b_2 - a_2b_1 \neq 0$. In this case, the two lines cross at a point, and hence there is exactly one solution: For instance, if the system is homogeneous, then the lines cross at the origin, so $(0,0)$ is the only solution. For a nonhomogeneous system, we may find the solution as follows: Express x in terms of y from the first equation, and then substitute it into the second equation (*i.e.*, eliminate the variable x from the second equation) to get

$$(b_2 - \frac{a_2}{a_1}b_1)y = c_2 - \frac{a_2}{a_1}c_1.$$

Since $a_1b_2 - a_2b_1 \neq 0$, this can be solved as

$$y = \frac{a_1c_2 - a_2c_1}{a_1b_2 - a_2b_1},$$

which is in turn substituted into one of the equations to find x and give a complete solution of the system. In detail, the process can be summarized as follows:

(1) Without loss of generality, we may assume $a_1 \neq 0$ since otherwise we can interchange the two equations. Then the variable x can be eliminated from the second equation by adding $-\frac{a_2}{a_1}$ times the first equation to the second, to get

$$\left\{\begin{array}{rcrcl} a_1x & + & b_1y & = & c_1 \\ & & (b_2 - \frac{a_2}{a_1}b_1)y & = & c_2 - \frac{a_2}{a_1}c_1. \end{array}\right.$$

(2) Since $a_1b_2 - a_2b_1 \neq 0$, y can be found by multiplying the second equation by a nonzero number $\frac{a_1}{a_1b_2 - a_2b_1}$ to get

$$\left\{\begin{array}{rcrcl} a_1x & + & b_1y & = & c_1 \\ & & y & = & \dfrac{a_1c_2 - a_2c_1}{a_1b_2 - a_2b_1}. \end{array}\right.$$

(3) Now, x is solved by substituting the value of y into the first equation, and we obtain the solution to the problem:

$$\begin{cases} x &= \dfrac{b_2c_1 - b_1c_2}{a_1b_2 - a_2b_1} \\ y &= \dfrac{a_1c_2 - a_2c_1}{a_1b_2 - a_2b_1}. \end{cases}$$

Note that the condition $a_1b_2 - a_2b_1 \neq 0$ is necessary for the system to have only one solution. $\square$

In this example, we have changed the original system of equations into a simpler one using certain operations, from which we can get the solution of the given system. That is, if (x, y) satisfies the original system of equations, then x and y must satisfy the above simpler system in (3), and vice versa.

It is suggested that the readers examine a system of three equations in three unknowns, each equation representing a plane in the 3-dimensional space $\mathbb{R}^3$, and consider the various possible cases in a similar way.

Problem 1.1 For a system of three equations in three unknowns

$$\begin{cases} a_{11}x + a_{12}y + a_{13}z = b_1 \\ a_{21}x + a_{22}y + a_{23}z = b_2 \\ a_{31}x + a_{32}y + a_{33}z = b_3, \end{cases}$$

describe all the possible types of the solution set in $\mathbb{R}^3$.

1.2 Gaussian elimination

As we have seen in Example 1.1, a basic idea for solving a system of linear equations is to change the given system into a simpler system, keeping the solutions unchanged; the example showed how to change a general system to a simpler one. In fact, the main operations used in Example 1.1 are the following three operations, called **elementary operations**:

(1) multiply a nonzero constant throughout an equation,

(2) interchange two equations,

(3) change an equation by adding a constant multiple of another equation.

After applying a finite sequence of these elementary operations to the given system, one can obtain a simpler system from which the solution can be derived directly.

Note also that each of the three elementary operations has its *inverse* operation which is also an elementary operation:

(1)′ divide the equation with the same nonzero constant,
(2)′ interchange two equations again,
(3)′ change the equation by subtracting the same constant multiple of the same equation.

By applying these inverse operations in reverse order to the simpler system, one can recover the original system. This means that a solution of the original system must also be a solution of the simpler one, and *vice versa*.

These arguments can be formalized in mathematical language. Observe that in performing any of these basic operations, only the coefficients of the variables are involved in the calculations and the variables $x_1, \ldots, x_n$ and the equal sign "=" are simply repeated. Thus, keeping the order of the variables and "=" in mind, we just extract the coefficients only from the equations in the given system and make a rectangular array of numbers:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{bmatrix}.$$

This matrix is called the **augmented matrix** for the system. The term *matrix* means just any rectangular array of numbers, and the numbers in this array are called the *entries* of the matrix. To explain the above operations in terms of matrices, we first introduce some terminology even though in the following sections we shall study matrices in more detail.

Within a matrix, the horizontal and vertical subarrays

$$[a_{i1}\ a_{i2}\ \cdots\ a_{in}\ b_i] \quad \text{and} \quad \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix}$$

are called the i-th *row* (matrix) and the j-th *column* (matrix) of the augmented matrix, respectively. Note that the entries in the j-th column are just the coefficients of j-th variable x_j, so there is a correspondence between columns of the matrix and variables of the system.

Since each row of the augmented matrix contains all the information of the corresponding equation of the system, we may deal with this augmented matrix instead of handling the whole system of linear equations.

The elementary operations to a system of linear equations are rephrased as the **elementary row operations** for the augmented matrix, as follows:

(1) multiply a nonzero constant throughout a row,

(2) interchange two rows,

(3) change a row by adding a constant multiple of another row.

The *inverse* operations are

(1)′ divide the row by the same constant,

(2)′ interchange two rows again,

(3)′ change the row by subtracting the same constant multiple of the other row.

Definition 1.1 Two augmented matrices (or systems of linear equations) are said to be **row-equivalent** if one can be transformed to the other by a finite sequence of elementary row operations.

If a matrix B can be obtained from a matrix A in this way, then we can obviously recover A from B by applying the inverse elementary row operations in reverse order. Note again that an elementary row operation does not alter the solution of the system, and we can formalize the above argument in the following theorem:

Theorem 1.1 *If two systems of linear equations are row-equivalent, then they have the same set of solutions.*

The general procedure for finding the solutions will be illustrated in the following example:

Example 1.2 Solve the system of linear equations:

$$\left\{\begin{array}{rcrcrcr} & & 2y & + & 4z & = & 2 \\ x & + & 2y & + & 2z & = & 3 \\ 3x & + & 4y & + & 6z & = & -1. \end{array}\right.$$

Solution: We could work with the augmented matrix alone. However, to compare the operations on systems of linear equations with those on the augmented matrix, we work on the system and the augmented matrix in parallel. Note that the associated augmented matrix of the system is

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

(1) Since the coefficient of x in the first equation is zero while that in the second equation is not zero, we interchange these two equations:

$$\left\{\begin{array}{rcrcrcr} x & + & 2y & + & 2z & = & 3 \\ & & 2y & + & 4z & = & 2 \\ 3x & + & 4y & + & 6z & = & -1 \end{array}\right. \qquad \begin{bmatrix} 1 & 2 & 2 & 3 \\ 0 & 2 & 4 & 2 \\ 3 & 4 & 6 & -1 \end{bmatrix}.$$

(2) Add -3 times the first equation to the third equation:

$$\left\{\begin{array}{rcrcrcr} x & + & 2y & + & 2z & = & 3 \\ & & 2y & + & 4z & = & 2 \\ & - & 2y & & & = & -10 \end{array}\right. \qquad \begin{bmatrix} 1 & 2 & 2 & 3 \\ 0 & 2 & 4 & 2 \\ 0 & -2 & 0 & -10 \end{bmatrix}.$$

The coefficient 1 of the first unknown x in the first equation (row) is called the **pivot** in this first elimination step.

Now the second and the third equations involve only the two unknowns y and z. Leave the first equation (row) alone, and the same elimination procedure can be applied to the second and the third equations (rows): The pivot for this step is the coefficient 2 of y in the second equation (row). To eliminate y from the last equation,

(3) Add 1 times the second equation (row) to the third equation (row):

$$\left\{\begin{array}{rcrcrcr} x & + & 2y & + & 2z & = & 3 \\ & & 2y & + & 4z & = & 2 \\ & & & & 4z & = & -8 \end{array}\right. \qquad \begin{bmatrix} 1 & 2 & 2 & 3 \\ 0 & 2 & 4 & 2 \\ 0 & 0 & 4 & -8 \end{bmatrix}.$$

The elimination process done so far to obtain this result is called a **forward elimination**: *i.e.*, elimination of x from the last two equations (rows) and then elimination of y from the last equation (row).

Now the pivots of the second and third rows are 2 and 4, respectively. To make these entries 1,

(4) Divide each row by the pivot of the row:

$$\begin{cases} x + 2y + 2z = 3 \\ \quad\; y + 2z = 1 \\ \qquad\quad z = -2 \end{cases} \qquad \begin{bmatrix} 1 & 2 & 2 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix}.$$

The resulting matrix on the right side is called a **row-echelon form** of the matrix, and the 1's at the leftmost entries in each row are called the **leading** 1's. The process so far is called a **Gaussian elimination.**

We now want to eliminate numbers above the leading 1's;

(5) Add -2 times the third row to the second and the first rows,

$$\begin{cases} x + 2y = 7 \\ \quad\; y = 5 \\ \qquad z = -2 \end{cases} \qquad \begin{bmatrix} 1 & 2 & 0 & 7 \\ 0 & 1 & 0 & 5 \\ 0 & 0 & 1 & -2 \end{bmatrix}.$$

(6) Add -2 times the second row to the first row:

$$\begin{cases} x = -3 \\ y = 5 \\ z = -2 \end{cases} \qquad \begin{bmatrix} 1 & 0 & 0 & -3 \\ 0 & 1 & 0 & 5 \\ 0 & 0 & 1 & -2 \end{bmatrix}.$$

This matrix is called the **reduced row-echelon form.** The procedure to get this reduced row-echelon form from a row-echelon form is called the **back substitution.** The whole process to obtain the reduced row-echelon form is called a **Gauss-Jordan elimination.**

Notice that the corresponding system to this reduced row-echelon form is row-equivalent to the original one and is essentially a solved form: *i.e.*, the solution is $x = -3,\ y = 5,\ z = -2$. □

In general, a matrix of **row-echelon form** satisfies the following properties.

