Henry C. Pinkham

Linear Algebra

July 10, 2015

Springer

Preface

This is a textbook for a two-semester course on Linear Algebra. Although the prerequisites for this book are a semester of multivariable calculus, in reality everything is developed from scratch and mathematical maturity is the real prerequisite. Traditionally linear algebra is the first course in the math curriculum where students are asked to understand proofs, and this book emphasizes this point: it gives the background to help students understand proofs and gives full proofs for all the theorems in the book.

Why write a textbook for a two semester course? First semester textbooks tend to focus exclusively on matrices and matrix manipulation, while second semester textbooks tend to dismiss matrices as inferior tools. This segregation of matrix techniques on one hand, and linear transformations of the other tends to obscure the intimate relationship between the two.

Students can enjoy the book without understanding all the proofs, as many numerically examples illustrate all the concepts.

As is the case for most elementary textbooks on linear algebra, we only study finite dimensional vector spaces and restrict the scalars to real or complex numbers. We emphasize complex numbers and hermitian matrices, since the complex case is essential in understanding the real case. However, whenever possible, rather than writing one proof for the hermitian case that also works for the real symmetric case, they are treated in separate sections, so the student who is mainly interested in the real case, and knows little about complex numbers, can read on, skipping the sections devoted to the complex case.

We spend more time that usual in studying systems of linear equations without using the matrix technology. This allows for flexibility that one loses when using matrices. We take advantage of this work to study families of linear inequalities, which is useful for the optional chapter on convexity and optimization at the end of the book.

In the second chapter, we study matrices and Gaussian elimination in the usual way, while comparing with elimination in systems of equations from the first chapter. We also spend more time than usual on matrix multiplication: the rest of the book shows how essential it is to understanding linear algebra.

Then we study vector spaces and linear maps. We give the classical definition of the rank of a matrix: the largest size of a non-singular square submatrix, as well as the standard ones. We also prove other classic results on matrices that are often omitted in recent textbooks. We give a complete change of basis presentation in Chapter 5.

In a portion of the book that can be omitted on first reading, we study duality and general bilinear forms. Then we study inner-product spaces: vector spaces with a positive definite scalar (or hermitian) product), in the usual way. We introduce the inner product late, because it is an additional piece of structure on a vector space. We replace it by duality in the early arguments where it can be used.

Next we study linear operators on inner product space, a linear operator being a linear transformation from a vector space to itself, we study important special linear operators: symmetric, hermitian, orthogonal and unitary operatrps, dealing with the real and the complex operators separately Finally we define normal operators.

Then with the goal of classifying linear operators we develop the important notion of polynomials of matrices. The elementary theory of polynomials in one variable, that most students will have already seen, is reviewed in an appendix. This leads us to the minimal polynomial of a linear operator, which allows us to establish the Jordan normal form in both the complex and real case.

Only then do we turn to determinants. This book shows how much of the elementary theory can be done without determinants, just using the rank and other similar tools. Our presentation of determinants is built on permutations, and our definition is the Leibnitz formula in terms of permutations. We then establish all the familiar theorems on determinants, but go a little further: we study the adjugate matrix and prove the classic Cauchy-Binet theorem.

Next we study the characteristic polynomial of a linear operator, and prove the Cayley-Hamilton theorem. We establish the classic meaning of all the coefficients of the characteristic polynomial, not just the determinant and the trace.

We conclude with the Spectral Theorem, the most important theorem of linear algebra. We have a few things to say about the importance of the computations of eigenvalues and eigenvectors. We derive all the classic tests for positive definite and positive semidefinite matrices.

Next there is an optional chapter on polytopes, polyhedra and convexity, a natural outgrowth of our study of inequalities in the first chapter. This only involves real linear algebra.

Finally, there is a chapter on the usefulness of linear algebra in the study of difference equations and linear ordinary differential equations. This only uses real linear algebra.

There are three appendices. the first is the summary of the notation used in the boof; the second gives some mathematical background that occasionally proves useful, especially the review of complex numbers. The last appendix on polynomials is very important if you have not seen the material in it before. Extensive use of it is made in the study of the minimal polynomial.

Leitfaden

There are several pathways through the book.

1. Many readers with have seen the material of the first three sections of Chapter 1; Chapters 2, 3, 4 and 5 form the core of the book and should be read carefully by everyone. I especially recommend a careful reading of the material on matrix multiplication in Chapter 2, since many of the arguments later on depend essentially on a good knowledge of it.
2. Chapter 6 on duality, and Chapter 7 on bilinear forms form an independent section that can be skipped in a one semester course.
3. Chapter 8 studies what we call inner-product spaces: either real vector spaces with a positive definite scalar product or complex vector spaces with a positive definite hermitian product. This begins our study of vector spaces equipped with a new bit of structure: an inner product. Chapter 9 studies operators on an inner product space. First it shows how to write all of them, and then it studies those that have a special structure with respect to the inner product. As already mentioned, the material for real vector spaces is presented independently for the reader who wants to focus on real vector spaces. These two chapter are essential.
4. In Chapter 9, we go back to the study of vector spaces without an inner product. The goal is to understand all operators, so in fact logically this could come before the material on operators on inner product spaces. After an introduction of the goals of the chapter, the theory of *polynomials of matrices* is developed. My goal is to convince the reader that there is nothing difficult here. The key result is the existence of the minimal polynomial of an operator. Then we can prove the primary decomposition and the Jordan canonical form, which allow us to decompose any linear operator into smaller building blocks that are easy to analyze.
5. Finally we approach the second main objective of linear algebra: the study of the eigenvalues and eigenvectors of a linear operator. This is done in three steps. First the determinant in Chapter 11, then the characteristic polynomial in Chapter 12, and finally the spectral theorem in Chapter 13. In the chapter concerning the spectral theorem we use the results on inner products and special operators of chapters 8 and 9 for the first time. It is essential to get to this material in a one semester course, which may require skipping items 2 and 4. Some applications show the importance of eigenvector computation.
6. Chapter 13 covers the method of least squares, one of the most important applications of linear algebra. This is optional for a one-semester course.
7. Chapter 14, another optional chapter considers first an obvious generalization of linear algebra: affine geometry. This is useful in developing the theory of iinear inequalities. From there is an a small step to get to the beautiful theory of convexity, with an emphasis on the complex bodies that come from linear inequalities: polyhedra and polytopes. This is ideal for the second semester of a linear algebra course, or for a one-semester course that only studies real linear algebra.
8. Finally the material on systems of differential equations forms a good applications for students who are familiar with multivariable calculus.
9. There are three appendices: first a catalog of the notation system used, then a brief review of some mathematics, including complex numbers, and what is most im-

portant for us, the roots of polynomials with real or complex coefficients. Finally the last appendix carefully reviews polynomials in one variable.

Recommended Books

Like generations of writers of linear algebra textbooks before me, I must disclaim any originality in the establishment of the results of this book, most of which are at least a century old. Here is a list of texts that I have found very helpful in writing this book and that I recommend.

- On the matrix side, I recommend three books:
  Gantmacher's classic two volume text [8], very thorough and perhaps somewhat hard to read;
  Franklin's concise and clear book [6].
  Denis Serre's beautiful book [24], very concise and elegant.
  Horn and Johnson's encyclopedic treatment of of matrices [13], which also shows how matrices and analysis can be interwoven.
- On the linear algebra side an excellent example of an older textbook is Minsky. More recently there is [12] - very complete.
- The classic textbook on the abstract side is Halmos's book [10]. For those who want to go even further in seeing how linear algebra is the first step in studying "abstract algebra", Michael Artin's text [1] is recommended, since he uses linear algebra as the first building block to abstract algebra.
- Linear algebra is very useful in studying advanced geometry. An excellent book that quite unusually combines the linear algebra with the geometry is Shafarevich. Even more advanced is Manin's book.
- There are two good self-described "second semester" linear algebra texts: Serge Lang's book [15] which suffers from its separation from his more elementary text that develops the matrix techniques, and then Sheldon Axler's beautifully written book [2].
- Finally there are books that focus on the computational side. It is because linear algebra algorithms can be implemented on computers is a central reason that linear algebra has come to occupy a central position in the mathematics curriculum. We do not do much of that in this book. The classic text is Golub-Van Loan [9]. There are books completely devoted to the computation of eigenvectors.

Comments, corrections, and other suggestions for improving these notes are welcome. Please email them to me at hcp3@columbia.edu.

HENRY C. PINKHAM
New York, NY
Draft of July 10, 2015

Contents

Chapter 1
Linear Equations

Abstract We define linear equations, both homogeneous and inhomogeneous, and describe what is certainly the oldest problem in linear algebra: finding the solutions of a system of linear equations. In the case of three or fewer variables we explain how elimination leads to the determinant - which we do not define in the general case. We do all this without introducing matrix notation. The sections 1.5 and 1.6 are optional. The chapter concludes with a short section about the history of the solution of linear equations

1.1 Linear Equations

The first problem of linear algebra is to solve a system of m linear equations in n unknowns $x_1, x_2, \ldots, x_n$. It was recognized early on that the case $n = m$ should be considered first. We will see why shortly.

Many readers may already familiar with linear equations. Still, since it is central to our concerns, here is the definition.

Definition 1.1.1. A system of equations is linear if it can be written

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m. \end{aligned} \tag{1.1}$$

The coefficients $a_{11}, \ldots, a_{mn}$ are numbers, as are $b_1, \ldots, b_m$ on the right hand side. The coefficients a_{ij} have a double index: the first one, i, designates the equation; the second one, j, designates the variable it is the coefficient. The coefficients b_i form the constant term of each equation and therefore have only one index designating the

equation. Thus a_{23} is the coefficient of the third unknown x_3 in the second equation, and b_3 is the constant term of the third equation. The unknowns are $x_1, x_2, \ldots, x_n$.

Definition 1.1.2. In this book, the coefficients a_{ij} of the x_j and the constants b_i are either real or complex numbers. For uniformity they are called *scalars*.

Using summations, we can write (1.1) as

$$\sum_{j=1}^{n} a_{ij}x_j = b_i, \quad 1 \le i \le m.$$

We will often use S to denote this system of linear equations. We will usually associate the running index i with the number of equations (usually m), and the running index j with the number of variables (usually n), as done above.

When we need to give the expression in the i-th equation of our system S a name, we call it f_i. So

$$f_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n - b_i. \tag{1.2}$$

Note that the constant b_i has been moved to the left-hand side of the equation, so that the right-hand side is always 0. Setting the expression f_i to 0 turns it into the i-th equation.

Definition 1.1.3. An equation is a *linear combination* of the equations in S if it can be written

$$c_1f_1 + c_2f_2 + \cdots + c_mf_m = 0$$

for some scalars $c_1, \ldots, c_m$.

Example 1.1.4. Take the 3×3 system

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

Then $3x_1 + 2x_2 = 4$ is a linear combination of the equations in this system. Indeed $c_1 = 1$, $c_2 = 1$ and $c_3 = 1$. This raises an interesting question: how do you find the c_i systematically? Set up the equations you need to solve: first the coefficients of each x_i must be the coefficients of the proposed linear combination, so you must solve

$$\begin{array}{rrrr} c_1 & +3c_2 & -c_3 = & 3 \\ -c_1 & -2c_2 & +5x_3 = & 2 \\ 2c_1 & -2c_2 & = & 0 \end{array}$$

The last step is to deal with the constant terms: we need $c_1 + 4c_2 - c_3 = 4$. This is a system of 4 linear equations in 3 unknowns, precisely the kind of system we will solve later in this chapter.

Definition 1.1.5. To solve the system S given by (1.1) means finding all the n-tuples of scalars $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$ that satisfy the system when the constants $\bar{x}_j$ are substituted

for the unknowns x_j, $1 \leq j \leq n$. We write $Z(S)$ for the set of all solutions of the system S: the letter Z stands for 'zeroes'.

Thus $Z(S)$ is the set of n-tuples $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_n)$ where for all i, $f_i(\bar{\mathbf{x}}) = 0$. This means that $Z(S)$ is the intersection

$$Z(S) = Z(f_1 = 0) \cap Z(f_2 = 0) \cap \cdots \cap Z(f_m = 0).$$

This just says that $Z(S)$ consists of the n-tuples that are simultaneously solutions to the first through the last equations.

When we have to evaluate the unknowns at more than one point we will also use upper indices to indicate constant values: for example $x_1^{(1)}$ and $x_1^{(2)}$ denote two different values for the unknown x_1.

In Example 1.1.4 it is easy to check that $\bar{x}_1 = 22/17$, $\bar{x}_2 = 1/17$, and $\bar{x}_3 = -2/17$ is a solution to the system: in fact the only solution, as we shall see shortly.

Definition 1.1.6. If all the right hand constants b_i, $1 \leq i \leq m$, are equal to 0, then the system is *homogeneous*. Otherwise it is *inhomogeneous*. If you set all the constants b_j in an inhomogeneous system 1.1.1 to zero, you get the homogeneous system *associated to* the inhomogeneous one.

So the homogeneous system associated to Example 1.1.4 is

$$\begin{aligned} x_1 \quad -x_2 \; +2x_3 &= 0 \\ 3x_1 \; -2x_2 \; -2x_3 &= 0 \\ -x_1 \; +5x_2 \qquad\; &= 0 \end{aligned}$$

The reason an inhomogeneous system and its associated homogeneous system are always considered together is:

Theorem 1.1.7. *If* $(x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)})$ *and* $(x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)})$ *are solutions of the inhomogeneous system 1.1, then their difference*

$$(x_1^{(1)} - x_1^{(2)}, x_2^{(1)} - x_2^{(2)}, \ldots, x_n^{(1)} - x_n^{(2)})$$

is a solution of the associated homogeneous system.

Proof. Just subtract the corresponding equations. The details are left to you as an exercise. □

Here is another example of three equations in three variables.

Example 1.1.8.

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

The corresponding homogeneous system of equations is

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

We will find the solutions later.

We are not interested in the system of equations S for itself, but only in its solutions $Z(S)$. The three fundamental questions are:

1. Does the system S have any solutions at all, or is the set of solutions empty? If it is empty we write $Z(S) = \emptyset$.
2. If there is a solution, is it unique? This means $Z(S)$ is a single element.
3. If there is more than one solution, what does the set of all solutions look like?

We can already answer the first question when the system is homogenous, since

$$x_1 = x_2 = \cdots = x_n = 0 \tag{1.3}$$

is clearly a solution., as we established in Theorem 1.1.7. So a homogeneous system always has the solution (1.3), called the trivial solution.

On the other hand, it is easy to produce an inhomogeneous system without solutions. The simplest example is perhaps

$$\begin{aligned} x_1 &= 0 \\ x_1 &= 1. \end{aligned} \tag{1.4}$$

Here $n = 1$ and $m = 2$. Similarly

$$\begin{aligned} 2x_1 - x_2 &= 1 \\ 2x_1 - x_2 &= 2. \end{aligned} \tag{1.5}$$

The solutions of the first equation of (1.5) are just the points of a line in the plane; the solutions of the second equation are the points of a distinct parallel line. Since the lines do not intersect, there is no solution to the system.

This can also be proved algebraically. If scalar values $(\bar{x}_1, \bar{x}_2)$ satisfy both equations, they they satisfy the difference of the two equations. But that says that $0 = -1$, a contradiction.

Definition 1.1.9. A system of linear equations S that does not have any solutions is *inconsistent*. Thus $Z(S) = \emptyset$. A system with at least one solution is *consistent*.

Corollary 1.1.10. *A consistent inhomogeneous system has exactly one solution if and only if the corresponding homogeneous solution has only one solution, which must be the trivial solution.*

Proof. This is an immediate consequence of Theorem 1.1.7 and the existence of the trivial solution for homogeneous systems of equations. □

Notice that the homogenous system associated to (1.5) is simply $2x_1 - x_2 = 0$, so that it has an infinite number of solutions of the form $(\bar{x}_1, 2\bar{x}_1)$, even though the inhomogeneous system of equation has no solutions at all.

We now make an important definition concerning systems of linear equations.

Definition 1.1.11. Two systems of linear equations in the same variables are *equivalent*, if their set of solutions is the same.

Given a system of linear equations, our goal is to transform it into an equivalent system that is easier to solve. Here is a useful tool.

Proposition 1.1.12. *If the equations in a linear system S_1 are all linear combinations of the equations in a linear system S, the $Z(S) \subset Z(S_1)$. Therefore if all the equations in S are also linear combinations of those in S_1, the two systems are equivalent: $Z(S_1) = Z(S)$.*

The proof is left to you.

1.2 Geometry Interpretation

When the scalars are the real numbers $\mathbb{R}$, and the number of variables is at most three, then it is important to interpret the linear equations geometrically. Here are some simple but useful remarks. Work out the easy case of only one variable on your own.

Example 1.2.1. Suppose you have two variables x_1 and x_2, so we are working in the plane.

First assume the system just consists of one linear equation $a_1x_1 + a_2x_2 = b$. If both a_1 and a_2 are 0, then we are left with the equation $0 = b$, which is inconsistent unless $b = 0$. So the set of solutions $X(S)$ is either empty or the entire plane $\mathbb{R}^2$. When at least one of a_1 and a_2 is not zero, the set of solutions is a line $a_1x_1 + a_2x_2 = b$, as you know. Thus we have an infinite number of solutions. If $b = 0$ the line goes through the origin.

What if there are two equations? The only interesting case occurs when each equation has at least one non-zero coefficient: then the solutions to each equation form a line. So the solutions to the system is just the intersection of two lines. What could happen? There are three cases: either the lines have different slopes so they intersect in a point. If they have the same slope, then they are either parallel and distinct, in which case there are no solutions, or they are the same line, in which case there is a line of solutions.

If there are three equations or more, then "usually" the set of solutions is empty, since usually the intersection of three lines in the plane is empty

Example 1.2.2. Finally the case $n = 3$. We are working in space. If there is only one equation $a_1x_1 + a_2x_2 + a_3x_3 = b$, then unless all the coefficients a_i are zero, the solutions form a plane in space.

If there are two equations, the set of solutions is the intersection of two planes in space. What can happen? In general they intersect in a line, which goes through the origin if the equations are homogeneous. But the planes could be parallel: what does that mean in terms of the coefficients? In that case there are no solutions unless the planes are the same plane. Check that two planes in $\mathbb{R}^3$ cannot intersect in a point.

Finally assume there are three equations. The set of solutions to each equation is a plane, so we are studying the intersection of three planes in space. If all goes well, we get the intersection of a plane and a line. You should convince yourself that this is usually a point, but in "degenerate" cases you could get a line or even a plane. Or as usual the intersection could be empty. But you cannot get more complicated sets, such an two lines, or two points.

If there are more than three equations, then "usually" the intersection is empty.

Our goal is to give a precise meaning to "usually" and "degenerate", and to analyze the set of solutions in higher dimensions in a systematic way, without using geometric intuition.

Exercise 1.2.3. Write down explicit numerical examples to illustrate all possible cases. For example two lines in the plane that intersect in a point: just take $x_1 = 0$ and $x_2 = 0$ for the two equations.

1.3 Elimination

The key to understanding the solutions of a system of linear equations S, which we always write as (1.1), is a process called Gaussian elimination. It is an algorithm that does two things:

- if the system S is not homogeneous, it determines whether it is consistent or not. This step is unnecessary for a homogeneous system which we already know is consistent.
- if the system is consistent, it determines all the solutions.

This is the most important algorithm of linear algebra. We will study it more algebraically when we have developed matrix technology in Chapter 2. Here we use some geometric ideas. For simplicity we only consider the case of real scalars, but the result goes through for complex scalars without change.

The key notion is that of the *projection* map from $\mathbb{R}^n$ to $\mathbb{R}^{n-1}$ obtained by omitting one of the coordinates. Any one of the coordinates could be omitted, but without loss of generality we may assume it is the last one. When $n = 3$ the projection maps (x_1, x_2, x_3) to (x_1, x_2). The image of the map is clearly all $\mathbb{R}^2$, and the inverse image[1] of any given point (a_1, a_2) is (a_1, a_2, x) for any $x \in \mathbb{R}$.

More generally the projection maps $(x_1, \ldots, x_{n-1}, x_n)$ to $(x_1, \ldots, x_{n-1})$, and the inverse image of any point is again $\mathbb{R}$. We will need to consider the case $n = 1$, the

[1] See §B.1 if you need to review this notion.

projection $\mathbb{R}^1$ to $\mathbb{R}^0$, so we need to define $\mathbb{R}^0$: it is a single element, written 0. If you compose a projection $p_n\colon \mathbb{R}^n \to \mathbb{R}^{n-1}$ with a projection $p_{n-1}\colon \mathbb{R}^{n-1} \to \mathbb{R}^{n-2}$ to get $p_{n-1} \circ p_n$ you get a new projection from $\mathbb{R}^n$ to $\mathbb{R}^{n-2}$, which omits two of the coordinates. More generally we can omit any number r of the coordinates and get a projection, which we can view as the composition of r projections that just omit one coordinate each. This is what we will do.

Assume we have a system S of m equations in n variables.

First we handle the trivial case: all the $a_{ij} = 0$. Therefore the left–hand side of all the equations is 0, so they take the form $0 = b_j$. If any one of the b_j is nonzero the system is obviously inconsistent, and if all the b_j are equal to 0, then the system imposes no conditions, so the set of solutions is $\mathbb{R}^n$. Our goal is to repeat an elimination step, which is described next, until we reach the trivial case.

So assume that there is a variable that occurs with non-zero coefficient in at least one of the equations. By renumbering the variables, we may assume the variable is x_n. This is the variable to be eliminated. We call it the *pivot variable* for this elimination step. By renumbering the equations we may assume that in f_m the coefficient a_{mn} of x_n is nonzero. Divide f_m by a_{mn}. Of course they may be several equations in which the pivot variable has a non-zero coefficient: just pick one. Clearly the new system is equivalent to the old one: for instance use Proposition 1.1.12. We continue to call the new system S, in other words we assume the coefficient of the pivot variable in f_m is 1.

Now replace S by the new system S_1 given by

$$\begin{aligned}
g_1(x_1,\dots,x_{n-1}) &= f_1(x_1,\dots,x_n) - c_1 f_m(x_1,\dots,x_n)\\
g_2(x_1,\dots,x_{n-1}) &= f_2(x_1,\dots,x_n) - c_2 f_m(x_1,\dots,x_n)\\
&\vdots\\
g_{m-1}(x_1,\dots,x_{n-1}) &= f_{m-1}(x_1,\dots,x_n) - c_{m-1} f_m(x_1,\dots,x_n)\\
f_m((x_1,\dots,x_{n-1},x_n) &= 0
\end{aligned}$$

where c_i is the coefficient of the pivot variable in f_i. By Proposition 1.1.12 again, the two systems are equivalent: they have the same solutions. In particular $Z(S) = \emptyset$ if and only if $Z(S_1) = \emptyset$. By construction, in all g_i the variable x_n appears with coefficient 0: we say these equations do not contain x_n. The collection of all the g_i is called the set of *residual*[2] equations. On the other hand the equation $f_m = 0$ contains x_n, the pivot variable. The process of replacing S by the equivalent S_1 process is called *eliminating* x_n, because x_n has been eliminated from the residual equations. Underlying this is the projection map $p_n\colon \mathbb{R}^n \to \mathbb{R}^{n-1}$ omitting the last coordinate. The linear system S lies in $\mathbb{R}^n$, while the residual equations of S_1 lie in $\mathbb{R}^{n-1}$.

Example 1.3.1. Let's apply this technique to the system S:

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

[2] This is not standard terminology.

Then $f_3 = x_1 - x_2 + x_3 - 1$, so by subtracting 3 times f_3 from the first listed equation, and adding it to the second, we get the system S_1

$$\begin{aligned} -x_1 \quad +x_2 \qquad &= 1 \\ x_1 \ +3x_2 \qquad &= 0 \\ x_1 \quad -x_2 \ +x_3 &= 1 \end{aligned}$$

It is easy to see by hand, or by repeating the process that there is only one solution to S_1: $x_1 = -3/4$, $x_2 = 1/4$. Putting these values into f_3 gives $x_3 = -x_1 + x_2 + 1 = 3/4 + 1/4 + 1 = 2$, and therefore S has the unique solution $(-3/4, 1/4, 2)$.

Exercise 1.3.2. Apply this technique to Example 1.1.8. Here again $m = n$.

The process of elimination consists in repeating the elimination step describe above as long as it is possible. The elimination step is only applied to the last collection of residual equations, which does not include any of the variables that have already been eliminated. So after each elimination step the number of variables in the residual equations decreases by one, since the pivot variables are all distinct. After k eliminations steps we are dealing with a projection $p\colon \mathbb{R}^n \to \mathbb{R}^{n-k}$. The process terminates when there are no more candidates for a pivot variable left in the residual equations. Note that when there is only one residual equation left with a variable with a non-zero coefficient, that variable a pivot variable so we can do one more elimination, after which the set of residual equations is empty.

We will see in Chapter 4, the number of eliminations steps does not depend on the order in which the pivot variables are chosen. Therefore we get an important invariant of the system of equations, called the *rank* of the system. We will not use that fact here. Start with a system S of m linear equations in n unknowns. Eliminate one variable at a time: each time the set of solutions gets projected in a one-to-one manner to the next lower dimension. After each elimination step, the number of variables decreases exactly by one, while the number of equations decreases by at least one. The elimination process terminates when

1. either the set of residual equations is non-empty, and they are all of the form $0 = b$, because there are no more candidates for a pivot variable. As we have already noted, in this case the system of equations could be inconsistent.
2. or the set of residual equations is empty. Then the system is consistent.

In both cases the variables that have not been eliminated are called the *free variables*. When the system is consistent, the set of solutions is in bijection with $\mathbb{R}^r$.

Thus we get two of the main theorems of linear equations.

Theorem 1.3.3. *A system of n linear equations in m variables, $m < n$, is either inconsistent or has an infinite number of solutions.*

Proof. The first statement follows from repeatedly applying Proposition 1.1.12. Next assume the system is consistent. In each projection there is one pivot variable, and they are all distinct. So after at most m elimination steps, the process terminates, because we run out of equations. Assume the actual number of steps is $l \leq m$. Then

we have at most $m-l$ residual equations that do not contain any of the l pivot variables. Then by the hypothesis $m < n$, we have $n-l > 0$ free variables Therefore the set of solutions is in bijection with $\mathbb{R}^{n-p}$. □

So this restricts the possibilities for the set of solutions of a system of linear equations: if it is finite, then it is just one point.

We record a special case of the theorem:

Corollary 1.3.4. *If there are k free variables left when elimination terminates, the set of solutions is either empty or in bijection with* $\mathbb{R}^k$.

It is also worth recording the

Corollary 1.3.5. *The system S is inconsistent if and only if there is a linear combination of its equations that reduces to* $0 = b \neq 0$.

At the end of the elimination process, we get a convenient representation of a system of equations equivalent to any system *S*:

Theorem 1.3.6. *Repeat the elimination process on the system of linear equations S until it terminates. The system S is equivalent to a system with r equations*

$$f_i(x_1, x_2, \ldots, x_n) = x_i + c_{i,i+1}x_{i+1} \cdots + c_{in}x_n - b_i = 0 \quad , \quad 1 \leq i \leq r. \tag{1.6}$$

for suitable scalars $c_{i,j}$, $j \leq i$ *and* b_i. *The* x_i, $1 \leq i \leq r$ *are the pivot variables, given in the order they are eliminated and r is the total number of elimination steps needed.*

This results is simply a translation of what we have already done, with a difference numbering of the pivots.

Example 1.3.7. Assume we are in $\mathbb{R}^3$, and that the system *S* has three equations, and that we can eliminate x_3. Then we get a new system S_1 with two residual equation in x_1 and x_2 only, and a single equation that can be written, after dividing by the coefficient of $_3$ as $x_3 = a_1x_1 + a_2x_2 - b$ for suitable real constants a_1, a_2' and b.

Consider the zeroes of the two residual equations in the plane with coordinates x_1 and x_2. We have the intersection of two lines, therefore a point, in general. However the intersection could be a line or empty, as we have already seen. Assume the intersecion is a point. Now consider the zeroes of the residual equation in $\mathbb{R}^3$: for each solution $\mathbf{p}$ of S_1 in the x_1x_2 plane we get the whole vertical line above p in $\mathbb{R}^3$, where the third coordinate is any real number. Then the remaining equation $x_3 = a_1x_1 + a_2x_2 - b$ picks out a unique point on this line which is a solution of the original equation.

Thus geometrically we have done the following. To find the zeroes of *S* in $\mathbb{R}^3$ we project them to $\mathbb{R}^2$. then the remaining equation, from which x_3 has not been eliminated picks out a unique point in above the locus of projection.

1.4 Examples of Elimination

Example 1.4.1. If S is the system

$$\begin{aligned} 2x_1 &= 1 \quad \text{so } f_1 = 2x_1 - 1, \\ x_2 &= 2 \quad \text{so } f_2 = x_2 - 2, \\ x_3 &= -1 \quad \text{so } f_3 = x_3 + 1, \end{aligned}$$

then it is clear consistent and the three equations are already the equations $f_1^{(0)} = 0$, $f_2^{(1)} = 0$, and $f_3^{(2)} = 0$, so $r = 3$. Of course this example is so simple we did not need elimination.

The slightly more complicated example

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

is solved the same way: no actual elimination occurs.

Example 1.4.2. Now we apply this process to Example 1.1.8, another inhomogeneous case.

$$\begin{aligned} 2x_1 - x_2 + 3x_3 &= 1 \\ 4x_1 - 2x_2 - x_3 &= -5 \\ 10x_1 - 5x_2 - 6x_3 &= -16 \end{aligned}$$

We eliminate x_1 using the first equation. Our system gets replaced by the system $S^{(1)}$, omitting the first equation of zeroes:

$$\begin{aligned} -7x_3 &= -7 \\ -21x_3 &= -21 \end{aligned}$$

We eliminate x_3 to get $S^{(2)}$, where all the coefficients are zero. So the original system is consistent, $r = 2$, and our new system of equivalent equations is:

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

Since there is only one unknown in the last equation, we can solve for it, getting $x_3 = 1$. Substituting this value in the first equation, we get

$$2x_1 - x_2 = -2 \tag{1.7}$$

which we recognize as the equation of a line in the plane. Thus there are an infinite number of solutions: for each value of x_2 we can find a solution

$$x_1 = \frac{x_2 - 2}{2}.$$

A particular solution is given by $(-1,0,1)$. Now let us consider the homogeneous equation corresponding to 1.7. :

$$\begin{aligned} 2x_1 - x_2 + 3x_3 &= 0 \\ -7x_3 &= 0 \end{aligned}$$

Thus $x_3 = 0$ and $x_2 = 2x_1$ is the most general solution of the homogeneous equation. By Theorem 1.1.7, any solution to the inhomogeneous equation can be written

$$(-1,0,1) + (x_1, 2x_1, 0).$$

You should check that this is what we found above.

Exercise 1.4.3. Show that the following system has a unique solution:

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

Since numerical computation is error-prone you should substitute the values you find into the original equations to confirm that you have not made a mistake.

Finally, you should check that the only solution to the corresponding homogeneous equation is $(0,0,0)$, so that Theorem 1.1.7 is again verified.

Example 1.4.4. Next we treat the general case of two equations in two variables. We use the usual notation

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2 \end{aligned}$$

for the system S.

1. If all four coefficients a_{ij} are zero, we are already in **Case 1**.

Otherwise we may assume that $a_{11} \neq 0$ by interchanging the variables x_1 and x_2, or the order of the equations if necessary. Then $f_1^{(0)}$ is just the first equation $a_{11}x_1 + a_{12}x_2 - b_1 = 0$, and the new system $S^{(1)}$ is

$$\begin{aligned} 0x_1 + 0x_2 &= 0 \\ (a_{22} - \frac{a_{21}}{a_{11}}a_{12})x_2 &= b_2 - \tfrac{a_{21}}{a_{11}}b_1 \end{aligned} \tag{1.8}$$

Examine carefully how the coefficients of the second equation are formed. The expressions become cleaner if we multiply it by the non-zero quantity a_{11}. We get for the second equation:

$$f_2^{(1)} = (a_{11}a_{22} - a_{12}a_{21})x_2 - a_{11}b_2 - b_1a_{21} = a_{22}^{(1)}x_2 - b_2^{(1)} = 0, \qquad (1.9)$$

using our usual notation.

2. If the coefficient of x_2 is zero, then we are in **Case 1**. In the original variables this coefficient is

$$a_{11}a_{22} - a_{12}a_{21} = 0. \qquad (1.10)$$

If this happens, then if the right hand side is also 0, then the system is consistent. If the right hand side is not 0, then we have a contradiction and the system is inconsistent, as in Example 1.5. In both cases the analysis is over.

3. The last case to consider is the case where the coefficient of x_2 in (1.9) is nonzero. Then we are still in **Case 2**, so we do elimination again to get $S^{(2)}$ which is just the trivial matrix.

Our general theorem says that the original system is equivalent to that formed by the two equations $f_1^{(0)} = 0$ and $f_2^{(1)} = 0$. From the second equation we get the unique solution

$$x_2 = \frac{a_{11}b_2 - b_1a_{21}}{a_{11}a_{22} - a_{12}a_{21}}$$

so substituting this value into the first equation, we get

$$a_{11}x_1 + a_{12}\frac{a_{11}b_2 - b_1a_{21}}{a_{11}a_{22} - a_{12}a_{21}} = b_1 = b_1\frac{a_{11}a_{22} - a_{12}a_{21}}{a_{11}a_{22} - a_{12}a_{21}}$$

or

$$a_{11}x_1 = \frac{-a_{12}a_{11}b_2 - a_{12}b_1a_{21} + b_1a_{11}a_{22} - b_1a_{12}a_{21}}{a_{11}a_{22} - a_{12}a_{21}} = \frac{-a_{12}a_{11}b_2 + b_1a_{11}a_{22}}{a_{11}a_{22} - a_{12}a_{21}}$$

which simplifies, since $a_{11} \neq 0$, to the unique solution

$$x_1 = \frac{b_1a_{22} - a_{12}b_2}{a_{11}a_{22} - a_{12}a_{21}}.$$

So the system has a unique solution. Notice that the expression of x_1 can be obtained from that for x_2 simply by interchanging the indices 1 and 2 everywhere.

Notice the key role played by the coefficient of x_2 in (1.9). This is called the determinant of the system, and is written

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \qquad (1.11)$$

Similarly the numerators in the expressions for x_1 and x_2 are written

$$\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix} \text{ and } \begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}$$

respectively. Notice how they arise from the determinant by replacing the appropriate 'column' of the determinant by the right hand side of the equation. You have probably seen these expressions for a previous mathematics class.

When the coefficients are the real numbers, our analysis has a geometric interpretation in the plane with coordinates x_1 and x_2 . We only deal with the case the equations are consistent. If all the coefficients are all zero, then the set of solutions is the entire plane. Otherwise, if the coefficients of one equation are all zero, we are just left with the other equation: the set of solutions is then a line. Otherwise, the locus where each one of the equations is satisfied in a line: call them L_1 and L_2. The locus where both are satisfied in the intersection of the two lines: what can that be?

1. The two lines could be the same: $L_1 = L_2$. Then the intersection is just this line, so the system has an infinite number of solutions.
2. The two lines could be distinct and parallel: then the intersection is empty. So the system is inconsistent.
3. If the two lines are not parallel, they meet in a point, giving a unique solution.

Exercise 1.4.5. How does the determinant change when two equations are interchanged?

Example 1.4.6. Finally we treat the general case of three equations in three variables. We first eliminate x_2 as in the two variable case, getting a system with only two variables, in addition to the equation in which x_1 can be solved in terms of x_2 and x_3. We write the two new equations as

$$\begin{aligned} a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 &= b_2^{(1)} \\ a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 &= b_3^{(1)}. \end{aligned}$$

where the coefficients $a_{22}^{(1)}$, $a_{23}^{(1)}$, $b_2^{(1)}$, $a_{32}^{(1)}$, $a_{33}^{(1)}$, $b_3^{(1)}$ can be written in terms of the original coefficients.

From the 2×2 case, we already know we can continue eliminating variables if and only if one of the four quantities $a_{ij}^{(1)}$ is different from 0. Otherwise the left-hand sides of the two remaining equations are 0. Then if one of the right-hand sides is different from 0, there is no solution; if both right-hand sides are 0, we are in fact dealing with only one equation. the solutions of which form a plane in three-space.

If one of the four quantities $a_{ij}^{(1)}$ is different from 0, by changing the numbering of the variables and the order of the equations, we may assume without loss of generality that $a_{22}^{(1)} \neq 0$. The final step is to eliminate the variable x_2 from the last equation, by adding to it a suitable multiple of the second equation. The last equation becomes

$$\left(a_{33}^{(1)} - a_{32}^{(1)}\frac{a_{23}^{(1)}}{a_{22}^{(1)}}\right)x_3 = b_3^{(1)} - a_{32}^{(1)}\frac{b_2^{(1)}}{a_{22}^{(1)}}. \tag{1.12}$$

To finish the analysis, we need to determine when the coefficient of x_3 is non-zero. Since we already know that $a_{22}^{(1)} \neq 0$, we can multiply (1.12) by that quantity, getting

$$\left(a_{22}^{(1)}a_{33}^{(1)} - a_{32}^{(1)}a_{23}^{(1)}\right)x_3 = a_{22}^{(1)}b_3^{(1)} - a_{32}^{(1)}b_2^{(1)}. \tag{1.13}$$

Not surprisingly, this is precisely parallel to (1.9). Now we substitute in the original coefficients and multiply by a suitable power of $a_{11} \neq 0$. Then we get for the coefficient of x_3, up to a non-zero constant:

$$\begin{aligned}\left(a_{11}a_{22} - a_{21}a_{12}\right)\left(a_{33} - \tfrac{a_{31}}{a_{11}}a_{13}\right) - \left(a_{11}a_{23} - a_{21}a_{13}\right)\left(a_{32} - \tfrac{a_{31}}{a_{11}}a_{12}\right)& \\ = a_{11}a_{22}a_{33} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33}& \\ -a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}.&\end{aligned} \tag{1.14}$$

It is called the *determinant* of the system: it was so named by Gauss in 1806. Gauss formalized this elimination process, which is now called Gaussian elimination. In the classic literature the determinant is always written

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \tag{1.15}$$

If it is non-zero, then the last equation gives a unique solution for x_3, the second equation a unique solution for x_2 and the first equation a unique expression for x_1. If the system is homogeneous, the only solution is $(0,0,0)$.

Exercise 1.4.7. In Example 1.4.3, determine the coefficients $a_{ij}^{(1)}$ obtained after one elimination, and then the coefficients obtained after the second elimination. In particular you should compute the determinant, which is 3.

Exercise 1.4.8. From the way we computed the unique solution in Example 1.4.6, it may seem that several conditions need to be satisfied. First we need to to find a a_{ij} that is non-zero. Then we need to find a $a_{kl}^{(1)}$ that is non-zero, where $k \neq i$ and $l \neq j$. Finally we need that the determinant (1.14) be non-zero. Show that if either of the first two conditions fails, then the determinant is zero. Thus the vanishing of the determinant by itself tells us if there is a unique solution to the system. So far we only have this for a 3×3 system, but, as we will see later, the result is true generally.

Exercise 1.4.9. Work out the right hand side of equation (1.13) exactly as we did for the coefficient of x_3 in (1.14).

Remark 1.4.10. What is the difference between doing the computation of the solution as in Example 1.4.3 and in Example 1.4.6? The first difference is that the method in Example 1.4.6 applies to any set of coefficients, and tells us exactly when the method fails. It does this by treating the coefficients as variables. This is a major advantage.

On the other hand, treating the coefficients as variables makes the computation more difficult. In fact, when there are a large number of variables, which often happens in applications, the general method is not feasible. The difference, in modern

computer terminology is that in Example 1.4.3 we are doing a numerical computation, while in Example 1.4.6 we are doing a computer algebra computation, which explodes as the number of variables becomes large.

Remark 1.4.11. The determinant is a function of the coefficients. If one randomly chooses values for the coefficients, the probability that the determinant vanishes is zero. Why?

Thus a non-zero determinant is the expected case. In fact, in the theory of equations, it was well known, from the earliest times that every time one imposes an equation on n variables, the degrees of freedom of the variables should go down by one. Thus the key case is the case of n equations, where there should be 0 degrees of freedom, meaning a finite number of solutions. As we already see that for the examples above, this is not always true, even in the case of linear equations.

The classic expression for the solution of a system of n equations in n variables was given by Cramer in 1750. We will return to it when we study determinants in Chapter 11.

1.5 Consequences of Linear Systems

We now record the following interesting corollary of elimination. This section is not used later in the book.

A *consequence* of the linear system S to be any linear equation :

$$g(\mathbf{x}) = d_1x_1 + \cdots + d_nx_n - e = 0 \tag{1.16}$$

that vanishes on the zeroes $Z(S)$. This definition is uninteresting when $Z(S) = \emptyset$, since any linear equation is then a consequence of S. So we will only use the definition when S is consistent.

Corollary 1.5.1. *Any consequence of a consistent system S can be written as a linear combination of the equations in S.*

Proof. It is enough to show that the consequence (1.16) is a linear combination of the f_i given in (1.6), since these are linear combinations of the equations in S. Here is how we do this. We first replace g by

$$g^{(1)} = g - d_1 f_1^{(0)}.$$

Because both g and $f_1^{(0)}$ vanish on $Z(S)$, so does their difference $g^{(1)}$. By construction $g^{(1)}$ does not contain the first pivot. Continuing in this way we can successively eliminate all the pivot variables until we get a $g^{(r)}$ that contains none of the pivot variables. It too is a consequent of S, so it vanishes of $Z(S)$. If $g^{(r)}$ contains any variables (which must be free variables) then it fails to vanish somewhere in the projected set of solutions. This is impossible since by elimination the inverse image

of any point in the projection of $Z(S)$ is non-empty. So $g^{(r)}$ contains no variable, so it is of the form $0 = b$. Since we assume the system S is consistent, b is 0, which says precisely that the equation g we started with is a linear combination of the equations in S. □

1.6 Diagonally Dominant Systems

This section is not used later in the book. It is a source of examples of $n \times n$ homogenous systems whose only solution is the trivial one.

Suppose we have a homogeneous system of n equations in n variables, which we write in the usual way as

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0, \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= 0 \end{aligned} \tag{1.17}$$

where the a_{ij} can be real or complex.

Assume that the system is *diagonally dominant*. This means that for each equation f_i in the system

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{rs}|, \quad \text{for } 1 \leq i \leq n. \tag{1.18}$$

Thus the absolute value of the 'diagonal coefficient' a_{ii} is greater than the sum of the absolute values of the other coefficients in the same equation. Then

Theorem 1.6.1. *The only solution to a diagonally dominant system is the trivial solution $x_j = 0$, $1 \leq j \leq n$.*

Proof. We prove this by contradiction. Assume there is a non-trivial solution $x_j = c_j$, for constants c_j. Some $|c_k|$ is greatest among the $|c_j|$, $1 \leq j \leq n$. Since the solution is non-trivial, $|c_k| > 0$. Because the c_j form a solution, all equations, in particular the k-th equation vanishes:

$$a_{kk}c_k = -\sum_{j=1, j\neq k}^{n} a_{kj}c_j,$$

so that

$$|a_{kk}||c_k| \leq \sum_{j=1, j\neq k}^{n} |a_{kj}||c_j|.$$

By choice of k, $|c_k| \geq |c_j|$ for all j, so

$$|a_{kk}||c_k| \leq \sum_{j=1, j\neq k}^{n} |a_{kj}||c_k|.$$

Divide by $|c_k| > 0$ to get

$$|a_{kk}| \leq \sum_{j=1, j\neq k}^{n} |a_{kj}|.$$

This contradicts the hypothesis that the system is diagonally dominant. □

So we have a way of producing $n \times n$ homogeneous systems that only have the trivial solution. The proof works (just by relabeling the variables) as long as there is one coefficient in each equation that satisfies the equivalent of (1.18), as long as that coefficient is that of a different variable in each equation.

Example 1.6.2. Consider the homogeneous $n \times n$ system with $a_{ii} = \pm n$ and all the other terms $a_{ij} = \pm 1$, $i \neq j$. This system is diagonally dominant, as you should check

Exercise 1.6.3. Write down some numerically explicit diagonally dominant 3×3 systems, especially where the diagonal coefficients are negative.

Exercise 1.6.4. Now take a diagonally dominant system with real coefficients. Assume that that all the diagonal terms are positive. Then do Gaussian elimination. Show that at each step the new system obtained is diagonally, so Gaussian elimination can continue without changing the order of the equations. Thus the system only has the trivial solution.

Exercise 1.6.5. Do Gaussian elimination on any example of diagonally dominant matrices you found in Exercise 1.6.3 that satisfies the hypotheses of Exercise 1.6.4.

1.7 History

Example 1.7.1. Babylonians already knew how to solve this problem: see Neugebauer [20], p. 181-183, in two variables at least. Here is a typical example. A field of area a is to be planted with two different grains, one where the yield per unit area is g_1, the other where the yield is g_2. The goal is to have a certain total yield b of both grains, and the question is how much surface area x_1 to plant in the first grain, and how much x_2 to plant in the second grain. If the area of the field is a, then we have the two inhomogeneous linear equations

$$\begin{aligned} x_1 + x_2 &= a; \\ g_1 x_1 + g_2 x_2 &= b. \end{aligned}$$

Here is the Babylonian method of solution. If you plant half the surface area in grain 1, and the rest in grain 2, you get a difference from the desired yield of

$$\begin{aligned}
&b-(\frac{a}{2}g_1+\frac{a}{2}g_2)\\
&=g_1x_1+g_2x_2-\frac{x_1+x_2}{2}(g_1+g_2)\\
&=g_1x_1+g_2x_2-\frac{1}{2}(g_1x_1+g_2x_2+g_1x_2+g_2x_1)\\
&=\frac{1}{2}(g_1x_1+g_2x_2-g_1x_2-g_2x_1)\\
&=\frac{1}{2}(g_1-g_2)(x_1-x_2)
\end{aligned}$$

This allows us to solve for $(x_1-x_2)/2$ on the right hand side:

$$\frac{x_1-x_2}{2}=b-(\frac{a}{2})\frac{g_1+g_2}{g_1-g_2}$$

Since the first equation gives us $\frac{x_1+x_2}{2}=\frac{a}{2}$, by adding and subtracting we can find x_1 and x_2.

This is actually more than just a linear algebra problem, in that we insist that the solutions $\bar{x}_1$ and $\bar{x}_2$ be non-negative. Therefore $x_i \le a$. We may assume that $g_1 \ge g_2$. Then we must have $g_1a \ge b$ and $g_2a \le b$. The numerical case that is actually treated is (in suitable units) $a=30$, $b=18.2$, $g_1=20$ and $g_2=15$.

However, if we forget about this positivity requirement, the only case where we will not get a unique solution is when $g_1=g_2$, which makes perfect sense. If the yields of the two grains are the same, it obviously does not matter how much we plant of one or the other. If $g_1=g_2$ then we only get a solution if $a=b/g_1$.

A shorter account of Babylonian Mathematics, in English, is given by Neugebauer in [19]. A more recent account in given in Robson [22]

Example 1.7.2. Chinese linear algebra is very well presented in Roger Hart's book [11].

Example 1.7.3. In a letter to De L'Hospital in 1693 Leibnitz indexed the coefficients of linear equations in the way we do above: the first index gives the equation the coefficient belongs to, and the second the 'letter' it belongs to. Then he shows how to eliminate the letters. His fundamental example is the inhomogeneous system of $n+1$ equations in n variables, which is our notation would be written:

$$\begin{aligned}
a_{10}+a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n&=0\\
a_{20}+a_{21}x_1+a_{22}x_2+\cdots+a_{2n}x_n&=0\\
&\vdots\\
a_{n+1,0}+a_{n+1,1}x_1+a_{n+1,2}x_2+\cdots+a_{n+1,n}x_n&=0
\end{aligned}$$

from which he concludes that the system has a solution if and only if the determinant of the square matrix of all the (a_{ij}), $1 \le i \le n+1$, $0 \le j \le n$. Leibnitz's letter was

only published in 1850, so his work did not have an effect on he development of linear algebra.

Instead, it is the 1750 work of Cramer that became famous because of its solution to (1.19), but where there are as many equations as variables. This rule is still referred to today in linear algebra books as Cramer's rule.

For more details on the early history of the solution of linear equations, see [18], Volume 1.

Example 1.7.4. We have seen that elimination from linear equations is rather easy. If we allow, instead of just equalities, a mixture of equalities and inequalities, the problem becomes more interesting. It was first considered by Fourier in 1826. We will consider this situation in Chapter 15. For comprehensive details also consult the first chapter of [27].

Chapter 2
Matrices

Abstract Matrices are the fundamental tool for computing in linear algebra. They are defined and studied in this chapter. After defining matrices themselves, our first task is to define the three fundamental matrix operations. The only difficult one is matrix multiplication. Then we focus on square matrices, the most interesting case we cause we can multiply two square matrices of the same size, and consider which ones have an inverse, a fundamental concept in linear algebra. In the two sections §2.4 and §2.9 on submatrices and block decomposition of matrices, we write matrices and parts of matrices in new ways. This will be useful later in the course. Then in a fundamental section we write systems of linear equations in terms of matrices, and we redo Gaussian elimination, studied in the first chapter, in the language of matrices. Using the same operations as in the first chapter, we put the matrix of coefficients in row echelon form, and then in reduced row echelon form, which is ideal for solving systems of linear equations. We see how this corresponds to row operations, that can be implemented by left multiplication of the matrix of coefficients by elementary matrices, which are very simple square matrices. Noting that elementary matrices are invertible, we show that any invertible matrix is a product of elementary matrices.

2.1 Matrices

The term matrix was proposed by Sylvester in his 1850 article [29] in the Philosophical Magazine. See Muir [18], Volume 2, p. 51. It is surprising to realize that matrices were only conceptualized as independent entities a hundred years after the theory of the solution of linear equations was understood.

A matrix of size $m \times n$ is a collection of mn scalars indexed in the following particular way:

$$a_{ij}, 1 \leq i \leq m, 1 \leq j \leq n.$$

These scalars are called the entries of the matrix.

We will write our matrices using capital roman letters, and their entries by the same lower case roman letter, with a double index. So for example, if A is a $m \times n$ matrix, we write $A = (a_{ij})$, where $1 \leq i \leq m$ and $1 \leq j \leq n$. We also write matrices out as rectangular arrays:

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \tag{2.1}$$

which allows us to talk about the rows and the columns of a matrix. We write the i-th row of the matrix A as $\mathbf{a}^i$ and the j-th column as $\mathbf{a}_j$.

So the 2×3 matrix

$$A = \begin{pmatrix} 1 & 2 & 4 \\ -1 & 3 & 5 \end{pmatrix}$$

has two rows and three columns, and

$$\mathbf{a}^2 = \begin{pmatrix} -1 & 3 & 5 \end{pmatrix} \text{ and } \mathbf{a}_3 = \begin{pmatrix} 4 \\ 5 \end{pmatrix}$$

Definition 2.1.1. A matrix of size $n \times 1$ is called a column vector of length n, or a n-column vector. A matrix of size $1 \times m$ is called a row vector of length m, or a m-row vector. Column vectors are written

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

but in the body of the text we will often write column vectors as a row vector, but with brackets: $\mathbf{x} = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}$. If we just say vector, we always mean a column vector.

Definition 2.1.2. We can define two simple operations on $m \times n$ matrices A and B.

1. First addition: $A + B = C$ where $C = (c_{ij})$ is the $m \times n$ matrix with $c_{ij} = a_{ij} + b_{ij}$ for all i, j. Thus the corresponding entries are added.
2. Then multiplication by a scalar c: $cA = (ca_{ij})$, so each entry of the matrix A is multiplied by the scalar c.

Definition 2.1.3. We can combine these two operations to form a *linear combination* of $m \times n$ matrices $A_1, \dots, A_k$, using scalars $c_1, \dots, c_k$: This is just the $m \times n$ matrix

$$A = c_1 A_1 + \cdots + c_k A_k.$$

This is the same concept as for systems of linear equations: see Definition 1.1.3.

Exercise 2.1.4. Determine the entry a_{ij} of the matrix A in terms of $c_1, \dots, c_k$ and the entry in position (i, j) of the matrices $A_1, \dots, A_k$.

Definition 2.1.5. Here are some special and important matrices to which we give names. First note that the diagonal of a square matrix A is the set of entries with equal indices: a_{ii}. The remaining elements are the off-diagonal terms.

- The $m \times n$ whose entries are all zero is written 0, or $0_{m\times n}$ if it is important to keep track of its size. The remaining definitions concern square matrices.
- The square matrix A is diagonal if all its off-diagonal terms are 0.
- The identity matrix I is the diagonal matrix with all diagonal terms equal to 1. If its size n needs to be recalled we write I_n. We usually write the entries of I as e_{ij}. So $e_{ii} = 1$ for all i, and $e_{ij} = 0$ if $i \neq j$.
- A is *upper-triangular* if all the terms below the diagonal are zero. In other words $a_{ij} = 0$ when $i > j$. Correspondingly A is lower triangular if $a_{ij} = 0$ when $i < j$.

Here are some examples in the 3×3 case.

$$0_{3\times 3} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.2}$$

The matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -2 & 0 \\ 0 & 0 & 5 \end{pmatrix} \tag{2.3}$$

is upper triangular and

$$B = \begin{pmatrix} 1 & 0 & 0 \\ 2 & -2 & 0 \\ 3 & 0 & 5 \end{pmatrix} \tag{2.4}$$

is lower triangular.

Definition 2.1.6. The *transpose* of a $m \times n$ matrix $A = (a_{ij})$ is the $n \times m$ matrix $B = (b_{ij})$ such that $b_{ij} = a_{ji}$. The transpose of A is written A^t. Thus the rows of A are the columns of S^t, and the columns of A are the rows of A^t. Obviously $(A^t)^t = A$. A square matrix that is equal to its transpose is called symmetric.

A row vector is the transpose of a column vector, and will usually be written $\mathbf{x}^t$ to make it explicit that we are dealing with a row vector.

The transpose of an upper triangular matrix is lower triangular. See the matrices A and B above..

Example 2.1.7. The transpose of the matrix

$$A = \begin{pmatrix} 1 & 2 & 4 \\ -1 & 3 & 5 \end{pmatrix} \text{ is } A^t = \begin{pmatrix} 1 & -1 \\ 2 & 3 \\ 4 & 5 \end{pmatrix}.$$

The matrix

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

is symmetric.

Now assume our matrix has complex entries. Then using complex conjugation (reviewed in Section B.5) we get the important:

Definition 2.1.8. The conjugate of the complex $m \times n$ matrix $A = (a_{ij})$ is $\overline{A} = (\overline{a_{ij}})$.

The conjugate transpose of A is the $n \times m$ matrix $B = (b_{ij})$ such that $b_{ij} = \overline{a_{ji}}$. The conjugate transpose of A is written A^*. A square matrix that is equal to its conjugate transpose is called hermitian. This is the most important kind of complex matrix. If A is real, its conjugate transpose is the same as its transpose, and to be hermitian is to be symmetric.

Example 2.1.9. The conjugate transpose of the square matrix

$$A = \begin{pmatrix} 1 & 2+\mathrm{i} \\ -1 & 3-\mathrm{i} \end{pmatrix} \text{ is } \quad A^* = \begin{pmatrix} 1 & -1 \\ 2-\mathrm{i} & 3+\mathrm{i} \end{pmatrix}.$$

The matrix

$$\begin{pmatrix} 1 & 2-\mathrm{i} \\ 2+\mathrm{i} & 3 \end{pmatrix}$$

is hermitian.

Problem 2.1.10. Show that for any square matrix A, the matrix $A + A^t$ is symmetric.

Problem 2.1.11. For any two matrices A and B of the same size, show that

$$(A+B)^t = A^t + B^t.$$

Problem 2.1.12. Show that for any square matrix A, the matrix $A + A^*$ is hermitian. Show that the diagonal elements of a hermitian matrix are real.

2.2 Matrix Multiplication

The fundamental matrix operation is multiplication of a $m \times n$ matrix A with a column vector $\mathbf{x}$ of length n to yield a column vector of length m.. Here is the all-important formula:

$$\boxed{\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{pmatrix}} \tag{2.5}$$

If we replace the column vector on the right hand side by **b**, then we have recreated (1.1.1) using matrix notation: $A\mathbf{x} = \mathbf{b}$.

Note the important special case where A is a row vector:

Definition 2.2.1.

$$\begin{pmatrix} a_1 & a_2 & \dots & a_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = a_1x_1 + a_2x_2 + \cdots + a_nx_n.$$

Calling the row vector **a** and the column vector **x**, we get **ax**.

Later in this book we will also call this the inner product, or scalar product of the two vectors **a** and **x**, written $\langle \mathbf{a}, \mathbf{x} \rangle$. See Chapter 8.

Example here.

Exercise 2.2.2. If A is any $m \times n$ matrix, show that $I_mA = A$ and $AI_n = A$.

Definition 2.2.3. The product $C = AB$ of a $m \times n$ matrix A multiplied on the right by a $n \times r$ matrix B is the $m \times r$ matrix $C = (c_{ik})$, where

$$c_{ik} = a_{i1}b_{1k} + a_{i2}b_{2k} + \cdots + a_{in}b_{nk}.$$

Using summation notation, we have

$$c_{ik} = \sum_{j=1}^{n} a_{ij}b_{jk}.$$

Note that as often in such cases we are summing over the repeated index j.

Remark 2.2.4. We can only form the product AB of a $m \times n$ matrix A by a $r \times s$ matrix B if $n = r$. In that case the product is a $m \times s$ matrix. This of course still works when B is a column vector of length n, the special case where $s = 1$, in which $C = AB$ is a column vector of length m.

Remark 2.2.5. In terms of the rows $\mathbf{a}^i$ of A and the columns $\mathbf{b}_k$ of B, matrix multiplication can be written using the notation of Definition 2.2.1 as:

$$AB = \begin{pmatrix} \mathbf{a}^1\mathbf{b}_1 & \mathbf{a}^1\mathbf{b}_2 & \dots & \mathbf{a}^1\mathbf{b}_r \\ \mathbf{a}^2\mathbf{b}_1 & \mathbf{a}^2\mathbf{b}_2 & \dots & \mathbf{a}^2\mathbf{b}_r \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}^m\mathbf{b}_1 & \mathbf{a}^m\mathbf{b}_2 & \dots & \mathbf{a}^m\mathbf{b}_r \end{pmatrix} \tag{2.6}$$

which we can write more compactly as $c_{ik} = \mathbf{a}^i\mathbf{b}_k$, $1 \le i \le m$, $1 \le k \le r$,

Exercise 2.2.6. Work out the matrix multiplication of a row vector A, namely a $1 \times n$ matrix, by a $n \times r$ matrix B.

Observe from (2.6) that the k-th column $\mathbf{c}_k$ of C only depends on k-th column $\mathbf{b}_k$ of B, while the i-th row $\mathbf{c}^i$ of C only depends on the i-th row $\mathbf{a}^i$ of A. In fact:

Proposition 2.2.7. *If $AB = C$, then $A\mathbf{b}_k = \mathbf{c}_k$, $1 \le k \le r$, and $\mathbf{a}^i B = \mathbf{c}^i$, $1 \le i \le m$.*

This can be reformulated:

$$\mathbf{c}_k = b_{1k}\mathbf{a}_1 + \cdots + b_{nk}\mathbf{a}_n = \sum_{j=1}^{n} b_{jk}\mathbf{a}_j \tag{2.7}$$

and

$$\mathbf{c}^i = a_{i1}\mathbf{b}^1 + \cdots + a_{in}\mathbf{b}^n = \sum_{j=1}^{n} a_{ij}\mathbf{b}^j, \tag{2.8}$$

showing that the columns $\mathbf{c}_k$ of C are linear combinations of the columns $\mathbf{a}_j$ of A, and the rows $\mathbf{c}^i$ of C are linear combinations of the rows $\mathbf{b}^j$ of B.

Proof. The first two equalities are the special cases of (2.6) when B has only one column and A has only one row.

For the reformulation,

$$\begin{aligned}
\mathbf{c}_k &= \begin{pmatrix} c_{1k} \\ c_{2k} \\ \cdots \\ c_{mk} \end{pmatrix} = \begin{pmatrix} a_{11}b_{1k} + a_{12}b_{2k} + \cdots + a_{1n}b_{nk} \\ a_{21}b_{1k} + a_{22}b_{2k} + \cdots + a_{2n}b_{nk} \\ \cdots \\ a_{m1}b_{1k} + a_{m2}b_{2k} + \cdots + a_{mn}b_{nk} \end{pmatrix} \\
&= \begin{pmatrix} a_{11}b_{1k} \\ a_{21}b_{1k} \\ \cdots \\ a_{m1}b_{1k} \end{pmatrix} + \begin{pmatrix} a_{12}b_{2k} \\ a_{22}b_{2k} \\ \cdots \\ a_{m2}b_{2k} \end{pmatrix} + \cdots + \begin{pmatrix} a_{1n}b_{nk} \\ a_{2n}b_{nk} \\ \cdots \\ a_{mn}b_{nk} \end{pmatrix} \\
&= b_{1k}\begin{pmatrix} a_{11} \\ a_{21} \\ \cdots \\ a_{m1} \end{pmatrix} + b_{2k}\begin{pmatrix} a_{12} \\ a_{22} \\ \cdots \\ a_{m2} \end{pmatrix} + \cdots + b_{nk}\begin{pmatrix} a_{1n} \\ a_{2n} \\ \cdots \\ a_{mn} \end{pmatrix} = b_{1k}\mathbf{a}_1 + \cdots + b_{nk}\mathbf{a}_n.
\end{aligned}$$

The second reformulation is proved the same way, so the proof is left to you. □

Exercise 2.2.8. Prove the second reformulation.

