SECOND EDITION

Linear Algebra and Its Applications

Gilbert Strang

LINEAR ALGEBRA AND ITS APPLICATIONS

SECOND EDITION

GILBERT STRANG
Massachusetts Institute of Technology

ACADEMIC PRESS, INC.
(Harcourt Brace Jovanovich, *Publishers*)
Orlando San Diego San Francisco New York London
Toronto Montreal Sydney Tokyo São Paulo

Academic Press, Inc
Orlando, Florida 32887

United Kingdom Edition Published by
Academic Press, Inc (London) Ltd
24/28 Oval Road, London NW1 7DX

ISBN: 0-12-673660-X
Library of Congress Catalog Card Number: 79-53993

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

I believe that the teaching of linear algebra has become too abstract. This is a sweeping judgment, and perhaps it is too sweeping to be true. But I feel certain that a text can explain the essentials of linear algebra, and develop the ability to reason mathematically, without ignoring the fact that *this subject is as useful and central and applicable as calculus*. It has a simplicity which is too valuable to be sacrificed.

Of course there are good reasons for the present state of courses in linear algebra: The subject is an excellent introduction to the precision of a mathematical argument, and to the construction of proofs. These virtues I recognize and accept (and hope to preserve); I enjoyed teaching in exactly this way. Nevertheless, once I began to experiment with alternatives at M.I.T., another virtue became equally important: Linear algebra allows and even encourages a very satisfying combination of both elements of mathematics—abstraction and application.

As it is, too many students struggle with the abstraction and never get to see the application. And too many others, especially those who are outside mathematics departments, never take the course. Even our most successful students tend to become adept at abstraction, but inept at any calculation—solving linear equations by Cramer's rule, for example, or understanding eigenvalues only as roots of the characteristic equation. There is a growing desire to make our teaching more useful than that, and more open.

I hope to treat linear algebra in a way which makes sense to a wide variety of students at all levels. This does not imply that the real mathematics is absent; the subject deserves better than that. It does imply less concentration on rigor for its own sake, and more on understanding—*we try to explain rather than to deduce*. Some definitions are formal, but others are allowed to come to the surface in the middle of a discussion. In the same way, some proofs are intended to be orderly and precise, but not all. In every case the underlying theory has to be there; it is the core of the subject, but it can be motivated and reinforced by examples.

One specific difficulty in constructing the course is always present, and is hard to postpone: How should it start? Most students come to the first class already knowing something about linear equations. Nevertheless, I am convinced that linear algebra must begin with the fundamental problem of n equations in n unknowns, and that it must teach the simplest and most useful method of solution—Gaussian elimination (not determinants!). Fortunately, even though this method is simple, there are a number of insights that are central to its understanding and new to almost every student. The most important is the equivalence between elimination and matrix factorization; the coefficient matrix is transformed into a product of triangular matrices. This provides a perfect introduction to matrix notation and matrix multiplication.

The other difficulty is to find the right speed. If matrix calculations are already familiar, then *Chapter 1 must not be too slow*; the next chapter is the one which demands hard work. Its goal is a genuine understanding, deeper than elimination can give, of the equation $Ax = b$. I believe that the introduction of four fundamental subspaces—the column space of A; the row space; and their orthogonal complements, the two nullspaces—is an effective way to generate examples of linear dependence and independence, and to illustrate the ideas of basis and dimension and rank. The orthogonality is also a natural extension to n dimensions of the familiar geometry of three-dimensional space. And of course those four subspaces are the key to $Ax = b$.

Chapters 1–5 are really the heart of a course in linear algebra. They contain a large number of applications to physics, engineering, probability and statistics, economics, and biology. (There is also the geometry of a methane molecule, and even an outline of factor analysis in psychology, which is the one application that my colleagues at M.I.T. refuse to teach!) At the same time, you will recognize that this text can certainly not explain every possible application of matrices. It is simply a first course in linear algebra. Our goal is not to develop all the applications, but to prepare for them—and that preparation can only come by understanding the theory.