(1) The first nonzero entry of each row is 1, called a leading 1.

(2) A row containing only 0's should come after all rows with some nonzero entries.

(3) The leading 1's appear from left to the right in successive rows. That is, the leading 1 in the lower row occurs farther to the right than the leading 1 in the higher row.

Moreover, the matrix of the **reduced row-echelon form** satisfies

(4) Each column that contains a leading 1 has zeros everywhere else, in addition to the above three properties.

Note that an augmented matrix has only one reduced row-echelon form while it may have many row-echelon forms. In any case, the number of nonzero rows containing leading 1's is equal to the number of columns containing leading 1's. The variables in the system corresponding to columns with the leading 1's in a row-echelon form are called the **basic variables**. In general, the reduced row-echelon form U may have columns that do not contain leading 1's. The variables in the system corresponding to the columns without leading 1's are called **free variables**. Thus the sum of the number of basic variables and that of free variables is precisely the total number of variables.

For example, the first two matrices below are in reduced row-echelon form, and the last two just in row-echelon form.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 & 5 & 0 & 6 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 2 & 3 & 2 \\ 0 & 1 & 4 & 5 \\ 0 & 0 & 1 & 7 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 & 2 & 6 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 3 \end{bmatrix}.$$

Notice that in an augmented matrix $[A\ \mathbf{b}]$, the last column $\mathbf{b}$ does not correspond to any variable. Hence, if we consider the four matrices above as augmented matrices for some systems, then the systems corresponding to the first and the last two augmented matrices have only basic variables but no free variables. In the system corresponding to the second augmented matrix, the second and the forth variables, x_2 and x_4, are basic, and the first and the third variables, x_1 and x_3, are free variables. These ideas will be used in later chapters.

In summary, by applying a finite sequence of elementary row operations, the augmented matrix for a system of linear equations can be changed to its reduced row-echelon form which is row-equivalent to the original one. From the reduced row-echelon form, we can decide whether the system has a solution, and find the solution of the given system if it has one.

Example 1.3 Solve the following system of linear equations by Gauss-Jordan elimination.

$$\left\{ \begin{array}{rcrcrcrcr} x_1 & + & 3x_2 & - & 2x_3 & & & = & 3 \\ 2x_1 & + & 6x_2 & - & 2x_3 & + & 4x_4 & = & 18 \\ & & x_2 & + & x_3 & + & 3x_4 & = & 10. \end{array} \right.$$

Solution: The augmented matrix for the system is

$$\begin{bmatrix} 1 & 3 & -2 & 0 & 3 \\ 2 & 6 & -2 & 4 & 18 \\ 0 & 1 & 1 & 3 & 10 \end{bmatrix}.$$

The Gaussian elimination begins with:

(1) Adding -2 times the first row to the second produces

$$\begin{bmatrix} 1 & 3 & -2 & 0 & 3 \\ 0 & 0 & 2 & 4 & 12 \\ 0 & 1 & 1 & 3 & 10 \end{bmatrix}.$$

(2) Note that the coefficient of x_2 in the second equation is zero and that in the third equation is not. Thus, interchanging the second and the third rows produces

$$\begin{bmatrix} 1 & 3 & -2 & 0 & 3 \\ 0 & 1 & 1 & 3 & 10 \\ 0 & 0 & 2 & 4 & 12 \end{bmatrix}.$$

(3) The pivot in the third row is 2. Thus, dividing the third row by 2 produces a row-echelon form

$$\begin{bmatrix} 1 & 3 & -2 & 0 & 3 \\ 0 & 1 & 1 & 3 & 10 \\ 0 & 0 & 1 & 2 & 6 \end{bmatrix}.$$

This is a row-echelon form, and we now continue the back-substitution:

(4) Adding -1 times the third row to the second, and 2 times the third row to the first produces

$$\begin{bmatrix} 1 & 3 & 0 & 4 & 15 \\ 0 & 1 & 0 & 1 & 4 \\ 0 & 0 & 1 & 2 & 6 \end{bmatrix}.$$

(5) Finally, adding -3 times the second row to the first produces the reduced row-echelon form:

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 3 \\ 0 & 1 & 0 & 1 & 4 \\ 0 & 0 & 1 & 2 & 6 \end{bmatrix}.$$

The corresponding system of equations is

$$\begin{cases} x_1 & & & + & x_4 & = & 3 \\ & x_2 & & + & x_4 & = & 4 \\ & & x_3 & + & 2x_4 & = & 6. \end{cases}$$

Since x_1, x_2, and x_3 correspond to the columns containing leading 1's, they are the basic variables, and x_4 is the free variable. Thus by solving this system for the basic variables in terms of the free variable x_4, we have the system of equations in a solved form:

$$\begin{cases} x_1 & = & 3 & - & x_4 \\ x_2 & = & 4 & - & x_4 \\ x_3 & = & 6 & - & 2x_4. \end{cases}$$

By assigning an arbitrary value t to the free variable x_4, the solutions can be written as

$$(x_1,\ x_2,\ x_3,\ x_4) = (3-t,\ 4-t,\ 6-2t,\ t),$$

for any $t \in \mathbb{R}$, where $\mathbb{R}$ denotes the set of real numbers. □

Remark: Consider a homogeneous system

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & 0 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & 0 \\ & & & & & & & \vdots & \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & 0, \end{cases}$$

with the number of unknowns greater than the number of equations: that is, $m < n$. Since the number of basic variables cannot exceed the number of rows, a free variable always exists as in Example 1.3, so by assigning an arbitrary value to each free variable we can always find infinitely many nontrivial solutions.

Problem 1.2 Suppose that the augmented matrix for a system of linear equations has been reduced to the reduced row-echelon form below by elementary row operations. Solve the systems:

$$(1) \begin{bmatrix} 1 & 0 & 0 & 5 \\ 0 & 1 & 0 & -2 \\ 0 & 0 & 0 & 4 \end{bmatrix}, \qquad (2) \begin{bmatrix} 1 & 0 & 0 & 4 & -1 \\ 0 & 1 & 0 & 2 & 6 \\ 0 & 0 & 1 & 3 & 2 \end{bmatrix}.$$

We note that if a row-echelon form of an augmented matrix has a row of the type $[\,0\ 0\ \cdots\ 0\ b\,]$ with $b \neq 0$, then it represents an equation of the form $0x_1 + 0x_2 + \cdots + 0x_n = b$ with $b \neq 0$. In this case, the system has no solution. If $b = 0$, then it has a row containing only 0's that can be neglected. Hence, when we deal with a row-echelon form, we may assume that the zero rows are deleted. Note also that, as in Example 1.3, if there exists at least one free variable in the row-echelon form, then the system has infinitely many solutions. On the other hand, if the system has no free variable, the system has a unique solution.

To study systems of linear equations in terms of matrices systematically, we will develop some general theories of matrices in the following sections.

Problem 1.3 Solve the following systems of equations by Gaussian elimination. What are the pivots?

$$(1) \left\{ \begin{array}{rcrcrcl} -x & + & y & + & 2z & = & 0 \\ 3x & + & 4y & + & z & = & 0 \\ 2x & + & 5y & + & 3z & = & 0. \end{array} \right. \qquad (2) \left\{ \begin{array}{rcrcrcl} & & 2y & - & z & = & 1 \\ 4x & - & 10y & + & 3z & = & 5 \\ 3x & - & 3y & & & = & 6. \end{array} \right.$$

$$(3) \left\{ \begin{array}{rcrcrcrcl} w & + & x & + & y & & & = & 3 \\ -3w & - & 17x & + & y & + & 2z & = & 1 \\ 4w & - & 17x & + & 8y & - & 5z & = & 1 \\ & - & 5x & - & 2y & + & z & = & 1. \end{array} \right.$$

Problem 1.4 Determine the condition on b_i so that the following system has a solution.

$$(1) \left\{ \begin{array}{rcrcrcl} x & + & 2y & + & 6z & = & b_1 \\ 2x & - & 3y & - & 2z & = & b_2 \\ 3x & - & y & + & 4z & = & b_3. \end{array} \right. \qquad (2) \left\{ \begin{array}{rcrcrcl} x & + & 3y & - & 2z & = & b_1 \\ 2x & - & y & + & 3z & = & b_2 \\ 4x & + & 2y & + & z & = & b_3. \end{array} \right.$$

1.3 Matrices

Rectangular arrays of real numbers arise in many real-world problems. Historically, it was the English mathematician A. Cayley who first introduced the word "matrix" in the year 1858. The meaning of the word is "that within which something originates," and he used matrices simply as a source for rows and columns to form squares.

In this section we are interested only in very basic properties of such matrices.

Definition 1.2 An m by n (written $m \times n$) **matrix** is a rectangular array of numbers arranged into m (horizontal) rows and n (vertical) columns. The

size of a matrix is specified by the number m of rows and the number n of columns.

In general, a matrix is written in the following form:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = [a_{ij}]_{m\times n},$$

or just $A = [a_{ij}]$ if the size of the matrix is clear from the context. The number a_{ij} is called the (i, j)-**entry** of the matrix A, and can be also written as $a_{ij} = [A]_{ij}$.

An $m\times 1$ matrix is called a **column (matrix)** or sometimes a **column vector**, and a $1\times n$ matrix is called a **row (matrix)**, or a **row vector**. These special cases are important, as we will see throughout the book. We will generally use capital letters like A, B, C for matrices and small boldface letters like $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ for columns or row vectors.