Exercise 2.2.9. For any $m \times n$ matrix A, compute the product $A0$, where 0 is the zero matrix of size $n \times r$. Also compute $0A$, where 0 is the zero matrix of size $s \times m$.

Theorem 2.2.10. *Let A be a $m \times n$ matrix, B a $n \times r$ matrix, and C a $r \times s$ matrix. Then*

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

Thus matrix multiplication is associative.

Proof. We first write down the (i,k)-th element of the matrix AB:

$$a_{i1}b_{1k} + a_{i2}b_{2k} + \cdots + a_{in}b_{nk} = \sum_{j=1}^{n} a_{ij}b_{jk}$$

Using this, we form the (i,l)-th element of the matrix $(AB)C$:

$$\sum_{k=1}^{r} \big(\sum_{j=1}^{n} a_{ij}b_{jk} \big) c_{kl} = \sum_{k=1}^{r} \big(\sum_{j=1}^{n} a_{ij}b_{jk}c_{kl} \big). \tag{2.9}$$

If instead we write down the (j,l)-th element of the matrix CD:

$$b_{j1}c_{1l} + b_{j2}c_{2l} + \cdots + b_{jr}c_{rl} = \sum_{k=1}^{r} b_{jk}c_{kl}$$

we can form the (i,l)-th element of the matrix $A(BC)$:

$$\sum_{j=1}^{n} a_{ij} \big(\sum_{k=1}^{r} b_{jk}c_{kl} \big) = \sum_{j=1}^{n} \big(\sum_{k=1}^{r} a_{ij}b_{jk}c_{kl} \big). \tag{2.10}$$

We need to convince ourselves that the sums in (2.9) and (2.10) are the same. This is true because we are summing the same terms over the same variables: we have changed the order of summation, but in finite sums that makes no difference. □

Example 2.2.11. When all the matrices are 2×2, for the triple product D we get, both ways

$$\begin{aligned}
d_{11} &= a_{11}b_{11}c_{11} + a_{11}b_{12}c_{21} + a_{12}b_{21}c_{11} + a_{12}b_{22}c_{21} \\
d_{12} &= a_{11}b_{11}c_{12} + a_{11}b_{12}c_{22} + a_{12}b_{21}c_{12} + a_{12}b_{22}c_{22} \\
d_{21} &= a_{21}b_{11}c_{11} + a_{21}b_{12}c_{21} + a_{22}b_{21}c_{11} + a_{22}b_{22}c_{21} \\
d_{22} &= a_{21}b_{11}c_{12} + a_{21}b_{12}c_{22} + a_{22}b_{21}c_{12} + a_{22}b_{22}c_{22}
\end{aligned}$$

Notice the beautiful pattern.

Exercise 2.2.12. In the two cases below, compute $(AB)C$ and $A(BC)$ and note they are equal.

1. A 2×2, B 2×2, C 2×2 with numeric values.
2. A 2×3, B 3×2, C 2×2 with numeric values.

Theorem 2.2.13. *Let A and B be $m \times n$ matrices, and let C and D be $n \times r$ matrices, and let $c \in F$ be a scalar. Then*

$$A(C+D) = AC + AD \text{ and } (A+B)C = AC + BC.$$

So matrix multiplication distributes over matrix addition, whenever the two operations are possible. Furthermore

$$(cA)(D) = c(AD) \text{ and } A(cD) = c(AD).$$

Proof. By Definition 2.2.3, the element of $A(C+D)$ in position (i,k) is

$$\begin{aligned} A(C+D)_{ik} &= a_{i1}(c_{1k}+d_{1k}) + a_{i2}(c_{2k}+d_{2k}) + \cdots + a_{in}(c_{nk}+d_{nk}) \\ &= a_{i1}c_{1k} + a_{i2}c_{2k} + \cdots + a_{in}c_{nk} + a_{i1}d_{1k} + a_{i2}d_{2k} + \cdots + a_{in}d_{nk} \\ &= (AC)_{ik} + (AD)_{ik} \end{aligned}$$

The other parts of the proof can be done the same way. Instead, we prove $(cA)(D) = c(AD)$ by noting that the i-th row of cA is $c\mathbf{a}^i$, so by Definition 2.2.1 the (i,k)-th entry of $(cA)(D)$ is

$$(c\mathbf{a}^i)\mathbf{d}_k = c(\mathbf{a}^i\mathbf{d}_k).$$

This is c times the (i,k)-th entry of AD, as required. □

Exercise 2.2.14. Prove $(A+B)C = AC+BC$ by writing the (i,k)-th entry of each side in terms of the appropriate rows of A and B and columns of C.

Proposition 2.2.15. *If A is a $m \times n$ matrix, and B a $n \times r$ matrices, then $(AB)^t = B^tA^t$.*

So the transpose of a product is the product of the transposes, but in the reverse order. Note that A^t is a $n \times m$ matrix, and B^t a $r \times n$ matrix, so the product B^tA^t is possible. The proof is easy and left to you. Write $C = AB$, and $D = B^tA^t$, and compare the (i,k)-th entry of C with the (k,i)-th entry of D.

Exercise 2.2.16. If A is a $m \times n$ matrix, B a $n \times r$ matrices, and C a $r \times s$ matrix, then $(ABC)^t = C^tB^tA^t$.

Exercise 2.2.17. For complex matrices, show that $\overline{AB} = \overline{A}\,\overline{B}$. Then show that for the conjugate transposes, $(AB)^* = B^*A^*$.

Definition 2.2.18 (Matrix Multiplication Algorithm). Here are the details for the computation of the product matrix $C = AB$, done in place, with the entries c_{ik} first being initiatized to 0 and then being updated:

```
for i = 1 : m,
        for k = 1 : r,
                for j = 1 : n,
                        c_ik = a_ij b_jk + c_ik
                end
        end
end
```

The first two loops just tell you which entry you are working with; with the inner loop producing the sum of products. in the formula.

2.3 Square Matrices

A square matrix is a matrix with the same number of rows as columns. Instead of saying a '$n \times n$ matrix', we will sometimes say a 'square matrix of size n'. The extra

feature that arises when dealing with two square matrices A and B of the same size it that we can form both products AB and BA. As a special case we can raise a square matrix A to any power, which we write A^n. In fact we can take polynomials in the square matrix A:

$$A^n + c_{n-1}A^{n-1} + \cdots + c_1A + c_0I.$$

Because matrix multiplication is associative and the matrices A and the identity matrix I commute ($AI = IA = A$), polynomials in matrices behave like ordinary polynomials. We will use this intensively later in this book.

Remark 2.3.1. Matrix multiplication is not necessarily commutative. If A and B are two square matrices of the same size n, so that AB and BA are both square matrices of size n, it is not necessarily the case that $AB = BA$. Proposition 2.2.7 shows how to construct examples systematically.

Give examples here.

Exercise 2.3.2. Let A and B be square matrices that commute: $AB = BA$. Using Theorem 2.2.13, show that

$$(A+B)^3 = A^3 + 3A^2B + 3AB^2 + B^3 \text{ and } (A+B)(A-B) = A^2 - B^2.$$

This shows that we can do algebra with squares matrices as with numbers, taking account, of course, that matrix multiplication is not generally commutative.

One key feature of multiplication of numbers is that there is a neutral element for multiplication, usually denoted 1. There also is a neutral element for matrix multiplication, the identity matrix I.

Continuing the analogy with multiplication of numbers, we may ask if every square matrix other than 0 has an inverse. What does that mean?

Definition 2.3.3. A square matrix A has an inverse if there exists a square matrix B, called the inverse of A, of the same size as A such that

$$AB = I = BA. \tag{2.11}$$

The equation $AB = I$ says that B is a right inverse of A, and $BA = I$ says that B is a left inverse of A.

So we require that A have a left inverse and a right inverse, and that they be the same. It is reasonable to require both, since matrix multiplication is not commutative. It is easy to show that we do not need all this.

Theorem 2.3.4. *If A has an inverse, then its inverse B is unique. If A has a left inverse B and a right inverse C, then they are equal and the inverse of A.*

Proof. Assume there is another matrix C satisfying (2.11) when C replaces B. Then

$$C = CI = C(AB) = (CA)B = IB = B. \tag{2.12}$$

This proves the first statement. The proof only uses that C is a left inverse and B a right inverse, so we get the last statement. □

Definition 2.3.5. If A has an inverse, it is said to be invertible or nonsingular. The unique inverse is written A^{-1}. A matrix that is not invertible is *singular*.

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

Exercise 2.3.6. Show by direct computation that the inverse of the upper triangular matrix

$$\begin{pmatrix} 1 & 1 & \dots & 1 & 1 \\ 0 & 1 & \dots & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} \quad \text{is} \quad \begin{pmatrix} 1 & -1 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & 1 & -1 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

Proposition 2.3.7. *Any product of invertible matrices A and B is invertible, and $(AB)^{-1} = B^{-1}A^{-1}$.*

Proof. Indeed just compute, using the associativity of matrix multiplication:

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

and

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

□

Proposition 2.3.8. *If a matrix A is invertible, its transpose A^t is too.*

Proof. This is easy. Let B be the inverse of A, so $AB = I = BA$. Take transposes using Proposition 2.2.15 to get $B^tA^t = I = A^tB^t$, so $C = A^t$ is invertible. □

Exercise 2.3.9. Show that the inverse of the transpose is the transpose of the inverse: $(A^t)^{-1} = (A^{-1})^t$. Hint: take the transpose of the identity $AA^{-1} = I$ and use the uniqueness of the inverse.

Exercise 2.3.10. If a complex matrix A is invertible its conjugate $\overline{A}$ is invertible. Its conjugate transpose A^* is also invertible, with inverse $(A^{-1})^*$.

One of the main questions of linear algebra is: which square matrices are invertible? Clearly not the zero matrix, since its product with any matrix is again the zero matrix. However there are many other square matrices that are non invertible. An important goal of this course is to develop criteria telling us when a matrix is invertible. In particular we will associate to every square matrix of size n a number $r \leq n$, called its rank, and prove that the matrix is invertible if and only if $r = n$. The rank of any matrix (not just square matrices) is defined in §5.6. Later we will show that a square matrix is invertible if and only if its determinant is non–zero: Corollary 11.3.13.

For triangular matrices we can give an criterion for invertibility. Later in this chapter we will give a less computational proof of the same statement, so you can skip the second half of the proof.

Proposition 2.3.11. *A diagonal matrix A is invertible if and only if all its diagonal entries are non-zero. Its inverse B is diagonal with diagonal elements* $b_{ii} = \frac{1}{a_{ii}}$.

An upper–triangular matrix A is invertible if and only if all its diagonal elements are non-zero. Its inverse B is also upper-triangular, with diagonal elements $b_{ii} = \frac{1}{a_{ii}}$. *Similarly for lower–triangular matrices.*

Proof. The first assertion is of course special case of the second one, but it is worth giving the one–line proof: just multiply A by the matrix B given by the statement of the proposition, and note that the product is the identity matrix.

Now assume A is upper–triangular, so $a_{ij} = 0$ when $i > j$. We will attempt to solve for its inverse B, and determine under which conditions on A and B this can be done. Assume B is the inverse of the upper–triangular A. Then last row of AB, which is supposed to be $\begin{bmatrix} 0 \dots 0 & 1 \end{bmatrix}$ is

$$\begin{bmatrix} a_{nn}b_{n1} & a_{nn}b_{n2} & \dots & a_{nn}b_{nn} \end{bmatrix}$$

showing $a_{nn} \neq 0$ and $b_{nn} = 1/a_{nn}$ since the last entry must be 1. Then $b_{nj} = 0$ for all the others. Using these values, we see that the first $n-1$ entries of the next-to-last row are

$$\begin{bmatrix} a_{n-1,n-1}b_{n-1,1} & a_{n-1,n-1}b_{n-1,2} & \dots & a_{n-1,n-1}b_{n-1,n-1} \end{bmatrix}$$

So in the same way, we see that $a_{n-1,n-1} \neq 0$ and $b_{n-1,n-1} = 1/a_{n-1,n-1}$ and therefore $b_{n-1,j} = 0$ for $j = 1, \dots, n-2$. Furthermore $b_{n-1,n-1} = \frac{1}{a_{n-1,n-1}}$. Continuing this argument by decreasing the row index, we see that all the diagonal entries of A are non-zero, and B is also upper-triangular with non-zero elements $b_{ii} = \frac{1}{a_{ii}}$ on the diagonal. This shows that an upper–triangular matrix is only invertible if all its diagonal elements are non–zero, and it computes the diagonal entries

Now prove the converse: assume A is upper–triangular with non–zero diagonal terms. Then we prove it it has a right inverse B. We have already shown that B is upper-triangular with diagonal terms $b_{ii} = \frac{1}{a_{ii}}$. Next we solve for the terms $b_{i,i+1}$. Take the product of the i-th row of A with the $i+1$-th column of B, we, get, for the $(i, i+1)$-th entry of the product, which must be 0:

$$a_{ii}b_{i,i+1} + a_{i,i+1}b_{1+1,i+1} = 0$$

We already know $b_{1+1,i+1}$, and a_{ii} is non–zero, so we can solve for all elements $b_{i,i+1}$, $1 \leq i \leq n-1$. This allows us to compute all the terms on the super diagonal. Continue in this way: next we solve for the elements $b_{i,i+2}$ in terms of the b_{jj} and $b_{j,j+1}$. in other words, by induction on $j-i$ we compute b_{ij}, $i < j$, by evaluating the ij term of the product AB, which must by 0:

$$a_{ii}b_{ij} + a_{i,i+1}b_{1+1,j} + \dots + a_{i,j-1}b_{j-1,j} + a_{ij}b_{jj} = 0$$

Thus $b_{i,j}$ can be solved in terms of $b_{k,j}$ with $j-k < j-i$. Thus A has a right inverse. By repeating the argument on CA, we see that A has a left inverse. So by Theorem 2.3.4, it is invertible.

For the lower-triangular case just reduce to the upper-triangular case by taking transposes. □

Example 2.3.12. Consider the $n \times n$ matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

with

$$a_{ij} = \begin{cases} 1, & \text{if } j = i+1; \\ 0, & \text{otherwise.} \end{cases}$$

A is not invertible, as we know. Furthermore, it satisfies $A^n = 0$, but $A^{n-1} \neq 0$. Work out the multiplication and determine what all the A^k are.

A matrix A for which a power A^k is the 0 matrix is *nilpotent*. Being nilpotent implies A is singular. Indeed, suppose by contradiction that A is invertible with inverse B, so

$$AB = I.$$

Let k be the smallest integer such that $A^k = 0$ Multiply the equation by the matrix A^{k-1} on the left, giving This gives $A^k B = A^{k-1}$, an impossibility since the left hand side is 0 and the right hand side is not.

How do you decide if a general square matrix A is invertible? Use Gaussian elimination, as described in §2.5 to transform A into a upper triangular matrix B. Then show that A is invertible if and only if B is: this is Proposition 2.8.8. Then apply Proposition 2.3.11 to B

2.4 Submatrices

This short section only contains definitions. Given a $m \times n$ matrix A, it is useful to refer to *submatrices* of A.

In this chapter we will only need very special cases of submatrices: first we may remove the first k rows of A. Therefore we are left with a $(m-r) \times n$ matrix B where $b_{ij} = a_{i+k,j}$. We might remove the first l columns, getting the matrix $m \times (n-l)$ matrix C with $c_{ij} = a_{i,j+l}$. Of course we could also do both simultaneously, to get a matrix D with $d_{ij} = a_{i+k,j+l}$. These are the submatrices we will need immediately.

For later purposes we need notation for the general case. No need to read this until you need it. Pick first a certain number k of rows of A: those with indices i_1, i_2, …, i_k. Then pick first a certain number l of rows of A: those with indices j_1, j_2, …, j_l. For short call I the collection of indices of rows, and J the collection of

indices of columns. Then we can form a $k \times l$ submatrix of A, indexed by the rows and columns:

$$A_J^I = A(i_1, \ldots, i_k; j_i, \ldots, j_l) = \begin{pmatrix} a_{i_1 j_1} & \ldots & a_{i_1 j_l} \\ \vdots & \ddots & \vdots \\ a_{i_k j_1} & \ldots & a_{i_k j_l} \end{pmatrix} \tag{2.13}$$

A *principal submatrix* arises from picking the same rows as columns. Our notation is

$$A_I = A(i_1, \ldots, i_k) = \begin{pmatrix} a_{i_1 i_1} & \ldots & a_{i_1 i_k} \\ \vdots & \ddots & \vdots \\ a_{i_k i_1} & \ldots & a_{i_k i_k} \end{pmatrix} \tag{2.14}$$

The *leading principal submatrix* of size k is $A(1,2,\ldots,k)$: pick the first k rows and the first k columns of A.

Principal submatrices are by definition square, while more general submatrices are only square if $k = l$.

Example 2.4.1. Assume A is the matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

Then the notation $A(1,2;2,3)$ means to take the first two rows of A, and the second and third columns.

$$A(1,2;2,3) = \begin{pmatrix} 2 & 3 \\ 5 & 6 \end{pmatrix}, \text{ while } A(1,2) = \begin{pmatrix} 1 & 2 \\ 4 & 5 \end{pmatrix}$$

is the leading principal submatrix of size 2 of A.

More numerical examples here.

2.5 Gaussian Elimination in Matrix Notation

We now redo elimination on systems of linear equations, studied in §1.3, using matrix notation and operations. Equation (1.1) can be written $A\mathbf{x} = \mathbf{b}$, where A is the $m \times n$ matrix (a_{ij}), $\mathbf{x}$ is the column vector of n unknowns, and $\mathbf{b}$ is the m-column vector of constants. Recall that we call this system of equations S, and its set of solutions $Z(S)$.

The key point is that the operations on linear equations used to derive Theorem 1.3.3 and others can be formulated in terms of matrix operations on A and $\mathbf{b}$. We will work with two different matrices: either the $m \times n$ matrix A, called the *coefficient matrix* or the matrix $m \times (n+1)$ whose first n columns form the coefficient matrix

A, and whose last column is $\mathbf{b}$. It is called the *augmented matrix* and usually written $\big(A \mid \mathbf{b}\big)$.

Example 2.5.1 (Special Case). We start with the simplest, but most important case: the matrix A is invertible matrix, and therefore square. So the system S has the same number of variables as of equations. Then, multiplying both sides of the equation by the inverse A^{-1} of A, we get the unique solution $\mathbf{x} = A^{-1}\mathbf{b}$ of the system. So $Z(S)$ is a point. In particular the system is consistent. To get the solution we need to compute A^{-1}. We will learn one way of doing this in §2.6.

The goal of Gaussian elimination in the general case is to replace the augmented $m \times (n+1)$ matrix $\big(A \mid \mathbf{b}\big)$ by a simpler matrix $\big(C \mid \mathbf{d}\big)$ of the same size, where the systems $A\mathbf{x} = \mathbf{b}$ and $C\mathbf{x} = \mathbf{d}$ are equivalent, meaning they have the same solutions. The i-th equation of the system $A\mathbf{x} = \mathbf{b}$ can be written $\mathbf{a}^i\mathbf{x} = b_i$, where $\mathbf{a}^i$ is the i-th row of A.

If we multiply the i-th equation by the non-zero scalar c, the i-th row $\mathbf{a}^i$ of A is replaced by $c\mathbf{a}^i$, b_i is replaced by cb_i, so the solutions $Z(S)$ do not change. We are using the operation of multiplication of a matrix (in this case a row vector) by a scalar, see Definition 2.1.2.

If we interchange two equations, say the first and the second equations, then we interchange the first and second row of $\big(A \mid \mathbf{b}\big)$. This also does not change $Z(S)$.

Finally, if we add to the second equation, a multiple c of the first equation, then we replace the second row of A by $\mathbf{a}^2 + c\mathbf{a}^1$, and leave the other rows unchanged. We replace the second element b_2 of $\mathbf{b}$ by $b_2 + cb_1$. Here we are using the operation of matrix addition (in the case of row vectors) in addition to multiplication by a scalar: again see Definition 2.1.2.

Motivated by these operations, we make the following definition. We only write it down for a matrix A that can stand for either the coefficient matrix or the augmented matrix.

Definition 2.5.2. Given a $m \times n$ matrix A, if we perform one of the following three rows operations on A, the new $m \times n$ matrix A' is *row equivalent* to A:

1. Multiply a row of A by a non-zero constant;
2. Interchange two rows of A;
3. Add to a row of A a multiple of a different row.

More generally if A can be transformed into A' by a finite sequence of row operations, A' is *row equivalent* to A.

You should now use the matrix in Example 1.4.3 to repeat the operations we did on the system of equations.

Theorem 2.5.3. *Row equivalence is an equivalence relation on $m \times n$ matrices.*

Proof. We must check the three properties of Definition B.2.1.

First A is row equivalent to itself: use, for example, the trivial row interchange.

Next we need to show that if A is row equivalent to A', then A' is equivalent to A. The key is to realize that each one of the three row operations has an inverse operation that undoes it. So it you multiply the i-th row of A by $c \neq 0$ to get A', then by multiplying the i-th row of A' by $1/c$ you get A back. The reader should find the inverse for the other two operations.

Finally we need to show that if A is row equivalent to B and B to C, then A is row equivalent to C. This is clear: just use the row operations that transform A to B followed by the row operations that transform B to C. □

Exercise 2.5.4. Fill in the details of the proof.

The beauty of having an equivalence relation is that $m \times n$ matrices are partitioned into equivalence classes: see Definition B.2.3 and Proposition B.2.4. This allows us to search for the most convenient matrix in each equivalence class: this is what we do next.

Here is a convenient form for A.

Definition 2.5.5. The $m \times n$ matrix A is in *row echelon form* if

1. All the rows of A that consist entirely of zeroes are below any row of A that has a non-zero entry;
2. If row $\mathbf{a}^i$ has its first non-zero entry in position j, then row $\mathbf{a}^{i+1}$ has its first non-zero entry in position $> j$. In other words if j_i denotes the column index of the first non-zero entry a_{ij} of $\mathbf{a}^i$, then $j_1 < j_2 < \cdots < j_m$, where we only take into account rows with a non-zero entry.

Example 2.5.6. The matrices

$$\begin{pmatrix} \mathbf{1} & 2 & 3 \\ 0 & \mathbf{4} & 0 \\ 0 & 0 & \mathbf{1} \end{pmatrix}, \begin{pmatrix} 0 & \mathbf{2} & 3 \\ 0 & 0 & \mathbf{2} \\ 0 & 0 & 0 \end{pmatrix}, \text{ and } \begin{pmatrix} \mathbf{-1} & 2 & 3 \\ 0 & \mathbf{-2} & 2 \\ 0 & 0 & 0 \end{pmatrix}$$

are in row echelon form, while

$$\begin{pmatrix} \mathbf{1} & 2 & 3 \\ 0 & \mathbf{4} & 0 \\ 0 & \mathbf{2} & 1 \end{pmatrix}, \begin{pmatrix} 0 & \mathbf{2} & 3 \\ 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} \end{pmatrix}, \text{ and } \begin{pmatrix} 0 & 0 & \mathbf{3} \\ 0 & 0 & \mathbf{2} \\ 0 & 0 & \mathbf{1} \end{pmatrix}$$

are not. In each matrix the first non-zero element of each row is marked in bold.

Remark 2.5.7. For A to be in row–echelon form is a generalization of being upper–triangular: it implies that the first non-zero entry a_{ij} of row $\mathbf{a}^i$ in a position $j \geq i$. Thus if A is in row echelon form, $a_{ij} = 0$ for all $j < i$. If A is square, it means that A is upper–triangular.

Exercise 2.5.8. Check that the matrix of coefficients of the left hand side of the system of equations produced by elimination in §1.3 is in row–echelon form.

The central theorem, which mimics Theorem 1.3.6 in the language of matrices, is:

Theorem 2.5.9. *Any matrix A can be put in row–echelon form by using only row operations.*

Proof. Consider the columns of the $m \times n$ matrix A moving from left to right, therefore starting with column $\mathbf{a}_1$.

1. If column $\mathbf{a}_1$ is the zero–vector, move to column 2. If $\mathbf{a}_2$ is still the zero–vector, keep moving right until you get to the first l such that $\mathbf{a}_l$ is not the zero–vector. If there is no such l then the matrix A is the zero matrix, and we are done.
2. Otherwise pick any row $\mathbf{a}^k$ with $a_{k,l} \neq 0$. If $k = 1$, do nothing. Otherwise interchange rows 1 and k. Continue to call the new matrix A.
3. Then subtract from each $\mathbf{a}^k$, $k > 1$, the appropriate multiple of $\mathbf{a}^1$ so that the $(1,k)$-th entry of the new matrix, still called A, is zero. Therefore A's first $l-1$ columns are zero, and its l-th column is zero except for entry $a_{1,l}$, which is definitely non–zero. This entry is called the first *pivot*.

Now repeat this operation to the $(m-1) \times (n-l)$ submatrix A_1 of A consisting of the last $m-1$ rows and the last $n-l$ columns. Notice that the first pivot, which is entry $a_{1,l}$, is not in A_1, so its position will not change in the repetitions, Keep repeating the procedure, say r times, until the new matrix A_r is the 0 matrix. Then you have r pivots, one in each of the first r rows. If a_{i,j_i} is the i-th pivot, then by construction $j_i < j_{i+1}$, for all i. Each pivot is the first non–zero entry in its row. The columns that contain pivots are called the pivot columns, and the remaining ones are called the free variables. □

Make sure you can relate this construction with that in Example 1.4.6.

Exercise 2.5.10. Reduce the matrices in Example 2.5.6 that are not already in row echelon form to row echelon form.

Now we get to the central result concerning systems of linear equations, an easy corollary of Theorem 2.5.9.

Theorem 2.5.11. *Let $A\mathbf{x} = \mathbf{b}$ be a system of m equations in n variables. Do row operations on the augmented matrix $\big(A \mid \mathbf{b}\big)$ to put it in row echelon form, Then the new system $C\mathbf{x} = \mathbf{d}$ is equivalent to the original system, so by definition it has the same solutions as the original system.* □

Many examples here.

The number r_A could conceivably depends on what choices of pivot are made to put A in row echelon form. We will see in §5.6 that this is not the case: r_A only depends on A, and is called the row rank of A. This number is the same number as the r from Theorem 1.3.6.

Example 2.5.12. If you happen to be able to write $A = LU$, the product of an invertible lower-triangular matrix L by an invertible upper-triangular matrix U, then A is invertible and the unique solution can be easily found.

With these hypotheses, it is equivalent to first solve $L\mathbf{y} = \mathbf{b}$, which since L is invertible gives us a unique $\mathbf{y} = L^{-1}\mathbf{b}$, and then solving $U\mathbf{x} = \mathbf{y}$. Since U is invertible,

this gives a unique $\mathbf{x}$, and by multiply the last equation on the left by L, we get $LU\mathbf{x} = L\mathbf{y}$, which is the equation we wanted to solve.

Why is this better than just knowing that A is invertible? It is almost trivial to solve the system $U\mathbf{x} = \mathbf{y}$, when $\mathbf{y}$ is known. Indeed the last row is $u_{nn}x_n = y_n$, so since $u_{nn} \neq 0$ by assumption, $x_n = \frac{y_n}{u_{nn}}$. The previous equation is $u_{n-1,n-1}x_{n-1} + u_{n-1,n}x_n = y_{n-1}$, which allows to to solve for the only unknown x_{n-1}, because $u_{n-1,n-1} \neq 0$. Continuing in this way we can solve for all the x_i, in a small number of steps. This simple process is called **back substitution**. Similarly we can easily solve $L\mathbf{y} = \mathbf{b}$, by an analogous process of **forward substitution**. Unfortunately it is not always possible to write an invertible matrix in LU form.

Remark 2.5.13. This is why it may not be possible to write a square matrix as LU, as in Example 2.5.1: because we allow row interchanges in Gaussian elimination, this may prevent it. We will see in §11.2 how to analyze this problem by analyzing the row interchanges.

2.6 Reduced Row–Echelon Form

Suppose the $m \times n$ coefficient matrix A is in row–echelon form. We now simplify it even further using row operations, to put it in reduced row echelon form. In the older literature this is called Hermite form. In the current literature it is called RREF form. The row operations needed to get from row–echelon form to RREF form are known as *back substitution*. This is is a generalization of what we did in Example 2.5.1.

As before, let r be the number of rows with a non–zero element in them, and for these rows let j_i be the smallest column index of a non-zero element in the i-th row. Thus ij_i, $1 \leq i \leq r$ are the coordinates of the i-pivot. Because A is in row–echelon form, the j_i increase strictly with i. Example here.

Then do the following elementary row operations.

- For $i = r$ down to 1, multiply the i–th row of A by $1/a_{i,j_i}$.
  - For $k = i-1$ down to $k = 1$, subtract a_{k,i_j} times the i-th row from the k-th row.

At each step you update the matrix A. Here is what is happening. First you make the last pivot take the value 1. Then you subtract a suitable multiple of this pivot row from all the rows above it to make the terms above the pivot 0. Then you repeat this on the next-to-last pivot. Notice that because of the previous operations, this does not affect the 0s already produced.

After these operations are complete each pivot column has a 1 for a pivot and a 0 in all other entries. No improvements is made to the free columns.

After this process the matrix is in *reduced row echelon form*.

Definition 2.6.1. A $m \times n$ matrix A in *reduced row–echelon form* if has two different kinds of columns.

1. For $i = 1$ to r, column $veca_{j_i}$ has a 1 is position (i, j_i) and zeroes every where else. Furthermore $j_i < j_{i+1}$, for $i < r$. They are called the pivot, or bound columns.
2. For any column $\mathbf{a}_l$, with $j_i < l < j_{i+1}$, then $a_{kl} = 0$ for $k > i$. These are the free columns. There are $n - r$ of them.

Therefore the rows $\mathbf{a}^i$ with $i \leq r$ have their first non–zero element in position j_i. The other rows are $\mathbf{0}$.

Need examples here.

Remark 2.6.2. When we studied elimination in systems of linear equations in Chapter 1, we arranged by changing the order of the variables to have all the bound columns on the left, and the rows with all coefficients zero at the bottom.

Example 2.6.3. In the following matrices, a, b, c, d, e, f denote arbitrary scalars. The matrices

$$A = \begin{pmatrix} 1 & a & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \text{ and } B = \begin{pmatrix} 1 & a & 0 & c \\ 0 & 0 & 1 & b \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

are in reduced row echelon form. Columns 1, 3 and 4 of A are bound. Columns 1 and 3 of B are bound. The matrices

$$C = \begin{pmatrix} 1 & a & 0 & 0 & b \\ 0 & 0 & 1 & 0 & c \\ 0 & 0 & 0 & 1 & d \end{pmatrix} \text{ and } D = \begin{pmatrix} 1 & 0 & 0 & a & b \\ 0 & 1 & 0 & c & d \\ 0 & 0 & 1 & e & f \end{pmatrix}$$

are also in reduced row echelon form. Columns 2 and 5 of C are free; columns 4 and 5 of D are free.

Exercise 2.6.4. Put the matrices of Example 2.5.6 into RREF form.

2.7 Solving Linear Systems of Equations

Next we record the following elementary consequence of matrix multiplication. Do not forget Example 2.5.1.

Theorem 2.7.1. *Consider the system of m equations in n variables, written in matrix notation as $A\mathbf{x} = \mathbf{b}$. As always A is the $m \times n$ matrix (a_{ij}), $\mathbf{x}$ is the n column vector with coordinates x_j and $\mathbf{b}$ is the m column vector with coordinates b_i.*

1. This system can be written as the vector equation

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \mathbf{b} \tag{2.15}$$

where $\mathbf{a}_j$ is the j-th column of A.

2. It can also be written as the system of m matrix products

$$\mathbf{a}^i\mathbf{x} = b_i\ ,\ 1 \le j \le n, \tag{2.16}$$

where $\mathbf{a}^j$ *is the* j*-th row of* A*.*

Recall that when $\mathbf{b}$ is not the zero vector, the inhomogeneous system may be inconsistent: no solutions at all. In §1.3 we showed how elimination allows us to determine effectively when this is the case.

Suppose we take a linear combination of the rows of $A\mathbf{x} = \mathbf{b}$. A moment's thought we tell you that this amounts to multiplying the system on the left by the row vector $\mathbf{y}^t$ of length m: then

$$\mathbf{y}^t A\mathbf{x} = \mathbf{y}^t\mathbf{b}. \tag{2.17}$$

This is a single equation that is the sum of y_1 times the first equation, y_2 times the second equation, ..., up to y_m times the last equation. It can also be written

$$y_1\mathbf{a}^1 + \cdots + y_m\mathbf{a}^m = y_1b_1 + \cdots + y_mb_m.$$

So it is a linear combinations of the rows of the system of equations, if the system has a solution, the equation (2.17) must have a solution. In fact we have:

Theorem 2.7.2. *The equation* $A\mathbf{x} = \mathbf{b}$ *has a solution* $\mathbf{x}$ *if and only if there is no vector* $\mathbf{y}$ *in* F^m *with*

$$\mathbf{y}^t A = \mathbf{0} \qquad \textit{and} \qquad \mathbf{y}^t\mathbf{b} \neq 0.$$

Proof. Assume there is a solution $\mathbf{x}$. Then just multiply the system $A\mathbf{x} = \mathbf{b}$ on the left by the row vector $\mathbf{y}^t$. We get

$$\mathbf{y}^t(A\mathbf{x}) = \mathbf{y}^t\mathbf{b}.$$

Since this is a linear combinations of the rows of the system, it must have a solution in $\mathbf{x}$. By the associativity of matrix multiplication, the left hand side can be written $(\mathbf{y}^tA)\mathbf{x}$. Thus for any $\mathbf{y}$ such that $\mathbf{y}^tA = \mathbf{0}$, we must have $\mathbf{y}^t\mathbf{b} = 0$, since otherwise we get the absurd equation $0 = \mathbf{y}^t\mathbf{b} \neq 0$.

To prove the converse, we must use row reduction. Rather than doing it again, refer to Corollary 1.3.5. It is a good exercise to rewrite the proof in matrix notation. □

This proves the result because Theorem 2.5.11 tells us that we only need to consider systems $A\mathbf{x} = \mathbf{b}$ where the matrix A is in reduced row echelon form, since the original system will have the same solutions as the one where A is in row reduced echelon form, and the corresponding row operations have been made on $\mathbf{b}$. Let us make that assumption, and see what we can deduce.

First we assume our original system is homogeneous, so $\mathbf{b} = \mathbf{0}$ after any row operation. We may throw out all the bottom rows of the system of the form $\mathbf{0} = \mathbf{0}$. The number of equations, which we still call m, may therefore be smaller than the number of equations we started out with.

Since A is in RREF form, in the i-th equation, the variable with non-zero coefficient of smallest index is x_{j_i}, where the indices j_i are strictly increasing in i, and thus all distinct. These variables are called the pivot variables or the bound variables, and they give the bound columns of Definition 2.6.1. These are the same pivot variables as defined in Chapter 1. The remaining $n-m$ variables are called the free variables, and they give the free columns. We can use each one of the m equations to write each x_{j_i} in terms of the free variables, which can take on arbitrary values. Thus we can solve the homogeneous system. If $m = n$ the system has a unique solution: the trivial solution. If $m < n$, then the $n-m$ free variables can take on arbitrary values, so the solution is not unique. We will not pursue this here: see Proposition 5.7.2.

Examples of homogeneous systems here. Mention that there many be several ways of choosing the bound variables.

Now we assume the system is inhomogeneous. Theorem 2.7.2 gives a criterion for the existence of a solution. So assume that the system has at least one solution.

Row reduction may create equations where both the left hand side and the right hand side are 0. Those we can just discard. Thus the number of rows in the system may decrease after row reduction.

Prove in matrix notation the very first theorem of the previous chapter:

Exercise 2.7.3. If $\mathbf{x}_1$ and $\mathbf{x}_2$ are distinct solutions of an inhomogeneous system, then $\mathbf{x}_1 - \mathbf{x}_2$ is a solution to the corresponding homogeneous equation.

Examples of inhomogeneous systems here. How to do this in practice.

2.8 Elementary Matrices

It is a remarkable fact that the three types of elementary row operations of Definition 2.5.2 can be achieved by left multiplication of the matrix A by suitable square matrices, which we call *elementary matrices*.

We introduce the three types of *elementary matrices*, and show they are invertible. We first define the matrix I_{rs} to be the square matrix with a $i_{rs} = 1$, and zeroes everywhere else. So for example in the 3×3 case

$$I_{23} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

Now we define the elementary matrices.

Definition 2.8.1. Elementary matrices E are square matrices, say $m \times m$. There are three types of elementary matrices.

1. $$E_r(c) := I + (c-1)I_{rr}.$$

   $E_r(c)$ is diagonal with a 1 for all diagonal elements except the r-th, where it has c.

2.
$$T_{rs}, r \neq s,$$
is the matrix that has

- a 1 in all the diagonal entries except those with index (r,r) and index (s,s);
- a 1 in the two entries with indices (r,s) and (s,r);
- a 0 in all other entries.

3. The matrix
$$E_{rs}(c) := I + cI_{rs}, r \neq s.$$

Proposition 2.8.2. *Here is how the elementary matrices E transform the matrix A by left multiplication: EA.*

1.
$$E_r(c) := I + (c-1)I_{rr}$$
multiplies the r-th row of A by c.
2. T_{rs} interchanges row r of A with row s and leaves the rest of A unchanged.
3. $E_{rs}(c) := I + cI_{rs}, r \neq s$, adds c times the s-th row of A to the r-th row of A.

Proof. This is a simple exercise in matrix multiplication, left to you.

Example 2.8.3. Here are some 3×3 examples of elementary matrices, acting on the 3×4 matrix
$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix}.$$

1. Since
$$E_1(c) = \begin{pmatrix} c & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$
matrix multiplication gives
$$E_1(c)A = \begin{pmatrix} ca_{11} & ca_{12} & ca_{13} & ca_{14} \\ a_{21} & a_{22} & a_{23}a_{24} & \\ a_{31} & a_{23} & a_{33}a_{34} & \end{pmatrix}.$$

2. Since
$$T_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$
we get
$$T_{23}A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{pmatrix}.$$

3. Since

$$E_{13}(c) = \begin{pmatrix} 1 & 0 & c \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

we get

$$E_{13}(c)A = \begin{pmatrix} a_{11}+ca_{31} & a_{12}+ca_{32} & a_{13}+ca_{33} & a_{14}+ca_{34} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix}.$$

Here is an essential feature of elementary matrices.

Theorem 2.8.4. *All elementary matrices are invertible.*

Proof. The proof is an simple computation in each case: For each type of elementary matrix E we write down an inverse, namely a matrix F such that $EF = I = FE$.

For $E_i(c)$ the inverse is $E_i(1/c)$. T_{ij} is its own inverse. Finally the inverse of $E_{ij}(c)$ is $E_{ij}(-c)$. □

Remark 2.8.5. The matrices $E_i(c)$ and T_{ij} are symmetric, while the transpose of $E_{ij}(c)$ is $E_{ji}(c)$.

We stated when two $m \times n$ matrices are row equivalent in 2.5.2. Now we prove

Theorem 2.8.6. *Two $m \times n$ matrices A and C are row equivalent if there is a product of elementary matrices E such that $C = EA$. Then the system of linear equations with augmented matrix $\big(A \mid \mathbf{b}\big)$ is transformed to $\big(EA \mid E\mathbf{b}\big)$.*

Proof. This is clear, since any row operation can be achieved by multiplication by an elementary matrix. □

Here is an important generalization. Again assume the coefficient matrix of a linear system is the $m \times n$ matrix A. Instead of taking an augmented matrix with only one column $\mathbf{b}$, take one with l columns. Call it B. Thus B is a $m \times l$ matrix. Because we only do row operations, the different columns do not interact with each other, so we can do row operations on each of the systems where the coefficient matrix is A, and the right-hand side is $\mathbf{b}_i$, $1 \leq i \leq l$, the columns of B. We write this augmented matrix as $\big(A \mid B\big)$. This allows us to solve l systems with the same coefficient matrix simultaneously.

Corollary 2.8.7. *If left multiplication by product of elementary matrices E puts A is row-echelon form then the augmented matrix is $\big(EA \mid EB\big)$. The same holds if E puts A in RREF form.*

This is a convenient way of solving several systems of linear equations with the same left-hand side simultaneously.

From now on we consider the important case where the coefficient matrix A is square.

Proposition 2.8.8. *If A is a square matrix, and if B is row equivalent to A, then A has an inverse if and only if B has an inverse.*

Proof. Any product of elementary matrices E is invertible by Theorem 2.8.4. If A is invertible, then by Proposition 2.3.7 $B = EA$ is invertible, with inverse $A^{-1}E^{-1}$. By the same argument, if B is invertible, so is $A = E^{-1}B$. □

Remark 2.8.9. So being invertible or not is a property of the equivalence class. An interesting question is: how many equivalence classes contain invertible matrices. This can be easily established by looking at reduced row–echelon form.

Exercise 2.8.10. Using RREF form, list all the equivalence classes of $n \times n$ matrices.

Theorem 2.8.11. *The square matrix A is either row–equivalent to the identity matrix I, and therefore invertible, or it is row–equivalent to a matrix with bottom row the zero vector, and not invertible.*

Proof. By Theorem 2.5.9 A is row–equivalent to a A' in row–echelon form. By Remark 2.5.7, A' is upper-triangular, so the only possible non-zero entry in the bottom row is a'_{nn}. If $a'_{nn} \neq 0$, then since A' is row reduced, all the previous diagonal elements are non-zero.

Under this hypothesis, if we put A in reduced row-echelon form, we may make all the diagonal elements 1, all the terms above the diagonal 0, so we get the identity matrix. Therefore A is row equivalent to I.

If $a'_{nn} = 0$, we are in the second case. □

Proposition 2.8.12. *Let A be a square matrix with one row equal to the zero vector. Then A is not invertible.*

Proof. Assume the i-th row of A is the zero vector. Multiply the matrix A by the column vector $\mathbf{e}_i$, which has a 0 in all entries except the i-th where is has a 1. Then matrix multiplication shows that $A\mathbf{v} = \mathbf{0}$. Assume A has an inverse B. Then by the associativity of matrix multiplication $\mathbf{v} = (BA)\mathbf{v} = B(A\mathbf{v}) = B\mathbf{0} = \mathbf{0}$, a contradiction. □

Finally, we get to a key result of this section: we only need to assume that A has an inverse on one side, for it to have an inverse.

Theorem 2.8.13. *Let A be a square matrix which has a right inverse B, meaning that $AB = I$. Then A is invertible and B is its inverse.*

Similarly, if A has a left inverse B, meaning that $BA = I$, the same conclusion holds.

Proof. Suppose first that $AB = I$. Perform row reduction on A. By Theorem 2.8.11, there are elementary matrices $E_1, E_2, \ldots, E_k$ so that the matrix $C = E_k \ldots E_1 A$ is in reduced row–echelon form. Write $E = E_k \ldots E_1$.

Then multiply by B on the right and use associativity:

$$CB = (EA)B = (E)(AB) = E.$$

This is invertible, because elementary matrices are invertible. Therefore all the rows of CB are non-zero by Proposition 2.8.12. Now if the i-th row of C were $\mathbf{0}$, then matrix multiplication shows that the i-th row of CB is $\mathbf{0}$. This is impossible since CB is invertible. So C is invertible. Since A is row-equivalent to C, it is invertible.

To do the direction $BA = I$, just interchange the role of A and B to find that B is invertible. But then $A = B^{-1}$ is invertible. □

Now we can establish a good method for computing the inverse of a square matrix. This is a special case of Corollary 2.8.7.

Theorem 2.8.14. *Let A be an invertible $n \times n$ matrix and I the identity matrix of size n. Then the product E of elementary matrices that reduce A to the identity matrix is the inverse of A.*

Proof. Since $EA = I, A^{-1} = E$. □

Here is how one usually sets up the computation. Do row-reduction on the augmented matrix $\big(A \mid I\big)$ until you have reached the identity matrix by row-reduction. Since A is invertible, by Theorem 2.8.11, A row-reduces to the identity, so the row reduction by E on $\big(A \mid I\big)$ gives $\big(EA \mid E\big)$. Therefore the inverse of A appears on the right-hand side of the augmented matrix when the left-hand side reaches int identity matrix.

Examples of this process here.

Exercise 2.8.15. Let A be a square matrix.

1. If $A^2 = 0$, then $I - A$ is invertible, with inverse $I + A$;
2. More generally, if $A^n = 0$ for some positive integer n, then $I - A$ is invertible;
3. More generally, if A satisfies an equation

$$c_n A^n + c_{n-1} A^{n-1} + \cdots + c_1 A = I$$

   where the c_i are scalars, then A is invertible. Hint: just factor the left hand side.
4. What is the inverse of $I + A$, where A is the matrix of Example 2.3.12?

Exercise 2.8.16. Find all 2×2 matrices such that $A^2 = 0$.

Exercise 2.8.17. Let

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

1. determine A^2, A^3, …, A^n using the trigonometric addition formulas and induction.
2. Let A act on $(x, y) \in \mathbb{R}^2$ by matrix multiplication:

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \end{pmatrix}$$

   What happens to $\mathbb{R}^2$?

3. Does this explain your result in the first part of this exercise?

Exercise 2.8.18. The trace $\mathrm{tr}A$ of a square matrix A of size n is defined as the sum or the diagonal terms of A:

$$\mathrm{tr}A = a_{11} + a_{22} + \cdots + a_{nn}.$$

1. Show $\mathrm{tr}A^t = \mathrm{tr}A$.
2. Show $\mathrm{tr}\overline{A} = \overline{\mathrm{tr}A}$ and $\mathrm{tr}A^* = \overline{\mathrm{tr}A}$.
3. If B is a second square matrix of the same size, show $\mathrm{tr}(A+B) = \mathrm{tr}A + \mathrm{tr}B$.
4. Prove that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.
5. If C is a third square matrix of the same size, show that $\mathrm{tr}(ABC) = \mathrm{tr}(CAB) = \mathrm{tr}(BCA)$. Give an example where $\mathrm{tr}(ABC) \neq \mathrm{tr}(ACB)$.
6. If B is invertible, then $\mathrm{tr}(B^{-1}AB) = \mathrm{tr}A$.

Remark 2.8.19. The first five chapters of Artin's book [1] form a nice introduction to linear algebra at a slightly higher level than here, with some group theory thrown in too. The main difference is that Artin allows his base field to be any field, including a finite field, while we only allow $\mathbb{R}$ and $\mathbb{C}$.

Exercise 2.8.20. Reduce the matrices in Example 2.5.6 either to a matrix with bottom row zero or to the identity matrix using left multiplication by elementary matrices.

For example, the first matrix

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

backsubstitutes to

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \text{ then } \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \text{ then } \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

2.9 Block Decomposition of Matrices

A good reference for this easy material is the classic [9]. As the authors say, facility with block matrix notation is crucial for matrix computation, which is why we study it here.

It is often convenient to think of a matrix as being made up of a grid of smaller submatrices. Here is the general procedure. There is nothing difficult except for the notation.

Definition 2.9.1. Let A be a $m \times n$ matrix. Write m as the sum of positive numbers $m_1, \ldots, m_s$ and n as the sum of positive integers $n_1, \ldots, n_t$.

Then we can write

$$A = \begin{pmatrix} A^{11} & A^{12} & \dots & A^{1t} \\ A^{21} & A^{22} & \dots & A^{2t} \\ \vdots & \vdots & \ddots & \vdots \\ A^{s1} & A^{s2} & \dots & A^{st} \end{pmatrix}$$

where A^{ij} is the $m_i \times n_j$ submatrix of A in the appropriate position. So there are st submatrices. By definition all blocks in a given column share the same columns of A, while all blocks in a given row share the same rows of A.

This is known as partitioning, or decomposing, the matrix into blocks.

Example 2.9.2. The 3×4 matrix

$$M = \begin{pmatrix} a_{11} & a_{12} & b_{13} & b_{14} \\ a_{21} & a_{22} & b_{23} & b_{24} \\ c_{31} & c_{32} & d_{33} & d_{34} \end{pmatrix}$$

can be partitioned into the blocks

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

where A and B are 2×2 matrices, and C and D are 1×2 matrices. So in this example $s = t = 2$, and $m_1 = n_1 = n_2 = 2$ while $m_2 = 1$.

Matrix multiplication behaves nicely with respect to block decomposition. So if some of the blocks are repeated or are simple (for example the identity matrix or the zero matrix) block multiplication can speed up the computation of the matrix product. Here is the main theorem of this section.

Theorem 2.9.3. *Let A be a $m \times n$ matrix block decomposed according to Definition 2.9.1. Let B be a $n \times p$ matrix block decomposed along its rows exactly as A is along its columns, and where $p = p_1 + \cdots + p_u$ is the block decomposition of its columns, so*

$$B = \begin{pmatrix} B^{11} & B^{12} & \dots & B^{1u} \\ B^{21} & B^{22} & \dots & B^{2u} \\ \vdots & \vdots & \ddots & \vdots \\ B^{t1} & B^{s2} & \dots & B^{tu} \end{pmatrix}.$$

Thus B^{jk} is a $n_j \times p_k$ submatrix of B. Then $AB = C$, where the $m \times p$ matrix C can be blocked decomposed as

$$C = \begin{pmatrix} C^{11} & C^{12} & \dots C^{1u} \\ C^{21} & C^{22} & \dots C^{2t} \\ \vdots & \vdots & \ddots & \vdots \\ C^{s1} & C^{s2} & \dots C^{su} \end{pmatrix}$$

where C^{ik} is a $m_i \times p_j$ matrix such that

$$C^{ik} = A^{i1}B^{1k} + A^{i2}B^{2k} + \cdots + A^{it}B^{tk} = \sum_{j=1}^{t} A^{ij}B^{jk} \tag{2.18}$$

Note that (2.18) is Definition 2.2.3 with blocks replacing numbers.

Proof. The details of the proof are left to the reader. First notice that the matrices on the right hand side of (2.18) are of the appropriate size to be multiplied and added.. Finally just check that for each entry of the matrix C^{ik} you have all the terms of the appropriate entry of C: all that is needed is Definition 2.2.3. □

An important special case occurs when the matrices A and B are square, meaning that $m = n = p$, and when the diagonal blocks are also square, implying that $s = t$, and $m_i = n_i$, $1 \leq i \leq n$. In this case, A^{ii} is an $n_i \times n_i$ matrix.

Example 2.9.4. Let A be an $m \times n$ matrix and let B be an $n \times p$ matrix. Let C be the product matrix AB of size $m \times p$. We block decompose A with

$$\begin{aligned} m &= m_1 + m_2; \\ n &= n, \end{aligned}$$

so there is no decomposition into columns. We block decompose B with

$$\begin{aligned} n &= n, \\ p &= p_1 + p_2, \end{aligned}$$

so there is no decomposition into rows. So

$$A = \begin{pmatrix} A^{11} \\ A^{21} \end{pmatrix}, \quad B = \begin{pmatrix} B^{11} & B^{12} \end{pmatrix} \tag{2.19}$$

Then C can be partitioned according to the partition of the rows of A and the columns of B so that

$$C = \begin{pmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{pmatrix} \tag{2.20}$$

with $C^{ij} = A^{i1}B^{1j}$.

Example 2.9.5. If A and B are decomposed in the other direction, with the common index n written as $n_1 + n_2$ for both matrices, and no decomposition of the other indices m and p, then we can write the matrix product as

$$\begin{pmatrix} A^{11} & A^{12} \end{pmatrix} \begin{pmatrix} B^{11} \\ B^{21} \end{pmatrix} = A^{11}B^{11} + A^{12}B^{21}$$

You should check that the matrix multiplications and the matrix addition on the right hand side are well defined.

Exercise 2.9.6. Let

$$A = \begin{pmatrix} 1 & -2 \\ -3 & 2 \\ -1 & 3 \end{pmatrix} \text{ and } B = \begin{pmatrix} 1 & -2 & 1 \\ -3 & 2 & 0 \end{pmatrix}$$

Break A into two blocks

$$A^{11} = \begin{pmatrix} 1 & -2 \\ -3 & 2 \end{pmatrix}, A^{21} = \begin{pmatrix} -1 & 3 \end{pmatrix}$$

Now break B into two blocks so that the decomposition of the column size ($3 = 2+1$) of A agrees with that of the row size ($3 = 2+1$) of B.

$$B^{11} = \begin{pmatrix} 1 & -2 \\ -3 & 2 \end{pmatrix}, B^{12} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

This allows block multiplication. Check that the formula of Example 2.9.4 applies by computing the matrix product two ways.

Definition 2.9.7. Assume that the matrix A is square of size n and that its diagonal blocks A^{ii} are square of sizes $n_1, n_2, \ldots, n_s$ with $n = n_1 + n_2 + \cdots + n_s$.

- Then A is *block diagonal* if A^{ij}, $i \neq j$, is the zero matrix:

$$A = \begin{pmatrix} A^{11} & 0 & \ldots & 0 \\ 0 & A^{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & A^{ss} \end{pmatrix} \tag{2.21}$$

- A is *block upper triangular* if A^{ij}, $i > j$, is the zero matrix:

$$A = \begin{pmatrix} A^{11} & A^{12} & \ldots & A^{1s} \\ 0 & A^{22} & \ldots & A^{2s} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & A^{ss} \end{pmatrix} \tag{2.22}$$

In the same way we can define block lower triangular.

Proposition 2.9.8. *Assume A and B are square matrices of size n, and and that blocks are of size $n_1, n_2, \ldots, n_s$ with $n = n_1 + n_2 + \cdots + n_s$.*

- *If they are both block diagonal, their product $C = AB$ is also block diagonal, with $C^{ii} = A^{ii}B^{ii}$. Furthermore*

$$A^k = \begin{pmatrix} (A^{11})^k & 0 & \ldots & 0 \\ 0 & (A^{22})^k & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & (A^{ss})^k \end{pmatrix} \tag{2.23}$$

- *If A and B are both block upper triangular, then so it their product.*

Proof. We prove Proposition 2.9.8 using the main theorem. The diagonal case is trivial, so let's just consider the upper triangular case. If $C = AB$ we must show that $C_{ik} = 0$ when $i > k$. By hypothesis $A^{it} = 0$ when $i > t$ and $B^{tk} = 0$ when $t > k$. By (2.18) this means that the only non-zero terms in the sum are those with $i \leq t \leq k$. Since $i > k$, there are no such terms. □

Example 2.9.9. A special case that will be important to us in the one where A and B are both square of size $n = r + s$ and decomposed as

$$\begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix} \text{ and } \begin{pmatrix} B^{11} & B^{12} \\ B^{21} & B^{22} \end{pmatrix}.$$

where A^{11} and B^{11} are $r \times r$ matrices,
A^{12} and B^{12} are $r \times s$ matrices,
A^{21} and B^{21} are $s \times r$ matrices,
A^{22} and B^{22} are $s \times s$ matrices. Then

$$AB = \begin{pmatrix} A^{11}B^{11} + A^{12}B^{21} & A^{11}B^{12} + A^{12}B^{22} \\ A^{21}B^{11} + A^{22}B^{21} & A^{21}B^{12} + A^{22}B^{22} \end{pmatrix}$$

If A and B are both block upper triangular, meaning that A^{21} and B^{21} are both the zero matrix, then their product AB is also block upper triangular. It is easier to check in this special case that the formula is correct.

Exercise 2.9.10. Let A be a 4×2 matrix and B be a 2×4 matrix, written in block form as in (2.19), where all the blocks are 2×2. Further assume that

$$A^{11} = A^{21} = \begin{pmatrix} 1 & 1 \\ -1 & -2 \end{pmatrix}, \text{ and } B^{11} = B^{12} = \begin{pmatrix} 2 & 1 \\ -1 & -1 \end{pmatrix};$$

Write out the matrices A and B, compute the product AB directly, and then compute it by block multiplication.

Exercise 2.9.11. If you have the block decomposition of a matrix A, write a decomposition for its transpose A^T.

2.10 Column Operations

This short section will only be used in § 7.5 and can be skipped. There is nothing special about row operations. We can also perform column operations. Because $(EA)^t = A^t E^t$, the matrix E^t , when multiplying A^t on the right, performs column operations on A^t. For simplicity we only write the result when A is symmetric. This is the only case we will need later. Then by analogy with Proposition 2.8.2 we have

Proposition 2.10.1. *Here is how the elementary matrices E transform the matrix A by right multiplication: AE.*

1. *$E_r(c)$ multiplies the r-th column of A by c.*
2. *T_{rs} interchanges columns r of A with column s and leaves the rest of A unchanged.*
3. *$E_{rs}(c), r \neq s$, adds c times the r-th column of A to the s-th column of A.*

Proof. This is also a simple exercise in matrix multiplication, left to you. Note the reversal of the roles of r and s in the third item. This is because the transpose of $E_{rs}(c)$ is $E_{sr}(c)$.

Chapter 3
Vector Spaces

Abstract The reader is presumably familiar with the definition of a vector in a space of dimension n: an ordered n-tuple of numbers. These n-tuples can be added, simply by adding corresponding entries, and can be multiplied by a number by multiplying each entry of the vector by that number: this last operation is called scalar multiplication. In this chapter we formalize these notions. First we look carefully at what properties of numbers we need to make the operations, say, of the last chapter. Because we allow different kinds of numbers (in this book only real and complex numbers) we refer to the numbers as scalars. Then we make the first fundamental definition of the book: that of a vector space. We simply axiomatize the rules of vector addition and scalar multiplication given above. Then we define a basis of a vector space, and define a finite dimensional vector space as a vector space that admits a basis with a finite number of elements. One of the most important and subtle theorems of linear algebra then tells us that all bases of a given vector space have the same number of elements. That number is called the dimension of the space. An easy corollary, using a basis, that a finite dimension vector space is in bijection with the set of ordered n tuples of scalars, so we recover the definition you know, but at the cost of choosing a basis. The chapter continues with one important method of deriving new vector spaces from old ones: direct sums.

3.1 Scalars

The coefficients of linear equations, and the entries of matrices are called *scalars*. In this book they will either be real numbers or complex numbers. It is worth writing down the properties of the scalars that we will use: they form a *field*, that we denote F. This means that they satisfy the following properties.

- Any two elements a and b of F can be added: $a+b \in F$ and multiplied: $ab \in F$. Both operations are commutative, meaning that $a+b = b+a$ and $ab = ba$. They are also associative: $(a+b)+c = a+(b+c)$ and $(ab)c = a(bc)$.

- The element 0 is the neutral element for addition, so that for every $a \in F$, $a+0=0$. Every $a \in F$ has an additive inverse $-a$, so that $a+(-a)=0$.
- The element $1 \neq 0$ is the neutral element for multiplication: for every $a \in F$, $1a = a$. Every element $a \in F$ other than 0 has a multiplicative inverse a^{-1} such that $aa^{-1} = 1$.

To establish these properties for complex numbers consult Section B.5.

Each one of our two fields of choice, the real numbers and the complex numbers, have an important property that we now recall.

The real numbers are *ordered*. This means that any two real numbers a and b can be compared: we either have $a < b$, $a = b$ or $a > b$. The complex numbers do not share this property.

The complex numbers are *algebraically closed*: every polynomial $f(t)$ with complex coefficients has a complex root, meaning that there is a complex number c such that $f(c) = 0$.

3.2 Vector Spaces

Now we can make the first of the two key definitions of this course: that of a *vector space* V over the field F. The elements of a vector space are called vectors, naturally.

Definition 3.2.1. A vector space is a set equipped with two operations:

1. Scalar multiplication, which associates to a scalar $a \in F$ and a $\mathbf{v} \in V$ an element written $a\mathbf{v} \in V$.
2. Addition, which associates to elements $\mathbf{v}$ and $\mathbf{w}$ in V the element $\mathbf{v}+\mathbf{w}$ in V.

These operations satisfy the following eight properties:

VS 1 Addition is associative, meaning that for any $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ in V,

$$(\mathbf{u}+\mathbf{v})+\mathbf{w} = \mathbf{u}+(\mathbf{v}+\mathbf{w}).$$

VS 2 There is a neutral element for addition, denoted $\mathbf{0}$, so that

$$\mathbf{0}+\mathbf{v} = \mathbf{v}+\mathbf{0} = \mathbf{v}.$$

VS 3 There is an additive inverse for any element $\mathbf{v} \in V$, written $-\mathbf{v}$, satisfying

$$\mathbf{v}+(-\mathbf{v}) = \mathbf{0}.$$

VS 4 Addition is commutative, so that for all $\mathbf{u}$ and $\mathbf{v}$ in V,

$$\mathbf{u}+\mathbf{v} = \mathbf{v}+\mathbf{u}.$$

VS 5 Scalar multiplication distributes over vector addition: if a is a scalar, then

$$a(\mathbf{u}+\mathbf{v}) = a\mathbf{u}+a\mathbf{v}.$$

VS 6 Scalar multiplication distributes over scalar addition: if a and b are scalars, then

$$(a+b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}.$$

VS 7 Multiplication of scalars and scalar multiplication are associative: If a and b are two scalars, then

$$(ab)\mathbf{v} = a(b\mathbf{v}).$$

VS 8 Normalization: if 1 denotes as usual the multiplicative neutral element of F,

$$1\mathbf{v} = \mathbf{v}.$$

The first four properties only concern vector addition in V: as we will learn later, the first three say that V is a group for addition, and the fourth that this group is commutative. The remaining four properties describe the interaction between vector addition and scalar multiplication.