This theory is well established. After the vector spaces of Chapter 2, we study projections and inner products in Chapter 3, determinants in Chapter 4, and eigenvalues in Chapter 5. I hope that engineers and others will look especially at Chapter 5, where we concentrate on the uses of diagonalization (including the spectral theorem) and save the Jordan form for an appendix. Each of these chapters is followed by an extra set of review exercises. In my own teaching I have regarded the following sections as optional: 3.4–3.5, 6.4–6.5, 7.1–7.4, and most of 1.6 and 2.6. I use the section on linear transformations in a flexible way, as a source of examples that go outside $\mathbf{R}^n$ and of a complementary approach to the theory; it illuminates in a new way what has been concretely understood. And I believe that even a brief look at Chapter 8 allows a worthwhile but relaxed introduction to linear programming and game theory—maybe my class is happier because it comes at the very end, without examination.

With this edition there is also a new Manual which I hope instructors will request from the publisher. It is a collection of ideas about the teaching of applied linear algebra, arranged section by section, and I very much hope it will grow; all suggestions and problems will be gratefully received (and promptly included). It also gives solutions to the review exercises, which now range from direct questions on the text to my favorite about the connectedness of the matrices with positive determinant.

I should like to ask one favor of the mathematician who simply wants to teach basic linear algebra. That is the true purpose of the book, and I hope he will not be put off by the "operation counts," and the other remarks about numerical computation, which arise especially in Chapter 1. From a practical viewpoint these comments are obviously important. Also from a theoretical viewpoint they have a serious purpose—to reinforce a detailed grasp of the elimination sequence, by actually counting the steps. In this edition there is also a new appendix on computer subroutines, including a full code for solving $Ax = b$. I hope that students will have a chance to experiment with it. But there is no need to discuss this or any other computer-oriented topic in class; any text ought to supplement as well as summarize the lectures.

In short, a book is needed that will permit the applications to be taught successfully, in combination with the underlying mathematics. That is the book I have tried to write.

Many readers have sent ideas and encouragement for this second edition, and I am tremendously grateful. The result is a better introduction to vector spaces, a large number of new exercises, and hundreds of changes and improvements and corrections. Nevertheless, the spirit of the book remains the same. My hope is to help construct a course with a real purpose. That is intangible but it underlies the whole book, and so does the support I have been given by my family; they are more precious than I can say. Beyond this there is an earlier debt, which I can never fully repay. It is to my parents, and I now dedicate the book to them, hoping that they will understand how much they gave to it: Thank you both.

GILBERT STRANG

Note to 5th Printing, September 1984. There is another manuscript in progress, *An Introduction to Applied Mathematics*, which will continue in the spirit of this one. Its goal is to understand the basic problems of science, engineering, and economics in the language of linear algebra (which is developed here) and of calculus. For information on *An Introduction to Applied Mathematics* you may contact me at 617-253-4383.

CONTENTS

1

GAUSSIAN ELIMINATION

INTRODUCTION ■ 1.1

The central problem of linear algebra is the solution of simultaneous linear equations. The most important case, and the simplest, is when the number of unknowns equals the number of equations. Therefore we begin with this problem: *n equations in n unknowns*.

Two ways of solving simultaneous equations are proposed, almost in a sort of competition, from high school texts on. The first is the method of ***elimination:*** Multiples of the first equation in the system are subtracted from the other equations, in such a way as to remove the first unknown from those equations. This leaves a smaller system, of $n - 1$ equations in $n - 1$ unknowns. The process is repeated over and over until there remains only one equation and one unknown, which can be solved immediately. Then it is not hard to go backward, and find all the other unknowns in reverse order; we shall work out an example in a moment. A second and more sophisticated way introduces the idea of ***determinants.*** There is an exact formula, called Cramer's rule, which gives the solution (the correct values of the unknowns) as a ratio of two n by n determinants. It is not always obvious from the examples that are worked in a textbook ($n = 3$ or $n = 4$ is about the upper limit on the patience of a reasonable human being) which way is better.