Definition 1.3 Let $A = [a_{ij}]$ be an $m\times n$ matrix. The **transpose** of A is the $n\times m$ matrix, denoted by A^T, whose j-th column is taken from the j-th row of A: That is, $A^T = [b_{ij}]$ with $b_{ij} = a_{ji}$.

For example, if $A = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$, then $A^T = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$.

In particular, the transpose of a column vector is a row vector and vice versa. For example, for an $n \times 1$ column vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

its transpose $\mathbf{x}^T = [x_1\ x_2\ \cdots\ x_n]$ is a row vector.

Definition 1.4 Let $A = [a_{ij}]$ be an $m \times n$ matrix.

(1) A is called a **square matrix of order** n if $m = n$.

In the following, we assume that A is a square matrix of order n.

(2) The entries a_{11}, a_{22}, $\ldots$, a_{nn} are called the **diagonal entries** of A.

(3) A is called a **diagonal** matrix if all the entries except for the diagonal entries are zero.

(4) A is called an **upper (lower) triangular** matrix if all the entries below (above, respectively) the diagonal are zero.

The following matrices U and L are the general forms of the upper triangular and lower triangular matrices, respectively:

$$U = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}, \quad L = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}.$$

Note that a matrix which is both upper and lower triangular must be a diagonal matrix, and the transpose of an upper (lower) triangular matrix is lower (upper, respectively) triangular matrix.

Definition 1.5 Two matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ are said to be **equal**, written $A = B$, if they have the same size and corresponding entries are equal: *i.e.*, $a_{ij} = b_{ij}$ for all i and j.

This definition allows us to write matrix equations. A simple example is $(A^T)^T = A$ by definition.

Let $M_{m\times n}(\mathbb{R})$ denote the set of all $m \times n$ matrices with entries of real numbers. Among the elements of $M_{m\times n}(\mathbb{R})$, we can define two operations, called scalar multiplication and the sum of matrices, as follows:

Scalar multiplication: Given an $m \times n$ matrix $A = [a_{ij}]$ and a *scalar* k (which is simply a real number), the scalar multiplication kA of k and A is defined to be the matrix $kA = [ka_{ij}]$: *i.e.*, in an expanded form:

$$k \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} ka_{11} & \cdots & ka_{1n} \\ \vdots & \ddots & \vdots \\ ka_{m1} & \cdots & ka_{mn} \end{bmatrix}.$$

Sum of matrices: If $A = [a_{ij}]$ and $B = [b_{ij}]$ are two matrices of the same size, then the **sum** $A + B$ is defined to be the matrix $A + B = [a_{ij} + b_{ij}]$:

i.e., in an expanded form:

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{mn} \end{bmatrix} = \begin{bmatrix} a_{11}+b_{11} & \cdots & a_{1n}+b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1}+b_{m1} & \cdots & a_{mn}+b_{mn} \end{bmatrix}.$$

Note that matrices of different sizes cannot be added. It is quite clear that $A+A=2A$, and $A+(A+A)=(A+A)+A=3A$. Thus, inductively we define $nA=(n-1)A+A$ for any positive integer n. If B is any matrix, then $-B$ is by definition the multiplication $(-1)B$. Moreover, if A and B are two matrices of the same size, then the difference $A-B$ is by definition the sum $A+(-1)B=A+(-B)$. A matrix whose entries are all zero is called a **zero matrix**, denoted by the symbol $\mathbf{0}$ (or $\mathbf{0}_{m\times n}$ when we emphasize the number of rows and columns).

Clearly, matrix addition has the same properties as the addition of real numbers. The real numbers in the context here are traditionally called **scalars** even though "numbers" is a perfectly good name and "scalar" sounds more technical. The following theorem lists some basic rules of these operations.

Theorem 1.2 *Suppose that the sizes of A, B and C are the same. Then the following rules of matrix arithmetic are valid:*

(1) $(A+B)+C=A+(B+C)$, *(written as $A+B+C$)* (Associativity),

(2) $A+\mathbf{0}=\mathbf{0}+A=A$,

(3) $A+(-A)=(-A)+A=\mathbf{0}$,

(4) $A+B=B+A$, (Commutativity),

(5) $k(A+B)=kA+kB$,

(6) $(k+\ell)A=kA+\ell A$,

(7) $(k\ell)A=k(\ell A)$.

Proof: We prove only (5) and the remaining are left for exercises. For any (i,j),

$$[k(A+B)]_{ij}=k[A+B]_{ij}=k([A]_{ij}+[B]_{ij})=[kA]_{ij}+[kB]_{ij}.$$

Consequently, $k(A+B)=kA+kB$. □

Definition 1.6 A square matrix A is said to be **symmetric** if $A^T = A$, or **skew-symmetric** if $A^T = -A$.

For example, the matrices A and B below

$$A = \begin{bmatrix} 1 & a & b \\ a & 3 & c \\ b & c & 5 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & 3 \\ -2 & -3 & 0 \end{bmatrix}$$

are symmetric and skew-symmetric, respectively. Notice here that all the diagonal entries of a skew-symmetric matrix must be zero, since $a_{ii} = -a_{ii}$.

By a direct computation, one can easily verify the following rules of the transpose of matrices:

Theorem 1.3 *Let A and B be $m \times n$ matrices. Then*

$$(kA)^T = kA^T, \quad \textit{and} \quad (A+B)^T = A^T + B^T.$$

Problem 1.5 Prove the remaining parts of Theorem 1.2.

Problem 1.6 Find a matrix B such that $A + B^T = (A - B)^T$, where

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

Problem 1.7 Find a, b, c and d such that

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = 2\begin{bmatrix} a & 3 \\ 2 & a+c \end{bmatrix} + \begin{bmatrix} 2+b & a+9 \\ c+d & b \end{bmatrix}.$$

1.4 Products of matrices

We introduced the operations sum and scalar multiplication of matrices in Section 1.3. In this section, we introduce the product of matrices. Unlike the sum of two matrices, the product of matrices is a little bit more complicated, in the sense that it is defined for two matrices of different sizes or for square matrices of the same order. We define the product of matrices in three steps:

(1) For a $1 \times n$ row matrix $\mathbf{a} = [a_1 \ \cdots \ a_n]$ and an $n \times 1$ column matrix $\mathbf{x} = [x_1 \ \cdots \ x_n]^T$, the **product ax** is a 1×1 matrix (*i.e.*, just a number) defined by the rule

$$\mathbf{ax} = [a_1 \ a_2 \ \cdots \ a_n] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = [a_1x_1 + a_2x_2 + \cdots + a_nx_n] = \left[\sum_{i=1}^{n} a_i x_i\right].$$

Note that the number of columns of the first matrix must be equal to the number of rows of the second matrix to have entrywise multiplications of the entries.

(2) For an $m \times n$ matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_m \end{bmatrix},$$

where $\mathbf{a}_i$'s denote the row vectors, and for an $n \times 1$ column matrix $\mathbf{x} = [x_1 \ \cdots \ x_n]^T$, the **product** $A\mathbf{x}$ is by definition an $m \times 1$ matrix defined by the rule:

$$A\mathbf{x} = \begin{bmatrix} \mathbf{a}_1\mathbf{x} \\ \mathbf{a}_2\mathbf{x} \\ \vdots \\ \mathbf{a}_m\mathbf{x} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} a_{1i}x_i \\ \sum_{i=1}^{n} a_{2i}x_i \\ \vdots \\ \sum_{i=1}^{n} a_{mi}x_i \end{bmatrix},$$

or in an expanded form

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{bmatrix},$$

which is just an $m \times 1$ column matrix of the form $[b_1 \, b_2 \, \cdots \, b_m]^T$.

Therefore, for a system of m linear equations in n unknowns, by writing the n unknowns as an $n \times 1$ column matrix $\mathbf{x}$ and the coefficients as an $m \times n$ matrix A the system may be expressed as a matrix equation $A\mathbf{x} = \mathbf{b}$. Notice that this looks just like the usual linear equation in one variable: $ax = b$.

(3) Product of matrices: Let A be an $m\times n$ matrix and B an $n\times r$ matrix. The **product** AB is defined to be an $m\times r$ matrix whose columns are the products of A and the columns of B in corresponding order.

Thus if A is $m\times n$ and B is $n\times r$, then B has r columns and each column of B is an $n\times 1$ matrix. If we denote them by $\mathbf{b}^1, \ldots, \mathbf{b}^r$, or $B = [\mathbf{b}^1 \cdots \mathbf{b}^r]$, then

$$\begin{aligned} AB &= \begin{bmatrix} A\mathbf{b}^1 & A\mathbf{b}^2 & \cdots & A\mathbf{b}^r \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{a}_1\mathbf{b}^1 & \mathbf{a}_1\mathbf{b}^2 & \cdots & \mathbf{a}_1\mathbf{b}^r \\ \mathbf{a}_2\mathbf{b}^1 & \mathbf{a}_2\mathbf{b}^2 & \cdots & \mathbf{a}_2\mathbf{b}^r \\ \vdots & & \ddots & \vdots \\ \mathbf{a}_m\mathbf{b}^1 & \mathbf{a}_m\mathbf{b}^2 & \cdots & \mathbf{a}_m\mathbf{b}^r \end{bmatrix}, \end{aligned}$$

which is an $m\times r$ matrix. Therefore, the (i,j)-entry $[AB]_{ij}$ of AB is the i-th entry of the j-th column matrix

$$A\mathbf{b}^j = \begin{bmatrix} \mathbf{a}_1\mathbf{b}^j \\ \mathbf{a}_2\mathbf{b}^j \\ \vdots \\ \mathbf{a}_m\mathbf{b}^j \end{bmatrix},$$

i.e., for $i = 1, \ldots, m$ and $j = 1, \ldots, r$, it is the product of i-th row and j-th column of A:

$$[AB]_{ij} = \mathbf{a}_i\mathbf{b}^j = \sum_{k=1}^{n} a_{ik}b_{kj}.$$

Example 1.4 Consider the matrices

$$A = \begin{bmatrix} 2 & 3 \\ 4 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 2 & 0 \\ 5 & -1 & 0 \end{bmatrix}.$$