Example 3.2.2. We start with a trivial example. Over every field, there is a vector space consisting just of the element $\mathbf{0}$. We could write it $\mathbb{R}^0$ or $\mathbb{C}^0$, and we call it the trivial vector space.

Example 3.2.3. The most important examples are the real vector space $\mathbb{R}^n$ of ordered n-tuples of real numbers, and the complex vector space $\mathbb{C}^n$ of ordered n-tuples of complex numbers. Here n is any positive integer. We write a vector in each one of these spaces as $\mathbf{v} = (v_1, v_2, \ldots, v_n)$, where each v_1 is a scalar. Scalar multiplication of a vector $\mathbf{v}$ with the scalar a is

$$a\mathbf{v} = (av_1, av_2, \ldots, av_n)$$

while addition of vectors is

$$\mathbf{v} + \mathbf{w} = (v_1 + w_1, v_2 + w_2, \ldots, v_n + w_n).$$

The neutral element for addition is clearly

$$\mathbf{0} = (0, 0, \ldots, 0).$$

In particular $\mathbb{R}^1 = \mathbb{R}$ is a vector space over the reals, and $\mathbb{C}^1 = \mathbb{C}$ is a vector space over the complex numbers.

Example 3.2.4. A special case of this example is the space of all $m \times n$ matrices. Similarly the set of all symmetric matrices is a vector space.

Example 3.2.5. The set of solutions of a homogeneous system of equations $A\mathbf{x} = \mathbf{0}$ is a vector space.

Example 3.2.6. The space of polynomials $F[t]$ is a vector space over F. By the definitions of §C.1 addition is just addition of coefficients of the same degree: if

$$g(t) = b_n t^n + b_{n-1} t^{n-1} + \cdots + b_1 t + b_0$$

is a second polynomial, then (assuming $n > m$)

$$f(t) + g(t) = b_n t^n + \ldots (a_m + b_m)t^m + \cdots + (a_0 + b_0).$$

Scalar multiplication is

$$cf(t) = ca_m t^m + ca_{m-1}t^{m-1} + \cdots + ca_1 t + ca_0$$

The next example is more abstract, but quite important.

Example 3.2.7. Let V be the set of all maps on any set S into a vector space W. Then we may add two such maps f and g in V by defining the sum map $f+g$ by

$$(f+g)(s) = f(s) + g(s), \text{ for all } s \in S.$$

In this equation the $+$ on the left hand side is the definition of addition in V in terms of the right hand side, where the $+$ is simply addition in W.

Similarly we define a scalar multiplication by letting af, $a \in F$ be the map whose value at $s \in S$ is

$$(af)(s) = a\big(f(s)\big).$$

The multiplication af on the left hand side is the definition of scalar multiplication in V in terms of the right hand side, where the multiplication is scalar multiplication in W of a by $f(s)$. The neutral element $\mathbf{0}$ of V is the map that takes every $s \in S$ to the neutral element of W. We then need to check the remaining properties of a vector space: for example the inverse of the map f is the map g such that $g(s) = -f(s)$ This makes V into a vector space, since all the other properties follow from the fact they are true in W.

Example 3.2.8. The complex numbers form a vector space over the real numbers. Indeed, any complex number can be written $a + b\mathrm{i}$, where a and b are real: for more details see Chapter 1. The two operations that give the complex numbers the structure of a real vector space are

- scalar multiplication is multiplication of a complex number by a real number c:

$$c(a + b\mathrm{i}) = ca + cb\mathrm{i}.$$

- vector addition is just the addition of two complex numbers:

$$(a + b\mathrm{i}) + (c + d\mathrm{i}) = (a + c) + (b + d)\mathrm{i}.$$

These operations are exactly those of $\mathbb{R}^2$. Note that we do not need the full strength of the multiplication of two complex numbers.

On the other hand, $\mathbb{R}$ is not a vector space over $\mathbb{C}$.

Next we prove the important cancelation rule for vector addition. Note that instead of writing $\mathbf{u} + (-\mathbf{v})$ we write $\mathbf{u} - \mathbf{v}$. Thus $\mathbf{v} - \mathbf{v} = \mathbf{0}$.

Theorem 3.2.9. *If* $\mathbf{u}$, $\mathbf{v}$ *and* $\mathbf{w}$ *are elements of a vector space* V*, and if*

$$\mathbf{u}+\mathbf{v}=\mathbf{w}+\mathbf{v},$$

then $\mathbf{u}=\mathbf{w}$.

Proof. We give all the details of the proof: add to each side of the equation the additive inverse $-\mathbf{v}$ of $\mathbf{v}$, which exists by VS 3. Then we have

$$(\mathbf{u}+\mathbf{v})-\mathbf{v}=(\mathbf{w}+\mathbf{v})-\mathbf{v}.$$

Next we use associativity (VS 1) to get

$$\mathbf{u}+(\mathbf{v}-\mathbf{v})=\mathbf{w}+(\mathbf{v}-\mathbf{v}).$$

Finally we use VS 2 to get the conclusion. □

Corollary 3.2.10. *The additive identity is unique. The additive inverse is unique.*

Proof. Assume there were two additive identities $\mathbf{0}$ and $\mathbf{0}'$. But then for any $\mathbf{v}\in V$,

$$\mathbf{v}+\mathbf{0}=\mathbf{v}=\mathbf{v}+\mathbf{0}'$$

so after cancelation $\mathbf{0}=\mathbf{0}'$. The same proof works for the additive inverse. □

Proposition 3.2.11. *We now deduce some additional properties of the operations of vector spaces.*

- *For all* $\mathbf{v}\in V$, $0\mathbf{v}=\mathbf{0}$. *Indeed,* $0\mathbf{v}+1\mathbf{v}=\mathbf{v}$ *by VS 6, and* $\mathbf{v}=\mathbf{0}+\mathbf{v}$ *by VS 2. Now finish by using VS 8 and then cancelation.*
- $(-1)\mathbf{v}=-\mathbf{v}$. *In other words, multiplying a vector by the number* -1 *gives its additive inverse. Thus we must show*

  $$\mathbf{v}+(-1)\mathbf{v}=\mathbf{0}.$$

  Since $\mathbf{v}=(1)\mathbf{v}$ *by VS 8,* $\mathbf{v}+(-1)\mathbf{v}=(1-1)\mathbf{v}$ *by VS 6 and so we are done by what we just proved.*

In the exercises below, V is a F-vector space.

Exercise 3.2.12. Show that $a\mathbf{0}=\mathbf{0}$ for all $a\in F$.

Exercise 3.2.13. If $a\neq 0$, then if $a\mathbf{v}=\mathbf{0}$, $\mathbf{v}=\mathbf{0}$.

3.3 Subspaces

A subspace W of a vector space V is a subset that is a vector space in its own right, using the operations of V. To check that W is a subspace, we must show that it is *closed* under the operations of V. In other words,

Definition 3.3.1. A subset W of the vector space V is a subspace of W if

1. For all $\mathbf{v}$ and $\mathbf{w}$ in W, $\mathbf{v}+\mathbf{w}$ is in W;
2. For all scalars $a \in F$ and all $\mathbf{w} \in W$, then $a\mathbf{w}$ is in W.

This implies that $\mathbf{0}$ is in W, since $\mathbf{0} = 0\mathbf{w}$, for any $\mathbf{w} \in W$.

Note that the trivial vector space (Example 3.2.2) is a subspace of any vector space. The space V is a subspace of itself. We call both of the subspaces the trivial subspaces of V.

Example 3.3.2. Check that the following subsets are actually subspaces.

- The subset of all triples in $\mathbb{R}^3$ where the last entry is 0: $(v_1, v_2, 0)$.
- The subset of all n-tuples in $\mathbb{R}^n$ where the last entry is 0: $(v_1, \ldots, v_{n-1}, 0)$.

Example 3.3.3. In the vector space of polynomials in t over F, consider the subset P_k of polynomials of degree at most k, for any integer k. Show P_k is a subspace of the vector space of polynomials over F. Explain why the polynomials of degree exactly n do not form a subspace.

Example 3.3.4. As before let V be the set of functions on a set S. Consider the subset V_s of functions in V that vanish at a fixed point $s \in S$. Show V_s is a subspace.

The key example for us is

Example 3.3.5. If A is a $m \times n$ matrix with coefficients in F, then the set of solutions of the homogeneous system of equations $A\mathbf{x} = \mathbf{0}$ is a subspace N_A of F^n.

Show that if the system is not homogenous, then the set of solutions is not a subspace.

See Theorem 4.2.2 for a restatement of this result: N_A will be called the nullspace of the linear map associated to A.

Definition 3.3.6. If V is a vector space, and $\mathbf{v}_1, \ldots, \mathbf{v}_r$ a collection of r elements of V, then any expression

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_r\mathbf{v}_r, \text{ where all } a_i \in F,$$

is called a *linear combination* of $\mathbf{v}_1, \ldots, \mathbf{v}_r$.

Proposition 3.3.7. *Given a collection $\mathbf{v}_1, \ldots, \mathbf{v}_r$ of elements in V, the set of all linear combinations of these vectors is a subspace of V, called the subspace* generated *by the elements $\mathbf{v}_1, \ldots, \mathbf{v}_r$.*

Proof. The only difficulty is understanding what needs to be proved. Let W be the space of all linear combinations. Thus if $\mathbf{v}$ and $\mathbf{w}$ are in W, we have

$$\mathbf{v} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_r\mathbf{v}_r$$
$$\mathbf{w} = b_1\mathbf{v}_1 + b_2\mathbf{v}_2 + \cdots + b_r\mathbf{v}_r$$

so that

$$\mathbf{v}+\mathbf{w}=(a_1+b_1)\mathbf{v}_1+(a_2+b_2)\mathbf{v}_2+\cdots+(a_r+b_r)\mathbf{v}_r$$

which is a linear combination, so in W. The other property is even easier to prove. □

Two important ways of producing new subspaces are by intersection and by sums of subspaces. If U and W are subspaces of a vector space V we let $U\cap W$ be their intersection. This is a purely set-theoretic construction. On the other hand, let

$$U+W=\{u+w \mid \forall u\in U, \forall w\in W\}$$

be their sum. This depends on having addition in V.

Proposition 3.3.8. *If U and W are both subspaces of the vector space V, then $U\cap W$ is a subspace of V.*

Proof. This is elementary set theory. If $\mathbf{u}$ is in $U\cap W$, then $\mathbf{u}$ is both in U and in W. Since U is a subspace, $c\mathbf{u}$ is in U for every scalar c; since V is a subspace, $c\mathbf{u}$ is in W for every scalar c. So $c\mathbf{u}$ is in $U\cap W$.

If $\mathbf{u}$ and $\mathbf{v}$ are in $U\cap W$, then $\mathbf{u}$ is both in U and in W, and $\mathbf{v}$ is both in U and in W. So $\mathbf{u}+\mathbf{v}$ in in U, because U is a subspace, and it is also in W, because W is a subspace. Thus $\mathbf{u}+\mathbf{v}$ is in $U\cap W$. □

Proposition 3.3.9. *If U and W are both subspaces of the vector space V, then $U+W$ is a subspace of V.*

Proof. Take two elements u_1+w_1 and u_2+w_2 in $U+W$. We must show that their sum is in $U+W$. This is clear because

$$(u_1+w_1)+(u_2+w_2)=(u_1+u_2)+(w_1+w_2)\in U+W.$$

Notice how we have used the associativity and commutativity of addition. The second property is even easier to prove, and left to you. □

Exercise 3.3.10. If U is the subspace generated by $\mathbf{u}_1,\ldots,\mathbf{u}_r$ and W is the subspace generated by $\mathbf{w}_1,\ldots,\mathbf{w}_s$, then $U+W$ is the subspace generated by

$$\mathbf{u}_1,\ldots,\mathbf{u}_r,\mathbf{w}_1,\ldots,\mathbf{w}_s.$$

Examples in $\mathbb{R}^3$.

It is important not to confuse affine subspaces with linear subspaces.

Example 3.3.11. In $\mathbb{R}^2$ with coordinates x_1 and x_2, let L be the line with equation:

$$a_1x_1+a_2x_2=b.$$

Assume that a_1 and a_2 are not both 0. Then if $b\neq 0$, L is not a linear subspace of $\mathbb{R}^2$, since $\mathbf{0}$ is not a point of L.

3.4 Bases

Before getting to the definition of a basis, we need two preliminary definitions.

Definition 3.4.1. V is again a vector space over F, and $\mathbf{v}_1, \ldots, \mathbf{v}_r$ elements of V. Then $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are *linearly dependent* if there are elements $a_1, a_2, \ldots, a_r$ in F, such that

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_1 + \cdots + a_r\mathbf{v}_r = \mathbf{0}\text{ , where not all the } a_i = 0. \tag{3.1}$$

Such an equation is called an equation of linear dependence. The key requirement is that the a_i cannot all be 0. Otherwise we could set all the a_i to be zero, and all sets of vectors would be linearly dependent.

If no equation of linear dependence exists, then the elements $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are *linearly independent*.

Example 3.4.2. If one of the $\mathbf{v}_i$ is the zero vector, then the collection of vectors is linearly dependent, since for any scalar a, $a\mathbf{0} = \mathbf{0}$. See Exercise 3.2.12. On the other hand, if $\mathbf{v}$ is not the zero vector, the set consisting just of $\mathbf{v}$ is linearly independent by Exercise 3.2.13.

Example 3.4.3. In F^3, let $\mathbf{i} = (1,0,0)$, $\mathbf{j} = (0,1,0)$ and $\mathbf{k} = (0,0,1)$. Prove these three vectors are linearly independent.

Example 3.4.4. Without difficulty we can generalize the previous example to F^n. For each i between 1 and n, let $\mathbf{e}_i$ be the i-th coordinate vector in F^n, meaning that it has a 1 in the i-th position, and a 0 everywhere else. So for example:

$$\mathbf{e}_1 = (1,0,\ldots,0), \mathbf{e}_2 = (0,1,0,\ldots,0), \ldots, \mathbf{e}_n = (0,\ldots,0,1).$$

Then the vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ are linearly independent.

Proof. Assume we have an equation of linear dependence:

$$a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + \cdots + a_r\mathbf{e}_n = \mathbf{0}$$

This can be written:

$$(a_1, a_2, \ldots, a_n) = (0,0,\ldots,0)$$

so all the a_i must be 0. Therefore our equation was not an equation of linear dependence. Contradiction. □

Here is a typical use of linear independence:

Theorem 3.4.5. *Let V be a vector space, and $\mathbf{v}_1, \ldots, \mathbf{v}_r$ a collection of linearly independent elements in V. Suppose that the following two linear combinations of the $\mathbf{v}_i$ are the same vector:*

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_r\mathbf{v}_r = b_1\mathbf{v}_1 + b_2\mathbf{v}_2 + \cdots + b_r\mathbf{v}_r$$

for scalars a_i and b_i, $1 \leq i \leq r$. Then $a_i = b_i$ for all i, so that they are in fact the same linear combination.

Proof. The equation yields:

$$(a_1 - b_1)\mathbf{v}_1 + (a_2 - b_2)\mathbf{v}_1 + \cdots + (a_r - b_r)\mathbf{v}_r = \mathbf{0}.$$

Linear independence of the $\mathbf{v}_i$ then says that all the coefficients are equal to zero, which is the desired result. □

Here is the second preliminary definition.

Definition 3.4.6. The vector space V is generated by $\mathbf{v}_1, \ldots, \mathbf{v}_r$ if every element in V can be written as a linear combination of the $\mathbf{v}_i$. We also say that the $\mathbf{v}_i$ span V.

Example 3.4.7. The vectors $\mathbf{e}_i$, $1 \leq i \leq n$ of Example 3.4.4 span F^n. However if you omit any one of them, the new collection does not span: why?

We can now make the fundamental definition of this section.

Definition 3.4.8. A *basis* of a vector space V is a set of linearly independent vectors that span V. If V has a basis with a finite number of elements, it is *finite-dimensional*.

Notice that we have defined what it means for a vector space to be finite dimensional without defining its dimension: that we will do in the next section.

A word of warning: zero-dimensional vector spaces do not have a basis. This means that zero-dimensional spaces have to be handled specially. Since they are trivial spaces (they only have one element: $\mathbf{0}$) this is not too much of a problem. We will usually be concerned with finite dimensional vector spaces, but we want a definition that applies to infinite dimensional vector spaces.

Example 3.4.9. The vectors $\mathbf{e}_i$, $1 \leq i \leq n$ of F^n in Example 3.4.4 form a basis. Indeed we have already seen that they are linearly independent, and that they obviously span. This basis is called the standard, or the natural basis of F^n.

Example 3.4.10. The vectors $1, t, t^2, \ldots, t^n, \ldots$ form an infinite basis for the polynomial ring $F[t]$. It is clear that they generate. Let us show they are linearly independent. This is always shown by contradiction: suppose there is an equation of linear dependence between a finite number of the basis element. This would imply that there is a polynomial of degree m:

$$f(t) = a_m t^m + a_{m-1} t^{m-1} + \cdots + a_1 t + a_0$$

that is identically equal to the zero polynomial. This cannot be.

Exercise 3.4.11. This is a continuation of Example 3.2.7. We now require that S be a finite set, and let V be the vector space of functions from S to $\mathbb{R}$. For any $s \in S$, let i_s be the function that takes the value 1 on s, and 0 on all the other points of S. Show that the i_s, $s \in S$ form a basis for V. So V is a finite dimensional vector space. For an arbitrary subset T of S, let

$$i_T = \sum_{s \in T} i_s,$$

so i_T is an element of V. Then

$$i_T(s) = \begin{cases} 1, & \text{if } s \in T; \\ 0, & \text{otherwise.} \end{cases}$$

which is why i_T is called the indicator function of T. Note that we have already written i_T as a linear combination of the basis elements: where?

Definition 3.4.12. By Theorem 3.4.5, any $\mathbf{v}$ in the finite dimensional vector space V can be written uniquely as

$$\mathbf{v} = a_1\mathbf{v}_1 + a_2\mathbf{v}_1 + \cdots + a_r\mathbf{v}_r$$

with respect to the basis $\mathbf{v}_1, \ldots, \mathbf{v}_r$. The a_i are called the *coordinates* of $\mathbf{v}$ with respect to the basis.

In F^r we write vectors as ordered r-tuples of F: $(a_1, a_2, \ldots, a_r)$. Using the standard basis of F^r given in Example 3.4.9, we see that the a_i are the coordinates with respect to the standard basis, justifying the terminology in Chapter 2.

Example 3.4.13. This yields one of the most important maps of linear algebra: the mapping $C_{\mathscr{B}}\colon V \to F^r$ that associates to any vector $\mathbf{v} \in V$ with basis $\mathscr{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$, the r-tuple of coordinates $(a_1, a_2, \ldots, a_r)$ of $\mathbf{v}$. We will have more to say about this mapping later. In particular we will show that it is a linear map (Example 4.1.9) and that it is injective and surjective. The injectivity follows from the uniqueness of the coordinates proved in Theorem 3.4.5, and the subjectivity then follows from the Rank-Nullity Theorem of Chapter 4.

Proposition 3.4.14. *Let $\mathbf{v}_1, \ldots, \mathbf{v}_r$ be a maximal subset of linearly independent elements in V, meaning that they are linearly independent, and that any element $\mathbf{w} \in V$ is linearly dependent on them. Then $\mathbf{v}_1, \ldots, \mathbf{v}_r$ is a basis of V.*

Proof. Since the $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are linearly independent, to show they form a basis we only have to show they generate V. Assume they do not. Then there is an element $\mathbf{w} \in V$ that cannot be written as a linear combination of the $\mathbf{v}_i$. By the hypothesis of maximality, we know that there is an equation of linear dependence:

$$a_1\mathbf{v}_1 + \cdots + a_r\mathbf{v}_r + b\mathbf{w} = \mathbf{0}.$$

Because $\mathbf{w}$ is not a linear combination of the $\mathbf{v}_i$, we must have $b = 0$. Then, because the $\mathbf{v}_i$ are linearly independent, all the a_i must be 0. Thus there is no equation of linear dependence, and we have the desired contradiction. □

Along the same lines we have:

Theorem 3.4.15. *Let V be a vector space of positive dimension. Assume that it is spanned by $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. Then a suitable subset of these elements forms a basis of V.*

Proof. If these is no equation of linear dependence between $\mathbf{v}_1, \ldots, \mathbf{v}_n$, they form a basis, so we are done. Otherwise there is an equation of linear dependence

$$a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n = \mathbf{0}.$$

Since not all the coefficients are equal to 0, we may assume that $a_n \neq 0$. Then we may solve for $\mathbf{v}_n$ in terms of the other generators. This shows that $\{\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}\}$ stills spans V. Continue eliminating generators on at a time in this way until there no longer is an equation of linear dependence. The remaining vectors form a basis. □

The assumption that V is positive dimensional is there only to exclude the trivial case $V = (0)$, in which case V does not have a basis. We will sometimes omit any mention of this case.

It is traditional to refer to this theorem by saying that one can extract a basis of any vector space from any set of generators.

3.5 Dimension

The key result of this section is that all bases of a finite dimensional vector space have the same number of elements, called its dimension. This is proved by the Steinitz exchange theorem:

Theorem 3.5.1. *Let V be a vector space. Assume that it is spanned by $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, and that $\{\mathbf{w}_1, \ldots, \mathbf{w}_r\}$ is a linearly independent set of vectors in V. Then $r \leq n$.*

Proof. Let $\mathscr{V}$ be the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. Since the collection of vectors

$$\mathscr{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_r\}$$

is linearly independent, $\mathbf{w}_1$ is non-zero, so we can write

$$\mathbf{w}_1 = a_1\mathbf{v}_1 + \ldots a_n\mathbf{v}_n$$

where not all the scalars a_i are 0. By changing the numbering of the variables, we may assume $a_1 \neq 0$. Dividing by a_1 and solving for $\mathbf{v}_1$ in terms of

$$\mathscr{V}_1 = \{\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\},$$

we see that $\mathscr{V}_1$ generates V. The set $\mathscr{W}_2 = \{\mathbf{w}_2, \ldots, \mathbf{w}_r\}$ is linearly independent so we repeat the argument: $\mathbf{w}_2$ is non-zero, so we can write

$$\mathbf{w}_2 = b_1\mathbf{w}_1 + c_2\mathbf{v}_2 + \ldots c_n\mathbf{v}_n.$$

At least one of the c_i, $2 \leq i \leq n$ must be non-zero because $\mathbf{w}_1$ and $\mathbf{w}_2$ are linearly independent. By renumbering we may assume it is c_2. Thus we can solve for $\mathbf{v}_2$ in

terms of

$$\mathscr{V}_2 = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{v}_3, \ldots, \mathbf{v}_n\}.$$

Thus $\mathscr{V}_2$ generates V.

Assume by contradiction that $n < r$. Continuing as above, replacing one element of $\mathscr{V}$ by an element of $\mathscr{W}$, we see that

$$\mathscr{V}_n = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n\}$$

spans. But then $\mathbf{w}_{n+1}$ can be written as a linear combination of the elements of $\mathscr{V}_n$, which contradicts the linear independence of $\mathscr{W}$, and we get the desired contradiction. □

Corollary 3.5.2. *Any two bases of a finite dimensional vector space have the same number of elements.*

Proof. Call the bases $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$. Since the vectors in both sets are linearly independent and span, we can apply the theorem in both directions: $n \leq m$ and $m \leq n$, so we are done. □

Definition 3.5.3. The *dimension* of a finite-dimensional vector space V is the number of elements in one (and therefore any) basis, assuming V has a basis. To the trivial vector space $\mathbf{0}$ we assign the dimension 0.

Example 3.5.4. The dimension of F^n is n. By convention this holds even for $n = 0$.

Exercise 3.5.5. Establish the dimensions of the following vector spaces by exhibiting a basis.

1. The dimension of the vector space of $m \times n$ matrices is mn.
2. The dimension of the space of diagonal matrices of size n is n.
3. The dimension of the space of upper-triangular matrices of size n is $\frac{n(n+1)}{2}$.
4. The dimension of the space of symmetric matrices of size n is $\frac{n(n+1)}{2}$.
5. The dimension of the space of skew-symmetric matrices of size n is $\frac{n(n-1)}{2}$. A skew-symmetric matrix is a square matrix such that $a_{ij} = -a_{ji}$ for all i and j.

Exercise 3.5.6. Show that any square matrix can be written as the sum of a symmetric matrix and a skew-symmetric matrix.

Corollary 3.5.7. *Let V be a vector space of dimension n. Suppose that $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent elements in V. Then they form a basis.*

Proof. By Proposition 3.4.14, if they do not form a basis, then we can find an element $\mathbf{v}_{n+1} \in V$ such that $\mathbf{v}_1, \ldots, \mathbf{v}_{n+1}$ are linearly independent. This contradicts the Steinitz Exchange Theorem 3.5.1 above. □

Corollary 3.5.8. *If V is a vector space of dimension n, and W a subspace of V of dimension n, then $V = W$.*

Proof. By the previous corollary, a basis for W, namely a set of n linearly independent elements of V, is a basis for V. □

Corollary 3.5.9. *Let V be an n dimensional vector space, and W a subspace of V that is not 0 dimensional. Then W has a basis with at most n elements.*

Proof. Since W has positive dimension, we can pick a non-zero vector $\mathbf{w}_1$ in W. This gives a linearly independent set. If it does not span W, we may find a $\mathbf{w}_2$ linearly independent from $\mathbf{w}_1$, etc... □

Corollary 3.5.10. *Let $\mathbf{v}_1, \ldots, \mathbf{v}_r$ be r linearly independent vectors in the n-dimensional vectors space V. Then $n-r$ elements $\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n$ can be added to that $\mathbf{v}_1, \ldots, \mathbf{v}_n$ forms a basis of V. We say any linearly independent subset of V can be completed to a basis.*

Proof. If $r = n$ we already have a basis by Corollary 3.5.8. Otherwise, by Corollary 3.5.9, $r < n$. So by Definition of the dimension, we can find a $\mathbf{v}_{r+1}$ that is linearly independent of the first r vectors. Then repeat the argument to the $r+1$ vectors if $r+1 < n$ until you get to n vectors. Then we have a basis. □

Other examples:

3.6 Products and Direct Sums

We discuss two ways of producing new vectors spaces from old ones. We are given two vector spaces U and W, both over the same field F. Considering U and W as sets, we can form the cartesian product $U \times W$: see §B.1.

Theorem 3.6.1. *$U \times W$ is a vector space, with the obvious operations:*

1. *Addition is component-wise*

$$(\mathbf{u}_1, \mathbf{w}_1) + (\mathbf{u}_2, \mathbf{w}_2) = (\mathbf{u}_1 + \mathbf{u}_2, \mathbf{w}_1 + \mathbf{w}_2)$$

2. *Scalar multiplication is*

$$c(\mathbf{u}, \mathbf{w}) = (c\mathbf{u}, c\mathbf{w})$$

The proof is an exercise for you.

Theorem 3.6.2. *If the dimension of U is m, and the dimension of W is n, then the dimension of $U \times W$ is $m+n$.*

Proof. We prove this by exhibiting a basis of $U \times W$, given a basis $\mathbf{u}_1, \ldots \mathbf{u}_m$ of U and a basis $\mathbf{w}_1, \ldots \mathbf{w}_n$ of W. It consists in the elements $(\mathbf{u}_1, \mathbf{0}), \ldots (\mathbf{u}_m, \mathbf{0})$ and $(\mathbf{0}, \mathbf{w}_1), \ldots (\mathbf{0}, \mathbf{w}_m)$.

As always we need to show these are linearly independent and span.

Assume they are not linearly independent. Then there is an equation of linear dependence:

$$a_1(\mathbf{u}_1,\mathbf{0})+\cdots+a_m(\mathbf{u}_m,\mathbf{0})+b_1(\mathbf{0},\mathbf{w}_1)+\cdots+b_n(\mathbf{0},\mathbf{w}_n)=(\mathbf{0},\mathbf{0})$$

Considering only the first coordinate, we get

$$a_1\mathbf{u}_1+\cdots+a_m\mathbf{u}_m=\mathbf{0}$$

which by linear independence of the $\mathbf{u}_i$ says that all the a_i are zero. doing the same thing for the second coordinate, we see that all the b_j are zero, so this is not an equation of linear dependence and we are done.

It is left to you to show that they span. □

Example 3.6.3. $\mathbb{R}^2$ is the product of $\mathbb{R}$ by $\mathbb{R}$, and $\mathbb{R}^n$ is the product of $\mathbb{R}^k$ by $\mathbb{R}^m$ for any positive integers k, m and n such that $k+m=n$.

Definition 3.6.4. The subspace of $U\times U$ of elements $(\mathbf{u},\mathbf{u})$, for all $\mathbf{u}\in U$ is called the diagonal. It has the same dimension as U.

Next assume we have a vector space V and two subspaces U and W of V.

Definition 3.6.5. We say that V is the *direct sum* of U and W if any element $\mathbf{v}\in V$ can be written uniquely as $\mathbf{u}+\mathbf{w}$, for $\mathbf{u}\in U$ and $\mathbf{w}\in W$. We then write $V=U\oplus W$.

This definition does not require that U or V be finite dimensional. If they both are, we have:

Exercise 3.6.6. Check that if $V=U\oplus W$, then $\dim U+\dim W=\dim V$. Indeed, $U\cap V$ must reduce to $(\mathbf{0})$.

We generalize this result in Theorem 4.2.11.

Problem 3.6.7. For vector spaces U and W, form the cartesian product $V=U\times W$. Then let U_1 be the subspace of V formed by all elements $(u,0)$, for $u\in U$. Let W_1 be the subspace of V formed by all elements $(0,w)$, for $w\in W$.

Then show that $V=U_1\oplus W_1$.

Chapter 4
Linear Maps

Abstract We now get to the second key definition of linear algebra: that of a linear map between vector spaces. These are the 'allowable maps' of linear algebra. The most important linear map, and, as we will see in Chapter 5, essentially the only example, is given by matrix multiplication: see Example 4.1.10. To a linear map we can associate two interesting new vector spaces: the nullspace and the range, defined in §4.2. Then we prove our first major theorem: the Rank-Nullity Theorem 4.2.8. Then we show that the composition of two linear maps, when it is defined, is linear. After studying the algebra of linear maps, we study invertible linear maps $L\colon U \to V$. They establish a bijection between the vector spaces U and V, and preserve the structure of vector space, as we show in §4.5.1: we say U and V are isomorphic. By the Rank-Nullity theorem, two vector spaces are isomorphic if and only if they have the same dimension.

4.1 Linear Maps

Definition 4.1.1. Let U and V be vector spaces over the field F. Then a linear map is a map L that satisfies the following two properties:

LM 1 $L(\mathbf{u}+\mathbf{v}) = L(\mathbf{u}) + L(\mathbf{v})$ for all $\mathbf{u}, \mathbf{v} \in U$. The addition on the left hand side of this equation is in U, while the addition on the right hand side in in V.

LM 2 $L(a\mathbf{u}) = aL(\mathbf{u})$ for all $\mathbf{u} \in U$ and all $a \in F$. The scalar multiplication on the left is in U, and on the right is in V.

Exercise 4.1.2. Prove that the property $L(\mathbf{0}) = \mathbf{0}$ follows from the definition. This is sometimes included in the definition of a linear map, but is not needed.

Exercise 4.1.3. Prove that $L(-\mathbf{u}) = -L(\mathbf{u})$.

First let's get two trivial examples of linear maps out of the way.

Example 4.1.4. If L takes every element of the vector space V to $\mathbf{0}$, then L is linear. It is called the zero map.

Example 4.1.5. The map $L\colon V \to V$ such that $L(\mathbf{v}) = \mathbf{v}$ is linear. It is called the identity map.

Remark 4.1.6. Note that a linear map can be defined for infinite dimensional vectors spaces U and V. This is, in fact, one of the reasons for making the definition

Now for a more interesting example.

Definition 4.1.7 (Projection). Let V be a vector space that is written as the direct sum of subspaces U and W: $V = U \oplus W$. See §3.6. So any $\mathbf{v} \in V$ can be written uniquely as $\mathbf{v} = \mathbf{u} + \mathbf{w}$, $\mathbf{u} \in U$ and $\mathbf{w} \in W$. Then $\mathbf{u}$ is called the component of $\mathbf{v}$ in U, and $\mathbf{w}$ is called the component of $\mathbf{v}$ in W. The linear map P_1 such that $P_1(\mathbf{v}) = \mathbf{u}$, its component $\mathbf{u}$ in U, is called the projection from V to U along W. Similarly we have a linear map $P_2\colon V \to W$, sending $\mathbf{v}$ to $\mathbf{w}$, the projection from V to W along U. If V is finite dimensional, then $\dim V = \dim U + \dim W$.

Let's show that P_1 is a linear map. For any scalar c, $c\mathbf{v} = c\mathbf{u} + c\mathbf{w}$, where $c\mathbf{u} \in U$ and $c\mathbf{w} \in W$ since they are subspaces. For the same reason, if $\mathbf{v}' = \mathbf{u}' + \mathbf{w}'$, with $\mathbf{u}' \in U$ and $\mathbf{w}' \in W$, then

$$\mathbf{v} + \mathbf{v}' = \mathbf{u} + \mathbf{w} + \mathbf{u}' + \mathbf{w}' = \mathbf{u} + \mathbf{u}' + \mathbf{w} + \mathbf{w}'$$

so that $\mathbf{v} + \mathbf{v}'$ is mapped to $\mathbf{u} + \mathbf{u}' \in U$.

We could think of this map as a linear map from V to U, but via the inclusion $U \subset V$ it is a map from V to V, and that will be our point of view.

In the same way, P_2 is a linear map from V to V. We define the sum of the maps P_1 and P_2 as

$$(P_1 + P_2)(\mathbf{v}) = P_1(\mathbf{v}) + P_2(\mathbf{v}).$$

This is the identity map: $(P_1 + P_2)(\mathbf{v}) = \mathbf{u} + \mathbf{w} = \mathbf{v}$.

Need two pictures here : in $\mathbb{R}^2$ with two different bases as shown, one the standard perpendicular basis, the other skew, show the two projections.

Example 4.1.8. The linear map $F^n \to F^n$, that sets to zero any set of $n - m$ coordinates in a projection. For example $F^4 \to F^4$ sending $(x_1, x_2, x_3, x_4) \mapsto (x_1, 0, x_3, 0)$ or the different projection sending $(x_1, x_2, x_3, x_4) \mapsto (0, x_2, 0, x_4)$.

When the subspace U is either the zero dimensional subspace or the full space V, P_1 is the zero map or the identity map, respectively.

For more about projections, see §4.6 .

Example 4.1.9 (Coordinate Map). For any vector space V with basis $\mathfrak{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, there is a linear map $V \to F^n$ associating to $\mathbf{v} = a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n$ in V its coordinates $(a_1, \ldots, a_n)$ in the basis. We write

$$[\mathbf{v}]_{\mathfrak{B}} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

This linear map will be useful in §5.3. To establish the linearity we must show

- $[c\mathbf{v}]_{\mathfrak{B}} = c[\mathbf{v}]_{\mathfrak{B}}$.
- $[\mathbf{v}+\mathbf{w}]_{\mathfrak{B}} = [\mathbf{v}]_{\mathfrak{B}} + [\mathbf{w}]_{\mathfrak{B}}$.

To do this, just write $\mathbf{v}$, $c\mathbf{v}$ and $\mathbf{w}$ in coordinates.

The following key example gives the link between linear maps and matrices in the finite dimensional case.

Example 4.1.10 (Matrix to Linear Map). For any $m \times n$ matrix A with coefficients in F, we get a linear map $L_A\colon F^n \to F^m$ that associates to the n- column vector $\mathbf{v}$ the matrix product $A\mathbf{v}$.

The linearity follows from Theorem 2.2.13 on matrix multiplication that establishes:

$$A(B+C) = AB+AC \text{ and } A(cB) = c(AB)$$

for matrices of the appropriate sizes. Let B and C be the n-column vectors $\mathbf{u}$ and $\mathbf{v}$ to get linearity. Then

$$L_A(\mathbf{u}+\mathbf{v}) = A(\mathbf{u}+\mathbf{v}) = A\mathbf{u}+A\mathbf{v} = L_A(\mathbf{u})+L_A(\mathbf{v})$$

as required. The second verification is left to the reader.

Theorem 4.1.11. *The set V of all linear maps from a vector space U to a vector space W is itself a vector space, denoted $\mathscr{L}(U,W)$.*

Proof. This is closely related to Example 3.2.7, that we rewrite in part. The vector space structure on V, and the neutral element are defined as in Example 3.2.7. The inverse of a linear transformation L in V is the map M such that $M(\mathbf{u}) = -L(\mathbf{u})$. You need to check that M is in V, namely that it is a linear transformation. This follows because L is a linear transformation:

$$M(\mathbf{u}+\mathbf{v}) = -L(\mathbf{u}+\mathbf{v}) = -L(\mathbf{u})-L(\mathbf{v}) = M(\mathbf{u})+M(\mathbf{v})$$

and

$$M(c\mathbf{u}) = -L(c\mathbf{u}) = -cL(\mathbf{u}) = cM(\mathbf{u})$$

as required. The other parts follow as in Example 3.2.7, and are left to you. □

We will use the following theorem many times to build linear maps.

Theorem 4.1.12. *Let V be a F-vector space of dimension n, and W a F-vector space of some dimension. Let $\{\mathbf{v}_1,\dots,\mathbf{v}_n\}$ be a basis for V, and $\{\mathbf{w}_1,\dots,\mathbf{w}_n\}$ any collection of n elements of W. There there is a unique linear map $L\colon V \to W$ such that $L(\mathbf{v}_i) = \mathbf{w}_i$, $1 \le i \le n$.*

Proof. Since any $\mathbf{v} \in V$ can be written uniquely as

$$\mathbf{v} = a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n \text{ , for suitable } a_i \in F$$

we define

$$L(\mathbf{v}) = a_1\mathbf{w}_1 + \cdots + a_n\mathbf{w}_n.$$

Thus we have a uniquely defined map L. We need to show that L is linear. First we pick a second element $\mathbf{v}' \in V$ written

$$\mathbf{v}' = b_1\mathbf{v}_1 + \cdots + b_n\mathbf{v}_n$$

and show that

$$L(\mathbf{v}+\mathbf{v}') = L(\mathbf{v}) + L(\mathbf{v}').$$

Indeed

$$L(\mathbf{v}+\mathbf{v}') = (a_1+b_1)\mathbf{w}_1 + \cdots + (a_n+b_n)\mathbf{w}_n = L(\mathbf{v}) + L(\mathbf{v}').$$

Then for any $c \in F$,

$$L(c\mathbf{v}) = L(ca_1\mathbf{v}_1 + \cdots + ca_n\mathbf{v}_n) = ca_1\mathbf{w}_1 + \cdots + ca_n\mathbf{w}_n = cL(\mathbf{v}),$$

which concludes the proof. □

Exercise 4.1.13. Let L be a linear map between a vector space V of dimension n and a vector space W of some dimension. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be any collection of elements of V, and $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ a linearly independent set of n elements of W. Assume that $L(\mathbf{v}_j) = \mathbf{w}_j$, $1 \leq j \leq n$. Prove that the $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ are linearly independent.

Hint: See the proof of the Rank-Nullity Theorem 4.2.8 below.

4.2 The Nullspace and the Range of a Linear Map

In this section we define the two most important subspaces associated to a linear map $L\colon U \to V$. They can be defined even when U and V are infinite dimensional.

Definition 4.2.1. The nullspace of L is the subset of $\mathbf{u} \in U$ such that $L(\mathbf{u}) = \mathbf{0}$. The nullspace is called the kernel in some books, but we will always use nullspace.

Theorem 4.2.2. *The nullspace N_L of L is a subspace of U.*

Proof. We must show that if $\mathbf{u}$ and $\mathbf{v}$ are in the nullspace, then $\mathbf{u}+\mathbf{v}$ and $a\mathbf{u}$ are in the nullspace, for any $a \in F$. By definition of a linear map

$$L(\mathbf{u}+\mathbf{v}) = L(\mathbf{u}) + L(\mathbf{v}) = \mathbf{0}$$

so $\mathbf{u}+\mathbf{v}$ is in the nullspace as required. Similarly

$$L(a\mathbf{u}) = aL(\mathbf{u}) = \mathbf{0}$$

so we are done. □

We already stated a special case of this result in Example 3.3.5:

Example 4.2.3. If L is the linear map of Example 4.1.10 for the $m \times n$ matrix A:

$$L(\mathbf{u}) = A\mathbf{u},$$

then the nullspace of L is the set of solutions of the homogeneous system of equations $A\mathbf{x} = \mathbf{0}$.

Since the nullspace is a subspace of U, it has a dimension, called the nullity of L.

Remark 4.2.4. If the nullity of L is 0, then L is injective.

Proof. Indeed, if $L(\mathbf{u}_1) = L(\mathbf{u}_2)$, then by linearity $L(\mathbf{u}_1 - \mathbf{u}_2) = \mathbf{0}$. This says that $\mathbf{u}_1 - \mathbf{u}_2$ is in N_L, and since the nullity is 0, it must be $\mathbf{0}$, so $\mathbf{u}_1 = \mathbf{u}_2$. □

More examples here.

Now we turn to the second subspace: as before $L\colon U \to V$ is a linear map.

Definition 4.2.5. The range R_L of L is the set of $\mathbf{v} \in V$ such that there is a $\mathbf{u} \in U$ with $L(\mathbf{u}) = \mathbf{v}$.

Theorem 4.2.6. *The range of L is a subspace of V.*

Proof. The proof proceeds in exactly the same way as for the nullspace. Assume that $\mathbf{v}_1$ and $\mathbf{v}_2$ are in the range, so that there are elements $\mathbf{u}_1$ and $\mathbf{u}_2$ in U with $L(\mathbf{u}_i) = \mathbf{v}_i$. Then by linearity,

$$L(\mathbf{u}_1 + \mathbf{u}_2) = L(\mathbf{u}_1) + L(\mathbf{u}_2) = \mathbf{v}_1 + \mathbf{v}_2$$

so that $\mathbf{v}_1 + \mathbf{v}_2$ is in the range, as required. The second part is left to you. □

Definition 4.2.7. The rank of L is the dimension of the range of L.

Examples here.

We now get to one of the most important theorems in linear algebra.

Theorem 4.2.8 (The Rank-Nullity Theorem). *If $L\colon U \to V$ is a linear map between finite dimensional vector spaces, if n is the nullity of L, r its rank, and d is the dimension of U, then*

$$n + r = d.$$

Proof. Pick a basis $\mathbf{v}_1, \dots \mathbf{v}_r$ of the range of L. By definition of the range, we may find elements $\mathbf{u}_1, \dots \mathbf{u}_r$ in U such that $L(\mathbf{u}_i) = \mathbf{v}_i$. Then the $\mathbf{u}_i$ are linearly independent (in U). Indeed, suppose not. Then there is an equation of linear dependence:

$$a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \cdots + a_r\mathbf{u}_r = \mathbf{0}.$$

Apply L. This gives

$$a_1L(\mathbf{u}_1) + a_2L(\mathbf{u}_2) + \cdots + a_rL(\mathbf{u}_r) = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_r\mathbf{v}_r = \mathbf{0},$$

an impossibility since the $\mathbf{v}_i$ are linearly independent.

Now let $\mathbf{w}_1, \dots \mathbf{w}_n$ be a basis of the nullspace. We claim that the $\mathbf{u}_i$ and the $\mathbf{w}_j$ form a basis of U.

To prove this, we first prove these vectors span U. Take an arbitrary $\mathbf{u} \in U$. Then $L(\mathbf{u})$ can be written in terms of the basis of the range:

$$L(\mathbf{u}) = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_r\mathbf{v}_r$$

for suitable scalars a_i. Then we see that

$$L(\mathbf{u} - a_1\mathbf{u}_1 - \cdots - a_r\mathbf{u}_r) = \mathbf{0}.$$

Thus $\mathbf{u} - a_1\mathbf{u}_1 - \cdots - a_r\mathbf{u}_r$ is an element of the nullspace, so it can be written

$$\mathbf{u} - a_1\mathbf{u}_1 - \cdots - a_r\mathbf{u}_r = b_1\mathbf{w}_1 + \cdots + b_r\mathbf{w}_n.$$

This shows that the $\mathbf{u}_i$ and the $\mathbf{w}_j$ span. To show that they form a basis, assume by contradiction that they satisfy an equation of linear dependence:

$$a_1\mathbf{u}_1 + \cdots + a_r\mathbf{u}_r + b_1\mathbf{w}_1 + \cdots + b_n\mathbf{w}_n = \mathbf{0}. \tag{4.1}$$

Apply L to get

$$a_1L(\mathbf{u}_1) + \cdots + a_rL(\mathbf{u}_r) = \mathbf{0}$$

since the remaining vectors are in the nullspace. Since the $\mathbf{v}_j = \mathbf{L}(u_j)$ are linearly independent, this forces all the a_i to be zero. Then (4.1) becomes

$$b_1\mathbf{w}_1 + \cdots + b_r\mathbf{w}_n = \mathbf{0},$$

which implies all the b_j are 0 since the $\mathbf{w}_j$ are linearly independent. Thus all the coefficients of (4.1) are zero, so it is not an equation of linear dependence.

Thus a basis for U has $r+n$ elements, and we are done. □

Corollary 4.2.9. *Let $L\colon U \to V$ be a linear map between finite dimensional vector spaces of the same dimension d. Then if the nullspace has dimension 0, or if the range is V, then L is bijective.*

Proof. We need to show that L is both injective and surjective.

First assume the nullspace of L has dimension 0. Then L is injective by Remark 4.2.4. By the Rank-Nullity Theorem, the range of L has dimension d, so by Corollary 3.5.8 it is all of V, so the map is surjective.

Next assume the range of L is V. Then L is surjective. The nullspace of L has dimension 0, so it is also injective. □

Example 4.2.10. Let V be a vector space that is written as the direct sum of subspaces U and W, so that $\dim V = \dim U + \dim W$. See §3.6. Then in Definition 4.1.7 we defined the projection P of V to U along W. By construction the range of P is U and the nullspace is W, as you should check.

As a corollary of the rank-nullity theorem, we get the following important formula.

Theorem 4.2.11. *Let U and W be subspaces of a finite dimensional vector space V. Then*

$$\dim U + \dim W - \dim U \cap W = \dim(U + W).$$

Proof. First recall that in §3.3 we defined the subspaces $U + W$ and $U \cap W$ of V. In §3.6 we also defined the direct sum $U \oplus W$ of two vector spaces. We build a linear map

$$L\colon U \oplus W \to V$$

by

$$L((\mathbf{u}, \mathbf{w})) = \mathbf{u} - \mathbf{w}.$$

You should first convince yourself that this is a linear map. Then notice that its nullspace is $U \cap W$. Its range is $U + W$. In Exercise 3.6.6 we established that $\dim U \oplus W = \dim U + \dim W$, so the theorem is a direct corollary of the Rank-Nullity Theorem. □

Next we prove a proposition that will be useful later.

Proposition 4.2.12. *Let $L\colon V \to W$ be a linear transformation with nullspace N of dimension ν, and U a subspace of V of dimension u. The dimension of V is n and the rank of L is r. Then*

$$\dim(L(U)) = \dim U - \dim(U \cap N).$$

Thus the dimension of $L(U)$ is at least $\min(0, u - \nu)$ *and at most* $\min(n, u + \nu) - \nu$.

Proof. We restrict the linear transformation L to U. The nullspace of the restriction is clearly $U \cap N$, so the equality just expresses the Rank-Nullity Theorem for the restriction. For the inequalities we use Theorem 4.2.11 applied to U and the nullspace N of L inside of V. The range of U will be as small as possible if U contains N, or at least as much of it as it can. That gives the first inequality. The range of U will be as large as possible by making the intersection of U and N as small as possible. That gives the second inequality.

Let's apply this result to some low dimensional examples. You may assume that F is the real numbers.

Example 4.2.13. Suppose that V is 3-dimensional. So the ambient space is ordinary space. Suppose that U and W are both surfaces: i.e. they have dimension two. So $\dim U + \dim W = 4$. Now $U + W$ is a subspace of a three dimensional space, so it has dimension at most three. On the other hand it has dimension at least 2: why? If $U + W$ has dimension 3, then by the theorem U and W intersect in a line: through the origin, since the intersection is a subspace. If $U + W$ has dimension 2, then U and W must be the same subspace of V: why?

Example 4.2.14. Suppose that V is 4-dimensional. Suppose that U and W are both surfaces: i.e. they have dimension two. So $\dim U + \dim W = 4$. Now $U + W$ is a subspace of a 4 dimensional space, so this time there is no restriction on its dimension.

Again it has dimension at least 2. If $U+W$ has dimension 4, then by the theorem U and W only intersect at the origin. If $U+W$ has dimension 3, they intersect in a line through the origin. If $U+W$ has dimension 2, then U and W must be the same subspace of V.

Example 4.2.15. The space M_n of square matrices of size n is a vector space of dimension n. As we noticed in Exercise 3.5.5 if has dimension n^2, while the subspace U of symmetric matrices has dimension $\frac{n(n+1)}{2}$ and the subspace V of skew-symmetric matrices has dimension $\frac{n(n-1)}{2}$. Convince yourself that $U \cap V = (0)$, so that every square matrix can be written uniquely as the sum of a symmetric matrix and a skew-symmetric matrix.

4.3 Composition of Linear Maps

Theorem 4.3.1. *If L is a linear map from U to V, and M a linear map from V to W, then the composition $M \circ L$ is a linear map from U to W.*

Proof. As always, we must show two things. $\mathbf{u}$ and $\mathbf{v}$ are arbitrary elements in U, and c is an arbitrary scalar.

$$(M \circ L)(\mathbf{u}+\mathbf{v}) = (M \circ L)(\mathbf{u}) + (M \circ L)(\mathbf{v}) \text{ and } (M \circ L)(c\mathbf{u}) = c(M \circ L)(\mathbf{u}).$$

Since L is linear,

$$L(\mathbf{u}+\mathbf{v}) = L(\mathbf{u}) + L(\mathbf{v}).$$

By linearity of M

$$M(L(\mathbf{u}) + L(\mathbf{v})) = M(L(\mathbf{u})) + M(L(\mathbf{v})).$$

By the definition of composition of maps this is $(M \circ L)(\mathbf{u})) + (M \circ L)(\mathbf{v}))$, so we are done. The second equality is even easier:

$$(M \circ L)(c\mathbf{u}) = M(L(c\mathbf{u})) = M(cL(\mathbf{u})) = cM(L(\mathbf{u})) = c(M \circ L)(\mathbf{c})$$

where we first use the linearity of L and then that of M □

We have shown in Theorem 4.1.11 that the linear maps from U to V form a vector space, which we denote $\mathscr{L}(U,V)$. We can now work out the interaction between the vector space operations and composition.

Theorem 4.3.2. *Let U, V and W be vectors spaces. Let L_1 and L_2 be linear maps from U to V, and M_1 and M_2 linear maps from V to W.*

1. Then following two equations are satisfied:

$$M_1 \circ (L_1 + L_2) = M_1 \circ L_1 + M_1 \circ L_2 \text{ and } (M_1 + M_2) \circ L_1 = M_1 \circ L_1 + M_2 \circ L_1.$$

The addition on the right hand side of these equations is that in $\mathscr{L}(U,W)$, while that on the left hand side is in $\mathscr{L}(U,V)$ for the first equation and $\mathscr{L}(V,W)$ for the second.

2. *If c is a scalar,*
$$M_1 \circ (cL_1) = (cM_1) \circ L_1 = c(M_1 \circ L_1).$$

Proof. The idea of the proof is simple: to prove that two linear maps are equal, we simply show that they give the same value when applied to an arbitrary element of the domain vector space. So let $\mathbf{u}$ be an arbitrary element of U. then to establish the first equation we need to show:

$$(M_1 \circ (L_1 + L_2))(\mathbf{u}) = M_1 \circ L_1(\mathbf{u}) + M_1 \circ L_2(\mathbf{u})$$

Use the associativity of composition to write the left hand side as $M_1((L_1 + L_2)(\mathbf{u})$, then the meaning of addition in $\mathscr{L}(U,W)$ to get $M_1(L_1(\mathbf{u}) + L_2(\mathbf{u}))$, then the linearity of M_1 to get $M_1(L_1(\mathbf{u})) + M_1(L_2(\mathbf{u}))$. This is the desired result.

The other results are proved the same way. □

Now that we know that $M \circ L$ is a linear transformation, what can be said about its rank and its nullity in the finite dimensional case? First some notation. Let n be the dimension of U, m the dimension of V and l the dimension of W.

Definition 4.3.3. We have five important subspaces associated to the composition of linear maps $M \circ L$, where $L\colon U \to V$ and $M\colon V \to W$.

1. In U, we have the nullspace N_L of L;
2. In V we have the range R_L of L, the nullspace N_M of M and their intersection $R_L \cap N_M$;
3. In W we have the range R_M of M.

The composition $M \circ L$ factors through R_L, so, denoting by $M_{|R_L}$ the restriction of M to the subspace R_L, we get:

$$R_{M\circ L} = R_{M_{|R_L}}$$

which yields, by the Rank-Nullity Theorem applied to $M_{|R_L}$:

$$\dim(R_{M\circ L}) = \dim(R_L) - \dim(R_L \cap N_M). \tag{4.2}$$

Therefore by the Rank-Nullity Theorem applied to $M \circ L$ and also to L we get

$$\begin{aligned} \dim(N_{M\circ L}) &= n - \dim(R_L) + \dim(R_L \cap N_M) \\ &= \dim(N_L) + \dim(R_L \cap N_M). \end{aligned} \tag{4.3}$$

Exercise 4.3.4. Convince yourself that both results make sense by drawing a picture in a low dimensional case, with different values for the dimensions of R_L, N_M, and $R_L \cap N_M$.

Numerical examples here.

We have the following classic theorems relating the ranks of linear maps and their compositions.

First an easy result.

Theorem 4.3.5. *Let L and M be linear maps from U to V with ranges of dimension $r(L)$ and $r(M)$. Let $r(L+M)$ be the dimension of the range of $L+M$. Then*

$$r(L+M) \leq r(L)+r(M)$$

Proof. The range of $L+M$ is spanned by the ranges of L and M, but these two spaces could well intersect non-trivially, giving the inequality. □

Exercise 4.3.6. With the notation of the theorem, show $r(L-M) \geq |r(L)-r(M)|$.

Hint: replace L by $L-M$ in the theorem.

Examples and exercises here.

Now for the main theorem:

Given linear maps $L\colon U \to V$ and $M\colon V \to W$, where n is the dimension of U, m the dimension of V and l the dimension of W.

Theorem 4.3.7 (Sylvester's Law of Nullity). *Given linear maps $L\colon U \to V$ and $M\colon V \to W$, where n be the dimension of U, m the dimension of V and l the dimension of W, then*

1. the nullity ν satisfies:

$$\nu(L) \leq \nu(M) \leq \nu(L)+\nu(M).$$

2. the rank r satisfies:

$$r(L)+r(M)-n \leq r(M\circ L) \leq \min(r(L),r(M)).$$

Proof. The inequalities concerning nullities are a direct consequence of (4.3), which also tells us when the extreme cases are realized.

Next we establish the right hand inequality for the rank, which is probably the most useful of the four inequalities. The range of $M\circ L$ is contained in the range of M, so $r(M\circ L) \leq r(M)$. On the other hand the nullspace of $M\circ L$ contains that of L as you see by applying both sides to an arbitrary element of L. So $\nu(M\circ L) \geq \nu(L)$. By (4.2) we get $r(M\circ L) \leq r(L)$, so we are done.

Finally we get the left hand inequality for the rank: the left side can be written

$$r(L)+r(M)-n = r(L)-\nu(M)$$

by the Rank-Nullity Theorem, so the inequality follows immediately the inequality $r(M\circ L) \leq r(L)$ that we have already used. □

Corollary 4.3.8. *If L and M are linear maps from U to U, a vector space of dimension n, and one of them has rank n, then the rank of the other is the rank of $M\circ L$. In particular, if L has rank n then the rank of $L^{-1}ML$ is the same as that of M.*

Proof. The first statement follows immediately from the formula for the ranks. For the second statement we need to know that if L has rank n, then its inverse map L^{-1} is a linear transformation. This will be established in Theorem 4.5.1. Then just apply the first part twice.

We will see the importance of the last statement in Theorems 5.5.1 and 5.5.4.

Exercise 4.3.9 (Frobenius's Inequality). If L, M and N are linear transformations that can be composed in the order $N \circ M \circ L$, then show

$$r(N \circ M) + r(N \circ A) \leq r(M) + r(N \circ M \circ L).$$

Hint: Just use the inequalities of Sylvester's Law of Nullity.

Exercise 4.3.10. L is a linear map from U to V, and M a linear map from V to W.

1. If L and M are both injective, then so is $M \circ L$.
2. If L and M are both surjective, then so is $M \circ L$.

4.4 Linear Operators

Composition of linear maps is even more interesting in the special case where V and W are the same vector space as U.

Definition 4.4.1. A linear map from a vector space to itself is called a linear operator.

We can form the power of the linear operator $L\colon U \to U$ with itself any number of times by composition. We write L^2 for $L \circ L$, L^3 for $L \circ L \circ L$, $L^r = L \circ L \circ \cdots \circ L$, r times, for any positive integer r. We also set L^0 to be the identity operator.

Exercise 4.4.2. Why is $L^n \circ L^m = L^{(n+m)}$ for all non-negative integers n and m?

A word of warning: if L and M are two linear operators on U it is not always the case that $L \circ M = M \circ L$. In fact an important part of what we will do later in this course is to determine when the two operators can be interchanged. For simplicity, when dealing with linear operators we often write LM for $L \circ M$.

Thus any polynomial in L:

$$a_n L^n + a_{n-1} L^{n-1} + a_1 L^1 + a_0 L^0 \text{ , } a_i \in F$$

is an operator. Here n is a positive integer. We sometimes omit the L^0 when writing the polynomial, or write I for L^0.

Example 4.4.3. Two operators M and N that are polynomials in the operator L commute: $MN = NM$.

We will exploit the idea of taking polynomials of operators systematically in Chapter 10.

Exercise 4.4.4. Assume $L^2 - I = 0$. where 0 is the operator that sends everything to zero. Then let

$$M = \frac{1}{2}(L+I) \text{ and } N = \frac{1}{2}(-L+I).$$

Show that $M+N = I$, $M^2 = M$, $N^2 = N$ and $MN = NM = 0$ simply by doing algebra in polynomials in L.

Exercise 4.4.5. Let L and M be linear operators on U. Assume they commute: $LM = ML$. Then, for example

$$(L+M)^2 = L^2 + 2LM + M^2 \text{ and } (L+M)((L-M) = L^2 - M^2.$$

4.5 Invertible Linear Maps

Our first goal is to show that if a linear map L from a vector space U to a vector space V has an inverse M, then M is itself a linear map. Recall that M is the inverse of L if $M \circ L$ is the identity map on U, and $L \circ M$ is the identity map on V.

First note that the dimensions of U and V must be equal, by the Rank-Nullity Theorem. Indeed L must be both injective and surjective. If $\dim U > \dim V$ then L cannot be injective, and if $\dim U < \dim V$, then L cannot be surjective.

Theorem 4.5.1. *If L is a linear map from U to V that has an inverse M, then M is a linear map from V to U*

Proof. As always, we must show that $M(\mathbf{v}_1 + \mathbf{v}_2) = M(\mathbf{v}_1) + M(\mathbf{v}_2)$ for all $\mathbf{v}_1$ and $\mathbf{v}_2$ in V, and that $M(c\mathbf{v}_1) = cM(\mathbf{v}_1)$ for all scalars c.

Because L is invertible, there is a unique $\mathbf{u}_1 \in U$ such that $L(\mathbf{u}_1) = \mathbf{v}_1$, and a unique $\mathbf{u}_2 \in U$ such that $L(\mathbf{u}_2) = \mathbf{v}_2$. Applying M to both equations gives $M(L(\mathbf{u}_1)) = M(\mathbf{v}_1)$, so since $M \circ L = I$, we get $\mathbf{u}_1 = M(\mathbf{v}_1)$ and of course $\mathbf{u}_2 = M(\mathbf{v}_2)$. So

$$M(\mathbf{v}_1 + \mathbf{v}_2) = M(L(\mathbf{u}_1) + L(\mathbf{u}_2)) = M(L(\mathbf{u}_1 + \mathbf{u}_2)) = \mathbf{u}_1 + \mathbf{u}_2 = M(\mathbf{v}_1) + M(\mathbf{v}_2)$$

by the linearity of L. □

Exercise 4.5.2. Provide the second part of the proof of Theorem 4.5.1 by showing that $M(c\mathbf{v}_1) = cM(\mathbf{v}_1)$ for all scalars c.