In fact, the more sophisticated formula involving determinants is a disaster, and elimination is the algorithm that is constantly used to solve large systems of simultaneous equations. Our first goal is to understand this algorithm. It is generally called *Gaussian elimination*.

The algorithm is deceptively simple, and in some form it may already be familiar to the reader. But there are four aspects that lie deeper than the simple mechanics of elimination, and which —together with the algorithm itself — we want to explain in this chapter. They are:

(1) The interpretation of the elimination method as a factorization of the coefficient matrix. We shall introduce *matrix notation* for the system of simultaneous equations, writing the n unknowns as a vector x and the n equations in the matrix shorthand $Ax = b$. Then *elimination amounts to factoring A into a product LU of a lower triangular matrix L and an upper triangular matrix U.* This is a basic and very useful observation.

Of course, we have to introduce matrices and vectors in a systematic way, as well as the rules for their multiplication. We also define the transpose A^{T} and the inverse A^{-1} of a matrix A.

(2) In most cases the elimination method works without any difficulties or modifications. In some exceptional cases it breaks down —either because the equations were originally written in the wrong order, which is easily fixed by exchanging them, or else because the equations $Ax = b$ fail to have a unique solution. In the latter case there may be no solution, or infinitely many. We want to understand how, at the time of breakdown, the elimination process identifies each of these possibilities.

(3) It is essential to have a rough count of the *number of arithmetic operations* required to solve a system by elimination. In many practical problems the decision of how many unknowns to introduce —balancing extra accuracy in a mathematical model against extra expense in computing —is governed by this operation count.

(4) We also want to see, intuitively, how sensitive to *roundoff error* the solution x might be. Some problems are sensitive; others are not. Once the source of difficulty becomes clear, it is easy to guess how to try to control it. Without control, a computer could carry out millions of operations, rounding each result to a fixed number of digits, and produce a totally useless "solution."

The final result of this chapter will be an elimination algorithm which is about as efficient as possible. It is essentially the algorithm that is in constant use in a tremendous variety of applications. And at the same time, understanding it in terms of matrices —the coefficient matrix, the matrices that carry out an elimination step or an exchange of rows, and the final triangular factors L and U —is an essential foundation for the theory.

1.2 ■ AN EXAMPLE OF GAUSSIAN ELIMINATION

The way to understand this subject is by example. We begin in three dimensions with the system

$$\begin{aligned} 2u + v + w &= 1 \\ 4u + v &= -2 \\ -2u + 2v + w &= 7. \end{aligned} \tag{1}$$

The problem is to find the unknown values of u, v, and w, and we shall apply Gaussian elimination. (Gauss is recognized as the greatest of all mathematicians, but certainly not because of this invention, which probably took him ten minutes. Ironically, however, it is the most frequently used of all the ideas that bear his name.) The method starts by *subtracting multiples of the first equation from the others, so as to eliminate u from the last two equations*. This requires that we

(a) subtract 2 times the first equation from the second;
(b) subtract -1 times the first equation from the third.

The result is an equivalent system of equations

$$\begin{aligned} 2u + v + w &= 1 \\ -1v - 2w &= -4 \\ 3v + 2w &= 8. \end{aligned} \tag{2}$$

The coefficient 2, which multiplied the first unknown u in the first equation, is known as the ***pivot*** in this first elimination step.

At the second stage of elimination, we ignore the first equation. The other two equations involve only the two unknowns v and w, and the same elimination procedure can be applied to them. *The pivot for this stage is* -1, and a multiple of this second equation will be subtracted from the remaining equations (in this case there is only the third one remaining) so as to eliminate v. We add 3 times the second equation to the third or, in other words, we

(c) subtract -3 times the second equation from the third.