The columns of AB are the product of A and each column of B:

$$\begin{aligned} \begin{bmatrix} 2 & 3 \\ 4 & 0 \end{bmatrix}\begin{bmatrix} 1 \\ 5 \end{bmatrix} &= \begin{bmatrix} 2\cdot 1+3\cdot 5 \\ 4\cdot 1+0\cdot 5 \end{bmatrix} = \begin{bmatrix} 17 \\ 4 \end{bmatrix}, \\ \begin{bmatrix} 2 & 3 \\ 4 & 0 \end{bmatrix}\begin{bmatrix} 2 \\ -1 \end{bmatrix} &= \begin{bmatrix} 2\cdot 2+3\cdot (-1) \\ 4\cdot 2+0\cdot (-1) \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \end{bmatrix}, \\ \begin{bmatrix} 2 & 3 \\ 4 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \end{bmatrix} &= \begin{bmatrix} 2\cdot 0+3\cdot 0 \\ 4\cdot 0+0\cdot 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \end{aligned}$$

Therefore, AB is

$$\begin{bmatrix} 2 & 3 \\ 4 & 0 \end{bmatrix}\begin{bmatrix} 1 & 2 & 0 \\ 5 & -1 & 0 \end{bmatrix} = \begin{bmatrix} 17 & 1 & 0 \\ 4 & 8 & 0 \end{bmatrix}.$$

Since A is a 2×2 matrix and B is a 2×3 matrix, the product AB is a 2×3 matrix. If we concentrate, for example, on the $(2,1)$-entry of AB, we single out the second row from A and the first column from B, and then we multiply corresponding entries together and add them up, *i.e.*, $4\cdot1+0\cdot5=4$. □

Note that the product AB of A and B is not defined if the number of columns of A and the number of rows of B are not equal.

Remark: In step (2), we could have defined for a $1\times n$ row matrix A and an $n\times r$ matrix B using the same rule defined in step (1). And then in step (3) an appropriate modification produces the same definition of the product of matrices. We suggest the readers verify this (see Example 1.6).

The **identity matrix** of order n, denoted by I_n (or I if the order is clear from the context), is a diagonal matrix whose diagonal entries are all 1, *i.e.*,

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

By a direct computation, one can easily see that $AI_n = A = I_nA$ for any $n\times n$ matrix A.

Many, but not all, of the rules of arithmetic for real or complex numbers also hold for matrices with the operations of scalar multiplication, the sum and the product of matrices. The matrix $\mathbf{0}_{m\times n}$ plays the role of the number 0, and I_n that of the number 1 in the set of real numbers.

The rule that does not hold for matrices in general is the commutativity $AB = BA$ of the product, while the commutativity of the matrix sum $A+B=B+A$ does hold in general. The following example illustrates the noncommutativity of the product of matrices.

Example 1.5 Let $A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Then,

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

Thus the matrices A and B in this example satisfy $AB \neq BA$. □

The following theorem lists some rules of ordinary arithmetic that do hold for matrix operations.

Theorem 1.4 *Let A, B, C be arbitrary matrices for which the matrix operations below are defined, and let k be an arbitrary scalar. Then*

(1) $A(BC) = (AB)C$, (*written as* ABC) (Associativity),

(2) $A(B+C) = AB + AC$, *and* $(A+B)C = AC + BC$, (Distributivity),

(3) $IA = A = AI$,

(4) $k(BC) = (kB)C = B(kC)$,

(5) $(AB)^T = B^T A^T$.

Proof: Each equality can be shown by direct calculations of each entry of both sides of the equalities. We illustrate this by proving (1) only, and leave the others to the readers.

Assume that $A = [a_{ij}]$ is an $m \times n$ matrix, $B = [b_{k\ell}]$ is an $n \times p$ matrix, and $C = [c_{st}]$ is a $p \times r$ matrix. We now compute the (i,j)-entry of each side of the equation. Note that BC is an $n \times r$ matrix whose (i,j)-entry is $[BC]_{ij} = \sum_{\lambda=1}^{p} b_{i\lambda} c_{\lambda j}$. Thus

$$[A(BC)]_{ij} = \sum_{\mu=1}^{n} a_{i\mu}[BC]_{\mu j} = \sum_{\mu=1}^{n} a_{i\mu} \sum_{\lambda=1}^{p} b_{\mu\lambda} c_{\lambda j} = \sum_{\mu=1}^{n} \sum_{\lambda=1}^{p} a_{i\mu} b_{\mu\lambda} c_{\lambda j}.$$

Similarly, AB is an $m \times p$ matrix with the (i,j)-entry $[AB]_{ij} = \sum_{\mu=1}^{n} a_{i\mu} b_{\mu j}$, and

$$[(AB)C]_{ij} = \sum_{\lambda=1}^{p} [AB]_{i\lambda} c_{\lambda j} = \sum_{\lambda=1}^{p} \sum_{\mu=1}^{n} a_{i\mu} b_{\mu\lambda} c_{\lambda j} = \sum_{\mu=1}^{n} \sum_{\lambda=1}^{p} a_{i\mu} b_{\mu\lambda} c_{\lambda j}.$$

This clearly shows that $[A(BC)]_{ij} = [(AB)C]_{ij}$ for all i, j, and consequently $A(BC) = (AB)C$ as desired. □

Problem 1.8 Prove or disprove: If A is not a zero matrix and $AB = AC$, then $B = C$.

Problem 1.9 Show that any triangular matrix A satisfying $AA^T = A^T A$ is a diagonal matrix.

Problem 1.10 For a square matrix A, show that

(1) AA^T and $A + A^T$ are symmetric,

(2) $A - A^T$ is skew-symmetric, and

(3) A can be expressed as the sum of symmetric part $B = \frac{1}{2}(A + A^T)$ and skew-symmetric part $C = \frac{1}{2}(A - A^T)$, so that $A = B + C$.

As an application of our results on matrix operations, we shall prove the following important theorem:

Theorem 1.5 *Any system of linear equations has either no solution, exactly one solution, or infinitely many solutions.*

Proof: We have already seen that a system of linear equations may be written as $A\mathbf{x} = \mathbf{b}$, which may have no solution or exactly one solution. Now assume that the system $A\mathbf{x} = \mathbf{b}$ of linear equations has more than one solution and let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two different solutions so that $A\mathbf{x}_1 = \mathbf{b}$ and $A\mathbf{x}_2 = \mathbf{b}$. Let $\mathbf{x}_0 = \mathbf{x}_1 - \mathbf{x}_2 \neq \mathbf{0}$. Since $A\mathbf{x}$ is just a particular case of a matrix product, Theorem 1.4 gives us

$$A(\mathbf{x}_1 + k\mathbf{x}_0) = A\mathbf{x}_1 + kA\mathbf{x}_0 = \mathbf{b} + k(A\mathbf{x}_1 - A\mathbf{x}_2) = \mathbf{b},$$

for any real number k. This says that $\mathbf{x}_1 + k\mathbf{x}_0$ is also a solution of $A\mathbf{x} = \mathbf{b}$ for any k. Since there are infinitely many choices for k, $A\mathbf{x} = \mathbf{b}$ has infinitely many solutions. $\square$

Problem 1.11 For which values of a does each of the following systems have no solution, exactly one solution, or infinitely many solutions?

$$(1)\ \left\{\begin{array}{rcrcrcl} x & + & 2y & - & 3z & = & 4 \\ 3x & - & y & + & 5z & = & 2 \\ 4x & + & y & + & (a^2-14)z & = & a+2. \end{array}\right.$$

$$(2)\ \left\{\begin{array}{rcrcrcl} x & - & y & + & z & = & 1 \\ x & + & 3y & + & az & = & 2 \\ 2x & + & ay & + & 3z & = & 3. \end{array}\right.$$

1.5 Block matrices

In this section we introduce some techniques that will often be very helpful in manipulating matrices. A **submatrix** of a matrix A is a matrix obtained from A by deleting certain rows and/or columns of A. Using a system of horizontal and vertical lines, we can partition a matrix A into submatrices, called **blocks**, of A as follows: Consider a matrix

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

divided up into four blocks by the dotted lines shown. Now, if we write

$$A_{11} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}, \quad A_{12} = \begin{bmatrix} a_{14} \\ a_{24} \end{bmatrix},$$

$$A_{21} = \begin{bmatrix} a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad A_{22} = \begin{bmatrix} a_{34} \end{bmatrix},$$

then A can be written as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

called a **block matrix.**

The product of matrices partitioned into blocks also follows the matrix product formula, as if the A_{ij} were numbers:

$$\begin{aligned} A &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \qquad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}; \\ AB &= \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}, \end{aligned}$$

provided that the number of columns in A_{ik} is equal to the number of rows in B_{kj}. This will be true only if the columns of A are partitioned in the same way as the rows of B.