Definition 4.5.3. A linear map L between two vector spaces U and V that is bijective (therefore both injective and surjective) is called an isomorphism. The vector spaces are then said to be isomorphic.

The previous theorem says that the inverse of L is also a linear map, and therefore also an isomorphism.

Theorem 4.5.4. *Two vector spaces of the same dimension are isomorphic. Vector spaces of different dimensions are not isomorphic.*

Proof. First assume that V and W have the same dimension n. Let $\{\mathbf{v}_1,\dots,\mathbf{v}_n\}$ be a basis for V, and $\{\mathbf{w}_1,\dots,\mathbf{w}_n\}$ a basis for W. By Theorem 4.1.12, we can build a linear map L from V to W sending $\mathbf{v}_i$ to $\mathbf{w}_i$. Because the $\mathbf{w}_i$ are linearly independent, L is injective; because the $\mathbf{w}_i$ span W, L is surjective, so the first statement is proved.

The converse follows from the Rank-Nullity Theorem, as already noted. □

As usual, in our proof we have not handled the trivial special case of vector spaces of dimension 0, which is left to the reader.

Exercise 4.5.5. Let $L\colon U \to U$ be a linear operator such that $L^r = 0$ for some positive r. Then show that $L - I$ is invertible by computing its inverse using the algebra of linear operators.

Exercise 4.5.6. Let $L\colon U \to U$ be a linear operator such that

$$a_n L^n + a_{n-1}L^{n-1} + a_1 L^1 + a_0 I = 0$$

where both a_n and a_0 are different from 0. Then show that L is invertible by displaying its inverse. As we will see in §10.3 we can alway find a polynomial in L that vanishes on L, and L is invertible is and only if its constant term is non-zero. The polynomial of smallest degree on which it vanishes is called the minimal polynomial. We study it in §10.3.

4.6 Projections

Projections are very important examples of linear operators. They are defined in Definition 4.1.7. They play a central role in the rest of this book. For one example see the proof of Theorem 10.5.1.

Before considering projections, let's look at a bigger class of linear operators. First recall that the Rank-Nullity Theorem says that for any subspace U of V on which the restriction of the operator L is injective, and if W is the nullspace of L, then $W \oplus U = V$. But it does not say that $W + L(U) = V$. If this happens to be true, then we say that the range and the nullspace span V. We have the following result.

Lemma 4.6.1. *Consider a linear operator L on V whose range and nullspace span V. Then the range of any power L^k of L is the range of L, and the nullspace of L^k is the nullspace of L*

Proof. Let $\mathbf{v}$ be in the nullspace of L^2. This means that $L\mathbf{v}$ is in the nullspace of L. But by the rank-nullity theorem and the hypothesis, the range and the nullspace of L have only 0 in common. Thus $L\mathbf{v} = 0$, so the nullspace of L and of L^2 are identical. Clearly the range of L^2 is contained in the range of L, but by the rank-nullity theorem and the first part of the proof, they have the same dimension, so they too are the same. We can continue in this way for any power of k. □

This is an extreme case of Definition 4.3.3 applied to L composed with L, as you should verify.

Now we turn to projections, which satisfy the hypothesis of the lemma.

Theorem 4.6.2. *Let $P\colon V \to V$ be the projection to the subspace U along the subspace W, where $V = U \oplus W$, U is the range of P and W its nullspace. Then*

$$P^2 = P.$$

Furthermore any linear operator satisfying this equation is a projection to its range along its nullspace.

Proof. First assume P is the projection to U along W. To any $\mathbf{v} \in V$, which can be written uniquely as $\mathbf{u}+\mathbf{w}$, $\mathbf{u} \in U$, $\mathbf{w} \in W$, $P\mathbf{v} = \mathbf{u}$ by definition. For any $\mathbf{u} \in U$, this unique representation is $\mathbf{u}+\mathbf{0}$. So $P^2\mathbf{v} = P\mathbf{u} = \mathbf{u} = P\mathbf{v}$ as required.

For the converse, just assume we have a linear operator P on V with $P^2 = P$. Obviously you can write any $\mathbf{v} \in V$ as

$$\mathbf{v} = P\mathbf{v} + (\mathbf{v} - P\mathbf{v}). \tag{4.4}$$

By definition $P\mathbf{v}$ is in the range of P, while $\mathbf{v} - P\mathbf{v}$ is in the nullspace of P, since

$$P(\mathbf{v} - P\mathbf{v}) = P\mathbf{v} - P^2\mathbf{v} = P\mathbf{v} - P\mathbf{v} = \mathbf{0}$$

by hypothesis. This shows that any $\mathbf{v}$ can be written as the sum of an element of the nullspace and the range of P. So $\dim V \leq \dim N_L + \dim R_L$. By the rank-nullity theorem we have equality, so V is the direct sum of the range U of P and the nullspace W of P. Thus (4.4) applied to an element in the range of P shows P is the identity map on U, since $\mathbf{v} - P\mathbf{v}$ is then $\mathbf{0}$. □

We can generalize this:

Corollary 4.6.3. *Assume that the vector space V is the direct sum of k subspaces $U_1, \ldots, U_k$. Then for every i we can define the projection $P_i\colon V \to V$ of V to U_i along $\sum_{j\neq i} U_j$. Then*

1. *$P_i \circ P_i = P_i$;*
2. *$P_i \circ P_j = 0$, when $i \neq j$;*
3. *$P_1 + \cdots + P_k = I$.*

Conversely if $P_1, \ldots, P_k$ are a family of linear operators on V such that the three conditions above are met, then letting $U_i = P_i(V)$, V is the direct sum of the U_i.

Proof. First the direct statement. Any $\mathbf{v} \in V$ can be written uniquely as

$$\mathbf{v} = \mathbf{u}_1 + \cdots + \mathbf{u}_k, \quad \mathbf{u}_i \in U_i.$$

Then $P_i(\mathbf{v}) = \mathbf{u}_i$. The conclusions either follow from the theorem or are obvious.

For the converse, pick any $\mathbf{v} \in V$. By hypothesis 3 it can be written

$$\mathbf{v} = P_1(\mathbf{v}) + \cdots + P_k(\mathbf{v}).$$

To show that the sum is direct, we must show this representation as a sum of elements from the U_i is unique. Write one such representation as

$$\mathbf{v} = \mathbf{u}_i + \cdots + \mathbf{u}_k \tag{4.5}$$

with $\mathbf{u}_i \in U_i$. So there is a collection of $\mathbf{w}_i \in V$ such that

$$\mathbf{u}_i = P_i(\mathbf{w}_i) \tag{4.6}$$

since U_i is the range of P_i. Then

$$\begin{aligned}
\mathbf{u}_i &= P_i(\mathbf{w}_i) && \text{by (4.6)} \\
&= P_i \circ P_i(\mathbf{w}_i) && \text{by hypothesis 1} \\
&= \sum_j P_i \circ P_j(\mathbf{w}_j) && \text{by hypothesis 2} \\
&= \sum_j P_i(\mathbf{u}_j)) && \text{by (4.6)} \\
&= P_i(\sum_j \mathbf{u}_j) && \text{by linearity} \\
&= P_i(\mathbf{v}) && \text{by (4.5)}
\end{aligned}$$

which shows that the $\mathbf{u}_j$ are uniquely determined. □

Graph here: 3 dimensional, with projections to the skew coordinate axes.

Example 4.6.4. First a simple example of a linear map L from V to itself where the nullspace and the range do not span the entire space, unlike the situation for projections. V is two-dimensional with basis $\mathbf{u}$ and $\mathbf{v}$. The linear operator L operates by $L(\mathbf{u}) = \mathbf{0}$ and $L(\mathbf{v}) = \mathbf{u}$. Notice that L^2 is the identity map.

Next assume V is three-dimensional, with basis $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$. The operator L acts by $L(\mathbf{u}) = \mathbf{0}$, $L(\mathbf{v}) = \mathbf{w}$, $L(\mathbf{w}) = \mathbf{v}$. Then the nullspace and the range of L span V, and yet the operator is not a projection.

Chapter 5
Representing Linear Maps by Matrices

Abstract In the second chapter of this book, we saw how matrices are used to represent systems of linear equations. In this chapter we see how they are used to represent linear maps between finite dimensional vector spaces. The important computation in the proof of Theorem 5.1.1: (5.3) shows that any linear map is given by matrix multiplication. We first do this for vector spaces with given bases, namely F^n and F^m, using their standard bases, and then in §5.3 we repeat the construction for general vector spaces, showing explicitly how the construction depends on the choice of bases. As part of our construction, we show that the set of linear maps between a vector space of dimension n and one of dimension m is itself a vector space in a natural way, and it has dimension mn. Next we discuss an equivalence relation on linear maps (§5.4) and then a much more important equivalence relation called similarity on linear operators in §5.5. Then we define the row rank and the column rank of a matrix, in terms of the rank of a linear map. We prove the important theorem that the row rank and the column rank are equal. We also give the classical definition of the rank of a matrix as the size of its biggest invertible submatrix. Next we apply the notion of rank to the matrix of coefficients to reformulate the theory of linear equations already studied in Chapter 2. After a section describing the relation between real and complex linear maps, we conclude with an example: nilpotent operators and their matrices.

5.1 The Matrix of a Linear Map

We start with a linear transformation $L\colon V \to W$, where V and W are vector spaces of dimension n and m respectively. Assume V and W are equipped respectively with bases

$$\mathfrak{B} = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\} \text{ and } \mathfrak{C} = \{\mathbf{f}_1, \ldots, \mathbf{f}_m\}.$$

To every linear map $L\colon V \to W$ with these bases, we associate a unique $m \times n$ matrix A, as follows. The j-th coordinate vector $\mathbf{e}_j$ is mapped by L to a linear

combination of the basis vectors of W, that we write

$$\boxed{L(\mathbf{e}_j) = a_{1j}\mathbf{f}_1 + a_{2j}\mathbf{f}_2 + \cdots + a_{mj}\mathbf{f}_m} \tag{5.1}$$

for each j, $1 \leq j \leq n$. This defines the scalars a_{ij}, $1 \leq i \leq m$ and $1 \leq j \leq n$. We let A be the $m \times n$ matrix with entries (a_{ij}). The matrix A is uniquely determined by L and the two bases.

Theorem 4.1.12 of Chapter 4 shows that to specify a linear map uniquely we only need to know what it does on a basis. So given a linear map $L\colon V \to W$, use the basis $\mathfrak{B}$ of V. An arbitrary vector $\mathbf{x}$ in V can be written uniquely as

$$\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n. \tag{5.2}$$

By the linearity of L, the image of (5.2) under L is

$$L(\mathbf{x}) = x_1L(\mathbf{e}_1) + x_2L(\mathbf{e}_2) + \cdots + x_nL(\mathbf{e}_n).$$

Now we use (5.1) to get:

$$\begin{aligned} L(\mathbf{x}) &= x_1L(\mathbf{e}_1) + x_2L(\mathbf{e}_2) + \cdots + x_nL(\mathbf{e}_n) \\ &= x_1(a_{11}\mathbf{f}_1 + a_{21}\mathbf{f}_2 + \cdots + a_{m1}\mathbf{f}_m) + x_2(a_{12}\mathbf{f}_1 + a_{22}\mathbf{f}_2 + \cdots + a_{m2}\mathbf{f}_m) \\ &+ \cdots + x_n(a_{1n}\mathbf{f}_1 + a_{2n}\mathbf{f}_2 + \cdots + a_{mn}\mathbf{f}_m) \\ &= (x_1a_{11} + x_2a_{12} + \cdots + x_na_{1n})\mathbf{f}_1 + (x_1a_{21} + x_2a_{22} + \cdots + x_na_{2n})\mathbf{f}_2 \\ &+ \cdots + (x_1a_{m1} + x_2a_{m2} + \cdots + x_na_{mn})\mathbf{f}_m \\ &= y_1\mathbf{f}_1 + y_2\mathbf{f}_2 + \cdots + y_m\mathbf{f}_m, \end{aligned} \tag{5.3}$$

where we have defined:

$$y_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n. \tag{5.4}$$

Since A is the $m \times n$ matrix with entries (a_{ij}), (5.4) is the matrix product of the i-th row of A with the column vector $\mathbf{x}$:

$$y_i = \mathbf{a}^i\mathbf{x}, \quad 1 \leq i \leq m.$$

So the computation (5.3) establishes the important:

Theorem 5.1.1. *Let L be the linear map from V to W defined by* (5.1)*. Then L maps the vector $\mathbf{x}$ with coordinates $(x_1, x_2, \ldots, x_n)$ in the $\mathfrak{B}$ basis to the vector $\mathbf{y}$ with coordinates $(y_1, y_2, \ldots, y_m)$ in the $\mathfrak{C}$ basis, where*

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

Definition 5.1.2. If V is a vector space of dimension n with basis $\mathfrak{B}$ and W is a vector space of dimension m with basis $\mathfrak{C}$, $1 \leq i \leq m$, and L a linear map between V and W defined by (5.1), then the $m \times n$ matrix A is the *matrix associated to L.*

5.2 The Linear Map of a Matrix

Conversely, given a $m \times n$ matrix A with coefficients in F, we showed in Example 4.1.10 how to build a linear map $L\colon F^n \to F^m$ using A: just map the n-tuple $\mathbf{v} \in F^n$ to the matrix product $A\mathbf{v}$, where $\mathbf{v}$ is a column vector. We call this linear map L_A when we need to make the dependence on A explicit, and we call it the linear map associated to A. Since we can write

$$A\mathbf{v} = v_1\mathbf{a}_1 + v_2\mathbf{a}_2 + \cdots + v_n\mathbf{a}_n,$$

where $\mathbf{a}_i$ is the i-th column of A, L_A takes the i-th coordinate vector of F^n to the vector $\mathbf{a}_i$. This is the inverse of the map constructed in Theorem 5.1.1.

Recall that $M_{m,n}$ is the vector space of all $m \times n$ matrices (see Chapter 3, Example 3.2.4) and $\mathscr{L}(F^n, F^m)$ is the vector space of linear maps from F^n to F^m: see Theorem 4.1.11 of Chapter 4. We know that $M_{m,n}$ has dimension mn.

Above we have constructed a bijection:

$$\mathscr{F}\colon M_{m,n} \to \mathscr{L}(F^n, F^m). \tag{5.5}$$

Even more is true.

Theorem 5.2.1. *The map $\mathscr{F}\colon M_{m,n} \to \mathscr{L}(F^n, F^m)$ of* (5.5) *is a linear map of vector spaces.*

Proof. Let A and B be two $m \times n$ matrices, and c a scalar. We need to show two things. They both follow from Theorem 2.2.13.

1. $\mathscr{F}(cA) = c\mathscr{F}(A)$.
   Applying the definition of $\mathscr{F}$, this gives $L_{cA} = cL_A$, which is clear, since $(cA)\mathbf{v} = c(A\mathbf{v})$ for all $\mathbf{v} \in F^n$.
2. $\mathscr{F}(A+B) = \mathscr{F}(A) + \mathscr{F}(B)$.
   Again applying the definition of $\mathscr{F}$, this gives $L_{A+B} = L_A + L_B$. This follows from $(A+B)\mathbf{v} = A\mathbf{v} + B\mathbf{v}$.

□

Therefore $\mathscr{F}$ is a bijective linear map, so by Definition 4.5.3 is an isomorphism. So the dimension of $\mathscr{L}(F^n, F^m)$ is that of M_{mn}:

Theorem 5.2.2. *The dimension of $\mathscr{L}(F^n, F^m)$ is mn.*

Theorem 5.2.3. *If L is the linear map from F^n to F^m with associated matrix A, and M the linear map from F^m to F^s with matrix B, then the matrix associated to the composite linear map $M \circ L$ is the matrix product BA.*

We first check that the matrix sizes work out. A has size $m \times n$, and B has size $s \times m$. Thus the product BA can be formed, and has size $s \times n$, the appropriate size for the matrix associated to a linear map from F^n to F^s.

Proof. We do this without computation. Let $\mathbf{x}$ be an n-vector. Then

$$(M \circ L)(\mathbf{x}) = M(L(\mathbf{x})) = B(A\mathbf{x}) = (BA)\mathbf{x}. \tag{5.6}$$

We use the associativity of composition of maps (Theorem B.1.5) in the second step, and the associativity of matrix multiplication in the fourth step (Theorem 2.2.10). In the third step we use Theorem 5.1.1 applied first to L to get $M(L(\mathbf{x})) = M(A\mathbf{x})$, and then to M evaluated at $\mathbf{y} = A\mathbf{x}$ to get $M(\mathbf{y}) = B\mathbf{y} = BA\mathbf{x}$. □

5.3 Change of Basis

In the previous sections we only consider linear maps from F^n to F^m, or, which amounts to the same thing, we assumed that the domain V and the target W were equipped with bases, which allow us to identity them with F^n and F^m respectively. In this section we reword the results of §5.1 in a new notation that gives the flexibility to change bases.

First we develop the notation. Let L be a linear map between V and W of dimension n, with bases, respectively $\mathfrak{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\mathfrak{C} = \{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$. Using the notation of Example 4.1.9, we write the vector of coordinates of $\mathbf{v}$ in the basis $\mathfrak{B}$ of V as $[\mathbf{v}]_{\mathfrak{B}}$, and the vector of coordinates of $\mathbf{w} \in W$ in the $\mathfrak{C}$ basis as $[\mathbf{w}]_{\mathfrak{C}}$. Then Theorem 5.1.1 says that the matrix A associated to L in these basis satisfies

$$[\mathbf{w}]_{\mathfrak{C}} = A[\mathbf{v}]_{\mathfrak{B}}.$$

To emphasize the dependence of A on the two bases, we write $A = [L]_{\mathfrak{C}}^{\mathfrak{B}}$. In the notation $[L]_{\mathfrak{C}}^{\mathfrak{B}}$, the basis $\mathfrak{B}$ of the domain is written as the superscript, while the basis $\mathfrak{C}$ of the target space is written as the subscript.

Remark 5.3.1. There is no generally accepted notation for the $n \times m$ matrix $[L]_{\mathfrak{C}}^{\mathfrak{B}}$ associated to a linear transformation. The notation we use here is adapted from [12], Chapter 3.4. See in particular Theorem 11.

In our new notation Theorem 5.1.1 says

$$[L(\mathbf{v}]_{\mathfrak{C}} = [L]_{\mathfrak{C}}^{\mathfrak{B}}\, [\mathbf{v}]_{\mathfrak{B}} \tag{5.7}$$

where

$$[L]_{\mathfrak{C}}^{\mathfrak{B}} = \begin{pmatrix} | & \cdots & | \\ [L(\mathbf{v}_1)]_{\mathfrak{C}} & \cdots & [L(\mathbf{v}_n)]_{\mathfrak{C}} \\ | & \cdots & | \end{pmatrix}$$

meaning that the i-th column of $[L]_{\mathfrak{C}}^{\mathfrak{B}}$ is $[L(\mathbf{v}_i)]_{\mathfrak{C}}$.

Example 5.3.2. For the identity map I on a vector space V with basis $\mathfrak{B}$, we get the identity matrix I:

$$[I]_{\mathfrak{B}}^{\mathfrak{B}} = I. \tag{5.8}$$

We record the following important special case of our computation (5.7).

Corollary 5.3.3 (Change of Basis). *Let V be an n-dimensional vector space with two bases $\mathfrak{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\mathfrak{C} = \{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$. Then*

$$[\mathbf{v}]_{\mathfrak{C}} = [I]_{\mathfrak{C}}^{\mathfrak{B}} [\mathbf{v}]_{\mathfrak{B}} . \tag{5.9}$$

Proof. In the computation above let W be the same space as V, let L be the identity map. Therefore $\mathfrak{B}$ and $\mathfrak{C}$ are two bases of V. □

The matrix $[I]_{\mathfrak{C}}^{\mathfrak{B}}$ is the *change of basis* matrix from the $\mathfrak{B}$ basis to the $\mathfrak{C}$ basis. Its j-th column is formed by the coefficients of $\mathbf{v}_i$ expressed in the $\mathfrak{C}$ basis, namely $[\mathbf{v}_i]_{\mathfrak{C}}$. The matrix $[I]_{\mathfrak{C}}^{\mathfrak{B}}$ is invertible with inverse $[I]_{\mathfrak{B}}^{\mathfrak{C}}$ as we see by exchanging the roles of the two bases. For more details see Corollary 5.3.7.

Example 5.3.4. A linear map is often given by equations such as (5.3). For example consider the linear map L from a 2-dimensional space to a 3-dimensional space that maps the bases as follows:

$$\begin{aligned} L(\mathbf{v}_1) &= 3\mathbf{w}_1 + 2\mathbf{w}_2 + \mathbf{w}_3 \\ L(\mathbf{v}_2) &= -\mathbf{w}_1 + 4\mathbf{w}_2 + 5\mathbf{w}_3 \end{aligned}$$

The matrix $A = [L]_{\mathfrak{C}}^{\mathfrak{B}}$ is

$$\begin{pmatrix} 3 & -1 \\ 2 & 4 \\ 1 & 5 \end{pmatrix}$$

so that the vector $x_1\mathbf{v}_1 + x_2\mathbf{v}_2$ gets mapped to the vector with coordinates:

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

as required.

We now rephrase Theorem 5.2.1 in this new notation .

Theorem 5.3.5. *Let V be a vector space of dimension n with basis $\mathfrak{B}$ and W a vector space of dimension m with basis $\mathfrak{C}$. Let L and M be linear maps from V to W. Then*

$$[L+M]_{\mathfrak{C}}^{\mathfrak{B}} = [L]_{\mathfrak{C}}^{\mathfrak{B}} + [M]_{\mathfrak{C}}^{\mathfrak{B}}$$

and

$$[cL]_{\mathfrak{C}}^{\mathfrak{B}} = c\,[L]_{\mathfrak{C}}^{\mathfrak{B}} .$$

So $[\bullet]_{\mathfrak{C}}^{\mathfrak{B}} : \mathscr{L}(V,W) \to M(m,n)$ is a linear map from the vector space of all linear maps $\mathscr{L}(V,W)$ to $m \times n$ matrices. Furthermore it is an isomorphism.

Next we turn to the multiplicative properties of the associated matrix. Here is the key result, which rephrases Theorem 5.2.3. It will allow us to write down the most general change of basis formula.

Theorem 5.3.6. *Let V, W and U be vector spaces of dimensions n, m and r. Let $\mathfrak{B}$, $\mathfrak{C}$, $\mathfrak{D}$ be bases for V, W, U respectively. Let*

$$L\colon V \to W \text{ and } M\colon W \to U$$

be linear maps. Then

$$[M]_{\mathfrak{D}}^{\mathfrak{C}}\,[L]_{\mathfrak{C}}^{\mathfrak{B}} = [M \circ L]_{\mathfrak{D}}^{\mathfrak{B}}\,, \tag{5.10}$$

where the left hand side is the matrix product of a matrix of size $r \times m$ by a matrix of size $m \times n$.

Warning: the theorem only applies when the basis $\mathfrak{C}$ of the vector space W in the middle is the same for both associated matrices.

Proof. We identity V, W and U with F^m, F^n and F^r respectively, once and for all, using the given bases, and we take the associated matrices in these bases. Then we write down (5.6) in our new notation. This is (5.10). □

Corollary 5.3.7. *Let V be a vector space and $\mathfrak{B}$ and $\mathfrak{C}$ two bases. Denote by I both the identity linear transformation and the identity matrix. Then*

$$[I]_{\mathfrak{C}}^{\mathfrak{B}}\,[I]_{\mathfrak{B}}^{\mathfrak{C}} = I = [I]_{\mathfrak{B}}^{\mathfrak{C}}\,[I]_{\mathfrak{C}}^{\mathfrak{B}}\,.$$

In particular $[I]_{\mathfrak{C}}^{\mathfrak{B}}$ is invertible.

Proof. Just let $V = U = W$, $L = M = I$, and $\mathfrak{D} = \mathfrak{B}$ in the theorem. Use (5.8), and we are done. □

Example 5.3.8. Return to the original setup of a linear map $L\colon V \to W$, where $\dim V = n$ with basis $\mathfrak{B}$ and $\dim W = m$ with basis $\mathfrak{C}$. Let I_n be the identity linear transformation on V, which now has a second basis $\mathfrak{D} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$. We consider I_n as a linear transformation from V with the basis $\mathfrak{D}$ to V with the basis $\mathfrak{B}$.

Then by Theorem 5.3.6,

$$[L]_{\mathfrak{C}}^{\mathfrak{D}} = [L \circ I_n]_{\mathfrak{C}}^{\mathfrak{D}} = [L]_{\mathfrak{C}}^{\mathfrak{B}}\,[I_n]_{\mathfrak{B}}^{\mathfrak{D}}\,.$$

This says explicitly how the matrix representing L changes when you change the basis of the domain: you multiply on the right by the change of basis matrix in the domain. Thus we multiply a $m \times n$ matrix on the right by a $n \times n$ matrix.

Exercise 5.3.9. In the same way, give W a second basis $\mathfrak{E} = \{\mathbf{u}^1, \ldots, \mathbf{u}^m\}$, and let I_m be the identity linear transformation on W. Then show that

$$[L]_{\mathfrak{E}}^{\mathfrak{B}} = [I_m]_{\mathfrak{E}}^{\mathfrak{C}}\,[L]_{\mathfrak{C}}^{\mathfrak{B}}$$

Finally, using the notation and the conclusion of Example 5.3.8 and Exercise 5.3.9, we get:

Theorem 5.3.10.

$$[L]_{\mathfrak{E}}^{\mathfrak{D}} = [I_m]_{\mathfrak{E}}^{\mathfrak{C}} [L]_{\mathfrak{C}}^{\mathfrak{B}} [I_n]_{\mathfrak{B}}^{\mathfrak{D}}$$

Notice that the right hand side is the product of a $m \times m$ matrix by a $m \times n$ matrix and a $n \times n$ matrix, which makes sense. By Corollary 5.3.7 the two change of basis matrices are invertible.

Numerical examples here.

In Theorem 5.5.1 we handle the most important case, where T is a linear transformation $V \to V$.

5.4 Equivalent Linear Maps

Suppose that $L\colon V \to W$ is a linear map of a n dimensional vector space V to a m dimensional vector space W. We allow ourselves to make an arbitrary change of basis in V, and an independent change of basis in W in order to simplify the matrix A of L in these bases. By Theorem 5.3.10 and Corollary 5.3.7 we see that we are allowed to multiply A by an arbitrary invertible $n \times n$ matrix D on the right and an arbitrary invertible $m \times m$ matrix C on the left. This section shows how to get the maximum simplification without computation, using the ideas in the proof of the Rank-Nullity Theorem.

If the rank of L is r, we can find linearly independent elements $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ whose images under L are linearly independent elements $\mathbf{w}_i = L(\mathbf{v}_i)$ that form a basis for the image of L. Then we can find elements $\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n$ in the nullspace of L so that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is a basis of V. Finally complete the $\mathbf{w}_i$ to a basis of W arbitrarily. It is clear that with this choice of bases, the matrix of L has r ones along the diagonal, and zeroes everywhere else. So we have proved:

Theorem 5.4.1. *There is a basis $\mathfrak{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of V and a basis $\mathfrak{E} = \{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ of W so that*

$$[L]_{\mathfrak{E}}^{\mathfrak{B}} = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \end{pmatrix} = \begin{pmatrix} I_r & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix} \tag{5.11}$$

where on the right we have written the block decomposition of the matrix into one identity matrix and three 0 *matrices.*

Definition 5.4.2. Given two $m \times n$ matrices A and B, we say that A is row-column equivalent to B if there are invertible matrices C of size m and D of size n such that $B = CAD$. We write $A \approx B$ if this is the case.

Theorem 5.4.3. *Row-column equivalence ($\approx$) is an equivalence relation on $m \times n$ matrices.*

Theorem 5.4.4. *Two $m \times n$ matrices are row-column equivalent if and only if they are the matrices associated to the same linear operator L in different bases for the domain and the codomain.*

Proof. Theorem 5.4.1 can be reformulated to say that any two $m \times n$ matrices associated to the same linear operator L in different bases for the domain and the codomain are row-column equivalent. On the other hand, any invertible matrix is a change of basis matrix, which gives the other inclusion. □

This shows that there are very few equivalence classes for row-column equivalence: they are characterized by the matrices (5.11), and therefore by the rank of the matrix. Note that we did not use the Rank-Nullity Theorem in the proof. However, the proof of Theorem 5.4.1 is essentially equivalent to that of the Rank-Nullity Theorem, so not much is gained.

Remark 5.4.5. There is a more interesting equivalence relation on $m \times n$ matrices that we will study later, once we give both spaces inner products, and have defined the notion of an orthogonal matrix (over $\mathbb{R}$) and unitary matrix (over $\mathbb{C}$). Here we will just deal with the real case. Then for any $m \times n$ matrix A, there is an orthogonal $m \times m$ matrix U and an orthogonal $n \times n$ matrix W so that $B = UAW$ is a $m \times n$ matrix whose only non zero elements are $\sigma_i = b_{ii}$, which can be arranged in weakly decreasing order:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$$

where $p = \min\{m,n\}$. This is called the Singular Value Decomposition of A, or SVD, since the σ_i are called the singular values. We will study it in 13.10 In particular we will show that the singular values partition $m \times n$ matrices into equivalence classes.

A similar, even more interesting decomposition holds for complex matrices.

For a good example of this material see Example 6.4.8 about the map for a vector space to its dual.

5.5 Equivalent Linear Operators

We now restrict to the case where L is a linear operator, a map from a vector space V to itself. In this case we require that the same basis be used on the domain and the range of L. In particular, if $\mathfrak{B}$ is a basis of V, we only consider associated matrices $[L]^{\mathfrak{B}}_{\mathfrak{B}}$.

Given another basis $\mathfrak{D}$ of V we want to compare $[L]^{\mathfrak{B}}_{\mathfrak{B}}$ and $[L]^{\mathfrak{D}}_{\mathfrak{D}}$. This is a special case of Theorem 5.3.10:

Theorem 5.5.1. *Let $L\colon V \to V$ be a linear operator, and let $\mathfrak{B}$ and $\mathfrak{D}$ be bases of V. Then there is an invertible matrix N such that*

$$[L]^{\mathfrak{D}}_{\mathfrak{D}} = N^{-1}\,[L]^{\mathfrak{B}}_{\mathfrak{B}}\,N$$

Proof. Indeed take $N = [I_n]^{\mathfrak{D}}_{\mathfrak{B}}$, which has inverse $[I_n]^{\mathfrak{B}}_{\mathfrak{D}}$ by Corollary 5.3.7. Then the right hand side is

$$[I_n]^{\mathfrak{B}}_{\mathfrak{D}}\,[L]^{\mathfrak{B}}_{\mathfrak{B}}\,[I_n]^{\mathfrak{D}}_{\mathfrak{B}} = [L]^{\mathfrak{D}}_{\mathfrak{D}}$$

by Theorem 5.3.10, □

Now we introduce some new terminology that will help us compare square matrices:

Definition 5.5.2. Given two $n \times n$ matrices A and B, we say that A is similar to B if there is an invertible $n \times n$ matrix N such that $B = N^{-1}AN$. We write $A \sim B$ if this is the case.

Theorem 5.5.3. *Similarity ($\sim$) is an equivalence relation on $n \times n$ matrices.*

Proof. By the definition of equivalence relation - see Chapter 1 - we need to establish the following three points:

1. $A \sim A$: Use the identity matrix for N.
2. if $A \sim B$, then $B \sim A$: If $A \sim B$, there is an invertible N such that $B = N^{-1}AN$. Then, multiplying both sides of the equation on the right by N^{-1} and on the left by N, and letting $D = N^{-1}$, we see that $A = D^{-1}BD$, so B is similar to A.
3. if $A \sim B$ and $B \sim D$, then $A \sim D$: The hypotheses mean that there are invertible matrices C_1 and C_2 such that $B = C_1^{-1}AC_1$ and $D = C_2^{-1}BC_2$, so, substituting from the first equation into the second, we get

$$D = C_2^{-1}C_1^{-1}AC_1C_2 = (C_1C_2)^{-1}A(C_1C_2),$$

so A is similar to D using the matrix C_1C_2.

□

Since similarity is an equivalence relation on $n \times n$ matrices, it partitions these matrices into equivalence classes.

Theorem 5.5.1 says is that two matrices that represent the same linear operator $F\colon V \to V$ in different bases of V are similar. We have an easy converse:

Theorem 5.5.4. *Assume that two $n \times n$ matrices A and B are similar, so $B = N^{-1}AN$, for an invertible matrix N. Then they represent the same linear operator L.*

Proof. Choose an n dimensional vector space V, a basis $\mathfrak{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ for V. Let L be the linear map represented by A in the $\mathfrak{B}$ basis, so that $B = [L]^{\mathfrak{B}}_{\mathfrak{B}}$. Construct a second basis $\mathfrak{D} = \{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ of V:

$$\mathbf{w}_j = n_{1j}\mathbf{v}_1 + n_{2j}\mathbf{v}_2 + \cdots + n_{nj}\mathbf{v}_n, \tag{5.12}$$

using the entries of the matrix $N = (n_{ij})$. This is possible because the matrix N is invertible, so we do get a basis. Then by definition $N = [I_n]^{\mathfrak{D}}_{\mathfrak{B}}$.

Then by Corollary 5.3.7

$$B = [I_n]^{\mathfrak{B}}_{\mathfrak{D}}\,[L]^{\mathfrak{B}}_{\mathfrak{B}}\,[I_n]^{\mathfrak{D}}_{\mathfrak{B}} = N^{-1}AN$$

as required. □

Any $n \times n$ matrix A can be viewed as the matrix $[L]^{\mathfrak{B}}_{\mathfrak{B}}$ of a linear operator L in the basis $\mathfrak{B}$ of the n-dimensional vector space V. Matrices in the same similarity class correspond to the same linear operator L, but expressed in different bases. One of the goals of the remainder of this course is to determine the common features of the matrices in a given similarity class. For example we will show that similar matrices have the same characteristic polynomial: see Chapter 12. We will also see that two matrices that have the same characteristic polynomial need not be similar: see Theorem 12.7.2. The simplest example is given by the matrices

$$\begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix} \text{ and } \begin{pmatrix} \alpha & 1 \\ 0 & \alpha \end{pmatrix}$$

for any complex number α.

Since our main goal is to study linear transformations L, not matrices, which are computational tools for understanding linear transformations, we will want to choose a basis in which the matrix of L is as simple as possible.

Exercise 5.5.5. Show that row equivalence (see Theorem 2.5.3) is an equivalence relation on $n \times m$ matrices.

We will study an equivalence relation on symmetric matrices called congruence in Definition 7.1.10.

5.6 The Rank of a Matrix

Recall that the rank of a linear transformation is the dimension of its range. We can now define the rank of a matrix. First we define the column rank and the row rank, and then we show they are equal: this is the rank of the matrix.

Definition 5.6.1. Let A be a $m \times n$ matrix.

1. The columns $\mathbf{a}_j$, $1 \le j \le n$ generate a subspace C_A in F^m, whose dimension c is the *column rank* of A.
2. Correspondingly, the rows $\mathbf{a}^i$, $1 \le i \le m$, of A generate a subspace R_A in F^n, whose dimension r is called the *row rank* of A.

The dimension c of C_A is at most m, since it is a subspace of F^m, and at most n, since it is generated by n elements. Similarly, the dimension r of R_A is at most n, since it is a subspace of F^n, and at most m, since it is generated by m elements.

Theorem 5.6.2. *If L_A is the linear map associated to A, then the rank of L_A is the column rank of A.*

Proof. Since any vector in F^n can be written $a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + \cdots + a_n\mathbf{e}_n$ for suitable scalars a_j and the standard unit coordinates vectors in F^n, any vector in the range can be written

$$a_1L(\mathbf{e}_1) + a_2L(\mathbf{e}_2) + \cdots + a_nL(\mathbf{e}_n).$$

On the other hand, matrix multiplication tells us that $L(\mathbf{e}_j)$ is the j-th column of A. So the range of L_A is spanned by the columns of A, so the rank of L_A is the column rank of A. □

Examples here.

Theorem 5.6.3. *Let A be an $n \times n$ matrix. Then A is an invertible matrix if and only if the linear map L_A is invertible.*

Proof. By definition A is invertible if and only if there is a $n \times n$ matrix B such that $AB = BA = I$. Then Theorem 5.2.3 says that $L_A \circ L_B = I$ and $L_B \circ L_A = I$, which says that L_A and L_B are inverse linear maps. □

Corollary 5.6.4. *Let A be an $n \times n$ matrix with columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$. Then A is invertible if and only if its columns are linearly independent.*

Proof. The key point is that

$$L_A(\mathbf{x}) = A\mathbf{x} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n.$$

So if the $\mathbf{a}_i$ are not linearly independent, we can find an element $\mathbf{x}$ in the nullspace of L_A, a contradiction. Conversely if the $\mathbf{a}_i$ are linearly independent, then the range of L_A has dimension n, so it is surjective and therefore an isomorphism by Corollary 4.2.9, for example. □

We can rework the proof of the corollary slightly. In Theorem 5.6.2 we show that the column rank of A is the rank of L_A. By the Rank-Nullity Theorem we know that if L_A is a linear map between vector spaces of size n, then it is invertible if and only if its rank is n. Then A is invertible by Theorem 5.6.3. So the corollary also follows from Theorem 5.6.2.

Theorem 5.6.5. *The row rank and the column rank of any matrix A are equal. We call this simply the rank of A.*

Proof. We can extract a basis of R_A from the rows $\mathbf{a}^i$ of A by Proposition 3.4.14. So by reordering the rows of the equations, we may assume that the first r rows of A form a basis of R_A. We only do this for convenience of notation only, as we

show in the exercise below. Let R be the $r \times n$ matrix formed by these rows. Then by definition of a basis, any row $\mathbf{a}^i$ of A can be written as a linear combination of the first r rows:

$$\mathbf{a}^i = b_{i1}\mathbf{a}^1 + b_{i2}\mathbf{a}^2 + \cdots + b_{ir}\mathbf{a}^r,\ 1 \leq i \leq m. \tag{5.13}$$

Let $B = (b_{ij})$ the $m \times r$ matrix formed by the scalars in (5.13). Note that for $i = 1$ to r we have $b_{ii} = 1$ and all the other entries in the first r rows are zero. Then as we noted in Proposition 2.2.7, (5.13) can be expressed by the matrix product: $A = BR$. The rows of A are linear combinations of the rows of R. On the other hand the same product says that the columns of A are linear combinations of the columns of B. Since B has r columns, the column rank of A is at most r, so $c \leq r$.

Now we just repeat the argument on the transpose A^t of A. Thus $r \leq c$, since the column rank of A^t is the row rank of A, etc. The two inequalities together give the conclusion. □

We can also get this result as a corollary of Theorem 5.4.4. Here is the proof.

Proof. The row and column ranks are obvious equal for the matrix in (5.11). On the other hand both the row rank and the column rank of a matrix A are properties of the linear transformation L_A. Indeed, the row rank is the dimension of the domain minus the dimension of the kernel of L_A, while the column rank is the dimension of the image of L_A. Since the bases of the domain and image of L can be chosen so that the matrix of L is these basis is the one in (5.11), the result is true for any matrix. □

Exercise 5.6.6. Here is the equivalent of (5.13) if a collection of rows numbered j_1, j_2, …, j_r form a basis for the row space of A. Then 5.13 becomes

$$\mathbf{a}^i = b_{ij_1}\mathbf{a}^{j_1} + b_{ij_2}\mathbf{a}^{j_2} + \cdots + b_{ij_r}\mathbf{a}^{j_r},\ 1 \leq i \leq m. \tag{5.14}$$

Let R be the $r \times n$ matrix whose i-th row, $1 \leq i \leq r$ is $\mathbf{a}^{j_i}$. Let C be the $m \times r$ matrix whose i-th column, $1 \leq i \leq r$ is the j_i-th column of the matrix B defined by the (b_{ij_k}) in (5.14). Then $A = CR$. Now conclude as before.

Exercise 5.6.7. Write down an $m \times n$ matrix, with $m > n$ which has several different collections of n rows that are linearly independent.

This exercise is important because it shows that even though it is convenient to give preference to the rows and columns with smaller indices, as we did in the proof of Theorem 5.6.5, with a little bit of extra indexing work it is possible to understand the situation without resorting to this preference. Row reduction is a key example where preference is given to the first row and first column of a matrix. The notation for submatrices of a matrix in (2.13) is a good example of the extra indexing work required. This indexing work can be represented by matrix multiplication by elementary matrices.

Exercise 5.6.8. Write down an 4×2 matrix (a_{ij}) whose bottom two rows are linearly independent. Find a product of elementary matrix that makes those rows the two top rows. What happens to the columns of the original matrix?

5.7 More on Linear Equations

Using the notion of rank of a matrix, we can revisit systems of linear equations and improve our results from §2.7. We first examine the homogeneous system $A\mathbf{x} = \mathbf{0}$, where A is as always a $m \times n$ matrix. The solutions of this system are simply the nullspace of the linear transformation L_A generated by the matrix A according to Example 4.1.10.

As noted in Theorem 2.7.1, the homogeneous system $A\mathbf{x} = \mathbf{0}$ can be written

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \mathbf{0},$$

where the columns $\mathbf{a}_j$ of A are vectors in F^m. So a nontrivial (i.e., a non-zero) solution $(x_1, \ldots, x_n)$ gives an equation of linear dependence between the vectors $\mathbf{a}_j$ in F^m. This leads us to:

Theorem 5.7.1. *If A is a $m \times n$ matrix and $N_A \subset F^n$ the subspace of solutions to the equation $A\mathbf{x} = \mathbf{0}$, then:*

1. *If $n > m$, N_A is positive dimensional, so there are nontrivial solutions.*
2. *If $m = n$, then N_A is zero dimensional (so the only solution is $\mathbf{0}$) if and only if the columns $\mathbf{a}_j$ form a basis of F^m, so that A has rank n and is nonsingular.*

Proof. By dimension considerations, if $n > m$ a nontrivial solution always exists: since F^m has dimension m, for $n > m$ a collection of n vectors cannot be linearly independent.

The second part just expresses the fact that m vectors in a m-dimensional space form a basis if and only if they are linearly independent. □

Now suppose that A has been row reduced. Discard the rows of the row reduced matrix that consist entirely of zeroes to get a $r \times n$ matrix R. By Theorem 2.5.11, the solutions of the system $A\mathbf{x} = \mathbf{0}$ are the same as those of the system $R\mathbf{x} = \mathbf{0}$. If there are no equations, the entire domain of A are solutions, so the dimension of the space of solutions is n. The next result is important and easy to remember: because the rows of R are linearly independent each equation in R imposes an independent condition of the solutions, making the dimension of the space of solutions go down by 1. So using Definition 2.6.1 we have $n - r$ free variables corresponding to the free columns. The following proposition finishes the proof.

Proposition 5.7.2. *We get a $n - r$ dimensional space of solutions to the system of equations $R\mathbf{x} = \mathbf{0}$ with basis $\mathbf{v}^j$, $1 \leq j \leq n - r$, obtained by letting the j-th free variable entry in $\mathbf{v}^j$ take the value 1, letting all the other free variable entries in $\mathbf{v}^j$ take the value 0, and solving for the bound variables.*

Proof. It is clear that the solutions $\mathbf{v}^j$ are linearly independent: imitate the proof in Example 3.4.4. So the only difficulty is showing that they span. Take any solution $\mathbf{w}$. By subtracting from it an appropriate linear combination of the solutions $\mathbf{v}^j$, we get a solution where all the free variable entries are 0. Then looking at the last equation and working up, we see that all the bound variables are 0, so we are done. □

Corollary 5.7.3. *The space of solutions of $A\mathbf{x} = \mathbf{0}$ has dimension $n - r$.*

This is just the Rank-Nullity Theorem.

In particular, no matter which method we use to row reduce, we will end up with the same number r of non-zero rows, since r is the dimension of the row space of A.

Definition 5.7.4. The augmented matrix associated to the system is the $m \times (n+1)$ matrix C whose first n columns are those of A, and whose last column is $\mathbf{b}$.

Theorem 5.7.5. *The system has a solution if and only if the rank of C is the same as that of A.*

The proof is left to you. Just phrase Theorem 2.7.2 in terms of rank. Note also that when the system is homogeneous, the rank of C is clearly that of A.

Next we look at the inhomogeneous equation

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

where A be a $m \times n$ matrix, and $\mathbf{b}$ a m-column vector. We started studying this equation in §2.7

Theorem 2.7.2 says that the inhomogeneous equation $A\mathbf{x} = \mathbf{b}$ can be solved if and only if any equation of linear dependence satisfied by all the rows, namely $\mathbf{y}^t A = \mathbf{0}$, implies the same linear relation between the right hand terms: $\mathbf{y}^t \mathbf{b} = 0$.

Example 5.7.6. Now a 3×3 example. We want to solve the system:

$$\begin{aligned} x_1 - x_2 &= b_1 \\ x_2 - x_3 &= b_2 \\ x_3 - x_1 &= b_3 \end{aligned}$$

So

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{pmatrix}.$$

Now A has rank 2, so up to a scalar, there is only one non-zero vector y such that $y^t A = 0$. To find y add the three equations. We get

$$0 = b_1 + b_2 + b_3.$$

This says that the scalar product of $(1,1,1)$ with b is 0. So by the theorem the system has a solution for all b such that $b_1 + b_2 + b_3 = 0$.

Let's work it out. Write $b_3 = -b_1 - b_2$. Then the third equation is a linear combination of the first two, so can be omitted. It is sufficient to solve the system:

$$\begin{aligned} x_1 - x_2 &= b_1 \\ x_2 - x_3 &= b_2 \end{aligned}$$

x_3 can be arbitrary, and then $x_2 = x_3 + b_2$ and

$$x_1 = x_2 + b_1 = x_3 + b_1 + b_2$$

so the system can be solved for any choice of x_3.

Remark 5.7.7. Here is what happens when one does Gaussian elimination on the inhomogeneous system, using the augmented matrix 5.7.4. Reduce the coefficient matrix A to row echelon form C, getting an equivalent system

$$C\mathbf{x} = \mathbf{d}$$

The matrix C may have rows of zeroes - necessarily the last rows. Assume it has p rows of zeroes. Then for the new system to have a solution, the last p components of $\mathbf{d}$ must be 0. The rank of C is $m-p$: just imitate the proof of Proposition 5.7.2. This can be at most n, since the row rank is the column rank. Then by the Rank-Nullity Theorem, any vector *vecd* whose last p components are 0 is in the image of C, and in that case the system has a solution. The matrix C has $m-p$ columns of index $\mu(i)$, $1 \leq i \leq m-p$, where $\mu(i)$ is a strictly increasing function of i, such that the entry $c_{i,\mu(i)}$ is the first non-zero coordinate of row C_i of C. The remaining columns correspond to the free variables x_i. Thus there are $n-(m-p)$ of them. For any choice of the free variables the system admits a unique solution in the remaining $(m-p)$ variables.

Example 5.7.8. Now we redo Example 5.7.6 via Gaussian elimination to illustrate the remark above. Here $n = m = 3$. The augmented matrix is

$$\begin{pmatrix} 1 & -1 & 0 & b_1 \\ 0 & 1 & -1 & b_2 \\ -1 & 0 & 1 & b_3 \end{pmatrix}.$$

We reduce A to row echelon form C:

$$\begin{pmatrix} 1 & -1 & 0 & b_1 \\ 0 & 1 & -1 & b_2 \\ 0 & 0 & 0 & b_1+b_2+b_3 \end{pmatrix}.$$

so $p = 1$. $\mu(1) = 1$, $\mu(2) = 2$, so x_3 is the only free variable. The only condition on b is that $b_1 + b_2 + b_3 = 0$.

5.8 Real and Complex Linear Maps

In Example 3.2.8, we noticed that $\mathbb{C}$ can be considered as a two dimensional vector space over $\mathbb{R}$, with the basis 1 and i. More generally $\mathbb{C}^n$ can be considered as the real vector space $\mathbb{R}^{2n}$. Indeed, if $\mathbf{e}_1, \ldots, \mathbf{e}_n$ is a basis for $\mathbb{C}^n$, then $\mathbf{e}_1, \mathbf{i}\mathbf{e}_1, \ldots, \mathbf{e}_n, \mathbf{i}\mathbf{e}_n$

is a basis for the space as a real vector space. So we have: if V is a complex vector space of dimension n, considered as a $\mathbb{R}$-vector space it has dimension $2n$.

Now we want to consider linear transformations. We start with the simplest case: a $\mathbb{C}$-linear transformation from $\mathbb{C}$ to $\mathbb{C}$. It is, of course just given by multiplication by a complex number $a+\mathrm{i}b$, where a and b are real. We can think of this as a $\mathbb{R}$-linear map from $\mathbb{R}^2$ to $\mathbb{R}^2$: what is its matrix? Since it sends the complex number $x+\mathrm{i}y$ to

$$(a+\mathrm{i}b)(x+\mathrm{i}y) = ax - by + \mathrm{i}bx + \mathrm{i}ay,$$

in terms of the basis $\{1,\mathrm{i}\}$ of $\mathbb{C}$ as a $\mathbb{R}$ vector space, the linear map $L_{\mathbb{R}}$ is

$$\begin{pmatrix} x \\ y \end{pmatrix} \longmapsto \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{5.15}$$

which gives us the 2×2 matrix representing multiplication by $a+\mathrm{i}b$ as a $\mathbb{R}$-linear map.

Exercise 5.8.1. Show that if $a+\mathrm{i}b \neq 0$, the map $L_{\mathbb{R}}$ has trivial nullspace by studying the 2×2 system of linear equations you get. Why is this obvious given that the map comes from multiplication by a complex number?

We could of course do the same thing for any $\mathbb{C}$-linear map from $\mathbb{C}^n$ to $\mathbb{C}^m$ with a $m\times n$ matrix A of complex numbers, getting a $\mathbb{R}$ linear map from $\mathbb{R}^{2n}$ to $\mathbb{R}^{2m}$ which is left to you to write down in the obvious bases. This process could be called *decomplexifying* a complex linear map.

Now we want to go the other way around. We start with a $\mathbb{R}$-linear map L represented by the real $m\times n$ matrix A. Because real numbers are contained in the complex numbers, we can view A as representing a $\mathbb{C}$-linear transformation from $\mathbb{C}^n$ to $\mathbb{C}^m$. We call this map $L_{\mathbb{C}}$, the *complexification* of L. Any vector in $\mathbb{C}^n$ can be written as $\mathbf{a}+\mathrm{i}\mathbf{b}$. Then $L_{\mathbb{C}}(\mathbf{a}+\mathrm{i}\mathbf{b}) = L(\mathbf{a})+\mathrm{i}L(\mathbf{b})$, as is easily checked since the matrix representing L is real.

Given a $\mathbb{R}$-linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$, which is just given by multiplication by a $m\times n$ real matrix A, we get a $\mathbb{C}$-linear map from $\mathbb{C}^n$ to $\mathbb{C}^m$ again just given by multiplication by A. We can now decomplexify the complexification. In the special case $m=n=1$, by (5.15) applied when $b=0$, the decomplexified 2×2 matrix of $L_{\mathbb{C}}$ in the usual basis is just the diagonal matrix aI_2.

We will use complexification in later chapter, sometimes without mention, because it is generally easier to study linear transformations over $\mathbb{C}$ than over $\mathbb{R}$, primarily because the field $\mathbb{C}$ is algebraically closed, meaning that every polynomial factors as a product of linear polynomials.

For that reason it is useful to determine that complexification is uniquely defined. In other words, given a real $m\times n$ matrix A, these is a unique $m\times n$ complex matrix B such that for $\mathbf{a}+\mathrm{i}\mathbf{b}$ as above, $B\mathbf{a} = A\mathbf{a}$. This forces B to be real, and then uniqueness follows from the uniqueness of the matrix representing a linear transformation.

In §9.2 we will study Hermitian matrices: square complex matrices A of size n: they are the matrices such that $a_{ji} = \overline{a_{ij}}$ for all i and j. Here we use them as an example of the interaction between real and complex structures on vector spaces.

We study the equations defining Hermitian matrices H inside the complex vector space $M_n(\mathbb{C})$ of all square complex matrices. First the diagonal elements of matrices in H are real. More generally the equations defining H in $M_n(\mathbb{C})$ are not linear over $\mathbb{C}$, but only linear over $\mathbb{R}$. So we can ask for the real dimensions of $M_n(\mathbb{C})$ and its $\mathbb{R}$-subspace H..

Theorem 5.8.2. *$M_n(\mathbb{C})$ has real dimension $2n^2$, and H has real dimension n^2.*

Proof. The result for $M_n(\mathbb{C})$ is easy. If we write every entry of the $n \times n$ complex matrix A as $a_{ij} = r_{ij} + \mathrm{i}c_{ij}$ with both r_{ij} and c_{ij} real, we can use the r_{ij} and c_{ij} as a basis for $M_n(\mathbb{C})$.

For H we get one real linear condition imposed on each diagonal term: $c_{ij} = 0$. For each term below the diagonal we get two conditions: $r_{ij} = r_{ji}$ and $c_{ij} = -c_{ji}$. Since all these conditions are independent, the $\mathbb{R}$-dimension $\dim_{\mathbb{R}}$ of H is

$$\dim_{\mathbb{R}} H = 2n^2 - n - n(n-1) = n^2.$$

□

The folllowing decomposition of a complex matrix into Hermitian parts will be useful later on.

Theorem 5.8.3. *Any square complex matrix can be written uniquely as $A = B + \mathrm{i}C$, where both B and C are Hermitian.*

Proof. To start, note that the dimensions are correct. Then note that we can get any complex number on the diagonal. Off the diagonal we simply need to solve the equations for the ij and ji term separately. If write $a_{ij} = a'_{ij} + \mathrm{i}a''_{ij}$, $b_{ij} = b'_{ij} + \mathrm{i}b''_{ij}$ and $c_{ij} = c'_{ij} + \mathrm{i}c''_{ij}$ then:

$$\begin{aligned} b'_{ij} - c''_{ij} &= a'_{ij}; \\ b''_{ij} + c'_{ij} &= a''_{ij}; \\ b'_{ij} + c''_{ij} &= a'_{ji}; \\ -b''_{ij} + c'_{ij} &= a''_{ji}. \end{aligned}$$

For each ij we get a system of 4 linear inhomogenous equations in the 4 real variables b'_{ij}, b''_{ij}, c'_{ij}, c''_{ij} with constants $(a'_{ij}, a''_{ij}, a'_{ji}, a''_{ji})$. In matrix notation we have:

$$\begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \end{pmatrix} \begin{pmatrix} b'_{ij} \\ b''_{ij} \\ c'_{ij} \\ c''_{ij} \end{pmatrix} = \begin{pmatrix} a'_{ij} \\ a''_{ij} \\ a'_{ji} \\ a''_{ji} \end{pmatrix}$$

This can by easily solved by Gaussian elimination, so there is a unique solution. □

Exercise 5.8.4. A matrix is skew-Hermitian if $a_{ji} = -\overline{a_{ij}}$ for all i and j. Show that the real dimension of the skew-Hermitian matrices is again n^2, and that we can write any complex matrix uniquely as a Hermitian matrix plus a skew-Hermitian matrix.

Hint: you can either just imitate the previous proof, or derive the result by just considering what king of matrix iA is, if A is Hermitian.

5.9 Nilpotent Operators

Recall that a linear map from a vector space to itself is called a linear operator: Definition 4.4.1.

In this section we study special operators that can be completely analyzed without too much difficulty. They give interesting examples of matrices representing operators. First some definitions.

Let $L\colon V \to V$ be any linear operator. Let $\mathbf{v} \in V$ be a non-zero vector. Let p be the smallest integer such that the vectors $\{\mathbf{v}, L\mathbf{v}, L^2\mathbf{v}, \ldots, L^p\mathbf{v}\}$ are linearly dependent. For dimension reasons $1 \leq p < \dim V$. The minimality of p implies that the coefficient of $L^p\mathbf{v}$ in the equation of linear dependence is non–zero. So we may write

$$L^p\mathbf{v} = a_0\mathbf{v} + a_1 L\mathbf{v} + \cdots + L^{p-1}\mathbf{v}$$

Under these hypotheses we say that $\mathbf{v}$ is a *cyclic vector* of *period* p.

Lemma 5.9.1. *The vectors* $\{\mathbf{v}, L\mathbf{v}, L^2\mathbf{v}, \ldots, L^{p-1}\mathbf{v}\}$ *form a basis of a subspace* W *of* V *of dimension* p*, invariant under* L*.* W *is a* cyclic subspace *for* L*. Then* W *has a* cyclic vector *for* L*.*

The only statement left to prove is the invariance under L: if $\mathbf{w}$ is in W, then $L\mathbf{w}$ is in W.

Proof. Write $\mathbf{w}$ in terms of the given basis:

$$\mathbf{w} = b_{p-1}L^{p-1}\mathbf{v} + \cdots + b_1 L\mathbf{v} + b_0\mathbf{v}$$

with coefficients b_i. Apply the operator L to the equation. Then

$$L\mathbf{w} = b_{p-1}L^p\mathbf{v} + \cdots + b_1 L^2\mathbf{v} + b_0 L\mathbf{v}.$$

Write out $L^p\mathbf{v}$ using (5.9) to see that $L\mathbf{w}$ is in W, which proves the invariance. □

In the basis of W given above, the matrix of L is

$$A = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 & a_0 \\ 1 & 0 & 0 & \dots & 0 & 0 & a_1 \\ 0 & 1 & 0 & \dots & 0 & 0 & a_2 \\ 0 & 0 & 1 & \ddots & 0 & 0 & a_3 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 & a_{p-2} \\ 0 & 0 & 0 & \dots & 0 & 1 & a_{p-1} \end{pmatrix} \tag{5.16}$$

as you can easy see, by interpreting the columns of this matrix as the coefficients of $L\mathbf{z}$, for the appropriate $\mathbf{z}$. In §10.4 we will see that this matrix is called the companion matrix of the polynomial $a_0 + a_1 t + \cdots + t^{p-1}$.

In this section we are only interested in the special case where all the a_i are zero.

Definition 5.9.2. Let $L\colon V \to V$ be any linear operator. Let $\mathbf{v}$ be a non-zero vector such that there is a non-negative integer p such that $L^p(\mathbf{v}) = 0$ and $L^{(p-1)}(\mathbf{v}) \neq \mathbf{0}$.

Proposition 5.9.3. *Under these hypotheses* $\mathbf{v}$ *has period* p *for* L*.*

Thus the vectors $\mathbf{v}$, $L\mathbf{v}$, $L^2\mathbf{v}$, …, $L^{(p-1)}\mathbf{v}$ are linearly independent, so they span a subspace W of dimension p. The operator L restricts to an operator on W.

Proof. We need to show that the p vectors given above are linearly independent. If not, there is an equation of linear dependence:

$$a_0\mathbf{v} + a_1 L\mathbf{v} + a_2 L^2\mathbf{v} + \cdots + a_{p-1}L^{(p-1)}\mathbf{v} = \mathbf{0}$$

Apply $L^{(p-1)}$ to this equation to get $a_0 L^{(p-1)}\mathbf{v} = \mathbf{0}$. This forces $a_0 = 0$. Then apply L^{p-2} to get $a_1 = 0$. Continuing in this way, all the a_i must be 0 so there is no equation of linear dependence. □

This result is most useful in the context of nilpotent operators, already mentioned in Example 2.3.12.

Definition 5.9.4 (Nilpotent Operators). Let L be a non-zero linear operator such that there is a power of L that is the zero–operator. Such an L is *nilpotent*. Let p be the smallest integer such that $L^p = 0$, but $L^{(p-1)} \neq 0$. This r is the *index of nullity* of L.

If L is nilpotent, then every non-zero vector in V has a finite period. Furthermore there is a vector $\mathbf{v}$ with period the index of nullity of L.

What else can we say about nilpotent operators? Pick a vector $\mathbf{v}$ of maximum period p, which is the index of nullity of L. If p is the dimension of V, then the vectors $\mathbf{v}$, $L\mathbf{v}$, $L^2\mathbf{v}$, …, $L^{(p-1)}\mathbf{v}$ form a basis for V, which is a *cyclic* space for L. The matrix of L in this basis is the $p \times p$ matrix:

$$N_p = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} \tag{5.17}$$

with ones on the subdiagonal (the entries $(i+1,i)$, and zeroes everywhere else. Not surprisingly this is the matrix (5.16) with all the a_i set to 0. We call this matrix the standard nilpotent matrix of size p.

We continue this analysis to give a similar description of all nilpotent operators.