The elimination process is now complete, at least in the "forward" direction, and leaves the simplified system

$$\begin{aligned} 2u + v + w &= 1 \\ -1v - 2w &= -4 \\ -4w &= -4. \end{aligned} \tag{3}$$

There is an obvious order in which to solve this system. The last equation gives $w = 1$; substituting into the second equation, we find $v = 2$; then the first equation gives $u = -1$. This simple process is called ***back-substitution***.

It is easy to understand how the elimination idea can be extended to n equations in n unknowns, no matter how large the system may be. At the first stage, we use multiples of the first equation to annihilate all coefficients below the first pivot. Next, the second column is cleared out below the second pivot; and so on. Finally, the last equation contains only the last unknown. Back-substitution yields the answer in the opposite order, beginning with the last unknown, then solving for the next to last, and eventually for the first.

EXERCISE 1.2.1 Apply elimination and back-substitution to solve

$$\begin{aligned} 2u - 3v \qquad &= 3 \\ 4u - 5v + \ w &= 7 \\ 2u - \ v - 3w &= 5. \end{aligned}$$

What are the pivots? List the three operations in which a multiple of one row is subtracted from another.

EXERCISE 1.2.2 Solve the system

$$\begin{aligned} 2u - v \qquad\qquad &= 0 \\ -u + 2v - \ w \qquad &= 0 \\ - v + 2w - z &= 0 \\ - w + 2z &= 5. \end{aligned}$$

We want to ask two questions. They may seem a little premature—after all, we have barely got the algorithm working—but their answers will shed more light on the method itself. The first question is whether this elimination procedure always leads to the solution. ***Under what circumstances could the process break down?*** The answer is: If none of the pivots are zero, there is only one solution to the problem and it is found by forward elimination and back-substitution. But if any of the pivots happens to be zero, the elimination technique has to stop, either temporarily or permanently.

If the first pivot were zero, for example, the elimination of u from the other equations would be impossible. The same is true at every intermediate stage. Notice that an intermediate pivot may become zero during the elimination process (as in Exercise 1.2.3 below) even though in the original system the coefficient in that place was not zero. Roughly speaking, *we do not know whether the pivots are nonzero until we try,* by actually going through the elimination process.

In most cases this problem of a zero pivot can be cured, and elimination can proceed to find the unique solution to the problem. In other cases, a breakdown is unavoidable since the equations have either no solution or infinitely many.

For the present, we trust all the pivots to be nonzero.

The second question is very practical, in fact it is financial. *How many separate arithmetical operations does elimination require for a system of n equations in n unknowns?* If n is large, a computer is going to take our place in carrying out the elimination (you may have such a program available, or you could use the Fortran codes in Appendix C). Since all the steps are known, we should be able to predict the number of operations a computer will take. For the moment, we ignore the right-hand sides of the equations, and count only the operations on the left. These operations are of two kinds. One is a division by the pivot in order to find out what multiple (say l) of the pivotal equation is to be subtracted from an equation below it. Then when we actually do this subtraction of one equation from another, we continually meet a "multiply-subtract" combination; the terms in the pivotal equation are multiplied by l, and then subtracted from the equation beneath it.

Suppose we agree to call each division, and each multiplication-subtraction, a single operation. At the beginning, when the first equation has length n, *it takes n operations for every zero we achieve in the first column* —one to find the multiple l, and the others to find the new entries along the row. There are $n - 1$ rows underneath the first one, so the first stage of elimination needs $n(n - 1) = n^2 - n$ operations. (Another approach to $n^2 - n$ is this: All n^2 entries need to be changed, except the n in the first row.) Now notice that later stages are faster because the equations are becoming progressively shorter; when the elimination is down to k equations, only $k^2 - k$ operations are needed to clear out the column below the pivot—by the same reasoning that applied to the first stage, when k equaled n. Altogether, the total number of operations on the left side of the equations is

$$P = (1^2 + \cdots + n^2) - (1 + \cdots + n) = \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)}{2} = \frac{n^3 - n}{3}.$$