It is not hard to see that the matrix product by blocks is correct. Suppose, for example, that we have a 3×3 matrix A and partition it as

$$A = \left[\begin{array}{cc|c} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \hline a_{31} & a_{32} & a_{33} \end{array}\right] = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

and a 3×2 matrix B which we partition as

$$B = \left[\begin{array}{cc} b_{11} & b_{12} \\ b_{21} & b_{22} \\ \hline b_{31} & b_{32} \end{array}\right] = \left[\begin{array}{c} B_{11} \\ B_{21} \end{array}\right].$$

Then the entries of $C = [c_{ij}] = AB$ are

$$c_{ij} = (a_{i1}b_{1j} + a_{i2}b_{2j}) + a_{i3}b_{3j}.$$

The quantity $a_{i1}b_{1j} + a_{i2}b_{2j}$ is simply the $(i\,j)$-entry of $A_{11}B_{11}$ if $i \le 2$, and the $(i\,j)$-entry of $A_{21}B_{11}$ if $i = 3$. Similarly, $a_{i3}b_{3j}$ is the $(i\,j)$-entry of $A_{12}B_{21}$ if $i \le 2$, and of $A_{22}B_{21}$ if $i = 3$. Thus AB can be written as

$$AB = \begin{bmatrix} C_{11} \\ C_{12} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} \\ A_{21}B_{11} + A_{22}B_{21} \end{bmatrix}.$$

In particular, if an $m \times n$ matrix A is partitioned into blocks of column vectors: *i.e.*, $A = [\mathbf{a}^1\ \mathbf{a}^2\ \cdots\ \mathbf{a}^n]$, where each block $\mathbf{a}^j$ is the j-th column, then the product $A\mathbf{x}$ with $\mathbf{x} = [x_1\ \cdots\ x_n]^T$ is the sum of the block matrices (or column vectors) with coefficients x_j's:

$$A\mathbf{x} = [\mathbf{a}^1\ \mathbf{a}^2\ \cdots\ \mathbf{a}^n] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1\mathbf{a}^1 + x_2\mathbf{a}^2 + \cdots + x_n\mathbf{a}^n,$$

where $x_j\mathbf{a}^j = x_j[a_{1j}\ a_{2j}\ \cdots\ a_{nj}]^T$.

Example 1.6 Let A be an $m \times n$ matrix partitioned into the row vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ as its blocks, and let B be an $n \times r$ matrix so that their product AB is well-defined. By considering the matrix B as a block, the product AB can be written

$$AB = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_m \end{bmatrix} B = \begin{bmatrix} \mathbf{a}_1 B \\ \mathbf{a}_2 B \\ \vdots \\ \mathbf{a}_m B \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1\mathbf{b}^1 & \mathbf{a}_1\mathbf{b}^2 & \cdots & \mathbf{a}_1\mathbf{b}^r \\ \mathbf{a}_2\mathbf{b}^1 & \mathbf{a}_2\mathbf{b}^2 & \cdots & \mathbf{a}_2\mathbf{b}^r \\ \vdots & & \ddots & \vdots \\ \mathbf{a}_m\mathbf{b}^1 & \mathbf{a}_m\mathbf{b}^2 & \cdots & \mathbf{a}_m\mathbf{b}^r \end{bmatrix},$$

where $\mathbf{b}^1, \mathbf{b}^2, \cdots, \mathbf{b}^r$ denote the columns of B. Hence, the row vectors of AB are the products of the row vectors of A and B.

Problem 1.12 Compute AB using block multiplication, where

$$A = \left[\begin{array}{rr|rr} 1 & 2 & 1 & 0 \\ -3 & 4 & 0 & 1 \\ \hline 0 & 0 & 2 & -1 \end{array}\right], \quad B = \left[\begin{array}{rr|r} 1 & 0 & 2 \\ 0 & 1 & 3 \\ \hline 2 & 3 & 4 \\ 3 & -2 & 1 \end{array}\right].$$

1.6 Inverse matrices

As we saw in Section 1.4, a system of linear equations can be written as $A\mathbf{x} = \mathbf{b}$ in matrix form. This form resembles one of the simplest linear equation in one variable $ax = b$ whose solution is simply $x = a^{-1}b$ when $a \neq 0$. Thus it is tempting to write the solution of the system as $\mathbf{x} = A^{-1}\mathbf{b}$. However, in the case of matrices we first have to have a precise meaning of A^{-1}. To discuss this we begin with the following definition.

Definition 1.7 For an $m \times n$ matrix A, an $n \times m$ matrix B is called a **left inverse** of A if $BA = I_n$, and an $n \times m$ matrix C is called a **right inverse** of A if $AC = I_m$.

Example 1.7 From a direct calculation for two matrices

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 0 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & -3 \\ -1 & 5 \\ -2 & 7 \end{bmatrix},$$

we have $AB = I_2$, and $BA = \begin{bmatrix} -5 & 2 & -4 \\ 9 & -2 & 6 \\ 12 & -4 & 9 \end{bmatrix} \neq I_3$.

Thus, the matrix B is a right inverse but not a left inverse of A, while A is a left inverse but not a right inverse of B. Since $(AB)^T = B^TA^T$ and $I^T = I$, a matrix A has a right inverse if and only if A^T has a left inverse. □

However, if A is a square matrix and has a left inverse, then we prove later (Theorem 1.8) that it has also a right inverse, and vice versa. Moreover, the following lemma shows that the left inverses and the right inverses of a square matrix are all equal. (This is not true for nonsquare matrices, of course).

Lemma 1.6 *If an $n \times n$ square matrix A has a left inverse B and a right inverse C, then B and C are equal, i.e., $B = C$.*

Proof: A direct calculation shows that

$$B = BI = B(AC) = (BA)C = IC = C.$$

Now any two left inverses must be both equal to a right inverse C, and hence to each other, and any two right inverses must be both equal to a left inverse B, and hence to each other. So there exist only one left and only one right inverse for a square matrix A ifit is known that A has both left and right inverses. Furthermore, the left and right inverses are equal. □

This theorem says that if a matrix A has both a right inverse and a left inverse, then they must be the same. However, we shall see in Chapter 3 that any $m \times n$ matrix A with $m \neq n$ cannot have both a right inverse and a left inverse: that is, a nonsquare matrix may have only a left inverse or only a right inverse. In this case, the matrix may have many left inverses or many right inverses.

Example 1.8 A nonsquare matrix $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$ can have more than one left inverse. In fact, for any x, $y \in \mathbb{R}$, one can easily check that the matrix $B = \begin{bmatrix} 1 & 0 & x \\ 0 & 1 & y \end{bmatrix}$ is a left inverse of A. □

Definition 1.8 An $n \times n$ square matrix A is said to be **invertible** (or **nonsingular**) if there exists a square matrix B of the same size such that

$$AB = I = BA.$$

Such a matrix B is called the **inverse** of A, and is denoted by A^{-1}. A matrix A is said to be **singular** if it is not invertible.

Note that Lemma 1.6 shows that if a square matrix A has both left and right inverses, then it must be unique. That is why we call B "the" inverse of A. For instance, consider a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. If $ad - bc \neq 0$, then it is easy to verify that

$$A^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} \dfrac{d}{ad-bc} & \dfrac{-b}{ad-bc} \\ \dfrac{-c}{ad-bc} & \dfrac{a}{ad-bc} \end{bmatrix},$$

since $AA^{-1} = I_2 = A^{-1}A$. (Check this product of matrices for practice!) Note that any zero matrix is singular.

Problem 1.13 Let A be an invertible matrix and k any nonzero scalar. Show that
(1) A^{-1} is invertible and $(A^{-1})^{-1} = A$;
(2) the matrix kA is invertible and $(kA)^{-1} = \frac{1}{k}A^{-1}$;
(3) A^T is invertible and $(A^T)^{-1} = (A^{-1})^T$.

Theorem 1.7 *The product of invertible matrices is also invertible, whose inverse is the product of the individual inverses in reverse order:*

$$(AB)^{-1} = B^{-1}A^{-1}.$$

Proof: Suppose that A and B are invertible matrices of the same size. Then $(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AIA^{-1} = AA^{-1} = I$, and similarly $(B^{-1}A^{-1})(AB) = I$. Thus AB has the inverse $B^{-1}A^{-1}$. □

We have written the inverse of A as "A to the power -1", so we can give the meaning of A^k for any integer k: Let A be a square matrix. Define $A^0 = I$. Then, for any positive integer k, we define the power A^k of A inductively as

$$A^k = A(A^{k-1}).$$

Moreover, if A is invertible, then the negative integer power is defined as

$$A^{-k} = (A^{-1})^k \quad \text{for } k > 0.$$

It is easy to check that with these rules we have $A^{k+\ell} = A^k A^\ell$ whenever the right hand side is defined. (If A is not invertible, $A^{3+(-1)}$ is defined but A^{-1} is not.)

Problem 1.14 Prove:
(1) If A has a zero row, so does AB.
(2) If B has a zero column, so does AB.
(3) Any matrix with a zero row or a zero column cannot be invertible.

Problem 1.15 Let A be an invertible matrix. Is it true that $(A^k)^T = (A^T)^k$ for any integer k? Justify your answer.

1.7 Elementary matrices

We now return to the system of linear equations $A\mathbf{x} = \mathbf{b}$. If A has a right inverse B such that $AB = I_m$, then $\mathbf{x} = B\mathbf{b}$ is a solution of the system since

$$A\mathbf{x} = A(B\mathbf{b}) = (AB)\mathbf{b} = \mathbf{b}.$$

In particular, if A is an invertible square matrix, then it has only one inverse A^{-1} by Lemma 1.6, and $\mathbf{x} = A^{-1}\mathbf{b}$ is the only solution of the system. In this section, we discuss how to compute A^{-1} when A is invertible.

Recall that Gaussian elimination is a process in which the augmented matrix is transformed into its row-echelon form by a finite number of elementary row operations. In the following, we will show that each elementary row operation can be expressed as a nonsingular matrix, called an *elementary matrix*, and hence the process of Gaussian elimination is simply multiplying a finite sequence of corresponding elementary matrices to the augmented matrix.

Definition 1.9 A matrix E obtained from the identity matrix I_n by executing only one elementary row operation is called an **elementary matrix**.

For example, the following matrices are three elementary matrices corresponding to each type of the three elementary row operations.