Theorem 5.9.5. *Let $L\colon V \to V$ be a nilpotent operator on the vector space V of dimension n. Then $V = \oplus V_i$, where V_i is a p_i-dimensional L-invariant cyclic subspace of size p_i, so that if L_i is the restriction of L to V_i, then there is a vector $\mathbf{v}_i \in V_i$ so that a basis for V_i is*

$$\{\mathbf{v}_i, L_i\mathbf{v}_i, L_i^2\mathbf{v}_i, \dots, L_i^{p_i-1}\mathbf{v}_i\}, \tag{5.18}$$

and the $p_i \times p_i$ matrix A_i for L_i in this basis is the standard nilpotent block N_{p_i}.

Therefore in that basis for each of the terms in the direct sum, you get a lower triangular matrix with all terms zero except some of the terms on the subdiagonal. For example you get, if the first nilpotent block is 3×3, and the second 2×2

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{5.19}$$

Therefore the matrix of L in these bases is block diagonal with matrices N_{p_i} along the diagonal.

Exercise 5.9.6. What is the index of nullity of L in terms of the block sizes $p_1, \dots, p_k$?

Proof (Theorem). We will prove this by induction on the dimension n of V. The result is trivial for $n = 1$. L has a non-trivial nullspace: just take any element $\mathbf{w}$ whose period is $m-1$, where m is the nullity of L. Then $\mathbf{v} = L^{m-1}\mathbf{w}$ is in the nullspace. Therefore by the Rank-Nullity Theorem, the dimension of the image $W = LV$ is at most $n-1$. L acts on W: since any element $\mathbf{w} \in W$ can be written $\mathbf{w} = L\mathbf{v}$, $\mathbf{v} \in V$, let $L\mathbf{w} = L^2\mathbf{v}$. In particular L is nilpotent on W with index of nullity $m-1$. By induction we can write W as a direct sum of L-invariant subspaces W_i, $1 \le i \le k$, each of which is cyclic, generated by a $\mathbf{w}_i$ of period $p_i - 1$, for some $p_i \ge 2$. Let $\mathbf{v}_i \in V$ be a vector such that $\mathbf{w}_i = L\mathbf{v}_i$. Then the subspace $V_i \subset V$ generated by

$$\{\mathbf{v}_i, L\mathbf{v}_i, \dots, L^{p_i}\mathbf{v}_i\}$$

is cyclic of dimension $p_i + 1$, and $\mathbf{v}_i$ has period $p_i + 1$.

Next we show that the subspace V_0 of V generated by $V_1, V_2, \ldots V_k$ is a direct sum of these subspaces. This is the main step of the proof. How do we show this? We have to show that any element $\mathbf{u} \in V_0$ can be written uniquely as a sum of elements from the V_i. By subtracting one representation from another, it is enough to show that if

$$\mathbf{0} = \mathbf{u}_1 + \mathbf{u}_2 + \cdots + \mathbf{u}_k, \text{for } \mathbf{u}_i \in V_i$$

then for all i, $\mathbf{u}_i = \mathbf{0}$. Apply L to this equation to get

$$\mathbf{0} = L\mathbf{u}_1 + L\mathbf{u}_2 + \cdots + L\mathbf{u}_k.$$

Since this sum in W is direct, each term must be in the nullspace of L, so the equation reduces to

$$\mathbf{0} = c_1 L^{p_1}\mathbf{v}_1 + c_2 L^{p_2}\mathbf{v}_2 + \cdots + c_k L^{p_k}\mathbf{v}_k,$$

which can be written

$$\mathbf{0} = c_1 L^{p_1-1}\mathbf{w}_1 + c_2 L^{p_2-1}\mathbf{w}_2 + \cdots + c_k L^{p_k-1}\mathbf{w}_k.$$

By using the fact that W is a direct sum, each one of the terms must be $\mathbf{0}$. Since we know that the period of $\mathbf{w}_i$ is $p_1 - 1$, this can only happen if all the c_i are 0. This shows that the sum is direct in V_0.

To finish the proof we deal with the case where V_0 is not all of V. Complete the basis of V_0 to a basis of V. All these new basis elements have period one, since they do not contribute to W. For each basis element, we get an additional direct sum component of dimension 1. □

Thus we have found a simple matrix representing any nilpotent transformation. This is an important step in establishing the Jordan Canonical Form in §10.6.

5.10 The Rank via Submatrices

Here is another way of defining the rank of a $m \times n$ matrix, using the notion of a square submatrix of A defined in §2.4. This is actually the classic definition of the rank. This material will not be used in the rest of the book.

Theorem 2.8.11 tells us that

Theorem 5.10.1. *The rank of any matrix A is the maximum over square submatrices B of A of the rank of B. It is also the size of the biggest nonsingular submatrix of A.*

Recall that nonsingular means invertible, which implies square.

Proof. We first show that the rank of any square submatrix of A is at most the rank r of A. Suppose not. Then we can find $s > r$ columns of the submatrix that are linearly independent. Then the corresponding columns of A are linearly independent, a contradiction.

To conclude the proof, we need to find a square submatrix of A of rank r. We essentially have done this in proving that row rank is column rank. Fix r linearly independent rows of A. The matrix formed by these rows has rank r. Therefore it has r linearly independent columns. The square submatrix formed by these r rows and r columns clearly has rank r. It is therefore nonsingular. □

Here is a useful theorem concerning the rank.

Theorem 5.10.2. *Assume that the matrix A has a square submatrix B of size r and of rank r. Let C be any square submatrix of size $r+1$ of A containing B. Assume any such C has rank r. Then A has rank r.*

Proof. As before we may assume that the submatrix B is in the upper left hand corner of A. This is simply for convenience of notation. We consider the square submatrix of A of size $r+1$ that can be written in block matrix form as

$$B_{pq} = \begin{pmatrix} B & \mathbf{d}_q \\ \mathbf{g}_p & a_{pq} \end{pmatrix}$$

for $r < p,q \leq n$. $\mathbf{d}_q$ is the column vector $[a_{1q},\ldots,a_{rq}]$ and $\mathbf{g}_p$ is the row vector $(a_{p1},\ldots,a_{pr})$. Because B has maximum rank r, any r-vector, in particular $\mathbf{g}_p$, can be written as a linear combination of the rows of B. Thus there are constants $c_1, \ldots, c_r$ such that

$$c_1\mathbf{b}^1 + \cdots + c_r\mathbf{b}^r = \mathbf{g}_p.$$

Let $\mathbf{c}^t$ be the r-row vector with entries $(c_1,c_2,\ldots,c_r)$. We multiply B_{pq} on the left by the invertible matrix E also written in block form, with the blocks of the same size as those of B_{pq}:

$$E = \begin{pmatrix} I_{r\times r} & \mathbf{0}_{r\times 1} \\ -\mathbf{c}^t & 1 \end{pmatrix}$$

Using Example 2.9.9, by block multiplication we get

$$EB_{pq} = \begin{pmatrix} B - \mathbf{0}_{r\times 1}\mathbf{c}^t & I_{r\times r}\mathbf{d}_q + \mathbf{0}_{r\times 1}a_{pq} \\ -\mathbf{c}^t B + \mathbf{g}_p & -\mathbf{c}^t\mathbf{d}_q + a_{pq} \end{pmatrix} = \begin{pmatrix} B & \mathbf{d}_q \\ \mathbf{0}_{1r} & -\mathbf{c}^t\mathbf{d}_q + a_{pq} \end{pmatrix}$$

Now $\mathbf{c}^t\mathbf{d}_q$ is the matrix product of the row vector $\mathbf{c}^t$ by the column vector $\mathbf{q}$, so it is a number, as required. Since E is invertible, EB_{pq} has the same rank as B_{pq}. This is only true if $-\mathbf{c}^t\mathbf{d}_q + a_{pq} = 0$. Since this is true for any q between $r+1$ and n, this implies that the p-th row of A is the linear combination

$$\mathbf{a}^p = c_1\mathbf{a}^1 + c_2\mathbf{a}^2 + \cdots + c_r\mathbf{a}^r$$

of the first r rows of A.

We can argue in the same way for every row $\mathbf{a}^s$ of A, $r < s \leq n$, which implies that A has rank r. □

Exercise 5.10.3. Show that the proof involves checking the rank of $(n-r)^2$ submatrices of A of size $(r+1)$.

Example 5.10.4. Consider Example 2.4.1, namely the matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 4 & 5 & 6 \end{pmatrix}.$$

The submatrix of size 2 formed by the first two rows and the first two columns is

$$\begin{pmatrix} 1 & 2 \\ 7 & 8 \end{pmatrix}$$

The two rows are not proportional, so this submatrix has rank 2. So the rows $(1,2)$ and $(7,8)$ form a basis for F^2, so we can write the first two entries of the third row of A as a linear combination of them. Indeed

$$(4,5) = \frac{1}{2}(1,2) + \frac{1}{2}(7,8).$$

So subtracting half of the first row and half of the second row from the third row, we get the matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & 0 \\ 7 & 8 & 9 \end{pmatrix}.$$

This matrix obviously has rank 2, confirming the theorem.

A 4×4 example with rank 2 needed here.

Now we can repeat this for a symmetric matrix. We prove an interesting variant of Theorem 5.10.1. Recall that we say that a square matrix of size n is non-singular if its rank r is equal to n, and is singular otherwise.

Theorem 5.10.5. *If a symmetric matrix has rank r, it has a non-singular principal submatrix of size r. In other words, the rank of a symmetric matrix is the maximum size of a non-singular principal submatrix of A.*

So to compute the rank of a symmetric matrix one only needs to look at the principal submatrices: see (2.14). This result can be false for matrices that are not symmetric. For example the matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

has rank 1, but its principal submatrices all have rank 0. On the other hand a symmetric matrix of rank r could well have a submatrix that is not principal of rank r. For example

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

Theorem 5.10.5 is a corollary of the following theorem, which is an improvement of Theorem 5.10.2. This theorem is useful because of the special role that principal submatrices play in the theory of symmetric bilinear forms.

Theorem 5.10.6. *Assume that the symmetric matrix A has a non-singular principal submatrix B of size r. Assume that every principal submatrix C of A containing B, of size $r+1$ or $r+2$, is singular. Then A has rank r.*

Remark 5.10.7. So we only have to check principal submatrices containing B: all the ones of size $r+1$ and $r+2$. So this means checking

$$n-r+\binom{n-r}{2} = \frac{(n-r+1)(n-r)}{2}$$

submatrices, while in Theorem 5.10.2, for a square matrix of size n and rank r you must check $(n-r)^2$ submatrices. So the saving is roughly a factor of 2, and is exactly what the symmetry of the matrix leads you to expect.

Proof. We first look at principal submatrices of size $r+1$, proceeding exactly as in the proof of Theorem 5.10.2. So we assume B is in the upper left hand corner of A, for convenience. We consider the principal submatrix of A of size $r+1$ that can be written in block matrix form as

$$B_{pp} = \begin{pmatrix} B & \mathbf{d}_p \\ \mathbf{d}_p^t & a_{pp} \end{pmatrix}$$

for $r < p \leq n$. $\mathbf{d}_p$ is the column vector $[a_{1p}, \ldots, a_{rp}]$. Because B has maximum rank r, any r-vector, in particular $\mathbf{d}_p^t$, can be written as a linear combination of the rows of B. Thus there are constants $c_1, \ldots, c_r$ such that

$$c_1\mathbf{b}^1 + \cdots + c_r\mathbf{b}^r = \mathbf{d}_p^t.$$

Let E be the invertible square matrix of size $r+1$, write in the same block form as B:

$$E = \begin{pmatrix} I_{r\times r} & \mathbf{0}_{r\times 1} \\ -\mathbf{c}^t & 1 \end{pmatrix}$$

and form

$$EB_{pp}E^t = \begin{pmatrix} B & \mathbf{0}_{r\times 1} \\ \mathbf{0}_{1\times r} & -\mathbf{c}^t\mathbf{d}_p + a_{pp} \end{pmatrix}$$

Now we can assume that A is congruent to a matrix

$$\begin{pmatrix} B & 0_{r\times s} \\ 0_{s\times r} & D \end{pmatrix}$$

where $s = n-r$ and D is a symmetric matrix with zeroes along the diagonal.

Now out hypothesis concerning principal submatrices of size $r+2$ containing B says that any submatrix

$$\begin{pmatrix} B & 0_{r\times 2} \\ 0_{2\times r} & D_{pq} \end{pmatrix} \tag{5.20}$$

has rank less than $r+2$, where D_{pq} is the matrix

$$\begin{pmatrix} 0 & a_{pq} \\ a_{qp} & 0 \end{pmatrix}$$

where $p \neq q$. Of course by symmetry $a_{pq} = a_{qp}$. The only way this can have rank less than $r+2$ is if $a_{pq} = 0$

Doing this for any choice $r < p < q \leq n$ shows that A is congruent to the matrix

$$\begin{pmatrix} B & 0_{r\times s} \\ 0_{s\times r} & 0_{s\times s} \end{pmatrix}$$

which obviously has rank r, so we are done. □

To finish the proof of Theorem 5.10.5, assume that the symmetric matrix A has a non-singular principal submatrix B of size r, and none of larger size. We show that all the submatrices of A containing B of size $r+1$ are singular by using Theorem 5.10.6, and then conclude by Theorem 5.10.2. Thus we need to show that all the submatrices B_{pq} are singular, using the notation of Theorem 5.10.2. B_{pq} is a submatrix of the submatrix of size $r+2$ where we adjoin to B the rows and columns of index p and q. The proof of Theorem 5.10.6 shows not only that this submatrix is singular, but that is only has rank r: see (5.20) and what follows. Then its submatrix B_{pq} can have rank at most r, and we are done. □

Chapter 6
Duality

Abstract The chapter studies on linear functionals and duality. In an introductory section we develop the terminology and notation of bilinear forms, which provide a useful interpretation of functionals and duality. This material will be used in the further discussion of bilinear forms in Chapter 7. These two chapters are independent from the rest of the book. They give a different point of view and are a generalization of material covered in Chapter 8.

6.1 The Dual Space

The field F is an F-vector space of dimension 1, so we make the following definition.

Definition 6.1.1. A *linear functional* on a F-vector space V is a linear map from V to F.

Assume V has dimension n. We have the linear functional $\mathbf{0}$ that is identically 0. Any other linear functional takes on non-zero values, so that by the Rank-Nullity Theorem its rank is 1 and its nullity is $n-1$.

Example 6.1.2. The most important example of a linear functional is definite integration. Fix a closed interval $I=[a,b]$ of real numbers, and let the vector space be the set V of continuous functions on I. This is indeed a vector space as you should check, but it is infinite dimensional. Then the map

$$f(x) \mapsto \int_a^b f(x)dx$$

is a linear functional on V. Determine what properties of integration you need to establish this. In §6.2 we will apply this to a finite-dimensional subspace of V.

Example 6.1.3. The trace of a square matrix of size n, already defined in Exercise 2.8.18:

$$\mathrm{tr}(A) = a_{11} + a_{22} + \cdots + a_{nn}$$

is a linear functional on the vector space M_{nn} of such matrices. Prove this.

A special case of Theorem 4.1.11 says that the set of linear functionals on V is a F-vector space. We call it V^* rather than the more cumbersome $\mathscr{L}(V,F)$. By Theorem 5.2.2 the dimension of V^* is n. V^* is called the dual space of V.

We construct a basis for V^* using a basis $\mathfrak{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of V.

Definition 6.1.4. For each i, $1 \leq i \leq n$, let $\mathbf{f}_i$ be the unique linear functional such that

$$\mathbf{f}_i(\mathbf{v}_j) = 0 \text{ if } i \neq j, \text{ and } \mathbf{f}_i(\mathbf{v}_i) = 1. \tag{6.1}$$

The linear functional $\mathbf{f}_i$ is *dual* to the basis element $\mathbf{v}_i$. These n functionals are also called the coordinate functions in the basis $\mathfrak{B}$.

The condition expressed in (6.1) comes up so often that we will write it using the Kronecker delta (Definition B.1.6): $\mathbf{f}_i(\mathbf{v}_j) = \delta_{ij}$. The existence and uniqueness of the $\mathbf{f}_i$ follows immediately from Theorem 4.1.12 applied to V and F as the space we are mapping to.

Example 6.1.5. When V is F^n, we can use as basis the standard unit coordinates vectors $\mathbf{e}_i$ of Example 3.4.4, in which case the dual linear functional $\mathbf{f}_i$ take the vector $(x_1, x_2, \ldots, x_n)$ to its i-th coordinate x_i.

Definition 6.1.6. The linear map $D_{\mathfrak{B}} : V \to V^*$ is defined by $D_{\mathfrak{B}}(\mathbf{v}_i) = \mathbf{f}_i$, $1 \leq i \leq n$.

By Theorem 4.1.12 again this is enough to define a linear map. We write $D_{\mathfrak{B}}$, because if the basis changes, the map changes. For example assume V is one-dimensional. Let $\mathbf{v}$ be a basis of V and $\mathbf{f}$ its dual, so that $\mathbf{f}(\mathbf{v}) = 1$. Then the functional $\mathbf{f}/k$ is the dual of $k\mathbf{v}$ for any scalar k. Compare the linear mapping $D_{\mathbf{v}} : \mathbf{v} \mapsto \mathbf{f}$ to the map $D_{k\mathbf{v}} : k\mathbf{v} \mapsto \mathbf{f}/k$. Since $D_{k\mathbf{v}}(\mathbf{v}) = \mathbf{f}/k^2$, they are the same if and only if $k^2 = 1$. We will generalize this computation in Example 6.4.8.

Theorem 6.1.7. *The $\mathbf{f}_i$, $1 \leq i \leq n$, form a basis $\mathfrak{B}^*$ of V^*, called the dual basis of the basis $\mathfrak{B}$ of V.*

Proof. First we show they are linearly independent. If there were an equation of linear dependence, we would have

$$c_1\mathbf{f}_1 + c_2\mathbf{f}_2 + \cdots + c_n\mathbf{f}_n = 0,$$

where not all of the c_i are equal to 0. Now evaluate at each $\mathbf{v}_i$: the equation becomes $c_i\mathbf{f}_i(\mathbf{v}_i) = 0$. But $\mathbf{f}_i(\mathbf{v}_i) = 1$, so $c_i = 0$. Therefore all the coefficients vanish and we do not have an equation of linear dependence. For dimension reasons the f_i clearly span. □

Theorem 6.1.8. *An arbitrary linear functional $\mathbf{f}$ is written in terms of the basis $\mathbf{v}_i$ and the dual basis $\mathbf{f}_i$ as*

$$\mathbf{f} = \sum_{i=1}^{n} \mathbf{f}(\mathbf{v}_i)\mathbf{f}_i. \tag{6.2}$$

Similarly an arbitrary vector $\mathbf{v}$ *is written*

$$\mathbf{v} = \sum_{i=1}^{n} \mathbf{f}_i(\mathbf{v})\mathbf{v}_i. \tag{6.3}$$

Proof. Evaluate the functionals on either side of (6.2) on every basis vector $\mathbf{v}_j$ to get, since the $\mathbf{f}_j$ are the dual basis:

$$\mathbf{f}(\mathbf{v}_j) = \sum_{i=1}^{n} \mathbf{f}(\mathbf{v}_i)\mathbf{f}_i(\mathbf{v}_j) = \mathbf{f}(\mathbf{v}_j)\mathbf{f}_j(\mathbf{v}_j) = \mathbf{f}(\mathbf{v}_j)$$

so these two functionals agree on a basis, so they agree everywhere. The last result is proved in the same way, this time applying the functional $\mathbf{f}_j$ to both sides of (6.3). □

Example 6.1.9. If V is F^n with its natural basis, so an element $\mathbf{v}$ in V has coordinates $(a_1, a_2, \ldots, a_n)$, then V^* is again F^n in the dual basis. Writing a $\mathbf{f}$ in the dual basis with coordinates $(b_1, b_2, \ldots, b_n)$, then the evaluation of the linear functional $\mathbf{f}$ on $\mathbf{v}$ is:

$$\mathbf{f}(\mathbf{v}) = a_1b_1 + a_2b_2 + \cdots + a_nb_n.$$

Note that this is the matrix product of the the row vector $\mathbf{b}$ with the column vector $\mathbf{a}$, a useful remark, as we shall soon see.

6.2 Application: Lagrange Interpolation

Let P_n be the vector space of polynomials $F[t]$ of degree at most n over F, which can be either $\mathbb{R}$ or $\mathbb{C}$. This is an $n+1$ dimensional vector space. One possible basis $\mathfrak{B}$ of P_n consists of the polynomials $\{1, t, t^2, \ldots, t^n\}$.

Consider the graph of the polynomial $f(t) \in P_n$ in F^2: this means the collection of points $(t, f(t))$. Take $n+1$ distinct points t_i, $0 \le i \le n$, in F and the corresponding $n+1$ points $(t_i, f(t_i))$ on the graph. The map $f(t) \mapsto (f(t_0), \ldots, f(t_n))$ is a linear map from P_n to F^{n+1}. Its matrix, for the basis $\mathfrak{B}$ is

$$V = \begin{pmatrix} 1 & t_0 & \ldots & t_0^n \\ 1 & t_1 & \ldots & t_1^n \\ \vdots & \vdots & \ldots & \vdots \\ 1 & t_n & \ldots & t_n^n \end{pmatrix}. \tag{6.4}$$

Indeed, writing the polynomial f as $f(t) = a_0 + a_1t + \cdots + a_nt^n$ and writing $\mathbf{y}$ for $(f(t_0, \ldots, f(t_n))$, the linear map is given by

$$V\mathbf{a} = \mathbf{y}. \tag{6.5}$$

The matrix of coefficients V of this system is called the Vandermonde matrix at the points $\{t_0, t_1, \ldots, t_n\}$. Any polynomial $f(t)$ that satisfies the equations (6.5) *interpolates* the $n+1$ points (t_i, y_i), $0 \leq i \leq n$. If the $(n+1) \times (n+1)$ matrix V is invertible, then there is a unique interpolating polynomial f. It is easy to see this is true if the points t_i are distinct. Indeed, a polynomial of degree $\leq n$ has at most n roots, except for the polynomial 0. Thus the nullspace of the linear map is trivial, and we are done. Later we will prove that V is invertible directly by showing that its determinant is non-zero in Example 11.5.5.

Here we will show this by using linear functionals and the dual basis.

Definition 6.2.1. For each $t_0 \in F$ we get a linear functional $\mathbf{e}_{t_0}$ on P_n given by

$$\mathbf{e}_{t_0} : f(t) \mapsto f(t_0),$$

for every $f \in P_n$. The functional $\mathbf{e}_{t_0}$ called the *evaluation* functional at t_0.

To check that $\mathbf{e}_{t_0}$ is a linear functional, we must show

$$\mathbf{e}_{t_0}(f(x) + g(x)) = \mathbf{e}_{t_0}(f) + \mathbf{e}_{t_0}(g), \text{ and } \mathbf{e}_{t_0}(cf(x)) = c\mathbf{e}_{t_0}(f(x)).$$

Both statements are obvious. Notice that we are in the general framework of Example 3.2.7.

Here we are only interested in the $n+1$ points t_i, and we will only consider the evaluation functionals $\mathbf{e}_{t_i}$, that we write for simplicity as $\mathbf{e}_i$.

Theorem 6.2.2. *For any collection of $n+1$ distinct points t_i, the evaluation functions $\mathbf{e}_i$ form a basis of the dual space P_n^*.*

Proof. Consider the polynomials

$$f_j(t) = \frac{(t-t_0)\ldots(t-t_{j-1})(t-t_{j+1})\ldots(t-t_n)}{(t_j-t_0)\ldots(t_j-t_{j-1})(t_j-t_{j+1})\ldots(t_j-t_n)}. \tag{6.6}$$

The numerator is chosen so that $f_j(t_i) = 0$ when $i \neq j$, and the denominator chosen so $f_j(t_j) = 1$. Clearly $f_j(t)$ is a polynomial of degree n, so it is in P_n. By Theorem 6.1.7, $f_j(t)$ form a basis for P_n. Indeed the $\mathbf{e}_i(f_j) = \delta_{ij}$. Thus the $\{\mathbf{e}_0(f), \ldots, \mathbf{e}_n(f)\}$ form the basis of P_n^* dual to the basis $\{f_0(t), \ldots, f_n(t)\}$ of P_n. □

As a corollary of Theorem 6.1.8, we can write an arbitrary polynomial $f(t) \in P_n$ in terms of the polynomials $f_i(t)$ of (6.6) as

$$f(t) = \sum_{i=0}^{n} \mathbf{e}_i(f) f_i(t) = \sum_{i=0}^{n} f(t_i) f_i(t).$$

Applying this to the polynomials t^i, we get

$$t^i = t_0^i f_0(t) + t_1^i f_1(t) + \cdots + t_n^i f_n(t)$$

We can write these equations in matrix form as

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ t_0 & t_1 & \dots & t_n \\ \vdots & \vdots & \ddots & \vdots \\ t_0^n & t_1^n & \dots & t_n^n \end{pmatrix} \begin{pmatrix} f_0(t) \\ f_1(t) \\ \vdots \\ f_n(t) \end{pmatrix} = \begin{pmatrix} 1 \\ t \\ \vdots \\ t^n \end{pmatrix}.$$

Notice that the matrix of coefficients is the transpose of the Vandermonde matrix V. Since the two column vectors form bases of the same vector space P_n, V^t is a change of basis matrix, and therefore invertible. This also allows us to bypass the Vandermonde determinant computation in Example 11.5.5.

We have established that there is a unique polynomial of degree less than or equal to n whose graph passes through the $n+1$ points. This is known as interpolating the $n+1$ points by a function of a specific type, here polynomials of degree at most n.

Thus we have proved the desired theorem, which gives us a formula for the unique interpolating polynomial.

Theorem 6.2.3 (The Lagrange interpolation formula). *The unique solution for the problem of interpolating the $n+1$ points (t_i, y_i), where the t_i are distinct, by a polynomial of degree at most n is*

$$f(t) = y_0 f_0(t) + y_1 f_1(t) + \cdots + y_n f_n(t),$$

for the functions $f_i(t)$ of (6.6)*.*

Now we return to Example 6.1.2. The vector space P_n is a finite dimensional subspace of the space of continuous function on any finite interval I. As we have already noticed the definite integral of any f in P_n over the interval is a linear functional. We leave it to you to prove it.

Exercise 6.2.4. Prove that integration is a linear functional. Write down carefully what needs to be established.

So the integral of f can be written as a linear combination of the evaluation functions $\mathbf{e}_i$, for any set of $n+1$ distinct points t_i.

Theorem 6.2.5 (Quadrature Formula). *There are scalars c_i, $0 \le i \le n$ such that for any polynomial $f(t) \in P_n$ we have*

$$\int_I f(t)dt = c_0\mathbf{e}_0(f) + c_1\mathbf{e}_1(f) + \cdots + c_n\mathbf{e}_n(f) = c_0 f(t_0) + c_1 f(t_1) + \cdots + c_n f(t_n).$$

It is amusing to see how integration can be reduced to evaluation at points. In particular, consider the polynomials $f_i(t)$ of (6.6). Then

$$\int_I f_i(t)dx = c_i f_i(t_i).$$

For the powers t^i we get

$$\int_I t^i dt = c_0 t_0^i + c_1 t_1^i + \cdots + c_n t_n^i.$$

Exercise 6.2.6. How can you reconcile this computation with what you know about the integral of t^i?

6.3 Bilinear Forms: the General Case

Let V be a vector space of dimension n, and W a vector space of dimension m, over the base field F.

Definition 6.3.1. A *bilinear form* on $V \times W$ is a map $b(\mathbf{v},\mathbf{w})\colon V \times W \to F$ such that

1. For each fixed $\mathbf{a} \in V$ the function $\mathbf{g}_\mathbf{a}(\mathbf{w})\colon W \to F$, defined by $\mathbf{g}_\mathbf{a}(\mathbf{w}) = b(\mathbf{a},\mathbf{w})$ is a linear functional on W;
2. For each fixed $\mathbf{b} \in W$ the function $\mathbf{f}_\mathbf{b}(\mathbf{v})\colon V \to F$, defined by $\mathbf{f}_\mathbf{b}(\mathbf{v}) = b(\mathbf{v},\mathbf{b})$ is a linear functional on V;

These conditions say that b is a linear map in each variable separately.

Our goal is to understand all bilinear forms. First an example, which, as we will see soon, contains all possible cases.

Example 6.3.2. Take a matrix A of size $m \times n$. Let $V = F^n$, with coordinates x_j, and $W = F^m$ with coordinates y_i. We get a scalar valued function $b\colon V \times W \to F$ by setting:

$$b(\mathbf{x},\mathbf{y}) = \mathbf{y}^t A \mathbf{x}.$$

For each fixed $\mathbf{a} \in F^n$, the function $\mathbf{g}_\mathbf{a}(\mathbf{y}) = \mathbf{y}^t A \mathbf{a}$ is a linear functional. Similarly the function $\mathbf{f}_\mathbf{b}(\mathbf{x}) = b(\mathbf{v},\mathbf{b}) = \mathbf{b}^t A \mathbf{x}$ is a linear functional. So $b(\mathbf{x},\mathbf{y})$ is a bilinear form.

Theorem 6.3.3. *A bilinear form $b(\mathbf{v},\mathbf{w})$ on $V \times W$ gives rise to two linear maps D_1 and D_2:*

$$D_1\colon \mathbf{w} \in W \mapsto \mathbf{f}_w \in V^*;$$
$$D_2\colon \mathbf{v} \in V \mapsto \mathbf{g}_v \in W^*;$$

for the functions $\mathbf{f}_\mathbf{w}$ and $\mathbf{g}_\mathbf{v}$ of Definition 6.3.1.

Proof. We only consider the case of D_1, since that of D_2 is nearly identical. First by hypothesis $\mathbf{f}_\mathbf{w}$ is a linear functional. The map D_1 is linear because of the linearly of $b(\mathbf{v},\mathbf{w})$ in its first variable: first,

$$b(\mathbf{v}_1 + \mathbf{v}_2, \mathbf{w}) = b(\mathbf{v}_1,\mathbf{w}) + b(\mathbf{v}_2,\mathbf{w}_2)$$

so $\mathbf{f}_{\mathbf{v}_1+\mathbf{v}_2}(\mathbf{w}) = \mathbf{f}_{\mathbf{v}_1}(\mathbf{w}) + \mathbf{f}_{\mathbf{v}_2}(\mathbf{w})$, and then $b(c\mathbf{v},\mathbf{w}) = cb(\mathbf{v},\mathbf{w})$ for any scalar c, so $f_{c\mathbf{v}}(\mathbf{w}) = c\mathbf{f}_\mathbf{v}(\mathbf{w})$. □

Our goal is to show that any bilinear form can be written using a matrix as in Example 6.3.2, once bases for V and W have been chosen. Thus we imitate what we did when we described linear maps by matrices in Theorem 5.1.1.

Theorem 6.3.4. *Let $b(\mathbf{x},\mathbf{y})$ be a bilinear form on $F^n \times F^m$. Use x_j for the coordinates on F^n, and y_i for the coordinates on F^m. Let $\mathbf{v}_j$ be the j-th unit coordinate vector on F^n and $\mathbf{w}_i$ be the i-th unit coordinate vector on F^m. Define the scalars a_{ij} by*

$$a_{ij} = b(\mathbf{v}_j, \mathbf{w}_i)$$

and let A be the $m \times n$ matrix (a_{ij}). Then the bilinear form $b(\mathbf{x},\mathbf{y})$ is written $\mathbf{y}^t A\mathbf{x}$, and the matrix A is uniquely determined by $b(\mathbf{x},\mathbf{y})$.

Proof. We reduce to Theorem 5.1.1. Consider the linear map D_2 that sends a vector $\mathbf{x} \in F^n$ to the linear functional $b(\mathbf{x},\bullet) \in W^*$, which has dimension m by Theorem 6.1.7. This is the functional that to every $\mathbf{y} \in W$ associates the scalar $b(\mathbf{x},\mathbf{y})$. We apply Theorem 5.1.1 to the linear map D_2: by choosing bases for V and for W^*, there is a unique $m \times n$ matrix A matrix representing D_2 in these bases. So the functional $b(\mathbf{x},\bullet)$ in W^* is the matrix product $A\mathbf{x}$, which is a m-vector. By Example 6.1.9 the value of this functional on $\mathbf{y} \in W$ is $\mathbf{y}^t A\mathbf{x}$. The uniqueness of A follows easily by letting $\mathbf{y}$ run through the standard basis of F^m, and $\mathbf{x}$ run through the standard basis of F^n. □

Corollary 6.3.5. *Consider instead the linear map $D_1 : F^m \to F^n$ which to a vector $\mathbf{y}$ associates the linear map $b(\bullet,\mathbf{y})$ in the dual of F^n. The associated matrix is A^t, the transpose of the matrix for D_2.*

Here is a corollary of Theorem 6.3.4 that reflects the uniqueness of A:

Corollary 6.3.6. *Let E_{ij} the the $m \times n$ matrix with 1 in position (i,i), and 0 everywhere else. Then the E_{ij}, $1 \le i \le m$, $1 \le j \le n$, form a basis of the matrices of bilinear forms on $V \times W$, such that*

$$\mathbf{w}_s^t E_{ij} \mathbf{v}_t = \begin{cases} 1 & \text{if } s = i \text{ and } t = j; \\ 0 & \text{otherwise.} \end{cases}$$

See §8.8 for some related results.

Of special interest to us is the case $W = V^*$, that we study next. Indeed to a $\mathbf{v} \in V$ and a $\mathbf{f} \in V^*$ we can associate the scalar $\mathbf{f}(\mathbf{v})$, the evaluation of $\mathbf{f}$ on $\mathbf{v}$. This is bilinear by definition. All bilinear forms on $V \times V^*$ can be interpreted in this way.

6.4 Annihilators

Definition 6.4.1. Let S be a subset of the vector space V. Then the annihilator S^a of S is the set of $\mathbf{f} \in V^*$ that vanish on S.

Exercise 6.4.2. The annihilator S^a of any subset $S \subset V$ is a vector subspace of V^*. Furthermore if W is the subspace of V spanned by S, then $W^a = S^a$.

Example 6.4.3. The annihilator of $\mathbf{0}$ is V^*; the annihilator of V is $\mathbf{0} \subset V^*$. Let W be a subspace of V of dimension $n-1$, where V has dimension n. Then W is called a *hyperplane* of V. W has a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}\}$ which can be extended by one vector $\mathbf{v}_n$ to a basis $\mathfrak{B}$ of V. Using the dual basis $\mathfrak{B}^*$ of V^*, we see that the annihilator W^a of W is spanned by $\mathbf{f}_n$, the functional dual to $\mathbf{v}_n$, so it is one-dimensional.

More generally we have:

Theorem 6.4.4. *If W is a subspace of dimension s of the n-dimensional vector space V, then its annihilator $W^a \subset V^*$ has dimension $n-s$.*

Proof. Use the same method as in the hyperplane example. □

Definition 6.4.5. If $W \subset V$ has dimension s, let $W^c \subset V$ denote any complement of W, namely any subspace of dimension $n-s$ such that $W \cap W^c = (\mathbf{0})$.

We can find a basis $\mathfrak{B}$ of V where $\mathbf{v}_1, \ldots, \mathbf{v}_s$ span W and $\mathbf{v}_{s+1}, \ldots, \mathbf{v}_n$ span W^c. Because W^c depends on the basis $\mathfrak{B}$, we could write $W^c_{\mathfrak{B}}$.

Theorem 6.4.6. *Under the isomorphism $D_{\mathfrak{B}}$, W^c is mapped to W^a, and W is mapped to $(W^c)^a$*

Proof. Using the basis $\mathfrak{B}$ defined above, it is clear that the dual basis elements $\mathbf{f}_{s+1}, \ldots, \mathbf{f}_n$ span W^a, and $\mathbf{f}_1, \ldots, \mathbf{f}_s$ span $(W^c)^a$ □

This theorem can be used to reprove an important result, already proved using row reduction in Theorem 5.6.5.

Theorem 6.4.7. *The row rank and the column rank of a matrix are equal.*

Proof. Let A be a $m \times n$ matrix. The rows of A span a subspace W of F^n of dimension s, the row rank of A. Let r be the column rank of A. This F^n will play the role of V above, so we will call it V.

Use the standard basis for V and use the dual basis for V^*. Then by Example 6.1.9 an element $\mathbf{f}$ of V^* is in W^a if and only if its coordinate vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ satisfies the equations

$$\mathbf{a}^i \mathbf{x} = 0, \text{ for all the rows } \mathbf{a}^i \text{ of } A.$$

This is equivalent to saying that $\mathbf{x}$ is in the nullspace N of the linear map $L_A : F^n \to F^m$ associated to the matrix A. So Theorem 6.4.4 tells us that $\dim N = n - s$. Combining with the Rank-Nullity Theorem, which says $\dim N + r = n$, we get $r = s$: the row rank of A is equal to its column rank. □

Example 6.4.8. Here is how the isomorphism $D_{\mathfrak{B}} : V \to V^*$ depends on the choice of basis in V, using the change of basis results of §5.3.

As usual we have a vector space V of dimension n with a basis $\mathfrak{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, and its dual space V^* with dual basis $\mathfrak{B}^* = \{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$. Then by construction the dual map $D_{\mathfrak{B}} : V \to V^*$ has the identity matrix as associated matrix $M^{\mathfrak{B}}_{\mathfrak{B}^*}(D_{\mathfrak{B}})$. Now take a second basis $\mathfrak{C} = \{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ for V with its dual basis $\mathfrak{C}^* = \{\mathbf{g}_1, \ldots, \mathbf{g}_n\}$

The bases $\mathfrak{B}$ and $\mathfrak{C}$ are related by the change of basis formula 5.9

$$\mathbf{v}^j = a_{1j}\mathbf{w}^1 + a_{2j}\mathbf{w}^2 + \cdots + a_{nj}\mathbf{w}^n, \quad 1 \leq j \leq n, \tag{6.7}$$

so the change of basis matrix $[I_V]^{\mathfrak{B}}_{\mathfrak{C}}$ is $A = (a_{ij})$. Here I_V is the identity mapping on V. Apply the functional $\mathbf{g}_k$ to (6.7) to get

$$\mathbf{g}_k(\mathbf{v}_j) = a_{kj} \tag{6.8}$$

Since the $\mathbf{f}_j$ are the dual basis to the $\mathbf{v}_j$, this implies that

$$\mathbf{g}_k = a_{k1}\mathbf{f}_1 + a_{k2}\mathbf{f}_2 + \cdots + a_{kn}\mathbf{f}_n.$$

Comparing this formula to (6.7) shows that $[I_V]^{\mathfrak{C}^*}_{\mathfrak{B}^*} = A^t$, because the order of the indices has been reversed. Finally by construction $[D_{\mathfrak{B}}]^{\mathfrak{B}}_{\mathfrak{B}^*} = I$, and $[D_{\mathfrak{C}}]^{\mathfrak{C}}_{\mathfrak{C}^*} = I$. This is simply because we have taken the dual bases. Now we write the dual map $D_{\mathfrak{C}}$ of $V \to V^*$, but express it in the $\mathfrak{B}$ basis.

By Theorem 5.3.6 the matrix of $D_{\mathfrak{C}}$ expressed in the bases $\mathfrak{B}$ for the domain, and $\mathfrak{B}^*$ for the target is

$$[I_V]^{\mathfrak{C}^*}_{\mathfrak{B}^*} \circ [D_{\mathfrak{C}}]^{\mathfrak{C}}_{\mathfrak{C}^*} \circ [I_V]^{\mathfrak{B}}_{\mathfrak{C}} = A^t I A = A^t A.$$

This is the identity matrix, and therefore the same as $M^{\mathfrak{B}}_{\mathfrak{B}^*}(D_{\mathscr{B}})$ if and only if $A^t A = I$. Such matrices are called orthogonal matrices. We will study them and give examples in §8.3.

Remark 6.4.9. The isomorphism $D_{\mathfrak{B}}$ depends on the basis $\mathfrak{B}$ and maps W^c to W^a by Theorem 6.4.6. This is why we can only make the weak statement that $W \oplus W^c = V$ in Theorem 6.4.4. Given a subspace W of dimension m inside a space V of dimension n, there are infinitely many subspaces W' of dimension $n-m$ such that $W \oplus W' = V$. By Example 6.4.8, each such W' is a $W^c_{\mathfrak{B}}$ for a suitable basis $\mathfrak{B}$.

6.5 The Double Dual

Given a n-dimensional vector space V, we have constructed its dual space V^*, which also has dimension n. The isomorphism between V and V^* depends on the choice of a basis for V as we showed in Example 6.4.8.

Next we can take the dual of V^*, the *double dual* of V, written V^{**}. Since V^* has dimension n, V^{**} again has dimension n. It also has an isomorphism with V that does not depend on a choice of bases, something V^* does not have. We now construct this isomorphism.

Definition 6.5.1. Pick a $\mathbf{v} \in V$. The map $\mathbf{e}_\mathbf{v} : V^* \to F$ given by:

$$\mathbf{e}_\mathbf{v}(\mathbf{f}) = \mathbf{f}(\mathbf{v}), \text{ for any } \mathbf{f} \in V^*$$

is called *evaluation at* $\mathbf{v}$. The map $\mathbf{e}_\mathbf{v}$ is easily seen to be a linear functional on V^*, so $\mathbf{e}_\mathbf{v} \in V^{**}$.

Exercise 6.5.2. Prove $\mathbf{e}_\mathbf{v} \in V^{**}$. You must show $\mathbf{e}_\mathbf{v}(\mathbf{f}_1 + \mathbf{f}_2) = \mathbf{e}_\mathbf{v}(\mathbf{f}_1) + \mathbf{e}_\mathbf{v}(\mathbf{f}_2)$ and $\mathbf{e}_\mathbf{v}(c\mathbf{f}) = c\mathbf{e}_\mathbf{v}(\mathbf{f})$ for any scalar c.

Theorem 6.5.3. *The map $D_2 : V \to V^{**}$ given by $\mathbf{v} \mapsto \mathbf{e}_\mathbf{v}$ is an isomorphism of V with V^{**}. It is called the* natural correspondence *between V and V^{**}.*

Proof. We first show D_2 is a linear map. The main point is that for two elements $\mathbf{v}$ and $\mathbf{w}$ of V,

$$\mathbf{e}_{\mathbf{v}+\mathbf{w}} = \mathbf{e}_\mathbf{v} + \mathbf{e}_\mathbf{w}.$$

To show this we evaluate $\mathbf{e}_{\mathbf{v}+\mathbf{w}}$ on any $\mathbf{f} \in V^*$:

$$\mathbf{e}_{\mathbf{v}+\mathbf{w}}(\mathbf{f}) = \mathbf{f}(\mathbf{v}+\mathbf{w}) = \mathbf{f}(\mathbf{v}) + \mathbf{f}(\mathbf{w}) = \mathbf{e}_\mathbf{v}(\mathbf{f}) + \mathbf{e}_\mathbf{w}(\mathbf{f})$$

just using the linearity of $\mathbf{f}$. Thus $D_2(\mathbf{v}+\mathbf{w}) = D_2(\mathbf{v}) + D_2(\mathbf{w})$. The remaining point $D_2(c\mathbf{v}) = cD_2(\mathbf{v})$ is left to you.

To show D_2 is an isomorphism, all we have to do is show D_2 is injective by the Rank-Nullity theorem. Suppose D_2 is not injective: then there is a $\mathbf{v}$ such that $\mathbf{e}_\mathbf{v}$ evaluates to 0 on all $\mathbf{f} \in V^*$, so $\mathbf{f}(\mathbf{v}) = 0$. But that is absurd: all functionals cannot vanish at a point. For example extend $\mathbf{v}$ to a basis of V and let $\mathbf{f}$ be the element in the dual basis that is dual to $\mathbf{v}$, so $\mathbf{f}(\mathbf{v}) = 1$. □

Thus we can identify V and V^{**} using the isomorphism D_2. This has a nice consequence:

Corollary 6.5.4. *Given an arbitrary basis $\mathfrak{B}^*$ of V^*, there exists a basis $\mathfrak{B}$ of V for which it is the dual.*

Proof. Just take the dual basis $\mathfrak{B}^{**}$ of V^{**}, and then use the isomorphism D_2 to get a basis $\mathfrak{B}$ of V of which $\mathfrak{B}^*$ is the dual. □

Remark 6.5.5. When we write $\mathbf{f}(\mathbf{v})$ we can either think of $\mathbf{f}$ as a functional acting on $\mathbf{v} \in V$ or $\mathbf{v}$ as a functional in V^{**} acting on $\mathbf{f} \in V^*$. This suggests that we use a more symmetric notation, say $(\mathbf{f}, \mathbf{v})$. Indeed, as we will define in Chapter 7, this is a bilinear form on the two spaces, because, as we have already checked:

- $(\mathbf{f}_1 + \mathbf{f}_2, \mathbf{v}) = (\mathbf{f}_1, \mathbf{v}) + (\mathbf{f}_2, \mathbf{v})$ and $(\mathbf{f}, \mathbf{v}_1 + \mathbf{v}_1) = (\mathbf{f}, \mathbf{v}_1) + (\mathbf{f}, \mathbf{v}_2)$.

- $(c\mathbf{f}, \mathbf{v}) = c(\mathbf{f}, \mathbf{v})$ and $(\mathbf{f}, c\mathbf{v}) = c(\mathbf{f}, \mathbf{v})$.

The annihilator of a subspace W of V is the subspace of functionals $\mathbf{f}$ such that $(\mathbf{f}, \mathbf{w}) = 0$ for all $\mathbf{w} \in W$. The annihilator of a subspace W of dimension r has dimension $n - r$. We can also consider the annihilator $(W^*)^a$ of a subspace W^* of V^*: $(W^*)^a$ is a subspace of V^{**}, but using the natural identification of V^{**} with V, we can view it as a subspace of V. The dimension result still holds.

6.6 Duality

We extend the results on linear functionals from §6.4 by applying the duality construction of §B.4.

Suppose we have a vector space V of dimension n, a vector space W of dimension m, and a linear map L between them:

$$L\colon V \to W.$$

To each linear functional $\mathbf{g} \in W^*$, which is a linear map $W \to F$, we can associate the composite linear map:

$$\mathbf{g} \circ L\colon V \to W \to F.$$

Exercise 6.6.1. Check that $\mathbf{g} \circ L$ is a linear functional on V.

Definition 6.6.2. The linear map $L^*\colon W^* \to V^*$ given by

$$L^*\colon \mathbf{g} \in W^* \longmapsto \mathbf{f} = \mathbf{g} \circ L \in V^*.$$

is called the *dual* or the transpose of L.

Theorem 6.6.3. *The annihilator $R(L)^a \subset W^*$ of the range $R(L) \subset W$ of L is the nullspace $N(L^*)$ of L^* in W^*. Similarly the annihilator $R(L^*)^a \subset V$ of the range $R(L^*) \subset V^*$ is the nullspace $N(L)$ of L in V. Furthermore L and L^* have the same rank.*

Proof. Let r be the rank of L so that the range $R(L) \subset W$ of L has dimension r and the nullspace $N(L)$ of L has dimension $n - r$ (by the Rank-Nullity theorem). The nullspace $N(L^*)$ of L^* is the collection of linear functionals $\mathbf{g} \in W^*$ such that $L^*(\mathbf{g})$ is the 0 functional on V, meaning that $L^*(\mathbf{g})(\mathbf{v}) = \mathbf{g} \circ L(\mathbf{v}) = 0$ for all $\mathbf{v} \in V$. This just says that $\mathbf{g}$ annihilates $L(\mathbf{v})$. So $N(L^*) = R(L)^a$.

Next, $\dim R(L^*) = m - \dim N(L^*)$ by Rank-Nullity again. By Theorem 6.4.4 $R(L)^a$ has dimension $m - r$, therefore $\dim R(L^*) = r$, so L and L^* have the same rank.

Finally a functional $\mathbf{f} \in R(L^*)$ is written $\mathbf{f} = \mathbf{g} \circ L$. Obviously any $\mathbf{v}$ in the nullspace of L is in the nullspace of the composite $\mathbf{g} \circ L$, so it is annihilated by $R(L^*)$. Thus $N(L) \subset R(L^*)^a$, which has dimension r, as we just established. By Theorem 6.4.4

again, applied in V, this annihilator has dimension $n-r$, so $\dim N \leq n-r$. The Rank-Nulllity theorem applied to L then tells us we have equality, so $N(L) = R(L^*)^a$.

This last result can be established in a way parallel to the first argument of the proof by identifying V^{**} with V, W^{**} with W, and L^{**} with L under the previous identifications. □

Now assume we are given a basis $\mathfrak{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of V and a basis $\mathfrak{C} = \{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ of W. What is the relationship between the $m \times n$ matrix $[L]_{\mathfrak{C}}^{\mathfrak{B}}$ associated to L and the $n \times m$ matrix $[L^*]_{\mathfrak{B}^*}^{\mathfrak{C}^*}$ of L^*, in the dual bases discussed in §6.4?

This is settled by the following theorem. It is a special case of Corollary 6.3.5, but we repeat the proof for convenience.

Theorem 6.6.4. *One matrix is the transpose of the other:*

$$[L^*]_{\mathfrak{B}^*}^{\mathfrak{C}^*} = \left([L]_{\mathfrak{C}}^{\mathfrak{B}}\right)^t.$$

Proof. Let $(x_1, x_2, \ldots, x_n)$ be the coordinates of a vector $\mathbf{v}$ in the $\mathfrak{B}$ basis of V. Then if A in the $m \times n$ matrix of L, $A\mathbf{x}$ is the vector of coordinates of $L(\mathbf{v})$ in the $\mathfrak{C}$ basis. Now apply a functional $\mathbf{g} \in W^*$ with coordinates $(y_1, \ldots, y_m)$ in the dual basis. Then by Example 6.1.9 it evaluates to

$$\mathbf{y}^t(A\mathbf{x}) = \mathbf{x}^t A^t \mathbf{y}$$

If B is the $n \times m$ matrix of L^* in the bases $\mathfrak{C}$ and $\mathfrak{B}$, then applied to $\mathbf{g} \in W^*$, you get the functional $B\mathbf{y} \in V^*$. By Example 6.1.9, evaluating this functional on $\mathbf{v}$ gives $\mathbf{x}^t B\mathbf{y}$. Since this is true for all $\mathbf{y}$ and $\mathbf{x}$, $B = A^t$ as we saw in Theorem 6.3.4. □

Exercise 6.6.5. Explain how the computation done in Example 6.4.8 is a special case of the theorem.

Exercise 6.6.6. Choose bases for V and W according to Theorem 5.4.1, and work out what we have done in this section completely for this case.

Now suppose we have a third vector space U of dimension p and a linear map $M\colon U \to V$. To each linear functional $\mathbf{g} \in W^*$, which is a linear map $W \to F$, we can associate the composite map:

$$\mathbf{g} \circ L\colon V \to W \to F.$$

Then we can compose the linear maps $L\colon V \to W$ and $M\colon U \to V$ to get $L \circ M$.

Theorem 6.6.7. *The transpose of the composite $(L \circ M)$ is the composite of the transposes, but in the opposite order:*

$$(L \circ M)^* = M^* \circ L^*.$$

Note that this makes sense at the level of the matrix representatives, since $(AB)^t = B^t A^t$.

Proof. On one hand we have the composite map $L \circ M \colon U \to W$, so that its transpose $(L \circ M)^* \colon W^* \to U^*$ maps $\mathbf{g} \in W^*$ to $\mathbf{g} \circ (L \circ M) \in U^*$.

On the other hand the transpose $L^* \colon W^* \to V^*$ maps $\mathbf{g}$ to $\mathbf{f} = \mathbf{g} \circ L$ and M^* maps $\mathbf{f} \in V^*$ to $\mathbf{e} \in U^*$ where $\mathbf{e} = \mathbf{f} \circ M$.

Putting this together we get, doing the maps one at a time,

$$\mathbf{g} \mapsto \mathbf{e} = \mathbf{f} \circ M = (\mathbf{g} \circ L) \circ M = \mathbf{g} \circ (L \circ M) = (L \circ M)^* \mathbf{g}$$

as claimed. Notice that other than Definition 6.6.2 all we used is the associativity of composition of maps in the very last step. □

Chapter 7
Bilinear Forms

Abstract Bilinear forms are a new kind of mapping on a vector space. We study them the same way we studied linear maps by representing them by matrices. The main goal is to classify bilinear forms in terms of a basis of the vector space. The result is simpler than the result we will derive for linear transformations later in this book in Chapter 10, since any bilinear form can be diagonalized by an appropriate choice of basis, while is not the case for linear transformations: see the Jordan canonical form in §10.6.

7.1 Bilinear Forms

Now we specialize the results of §6.3 to the case $W = V$. We simplify the notation $b(\mathbf{v},\mathbf{w})$ for the bilinear form to $(\mathbf{v},\mathbf{w})$. We will mainly be interested in bilinear forms that satisfy an additional property.

Definition 7.1.1. A bilinear form $(\mathbf{v},\mathbf{w})$ on $V \times V$ is symmetric if it satisfies

$$(\mathbf{v},\mathbf{w}) = (\mathbf{w},\mathbf{v}) \text{ for all } \mathbf{v} \text{ and } \mathbf{w} \text{ in } V.$$

If $(\mathbf{v},\mathbf{w})$ is symmetric then the two linear maps D_1 and D_2 of Theorem 6.3.3 are identical. We will just call it D.

If the matrix A is symmetric in Example 6.3.2, then $(\bullet,\bullet)$ is a symmetric bilinear form, since if $A^t = A$,

$$(\mathbf{v},\mathbf{w}) = \mathbf{y}^t A\mathbf{x} = \mathbf{y}^t A^t\mathbf{x} = (\mathbf{y}^t A^t)\mathbf{x} = \mathbf{x}^t A\mathbf{y} = (\mathbf{w},\mathbf{v}).$$

Example 7.1.2. Let A be the square matrix of size 2:

$$A = \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix},$$

then

$$\mathbf{y}^t A\mathbf{x} = \begin{pmatrix} y_1 & y_2 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} y_1 & y_2 \end{pmatrix} \begin{pmatrix} 2x_1 + x_2 \\ 3x_1 + 4x_2 \end{pmatrix} = 2x_1y_1 + x_2y_1 + 3x_1y_2 + 4x_2y_2.$$

Check that this bilinear form is not symmetric.

Example 7.1.3. If we make the matrix A symmetric by averaging its off-diagonal terms, we get the symmetrio bilinear form:

$$\mathbf{y}^t A\mathbf{x} = \begin{pmatrix} y_1 & y_2 \end{pmatrix} \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 2x_1y_1 + 2x_2y_1 + 2x_1y_2 + 4x_2y_2$$

Notice that the coefficients of the crossterms x_2y_1 and x_1y_2 are the same, as they always will be if A is symmetric.

Using Theorem 6.3.4, we see that a bilinear form on $V \times V$ is symmetric if and only if the matrix A representing it is symmetric. We use, of course, the same basis $\mathfrak{B}$ on both copies of V. We say that the bilinear form b is represented by the matrix A in the basis $\mathfrak{B}$.

Definition 7.1.4. The symmetric bilinear form $(\mathbf{v}, \mathbf{w})$ is *non-degenerate* if the only $\mathbf{v} \in V$ for which

$$(\mathbf{v}, \mathbf{w}) = 0 \text{ for all } \mathbf{w} \in V,$$

is $\mathbf{v} = \mathbf{0}$.

Definition 7.1.5. We say that $\mathbf{v}$ is perpendicular, or orthogonal, to $\mathbf{w}$ if

$$(\mathbf{v}, \mathbf{w}) = 0.$$

We write $\mathbf{v} \perp \mathbf{w}$ if this is the case. For any subset S of V, we let $S^\perp$ be the set of elements of V perpendicular to all the elements of S. We call $S^\perp$ the *orthogonal complement* of S. This notion of course depends on the choice of the bilinear form.

Exercise 7.1.6. Prove $S^\perp$ is a subspace of V. Let U be the subspace of V generated by the elements of S. Show that $S^\perp = U^\perp$.

Definition 7.1.7. The *radical* of $(\mathbf{v}, \mathbf{w})$ is the orthogonal complement $V^\perp$ of the full space V. We use the notation ν for $\dim V^\perp$.

The radical is $(\mathbf{0})$ if and only if the bilinear form is non-degenerate, by Definition 7.1.4. We can describe the radical in terms of the linear map $D\colon V \to V^*$ given by $D(\mathbf{v}) = (\mathbf{v}, \bullet)$.

Theorem 7.1.8. *The linear map D is an isomorphism if and only if $(\mathbf{v}, \mathbf{w})$ is non-degenerate. The nullspace of D is the radical $V^\perp$. The rank of D is $n - \dim V^\perp$, and is called the rank of $(\mathbf{v}, \mathbf{w})$.*

Proof. An element $\mathbf{v}$ is in the nullspace of D if and only if the linear functional $\mathbf{f}_\mathbf{v}$ is identically 0, meaning that for all $\mathbf{w} \in V$, $(\mathbf{v}, \mathbf{w}) = 0$. Thus the nullspace of D is the radical of V. Because V and V^* have the same dimension, D is an isomorphism if and only if the nullspace is reduced to $(\mathbf{0})$. The last statement is just the Rank-Nullity Theorem applied to D. □

By Theorem 6.3.4 a symmetric bilinear form form gives rise to a symmetric matrix. In fact it gives rise to many symmetric matrices, depending on the basis used for V. We want to know how the matrix varies in terms of the basis.

Theorem 7.1.9. *If the symmetric quadratic form* $(\mathbf{v},\mathbf{w})$ *is written* $\mathbf{y}^tA\mathbf{x}$ *in the basis* $\mathfrak{B}=\{\mathbf{v}^1,\dots,\mathbf{v}^n\}$*, then it is written* $\mathbf{y}^tC^tAC\mathbf{x}$ *in the basis* $\mathfrak{C}=\{\mathbf{w}^1,\dots,\mathbf{w}^n\}$*, where* C *is the change of basis matrix* C *from the* $\mathfrak{B}$ *basis to the* $\mathfrak{C}$ *basis.*

Proof. By Corollary 5.3.3

$$[\mathbf{v}]_{\mathfrak{C}}=[I]_{\mathfrak{C}}^{\mathfrak{B}}[\mathbf{v}]_{\mathfrak{B}}, \tag{7.1}$$

Here $C=[I]_{\mathfrak{C}}^{\mathfrak{B}}$. Because $\mathbf{x}=C\mathbf{y}$, and therefore $\mathbf{y}^tC^t=\mathbf{x}^t$, the symmetric matrix A representing the bilinear form $(\mathbf{v},\mathbf{w})$ in the basis $\mathfrak{B}$ is replaced by the symmetric matrix C^tAC representing the form in the basis $\mathfrak{C}$. Corollary 5.3.7 shows C is invertible. To show that C^tAC is symmetric, just take its transpose. □

This yields a equivalence relation on symmetric matrices of size n, known as congruence:

Definition 7.1.10. The symmetric matrix A of size n is *congruent* to a symmetric matrix B of the same size if there is an invertible matrix C such that $B=C^tAC$.

Example 7.1.11. Let A be the symmetric matrix

$$\begin{pmatrix}2&0\\0&3\end{pmatrix}$$

and C the invertible matrix

$$\begin{pmatrix}1&1\\0&2\end{pmatrix}$$

Then the matrix

$$B=C^tAC=\begin{pmatrix}1&0\\1&2\end{pmatrix}\begin{pmatrix}2&0\\0&3\end{pmatrix}\begin{pmatrix}1&1\\0&2\end{pmatrix}=\begin{pmatrix}2&2\\2&14\end{pmatrix}$$

is congruent to A.

Proposition 7.1.12. *Congruence is an equivalence relation on symmetric matrices.*

Proof. Use the identity $(C^t)^{-1}=(C^{-1})^t$: the inverse of the transpose is the transpose of the inverse. See Exercise 2.3.9. The rest of the proof is nearly identical to that of Theorem 5.5.3, and is left to the reader. □

Congruence partitions symmetric matrices into *congruence classes* of congruent matrices. Our goal is to find the simplest matrix is each congruence class: we will see that there is a diagonal matrix.

Do not confuse congruence with similarity (Definition 5.5.2). Once we have established the Spectral Theorem 13.3.1, we will see the connection between these equivalence relations.

In Theorem 7.1.8 we defined the rank of a symmetric bilinear form as the rank of the linear map $D\colon V \to V^*$ it induces. We could also define it as the rank of any matrix A that represents it. This implies:

Theorem 7.1.13. *The rank of a symmetric matrix is an invariant of its congruence class: in other words, if two matrices are congruent, they have the same rank.*

Proof. There is nothing to do, since all matrices representing the same linear form have the same rank.. We can confirm this by computation. Let $B = C^t AC$ be a matrix congruent to A, so C is invertible by definition. The theorem follows from Corollary 4.3.8 applied to the matrix $(C^t)^{-1}B = AC$. Because C and $(C^t)^{-1}$ are invertible, the corollary says that A and B have the same rank. □

Example 7.1.14 (Hankel Forms). We can produce a symmetric matrix A of size n from $2n-1$ numbers $s_0, \ldots, s_{2n-2}$, by letting $a_{ij} = s_{i+j-2}$. Written out, this gives a symmetric matrix

$$A = \begin{pmatrix} s_0 & s_1 & s_2 & \cdots & s_{n-1} \\ s_1 & s_2 & s_3 & \cdots & s_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{n-1} & s_n & s_{n+1} & \cdots & s_{2n-2} \end{pmatrix} \tag{7.2}$$

called a *Hankel form*.