If n is at all large, ***a good estimate for the number of operations is $P \approx n^3/3$.***

Back-substitution is considerably faster. The last unknown is found in one operation (a division by the last pivot), the second to last unknown requires two, and so on. The total for back-substitution is

$$Q = 1 + 2 + \cdots + n = \frac{n(n+1)}{2} \approx \frac{n^2}{2}.$$

A few years ago, almost every mathematician would have guessed that these numbers were essentially optimal, in other words that a general system of order n could not be solved with much fewer than $n^3/3$ multiplications. (There were even theorems to demonstrate it, but they did not allow for all possible methods.) Astonishingly, that guess has been proved wrong, and *there now exists a method that requires only $Cn^{\log_2 7}$ operations!* Fortunately for elimination, the constant C is by comparison so large, and so many more additions are required, and the computer programming is so awkward, that the new method is largely of theoretical interest. It seems to be completely unknown whether the exponent can be made any smaller.†

EXERCISE 1.2.3 Apply elimination to the system

$$\begin{aligned} u + v + w &= -2 \\ 3u + 3v - w &= 6 \\ u - v + w &= -1. \end{aligned}$$

When a zero pivot arises, exchange that equation for the one below it, and proceed. What coefficient of v in the third equation, in place of the present -1, would make it impossible to proceed—and force the elimination method to break down?

EXERCISE 1.2.4 Solve by elimination the system of two equations

$$\begin{aligned} x - y &= 0 \\ 3x + 6y &= 18. \end{aligned}$$

† Added in the second edition: The exponent is still coming down, thanks to a group at IBM. It just went below 2.5.

Draw a graph representing each equation as a straight line in the x-y plane; the lines intersect at the solution. Also, add one more line—the graph of the new second equation which arises after elimination.

EXERCISE 1.2.5 With reasonable assumptions on computer speed and cost, how large a system can be solved for \$1, and for \$1000? Use $n^3/3$ as the operation count, and you might pay \$1000 an hour for a computer that could average a million operations a second.

EXERCISE 1.2.6 (very optional) Normally the multiplication of two complex numbers

$$(a+ib)(c+id) = (ac-bd) + i(bc+ad)$$

involves the four separate multiplications ac, bd, bc, ad. Ignoring i, can you compute the quantities $ac - bd$ and $bc + ad$ with only three multiplications? (You may do additions, such as forming $a + b$ before multiplying, without any penalty.)

EXERCISE 1.2.7 Use elimination to solve

$$\begin{aligned} u + v + w &= 6 \\ u + 2v + 2w &= 11 \\ 2u + 3v - 4w &= 3. \end{aligned}$$

To get some experience in *setting up* linear equations, suppose that

(a) of those who start a year in California, 80 percent stay in and 20 percent move out;
(b) of those who start a year outside California, 90 percent stay out and 10 percent move in.

If we know the situation at the beginning, say 200 million outside and 30 million in, then it is easy to find the numbers u and v that are outside and inside at the end:

$$\begin{aligned} .9\,(200{,}000{,}000) + .2\,(30{,}000{,}000) &= u \\ .1\,(200{,}000{,}000) + .8\,(30{,}000{,}000) &= v \end{aligned}$$

The real problem is to go backwards, and compute the start from the finish.

EXERCISE 1.2.8 If $u = 200$ million and $v = 30$ million at the end, set up (without solving) the equations to find u and v at the beginning.

EXERCISE 1.2.9 If u and v at the end are the same as u and v at the beginning, what equations do you get? What is the ratio of u to v in this "steady state"?

1.3 ■ MATRIX NOTATION AND MATRIX MULTIPLICATION

So far, with our 3 by 3 example, we have been able to write out all the equations in full. We could even list in detail each of the elimination steps, subtracting a multiple of one row from another, which puts the system of equations into a simpler form. For a large system, however, this way of keeping track of the elimination would be hopeless; a much more concise record is needed. We shall now introduce matrix notation to describe the original system of equations, and matrix multiplication to describe the operations that make it simpler.