(1) $\begin{bmatrix} 1 & 0 \\ 0 & -5 \end{bmatrix}$: multiply the second row of I_2 by -5;

(2) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$: interchange the second and the fourth rows of I_4;

(3) $\begin{bmatrix} 1 & 0 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$: add 3 times the third row to the first row of I_3.

It is an interesting fact that, if E is an elementary matrix obtained by executing a certain elementary row operation on the identity matrix I_m, then for any $m \times n$ matrix A, *the product EA is exactly the matrix that is obtained when the same elementary row operation in E is executed on A.* The following example illustrates this argument. (Note that AE is not what we want. For this, see Problem 1.17).

Example 1.9 For simplicity, we work on a 3×1 column matrix $\mathbf{b}$. Suppose that we want to do the operation "adding $(-2) \times$ the first row to the second row" on matrix $\mathbf{b}$. Then, we execute this operation on the identity matrix I first to get an elementary matrix E:

$$E = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Multiplying the elementary matrix E to $\mathbf{b}$ on the left produces the desired result:

$$E\mathbf{b} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 - 2b_1 \\ b_3 \end{bmatrix},$$

Similarly, the operation "interchanging the first and third rows" on the matrix $\mathbf{b}$ can be achieved by multiplying a *permutation matrix* P, which is an elementary matrix obtained from I_3 by interchanging two rows, to $\mathbf{b}$ on the left:

$$P\mathbf{b} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} b_3 \\ b_2 \\ b_1 \end{bmatrix}.$$

□

Recall that each elementary row operation has an inverse operation, which is also an elementary operation, that brings the matrix back to the original one. Thus, suppose that E denotes an elementary matrix corresponding to an elementary row operation, and let E' be the elementary matrix corresponding to its "inverse" elementary row operation in E. Then,

(1) if E multiplies a row by $c \neq 0$, then E' multiplies the same row by $\frac{1}{c}$;

(2) if E interchanges two rows, then E' interchanges them again;

(3) if E adds a multiple of one row to another, then E' subtracts it back from the same row.

Thus, for any $m \times n$ matrix A, $E'EA = A$, and $E'E = I = EE'$. That is, *every elementary matrix is invertible so that $E^{-1} = E'$, which is also an elementary matrix.*

For instance, if

$$E_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & 1 \end{bmatrix}, \; E_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}, \; E_3 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \text{ then}$$

$$E_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/c & 0 \\ 0 & 0 & 1 \end{bmatrix}, \; E_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix}, \; E_3^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Definition 1.10 A **permutation matrix** is a square matrix obtained from the identity matrix by permuting the rows.

Problem 1.16 Prove:

(1) A permutation matrix is the product of a finite number of elementary matrices each of which is corresponding to the "row-interchanging" elementary row operation.
(2) Any permutation matrix P is invertible and $P^{-1} = P^T$.
(3) The product of any two permutation matrices is a permutation matrix.
(4) The transpose of a permutation matrix is also a permutation matrix.

Problem 1.17 Define the **elementary column operations** for a matrix by just replacing "row" by "column" in the definition of the elementary row operations. Show that if A is an $m \times n$ matrix and if E is an elementary matrix obtained by executing an elementary column operation on I_n, then AE is exactly the matrix that is obtained from A when the same column operation is executed on A.

The next theorem establishes some fundamental relationships between $n \times n$ square matrices and systems of n linear equations in n unknowns.

Theorem 1.8 *Let A be an $n \times n$ matrix. The following are equivalent:*

(1) *A has a left inverse;*

(2) *$A\mathbf{x} = \mathbf{0}$ has only the trivial solution $\mathbf{x} = \mathbf{0}$;*

(3) *A is row-equivalent to I_n ;*

(4) *A is a product of elementary matrices;*

(5) *A is invertible;*

(6) *A has a right inverse.*

Proof: **(1)** $\Rightarrow$ **(2)** : Let $\mathbf{x}$ be a solution of the homogeneous system $A\mathbf{x} = \mathbf{0}$, and let B be a left inverse of A. Then

$$\mathbf{x} = I_n\mathbf{x} = (BA)\mathbf{x} = BA\mathbf{x} = B\mathbf{0} = \mathbf{0}.$$

(2) $\Rightarrow$ (3) : Suppose that the homogeneous system $A\mathbf{x} = \mathbf{0}$ has only the trivial solution $\mathbf{x} = \mathbf{0}$:

$$\left\{\begin{array}{cccccc} x_1 & & & & = & 0 \\ & x_2 & & & = & 0 \\ & & \ddots & & & \\ & & & x_n & = & 0. \end{array}\right.$$

This means that the augmented matrix $[A\ \mathbf{0}]$ of the system $A\mathbf{x} = \mathbf{0}$ is reduced to the system $[I_n\ \mathbf{0}]$ by Gauss-Jordan elimination. Hence, A is row-equivalent to I_n.

(3) $\Rightarrow$ (4) : Assume A is row-equivalent to I_n, so that A can be reduced to I_n by a finite sequence of elementary row operations. Thus, we can find elementary matrices $E_1, E_2, \ldots, E_k$ such that

$$E_k \cdots E_2 E_1 A = I_n.$$

Since $E_1, E_2, \ldots, E_k$ are invertible, by multiplying both sides of this equation on the left successively by $E_k^{-1}, \ldots, E_2^{-1}, E_1^{-1}$, we obtain

$$A = E_1^{-1} E_2^{-1} \cdots E_k^{-1} I_n = E_1^{-1} E_2^{-1} \cdots E_k^{-1},$$

which expresses A as the product of elementary matrices.

(4) $\Rightarrow$ (5) is trivial, because any elementary matrix is invertible. In fact, $A^{-1} = E_k \cdots E_2 E_1$.

(5) $\Rightarrow$ (1) and **(5) $\Rightarrow$ (6)** are trivial.

(6) $\Rightarrow$ (5) : If B is a right inverse of A, then A is a left inverse of B and we can apply (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (4) $\Rightarrow$ (5) to B and conclude that B is invertible, with A as its unique inverse. That is, B is the inverse of A and so A is invertible. $\square$

This theorem shows that a square matrix is invertible if it has a one-side inverse. In particular, if a square matrix A is invertible, then $\mathbf{x} = A^{-1}\mathbf{b}$ is a unique solution to the system $A\mathbf{x} = \mathbf{b}$.

Problem 1.18 Find the inverse of the product

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -c & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -b & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -a & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

As an application of the preceding theorem, we give a practical method for finding the inverse A^{-1} of an invertible $n \times n$ matrix A. If A is invertible, there are elementary matrices $E_1,\ E_2,\ \ldots,\ E_k$ such that

$$E_k \cdots E_2 E_1 A = I_n.$$

Hence,

$$A^{-1} = E_k \cdots E_2 E_1 = E_k \cdots E_2 E_1 I_n.$$

It follows that *the sequence of row operations that reduces an invertible matrix A to I_n will resolve I_n to A^{-1}*. In other words, let $[A \mid I]$ be the augmented matrix with the columns of A on the left half, the columns of I on the right half. A Gaussian elimination, applied to both sides, by some elementary row operations reduces the augmented matrix $[A \mid I]$ to $[U \mid K]$, where U is a row-echelon form of A. Next, the back substitution process by another series of elementary row operations reduces $[U \mid K]$ to $[I \mid A^{-1}]$:

$$\begin{aligned} [A \mid I] & \rightarrow [E_\ell \cdots E_1 A \mid E_\ell \cdots E_1 I] = [U \mid K] \\ & \rightarrow [F_k \cdots F_1 U \mid F_k \cdots F_1 K] = [I \mid A^{-1}], \end{aligned}$$

where $E_\ell \cdots E_1$ represents a Gaussian elimination and $F_k \cdots F_1$ represents the back substitution. The following example illustrates the computation of an inverse matrix.

Example 1.10 Find the inverse of

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 5 \\ 1 & 0 & 2 \end{bmatrix}.$$

We apply Gauss-Jordan elimination to

$$[A \mid I] = \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 2 & 3 & 5 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 & 0 & 1 \end{array}\right] \begin{array}{l} \\ (-2)\text{row } 1 + \text{row } 2 \\ (-1)\text{row } 1 + \text{row } 3 \end{array}$$

$$\rightarrow \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -1 & -1 & -2 & 1 & 0 \\ 0 & -2 & -1 & -1 & 0 & 1 \end{array}\right] \ (-1)\text{row } 2$$

$$\rightarrow \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 1 & 2 & -1 & 0 \\ 0 & -2 & -1 & -1 & 0 & 1 \end{array}\right] \ (2)\text{row } 2 + \text{row } 3$$

$$\to \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 1 & 2 & -1 & 0 \\ 0 & 0 & 1 & 3 & -2 & 1 \end{array}\right].$$

This is $[U \mid K]$ obtained by Gaussian elimination. Now continue the back substitution to reduce $[U \mid K]$ to $[I \mid A^{-1}]$

$$\begin{array}{rcll} [U \mid K] & = & \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 1 & 2 & -1 & 0 \\ 0 & 0 & 1 & 3 & -2 & 1 \end{array}\right] & \begin{array}{l} (-1)\text{row } 3 + \text{row } 2 \\ (-3)\text{row } 3 + \text{row } 1 \end{array} \\ & \to & \left[\begin{array}{ccc|ccc} 1 & 2 & 0 & -8 & 6 & -3 \\ 0 & 1 & 0 & -1 & 1 & -1 \\ 0 & 0 & 1 & 3 & -2 & 1 \end{array}\right] & (-2)\text{row } 2 + \text{row } 1 \\ & \to & \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -6 & 4 & -1 \\ 0 & 1 & 0 & -1 & 1 & -1 \\ 0 & 0 & 1 & 3 & -2 & 1 \end{array}\right] & = \left[I \mid A^{-1}\right]. \end{array}$$

Thus, we get

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

(The reader should verify that $AA^{-1} = I = A^{-1}A$.) □

Note that if A is not invertible, then, at some step in Gaussian elimination, a zero row will show up on the left side in $[U \mid K]$. For example, the matrix $A = \begin{bmatrix} 1 & 6 & 4 \\ 2 & 4 & -1 \\ -1 & 2 & 5 \end{bmatrix}$ is row-equivalent to $\begin{bmatrix} 1 & 6 & 4 \\ 0 & -8 & -9 \\ 0 & 0 & 0 \end{bmatrix}$ which is a noninvertible matrix.