Notice that each square submatrix of a Hankel form is again a Hankel form. Hankel forms were investigated by the German mathematician Frobenius in the late nineteenth century: a good reference for his work is Gantmacher [8], V. 1, X.10. Frobenius showed how to compute the rank of a Hankel form in most circumstances.

Exercise 7.1.15. Given any symmetric matrix A, and any square matrix B of the same size, we have seen that the matrix B^tAB is symmetric. What can you say about the rank of B^tAB, if you know the rank of A and of B?

Hint: use Frobenius's Inequality, Exercise 4.3.9, which builds on Sylvester's Law of Nullity, Theorem 4.3.7.

Now we carry this analysis one step further.

Theorem 7.1.16. *The set of all bilinear forms on V is a vector space.*

Proof. Use Example 3.2.7. We add two bilinear maps $f(\mathbf{v},\mathbf{w})$ and $g(\mathbf{v},\mathbf{w})$ by setting $f+g$ to be the map such that

$$(f+g)(\mathbf{v},\mathbf{w}) = f(\mathbf{v},\mathbf{w}) + g(\mathbf{v},\mathbf{w})$$

and for any scalar c, $(cf)(\mathbf{v},\mathbf{w}) = c(f(\mathbf{v},\mathbf{w}))$. You should check that these maps are bilinear. This is the vector space structure. □

Recall that $\mathscr{L}(V,V^*)$ is the vector space of linear maps from V to its dual space V^*, according to Theorem 4.1.11. The dual space V^* of V is discussed in §6.4. Next we define a linear map M from the vector space B of bilinear forms on V to

$\mathscr{L}(V,V^*)$. This is the map we have been implicitly discussing above: to the bilinear form $(\mathbf{v},\mathbf{w})$ of a point in B we associate the linear map from V to V^* that maps $\mathbf{v}\in V$ to its $\mathbf{g}_\mathbf{v}\in V^*$, where $\mathbf{g}_\mathbf{v}(\mathbf{w})=(\mathbf{v},\mathbf{w})$.

Theorem 7.1.17. *The linear map $M\colon B\to\mathscr{L}(V,V^*)$ is an isomorphism.*

Proof. It is easy to see that M is linear: this is left to you. We construct an inverse to M. So we start with a linear map $\mathbf{g}\colon V\to V^*$ that maps any $\mathbf{v}\in V$ to a linear functional $\mathbf{g}_\mathbf{v}\in V^*$. This functional $\mathbf{g}_\mathbf{v}$ can be evaluated at any $\mathbf{w}\in V$, giving $\mathbf{g}_\mathbf{v}(\mathbf{w})$. This gives the bilinear form $(\mathbf{v},\mathbf{w})$ with values $\mathbf{g}_\mathbf{v}(\mathbf{w})$. This is the inverse of M, so it is automatically linear and we are done. □

Remark 7.1.18. The proof is written without resorting to a basis for V and V^*. In coordinates the proof becomes easier. You should write it down.

7.2 Quadratic Forms

To each symmetric bilinear form $(\mathbf{w},\mathbf{v})$ on $V\times V$ we associate the scalar valued function on V given by

$$q(\mathbf{v})=(\mathbf{v},\mathbf{v}).$$

This function is called the quadratic form on V associated to $(\bullet,\bullet)$. We have:

Lemma 7.2.1. *The quadratic form $q(\mathbf{v})$ associated to $(\mathbf{v},\mathbf{w})$ satisfies:*

1. The ***polarization identities***

$$(\mathbf{v},\mathbf{w})=\frac{q(\mathbf{v}+\mathbf{w})-q(\mathbf{v}-\mathbf{w})}{4}. \tag{7.3}$$

and

$$(\mathbf{v},\mathbf{w})=\frac{q(\mathbf{v}+\mathbf{w})-q(\mathbf{v})-q(\mathbf{w})}{2}. \tag{7.4}$$

2. an identity only involving q:

$$q(\mathbf{v})+q(\mathbf{w})=\frac{q(\mathbf{v}+\mathbf{w})+q(\mathbf{v}-\mathbf{w})}{2} \tag{7.5}$$

3. For any scalar c, $q(c\mathbf{x})=c^2q(c\mathbf{x})$.

Proof. For 1) just expand the right hand side using the bilinearity and symmetry of $(\mathbf{v},\mathbf{w})$. For example, for (7.3) use the bilinearity and the symmetry of $(\mathbf{v},\mathbf{w})$ to get

$$\begin{aligned}q(\mathbf{v}+\mathbf{w})&=(\mathbf{v}+\mathbf{w},\mathbf{v}+\mathbf{w})\\&=(\mathbf{v},\mathbf{v})+(\mathbf{v},\mathbf{w})+(\mathbf{w},\mathbf{v})+(\mathbf{w},\mathbf{w})\\&=(\mathbf{v},\mathbf{v})+2(\mathbf{v},\mathbf{w})+(\mathbf{w},\mathbf{w})\end{aligned}$$

and similarly

$$q(\mathbf{v}-\mathbf{w}) = (\mathbf{v},\mathbf{v}) - 2(\mathbf{v},\mathbf{w}) + (\mathbf{w},\mathbf{w}).$$

Then subtract, and divide by 4.

For 2) subtract (7.4) from (7.3) and rearrange. For 3) by linearity in each factor, again:

$$q(c\mathbf{v}) = (c\mathbf{v},c\mathbf{v}) = c(\mathbf{v},c\mathbf{v}) = c^2(\mathbf{v},\mathbf{v}) = c^2 q(\mathbf{v}).$$

□

The third property is the reason why these functions are called quadratic: they are homogeneous functions of degree 2. The second property is verified for all homogeneous polynomials of degree 2, as you should check.

Equip V with the basis $\mathfrak{B} = \{\mathbf{v}_1,\ldots,\mathbf{v}_n\}$. If the symmetric bilinear form has the symmetric matrix A in this basis, then the associated quadratic form is written:

$$q(\mathbf{v}) = \mathbf{v}^t A\mathbf{v} = \sum_{i=1}^{n}\Big(\sum_{j=1}^{n} a_{ij}x_i x_j\Big), \tag{7.6}$$

where the x_i are the coefficients of $\mathbf{v}$ with respect to the basis. Thus $q(\mathbf{v})$ is a polynomial of degree 2 in the coefficients x_i.

By the polarization identities, we can reconstruct all the entries of a matrix A associated to the symmetric bilinear form from the values of its quadratic form q. Indeed,

$$a_{ij} = (\mathbf{v}_i,\mathbf{v}_j) = \frac{q(\mathbf{v}_i+\mathbf{v}_j) - q(\mathbf{v}_i-\mathbf{v}_j)}{4}.$$

Thus if we know the quadratic form, we know the associated biinear form. Quadratic forms are sometimes easier to deal with, since they only depend on one set of variables.

We often write quadratic forms as functions of the coefficients in the chosen basis. For example:

Example 7.2.2. Let $q(x_1,x_2,x_3) = x_1^2 + 2x_1x_2 - x_1x_3 - x_2^2 + x_2x_3 + 4x_3^2$. The associated matrix A is

$$\begin{pmatrix} 1 & 1 & -1/2 \\ 1 & -1 & 1/2 \\ -1/2 & 1/2 & 4 \end{pmatrix}$$

as you should check by carrying out the matrix multiplication $\mathbf{x}^T A\mathbf{x}$.

Remark 7.2.3. In this example, note that the off-diagonal terms in the matrix are half the coefficients in the quadratic polynomial. This is because we have not written separate coefficients for x_ix_j and x_jx_i in the polynomial, as we have in the sum in (7.6). If we write the summation differently, by starting the inner summation at i, we would have:

$$q(\mathbf{x}) = \sum_{i=1}^{n}\Big(\sum_{j=i}^{n} b_{ij}x_i x_j\Big).$$

For $i \neq j$, $b_{ij} = 2a_{ij}$, while $b_{ii} = a_{ii}$.

Exercise 7.2.4. In Example 7.1.3 a basis has already been chosen, and the associated quadratic form is $2x_1^2 + 4x_1x_2 + 4x_2^2$. Reconstruct the symmetric bilinear form from this quadratic form using a polarization identity. Note that the quadratic form associated to the non-symmetric bilinear form in Example 7.1.2 is the same.

7.3 Decomposition of a Symmetric Bilinear Form

Recall that the annihilator U^a of a subspace U of V is the subspace of elements of V^* that vanish on U: see §6.4 for details. On the other hand by Definition 7.1.5 $U^\perp$ is the subspace of elements $\mathbf{w} \in V$ such that $(\mathbf{u}, \mathbf{w}) = 0$ for all $\mathbf{u} \in U$. $U^\perp$ depends on the chosen bilinear form.

Theorem 7.3.1. *Let V be a vector space with a non-degenerate symmetric bilinear form $(\mathbf{v}, \mathbf{w})$, and U a subspace of V. Then the space $U^\perp \subset V$ associated to $(\mathbf{v}, \mathbf{w})$ is isomorphic to the annihilator $U^a \subset V^*$ of U via the linear map $D\colon V \to V^*$ defined by $D(\mathbf{v}) = (v, \bullet)$.*

Proof. By Theorem 7.1.8 D is an isomorphism if and only if the form $(\mathbf{v}, \mathbf{w})$ is non-degenerate. Since $U^\perp = \{\mathbf{v} \in V \mid (\mathbf{u}, \mathbf{v}) \forall \mathbf{u} \in U\}$ and $U^a = \{\mathbf{f} \in V^* \mid \mathbf{f}(\mathbf{u}) = 0 \forall \mathbf{u} \in U\}$ and the isomorphism D maps $\mathbf{v}$ to the functional $(v, \bullet)$ it is clear that $U^\perp$ is isomorphic to U^a under D. □

Notice the connection with Theorem 6.4.6.

Recall that the rank of a symmetric bilinear form is the rank of the linear map D, or the rank of any matrix A representing it: see Theorem 7.1.13.

Theorem 7.3.2. *If the radical of $(\mathbf{v}, \mathbf{w})$ on V has dimension ν, then the rank of $(\mathbf{v}, \mathbf{w})$ is $n - \nu$. In particular if the bilinear form $(\mathbf{v}, \mathbf{w})$ is non-degenerate, it has rank n.*

Proof. Since the rank does not depend of the choice of basis by Theorem 7.1.13, pick any basis $\mathfrak{B} = \{\mathbf{v}^1, \ldots, \mathbf{v}^n\}$ of V, in which the first ν basis elements form a basis of the radical. Then in this basis the matrix A looks like, in block notation:

$$\begin{pmatrix} 0 & 0 \\ 0 & B \end{pmatrix}$$

where B is a symmetric matrix of size $n - \nu$. The matrix B represents a non-degenerate bilinear form by construction. This reduction shows that it is enough to prove that the bilinear form $(\mathbf{v}, \mathbf{w})$ has rank n when it is non-degenerate.

The non-degeneracy assumption means that there is no non-zero $\mathbf{v} \in V$ such that $(\mathbf{v}, \mathbf{w}) = 0$ for all $\mathbf{w} \in V$. Writing $\mathbf{v}$ in coordinates for an arbitrary basis as $(x_1, \ldots, x_n)$ and $\mathbf{w}$ as $(y_1, \ldots, y_n)$, this means that there is no non-zero $\mathbf{x}$ such that

$$\mathbf{y}^t A \mathbf{x} = 0, \text{ for all } \mathbf{y} \in F^n. \tag{7.7}$$

If there happens to be a non-zero $\mathbf{x}$ with $A\mathbf{x} = \mathbf{0}$ then (7.7) fails. So the nullspace of the linear map $\mathbf{x} \mapsto A\mathbf{x}$ is trivial, so that A has rank n as desired. □

Remark 7.3.3. Let W be the subspace of V of dimension $n-\nu$ spanned by the vectors $\{\mathbf{v}^{\nu+1},\dots,\mathbf{v}^n\}$ constructed in the proof. Then clearly

$$V = V^\perp \oplus W. \tag{7.8}$$

so by Theorem 7.3.2 the rank of the $(\mathbf{v},\mathbf{w})$ restricted to W is $n-\nu$.

More generally we have a theorem that does not depend on Theorem 7.3.2. The method of proof is similar.

Theorem 7.3.4. *Let W be any subspace of V on which the restriction of the bilinear form $(\mathbf{v},\mathbf{w})$ has rank equal to* $\dim W$. *Then $V = W \oplus W^\perp$.*

Proof. Let m denote the dimension of W. Pick an arbitrary vector $\mathbf{v} \in V$. We must show that it can be written uniquely as $\mathbf{v} = \mathbf{w} + \mathbf{u}$, with $\mathbf{w} \in W$, and $\mathbf{u} \in W^\perp$. It is enough to show that $\mathbf{v} - \mathbf{w}$ is orthogonal to W, i.e. that

$$(\mathbf{v}-\mathbf{w},\mathbf{w}_i) = 0\ ,\ 1 \le i \le m, \text{ for a basis } \mathbf{w}_1,\dots,\mathbf{w}_m \text{ of } W. \tag{7.9}$$

Complete this basis of W to a basis of V. Write the unknown vector $\mathbf{w}$ as $x_1\mathbf{w}_1 + \cdots + x_m\mathbf{w}_m$ in this basis. We must solve the inhomogeneous system of m linear equations in the m variables x_i,

$$x_1(\mathbf{w}_1,\mathbf{w}_i) + x_2(\mathbf{w}_2,\mathbf{w}_i) + \cdots + x_m(\mathbf{w}_m,\mathbf{w}_i) = (\mathbf{v},\mathbf{w}_i),\ 1 \le i \le m.$$

Writing a_{ij} as usual for $(\mathbf{w}_i,\mathbf{w}_j)$, but this time just for $1 \le i,j \le m$, and $\mathbf{b}$ for the known vector $\big((\mathbf{v},\mathbf{w}_i)\big)$, $1 \le i \le m$, on the right hand side, we have the linear system in matrix notation

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

The square matrix A of this system has maximal rank by hypothesis, so the system can be solved uniquely for any $\mathbf{b}$. □

The key example to which we will apply this theorem in the next section is any subspace W of dimension 1 generated by a vector $\mathbf{w}$ with $(\mathbf{w},\mathbf{w}) \neq 0$. Here is another example that follows immediately from Remark 7.3.3.

Corollary 7.3.5. *Let $V^\perp$ denote the radical of V. Then if W is any subspace of V such that $V = V^\perp \oplus W$, then $(\mathbf{v},\mathbf{w})$ restricted to W has maximum rank.*

Therefore by Theorem 7.3.4, the orthogonal complement $W^\perp$ of W satisfies $V = W \oplus W^\perp$. Clearly $W^\perp = V^\perp$.

Remark 7.3.6. The results of this section can be proved using Theorem 5.10.5. Indeed, since a symmetric matrix represents a symmetric bilinear form in a basis $\mathfrak{B} = \{\mathbf{v}_1,\dots,\mathbf{v}_n\}$, the fact that the quadratic form has a rank r equal to the rank of a principal submatrix of size r shows that the symmetric quadratic form restricted to the basis vectors corresponding to the principal submatrix has maximal rank. From this you can easily show that the radical has dimension $n-r$, etc.

Definition 7.3.7 (Orthogonal Projection). Assume that the symmetric bilinear form has maximal rank on W, so that $V = W \oplus W^\perp$. Then the projection (see Definition 4.1.7) P from V to W along $W^\perp$ is defined by the unique solution to (7.9). It is called the *orthogonal projection* of V to W.

This follows from Theorem 7.3.4. By Theorem 4.6.2, $P^2 = P$.

Remark 7.3.8. To define an ordinary projection, we need two subspaces U and W such that $V = W \oplus U$, where W is the space you are projecting to: see §4.6. Different choices of U give rise to different projections to W. In the current situation the subspace U is uniquely determined by W and the symmetric bilinear form. So $W^\perp$ is its nullspace of the projection, and plays the role of U. The projection exists if the symmetric bilinear form b has maximal rank on W.

7.4 Diagonalization of Symmetric Bilinear Forms

Using the results of §7.2 and 7.3, we prove Theorem 7.4.1, one of the most important theorems in linear algebra. The following proof is not constructive, but it is very simple. Later we give a constructive proof using the Lagrange Algorithm 7.5.3.

Theorem 7.4.1. *Any symmetric matrix is congruent to a diagonal matrix.*

Proof. Assume that the symmetric matrix A has size n, and acts on a vector space V of dimension n with a given basis. This allows us to construct a symmetric bilinear form $(\mathbf{v}, \mathbf{w})$ in the usual way.

If the radical $V^\perp$ of $(\mathbf{v}, \mathbf{w})$ has dimension n, the matrix A is the zero matrix, so we are done.

So we may assume the $V^\perp$ has dimension $\nu < n$. Let W be a subspace of V of dimension $n - \nu$ such that $V = V^\perp \oplus W$. Then by Corollary 7.3.5, $(\mathbf{v}, \mathbf{w})$ restricted to W has maximum rank. Then we can find a vector $\mathbf{w}_1 \in W$ such that $(\mathbf{w}_1, \mathbf{w}_1) \neq 0$. Indeed if this were not true, the polarization identities would show us that $(\mathbf{u}, \mathbf{w})$ is identically 0 on W. This cannot be the case, since that would imply that W is part of the radical. So let W_1 be the orthogonal complement of $\mathbf{w}_1$ in W. By the same argument as before, we can find a $\mathbf{w}_2$ in W_1 with $(\mathbf{w}_2, \mathbf{w}_2) \neq 0$. Continuing in this way, we have found a collection of mutually perpendicular $\mathbf{w}_1$, $\mathbf{w}_2$, …, $\mathbf{w}_{n-r}$, all non-zero. In this basis the matrix of the symmetric bilinear form is written

$$\begin{pmatrix} 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & d_1 & \dots & 0 \\ \vdots & \ddots & \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & 0 & \dots & d_{n-\nu} \end{pmatrix} \tag{7.10}$$

where all the diagonal elements $d_1, d_2, \dots, d_{n-\nu}$ are non zero. □

The proof is not constructive, because there it does not give an algorithm for finding $\mathbf{w}_1$, $\mathbf{w}_2$, etc.

7.5 Lagrange's Diagonalization Algorithm

We now give a constructive, algorithmic proof of Theorem 7.4.1. We simplify symmetric matrices A by conjugation using equivalent row and column operations. Here is what equivalent means: on the left-hand side multiply A by a product of elementary matrices, which we call E. Thus we can row-reduce as in Gaussian elimination. On the right-hand side multiply A by the transpose E^t. Since A is symmetric, so is EAE^t, which is what we want. As we saw in Proposition 2.10.1, this means that we are column reducing A Finally if $E = E_nE_{n-1}\dots E_1$, then

$$EAE^t = E_nE_{n-1}\dots E_1AE_1\dots E_{n-1}E_n$$

so if we set $A_1 = E_1AE_1^t.\dots, A_k = E_kA_{k-1}E_k^t$, at each step we set a symmetric matrix A_k. So the goal is to find a suitable collection of E_n that achieve diagonallization. This is the content of Algorithm 7.5.3 below. We will use the elementary matrices T_{rs}, and $E_{rs}(c)$. We will also use a new elementary matrix

$$S_{ab} := E_{ba}(1)E_b(-2)E_{ab}(1).$$

Conjugation by this matrix is useful when the diagonal elements of A in rows r and s are both 0, and the element in position (r,s) (and therefore (s,r)) is non-zero. Note S_{ab} is symmetric.

First some examples.

Example 7.5.1. Start with the symmetric matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 5 \\ 3 & 5 & 9 \end{pmatrix}$$

To diagonalize, we first use $E_{21}(-2)$ which adds to the second row the first row multiplied by -2 to get

$$A_1 = E_{21}(-2)AE_{21}(-2)^t = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 5 \\ 3 & 5 & 9 \end{pmatrix} \begin{pmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 3 \\ 0 & -1 & -1 \\ 3 & -1 & 9 \end{pmatrix}.$$

We follow with $E_{31}(-3)$:

$$A_2 = E_{31}(-2)A_1E_{31}(-3)^t = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 3 \\ 0 & -1 & -1 \\ 3 & -1 & 9 \end{pmatrix} \begin{pmatrix} 1 & 0 & -3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & -1 \\ 0 & -1 & 0 \end{pmatrix}.$$

Finally conjugate by $E_{32}(-1)$:

$$A_3 = E_{32}(-1)A_2E_{32}(-1)^t = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & -1 \\ 0 & -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Check all these computations and notice how the matrix stays symmetric at each step.

Example 7.5.2. This example illustrates diagonalization using the new kind of elementary matrix. Let

$$A = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix} \quad \text{and} \quad E = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Then

$$EAE^t = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$$

as you should check. E is the product of three elementary matrices:

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = E_{21}(1)E_2(-2)E_{12}(1)$$

in terms of the elementary matrices, as you should also check.

Here is the algorithm for diagonalizing symmetric matrices. We will use the elementary matrices from Definition 2.8.1, and a generalization of the matrix E from Example 7.5.2. This is called Lagrange's method.

Algorithm 7.5.3 (Diagonalization algorithm for symmetric matrices) *The symmetric matrix $A = (a_{ij})$ is of size n. This algorithm diagonalizes A by repeatedly replacing it by a congruent matrix EAE^t, where E is invertible. For convenience set $j_0 = i_0$. Fix an integer $i_1 > i_0$ in the j_0-th row. We assume that we have the following situation. Normally we would start at $i_0 = 1$ and $i_1 = 2$, but this description is written in a way that it also describes all the intermediate steps.*

- *A has zeroes in all its off-diagonal elements of its first $i_0 - 1$ rows (and columns since it is symmetric). Thus if $i \neq j$, $a_{ij} = 0$ for $i < i_0$ and therefore for $j < j_0$.*
- *$a_{i,j_0} = 0$ when $i_0 < i < i_1$. Since A is symmetric, $a_{i_0,j} = 0$ when $j_0 < j < i_1$.*

Our goal is to increase the number of zeroes in the region given by the assumption, by incrementing (i.e., increasing by 1) i_1 until n is reached, and then incrementing i_0, keeping the matrix symmetric at each step. Here are the cases we need to consider:

1. *$i_1 = n$ and $a_{i_1,j_0} = 0$. We are at the bottom of the column. Just increment i_0.*
2. *$i_1 < n$ and $a_{i_1,j_0} = 0$. We are not at the bottom of the column. Just increment i_i.*
3. *$a_{i_1,j_0} \neq 0$ and $a_{i_0,i_0} \neq 0$. Then let*

$$c = \frac{a_{i_1,j_0}}{a_{i_0,j_0}}, \text{ and use the elementary matrix } E_{i_1,i_0}(-c) \text{ to clear the entry in } (i_1, j_0).$$

The new symmetric matrix $E_{i_1,i_0}(-c)A(E_{i_1,i_0}(-c))^t$ still has zeroes in all the entries indicated in the assumption and a new 0 in position (i_1, j_0). So we can increment i_1 if $i_1 < n$ or i_0 if $i_1 = n$.

4. *$a_{i_1,j_0} \neq 0$, $a_{i_0,i_0} = 0$ and there is diagonal entry $a_{i_2,i_2} \neq 0$ with $i_2 > i_0$. Then interchange the rows and columns i_0 and i_2 using the transposition T_{i_0,i_2}, namely taking the matrix $T_{i_0,i_2}AT_{i_0,i_2}$ since a transposition is its own transpose.*
5. *$a_{i_1,j_0} \neq 0$, $a_{i_0,i_0} = 0$ and there is no diagonal entry $a_{i_2,i_2} \neq 0$ with $i_2 > i_0$. This is the most difficult case, illustrated by Example 7.5.2. We have to diagonalize using a new matrix E which has a 1 along the diagonal except for $e_{i_2,i_2} = -2$. All the off-diagonal terms of E are 0, except $e_{i_0,i_2} = e_{i_2,i_0} = 1$. Note that E is symmetric and invertible. Then EAE has an extra 0 in position (i_0, i_2), and a non-zero element in positions (i_0, i_0) and (i_2, i_2).*

The algorithm terminates in a finite number of steps, since each step increases the number of 0 in the entries given by the initial assumption.

This algorithm is the analog of Gaussian elimination for a symmetric matrix. Note that we can read the rank of A from the congruent diagonal matrix A' obtained: it is just the number of non-zero elements on the diagonal.

7.6 Skew Symmetric Linear Forms

It is also useful to consider skew symmetric bilinear forms.

Definition 7.6.1. A bilinear form $(\mathbf{v}, \mathbf{w})$ on a vector space V is skew symmetric if

$$(\mathbf{v}, \mathbf{w}) = -(\mathbf{w}, \mathbf{v}) \text{ for all } \mathbf{v} \text{ and } \mathbf{w} \text{ in } V.$$

We have the analog of Theorem 6.3.4 which says that we can associate to an skew symmetric $(\mathbf{v}, \mathbf{w})$ a skew symmetric matrix A for each basis of V. Notice that the diagonal entries of A must be 0, since when $i = j$, $(i, j) = -(j, i)$. Thus the quadratic form associated to a skew symmetric form is identically 0.

Congruence forms an equivalence relation on skew symmetric matrices, so we can look for the matrix with the simplest form in each congruence class.

Theorem 7.6.2. *A skew symmetric bilinear form can be written in an appropriate basis as the block diagonal matrix*

$$\begin{pmatrix} A_1 & 0 & \dots & 0 & 0 \\ 0 & \ddots & 0 & \dots & 0 \\ 0 & \dots & A_k & 0 & \dots \\ \vdots & \dots & \dots & \vdots & 0 \\ 0 & \dots & \dots & \dots & 0 \end{pmatrix}$$

where each A_i is the 2×2 block

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Thus the rank of A is $2k$, twice the number of blocks A_i, and the last $n-2k$ rows and columns of A are 0.

The proof imitates that of Theorem 7.4.1. Say a little more.

7.7 Sylvesters Law of Inertia

In this section we improve the Diagonalization Theorem 7.4.1 when the scalars are the real numbers.

Recall that we write diagonal matrices as $D(d_1,\dots,d_n)$, where the d_i are the diagonal elements. So

Example 7.7.1. Let $q(\mathbf{x})$ be the quadratic form associated to the diagonal matrix $D(d_1,\dots,d_n)$. Then

$$q(\mathbf{x}) = d_1x_1^2 + d_2x_2^2 + \cdots + d_nx_n^2,$$

as you should check.

Exercise 7.7.2. By using the diagonal matrix

$$C = D(\sqrt{|d_1|},\dots,\sqrt{|d_n|})$$

verify that $D(d_1,\dots,d_n)$ is congruent to the diagonal matrix

$$B = CD(d_1,\dots,d_n)C^t,$$

where all the diagonal terms of B are either 0, 1 or -1. We get 1 when d_i is positive, -1 when it is negative, and 0 when it is 0.

Sylvester's Law of Inertia 7.7.6[1] shows that the following three numbers associated to a diagonal matrix D are congruence invariants of D, even though the diagonal entries d_i themselves are not.

Definition 7.7.3. Let B be an $n\times n$ diagonal real matrix with diagonal entries b_1, b_2, $\dots$, b_n. Then

- p is the number of positive $b_i, 1\le i\le n$.
- ν is the number of zero $b_i, 1\le i\le n$.
- m is the number of negative $b_i, 1\le i\le n$.

The triple of integers (p,ν,m) is called the *inertia* of B.

Note that $p+\nu+m=n$. The dimension of the radical of B is ν, so $n-\nu$ is the rank of B. Theorem 7.1.13 says ν is a congruence invariant of B.

[1] Published by J. J. Sylvester in 1852 - [29].

Example 7.7.4. If D is the diagonal matrix D(7, -1, 0, 3, 3, -2), then $p = 3$, $\nu = 1$, and $m = 2$.

Definition 7.7.5. The *signature* of a diagonal matrix B is the number $p - m$. If $p + m = n$, B is non-degenerate (sometimes called *non-singular*). If $p + m < n$, B is called degenerate or singular.

Theorem 7.7.6 (Sylvester's Law of Inertia). *Let A be a symmetric $n \times n$ matrix. By Theorem 7.4.1 it is congruent to a diagonal matrix B, which has an inertia. The inertia is a congruence invariant of A: it is the same for any diagonal matrix congruent to A. Conversely any diagonal matrix with the same inertia as B is congruent to B.*

Proof. Work with a real vector space V of dimension n. Since the radical of V is well defined independently of the basis, it is enough to prove the theorem for any subspace W of V such that $V = V^{\perp} \oplus W$. Then by Theorem 7.3.2, we may assume that $\nu = 0$. Assume we have two coordinate systems $\mathbf{e}$ and $\mathbf{f}$ in which the quadratic form q is diagonal. Let V_p and V_m be the subspaces of V spanned by the basis elements of $\mathbf{e}$ on which the quadratic from is positive and negative, respectively, and let W_p and W_m be the analogous subspaces for the $\mathbf{f}$-basis. Let p_V, m_V be the dimensions of V_p and V_m, and p_W, m_W the dimensions of W_p and W_m. Clearly $p_V + m_V = p_W + m_W = n$. We will show that $p_V = p_W$, from which it will follow that $m_V = m_W$.

We claim that the linear subspaces V_p and W_m of V do not intersect except at the origin. Suppose they did at a point $\mathbf{p} \neq \mathbf{0}$. Because $\mathbf{p} \in V_p$, we have $q(\mathbf{p}) > 0$, but because $\mathbf{p} \in W_m$, $q(\mathbf{p}) < 0$, a contradiction, so the claim is established.

This shows that $p_V \leq n - m_W = p_W$. Indeed, the $\mathbf{e}$-basis vectors spanning V_p, and the $\mathbf{f}$-basis vectors spanning W_m can be extended, by the claim, to a basis for V. Indeed, suppose not: then we would have an equation of linear dependence, which would express an element of V_p as an element of W_m, and this is precisely what we ruled out.

Exchanging the role of the V's and W's, we get $p_W \leq p_V$, so they are equal. This concludes the proof that (p, k, m) are congruence class invariants.

The converse follows easily: using the notation above, construct linear maps between V_p and W_p, between V_k and W_k, and between V_m and W_m sending basis elements to basis elements. This is possible since there are the same number of basis elements in all three cases. This gives the desired change of basis. The theorem is proved. □

The Law of Inertia allows us to talk about the signature of q: it is the signature of any diagonal matrix representing q.

Here are the main definitions concerning quadratic forms over $\mathbb{R}$.

Definition 7.7.7. The quadratic form $q(\mathbf{x})$ with matrix A is *definite* if $q(\mathbf{x}) \neq 0$ for all $\mathbf{x} \neq \mathbf{0}$.

We can refine this classification as follows.

Definition 7.7.8. The quadratic form $q(\mathbf{x})$ with matrix A is

- *Positive definite* if $\forall \mathbf{x} \neq \mathbf{0}$, $q(\mathbf{x}) > 0$, or, equivalently, $\mathbf{x}^t A\mathbf{x} > 0$;
- *Positive semidefinite* if $\forall \mathbf{x}$, $q(\mathbf{x}) \geq 0$, or, equivalently, $\mathbf{x}^t A\mathbf{x} \geq 0$;
- *Negative definite* if $\forall \mathbf{x} \neq \mathbf{0}$, $q(\mathbf{x}) < 0$, or, equivalently, $\mathbf{x}^t A\mathbf{x} < 0$;
- *Negative semidefinite* if $\forall \mathbf{x}$, $q(\mathbf{x}) \leq 0$, or, equivalently, $\mathbf{x}^t A\mathbf{x} \leq 0$;
- *Indefinite* if it does not fall into one of the four previous cases. Then it is not definite.

Example 7.7.9. The matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

associated to the quadratic form $q = x_1^2 - x_2^2$ is indefinite, because

$$\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 1, \text{ while } \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = -1$$

We pursue this in Example 7.7.10.

Example 7.7.10. This is a continuation of Example 7.7.9. Let V be a two-dimensional vector space with basis $\mathbf{e}_1$, $\mathbf{e}_2$, and write an element $\mathbf{v}$ of V as $x_1\mathbf{e}_1 + x_2\mathbf{e}_2$. Assume that the quadratic form q is represented in the $\mathbf{e}$-basis as $q(x_1, x_2) = x_1x_2$, so its matrix is

$$A = \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix}.$$

The bilinear form associated to q is

$$(\mathbf{x}, \mathbf{y}) = \frac{(x_1 + y_1)(x_2 + y_2) - (x_1 + y_1)(x_2 + y_2)}{4} = \frac{x_1y_2 + y_1x_2}{2},$$

by (7.3). We construct a diagonalizing basis as in Algorithm 7.5.3: we choose $\mathbf{f}_1 = a_1\mathbf{e}_1 + a_2\mathbf{e}_2$ with $q(\mathbf{f}_1) = a_1a_2 \neq 0$. So both a_1 and a_2 must be non-zero. We could normalize $\mathbf{f}_1$ so that $q(\mathbf{f}_1) = \pm 1$, by dividing by $\sqrt{a_1^2 + a_2^2}$, but we will not bother, to avoid burdening the computation. Then, following Algorithm 7.5.3, we consider the linear form $(\mathbf{x}, vf_1)$ and find an element $\mathbf{f}_2$ in its nullspace. This means solving for $\mathbf{x}$ in the equation $x_1a_2 + x_2a_1 = 0$. Up to multiplication by a non-zero scalar, we can take $\mathbf{x} = (a_1, -a_2)$, so that the second basis vector $\mathbf{f}_2 = a_1\mathbf{e}_1 - a_2\mathbf{e}_2$. If z_1 and z_2 are the coordinates in the $\mathbf{f}$-basis, the i-th column of the change of basis matrix satisfying $\mathbf{x} = E\mathbf{z}$ is the vector of coefficients of $\mathbf{f}_i$ in the $\mathbf{e}$-basis, so

$$E = \begin{pmatrix} a_1 & a_1 \\ a_2 & -a_2 \end{pmatrix}.$$

E is invertible because its determinant $-2a_1a_2 \neq 0$ by our choice of $\mathbf{f}^1$.

Then the matrix representing our quadratic form in the $\mathbf{f}$-basis is

$$B = E^T AE = \begin{pmatrix} a_1 & a_2 \\ a_1 & -a_2 \end{pmatrix} \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix} \begin{pmatrix} a_1 & a_1 \\ a_2 & -a_2 \end{pmatrix} = \begin{pmatrix} a_1a_2 & 0 \\ 0 & -a_1a_2 \end{pmatrix},$$

so, as predicted, it is diagonal, but with entries along the diagonal depending on a_1 and a_2. This shows there are infinitely many bases for V in which the quadratic form is diagonal. Even if one normalizes $\mathbf{f}_1$ and $\mathbf{f}_2$ to have length one, there is more than one choice. Our computation shows that in all of them, one of the diagonal entries is positive and the other is negative. The Law of Inertia 7.7.6 generalizes this computation.

Corollary 7.7.11. *A quadratic form q in $\mathbb{R}^n$ is:*
Positive definite, if its signature is n, which forces the rank to be n;
Positive semidefinite, if its signature is m, $m \leq n$, and its rank m;
Negative definite, if its signature is $-n$, which forces the rank to be n;
Negative semidefinite, if its signature is $-m$, $m \leq n$, and its rank m;
Indefinite, if its signature is less than the rank, so both p and m are positive.

Proof. Call the signature s and the rank r. Then $s = p - m$, $r = p + m$. Referring back to Definition 7.7.8, the proof is immediate.

Here is a second example showing what happens when the quadratic form does not have maximum rank.

Example 7.7.12. Using the same notation as in the previous example, assume that q can be written in the **e**-basis as $q(x_1, x_2) = x_1^2$, so its matrix is

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

The bilinear form associated to q is $(\mathbf{x}, \mathbf{y}) = x_1 y_1$, as per (7.3). Pick any vector $\mathbf{f}_1 = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 \in V$, so that $q(\mathbf{f}_1) \neq 0$. This just says that $a_1 \neq 0$. In this case we divide by a_1, and write $\mathbf{f}_1 = \mathbf{e}_1 + a\mathbf{e}_2$. Then, following Algorithm 7.5.3, consider the linear form $(\mathbf{x}, \mathbf{a}) = x_1$ and find a non-zero element $\mathbf{f}_2$ in its nullspace. Take $\mathbf{f}_2 = c\mathbf{e}_2$, for $c \neq 0$ Let

$$D = \begin{pmatrix} 1 & a \\ 0 & c \end{pmatrix}$$

be the change of basis matrix from the **e**-basis to the **f**-basis. D is invertible because its determinant $c \neq 0$ by choice of $\mathbf{f}_1$ and $\mathbf{f}_2$. Then we have

$$\begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{pmatrix} = D \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix} \text{ and } \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = D^T \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$

Then the matrix of our quadratic form in the **f**-basis is

$$B = DAD^T = \begin{pmatrix} 1 & a \\ 0 & c \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ a & c \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

so, as predicted, it is diagonal. In this example, because we normalized the length of the first new basis vector $\mathbf{f}_1$, then entries of the new diagonal matrix are the same as the ones we started with.

The form in Example 7.7.4 has signature 1. It is degenerate and indefinite.

Example 7.7.13. We compute the signature of the $n \times n$ symmetric matrix M_n with all diagonal terms equal to $n-1$ and all off diagonal terms equal to -1:

$$M_n = \begin{pmatrix} n-1 & -1 & \dots & -1 \\ -1 & n-1 & \dots & -1 \\ \dots & \dots & \dots & \dots \\ -1 & -1 & \dots & n-1 \end{pmatrix}$$

We will show that the signature and the rank are $n-1$, so that the form is positive semidefinite. We do this by first computing the signature for $n = 2$ and then setting up a proof by induction. Letting $n = 2$, we get

$$M_2 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

By using symmetric Gaussian elimination we can transform this to the diagonal matrix $(1,0)$, so $p = 1$, $k = 1$ and $m = 0$. We are done. Next

$$M_3 = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

By symmetric Gaussian elimination again, this transforms our matrix into the congruent matrix: We get

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{3}{2} & -\frac{3}{2} \\ 0 & -\frac{3}{2} & \frac{3}{2} \end{pmatrix}$$

and the 2×2 matrix in the bottom right is just M_2 multiplied by $\frac{3}{2}$. The 1 in upper left-hand corner just adds 1 to the signature we found in the case $n = 2$, so the signature is $(2,0)$. This suggests the general strategy: we prove by induction that the signature of M_n is $n-1$ and the rank $n-1$. By row reduction, first dividing the top row by $n-1$, and then clearing the first column, you get

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \frac{n(n-2)}{n-1} & \dots & -\frac{n}{n-1} \\ \dots & \dots & \dots & \dots \\ 0 & -\frac{n}{n-1} & \dots & \frac{n(n-2)}{n-1} \end{pmatrix}$$

The bottom right matrix of size $(n-1) \times (n-1)$ is $\frac{n}{n-1}$ times the matrix M_{n-1}. By induction we know that the signature and the rank of M_{n-1} are both $n-2$ and we are done. Note that we are using Sylvester's law of inertia 7.7.6 to say that this matrix is congruent to M_n.

Some authors call a matrix A *degenerate*, when two of its eigenvalues are the same (as is the case in the example above). This is a relatively rare phenomenon, as

explained in [16], p.112-113. This use of the word degenerate conflicts with standard usage for quadratic forms: see Definition 7.7.5.

Exercise 7.7.14. Show that the matrix of the quadratic form

$$q(x_1,x_2,x_3) = x_1^2 + x_1x_2 + x_1x_3 + x_2^2 + x_2x_3 + x_3^2 \tag{7.11}$$

is

$$A = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix} \tag{7.12}$$

since $\mathbf{x}^t A\mathbf{x} = q(\mathbf{x})$. Show this matrix is positive-definite, so its signature is 3, by the same method as the previous example.

Corollary 7.7.15. *There are as many congruence classes of quadratic forms on an n-dimensional real vector space as there are different signatures.*

Problem 7.7.16. Count the number of signatures when n = 2, 3, 4.

We will develop more tests for positive definiteness, negative definiteness and the like in §13.5 , but first we must prove the most important theorem connected to real symmetric matrices: the Spectral Theorem. We do this in a later chapter.

7.8 Hermitian Forms

Bilinear forms are extremely useful over $\mathbb{R}$, as we have just seen, but over $\mathbb{C}$ it is better to use Hermitian forms. First some definitions. V is a complex vector space, and unless otherwise mentioned, the vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ are in V

Definition 7.8.1. A complex valued function $f(\mathbf{v})$ on V is *conjugate linear* if

1. $f(\mathbf{v}+\mathbf{w}) = f(\mathbf{v}) + f(\mathbf{w})$
2. $f(c\mathbf{v}) = \overline{c}f(\mathbf{v})$, where $\overline{c}$ is the complex conjugate of the complex number c.

This motivates the next definition.

Definition 7.8.2. A form $(\mathbf{v},\mathbf{w})$ on V is *sesquilinear* if

$$\begin{aligned} (\mathbf{u}+\mathbf{v},\mathbf{w}) &= (\mathbf{u},\mathbf{w}) + (\mathbf{v},\mathbf{w}); \\ (\mathbf{u},\mathbf{v}+\mathbf{w}) &= (\mathbf{u},\mathbf{v}) + (\mathbf{u},\mathbf{w}); \\ (c\mathbf{v},\mathbf{w}) &= c(\mathbf{v},\mathbf{w}); \\ (\mathbf{v},c\mathbf{w}) &= \overline{c}(\mathbf{v},\mathbf{w}). \end{aligned}$$

Thus the form is linear in its first variable and conjugate linear in the second variable. One could have done this the other way around, and indeed physicists do:

for them a sequilinear form is conjugate linear in the first variable, and linear in the second. Most mathematicians do it as described here. The word 'sesquilinear' means one and a half times linear, which is about right.

Finally we get to the definition we are really interested in:

Definition 7.8.3. A form $(\mathbf{v},\mathbf{w})$ is *Hermitian* if it is sesquilinear and is also conjugate symmetric:

$$(\mathbf{v},\mathbf{w}) = \overline{(\mathbf{w},\mathbf{v})}. \tag{7.13}$$

Exercise 7.8.4. Show that if a form $(\mathbf{v},\mathbf{w})$ is linear in its first variable and satisfies (7.13), then it is conjugate linear in the second variable.

Following Theorem 6.3.4, we may associate a $n \times n$ matrix A to a Hermitian form in n variables, and the analog of the matrix of a symmetric bilinear form being symmetric is that the matrix of a Hermitian form equal to its conjugate transpose: $A = A^*$. We say that such a matrix is Hermitian.

Exercise 7.8.5. Prove this last statement.

Next we decompose the Hermitian form which we now write $h(\mathbf{v},\mathbf{w})$, since other forms will be introduced, into its real and imaginary parts.

$$h(\mathbf{v},\mathbf{w}) = s(\mathbf{v},\mathbf{w}) + \mathrm{i}a(\mathbf{v},\mathbf{w}) \tag{7.14}$$

where $s(\mathbf{v},\mathbf{w})$ and $a(\mathbf{v},\mathbf{w})$ are real-valued form. Write $V_{\mathbb{R}}$ for the vector space V considered as an $\mathbb{R}$ vector space. See §5.8 for details. Recall that a form $a(\mathbf{v},\mathbf{w})$ is antisymmetric if $a(\mathbf{v},\mathbf{w}) = -a(\mathbf{w},\mathbf{v})$.

Theorem 7.8.6. *If $h(\mathbf{v},\mathbf{w})$ be a Hermitian form, decomposed as in* (7.14)*, then*

1. *$s(\mathbf{v},\mathbf{w})$ and $a(\mathbf{v},\mathbf{w})$ are bilinear forms on $V_{\mathbb{R}}$.*
2. *$s(\mathbf{v},\mathbf{w})$ is symmetric and $a(\mathbf{v},\mathbf{w})$ is antisymmetric.*
3. *$s(\mathbf{v},\mathbf{w}) = a(\mathrm{i}\mathbf{v},\mathbf{w}) = -a(\mathbf{v},\mathrm{i}\mathbf{w})$ and $a(\mathbf{v},\mathbf{w}) = -s(\mathrm{i}\mathbf{v},\mathbf{w}) = s(\mathbf{v},\mathrm{i}\mathbf{w})$.*
4. *$s(\mathrm{i}\mathbf{v},\mathrm{i}\mathbf{w}) = s(\mathbf{v},\mathbf{w})$ and $a(\mathrm{i}\mathbf{v},\mathrm{i}\mathbf{w}) = a(\mathbf{v},\mathbf{w})$.*

Proof. 1. is trivial: in (7.14) just replace $\mathbf{v}$ by $c\mathbf{v}$, where c is a real number, and then $\mathbf{w}$ by $c\mathbf{w}$. This plus the additively of h gives the result.

2. comes from interchanging $\mathbf{v}$ and $\mathbf{w}$ in (7.14). Because the Hermitian is conjugate symmetric, we get
$$h(\mathbf{w},\mathbf{v}) = s(\mathbf{w},\mathbf{v}) - \mathrm{i}a(\mathbf{w},\mathbf{v})$$
which is precisely what we need.

3. is obtained by replacing $\mathbf{v}$ by $\mathrm{i}\mathbf{v}$ in (7.14). Because h is $\mathbb{C}$-linear in the first variable, we get
$$\mathrm{i}h(\mathbf{v},\mathbf{w}) = h(\mathrm{i}\mathbf{v},\mathbf{w}) = s(\mathrm{i}\mathbf{v},\mathbf{w}) + \mathrm{i}a(\mathrm{i}\mathbf{v},\mathbf{w})$$
Equating this to (7.14) multiplied by i, we get
$$\mathrm{i}s(\mathbf{v},\mathbf{w}) - a(\mathbf{v},\mathbf{w}) = s(\mathrm{i}\mathbf{v},\mathbf{w}) + \mathrm{i}a(\mathrm{i}\mathbf{v},\mathbf{w}).$$

Equating the real and imaginary parts, we get $s(i\mathbf{v},\mathbf{w}) = -a(\mathbf{v},\mathbf{w})$ and $s(\mathbf{v},\mathbf{w}) = a(\mathrm{i}\mathbf{v},\mathbf{w})$.

Similarly, replace $\mathbf{w}$ by $\mathrm{i}\mathbf{w}$ in (7.14), and use the fact that h is conjugate linear in the second variable. So $h(\mathbf{v},\mathrm{i}\mathbf{w}) = -\mathrm{i}h(\mathbf{v},\mathbf{w}) = -\mathrm{i}s(\mathbf{v},\mathbf{w}) + a(\mathbf{v},\mathbf{w})$ and this is also $s(\mathbf{v},\mathrm{i}\mathbf{w}) + \mathrm{i}a(\mathbf{v},\mathrm{i}\mathbf{w})$, so:

$$-\mathrm{i}s(\mathbf{v},\mathbf{w}) + a(\mathbf{v},\mathbf{w}) = s(\mathbf{v},\mathrm{i}\mathbf{w}) + \mathrm{i}a(\mathbf{v},\mathrm{i}\mathbf{w}).$$

Equating real and imaginary parts, we get $s(\mathbf{v},\mathrm{i}\mathbf{w}) = a(\mathbf{v},\mathbf{w})$ and $s(\mathbf{v},\mathbf{w}) = -a(\mathbf{v},\mathrm{i}\mathbf{w})$, as required.

4. works in exactly the same way, but this time using $h(\mathrm{i}\mathbf{v},\mathrm{i}\mathbf{w}) = h(\mathbf{v},\mathbf{w})$. The details are left to you.

□

Unfortunately, if h is conjugate linear in the first variable, and linear in the second variable, the formulas in 3. change.

Just as we associated a quadratic form to a symmetric bilinear form, we can associate the form

$$q(\mathbf{v}) = h(\mathbf{v},\mathbf{v})$$

to a Hermitian form.

Let's establish the properties of $q(\mathbf{v})$.

Theorem 7.8.7. *Writing the Hermitian form $h(\mathbf{v},\mathbf{w})$ in terms of its real and imaginary parts $s(\mathbf{v},\mathbf{w})$ and $a(\mathbf{v},\mathbf{w})$ as in* (7.14)*, we get*

1.
$$q(\mathbf{v}) \text{ is real valued and } q(c\mathbf{v}) = \|c\|^2 q(\mathbf{v}).$$

2.
$$s(\mathbf{v},\mathbf{w}) = \frac{q(\mathbf{v}+\mathbf{w}) - q(\mathbf{v}) - q(\mathbf{w})}{2}.$$

3.
$$a(\mathbf{v},\mathbf{w}) = \frac{q(\mathbf{v}+\mathrm{i}\mathbf{w}) - q(\mathbf{v}) - q(\mathrm{i}\mathbf{w})}{2}.$$

Proof. Since the defining property of a Hermitian form says that $h(\mathbf{v},\mathbf{v}) = \overline{h(\mathbf{v},\mathbf{v})}$, all the values of $q(\mathbf{v})$ are real.

For any complex number c, we have

$$q(c\mathbf{v}) = b(c\mathbf{v},c\mathbf{v}) = c\bar{c}b(\mathbf{v},\mathbf{v}) = \|c\|^2 q(\mathbf{v}),$$

so the first item is established.

$$\begin{aligned} q(\mathbf{u}+\mathbf{v}) &= h(\mathbf{u}+\mathbf{v},\mathbf{u}+\mathbf{v}) = h(\mathbf{u},\mathbf{u}) + h(\mathbf{u},\mathbf{v}) + h(\mathbf{v},\mathbf{u}) + h(\mathbf{v},\mathbf{v}) \\ &= q(\mathbf{u}) + h(\mathbf{u},\mathbf{v}) + \overline{h(\mathbf{u},\mathbf{v})} + q(\mathbf{v}), \end{aligned} \tag{7.15}$$

so

$$h(\mathbf{u},\mathbf{v}) + \overline{h(\mathbf{u},\mathbf{v})} = q(\mathbf{u}+\mathbf{v}) - q(\mathbf{u}) - q(\mathbf{v}).$$

If we replace $h(\mathbf{u},\mathbf{v})$ in this expression by (7.14), we get

$$c(\mathbf{v},\mathbf{w}) + \mathrm{i}d(\mathbf{v},\mathbf{w}) + c(\mathbf{v},\mathbf{w}) - \mathrm{i}d(\mathbf{v},\mathbf{w}) = q(\mathbf{u}+\mathbf{v}) - q(\mathbf{u}) - q(\mathbf{v})$$

or

$$2c(\mathbf{v},\mathbf{w}) = q(\mathbf{u}+\mathbf{v}) - q(\mathbf{u}) - q(\mathbf{v}).$$

which gives the second item.

Finally replace $\mathbf{v}$ in (7.15) by $\mathrm{i}\mathbf{v}$ to get

$$\begin{aligned} q(\mathbf{u}+\mathrm{i}\mathbf{v}) &= q(\mathbf{u}) + h(\mathbf{u},\mathrm{i}\mathbf{v}) + \overline{h(\mathbf{u},\mathrm{i}\mathbf{v})} + q(\mathrm{i}\mathbf{v}) \\ &= q(\mathbf{u}) - \mathrm{i}h(\mathbf{u},\mathbf{v}) - \overline{\mathrm{i}h(\mathbf{u},\mathbf{v})} + q(\mathrm{i}\mathbf{v}) \end{aligned} \tag{7.16}$$

or

$$-\mathrm{i}(h(\mathbf{u},\mathbf{v}) + \overline{\mathrm{i}h(\mathbf{u},\mathbf{v})} = q(\mathbf{u}+\mathrm{i}\mathbf{v}) - q(\mathbf{u}) - q(\mathrm{i}\mathbf{v})$$

As before substitute out h using (7.14) to get

$$2d(\mathbf{v},\mathbf{w}) = q(\mathbf{u}+\mathrm{i}\mathbf{v}) - q(\mathbf{u}) - q(\mathrm{i}\mathbf{v})$$

proving the last item. □

Definition 7.8.8. A form $q(\mathbf{v})$ is a *Hermitian quadratic form* if it takes real values and satisfies $q(c\mathbf{v}) = \|c\|^2 q(\mathbf{v})$.

Just as in the symmetric bilinear case (see Lemma 7.2.1) we can recover the Hermitian form from its Hermitian quadratic from 'polarization identities'.

Theorem 7.8.9. *Let $q(\mathbf{v})$ be a Hermitian quadratic form. Then there is a unique Hermitian form $h(\mathbf{v},\mathbf{w})$ for which $q(\mathbf{v})$ is the Hermitian quadratic form.*

Proof. Indeed, it is given as $h(\mathbf{v},\mathbf{w}) = s(\mathbf{v},\mathbf{w}) + is(\mathbf{v},\mathbf{w})$, which can both be expressed in terms of $q(\mathbf{v})$. □

7.9 Diagonalization of Hermitian Forms

We can derive properties of Hermitian forms exactly as we did for symmetric bilinear forms over $\mathbf{R}$: we can reduce them to diagonal form, prove Syslvester's law of inertia and then define the various categories of Hermitian forms: positive definite being the most important.

Examples here.

Chapter 8
Inner Product Spaces

Abstract The informed reader will be surprised that the notion of scalar product of vectors is only introduced now. There are several reasons for the delay. The first is that introducing a scalar product is adding an additional piece of data to the structure of a vector space, so it is useful to understand first what that can be done without it. The second reason is that a scalar product, which can be defined on a complex vector space, is not the most useful concept there. On a complex vector space the most useful concept is that of a Hermitian product. In this chapter we only consider positive definite scalar products in the real case, and positive definite Hermitian products in the complex case. We refer to both as inner products. In the first four sections we develop the theory of the inner product, and its applications to real vector spaces; in §8.6 we define the Hermitian product in much the same way, and derive the parallel applications in complex vector spaces. Then we go back to results that can be treated simultaneously for real and complex vector spaces. We improve the statements concerning a linear map and its transpose when the vectors spaces have an inner product or a Hermitian product. Finally we show how to put an inner product on the space of all matrices, and the space of all symmetric matrices.

8.1 Scalar Products

A scalar product is just a symmetric bilinear form on a vector space. It is given a new name because we think of the scalar product as being permanently associated to the vector space. The scalar product we consider here will be positive definite, which is the standard case considered in calculus and geometry. There are interesting examples of other kinds of scalar products, especially in physics. The following definition is most useful when the scalars are the real numbers. Although the definition makes sense for the complex numbers, it is less useful than the Hermitian product, as we will see.

We also give a new definition because we use a different notation for the scalar product.

Definition 8.1.1. Let V be a vector space over F. A scalar product on V associates a scalar to any pair $\mathbf{v}$, $\mathbf{w}$ of elements of V. It is written $\langle \mathbf{v}, \mathbf{w} \rangle$. It satisfies the following three properties for all $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ in V:

SP 1 Commutativity: $\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle$;
SP 2 Additivity: $\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{x}, \mathbf{w} \rangle$;
SP 3 For all scalars a, $\langle a\mathbf{v}, \mathbf{w} \rangle = a\langle \mathbf{v}, \mathbf{w} \rangle$.

Exercise 8.1.2. Prove that the definition implies

$$\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle \quad \text{and } \langle \mathbf{v}, a\mathbf{w} \rangle = a\langle \mathbf{v}, \mathbf{w} \rangle, \text{ for all scalars } a,$$

and therefore $\langle \mathbf{0}, \mathbf{w} \rangle = 0$.

Exercise 8.1.3. Convince yourself that a scalar product is nothing more than a symmetric bilinear form.

Example 8.1.4. The prototypical example is F^n with scalar product

$$\langle \mathbf{x}, \mathbf{y} \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n. \tag{8.1}$$

This is called the standard scalar product on F^n.

Exercise 8.1.5. Prove that this is a scalar product according to our definition.

A scalar product is non-degenerate if it is non-degenerate as a symmetric bilinear form: See Definition 7.1.4. Thus the only $\mathbf{v} \in V$ for which $(\mathbf{v}, \mathbf{w}) = 0$ for all $\mathbf{w} \in V$ is the origin.

Exercise 8.1.6. Show that the standard scalar product on $\mathbb{R}^n$ is non-degenerate.

In the next definition, for the first time we require that the scalars be $\mathbb{R}$.

Definition 8.1.7. A scalar product is *positive definite* if for all non-zero $\mathbf{v} \in V$

$$\langle \mathbf{v}, \mathbf{v} \rangle > 0.$$

Exercise 8.1.8. Show that a positive definite scalar product is non-degenerate.

Definition 8.1.9. A real vector space with a positive definite scalar product is called a *Euclidean space*.

Let V be a n-dimensional real vector space, and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ any basis of V. Then we can make V into a Euclidean space by taking as the positive definite scalar product the one defined by

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \delta_{ij}.$$

Thus a vector space can be made into a Euclidean space in many different ways, depending on the choice of basis.

To conclude this section we prove some results about orthogonality in Euclidean spaces. They generalize without difficulty to Hermitian spaces, as we will see in §8.6. Two vectors $\mathbf{v}$ and $\mathbf{w}$ are orthogonal if $\langle \mathbf{v}, \mathbf{w} \rangle = 0$. We write $\mathbf{v} \perp \mathbf{w}$ if this is the case. So this is the same definition as for symmetric bilinear forms. For a subspace U of V we define $U^\perp$, the *orthogonal complement* of U, as the subspace of vectors that are perpendicular to all vectors in U.

Theorem 8.1.10. *V is a Euclidean space. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ be non-zero vectors that are mutually orthogonal, meaning that*

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0, \text{ whenever } i \neq j.$$

Then the $\mathbf{v}_i$ are linearly independent. Thus if W is the span of the $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$, they form a basis of W.

Proof. Assume there is an equation of linear dependence:

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_r\mathbf{v}_r = \mathbf{0}.$$

Take the scalar product of this expression with $\mathbf{v}_i$, $1 \leq i \leq r$, to get

$$a_i \langle \mathbf{v}_i, \mathbf{v}_i \rangle = 0.$$

Since $\langle \mathbf{v}_i, \mathbf{v}_i \rangle \neq 0$ since the scalar product is positive definite, this forces $a_i = 0$ for all i. Thus there is no equation of linear dependence. □

Definition 8.1.11. The orthogonal projection of $\mathbf{v}$ to the line through the origin with basis the non-zero vector $\mathbf{w}$ is the vector

$$\frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle} \mathbf{w}.$$

Suppose we try to find a scalar c such that

$$\langle \mathbf{v} - c\mathbf{w}, \mathbf{w} \rangle = 0 \tag{8.2}$$

The linearity of the scalar product implies $\langle \mathbf{v}, \mathbf{w} \rangle = c\langle \mathbf{w}, \mathbf{w} \rangle$, so

$$c = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle}. \tag{8.3}$$

The denominator is non-zero since we assume the vector $\mathbf{w} \neq 0$ and the inner product is positive definite. If you have read Chapter 7, you will notice that (8.2) is a special case of the orthogonal projection (7.9) defined for a more general bilinear form.

Definition 8.1.12. The scalar c in (8.3) is called the *component* of $\mathbf{v}$ along $\mathbf{w}$. In the important special case where $\mathbf{w}$ is a unit vector, $c = \langle \mathbf{v}, \mathbf{w} \rangle$.

This is illustrated by the following graph in the plane generated by $\mathbf{v}$ and $c\mathbf{w}$.

Example 8.1.13. Let $\mathbf{e}_i$, $1 \leq i \leq n$ be the standard basis of $\mathbb{R}^n$. So $\mathbf{e}_1 = (1,0,\ldots,0)$, etc. Let

$$\mathbf{v} = (v_1, v_2, \ldots, v_n)$$

be any vector in $\mathbb{R}^n$. Then v_i is the component of $\mathbf{v}$ along $\mathbf{e}_i$, and $v_i\mathbf{e}_i$ is the projection of $\mathbf{v}$ along $\mathbf{e}_i$.

Theorem 8.1.14. *Consider a Euclidean space V of dimension n, and a non-zero vector $\mathbf{w} \in V$, spanning a subspace W of dimension 1. Then the orthogonal complement $W^\perp$ of W has dimension $n-1$, and $V = W \oplus W^\perp$.*

Proof. The linear map $\mathbf{v} \mapsto \langle \mathbf{v}, \mathbf{w} \rangle$ has range of dimension 1, since $< \mathbf{w}, \mathbf{w} > 0$, so its nullspace has dimension $n-1$. It is obviously $W^\perp$. Since $\mathbf{w}$ is not in $W^\perp$, we get the last statement. □

From this follows:

Corollary 8.1.15. *Any Euclidean space has an orthogonal basis.*

Proof. Call the Euclidean space V. We prove the result by induction on the dimension n of V. If the dimension is 1, the result is trivial. Assume the result is true for dimension $n-1$. Consider a Euclidean space V of dimension n. Pick a non-zero vector $\mathbf{v}_1 \in V$. By Theorem 8.1.14, the orthogonal complement V_1 of $\mathbf{v}_1$ has dimension $n-1$. Therefore by induction V_1 has an orthogonal basis $\{\mathbf{v}_2, \ldots, \mathbf{v}_n\}$. Furthermore $\mathbf{v}_1$ and V_1 span V. Thus $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is an orthogonal basis of V, so we are done. □

A constructive proof of this result is given by the Gram-Schmidt process that we will study in §8.3. Finally we get, by an easy extension of the proof of the corollary:

Theorem 8.1.16. *Let W be a subspace of the Euclidean space V. Then $V = W \oplus W^\perp$. Thus $W^\perp$ has dimension* $\dim V - \dim W$.