Notice that in our example

$$\begin{aligned} 2u + v + w &= 1 \\ 4u + v &= -2 \\ -2u + 2v + w &= 7 \end{aligned}$$

three different types of quantities appear. There are the unknowns u, v, w; there are the right sides 1, −2, 7; and finally, there is a set of nine numerical coefficients on the left side (one of which happens to be zero). For the column of numbers on the right side—the *inhomogeneous terms* in the equations—we introduce the vector notation

$$b = \begin{bmatrix} 1 \\ -2 \\ 7 \end{bmatrix}.$$

This is a ***three-dimensional column vector***. It is represented geometrically in Fig. 1.1, where the three components 1, −2, and 7 are the coordinates of a point in three-dimensional space. Any vector b can be identified in this way with a point in space; there is a perfect match between points and vectors.†

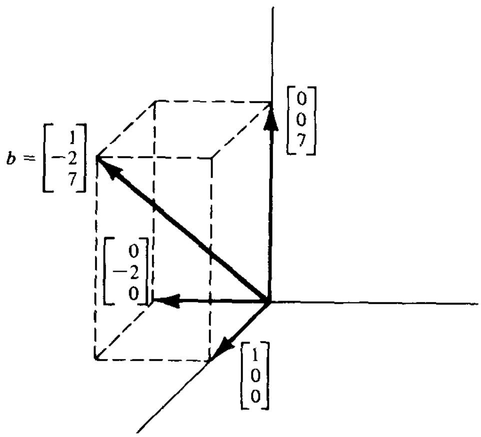

Fig. 1.1. A vector in three-dimensional space.

The basic operations are the addition of two such vectors and the multiplication of a vector by a scalar. Geometrically, $2b$ is a vector in the same direction as b but twice as

† Some authors prefer to say that the arrow is really the vector, but I think it doesn't matter; you can choose the arrow, the point, or the three numbers. (Note that the arrow starts at the origin.) In six dimensions it is probably easiest to choose the six numbers.

long, and $-2b$ goes in the opposite direction:

$$2b = \begin{bmatrix} 2 \\ -4 \\ 14 \end{bmatrix}, \qquad -2b = \begin{bmatrix} -2 \\ 4 \\ -14 \end{bmatrix}.$$

Addition of vectors is also carried out on each component separately; in Fig. 1.1 we have

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -2 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 7 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \\ 7 \end{bmatrix}.$$

This example seems special (the vectors are in the coordinate directions), so we give a more typical addition in Fig. 1.2. Once again, the addition is done a component at a time,

$$b + c = \begin{bmatrix} 1 \\ -2 \\ 7 \end{bmatrix} + \begin{bmatrix} 0 \\ 4 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 8 \end{bmatrix},$$

and geometrically this produces the famous parallelogram.

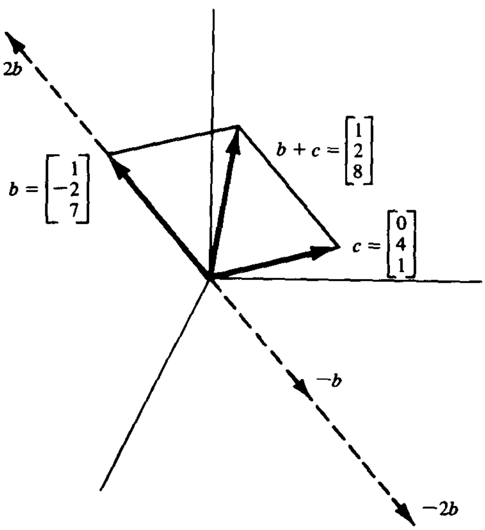

Fig. 1.2. Vector addition and scalar multiplication.

Two vectors can be added only if they have the same dimension, that is, the same