Problem 1.19 Write A^{-1} as a product of elementary matrices for A in Example 1.10.

of A by using Gaussian elimination.

Problem 1.20 Find the inverse of each of the following matrices:

$$A = \begin{bmatrix} 1 & -1 & 2 \\ -1 & 0 & 2 \\ -6 & 4 & 11 \end{bmatrix}, B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 2 & 4 & 0 \\ 1 & 2 & 4 & 8 \end{bmatrix}, C = \begin{bmatrix} k & 0 & 0 & 0 \\ 1 & k & 0 & 0 \\ 0 & 1 & k & 0 \\ 0 & 0 & 1 & k \end{bmatrix} \quad (k \neq 0).$$

Problem 1.21 When is a diagonal matrix $D = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{bmatrix}$ nonsingular, and what is D^{-1}?

From Theorem 1.8, a square matrix A is nonsingular if and only if $A\mathbf{x} = \mathbf{0}$ has only the trivial solution. That is, a square matrix A is singular if and only if $A\mathbf{x} = \mathbf{0}$ has a nontrivial solution, say $\mathbf{x}_0$. Now, for any column vector $\mathbf{b} = [b_1 \ \cdots \ b_n]^T$, if $\mathbf{x}_1$ is a solution of $A\mathbf{x} = \mathbf{b}$ for a singular matrix A, then so is $k\mathbf{x}_0 + \mathbf{x}_1$ for any k:

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

This argument strengthens Theorem 1.5 as follows when A is a square matrix:

Theorem 1.9 *If A is an invertible $n \times n$ matrix, then for any column vector $\mathbf{b} = [b_1 \ \cdots \ b_n]^T$, the system $A\mathbf{x} = \mathbf{b}$ has exactly one solution $\mathbf{x} = A^{-1}\mathbf{b}$. If A is not invertible, then the system has either no solution or infinitely many solutions according to whether or not the system is inconsistent.* □

Problem 1.22 Write the system of linear equations

$$\left\{ \begin{array}{rcrcrcr} x & + & 2y & + & 2z & = & 10 \\ 2x & - & 2y & + & 3z & = & 1 \\ 4x & - & 3y & + & 5z & = & 4 \end{array} \right.$$

in matrix form $A\mathbf{x} = \mathbf{b}$ and solve it by finding $A^{-1}\mathbf{b}$.

1.8 *LDU* factorization

Recall that a basic method of solving a linear system $A\mathbf{x} = \mathbf{b}$ is by Gauss-Jordan elimination. For a fixed matrix A, if we want to solve more than one system $A\mathbf{x} = \mathbf{b}$ for various values of $\mathbf{b}$, then the same Gaussian elimination on A has to be repeated over and over again. However, this repetition may be avoided by expressing Gaussian elimination as an invertible matrix which is a product of elementary matrices.

We first assume that no permutations of rows are necessary throughout the whole process of Gaussian elimination on $[A \ \mathbf{b}]$. Then the forward elimination is just to multiply finitely many elementary matrices $E_k, \ldots, E_1$ to the augmented matrix $[A \ \mathbf{b}]$: that is,

$$[E_k \cdots E_1 A \quad E_k \cdots E_1 \mathbf{b}] = [U \ \mathbf{c}],$$

where each E_i is a lower triangular elementary matrix whose diagonal entries are all 1's and $[U\ \mathbf{c}]$ is the augmented matrix of the system obtained after forward elimination on $A\mathbf{x} = \mathbf{b}$ (Note that U need not be an upper triangular matrix if A is not a square matrix). Therefore, if we set $L = (E_k \cdots E_1)^{-1} = E_1^{-1} \cdots E_k^{-1}$, then $A = LU$ and

$$\mathbf{c} = U\mathbf{x} = E_k \cdots E_1 A\mathbf{x} = E_k \cdots E_1 \mathbf{b} = L^{-1}\mathbf{b}.$$

Note that L is a lower triangular matrix whose diagonal entries are all 1's (see Problem 1.24). Now, for any column matrix $\mathbf{b}$, the system $A\mathbf{x} = LU\mathbf{x} = \mathbf{b}$ can be solved in two steps: first compute $\mathbf{c} = L^{-1}\mathbf{b}$ which is a forward elimination, and then solve $U\mathbf{x} = \mathbf{c}$ by the back substitution.

This means that, to solve the ℓ-systems $A\mathbf{x} = \mathbf{b}_i$ for $i = 1, \ldots, \ell$, we first find the matrices L and U such that $A = LU$ by performing forward elimination on A, and then compute $\mathbf{c}_i = L^{-1}\mathbf{b}_i$ for $i = 1, \ldots, \ell$. The solutions of $A\mathbf{x} = \mathbf{b}_i$ are now those of $U\mathbf{x} = \mathbf{c}_i$.

Example 1.11 Consider the system of linear equations

$$A\mathbf{x} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 1 & 0 & 1 \\ -2 & 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \\ 7 \end{bmatrix} = \mathbf{b}.$$

The elementary matrices for Gaussian elimination of A are easily found to be

$$E_1 = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \ E_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \text{ and } \ E_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & 1 \end{bmatrix},$$

so that

$$E_3E_2E_1A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 0 & -1 & -2 & 1 \\ 0 & 0 & -4 & 4 \end{bmatrix} = U.$$

Note that U is the matrix obtained from A after forward elimination, and $A = LU$ with

$$L = E_1^{-1}\, E_2^{-1}\, E_3^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -3 & 1 \end{bmatrix},$$

which is a lower triangular matrix with 1's on the diagonal. Now, the system

$$L\mathbf{c} = \mathbf{b}: \left\{ \begin{array}{rcrcrcr} c_1 & & & & & = & 1 \\ 2c_1 & + & c_2 & & & = & -2 \\ -c_1 & - & 3c_2 & + & c_3 & = & 7 \end{array} \right.$$

resolves to $\mathbf{c} = (1, -4, -4)$ and the system

$$U\mathbf{x} = \mathbf{c}: \quad \left\{ \begin{array}{rcrcrcrcr} 2x_1 & + & x_2 & + & x_3 & & & = & 1 \\ & - & x_2 & - & 2x_3 & + & x_4 & = & -4 \\ & & & - & 4x_3 & + & 4x_4 & = & -4 \end{array} \right.$$

resolves to

$$\mathbf{x} = \begin{bmatrix} -1 + t \\ 2 + 3t \\ 1 - t \\ t \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ 3 \\ -1 \\ 1 \end{bmatrix},$$

for $t \in \mathbb{R}$. It is suggested that the readers find the solutions for various values of $\mathbf{b}$. □

Problem 1.23 Determine an LU decomposition of the matrix

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

and then find solutions of $A\mathbf{x} = \mathbf{b}$ for (1) $\mathbf{b} = [1\ 1\ 1]^T$ and (2) $\mathbf{b} = [2\ 0\ -1]^T$.

Problem 1.24 Let A, B be two lower triangular matrices. Prove that
(1) their product is also a lower triangular matrix;
(2) if A is invertible, then its inverse is also a lower triangular matrix;
(3) if the diagonal entries are all 1's, then the same holds for their product and their inverses.
Note that the same holds for upper triangular matrices, and for the product of more than two matrices.

Now suppose that A is a *nonsingular square* matrix with $A = LU$ in which no row interchanges were necessary. Then the pivots on the diagonal of U are all nonzero, and the diagonal of L are all 1's. Thus, by dividing each i-th row of U by the nonzero pivot d_i, the matrix U is factorized into a diagonal matrix D whose diagonals are just the pivots $d_1, d_2, \ldots, d_n$ and a new upper triangular matrix, denoted again by U, whose diagonals are all 1's so that $A = LDU$. For example,

$$\begin{bmatrix} d_1 & r & \cdots & s \\ 0 & d_2 & & t \\ \vdots & & \ddots & u \\ 0 & & \cdots & d_n \end{bmatrix} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & & 0 \\ \vdots & & \ddots & 0 \\ 0 & & \cdots & d_n \end{bmatrix} \begin{bmatrix} 1 & r/d_1 & & s/d_1 \\ 0 & 1 & & t/d_2 \\ \vdots & & \ddots & u/d_{n-1} \\ 0 & & \cdots & 1 \end{bmatrix}.$$

This decomposition of A is called the LDU **factorization** of A. Note that, in this factorization, U is just a row-echelon form of A (with leading 1's on the diagonal) after Gaussian elimination and before back substitution.

In Example 1.11, we found a factorization of A as

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

This can be further factored as $A = LDU$ by taking

$$\begin{bmatrix} 2 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & -4 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -4 \end{bmatrix} \begin{bmatrix} 1 & 1/2 & 1/2 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} = DU.$$

Suppose now that during forward elimination row interchanges are necessary. In this case, we can first do all the row interchanges before doing any other type of elementary row operations, since the interchange of rows can be done at any time, before or after the other operations, with the same effect on the solution. Those "row-interchanging" elementary matrices altogether form a permutation matrix P so that no more row interchanges are needed during Gaussian elimination of PA. So PA has an LDU factorization.

Example 1.12 Consider a square matrix $A = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$. For Gaussian elimination, it is clearly necessary to interchange the first row with the third row, that is, we need to multiply the permutation matrix $P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ to A so that

$$PA = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} = LU.$$

□

Of course, if we choose a different permutation P', then the LDU factorization of $P'A$ may be different from that of PA, even if there is another permutation matrix P'' that changes $P'A$ to PA. However, if we fix a permutation matrix P when it is necessary, the uniqueness of the LDU factorization of A can be proved.