Proof. This generalizes Theorem 8.1.14 to the case where W has dimension greater than one. Only the first statement needs proof, since the second statement follows from Definition 3.6.5. We prove this by induction on the dimension m of W. The case $m = 1$ is Theorem 8.1.14. Assume the result is true for $m-1$. Pick a non-zero $\mathbf{v} \in W$. Then its orthogonal W_0 in W has dimension $m-1$ in W. If $n = \dim V$, then the orthogonal V_0 of $\mathbf{v}$ in V_0 of V has dimension $n-1$. Thus the orthogonal $W_0^\perp$ of W_0 in V_0 has dimension $n-1-(m-1) = n-m$. By Corollary 8.1.15 pick an orthogonal basis $\mathbf{w}_2, \ldots, \mathbf{w}_m$ of W_0 so that $\mathbf{v}, \mathbf{w}_2, \ldots, \mathbf{w}_m$ is a basis of W. It is then clear that $W_0^\perp$ is the orthogonal of W in V. □

8.2 The Geometry of Euclidean Spaces

Throughout this section V is a Euclidean space: see Definition 8.1.9. On first reading, you should think of $\mathbb{R}^n$ with the standard scalar product, even though all the results are valid even for an infinite dimensional V.

If W is a subspace of V, it inherits a positive definite scalar product from V.

The *norm*, also called the *length*, of the vector $\mathbf{v}$ is:

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}. \tag{8.4}$$

When V is $\mathbb{R}^n$, with the standard scalar product, then

$$\|\mathbf{v}\| = \sqrt{v_1^2 + \cdots + v_n^2}.$$

Exercise 8.2.1. Show

$$\|c\mathbf{v}\| = |c|\|\mathbf{v}\| \text{ for all } c \in \mathbb{R}.$$

The vector $\mathbf{v}$ is a unit vector if its length is 1. Any non-zero vector $\mathbf{v}$ can be *normalized* to length 1 by replacing it by

$$\frac{\mathbf{v}}{\|\mathbf{v}\|}.$$

The *distance* $d(\mathbf{v}, \mathbf{w})$ between the two vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^n$ is the norm of the difference vector:

$$d(\mathbf{v}, \mathbf{w}) = \|\mathbf{v} - \mathbf{w}\|. \tag{8.5}$$

Exercise 8.2.2. Show that $d(\mathbf{v}, \mathbf{w}) = d(\mathbf{w}, \mathbf{v})$.

Theorem 8.2.3 (The Pythagorean Theorem). *The vectors* $\mathbf{v}$ *and* $\mathbf{w}$ *are perpendicular if and only if*

$$\|\mathbf{v} + \mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2.$$

Proof. This is easy. We first use the definition of the scalar product

$$\begin{aligned} \|\mathbf{v} + \mathbf{w}\|^2 &= \langle \mathbf{v} + \mathbf{w}, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{v} \rangle + 2\langle \mathbf{v}, \mathbf{w} \rangle + \langle \mathbf{w}, \mathbf{w} \rangle \\ &= \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 \end{aligned}$$

if and only if $\langle \mathbf{v}, \mathbf{w} \rangle = 0$. □

Example 8.2.4. Apply the theorem to the vectors $(1, -1)$ and $(2, 2)$ in $\mathbb{R}^2$. Make a graph.

Theorem 8.2.5 (The Parallelogram Law). *For all* $\mathbf{v}$ *and* $\mathbf{w}$ *in* V

$$\|\mathbf{v} + \mathbf{w}\|^2 + \|\mathbf{v} - \mathbf{w}\|^2 = 2\|\mathbf{v}\|^2 + 2\|\mathbf{w}\|^2.$$

Proof. Just expand the left hand side as in the proof of the Pythagorean Theorem. □

Example 8.2.6. Apply to the vectors $\mathbf{v} = (1, 1)$ and $\mathbf{w} = (1, 2)$. Make a graph.

Theorem 8.2.7 (The Cauchy-Schwarz Inequality). *For any two vectors* $\mathbf{v}$ *and* $\mathbf{w}$ *in* V*,*

$$|\langle \mathbf{v}, \mathbf{w} \rangle| \leq \|\mathbf{v}\| \|\mathbf{w}\| \tag{8.6}$$

with equality only if one vector is $\mathbf{0}$ *or if the vectors are proportional—namely,* $\mathbf{w} = c\mathbf{v}$ *for a scalar* c*.*

Proof. First, the result is trivial if either $\mathbf{v}$ or $\mathbf{w}$ is the zero vector, since then both sides are 0.

Next assume $\mathbf{w}$ has length 1. Then the component c of $\mathbf{v}$ along $\mathbf{w}$ is $\langle \mathbf{v}, \mathbf{w} \rangle$, as per Definition 8.1.12. The Cauchy-Schwarz inequality becomes

$$|c| \leq \|\mathbf{v}\|. \tag{8.7}$$

Now $\mathbf{v} - c\mathbf{w}$ is perpendicular to $\mathbf{w}$, and therefore to $c\mathbf{w}$. Since

$$\mathbf{v} = \mathbf{v} - c\mathbf{w} + c\mathbf{w},$$

the Pythagorean Theorem says

$$\|\mathbf{v}\|^2 = \|\mathbf{v} - c\mathbf{w}\|^2 + \|c\mathbf{w}\|^2 = \|\mathbf{v} - c\mathbf{w}\|^2 + |c|^2.$$

Since the first term on the right hand side is non-negative, this implies $|c|^2 \leq \|\mathbf{v}\|^2$, which is equivalent to (8.7), so we are done. We get equality when $\|\mathbf{v} - c\mathbf{w}\| = \mathbf{0}$, namely when $\mathbf{v}$ is proportional to $\mathbf{w}$.

Finally, given that the result holds for a $\mathbf{w}$ of length 1, it holds for any $c\mathbf{w}$, $c \in \mathbb{R}$. Indeed, just substitute $c\mathbf{w}$ for $\mathbf{w}$ in (8.6), and note that the positive factor $|c|$ appears on both side. □

Theorem 8.2.8 (The Triangle Inequality). *For all* $\mathbf{u}$ *and* $\mathbf{v}$ *in* V*,*

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|.$$

Proof. Square the left hand side:

$$\|\mathbf{u} + \mathbf{v}\|^2 = \langle \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v} \rangle = \|\mathbf{u}\|^2 + 2\langle \mathbf{u}, \mathbf{v} \rangle + \|\mathbf{v}\|^2.$$

Now we use the Cauchy-Schwarz inequality to replace $2\langle \mathbf{u}, \mathbf{v} \rangle$ by the larger term $2\|\mathbf{u}\|\|\mathbf{v}\|$:

$$\|\mathbf{u}\|^2 + 2\langle \mathbf{u}, \mathbf{v} \rangle + \|\mathbf{v}\|^2 \leq \|\mathbf{u}\|^2 + 2\|\mathbf{u}\|\|\mathbf{v}\| + \|\mathbf{v}\|^2.$$

We recognize the right-hand side as the square of $\|\mathbf{u}\| + \|\mathbf{v}\|$, so we get

$$\|\mathbf{u} + \mathbf{v}\|^2 \leq (\|\mathbf{u}\| + \|\mathbf{v}\|)^2.$$

Taking the square root of both sides, we are done. □

Exercise 8.2.9. Show that the triangle inequality implies that the length of a side of a triangle in a real vector space is less than or equal to the sum of the lengths of the other two sides. This explains the name of the result.

8.3 Gram-Schmidt Orthogonalization

In this section, V is again a Euclidean space: see Definition 8.1.7. As we saw in Theorem 8.1.10, it is convenient to have a basis for V where the basis elements are mutually perpendicular. We say the basis is orthogonal. In this section we give an algorithm for constructing an orthogonal basis $\{\mathbf{w}_1,\dots,\mathbf{w}_n\}$ of V, starting from any basis $\{\mathbf{v}_1,\dots,\mathbf{v}_n\}$ of V.

First we explain the idea behind the algorithm. Define V_k to be the subspace of V spanned by the first k vectors in the basis, namely $\{\mathbf{v}_1,\dots,\mathbf{v}_k\}$. Then obviously V_k has dimension k, and for any $j<k$, $V_j \subsetneqq V_k$. So here is how we build the orthogonal basis. Start with $\mathbf{w}_1=\mathbf{v}_1$ as our first basis element. Then in V_2 take a non-zero vector $\mathbf{w}_2$ orthogonal to V_1. Because the dimension of V_2 is only one more than that of V_1, the orthogonal complement of V_1 in V_2 has dimension 1, so $\mathbf{w}_1$ and $\mathbf{w}_2$ form a basis of V_2 Next in V_3 take a non-zero vector $\mathbf{w}_3$ orthogonal to V_2, and continue in this way.

The general case is: in V_k take a non-zero vector $\mathbf{w}_k$ orthogonal to V_{k-1}. Then $\mathbf{w}_1$, $\dots$, $\mathbf{w}_k$ form a basis for V_k They form a basis because they are non-zero and mutually orthogonal by construction. So we have found the desired orthogonal basis of V.

Our goal is to write down this method as a computational algorithm. The main step is to compute a vector in V_k that is orthogonal to V_{k-1}. We already know how to do this, since we have already computed an orthogonal basis for V_{k-1}: it is sufficient to modify $\mathbf{v}_k$ so that it is orthogonal to $\mathbf{w}_1$, $\dots$, $\mathbf{w}_{k-1}$. For this, computing the component of $\mathbf{v}_k$ along each one of these $\mathbf{w}_i$ is the appropriate tool. See Definition 8.1.12 and (8.3). So we write

$$c_{jk}=\frac{\langle \mathbf{v}_k,\mathbf{w}_j\rangle}{\langle \mathbf{w}_j,\mathbf{w}_j\rangle}$$

The computation in V_k is given by

$$\mathbf{w}_k=\mathbf{v}_k-\sum_{j=1}^{k-1}\frac{\langle \mathbf{v}_k,\mathbf{w}_j\rangle}{\langle \mathbf{w}_j,\mathbf{w}_j\rangle}\mathbf{w}_j=\mathbf{v}_k-\sum_{j=1}^{k-1}c_{jk}\mathbf{w}_j.$$

Just check that when you dot this by any $\mathbf{w}_j$, $j<k$, you get 0. This completes the construction that we now state.

Theorem 8.3.1 (Gram-Schmidt Orthogonalization Process). *If V is a Euclidean space of dimension n with basis $\{\mathbf{v}_1,\dots,\mathbf{v}_n\}$, then V has a basis of mutually orthogonal vectors $\mathbf{w}_1$, $\dots$, $\mathbf{w}_n$ constructed as follows:*

$$\begin{aligned}
&\mathbf{w}_1=\mathbf{v}_1;\\
&\mathbf{w}_2=\mathbf{v}_2-c_{12}\mathbf{w}_1;\\
&\mathbf{w}_3=\mathbf{v}_3-c_{23}\mathbf{w}_2-c_{13}\mathbf{w}_1;\\
&\quad\vdots\\
&\mathbf{w}_n=\mathbf{v}_n-c_{n-1,n}\mathbf{w}_{n-1}-c_{n-2,n}\mathbf{w}_{n-2}-\cdots-c_{1n}\mathbf{w}_1.
\end{aligned}$$

Furthermore any set of non-zero mutually orthogonal vectors $\mathbf{w}_1, \ldots, \mathbf{w}_k$ *can be extended to a basis of mutually orthogonal vectors by the same process.*

Proof. As above let V_k be the span of the vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. Since $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is a basis of V_k, $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ forms another basis, so all the basis vectors $\mathbf{w}_i$ are non-zero. As noted above the $\mathbf{w}_i$ are mutually orthogonal.

The main point is to describe the computational algorithm. We first compute $\mathbf{w}_1$ using the first equation, then $\mathbf{w}_2$ using the second equation and the computed value for $\mathbf{w}_1$, and so on. This can be done, since the equation defining $\mathbf{w}_k$ only involves the known $\mathbf{v}_i$, and the $\mathbf{w}_j$, with $j < k$, which have already been computed.

This concludes the proof of the first part of the Gram-Schmidt Theorem. The last part is easy. Just complete the set of mutually perpendicular non-zero vectors $\mathbf{w}_1$, $\ldots \mathbf{w}_k$ to a basis of V using Proposition 3.4.14: keep adding linearly independent elements $\mathbf{v}_{k+1}, \ldots, \mathbf{v}_n$ to the $\mathbf{w}_i$ until you get a basis. Then apply the Gram-Schmidt process to this basis. □

Corollary 8.3.2. *Write the equations in Gram-Schmidt as*

$$\begin{aligned}
\mathbf{v}_1 &= \mathbf{w}_1; \\
\mathbf{v}_2 &= \mathbf{w}_2 + c_{12}\mathbf{w}_1; \\
\mathbf{v}_3 &= \mathbf{w}_3 + c_{23}\mathbf{w}_2 + c_{13}\mathbf{w}_1; \\
&\vdots \\
\mathbf{v}_n &= \mathbf{w}_n + c_{n-1,n}\mathbf{w}_{n-1} - c_{n-2,n}\mathbf{w}_{n-2} + \cdots + c_{1n}\mathbf{w}_1.
\end{aligned}$$

Let X be the matrix whose columns are the coefficients of $\mathbf{v}_k$ in the standard basis. Let W be the matrix whose columns are the coefficients of $\mathbf{w}_k$ in the standard basis. Finally let C be the upper-triangular matrix with 1 down the diagonal, c_{ij} above the diagonal (so $i < j$), and of course 0 below the diagonal. Then our equations express the matrix product $X = WC$ by the ever useful Proposition 2.2.7.

Example 8.3.3. Assume the three vectors $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ are the columns, in order, of the matrix

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

It is easy to see that they form a basis for $\mathbb{R}^3$. We start the Gram-Schmidt process: $\mathbf{w}_1 = \mathbf{v}_1$. Note that $\langle \mathbf{w}_1, \mathbf{w}_1 \rangle = 3$. Also $\langle \mathbf{w}_1, \mathbf{v}_2 \rangle = 4$. So $\mathbf{w}_2 = \mathbf{v}_2 - 4/3\mathbf{w}_1$, which is the column vector $(-1/3, 2/3, -1/3)$. Check it is orthogonal to $\mathbf{w}_1$. Also $\langle \mathbf{w}_2, \mathbf{w}_2 \rangle = \frac{2}{3}$. Finally $\langle \mathbf{v}_3, \mathbf{w}_1 \rangle = 5$ and $\langle \mathbf{v}_3, \mathbf{w}_2 \rangle = -\frac{2}{3}$. So

$$\mathbf{w}_3 = \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} - \frac{5}{3}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1/3 \\ -2/3 \\ -1/3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

which is orthogonal to $\mathbf{w}_1$ and $\mathbf{w}_2$. Later we will need $\langle \mathbf{w}_3, \mathbf{w}_3 \rangle = 2$.

Remark 8.3.4. Once one has an orthogonal basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ of a vector space, we get an orthonormal basis by dividing by the length of each basis vector:

$$\{\frac{1}{\|\mathbf{w}_1\|}\mathbf{w}_1, \ldots, \frac{1}{\|\mathbf{w}_n\|}\mathbf{w}_n\}.$$

It is often useful to do this, when dealing with the *QR* factorization, for example.

Remark 8.3.5. As we showed in Corollary 8.1.15, an easier proof for the existence of an orthogonal basis is available, if you only want to show that an orthogonal basis of Euclidean space exists. All we used is the elementary Theorem 5.7.1 on solutions of homogeneous linear equations. However this does not give us an algorithm for constructing a solution, which is the true importance of the Gram-Schmidt process.

Our next topic is the *QR* factorization, which is a way of expressing the Gram-Schmidt process as a matrix factorization. For that we will need orthogonal matrices, that we now define. We will study them in more detail later.

Definition 8.3.6. A matrix Q is orthogonal it is is invertible and if its inverse is its transpose: $Q^t = Q^{-1}$. So $Q^tQ = I = QQ^t$.

In particular the columns $\mathbf{q}_i$ of Q are mutually perpendicular, and each column has length 1: $\|\mathbf{q}_i\| = 1$. Conversely, if you take an orthogonal basis for V, and normalize all the basis vector so they have length 1, then the matrix whose columns are the basis elements, in any order, is an orthogonal matrix.

Example 8.3.7 (Rotation Matrix). The rotation matrix

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

is orthogonal, as you can easily check.

Example 8.3.8. A permutation matrix is orthogonal. We will define permutation matrices and prove this result in Theorem 11.2.5.

From the Gram-Schmidt orthogonalization process, we get an interesting factorization theorem for invertible matrices..

Theorem 8.3.9 (*QR* factorization). *Any real invertible matrix A of size n can be written as the product of an orthogonal matrix Q and an upper triangular matrix R:*

$$A = QR$$

Proof. This will drop straight out of Gram-Schmidt, especially Corollary 8.3.2. To keep with the notation there, let the columns of the matrix A be written $\mathbf{v}_1, \ldots, \mathbf{v}_n$, and let $\mathbf{w}_i$ be the basis found using Gram-Schmidt. Let

$$\mathbf{q}_1 = \frac{\mathbf{w}_1}{\|\mathbf{w}_1\|}.$$

Then the second Gram-Schmidt equation can be written

$$\mathbf{w}_2 = \mathbf{v}_2 - \langle \mathbf{v}_2, \mathbf{q}_1 \rangle \mathbf{q}_1.$$

Set

$$\mathbf{q}_2 = \frac{\mathbf{w}_2}{\|\mathbf{w}_2\|}.$$

The third Gram-Schmidt equation becomes

$$\mathbf{w}_3 = \mathbf{v}_3 - \langle \mathbf{v}_3, \mathbf{q}_2 \rangle \mathbf{q}_2 - \langle \mathbf{v}_3, \mathbf{q}_1 \rangle \mathbf{q}_1.$$

Then set

$$\mathbf{q}_3 = \frac{\mathbf{w}_3}{\|\mathbf{w}_3\|}.$$

and continue in this way. Recalling the ever useful (2.7), we see that if we write Q for the matrix whose columns are $\mathbf{q}_1, \ldots, \mathbf{q}_n$, and $R = (r_{jk})$, where

$$r_{jk} = \begin{cases} \langle \mathbf{v}_k, \mathbf{q}_j \rangle, & \text{if } j < k; \\ \|\mathbf{w}_j\|, & \text{if } j = k; \\ 0, & \text{if } j > k. \end{cases}$$

we have $A = QR$ as claimed. The proof only involved untangling a matrix product. □

Notice that when $j < k$, $r_{jk} = \frac{c_{jk}}{\|\mathbf{w}_j\|}$ with c_{jk} as defined above.

Example 8.3.10. We now find the QR factorization for Example 8.3.3 First we find the orthonormal basis, by dividing each $\mathbf{w}_i$ by its length. We get the matrix

$$Q = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{2}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \end{pmatrix}$$

Using the formulas above, we get

$$R = \begin{pmatrix} \sqrt{3} & \frac{4}{\sqrt{3}} & \frac{5}{\sqrt{3}} \\ 0 & \sqrt{2/3} & -\sqrt{2/3} \\ 0 & 0 & \sqrt{2} \end{pmatrix}$$

You should check that $A = QR$ to verify the computation.

In the same way we can prove

Theorem 8.3.11. *Let A be a $m \times n$ matrix with linearly independent columns. Then A can be written as the product of a $m \times n$ matrix Q with mutual perpendicular columns of length 1 and an upper triangular $n \times n$ matrix R: $A = QR$.*

Proof. The hypotheses imply $m \geq n$ and the QR theorem is the case $m = n$. This theorem corresponds to the construction of a partial orthonormal basis with n elements in a Euclidean space V of dimension m. The columns of A need to be linearly independent because they correspond to the partial basis $\mathbf{v}_1, \ldots \mathbf{v}_n$ of V. □

Example 8.3.12. Suppose

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

The columns $\mathbf{v}_1$ and $\mathbf{v}_2$ are obviously linearly independent. Set $\mathbf{w}_1 = \mathbf{v}_1$;

$$\mathbf{w}_2 = \mathbf{v}_2 - \frac{\langle \mathbf{v}_2, \mathbf{w}_1 \rangle}{\langle \mathbf{w}_1, \mathbf{w}_1 \rangle} \mathbf{w}_1 = \mathbf{v}_2 - \mathbf{v}_1.$$

Then

$$Q = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ -1 & -1 \\ -1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad R = \begin{pmatrix} 2 & 2 \\ 0 & 2 \end{pmatrix}$$

Check $A = QR$, as expected.

Example 8.3.13. The matrix Q from Example 8.3.12 is connected to the 4×4 matrix

$$H_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

known as a Hadamard matrix, namely a matrix whose entries are either 1 or -1 and whose columns are mutually orthogonal. If you divide this matrix by 2 you get an orthogonal matrix. By choosing the columns in the order we did, we get additional properties. Let

$$H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Then in block form

$$H_4 = \begin{pmatrix} H_2 & H_2 \\ H_2 & -H_2 \end{pmatrix}$$

and H_4 is a symmetric matrix, since H_2 is.

The QR factorization of an arbitrary invertible real matrix as a product of an orthogonal matrix and an upper-triangular matrix is the basis is used in important algorithms for finding eigenvalues of matrices, as we will see later.

8.4 Orthogonal Projection in Euclidean Spaces

This section defines orthogonal projections in Euclidean spaces, and establishes the most important properties. We have already considered the case of projection to a line in Definition 8.1.11. We repeat material from §7.3 with improved and simpler results due to the stronger hypothesis: the scalar product is positive definite.

Start with a n-dimensional Euclidean space V. Take a subspace $U \subset V$ of dimension $m > 0$. By Theorem 8.1.16 its orthogonal complement $U^\perp$ has dimension $r = n - m$ and $V = U \oplus W$.

Then by Definition 4.1.7 we have a linear map P called projection of V to U along $U^\perp$. Because the nullspace of P is $U^\perp$, we call P an orthogonal projection as in §7.3. In our new notation,

Definition 8.4.1. A linear transformation $L\colon V \to V$ is a orthogonal projection of V to $U = L(V)$ for the inner product if

$$\langle \mathbf{v} - L(\mathbf{v}), \mathbf{u} \rangle = 0, \forall \mathbf{u} \in U. \tag{8.8}$$

This says that the vector from $\mathbf{v}$ to its projection $L\mathbf{v}$ is perpendicular to the entire range of L. Notice that (8.8) generalizes Definition 8.1.11, the case of projection to a line. By linearity it is enough to check (8.8) for a basis of U.

Let's work out what a projection P looks like in coordinates. Recall Definition 8.1.12 of the component of a vector along a non-zero vector.

Corollary 8.4.2. *Let $\mathbf{u}_1, \ldots, \mathbf{u}_m$ be an orthogonal basis of the subspace U. Consider the orthogonal projection P of V to U. Then for any $\mathbf{v} \in V$,*

$$P(\mathbf{v}) = c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + \cdots + c_k\mathbf{u}_m$$

where c_i is the component of $\mathbf{v}$ along $\mathbf{u}_i$.

Proof. The vector

$$\mathbf{v} - c_1\mathbf{u}_1 - c_2\mathbf{u}_2 - \cdots - c_k\mathbf{u}_m$$

is in $U^\perp$, as an easy computation shows. □

In addition to the orthogonal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ of U, choose an orthogonal basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_{n-m}\}$ of $U^\perp$. The two together form an orthogonal basis of V.

Write any $\mathbf{v} \in V$ in terms of this orthogonal basis:

$$\mathbf{v} = c_1\mathbf{u}_1 + \cdots + c_m\mathbf{u}_m + d_1\mathbf{w}_1 + \cdots + d_{n-m}\mathbf{w}_{n-m} \tag{8.9}$$

Then c_i is the component of $\mathbf{v}$ along $\mathbf{u}_i$, and d_j is the component of $\mathbf{v}$ along $\mathbf{w}_j$.

As noted, the orthogonal projection P of V to U maps $\mathbf{v}$ to $c_1\mathbf{u}_1 + \cdots + c_m\mathbf{u}_m$, while the orthogonal projection Q of V to $U^\perp$ maps $\mathbf{v}$ to $d_1\mathbf{w}_1 + \cdots + d_{n-m}\mathbf{w}_{n-m}$.

By Definition 4.1.7, P has nullspace $U^\perp$ and range U. Also $P^2 = P$. Similarly, Q has nullspace U and range $U^\perp$. Also $Q^2 = Q$. Note that $P + Q$ is the identity linear transformation.

Finally divide each basis element by its length. Then the matrix of P in the orthonormal basis obtained can be written in block form as

$$A_m = \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix} \tag{8.10}$$

where I_m is the $m \times m$ identity matrix, and the other matrices are all zero matrices. In particular it is a symmetric matrix.

Conversely we can establish:

Theorem 8.4.3. *Any square matrix P that*

- *is symmetric ($P^t = P$),*
- *and satisfies $P^2 = P$;*

is the matrix of the orthogonal projection to the range U of P with respect to the standard inner product.

Proof. We establish Definition 8.4.1 just using the two hypotheses. Any $\mathbf{u} \in U$ can be written $\mathbf{u} = P(\mathbf{v}')$ for some $\mathbf{v}' \in V$. For all $\mathbf{v}$ in V and $\mathbf{u}$ in U , we need to establish $\langle \mathbf{v} - P(\mathbf{v}), \mathbf{u} \rangle = 0$. Substituting out $\mathbf{u}$, we get

$$\langle \mathbf{v} - P\mathbf{v}, P\mathbf{v}' \rangle = (\mathbf{v} - P\mathbf{v})^t P\mathbf{v}' = \mathbf{v}^t P\mathbf{v}' - \mathbf{v}^t P^t P\mathbf{v}' = \mathbf{v}^t P\mathbf{v}' - \mathbf{v}^t P^2 \mathbf{v}' = \mathbf{v}^t P\mathbf{v}' - \mathbf{v}^t P\mathbf{v}' = 0.$$

In the next-to-the-last step we replaced $P^t P$ by P^2 because P is symmetric, and then in the last step we used $P^2 = P$. □

Now we establish a minimization result, which states the intuitively clear fact that dropping the perpendicular from a point $\mathbf{v}$ to a subspace W gives the point $\mathbf{p}$ that is at the minimum distance of $\mathbf{v}$ to W. 'Dropping the perpendicular' means taking the orthogonal projection to the subspace.

Theorem 8.4.4. *Let V be a Euclidean space, and U a subspace of smaller dimension. Let $\mathbf{v} \in V$, and let $\mathbf{p} = P(\mathbf{v})$ be its orthogonal projection to U. Let $\mathbf{u}$ be any point in U different from $\mathbf{p}$. Then $\|\mathbf{v} - \mathbf{p}\| < \|\mathbf{v} - \mathbf{u}\|$.*

Proof. Write

$$\mathbf{v} - \mathbf{u} = (\mathbf{v} - \mathbf{p}) + (\mathbf{p} - \mathbf{u}).$$

The vector $\mathbf{p} - \mathbf{u}$ is in U, and by definition of $\mathbf{p}$, $\mathbf{v} - \mathbf{p} \in U^\perp$. Therefore by the Pythogorean Theorem 8.2.3,

$$\|\mathbf{v} - \mathbf{u}\|^2 = \|\mathbf{v} - \mathbf{p}\|^2 + \|\mathbf{p} - \mathbf{u}\|^2,$$

so we get the strict inequality $\|\mathbf{v} - \mathbf{p}\| < \|\mathbf{v} - \mathbf{u}\|$ unless $\|\mathbf{p} - \mathbf{u}\| = \mathbf{0}$, which means that $\mathbf{u} = \mathbf{p}$. □

8.5 Solving the Inconsistent Inhomogeneous System

Consider the inhomogeneous linear system in a Euclidean space:

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

where A is a $m \times n$ matrix. We assume that m is larger than n, and that the rank of A is n, so that its columns are linearly independent. Therefore the linear map $L_A : \mathbb{R}^n \to \mathbb{R}^m = V$ is injective, so the range of L_A is a proper subspace U of dimension n of V by the Rank-Nullity theorem. Then

Proposition 8.5.1. *The square matrix A^tA is invertible if and only if A has maximum rank.*

Proof. First assume that A^tA is invertible. Then for any non-zero $\mathbf{v} \in \mathbb{R}^n$, $A^tA\mathbf{v} \neq \mathbf{0}$. Assume that $\mathbf{v}^tA^tA\mathbf{v} = 0$. This can be rewritten $\langle A\mathbf{v}, A\mathbf{v}\rangle = 0$. Because the scalar product is Euclidean, this can only happen if $A\mathbf{v} = \mathbf{0}$, which is impossible since A has maximal rank, as we just saw.

Now assume A has maximum rank, so for a non-zero $\mathbf{v}$, $A\mathbf{v} \neq 0$. As before the inner product of $A\mathbf{v}$ with itself can be written $\mathbf{v}^tA^tA\mathbf{v}$, which is non-zero. This is impossible unless $A\mathbf{v} = 0$, a contradiction. □

Typically, a right hand vector $\mathbf{b} \in \mathbb{R}^m$ does not lie in the range U of L_A, so the equation $A\mathbf{x} = \mathbf{b}$ is not solvable. If this is the case, we say the system is inconsistent: see Definition 1.1.9. Still, we want to find the best possible approximate solution of the system of linear equations, in some sense.

Here is the approach of the previous section. Project $V = \mathbb{R}^m$ to the range U of L_A. The image $\mathbf{p}$ of $\mathbf{b}$ in U is the point in U that is closest to $\mathbf{b}$, as we showed in Theorem 8.4.4.

So given any $\mathbf{b} \in \mathbb{R}^m$, we need to compute its orthogonal projection $\mathbf{p}$ in the range U. Then instead of solving $A\mathbf{x} = \mathbf{b}$, which is impossible, we will solve $A\mathbf{x} = \mathbf{p}$.

Theorem 8.5.2. *Assuming that A is a $m \times n$ matrix, $m > n$, of rank n, then the orthogonal projection $\mathbf{p}$ of any point $\mathbf{b}$ in $V = \mathbb{R}^m$ to the range $U \subset V$ of the map $\mathbf{x} \to A\mathbf{x}$ is given by*

$$\mathbf{p} = A(A^tA)^{-1}A^t\mathbf{b}$$

Proof. The columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$ of A form a basis of the range U, since the rank of A is n, so the projection $\mathbf{p}$ of $\mathbf{b}$ can be written uniquely as a linear combination

$$\mathbf{p} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n \tag{8.11}$$

for real constants x_i. Equation (8.11) is the matrix product

$$\mathbf{p} = A\mathbf{x} \tag{8.12}$$

as you should convince yourself. Our goal is to determine the x_i.

By definition of the orthogonal projection to U, $\mathbf{b}-\mathbf{p}$ is orthogonal to all the vectors in U. Since U is spanned by the columns $\mathbf{a}_i$ of A, this is equivalent to

$$\langle \mathbf{a}_i, \mathbf{b}-\mathbf{p}\rangle = 0, \quad 1 \le i \le n.$$

This system of equations can be rewritten as the matrix product $A^t(\mathbf{b}-\mathbf{p}) = \mathbf{0}$. Replacing $\mathbf{p}$ by $A\mathbf{x}$ as in (8.12) we get the key condition:

$$A^t(\mathbf{b}-A\mathbf{x}) = 0, \quad \text{or} \quad A^tA\mathbf{x} = A^t\mathbf{b}. \tag{8.13}$$

This is a system of n equations in n variables. Because A^tA is invertible by hypothesis, we can solve for the unknowns $\mathbf{x}$ and get the **normal equations**:

$$\boxed{\mathbf{x} = (A^tA)^{-1}A^t\mathbf{b}} \tag{8.14}$$

Finally we can find the projection point:

$$\mathbf{p} = A\mathbf{x} = A(A^tA)^{-1}A^t\mathbf{b}. \tag{8.15}$$

□

So the linear transformation with matrix $P = A(A^tA)^{-1}A^t$ takes any $\mathbf{b} \in V$ to its projection $\mathbf{p} \in U$. Since A is a $m \times n$ matrix , P is a $m \times m$ matrix. Thus $\mathbf{p}$ is a vector in V that lies in the range U of A.

Remark 8.5.3. The basis of the range of L_A given by the columns of A is not generally orthonormal. However using the QR factorization of Theorem 8.3.9 we can see what happens when we choose an orthonormal basis. This is an amusing computation. Just replace A by QR in (8.15), using the fact that $Q^tQ = I$ and R is square to get

$$\begin{aligned} A(A^tA)^{-1}A^t &= QR(R^tQ^tQR)^{-1}R^tQ^t = QR(R^tR)^{-1}R^tQ^t = QRR^{-1}(R^t)^{-1}R^tQ^t \\ &= QQ^t \end{aligned}$$

which is the projection formula we derived in Chapter 8. A word of warning: Q is, like A, a $m \times n$ matrix, with $m > n$. Q^tQ is the identity matrix, but QQ^t is not necessarily diagonal. The matrix A in Example 8.5.5 has orthogonal but not orthonormal columns. Dividing by their length makes it a Q matrix. Compute that QQ^T is not the diagonal matrix.

Remark 8.5.4. Let's confirm Theorem 8.4.3. First notice that $P^2 = P$:

$$P^2 = A(A^tA)^{-1}A^tA(A^tA)^{-1}A^t = A(A^tA)^{-1}A^t = P$$

by cancellation of one of the $(A^tA)^{-1}$ by A^tA in the middle. Also notice that P is symmetric by computing its transpose:

$$P^t = (A^t)^t((A^tA)^{-1})^tA^t = A((A^tA)^t)^{-1}A^t = A(A^tA)^{-1}A^t = P.$$

We used $((A^tA)^t)^{-1} = ((A^tA)^{-1})^t$ and of course we used $(A^t)^t = A$. So we have shown (no surprise, since it is a projection matrix): the matrix P satisfies the hypotheses of Theorem 8.4.3: it is symmetric and $P^2 = P$.

Problem 8.5.5. Compute A^tA for the rank 2 matrix

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$$

and show that it is positive definite.

$$A^tA = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$$

This is obviously positive definite. In this case it is easy to work out the projection matrix $A(A^tA)^{-1}A^t$:

$$P = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1/3 & 0 \\ 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 5/6 & 2/6 & -1/6 \\ 2/6 & 2/6 & 2/6 \\ -1/6 & 2/6 & 5/6 \end{pmatrix}$$

which is of course symmetric, and $P^2 = P$ as you should check.

In conclusion, given an inconsistent system $A\mathbf{x} = \mathbf{b}$, the technique explained above shows how to replace it by the solvable system $A\mathbf{x} = \mathbf{p}$ that matches it most closely.

Orthogonal projections will be used when we study the method of least squares.

We proved Theorem 8.5.2 using Theorem 8.4.4 a distance minimizing result proved just using the Pythagorean Theorem 8.2.3. Here is an alternate approach for those of you who enjoy optimization techniques using multivariable calculus.

Remark 8.5.6. As before our goal is to minimize the expression

$$\|A\mathbf{x} - \mathbf{b}\|$$

as a function of the n variables x_i, using the standard minimization technique from multivariable calculus. First, to have an easier function to deal with, we take its square, which we write as a matrix product:

$$f(\mathbf{x}) = (\mathbf{x}^tA^t - \mathbf{b}^t)(A\mathbf{x} - \mathbf{b}) = \mathbf{x}^tA^tA\mathbf{x} - \mathbf{x}^tA^t\mathbf{b} - \mathbf{b}^tA\mathbf{x} + \mathbf{b}^t\mathbf{b}.$$

Notice that each term is a number: check the size of the matrices and the vectors involved. Calculus tells us $f(\mathbf{x})$, which is a quadratic polynomial in the x_i, has an extremum (minimum or maximum) only when all the partial derivatives with respect

to the x_i vanish. It is an exercise to see that the vector of partial derivatives, namely the gradient ∇f of f in $\mathbf{x}$ is

$$2A^tA\mathbf{x} - 2A^t\mathbf{b},$$

so setting this to $\mathbf{0}$ gives the key condition (8.13) back. No surprise. We can finish the problem by computing the second derivative, called the Hessian. It is the constant matrix $2A^tA$. By hypothesis it is invertible, so Proposition 8.5.1 tells us that A^tA is positive definite, which guarantees that the point found is a minimum.

Theorem 8.5.2 shows that it is possible to bypass this calculus approach by using perpendicularity.

8.6 Hermitian Products

When the scalars are the complex numbers, we modify the notion of the scalar product to preserve positivity.

Definition 8.6.1. Let V be a vector space over $\mathbb{C}$. A Hermitian product on V associates to any pair $\mathbf{v}$, $\mathbf{w}$ of elements of V a complex number written $\langle \mathbf{v}, \mathbf{w} \rangle$ which satisfies the following three properties for all $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ in V:

HP 1 $\langle \mathbf{v}, \mathbf{w} \rangle = \overline{\langle \mathbf{w}, \mathbf{v} \rangle}$, where $\overline{\langle \mathbf{w}, \mathbf{v} \rangle}$ denotes the complex conjugate of $\langle \mathbf{w}, \mathbf{v} \rangle$;
HP 2 Additivity: $\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{x}, \mathbf{w} \rangle$;
HP 3 For $a \in \mathbb{C}$, $\langle a\mathbf{v}, \mathbf{w} \rangle = a\langle \mathbf{v}, \mathbf{w} \rangle$.

Exercise 8.6.2. For two complex numbers a and b, prove that $\overline{ab} = \overline{a}\overline{b}$: the complex conjugate of the product is the product of the complex conjugates.

Exercise 8.6.3. Prove that the definition implies

1. Additivity in the first variable:

$$\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle;$$

Hint: justify

$$\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \overline{\langle \mathbf{w}, \mathbf{u} + \mathbf{v} \rangle} = \overline{\langle \mathbf{w}, \mathbf{u} \rangle} + \overline{\langle \mathbf{w}, \mathbf{v} \rangle} = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle.$$

2. Conjugate linearly in the second variable:

$$\langle \mathbf{v}, a\mathbf{w} \rangle = \overline{a}\langle \mathbf{v}, \mathbf{w} \rangle \text{ for all } a \in \mathbb{C};$$

Hint: justify

$$\langle \mathbf{v}, a\mathbf{w} \rangle = \overline{\langle a\mathbf{w}, \mathbf{v} \rangle} = \overline{a\langle \mathbf{w}, \mathbf{v} \rangle} = \overline{a}\overline{\langle \mathbf{w}, \mathbf{v} \rangle} = \overline{a}\langle \mathbf{v}, \mathbf{w} \rangle. \tag{8.16}$$

Thus Hermitian products are not linear in their second variable.

3. Finally show the easy:

$$\langle \mathbf{0}, \mathbf{w} \rangle = 0.$$

Remark 8.6.4. The elementary, but key remark that motivates this definition is that for any vector $\mathbf{v} \in V$,

$$\langle \mathbf{v}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{v} \rangle} \text{ by the first property,}$$

so that $\langle \mathbf{v}, \mathbf{v} \rangle$ is always real.

The key example of a Hermitian product is

Example 8.6.5. On $\mathbb{C}^n$, the product given by

$$\langle \mathbf{v}, \mathbf{w} \rangle = v_1\overline{w_1} + v_2\overline{w_2} + \cdots + v_n\overline{w_n}$$

is Hermitian.

Definition 8.6.6. A Hermitian product is positive definite if for all non-zero $\mathbf{v} \in V$

$$\langle \mathbf{v}, \mathbf{v} \rangle > 0.$$

Exercise 8.6.7. Show that the Hermitian product of Example 8.6.5 is positive definite.

Definition 8.6.8. Let V be a complex vector space with Hermitian product $\langle \mathbf{v}, \mathbf{w} \rangle$. We say that $\mathbf{v}$ is perpendicular, or orthogonal to $\mathbf{w}$ if

$$\langle \mathbf{v}, \mathbf{w} \rangle = 0.$$

We write $\mathbf{v} \perp \mathbf{w}$ if this is the case. For any subset S of V, we let $S^\perp$ be the set of elements of V perpendicular to all the elements of S.

Note the potential confusion with the similar notations of positive definiteness and orthogonality for scalar products in the real case.

Exercise 8.6.9. Given a complex vector space V of dimension n with a Hermitian product $\langle \mathbf{v}, \mathbf{w} \rangle$, write the form in terms of its real and imaginary parts:

$$\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle_{\mathbb{R}} + \mathrm{i}\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbb{C}}$$

What can we say about these parts? The form $\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbb{R}}$ is real valued and symmetric; $\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbb{C}}$ is real valued and alternating, meaning that $\langle \mathbf{w}, \mathbf{v} \rangle_{\mathbb{C}} = -\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbb{C}}$

8.7 The Geometry of Hermitian Spaces

In this section we modify the results of §8.2 to the case of a complex vector space V with a positive definite Hermitian product: see Definition 8.6.1. On first reading, you should think of $\mathbb{C}^n$ with the positive definite Hermitian product of Example 8.6.5.

As in the real case, any subspace W of V, inherits a positive definite Hermitian product from V.

The *norm* of $\mathbf{v}$ is:

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}. \tag{8.17}$$

where here $\langle\rangle$ denotes the Hermitian product When V is $\mathbb{C}^n$, with the standard Hermitian product, then

$$\|\mathbf{v}\| = \sqrt{v_1 \overline{v_1} + \cdots + v_n \overline{v_n}}.$$

As in the real case we have

$$\langle \mathbf{v}, \mathbf{v} \rangle = 0 \text{ if and only if } \mathbf{v} = \mathbf{0}.$$

and

$$\|c\mathbf{v}\| = |c| \|\mathbf{v}\| \text{ for all } c \in \mathbb{C}.$$

The distance between two vectors is defined in the same way as the real case.

The Cauchy-Schwarz Inequality still holds: for any two vectors $\mathbf{v}$ and $\mathbf{w}$ in V,

$$|\langle \mathbf{v}, \mathbf{w} \rangle| \leq \|\mathbf{v}\| \|\mathbf{w}\|$$

and as in the real case it implies the triangle inequality: for all $\mathbf{u}$ and $\mathbf{v}$ in V,

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|.$$

for which we write the proof to show the difference with scalar products.

Proof (Triangle Inequality). Square the left hand side:

$$\|\mathbf{u} + \mathbf{v}\|^2 = \langle \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v} \rangle = \|\mathbf{u}\|^2 + \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{u} \rangle + \|\mathbf{v}\|^2.$$

Since

$$\langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{u} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \overline{\langle \mathbf{u}, \mathbf{v} \rangle}$$

we are adding two complex conjugate numbers, so the sum is real. It is at most $2|\langle \mathbf{u}, \mathbf{v} \rangle|$. Use the Cauchy-Schwarz inequality to replace $2|\langle \mathbf{u}, \mathbf{v} \rangle|$ by the larger term $2\|\mathbf{u}\|\|\mathbf{v}\|$ to get

$$\|\mathbf{u}\|^2 + 2\langle \mathbf{u}, \mathbf{v} \rangle + \|\mathbf{v}\|^2 \leq \|\mathbf{u}\|^2 + 2\|\mathbf{u}\|\|\mathbf{v}\| + \|\mathbf{v}\|^2.$$

We recognize the right-hand side as the square of $\|\mathbf{u}\| + \|\mathbf{v}\|$, so we get

$$\|\mathbf{u} + \mathbf{v}\|^2 \leq (\|\mathbf{u}\| + \|\mathbf{v}\|)^2.$$

Taking the square root of both sides, we are done. □

Exercise 8.7.1. For any complex number a written in terms of its real and imaginary parts as $a = b + \mathrm{i}c$, show that $a + \overline{a} \leq 2|a|$, where $|a|$ of course is $\sqrt{b^2 + c^2}$. We used this simple result in the proof above.

Exactly as in the real case we can prove:

Theorem 8.7.2 (Gram-Schmidt). *If V is a complex vector space of dimension n with a positive definite Hermitian product, then V has a basis $\mathbf{w}_1, \ldots, \mathbf{w}_n$, where the basis vectors are mutually perpendicular: $\langle \mathbf{w}_i, \mathbf{w}_j \rangle = 0$ whenever $i \neq j$.*

Furthermore any set of non-zero vectors $\mathbf{w}_1, \ldots, \mathbf{w}_k$ that are mutually perpendicular can be extended to a basis where all the basis vectors are mutually perpendicular.

From complex Gram-Schmidt orthogonalization, we again get a QR factorization. First we define the analog of an orthogonal matrix in the real case.

Definition 8.7.3. A matrix Q is unitary it is is invertible and if its inverse is its conjugate transpose: $Q^* = Q^{-1}$. So $Q^*Q = I = QQ^*$.

In particular the columns $\mathbf{q}_i$ of Q are mutually perpendicular, and each column has length 1: $\|\mathbf{q}_i\| = 1$. Conversely if you have an orthonormal basis for V, each vector normalized to length 1, then the matrix whose columns are the normalized basis elements, in any order, is unitary.

Theorem 8.7.4. *Any complex invertible matrix A of size n can be written as the product of a unitary matrix Q and an upper triangular matrix R:*

$$A = QR$$

As in the real case this follows immediately from Gram-Schmidt orthonormalization.

As in the real case we get:

Theorem 8.7.5. *Let V be a complex vector space with an positive definite Hermitian product, and W a subspace of V. Then V is the direct sum of W and its orthogonal complement $W^\perp$.*

$$V = W \oplus W^\perp.$$

8.8 Scalar Product on Spaces of Matrices

Definition 8.8.1. Let M_n denote the vector space of all real $n \times n$ matrices.

As we know, M_n is a vector space of dimension n^2: here is a basis. Let E_{ij} be the element of M_n with 0 in every position except the (ij)-th position where it has a 1. The E_{ij}, $1 \leq i, j \leq n$, form a basis for M_n, confirming that M has dimension n^2.

Consider the following scalar product on M_n. As we will see, the E_{ij} form an orthonormal basis for this inner product.

Theorem 8.8.2. *For any two matrices A and B in M_n, let $\langle A, B \rangle = \operatorname{tr}(AB^t)$. This is an inner product on M_n for which the E_{ij} form an orthonormal basis.*

Proof. First note that $\langle A,B\rangle$ is linear in each variable, since the trace is, and that $\langle A,B\rangle = \langle B,A\rangle$. For this last point we need to show that $\mathrm{tr}\,(AB^t) = \mathrm{tr}\,(BA^t)$. But the two matrices AB^t and BA^t are just transposes of each other, and transposes have the same trace, since the trace is just the sum of the diagonal elements. The final point is just the computation that the E_{ij} form an orthonormal basis. Indeed $E_{ij}E_{ij}^t = E_{ii}$, which has trace 1, and $E_{ij}E_{kl}^t$ is the 0-matrix unless $i = k$ and $j = l$. □

Since we have a positive definite scalar product, we can form the norm on M_n:

$$\|A\| = \sqrt{\langle A,A\rangle} = \sqrt{\mathrm{tr}\,(A^2)}$$

and the distance function

$$d(A,B) = \|A-B\|.$$

Exercise 8.8.3. On the space M_{mn} of real $m \times n$ matrices, consider the expression $\mathrm{tr}(AB^t)$. Show that it is an inner product. Find an orthonormal basis for this inner product. Show that $\mathrm{tr}(A^tB)$ is also an inner product, and find an orthonormal basis.

Inside M_n we look at the symmetric matrices, which form a vector subspace S_n of dimension $n(n+1)/2$, as you can determine by counting the diagonal entries (n of them) and the above diagonal entries ($(n^2-n)/2$ of them).

We now let $A = (a_{ij})$ and $B = (b_{ij})$ be two matrices in S_n. We get an inner product on S_n, by restricting the one on M_n.

Definition 8.8.4. The inner product on S_n is given by

$$\langle A,B\rangle = \mathrm{tr}\,(AB) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ji}.$$

The norm and the distance function are as before.

Just as in Theorem 8.8.2, we write an orthonormal basis of S_n for this product. Not surprisingly, we just modify the basis of M_n to get symmetric matrices. Using the notation of Theorem 8.8.2, we take the diagonal matrices E_{ii}, and then the symmetric matrices $(E_{ij}+E_{ji})/\sqrt{2}$, $1 \le i < j \le n$. They form an orthonormal basis, so we have an inner product space.

Exercise 8.8.5. Using the inner product for M_n given above, project it orthogonally to the subspace S_n of symmetric matrices. What is the nullspace of this linear map? In other words, what matrices are orthogonal to all symmetric matrices? Find an orthonormal basis for the nullspace: be sure to check that it has the right dimension.

Proposition 8.8.6. *This norm interacts nicely with multiplication of matrices, namely*

$$\|AB\| \le \|A\|\|B\|$$

Proof. We need to prove that

$$\operatorname{tr}(ABAB) \leq \operatorname{tr}(A^2)\operatorname{tr}(B^2) \tag{8.18}$$

Write C for the product matrix AB. The the entries c_{ij} of C are

$$c_{ij} = \sum_k a_{ik}b_{ik}$$

so then

$$\operatorname{tr}(ABAB) = \operatorname{tr}(C^2) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(\sum_{k=1}^{n} a_{ik}b_{ik})^2$$

We need to show that this is less than or equal to

$$(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2)(\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}^2)$$

Notice that both sides are sums of terms of the form $a_{ij}^2 b_{kl}^2$, and on the left-hand side there are many fewer terms (only terms with an index in common) that on the right-hand side (all possible indices). □

How to we use this additional property? See [13], §5.6. First note that $\|I_n\| = n$ and then, if A is invertible, that

$$\|A^{-1}\| \geq \frac{n}{\|A\|}.$$

Example 8.8.7. If A is a symmetric matrix with $\|A\| < 1$, then A^i, the i-th power of A, tends to the zero matrix as i tends to ∞.

Example 8.8.8. If A is a symmetric matrix, then the infinite series of matrices

$$\sum_{i=1}^{\infty} c_i A^i$$

converges if the infinite numerical series

$$\sum_{i=1}^{\infty} c_i \|A\|^i$$

converges.

Now we consider the complex case, so let $M_n(\mathbb{C}$ be the complex vector space of $n \times n$ complex matrices. It has dimension n^2, as in the real case. We define a Hermitian basis by setting $\langle A, B\rangle = \operatorname{tr}(AB^*)$, so we take the conjugate transpose of the second matrix. This inner product is clearly Hermitian. The basis we defined in the real case is still an orthonormal basis, and the form is positive definite. The remaining details are left to you.

Example 8.8.9. We construct a linear map from the $\binom{n}{2}$ dimensional space S_n to the $2n-1$-dimensional vector space V of polynomials of degree at most $2n-2$ in the variable t . To the symmetric matrix $A = (a_{ij}) \in S_n$ in the variables a_{ij} and the n-vector $\mathbf{t} = (1, t, t^2, t^3, \ldots, t^{n-1})$, we associate the polynomial in t of degree $2n-2$:

$$\mathbf{t}^t A \mathbf{t} = \sum_{k=0}^{2n-2} r_k t^k \quad \text{with} \quad r_k = \sum_{i+j=k+2} a_{ij}.$$

The $2n-1$ coefficients of this polynomial, of degree 1 in the a_{ij} give a linear map $\phi\colon S_n \to \mathbb{R}^{2n-1}$. It is easy to see that this map is surjective, so that the nullspace of ϕ has dimension $n(n+1)/2-(2n-1) = (n-1)(n-2)/2 = \binom{n-1}{2}$.

Now a Hankel form associates to any vector $\mathbf{s} = (s_0, s_1, \ldots, s_{2n-2})$ in $\mathbb{R}^{2n-1}$ a symmetric matrix $A_\mathbf{s}$ (7.2). This is also a linear map, call it $\psi\colon \mathbb{R}^{2n-1} \to S_n$. It has trivial null space.

Thus we get a composite linear map $\phi \circ \psi\colon \mathbb{R}^{2n-1} \to \mathbb{R}^{2n-1}$. It associates to a vector $\mathbf{s} = (s_0, s_1, \ldots, s_{2n-2})$ the coefficients of the polynomial $f_\mathbf{s}(t) = \sum_{k=0}^{2n-2} r_k t^k$, where $r_k = s_k(\sum_{i+j=k} 1) = \ell(k) s_k$, where $\ell(k)$ counts the (constant) antidiagonal terms in the k-th antidiagonal of an $n \times n$ matrix, starting the numbering at 0. So when $n = 3$, the vector $(s_0, \ldots, s_4)$ gets mapped by ψ to the matrix

$$A_s = \begin{pmatrix} s_0 & s_1 & s_2 \\ s_1 & s_2 & s_3 \\ s_2 & s_3 & s_4 \end{pmatrix}$$

which gets mapped by ϕ to the polynomial

$$f_\mathbf{s}(t) = s_0 + 2s_1 t + 3s_2 t^2 + 2s_3 t^3 + s_4 t^4.$$

and $\ell(0) = \ell(4) = 1$, $\ell(1) = \ell(3) = 2$, $\ell(2) = 3$. This computation obviously generalizes to all n.

Thus the matrix of the composite map $\phi \circ \psi$ from the $\{s_i\}$ basis to the $\{t^j\}$ basis is diagonal: in the case $n = 3$ the diagonal entries are $1, 2, 3, 2, 1$. So the map is surjective, and any polynomial of degree at most $2n-2$ can be written uniquely as the range of a Hankel form of size n. So we can use polynomials to study Hankel forms.

It turns out this can be used to estimate the rank of a Hankel matrix. For example, suppose that the polynomial $f_\mathbf{s}(t)$ associated to $\mathbf{s}$ has a real root t_0. If $\mathbf{t}_0 = (1, t_0, t_0^2, \ldots, t_0^{n-1})$, then by construction the quadratic form $\mathbf{t}_0^t A_\mathbf{s} \mathbf{t}_0 = 0$, so the nullspace of the Hankel form $A_\mathbf{s}$ is at at least one dimensional. In particular this form is not positive definite.

Chapter 9
Operators on Inner Product Spaces

Abstract we start out by giving the classical definition of the adjoint of a linear operator, first in the real case and then in the complex case. Using it, we list the linear operators with special properties, some of which we have already encountered. We also explain where we will encounter them later in this book.

9.1 Adjoints on Real Spaces and Symmetric Matrices

Let V be a Euclidean space. Thus V is a $\mathbb{R}$-vector space with a positive-definite inner product $\langle \mathbf{v}, \mathbf{w} \rangle$. Fix a vector $\mathbf{w}$, and let $f_{\mathbf{w}} \colon V \to \mathbb{R}$ be the function $f_{\mathbf{w}}(\mathbf{v}) = \langle \mathbf{v}, \mathbf{w} \rangle$. The function $f_{\mathbf{w}}$ is a linear map because the inner product is linear in its first variable $\mathbf{v}$. So $f_{\mathbf{w}}$ is a linear functional, as defined in (6.1.1). To avoid confusion, instead of giving the variable of $f_{\mathbf{w}}$ a name, we sometimes replace by a dot, so we write $f_{\mathbf{w}}(\cdot) = \langle \cdot, \mathbf{w} \rangle$.

The next theorem is just a restatement of Theorem 7.1.8, which says that a non-degenerate bilinear form gives an isomorphism between V and its dual V^*, the vector space of linear functionals. In particular it is not necessary to assume that the scalar product is positive-definite: it only need be non-degenerate. Here is a simple independent proof when the inner product is positive definite..

Theorem 9.1.1. *Let V be a Euclidean space, and $f(\mathbf{v})$ any linear functional on V. Then there is a unique $\mathbf{w} \in V$ such that $f(\mathbf{v}) = \langle \mathbf{v}, \mathbf{w} \rangle$. Thus the map $\mathrm{d} \colon \mathbf{w} \mapsto f_{\mathbf{w}}$, where $f_{\mathbf{w}}(\mathbf{v}) = \langle \mathbf{v}, \mathbf{w} \rangle$, is an isomorphism of V with its dual V^*.*

Proof. First we need to show that the map $\mathrm{d} \colon \mathbf{w} \mapsto f_{\mathbf{w}}$ is linear. Indeed

$$\mathbf{w} + \mathbf{z} \mapsto \langle \cdot, \mathbf{w} + \mathbf{z} \rangle = \langle \cdot, \mathbf{w} \rangle + \langle \cdot, \mathbf{z} \rangle$$

and

$$c\mathbf{w} \mapsto \langle \cdot, c\mathbf{w} \rangle = c\langle \cdot, \mathbf{w} \rangle$$

using the linearity of the inner product in the second variable. Since V and V^* have the same dimension it is enough to show that this map is injective. If not that would mean that there is a non-zero element $\mathbf{w}$ such that $\langle \cdot, \mathbf{w} \rangle$ is the zero functional. This is impossible since $\langle \mathbf{w}, \mathbf{w} \rangle \neq 0$. □

Now we pick an arbitrary linear operator $L\colon V \to V$, and consider the mapping $h_{\mathbf{w}}\colon V \to \mathbb{R}$ given by

$$h_{\mathbf{w}}(\mathbf{v}) = \langle L\mathbf{v}, \mathbf{w} \rangle.$$

Since $h_{\mathbf{w}}$ is the composition of the linear maps:

$$\mathbf{v} \in V \longmapsto L\mathbf{v} \longmapsto \langle L\mathbf{v}, \mathbf{w} \rangle \in \mathbb{R}$$

$h_{\mathbf{w}}(\mathbf{v})$ is a linear functional. So by Theorem 9.1.1, there is a unique vector $\mathbf{z} \in V$ such that $h_{\mathbf{w}}(\mathbf{v}) = \langle \mathbf{v}, \mathbf{z} \rangle$. Therefore $\langle L\mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{z} \rangle$. So holding L and $\mathbf{v}$ fixed, we can view $\mathbf{z}$ as a function M of $\mathbf{w}$. The key point is that $M\colon V \to V$ is a linear function. To prove this first compute

$$\begin{aligned}
\langle \mathbf{v}, M(\mathbf{w}_1 + \mathbf{w}_2) \rangle &= \langle L\mathbf{v}, \mathbf{w}_1 + \mathbf{w}_2 \rangle && \text{by definition of } M \\
&= \langle L\mathbf{v}, \mathbf{w}_1 \rangle + \langle L\mathbf{v}, \mathbf{w}_2 \rangle && \text{by linearity of } \langle , \cdot \rangle \\
&= \langle \mathbf{v}, M\mathbf{w}_1 \rangle + \langle \mathbf{v}, M\mathbf{w}_2 \rangle && \text{by definition of } M \\
&= \langle \mathbf{v}, M\mathbf{w}_1 + M\mathbf{w}_2 \rangle && \text{by linearity of } \langle , \cdot \rangle
\end{aligned}$$

Since this is true for any $\mathbf{v}$, we get $M(\mathbf{w}_1 + \mathbf{w}_2) = M(\mathbf{w}_1) + M(\mathbf{w}_2)$, the first of the two equalities needed to prove linearity.

Similarly, for any scalar c,

$$\langle \mathbf{v}, M(c\mathbf{w}) \rangle = \langle L\mathbf{v}, c\mathbf{w} \rangle = c\langle L\mathbf{v}, \mathbf{w} \rangle = c\langle \mathbf{v}, M(\mathbf{w}) \rangle = \langle \mathbf{v}, cM(\mathbf{w}) \rangle$$

so by the same argument as before $M(c\mathbf{w}) = cM(\mathbf{w})$. We have proved:

Theorem 9.1.2. *For any linear operator L on the Euclidean space V there is a unique linear operator M on V, called the* adjoint *of L, such that*

$$\langle L\mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}, M\mathbf{w} \rangle.$$

The adjoint of L is written L^t.

Obviously the adjoint of the identity map I is itself.

Definition 9.1.3. The operator L is *self-adjoint* or *symmetric* on the Euclidean space V if $L = L^t$. Then

$$\langle L\mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}, L\mathbf{w} \rangle.$$

Theorem 9.1.4. *Let L and M be linear operators on the real inner product space V. Then the adjoint with respect to this inner product satisfies:*

$$\begin{aligned}(L+M)^t &= L^t + M^t;\\ (L\circ M)^t &= M^t \circ L^t;\\ (rL)^t &= rL^t, \forall r \in \mathbb{R};\\ (L^t)^t &= L.\end{aligned}$$

Proof. Here is a proof of the second identity.

$$\langle (L\circ M)\mathbf{v}, \mathbf{w}\rangle = \langle M\mathbf{v}, L^t\mathbf{w}\rangle = \langle \mathbf{v}, M^t \circ L^t\mathbf{w}\rangle.$$

Since this is true for all $\mathbf{v}$ and $\mathbf{w}$ in V, by definition we see that $(L\circ M)^t = M^t \circ L^t$. The other identities are even simpler to prove. □

Corollary 9.1.5. *Assume L is a self adjoint operator on V, and that C is an arbitrary operator on V. Then $M = C^tLC$ is self adjoint. Furthermore, if C is invertible, and C^tLC is self adjoint, then L is also self-adjoint.*

Proof. We use the previous theorem several times. The adjoint of C^tLC is C^tL^tC, so if L is self adjoint, so is C^tLC. Furthermore if C is invertible, so is C^t: indeed let D be the inverse of C, so $DC = I$. Then $(DC)^t = C^tD^t = I$. Finally assume C^tLC is self-adjoint, so

$$C^tLC = (C^tLC)^t = C^tL^tC.$$

Then just compose with the operator $(C^t)^{-1}$ on the left, and C^{-1} on the right to get $L = L^t$. □

Theorem 9.1.6. *An operator L on a Euclidean space is the zero operator if and only if $\langle L\mathbf{v}, \mathbf{w}\rangle = 0$ for all $\mathbf{v}$ and $\mathbf{w}$ in V.*

Proof. This is a special case of Theorem 6.3.4, but it is easy to provide a simpler proof. One direction is obvious: if L is the zero operator, $\langle L\mathbf{v}, \mathbf{w}\rangle = 0$.