Theorem 1.10 *For an invertible matrix A, the LDU factorization of A is unique up to a permutation: that is, for a fixed P the expression $PA = LDU$ is unique.*

Proof: Suppose that $A = L_1D_1U_1 = L_2D_2U_2$, where the L's are lower triangular, the U's are upper triangular, all with 1's on the diagonal, and the D's are diagonal matrices with no zeros on the diagonal. We need to show $L_1 = L_2$, $D_1 = D_2$, and $U_1 = U_2$.

Note that the inverse of a lower (upper) triangular matrix is also a lower (upper) triangular matrix. And the inverse of a diagonal matrix is also diagonal. Therefore, by multiplying $(L_1D_1)^{-1} = D_1^{-1}L_1^{-1}$ on the left and U_2^{-1} on the right, our equation $L_1D_1U_1 = L_2D_2U_2$ becomes

$$U_1U_2^{-1} = D_1^{-1}L_1^{-1}L_2D_2.$$

The left side is an upper triangular matrix, while the right side is a lower triangular matrix. Hence, both sides must be diagonal. However, since the diagonal entries of the upper triangular matrix $U_1U_2^{-1}$ are all 1's, it must be the identity matrix I (see Problem 1.24). Thus $U_1U_2^{-1} = I$, *i.e.*, $U_1 = U_2$. Similarly, $L_1^{-1}L_2 = D_1D_2^{-1}$ implies that $L_1 = L_2$ and $D_1 = D_2$. □

In particular, if A is symmetric (*i.e.*, $A = A^T$), and if it can be factored into $A = LDU$ without row interchanges, then we have

$$LDU = A = A^T = (LDU)^T = U^TD^TL^T = U^TDL^T,$$

and thus, by the uniqueness of factorizations, we have $U = L^T$ and $A = LDL^T$.

Problem 1.25 Find the factors L, D, and U for $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$.

What is the solution to $A\mathbf{x} = \mathbf{b}$ for $\mathbf{b} = [1\ 0\ -1]^T$?

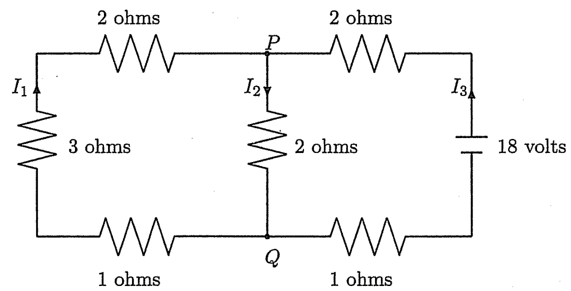
Problem 1.26 For all possible permutation matrices P, find the LDU factorization of PA for $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 2 \\ 1 & 1 & 1 \end{bmatrix}$.

1.9 Application: Linear models

(1) In an electrical network, a simple current flow may be illustrated by a diagram like the one below. Such a network involves only voltage sources, like batteries, and resistors, like bulbs, motors, or refrigerators. The voltage is measured in *volts*, the resistance in *ohms*, and the current flow in amperes (*amps*, in short). For such an electrical network, current flow is governed by the following three laws:

- **Ohm's Law:** The voltage drop V across a resistor is the product of the current I and the resistance R: $V = IR$.
- **Kirchhoff's Current Law (KCL):** The current flow into a node equals the current flow out of the node.
- **Kirchhoff's Voltage Law (KVL):** The algebraic sum of the voltage drops around a closed loop equals the total voltage sources in the loop.

Example 1.13 Determine the currents in the network given in the above figure.



Solution: By applying KCL to nodes P and Q, we get equations

$$\begin{aligned} I_1 + I_3 &= I_2 \text{ at } P, \\ I_2 &= I_1 + I_3 \text{ at } Q. \end{aligned}$$

Observe that both equations are the same, and one of them is redundant. By applying KVL to each of the loops in the network clockwise direction, we get

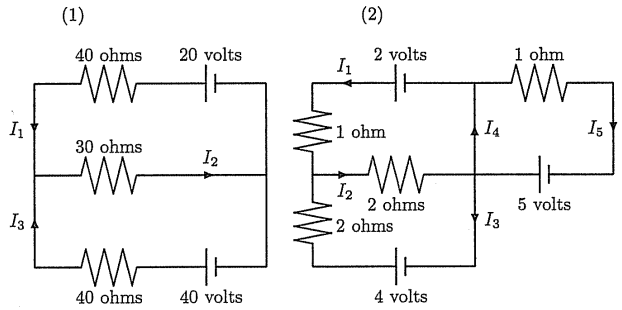
$$\begin{aligned} 6I_1 + 2I_2 &= 0 \text{ from the left loop,} \\ 2I_2 + 3I_3 &= 18 \text{ from the right loop.} \end{aligned}$$

Collecting all the equations, we get a system of linear equations:

$$\begin{cases} I_1 - I_2 + I_3 = 0 \\ 6I_1 + 2I_2 = 0 \\ 2I_2 + 3I_3 = 18. \end{cases}$$

By solving it, the currents are $I_1 = -1$ amp, $I_2 = 3$ amps and $I_3 = 4$ amps. The negative sign for I_1 means that the current I_1 flows in the direction opposite to that shown in the figure. □

Problem 1.27 Determine the currents in the following networks.



(2) Cryptography is the study of sending messages in disguised form (secret codes) so that only the intended recipients can remove the disguise and read the message; modern cryptography uses advanced mathematics. As another application of invertible matrices, we introduce a simple coding. Suppose we associate a prescribed number with every letter in the alphabet; for example,

$$\begin{array}{cccccccccccc} A & B & C & D & \cdots & X & Y & Z & \text{Blank} & ? & ! \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow \\ 0 & 1 & 2 & 3 & \cdots & 23 & 24 & 25 & 26 & 27 & 28. \end{array}$$

Suppose that we want to send the message "GOOD LUCK". Replace this message by

$$6,\ 14,\ 14,\ 3,\ 26,\ 11,\ 20,\ 2,\ 10$$

according to the preceding substitution scheme. A code of this type could be cracked without difficulty by a number of techniques of statistical methods, like the analysis of frequency of letters. To make it difficult to crack the code, we first break the message into six vectors in $\mathbb{R}^3$, each with 3 components (optional), by adding extra blanks if necessary:

$$\begin{bmatrix} 6 \\ 14 \\ 14 \end{bmatrix}, \quad \begin{bmatrix} 3 \\ 26 \\ 11 \end{bmatrix}, \quad \begin{bmatrix} 20 \\ 2 \\ 10 \end{bmatrix}.$$

Next, choose a nonsingular 3×3 matrix A, say

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

which is supposed to be known to *both* sender and receiver. Then as a linear transformation A translates our message into

$$A \begin{bmatrix} 6 \\ 14 \\ 14 \end{bmatrix} = \begin{bmatrix} 6 \\ 26 \\ 34 \end{bmatrix}, \quad A \begin{bmatrix} 3 \\ 26 \\ 11 \end{bmatrix} = \begin{bmatrix} 3 \\ 32 \\ 40 \end{bmatrix}, \quad A \begin{bmatrix} 20 \\ 2 \\ 10 \end{bmatrix} = \begin{bmatrix} 20 \\ 42 \\ 32 \end{bmatrix}.$$

By putting the components of the resulting vectors consecutively, we transmit

$$6,\ 26,\ 34,\ 3,\ 32,\ 40,\ 20,\ 42,\ 32.$$

To decode this message, the receiver may follow the following process. Suppose that we received the following reply from our correspondent:

$$19,\ 45,\ 26,\ 13,\ 36,\ 41.$$

To decode it, first break the message into two vectors in $\mathbb{R}^3$ as before:

$$\begin{bmatrix} 19 \\ 45 \\ 26 \end{bmatrix}, \quad \begin{bmatrix} 13 \\ 36 \\ 41 \end{bmatrix}.$$

We want to find two vectors $\mathbf{x}_1$, $\mathbf{x}_2$ such that $A\mathbf{x}_i$ is the i-th vector of the above two vectors: *i.e.*,

$$A\mathbf{x}_1 = \begin{bmatrix} 19 \\ 45 \\ 26 \end{bmatrix}, \quad A\mathbf{x}_2 = \begin{bmatrix} 13 \\ 36 \\ 41 \end{bmatrix}.$$

Since A is invertible, the vectors $\mathbf{x}_1$, $\mathbf{x}_2$ can be found by multiplying the inverse of A to the two vectors given in the message. By an easy computation, one can find

$$A^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix}.$$

Therefore,

$$\mathbf{x}_1 = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 19 \\ 45 \\ 26 \end{bmatrix} = \begin{bmatrix} 19 \\ 7 \\ 0 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 13 \\ 10 \\ 18 \end{bmatrix}.$$

The numbers one obtains are

$$19, \ 7, \ 0, \ 13, \ 10, \ 18.$$

Using our correspondence between letters and numbers, the message we have received is "THANKS".

Problem 1.28 Encode "TAKE UFO " using the same matrix A used in the above example.

(3) Another significant application of linear algebra is to a mathematical model in economics. In most nations, an economic society may be divided into many sectors that produce goods or services, such as the automobile industry, oil industry, steel industry, communication industry, and so on. Then a fundamental problem in economics is to find the *equilibrium* of the supply and the demand in the economy.

There are two kinds of demands for goods: the *intermediate demand* from the industries themselves (or the sectors) that are needed as inputs for their own production, and the *extra demand* from the consumer, the governmental use, surplus production, or exports. Practically, the interrelation between the sectors is very complicated, and the connection between the