For the reverse implication, if all the inner products vanish, then choosing for $\mathbf{w}$ the vector $L\mathbf{v}$, we would have $\langle L\mathbf{v}, L\mathbf{v}\rangle = 0$, which by positive definiteness can only happen if $L\mathbf{v} = \mathbf{0}$. □

Here is another result we have already proved by bilinear form techniques.

Theorem 9.1.7. *A self-adjoint operator L on a Euclidean space is the zero operator, if and only if $\langle L\mathbf{v}, \mathbf{v}\rangle = 0$ for all $\mathbf{v}$ in V.*

Proof. As in the previous theorem, one implication in obvious. For the other, just establish directly a polarization identity of Lemma 7.2.1:

$$\langle L\mathbf{v}, \mathbf{w}\rangle + \langle L\mathbf{w}, \mathbf{v}\rangle = \langle L(\mathbf{v}+\mathbf{w}), \mathbf{v}+\mathbf{w}\rangle - \langle L\mathbf{v}, \mathbf{v}\rangle - \langle L\mathbf{w}, \mathbf{w}\rangle \tag{9.1}$$

by expanding, using linearity, $\langle L(\mathbf{v}+\mathbf{w}), \mathbf{v}+\mathbf{w}\rangle$. Then use the fact that L is self-adjoint:

$$\langle L\mathbf{v}, \mathbf{w}\rangle = \langle L\mathbf{w}, \mathbf{v}\rangle.$$

Since the hypothesis says that all the terms on the right hand side of (9.1) are 0, we get $\langle L\mathbf{v}, \mathbf{w}\rangle = 0$, so by Theorem 9.1.6, L is the zero operator. □

Exercise 9.1.8. Show that any operator L can be written uniquely as the sum of a self adjoint operator and a skew adjoint operator, meaning that $L^t = -L$.

Theorem 9.1.9. *Let V be a Euclidean space and $\mathfrak{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ an orthonormal basis of V. Let L be any operator on V, and let A be the matrix of L associated to L in the basis $\mathfrak{B}$. Then the matrix of the adjoint L^t is the transpose of A. Therefore L is self adjoint if and only if A is symmetric.*

This is why we write the adjoint of a linear transformation with the same symbol as that of the transpose of a matrix.

Proof. Write $\mathbf{v} = x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n$ and $\mathbf{w} = y_1\mathbf{v}_1 + \cdots + y_n\mathbf{v}_n$ in the basis $\mathfrak{B}$. Thus the inner product $\langle \mathbf{v}, \mathbf{w} \rangle = x_1y_1 + \ldots x_ny_n$, namely the matrix product $\mathbf{x}^t\mathbf{y}$. Since A is the matrix associated to L in the basis $\mathfrak{B}$, L maps the vector with coordinates (x_i) to the vector with coordinates (z_i) given by $\mathbf{z} = A\mathbf{x}$ according to Definition 5.1.2. Then because $\mathfrak{B}$ is orthonormal, $\langle L\mathbf{v}, \mathbf{w} \rangle = (A\mathbf{x})^t\mathbf{y}$. By the properties of matrix multiplication $(A\mathbf{x})^t\mathbf{y} = \mathbf{x}^tA^t\mathbf{y} = \mathbf{x}^t(A^t\mathbf{y}) = \langle \mathbf{v}, M\mathbf{w} \rangle$ where M is the linear transformation with matrix A^t. M is clearly the adjoint of L, so the matrix associated to the adjoint L^t is the transpose of A. Finally L is self adjoint if and only if A is symmetric. □

Remark 9.1.10. This construction is closely tied to the construction in §6.6, in the special case where $W = V$. In the current situation, all elements of V^* can be written as $h_{\mathbf{w}}(\cdot) = \langle L(\cdot), \mathbf{w} \rangle$, as $\mathbf{w}$ varies over V. The map $L^*\colon V^* \to V^*$ of Definition 6.6.2 maps the linear functional $h_{\mathbf{w}}(\cdot)$ to the composite $h_{\mathbf{w}}(L(\cdot))$. By Theorem 9.1.2 we can write this as $h_{M(\mathbf{w})}(\cdot)$ and M is what we called the transpose L^* in §6.6. Not surprisingly, this is our adjoint, which as we noted in Theorem 6.6.4 in suitable bases has for matrix the transpose of the original matrix.

should I put exercises about positive definiteness, etc, here or elsewhere? Perhaps move some of the material from Bilinear Forms to this chapter?

9.2 Adjoints for Hermitian Products and Hermitian Matrices

We now imitate what we did in §9.1 for V is a complex vector space with a positive-definite Hermitian product $\langle \mathbf{v}, \mathbf{w} \rangle$. We establish the same results as in first section. We need to be a little more careful, because the Hermitian product is linear only in the first variable, while conjugate linear in the second variable. So what we did in §9.1 is a special case of what we do here.

Theorem 9.1.1 has to be modified, because the Hermitian inner product is only conjugate linear in the second variable, so the map $V \to V^*$ is not complex linear, but only conjugate linear, since $\langle \mathbf{v}, c\mathbf{w} \rangle = \overline{c}\langle \mathbf{v}, \mathbf{w} \rangle$. So we only have

Theorem 9.2.1. *Let V be a complex inner product space, and $f(\mathbf{v})$ any linear functional on V. Then there is a unique $\mathbf{w} \in V$ such that $f(\mathbf{v}) = \langle \mathbf{v}, \mathbf{w} \rangle$. The map $\mathbf{w} \in V \mapsto f_{\mathbf{w}} \in V^*$, where $f_{\mathbf{w}}(\mathbf{v}) = \langle \mathbf{v}, \mathbf{w} \rangle$, satisfies $f_{c\mathbf{w}} = \overline{c}f_{\mathbf{w}}$.*

The analog of Theorem 9.1.2 holds:

Theorem 9.2.2. *For any linear operator L on the complex inner product space V there is a unique linear operator M on V, called the* adjoint *of L, such that*

$$\langle L\mathbf{v}, \mathbf{w}\rangle = \langle \mathbf{v}, M\mathbf{w}\rangle.$$

The adjoint is written L^* in the complex case.

The proof is the same, and therefore is left to you. It is worth noticing that L^* is complex linear and not conjugate linear. Indeed

$$\langle \mathbf{v}, L^*(c\mathbf{w})\rangle = \langle L\mathbf{v}, c\mathbf{w}\rangle = \overline{c}\langle L\mathbf{v}, \mathbf{w}\rangle = \overline{c}\langle \mathbf{v}, L^*\mathbf{w}\rangle = \langle \mathbf{v}, cL^*\mathbf{w}\rangle$$

because the Hermitian form is conjugate linear in the second variable, and we use that fact twice.

We say the operator L is self-adjoint, or *Hermitian* if $L = L^*$.

Theorem 9.1.4 remains true with the obvious change of notation:

Theorem 9.2.3. *Let L and M be linear operators on the complex inner product space V. Then the adjoint with respect to this inner product satisfies:*

$$\begin{aligned}(L+M)^* &= L^* + M^*;\\ (L\circ M)^* &= M^*\circ L^*;\\ (cL)^* &= \overline{c}L^*, \forall c \in \mathbb{C};\\ (L^*)^* &= L.\end{aligned}$$

Proof. Since the proof is very similar, it is left to you, as are the proofs of the analogs of Corollary 9.1.5 and Theorem 9.1.6. We just prove the third equality:

$$\langle (cL)\mathbf{v}, \mathbf{w}\rangle = c\langle L\mathbf{v}, \mathbf{w}\rangle = c\langle \mathbf{v}, L^*\mathbf{w}\rangle = \langle \mathbf{v}, \overline{c}L^*\mathbf{w}\rangle.$$

□

Next an easy but useful theorem that applies to all operators in an inner product space. It illustrates proof techniques for adjoints.

Theorem 9.2.4. *Let V be an inner product space, L any operator on V, Let W be a subspace of V invariant under L. Then $W^\perp$ is invariant under the adjoint L^*.*

Proof. The fact that W is invariant under L means that for all $\mathbf{w} \in W$, $L\mathbf{w} \in W$. Therefore $\langle L\mathbf{w}, \mathbf{u}\rangle = 0$ for all $\mathbf{u} \in W^\perp$. Then by definition of the adjoint

$$\langle L\mathbf{w}, \mathbf{u}\rangle = \langle \mathbf{w}, L^*\mathbf{u}\rangle = 0$$

for all $\mathbf{u} \in W^\perp$ and all $\mathbf{w} \in W$. This says precisely the $L^*\mathbf{u}$ is perpendicular to W, so $L^*\mathbf{u} \in W^\perp$. □

After this survey of results from §9.1 and above that hold for both $\mathbb{R}$ and $\mathbb{C}$, we prove two important results that only hold over $\mathbb{C}$.

Theorem 9.2.5. *Any operator L on a Hermitian inner product space is the zero operator if and only if $\langle L\mathbf{v},\mathbf{v}\rangle = 0$ for all $\mathbf{v} \in V$.*

Thus this improves Theorem 9.1.6.

Proof. If L is the zero operator, the implication is obvious. So assume $\langle L\mathbf{v},\mathbf{v}\rangle = 0$ for all $\mathbf{v} \in V$, and show that L is the zero operator. We use (9.1), which is still true in the complex case, from which we get

$$\langle L\mathbf{v},\mathbf{w}\rangle + \langle L\mathbf{w},\mathbf{v}\rangle = 0. \tag{9.2}$$

Because we do not assume that L is self-adjoint, we cannot go further without a new idea involving $\mathbb{C}$. In (9.2) replace $\mathbf{v}$ by $\mathrm{i}\mathbf{v}$. Then it becomes

$$\mathrm{i}\langle L\mathbf{v},\mathbf{w}\rangle - \mathrm{i}\langle L\mathbf{w},\mathbf{v}\rangle = \langle L(\mathrm{i}\mathbf{v}+\mathbf{w}),\mathrm{i}\mathbf{v}+\mathbf{w}\rangle + \langle L\mathbf{v},\mathbf{v}\rangle - \langle L\mathbf{w},\mathbf{w}\rangle. \tag{9.3}$$

Check this carefully, especially the signs. Since by hypothesis the right hand side is 0, we get

$$\mathrm{i}\langle L\mathbf{v},\mathbf{w}\rangle = \mathrm{i}\langle L\mathbf{w},\mathbf{v}\rangle \quad \text{or} \quad \langle L\mathbf{v},\mathbf{w}\rangle = \langle L\mathbf{w},\mathbf{v}\rangle.$$

Comparing this equation to (9.2) shows $\langle L\mathbf{v},\mathbf{w}\rangle = 0$ for all $\mathbf{v}$ and $\mathbf{w}$ in V, so by Theorem 9.1.6 we are done. □

We conclude with a useful result.

Theorem 9.2.6. *Let L be an operator on a Hermitian inner product space. Then L is Hermitian (self-adjoint) if and only if $\langle L\mathbf{v},\mathbf{v}\rangle$ is real for all $\mathbf{v} \in V$.*

Proof. First assume L is Hermitian. Then

$$\langle L\mathbf{v},\mathbf{v}\rangle = \langle \mathbf{v},L\mathbf{v}\rangle = \overline{\langle L\mathbf{v},\mathbf{v}\rangle}$$

which means that $\langle L\mathbf{v},\mathbf{v}\rangle$ is equal to its conjugate, therefore real.

Now we do the other implication, so assume $\langle L\mathbf{v},\mathbf{v}\rangle$ is real. Then

$$\langle L\mathbf{v},\mathbf{v}\rangle = \overline{\langle L\mathbf{v},\mathbf{v}\rangle} = \langle \mathbf{v},L\mathbf{v}\rangle = \langle L^*\mathbf{v},\mathbf{v}\rangle.$$

First we used the reality of $\langle L\mathbf{v},\mathbf{v}\rangle$, then the fact that the inner product is Hermitian, thus conjugate symmetric. Finally we used the definition of the adjoint.

Now by linearity in the first variable $\langle (L-L^*)\mathbf{v},\mathbf{v}\rangle = 0$ for all $\mathbf{v} \in V$. Theorem 9.2.5 then tells us that $L = L^*$, the desired conclusion. □

The analog of Theorem 9.1.9 is

Theorem 9.2.7. *Let V be a finite dimensional Hermitian space and $\{\mathbf{v}_1,\ldots,\mathbf{v}_n\}$ an orthonormal basis. Let L be any operator on V, and let A be the matrix of L in the given basis. Then the matrix of L^* is the conjugate transpose of A. Therefore if L is self adjoint, A is conjugate symmetric: $a_{ji} = \overline{a_{ij}}$.*

The proof is left to you.

A square matrix is Hermitian if it is equal to its conjugate transpose: $A^* = A$. Obviously if A is real, then to be Hermitian is just to be symmetric. For that reason most results on symmetric matrices are special cases of results on Hermitian matrices.

9.3 Positive Definite Operators and Matrices

In this section V denotes either a Euclidean space or a Hermitian space, with the inner product written $\langle \mathbf{v}, \mathbf{w} \rangle$ as usual. By the Gram-Schmidt orthonormalization process this guarantees that there exists an orthonormal basis $\mathfrak{B} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ for V. We can do this over $\mathbb{R}$ or $\mathbb{C}$.

A self-adjoint operator L is *positive semidefinite*

$$\langle L\mathbf{v}, \mathbf{v} \rangle \geq 0, \text{ for all } \mathbf{v} \in V. \tag{9.4}$$

Similarly a self-adjoint operator L is *positive definite* if

$$\langle L\mathbf{v}, \mathbf{v} \rangle > 0, \text{ for all } \mathbf{v} \neq \mathbf{0}. \tag{9.5}$$

In particular, positive definite operators are a subset of positive semidefinite operators. By Theorem 9.2.6 in the complex case $\langle L\mathbf{v}, \mathbf{v} \rangle$ is real.

By Theorems 9.1.9 and 9.2.7 the matrix for L in an orthonormal basis is symmetric in the real case, and Hermitian in the complex case. In both cases they are called positive definite (and positive semidefinite) matrices. We could make a parallel definition for negative definite and negative semidefinite operators and matrices: that is left to you as an exercise.

Positive definite matrices are among the most important in linear algebra, and after proving the Spectral Theorem in Chapter 13, we will develop many ways of testing when a matrix is positive definite. Here is a theorem we can prove with the tools at hand.

Theorem 9.3.1. *If L is positive definite, then it is invertible, and its inverse is positive definite.*

Proof. By hypothesis $\langle L\mathbf{v}, \mathbf{v} \rangle > 0$ when $\mathbf{v} \neq \mathbf{0}$. By the Cauchy-Schwarz inequality, $|\langle L\mathbf{v}, \mathbf{v} \rangle \leq \|L\mathbf{v}\| \|\mathbf{v}\|$, so $\|L\mathbf{v}\| \neq 0$, and L is injective. Therefore L is an isomorphism. Its inverse L^{-1} is also positive definite, as we now show. Since any $\mathbf{v} \in V$ can be written $L\mathbf{w}$, we have

$$\langle L^{-1}\mathbf{v}, \mathbf{v} \rangle = \langle L^{-1}L\mathbf{w}, L\mathbf{w} \rangle = \langle \mathbf{w}, L\mathbf{w} \rangle = \langle L\mathbf{w}, \mathbf{w} \rangle > 0.$$

□

Next we make a construction that will be useful later. We only write the details in the real case, and leave the complex case to you. Let A be a $m \times n$ matrix, and B

its transpose in the real case, or its conjugate transpose in the complex case. Then we consider the square matrix $G = BA$ of size n. Thus in the real case $g_{ij} = \langle \mathbf{a}_i, \mathbf{a}_j \rangle$. G is called the Gram matrix of A.

Theorem 9.3.2. *The rank of G equals the rank of A.*

Proof. Here is the proof in the real case: for the complex case just replace the transpose by the conjugate transpose. Let $\mathbf{v}$ be a vector in the nullspace of G, so that $A^tA\mathbf{v} = \mathbf{0}$. Obviously then $\mathbf{v}^tA^tA\mathbf{v} = 0$, so $(A\mathbf{v})^tA\mathbf{v} = 0$. Because the product on the vector space V is positive definite, this means $A\mathbf{v} = \mathbf{0}$, so $\mathbf{v}$ is in the nullspace of A. The other implication is obvious. Now we just use the Rank-Nullity Theorem applied to the linear maps L_A and L_G. Because they have the same nullity and the source space has dimension n in both cases, they have the same rank. □

Corollary 9.3.3. *The Gram matrix G of any matrix A is positive semidefinite. Furthermore G is invertible if and only if A has rank n. G is invertible if and only if it is positive definite.*

Proof. First G is clearly symmetric. We need to show that for any $\mathbf{v} \in V$, $\mathbf{v}^tG\mathbf{v} \geq 0$. Replacing G by A^tA as in the previous theorem, we see that we need

$$\mathbf{v}^tA^tA\mathbf{v} = (A\mathbf{v})^tA\mathbf{v} \geq 0$$

which is clear because the inner product on V is positive definite.The last point is left to you. □

Remark 9.3.4. If $m < n$, obviously A can have at most rank m, and therefore G has at most rank m. As we will see, the Gram construction is mainly interesting when $m \geq n$.

9.4 Orthogonal Operators

Let V be a Euclidean space of dimension n. Recall that in Definition 8.3.6 we defined orthogonal matrices.

Definition 9.4.1. An operator $L\colon V \to V$ is *orthogonal* for the positive definite inner product on V if

$$\langle L\mathbf{v}, L\mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle \text{ for all } \mathbf{v}, \mathbf{w} \in V. \tag{9.6}$$

As usual, we write $\|\mathbf{v}\|$ for $\sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$. Then

Theorem 9.4.2. *The operator L is orthogonal if and only if* $\|L\mathbf{v}\| = \|\mathbf{v}\|$ *for all* $\mathbf{v} \in V$.

Thus L is orthogonal if and only if it preserves distance.

Proof. If L is orthogonal then it is obvious that $\|L\mathbf{v}\| = \|\mathbf{v}\|$. The reverse implication is more interesting. We prove it by a polarization argument already used when studying bilinear forms. By linearity we have

$$\langle \mathbf{v}+\mathbf{w}, \mathbf{v}+\mathbf{w}\rangle - \langle \mathbf{v}-\mathbf{w}, \mathbf{v}-\mathbf{w}\rangle = 4\langle \mathbf{v}, \mathbf{w}\rangle$$
$$\langle L(\mathbf{v}+\mathbf{w}), L(\mathbf{v}+\mathbf{w})\rangle - \langle L(\mathbf{v}-\mathbf{w}), L(\mathbf{v}-\mathbf{w})\rangle = 4\langle L\mathbf{v}, L\mathbf{w}\rangle$$

Rewriting in terms of the norm, we get

$$\|\mathbf{v}+\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2 = 4\langle \mathbf{v}, \mathbf{w}\rangle$$
$$\|L(\mathbf{v}+\mathbf{w})\|^2 - \|L(\mathbf{v}-\mathbf{w})\|^2 = 4\langle L\mathbf{v}, L\mathbf{w}\rangle$$

Since the left hand sides are equal by hypothesis, the right hand sides are too, and this is what we needed to prove. □

In the theorem we only need to know that $\|L\mathbf{v}\| = \|\mathbf{v}\|$ for vectors $\mathbf{v}$ of length 1, since it is then obviously true for $r\mathbf{v}$, for any $r \in \mathbb{R}$.

Note that orthogonal operators preserve angles, in particular perpendicularity. The last point is obvious since for an orthogonal L, since by definition $\langle \mathbf{v}, \mathbf{w}\rangle = 0$ if an only if $\langle L\mathbf{v}, L\mathbf{w}\rangle = 0$.

Exercise 9.4.3. Write down an explicit linear operator on $\mathbb{R}^2$ that preserves perpendicularity, but not distance.

Exercise 9.4.4. What does it mean to say that L preserves angles? As usual we define the angle between vectors $\mathbf{v}$ and $\mathbf{w}$ to be the angle θ whose cosine is

$$\frac{\langle \mathbf{v}, \mathbf{w}\rangle}{\|\mathbf{v}\|\|\mathbf{w}\|}.$$

By the Cauchy-Schwarz inequality (8.6), this number is between -1 and 1, so it is the cosine of a well-defined angle θ. So a linear operator L preserves angles if the angle between $L\mathbf{v}$ and $L\mathbf{w}$ is the same as that between $\mathbf{v}$ and $\mathbf{w}$ for all $\mathbf{v}$, $\mathbf{w}$ in V. What is L?

Theorem 9.4.5. *An operator $L\colon V \to V$ is orthogonal if and only if it is invertible and $L \circ L^t = I$, so that its adjoint is its transpose.*

Proof. Indeed, by definition of the adjoint $\langle L\mathbf{v}, L\mathbf{w}\rangle = \langle \mathbf{v}, L^t \circ L\mathbf{w}\rangle$. On the other hand, by definition of an orthogonal operator $\langle L\mathbf{v}, L\mathbf{w}\rangle = \langle \mathbf{v}, \mathbf{w}\rangle$ for all $\mathbf{w}$.

So $\langle \mathbf{v}, L^t \circ L\mathbf{w}\rangle = \langle \mathbf{v}, \mathbf{w}\rangle$ for all $\mathbf{v}$ and all $\mathbf{w}$. This implies $L^t \circ L(\mathbf{w}) = \mathbf{w}$ and therefore $L^t \circ L = I$. In particular L is invertible and L^t is its inverse. The converse is immediate. □

Theorem 9.4.6. *Let V be a Euclidean space and $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ an orthonormal basis. Let L be an operator on V, and let A be the matrix of L in the given basis. L is orthogonal if and only if $A^{-1} = A^t$.*

Therefore by Definition 8.3.6 A is an orthogonal matrix. Orthogonal matrices are especially pleasant to deal with because their inverses can be computed trivially. As we have already noted:

Theorem 9.4.7. *In an orthogonal matrix A, we have $\langle \mathbf{a}^i, \mathbf{a}_j \rangle = \delta_{ij}$, where $\mathbf{a}^i$ is the i-th row of A, $\mathbf{a}_j$ the j-th column, and δ_{ij} is the Kronecker delta.*

Example 9.4.8. Let P be a $n \times n$ matrix with exactly one 1 in each row and each column, and 0 everywhere else. Then P is orthogonal.

Explicit examples here.

We will study permutation matrices in §11.2.

Example 9.4.9. Rotations and reflections in $\mathbb{R}^2$. Refer to §13.7. Similarly for $\mathbb{R}^3$.

Similarity with orthogonal matrices = conjugation.

9.5 Unitary Operators

Now let V be a Hermitian space of dimension n. We develop the complex analog of orthogonal operators. You should review the notation at the beginning of §8.7 before reading on.

Definition 9.5.1. An operator $L\colon V \to V$ is *unitary* for the Hermitian product on V if

$$\langle L\mathbf{v}, L\mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle \text{ for all } \mathbf{v}, \mathbf{w} \in V. \tag{9.7}$$

Theorem 9.5.2. *The operator $L\colon V \to V$ is unitary if and only if $\|L\mathbf{v}\| = \|\mathbf{v}\|$ for all $\mathbf{v} \in V$.*

The proof is identical to that of Theorem 9.4.2 in the real case. In the theorem we only need to know that $\|L\mathbf{v}\| = \|\mathbf{v}\|$ for vectors $\mathbf{v}$ of length 1, since it is then obviously true for $c\mathbf{v}$, for any $c \in \mathbb{C}$.

Theorem 9.5.3. *An operator $L\colon V \to V$ is unitary if and only if it is invertible and its inverse is its adjoint: $L \circ L^* = L^* \circ L = I$*

Theorem 9.5.4. *Let V be a complex inner product space and $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ an orthonormal basis. Let L be an operator on V, and let A be the matrix of L in the given basis. Then L is unitary if and only if $A^{-1} = A^*$.*

In Definition 8.7.3 we learned that a square matrix A is unitary if $A^{-1} = A^*$. Therefore a square matrix A is unitary if it is invertible and its inverse is its conjugate transpose. So $A^*A = I = AA^*$. Unitary matrices are as pleasant as orthogonal matrices. As we have already noted:

Theorem 9.5.5. *In a unitary matrix A, we have $\langle \mathbf{a}^i, (\mathbf{a}_j)^* \rangle = \delta_{ij}$, where $\mathbf{a}^i$ is the i-th row of A, $\mathbf{a}_j$ the j-th column, and δ_{ij} is the Kronecker delta.*

Let's recall here Theorem 8.7.4, which say that any complex invertible matrix can be written as the product of a unitary matrix Q and an upper triangular matrix R:

$$A = QR.$$

Similarity with unitary matrices = conjugation.

9.6 Normal Operators

A linear operator L is normal if L commutes with its conjugate transpose L^*:

$$LL^* = L^*L.$$

Passing to the matrix associated to L in an orthonormal basis, we say that a square matrix A is normal if A commutes with its conjugate transpose A^*: $AA^* = A^*A$. If A is a real matrix this means that A commutes with its transpose.

Since a matrix always commutes with itself and with its inverse (by definition) we see that normal matrix encompass the classes of symmetric matrices, skew symmetric matrices, Hermitian matrices, skew Hermitian matrices, orthogonal matrices and unitary matrices. Scalar multiples of such matrices are obviously also normal.

Exercise 9.6.1. Verify the skew-symmetric case and the skew-Hermitian cases by direct computation of the matrix products.

In fact it is worth asking if there are any matrices other than those in the classes mentioned above that commute with their adjoint.

Exercise 9.6.2. In the 2×2 real case show all normal matrices are of the form described above.

Exercise 9.6.3. Do the same thing for 2×2 complex matrices.

Remark 9.6.4. Once we have studied polynomials in operators and square matrices in the next chapter, we will see that any operator L commutes with any polynomial in L, and if it is invertible, and polynomial in L^{-1}. So if the adjoint of L can be expressed in the way, as in the examples above, L is normal.

Here is an example of how to normality is used..

Theorem 9.6.5. *If L is a normal operator with adjoint L^*, then for any vector* $\mathbf{v}$*,*

$$\|L\mathbf{v}\|^2 = \langle L\mathbf{v}, L\mathbf{v}\rangle = \langle L^*\mathbf{v}, L^*\mathbf{v}\rangle = \|L^*\mathbf{v}\|^2.$$

Proof. The two outside equalities are just the definition of the length of a vector. For the rest

$$\begin{aligned}\langle L\mathbf{v}, L\mathbf{v}\rangle &= \langle \mathbf{v}, L^*L\mathbf{v}\rangle && \text{by definition of the adjoint}\\ &= \langle \mathbf{v}, LL^*\mathbf{v}\rangle && \text{by normality of } L\\ &= \langle L^*\mathbf{v}, L^*\mathbf{v}\rangle && \text{by definition of the adjoint again.}\end{aligned}$$

□

Exercise 9.6.6. Show directly that this result is true for self-adjoint, orthogonal and unitary matrices.

Theorem 9.6.7. *Let L be a normal operator on an inner product space V. Then the nullspace and the range of L are mutually perpendicular. In particular their intersection is* (0).

Proof. Assume a vector $\mathbf{v} \in V$ is orthogonal to the range of L, namely $\langle \mathbf{v}, L\mathbf{w}\rangle = 0$ for all $\mathbf{w} \in V$. So $\langle L^*\mathbf{v}, \mathbf{w}\rangle = 0$ for all $\mathbf{w} \in V$. This means that $L^*\mathbf{v}$ is the zero vector. By Theorem 9.6.5, $L\mathbf{v} = 0$, so $\mathbf{v}$ is in the nullspace of L. This shows that all the vectors orthogonal to the range of L are in the nullspace of L. By the rank-nullity theorem this must be the full nullspace, so we are done. □

There are a number of other interesting theorems that we can prove for the large class of normal matrices: all of them require understanding eigenvalues and eigenvectors so they will have to be postponed to Chapter 13, devoted to the Spectral Theorem.

9.7 The Four Subspaces Associated to a Matrix

This section says more about the four key spaces[1] associated to a $m \times n$ matrix A using standard inner product on $\mathbb{R}^n$ and $\mathbb{R}^m$ or the standard Hermitian product on $\mathbb{C}^n$ and $\mathbb{C}^m$. In both cases we write it $\langle \mathbf{v}_1, \mathbf{v}_2\rangle$.

The four subspaces are the range $R(A)$ and nullspace $N(A)$ of A, and the nullspace $N(A^t)$ and the range $R(A^t)$ of A^t. We summarize what we have obtained so far, from the Rank-Nullity Theorem and the fact that row rank equals column rank:

Theorem 9.7.1.

$$\begin{aligned}\dim N(A) + \dim R(A) &= n;\\ \dim N(A^t) + \dim R(A^t) &= m;\\ \dim R(A) &= \dim R(A^t).\end{aligned}$$

Obviously the rank of A (and A^t) is at most the smaller of m and n.

We have already covered some of the material in this section using duality in §6.6. In particular, if you have studied duality, compare with Theorem 6.6.3.

[1] An excellent reference for this material is Strang [28], whose entire presentation in §2.4 and §3.1 is organized around this approach (especially 3C on p. 136).

Definition 9.7.2. Two subspaces V_1 and V_2 of F^n are *mutually orthogonal* if for any $\mathbf{v}_1 \in V_1$ and any $\mathbf{v}_2 \in V_2$, $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = 0$. We write $V_1 \perp V_2$ to indicate that the spaces are mutually orthogonal. V_1 and V_2 have *complementary dimensions* in F^n if $\dim V_1 + \dim V_2 = n$. By Theorem 8.1.16 and the analogous theorem in the Hermitian case we know that in an inner product space, the orthogonal complement $U^\perp$ of any subspace U has complementary dimension, and furthermore $F^n = U \oplus U^\perp$.

Theorem 9.7.1 shows that $N(A)$ and $R(A^t)$ have complementary dimensions in F^n, and $N(A^t)$ and $R(A)$ have complementary dimensions in F^m.

Theorem 9.7.3 (The Four Subspaces Theorem).

- *$N(A) \perp R(A^t)$ in the domain F^n of A. In particular any element of F^n can be written uniquely as the sum of a vector of $N(A)$ and a vector of $R(A^t)$.*
- *$N(A^t) \perp R(A)$ in the domain F^m of A^t. In particular any vector of F^m can be written uniquely as the sum of a vector of $N(A^t)$ and a vector of $R(A)$.*

Proof. Take an element $\mathbf{x}_0$ in the nullspace of A, so $A\mathbf{x}_0 = 0$, and an element $\mathbf{x}_1$ in the range of B, so there exists $\mathbf{y}$ such that $\mathbf{x}_1 = B\mathbf{y}$. We compute the inner product

$$\langle \mathbf{x}_1, \mathbf{x}_0 \rangle = \mathbf{x}_1^t \mathbf{x}_0 = (\mathbf{y}^t B^t)\mathbf{x}_0 = (\mathbf{y}^t A)\mathbf{x}_0 = \mathbf{y}^t (A\mathbf{x}_0) = 0$$

so that they are orthogonal. It is enough to prove the first statement, since $B^t = A$. The first and the third equations of Theorem 9.7.1 show that a basis for F^n can be formed by adjoining a basis of $R(B)$ to a basis of $N(A)$, if we can show that the basis elements of the two spaces are linearly independent. This is what we just established, so we are done. □

We get Theorem 2.7.2 as an easy corollary of this result. Here is a proof that uses the positive definite inner product over $\mathbb{R}$.

Corollary 9.7.4. *The equation $A\mathbf{x} = \mathbf{b}$ has a solution if and only if there is no vector $\mathbf{y}$ in $\mathbb{R}^m$ with*

$$\mathbf{y}^t A = \mathbf{0} \qquad \text{and} \qquad \mathbf{y}^t \mathbf{b} \neq 0.$$

Proof. If $A\mathbf{x} = \mathbf{b}$ has a solution, then $\mathbf{b} \in R(A)$. If $\mathbf{y}^t A = 0$, then $\mathbf{y} \in N(B)$. According to the Four Subspaces Theorem 9.7.3, $N(B) \perp R(A)$. This is contradicted by the condition that there exists a $\mathbf{y}$ with $\mathbf{y}^t \mathbf{b} = \langle \mathbf{y}, \mathbf{b} \rangle \neq 0$ so the existence of both $\mathbf{x}$ and $\mathbf{y}$ satisfying the hypotheses of the theorem is impossible. On the other hand, one always exists. If $\mathbf{b} \in R(A)$, then we are in case 1. If not, then $\mathbf{b}$ can be written uniquely as $\mathbf{b}' + \mathbf{y}$, with $\mathbf{b}' \in R(A)$ and $\mathbf{y}$ a non-zero element in $N(B)$. Furthermore $\mathbf{b}' \perp \mathbf{y}$ by the Four Subspaces Theorem. Then

$$\mathbf{y}^t \mathbf{b} = \mathbf{y}^t (\mathbf{b}' + \mathbf{y}) = \mathbf{y}^t \mathbf{y} \neq 0,$$

so we have the desired $\mathbf{y}$. We used positive definiteness to get the very last step. □

Remark 9.7.5. By replacing $\mathbf{y}$ by a scalar multiple, we can assume that $\mathbf{y}^t \mathbf{b} = 1$ in the statement of Corollary 9.7.4.

We state the next result for $\mathbb{R}$ only, for convenience.

Proposition 9.7.6. *In terms of the inner product $\langle *,*\rangle_1$ on $\mathbb{R}^n$, and the inner product $\langle *,*\rangle_2$ on $\mathbb{R}^m$, we have*

$$\langle \mathbf{y}, A\mathbf{x}\rangle_2 = \langle A^t\mathbf{y}, \mathbf{x}\rangle_1 \tag{9.8}$$

Proof.

$$\begin{aligned}
\langle \mathbf{y}, A\mathbf{x}\rangle_2 &= \mathbf{y}^t A\mathbf{x} && \text{(switching to matrix multiplication in } F^m\text{)},\\
&= (A^t\mathbf{y})^t\mathbf{x} && \text{(since transposition reverses order)},\\
&= \langle \mathbf{A}^t\mathbf{y}, \mathbf{x}\rangle_1 && \text{(switching back to dot product in } F^n\text{)},
\end{aligned}$$

so we are done. □

When $m = n$, we recover the result concerning the matrix of self-adjoint linear transformations in the standard basis.

Example 9.7.7. Assume $m < n$, and let A be a $m \times n$ matrix of rank m, the maximum possible. Then the nullspace of A has dimension $n - m$, and the nullspace of B has dimension 0.

Chapter 10
The Minimal Polynomial

Abstract After a general discussion of our goal, this chapter applies the results on polynomials given in Appendix C to linear operators, therefore to square matrices. The main idea is that it is possible to substitute a linear operator or square matrix for the variable of a polynomial. This allows us to define the minimal polynomial of a linear operator, and compute it in some cases. Then we prove the Primary Decomposition Theorem which explains how to decompose a vector space as a direct sum of subspaces, each invariant under the linear operator: these subspaces corresponding to irreducible factors of the minimum polynomial, which we compute in a few special cases. Finally we derive the Jordan Canonical Form for any linear operator over the complex numbers. Indeed we do this for any operator whose minimal polynomial factors as a product of linear polynomials. We show that the Jordan Canonical Form is unique and then show how to find a canonical form for linear operators over the real numbers.

10.1 Linear Operators: the Problem

Before starting out, note that we are going back to vector spaces without inner products. Here are two questions we want to answer.

1. Given a linear operator L , what is the best basis to put on V to make L appear as simple as possible?
2. What is a complete list of all linear operators on V?

For the first question, if A is the matrix of the linear transformation in one basis, and B in a second basis, then there is an invertible change of basis matrix C such that $B = C^{-1}AC$ as we saw in Theorem 5.5.4. Because every invertible matrix is a product of elementary matrices, the direct approach to solving this problem would be to find suitable elementary matrices E so that EAE^{-1} is simpler than A. Check with the three kinds of elementary matrices in Definition 2.8.1 and their inverses in Theorem 2.8.4 to see how difficult it would be to make this direct method work.

The second question requires some additional tools to answer. We develop these tools in this chapter and the next. In this chapter we use tools that work for any base field F. In the next chapter we look at tools that work best for $\mathbb{C}$, and sometimes do not work for $\mathbb{R}$ and $\mathbb{Q}$,

First let's look at some examples.

Definition 10.1.1. A linear operator $L\colon V \to V$ is *diagonalizable* if there is a basis for V such that for every element $\mathbf{v}$ of the basis, $\mathbf{v} = \lambda\mathbf{v}$, where λ is a scalar. This means that the matrix for L in that basis is diagonal. If we start with a matrix A for the linear operator L in another basis, there is a change of basis matrix C so that $C^{-1}AC$ is diagonal.

This is certainly the most important (and desirable) kind of linear operator, because it says that we can choose the coordinates of V so that each variable in the output of L only depends on one variable in the input. We say that we have decoupled the variables. This is the result we achieved for symmetric bilinear forms in Chapter 7. Unfortunately, as we will see, not all linear operators are diagonalizable.

Definition 10.1.2. A subspace W of V is *invariant* for L if for all $\mathbf{w} \in W$, $L\mathbf{w}$ is in W.

If the invariant subspace W is one-dimensional, so a basis consists of one element, say $\mathbf{w}$, then $L\mathbf{w} = \lambda\mathbf{w}$ for some λ. We have special names for this case.

Definition 10.1.3. If $\mathbf{w}$ is a non-zero vector, and $L\mathbf{w} = \lambda\mathbf{w}$, then $\mathbf{w}$ is an *eigenvector* of L and the scalar λ is its associated *eigenvalue*.

It is traditional to write eigenvalues with lower case Greek letters. Note carefully that eigenvectors must be non-zero by definition.

If W is an invariant subspace, then L restricts to a linear operator L_W on W. The requirement that W be non-trivial just excludes the cases $W = (0)$ and $W = V$, where the result is always true. Then the dimension of W is less than that of V, so it is usually easier to study L_W than L. Let r be the dimension of W. We can build a basis for V such that the first r basis elements $\mathbf{w}_1, \ldots, \mathbf{w}_r$ form a basis for W, and the last $n-r$ basis elements $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_n$ are chosen arbitrarily. Then the matrix A for L is this basis can be written in block form as

$$\begin{pmatrix} A_W & C \\ 0 & B \end{pmatrix}$$

where A_W is the $r \times r$ matrix of L_W. The size of the zero matrix 0 is $(n-r) \times r$, and C is $r \times (n-r)$. The matrix B is square of size $n-r$. The ideal situation occurs when we can choose C to be the 0 matrix. The the subspace U spanned by the vectors $\mathbf{u}_{r+1}, \ldots \mathbf{u}_n$ is invariant under L. So we can write the restriction of L to U as L_U, which has matrix A_U on U, so

$$A = \begin{pmatrix} A_W & 0 \\ 0 & A_U \end{pmatrix}$$

In this case V as the direct sum $W \oplus U$ of two subspaces invariant under L, so that if we understand L_W and L_U, we understand L. This leads us to the following definition.

Definition 10.1.4. A operator L can be block-diagonalized if V is the direct sum of subspaces V_i that are each invariant under L.

If each block has dimension 1 so that there are $n = \dim V$ blocks, the L is diagonalizable.

There is one special case in the classification of linear operators that we can handle immediately.

Theorem 10.1.5. *Assume V has dimension n, and that $L\colon V \to V$ has k eigenvectors $\mathbf{v}_i$ whose eigenvalues λ_i, $1 \le i \le k$, are all distinct. Then the $\mathbf{v}_i$ are linearly independent. In particular, if $k = n$ they form a basis for V, and in that basis the matrix of L is diagonal with the elements $\lambda_1, \ldots, \lambda_n$ down the diagonal.*

Proof. Assume by contradiction that the $\mathbf{v}_i$ are not linearly independent, so that there is an equation of linear dependence

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_k\mathbf{v}_k = \mathbf{0}, \tag{10.1}$$

where at least one $a_i \neq 0$. Apply the operator L and use the fact that the $\mathbf{v}_i$ are eigenvectors with eigenvalue λ_i:

$$a_1L\mathbf{v}_1 + a_2L\mathbf{v}_2 + \cdots + a_kL\mathbf{v}_k = a_1\lambda_1\mathbf{v}_1 + a_2\lambda_2\mathbf{v}_2 + \cdots + a_k\lambda_k\mathbf{v}_k = 0 \tag{10.2}$$

By induction we show that this cannot happen. We start the induction at $k = 1$. Because by definition all eigenvectors are non-zero, then $a_1\mathbf{v}_1 = \mathbf{0}$ implies $a_1 = 0$, which is not an equation of linear dependence. So we assume the result is true for $k-1$, and must prove it is true for k. Note that induction implies that all the coefficients a_i in (10.1) must be non-zero. Because the λ_i are distinct, since $n \ge 2$ this implies that at least one of them is non-zero: we may assume it is λ_1. Then subtract (10.2) from λ_1 times (10.2) to get an equation of linear dependence involving only $k-1$ vectors: thus all the coefficients a_i must be 0 for $i \ge 2$. But then $a_1 = 0$ so we do not have an equation of linear dependence, and we are done.

The statement for $k = n$ follows trivially. □

We have the following pretty variation on Theorem 10.1.5.

Theorem 10.1.6. *The $n \times n$ matrix A is similar to a diagonal matrix if and only if A has n linearly independent eigenvectors.*

Proof. If A is similar to a diagonal matrix, then $C^{-1}AC = D$, where D is a diagonal matrix with diagonal entries $d_1, \ldots, d_n$. Multiply this equation by C to get $AC = CD$. We now equate the i-th column on both sides. On the right side, recalling that the i-th column of C is written $\mathbf{c}_i$, the i-th column of CD is $d_i\mathbf{c}_i$. On the left side, the i-th column of AC is $A\mathbf{c}_i$. See Proposition 2.2.7. So $A\mathbf{c}_i = d_i\mathbf{c}_i$, which says that $\mathbf{c}_i$ is an eigenvector for A with eigenvalue d_i.

Now assume A has n linearly independent eigenvectors. We could just invoke the end of the proof of Theorem 10.1.5 to conclude, but instead we will use an argument that is parallel to the first part. Write the eigenvectors as $\mathbf{c}_1, \ldots, \mathbf{c}_n$. Write the eigenvalue of $\mathbf{c}_i$ as d_i, and let D be the diagonal matrix with diagonal entries the d_i. Let C be the matrix whose columns are the $\mathbf{c}_i$. This matrix is invertible because its columns are linearly independent by hypothesis. We can just reverse the argument of the first part to show that $AC = CD$. Multiply on the left by C^{-1} to get $C^{-1}AC = D$, so A is diagonalizable. □

Definition 10.1.7. A linear operator is *triangulable* if there is a basis of V so that the matrix of L is this basis is triangular, either upper triangular or lower triangular.

Proposition 10.1.8. *If a linear operator is triangulable, it has an eigenvector.*

Proof. Assume that the $n \times n$ matrix A of the linear transformation is lower triangular. Then the last column $\mathbf{a}_n$ of A has all entries 0 except for the last entry, which is a_{nn}. An easy computation shows that $A\mathbf{a}_n = a_{nn}\mathbf{a}_n$, so we are done. □

We will show later that every operator over $\mathbb{C}$ is triangulable..

Finally consider the matrix of a rotation of $\mathbb{R}^2$. A rotation other than one of 0 radians or π radians, clearly has no eigenvectors, so is not triangulable.

In the next few sections we study polynomials in order to develop the tools we need. We will use these tools to study linear operators at the end of the chapter.

10.2 Polynomials of Matrices

A key use of polynomials in linear algebra is to take a polynomial $f(x)$ and to substitute for x either a linear transformation or a square matrix of some size n with coefficients in the same field F as the coefficients of the polynomial. Notice that we have started this for linear operators in §4.6 and 4.5. We now study this for matrices using the results of Appendix C.

So the first thing to do is to make sure that given a polynomial

$$f(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0$$

with coefficients in F, it makes sense to write down the expression

$$f(A) = a_m A^m + a_{m-1}A^{m-1} + \cdots + a_1 A + a_0 I$$

where I is the identity matrix of size n. Because we can add matrices and multiply them by scalars, this does make sense. Because the only matrices that will appear are powers of A and I, they all commute.

Next we need to show that the operations on polynomials get reflected in the operations on matrices. More precisely

Theorem 10.2.1. *Let $f(x)$ and $g(x)$ be two polynomials in $F[x]$, written as in* (C.1) *and* (C.2)*. Let A be any square matrix with entries in F. Then*

$$\begin{aligned}(cf)(A) &= c(f(A), \quad \forall c \in F\\ (f+g)(A) &= f(A)+g(A)\\ (fg)(A) &= f(A)g(A)\end{aligned}$$

Thus on the left hand side of this three equations, the operation (scalar multiplication, addition and multiplication) is performed on polynomials, while on the right hand side the operation is performed on matrices. This is why we need square matrices: we need to be able to multiply two of them and get a matrix of the same size.

Proof. • For all $c \in K$, $(cf)(A) = ca_nA^n + \cdots + ca_1A + ca_0 = cf(A)$.

- For addition, we may assume by symmetry that $\deg f \geq \deg g$, and set all the coefficients b_k, for $m+1 \leq k \leq n$, to 0, for convenience in writing the formula. Then

$$\begin{aligned}(f+g)(A) &= (a_n+b_n)A^n + \cdots + (a_1+b_1)A + (a_0+b_0)I\\ &= a_nA^n + \cdots + a_1A + a_0I + b_nA^n + \cdots + b_1A + b_0I\\ &= f(A)+g(A).\end{aligned}$$

- For multiplication, using the notation of (C.3), we get

$$(fg)(A) = c_{n+m}A^{n+m} + \cdots + c_1A + c_0I. \tag{10.3}$$

On the other hand,

$$\begin{aligned}f(A) &= a_nA^n + \cdots + a_1A + a_0I\\ g(A) &= b_nA^m + \cdots + b_1A + b_0I\end{aligned}$$

Multiplying them together, we get (10.3) by collecting all the terms that have the same power in A. For example the leading term of the product is

$$a_nA^nb_mA^m = a_nb_mA^nA^m = c_{n+m}A^{n+m}.$$

Remark 10.2.2. Instead of doing this for a matrix A, we could do it instead for a linear transformation $L\colon V \to V$. We replace the product A^k by the composition of the linear map k times. The statement of Theorem 10.2.1 still makes sense and the proof follows the same lines as that for matrices. The advantage of using linear transformations is that the result is then independent of the choice of basis.

We will often look at equations of the form

$$f(A) = a_mA^m + a_{m-1}A^{m-1} + \cdots + a_1A + a_0I = 0 \tag{10.4}$$

where the 0 is the $n \times n$ zero matrix. All the terms on the left side are also $n \times n$ matrices, so the equations says that the sum of all particular entries are zero.

Definition 10.2.3. A polynomial $f(x)$ vanishes on the matrix A if $f(A) = 0$. The right hand side is the zero $n \times n$ matrix.

We are primarily interested in linear transformations, and only use matrices as a computational tool. We should verify that if (10.4) is true for a matrix A, it is true for any matrix $C^{-1}AC$ that is similar to A, since it represents the same linear transformation expressed in a different basis. This means that we can talk about a polynomial $f(x)$ vanishing on a linear transformation.

Theorem 10.2.4. *For any invertible matrix C, and any polynomial $f(x)$,*

$$C^{-1}f(A)C = f(C^{-1}AC).$$

Proof. Write the polynomial as in (10.4). Then

$$C^{-1}f(A)C = a_m C^{-1}A^m C + a_{m-1}C^{-1}A^{m-1}C + \cdots + a_1 C^{-1}AC + a_0 C^{-1}IC$$

and for each k

$$C^{-1}A^k C = C^{-1}A(CC^{-1})A(CC^{-1})\ldots(CC^{-1})AC = \left(C^{-1}AC\right)^k$$

so we are done. □

Thus if an equation (10.4) is true for a matrix A, it is true for any matrix similar to it, as promised.

10.3 The Minimal Polynomial

We fix a square matrix A of size n. First we show that there always exists a polynomial $f(x)$ that vanishes on A.

Theorem 10.3.1. *There is a polynomial $f(x)$ of degree m less than or equal to n^2 such that $f(A) = 0$.*

Proof. The vector space of square matrices of size n has dimension n^2, as we have seen. We are trying to find a certain polynomial, so we let its coefficients be the unknowns y_i, $0 \leq i \leq m$ of our problem, for some m to be determined. Write

$$f(x) = y_m x^m + y_{m-1}x^{m-1} + \cdots + y_1 x + y_0$$

so we are trying to solve the linear system in the y_i given by

$$f(A) = y_m A^m + y_{m-1}A^{m-1} + \cdots + y_1 A + y_0 I = 0.$$

The coefficients of A and all its powers are known to us. Write $A^k = (a_{ij}^{(k)})$ so we have a name for the entries of all the powers, including $I = A^0$. So for example

$$a_{ij}^{(0)} = \begin{cases} 1 & \text{if } i = j; \\ 0 & \text{otherwise.} \end{cases} \qquad \text{and } a_{ij}^{(1)} = a_{ij}.$$

So we have a homogeneous system of n^2 linear equations in $m+1$ variables, one for each entry (i, j):

$$y_0 a_{ij}^{(0)} + y_1 a_{ij}^{(1)} + \cdots + y_m a_{ij}^{(m)} = 0. \tag{10.5}$$

We want a non-zero solution. We can only guarantee one if the number of variables is strictly greater than the number of equations, so $m > n^2$. □

Of course they could be a polynomial of smaller degree vanishing on A. In fact there always is one of degree at most n, as we will see using the characteristic polynomial in Chapter 12. In this chapter we will prove this for many matrices A.

Definition 10.3.2. Call the smallest degree of a non-trivial polynomial vanishing on A the *minimal degree*.

Theorem 10.3.1 guarantees that there is a minimal degree. Then

Theorem 10.3.3. *All polynomials of minimal degree vanishing on A differ by a scalar multiple. The one that is monic (i.e. has leading coefficient 1) is called the minimal polynomial of A. Any other polynomial $g(x)$ vanishing on A is divisible by the minimal polynomial.*

Proof. Both parts of the theorem can be proved at once. Let $f(x)$ be a monic polynomial of minimal degree vanishing on A, and let $g(x)$ be any other polynomial vanishing on A. Then do long division: $g(x) = q(x)f(x) + r(x)$, with $\deg r(x) < \deg f(x)$. By hypothesis $g(A) = 0$ and $f(A) = 0$, so $q(A)f(A) = 0$ too. Thus $r(A) = 0$. This forces $r(A)$ to be the zero matrix, since its degree is otherwise too small. This proves the theorem. □

Theorem 10.3.4. *All similar matrices have the same minimal polynomial. Therefore we can talk about the minimal polynomial of a linear transformation.*

Proof. This is an immediate corollary of Theorem 10.2.4. □

Example 10.3.5. Suppose that the diagonal $n \times n$ matrix A has k distinct diagonal elements $a_1, \ldots, a_k$. Then the minimal polynomial of A is $\prod_{i=1}^{k}(x - a_i)$.

Proof. It is obvious that the minimal polynomial divides $\prod_{i=1}^{k}(x - a_i)$, since this polynomial vanishes on the space. So assume one of the factors $x - a_i$ is missing. On the other hand, if one of the diagonal elements does not appear in the product, then the simple fact that a product of diagonal matrices is just the diagonal matrix that is the product of the diagonal entries shows that the product does not vanish on the matrix. □

In one case it is easy to get an upper good bound on the minimal degree. The result follows from the Cayley Hamilton theorem proved in §12.4, but it is may be interesting to see a direct proof. We will soon prove in Corollary 10.6.1 that all linear operators over $\mathbb{C}$ are triangulable, so the result below holds for any linear operators over $\mathbb{C}$. In particular it gives an alternate proof of the Cayley-Hamilton Theorem.

Theorem 10.3.6. *Assume L is a linear operator on V that is triangulable, so that there is a basis of V in which the matrix A of L is upper-triangular. The diagonal elements of A are written $a_{11}, \ldots, a_{nn}$ as usual. Then the polynomial*

$$\prod_{i=1}^{n}(x-a_{ii})$$

vanishes on A, so the minimal polynomial divides this polynomial of degree n.

Proof. We need to show that $\prod_{i=1}^{n}(A-a_{ii}I)=0$. We prove this by induction on n. We start the induction at $n=1$, in which case $A=(a_{11})$, the polynomial is $(x-a_{11})$. Substituting a_{11} for x the result is obvious.

So assume the result is true for $n-1$. Write A in block triangular form as

$$\begin{pmatrix} A^{11} & \star \\ 0 & a_{nn} \end{pmatrix}$$

where A^{11} is square and upper–triangular and $\star$ is a matrix that will not affect the computation. By induction we know that

$$\prod_{i=1}^{n-1}(A^{11}-a_{ii}I_{n-1})=0.$$

Then by block multiplication

$$\prod_{i=1}^{n-1}(A-a_{ii}I_n)=\begin{pmatrix} \prod_{i=1}^{n-1}(A^{11}-a_{ii}I_{n-1}) & \star \\ 0 & \prod_{i=1}^{n-1}(a_{nn}-a_{ii}) \end{pmatrix}=\begin{pmatrix} 0 & \star \\ 0 & \star \end{pmatrix}.$$

To get this use the induction hypothesis and the fact that A is upper–triangular. We multiply this by $A-a_{nn}I$, which in the same block form is

$$\begin{pmatrix} \star & \star \\ 0 & 0 \end{pmatrix}$$

By block–multiplication the product of these two matrices is the zero matrix, so we are done. □

10.4 Cyclic Vectors

First we construct the minimal polynomial on certain subspaces invariant under a linear operator $L\colon V \to V$. We consider polynomials $f(x)$ such that $f(A)$ vanishes on the cyclic subspace generated by any non–zero $\mathbf{v} \in V$, rather than all of V. We use the definitions of §5.9: the powers of $\mathbf{v}$ under L generate a subspace W of V invariant under L of dimension p, called the period of $\mathbf{v}$. Therefore L restricts to an operator $L_{\mathbf{v}}$ on W. Lemma 5.9.1 gives the matrix A of $L_{\mathbf{v}}$ in the basis of powers of $\mathbf{v}$.

Definition 10.4.1. Any polynomial $g(x)$ that satisfies $g(A)\mathbf{v} = \mathbf{0}$ is an *A–annihilator* of $\mathbf{v}$ in V.

Theorem 10.4.2. *There is a unique monic polynomial $f(x)$ of smallest degree that annihilates $\mathbf{v}$. It divides any other polynomial that A–annihilates $\mathbf{v}$. $f(x)$ is the minimal polynomial of $L_{\mathbf{v}}$ on W.*

Proof. This follows from the existence of the minimal polynomial of $L_{\mathbf{v}}$ on W. □

For a moment assume that the iterations of $L_{\mathbf{v}}$ on $\mathbf{v}$ generate all of V, so the A–annihilator of $\mathbf{v}$ is the minimal polynomial of A on V. Then

Theorem 10.4.3. *The degree of the minimal polynomial of A is the dimension of V.*

Proof. Let $n = \dim V$. Then for dimension reasons there is an equation of linear dependence between $\{\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \ldots, A^n\mathbf{v}\}$:

$$b_0\mathbf{v} + b_1A\mathbf{v} + \cdots + b_{n-1}A^{n-1}\mathbf{v} + A^n\mathbf{v} = \mathbf{0}.$$

The last coefficient can be taken to be 1 since we know that the previous powers are linearly independent. Set M to be the matrix

$$b_0I + b_1A + \cdots + b_{n-1}A^{n-1} + A^n.$$

By construction M vanishes on $\mathbf{v}$, and it vanishes on all the $A^i\mathbf{v}$ since, for instance, $MA^i = A^iM$. So M vanishes on V. Any polynomial in A of degree less than n would lead to a relation of linear dependence between the powers $\{\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \ldots, A^{n-1}\mathbf{v}\}$ of $\mathbf{v}$. This contradicts Lemma 5.9.1. □

In the basis of compositions of $A^i\mathbf{v}$ on V, the matrix for A is given by (5.16). This matrix is called the *companion matrix* of the annihilator $f(A)$ of $\mathbf{v}$ In particular any monic polynomial is the minimal polynomial of a suitable operator, namely the operator with matrix A.

Back to the general situation. We can construct the A–annihilating polynomial for every vector $\mathbf{v} \in V$. Then finally

Theorem 10.4.4. *The minimal polynomial of A on V is divisible by the annihilator of the cyclic subspace generated by any non-zero $\mathbf{v} \in V$. More generally it is divisible by the minimal polynomial of the restriction of L to any L-invariant subspace W of V.*

Proof. It suffices to prove the last statement. Just use the fact that the minimal polynomial on V vanishes on W.

Thus any cyclic vector allows us to find factors of the minimal polynomial of A on V. If $\mathbf{v}$ is a non-zero cyclic vector of period 1, then $A\mathbf{v} = \lambda\mathbf{v}$ for some scalar λ, and the minimal polynomial of A restricted to the one-dimensional subspace generated by $\mathbf{v}$ is obviously $x-\lambda$. So $\mathbf{v}$ is an eigenvector for A, and λ its associated eigenvalue. The results above show that $x-\lambda$ divides the minimal polynomial of A on V.

Example 10.4.5. Consider the special case where A is the permutation matrix

$$A = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} \tag{10.6}$$

where the only nonzero entries are a_{1n} and $a_{i,i-1}$ for $i \geq 2$. An easy computation (do it) shows $A^n = I$, so $x^n - 1$ is a A–annihilating polynomial. By Theorem 10.4.3 the minimal polynomial is $x^n - 1$. Write $\lambda = e^{\mathrm{i}2\pi/n}$. Then the roots of this polynomial are the n distinct complex numbers λ^k, $0 \leq k \leq n-1$, called the n-th roots of unity, because $(\lambda^k)^n = 1$. Consider the vectors

$$\mathbf{v}_j = \lambda^j\mathbf{e}_1 + \lambda^{2j}\mathbf{e}_2 + \cdots + \lambda^{nj}\mathbf{e}_n.$$

Then

$$\begin{aligned} A\mathbf{v}_j &= \lambda^j A\mathbf{e}_1 + \lambda^{2j}A\mathbf{e}_2 + \cdots + A\mathbf{e}_n \\ &= \lambda^j\mathbf{e}_2 + \lambda^{2j}\mathbf{e}_3 + \cdots + \mathbf{e}_1 \\ &= \mathbf{e}_1 + \lambda^j\mathbf{e}_2 + \cdots + \lambda^{(n-1)j}\mathbf{e}_n \\ &= \lambda^{-j}\left(\lambda^j\mathbf{e}_1 + \lambda^{2j}\mathbf{e}_2 + \cdots + \lambda^{nj}\mathbf{e}_n\right) \\ &= \lambda^{-j}\mathbf{v}_j = \lambda^{n-j}\mathbf{v}_j. \end{aligned}$$

so we have found an eigenvector $\mathbf{v}_j$ with eigenvalue λ^{n-j} for each one of the factors of the minimal polynomial.

Exercise 10.4.6. Take any other permutation σ that is a cycle of period n. Write down the matrix that corresponds to A, and find the eigenvalues and eigenvectors.

In §5.9 we analyzed nilpotent operators. In the notation of Theorem 5.9.5 the minimal polynomial of a nilpotent L is the maximum of the p_i.

Exercise 10.4.7. Prove this.

10.5 The Primary Decomposition Theorem

Assume L is a linear operator on V and $f(x)$ a polynomial such that the operator $f(L)$ vanishes on V. The important primary decomposition theorem allows us to write V as a direct sum of subspaces corresponding to the primary factors of the minimal polynomial of the linear operators acting on V. Its proof only uses the polynomial techniques developed in Appendix C. We first prove a special case that we will use in the induction needed in the general case.

Theorem 10.5.1. *Let L be a linear operator acting on the vector space V, and let $f(x)$ a polynomial that vanishes on L: $f(L) = 0$. Assume that $f(x) = f_1(x)f_2(x)$, where $f_1(x)$ and $f_2(x)$ are polynomials of positive degree (therefore not constants) that are relatively prime. Let N_1 be the nullspace of $f_1(L)$ and N_2 be the nullspace of $f_2(L)$. Then $V = N_1 \oplus N_2$, where both N_1 and N_2 are invariant under L.*

If V is finite dimensional, let A be a matrix representing L in a basis where the first basis elements are a basis for N_1 and the remaining ones a basis for N_2. Then

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

so it is written in block diagonal form.

If V is finite dimensional, and $f(x)$ is the minimal polynomial of L on V, then $f_1(x)$ is the minimal polynomial of L restricted to N_2 and $f_2(x)$ is the minimal polynomial of L restricted to N_1.

Proof. By Corollary C.4.4 there are polynomials $c_1(x)$ and $c_2(x)$ such that $c_1(x)f_1(x) + c_2(x)f_2(x) = 1$. So

$$c_1(L)f_1(L) + c_2(L)f_2(L) = I \tag{10.7}$$

where I is the identity operator. Let P_1 be the operator $c_1(L)f_1(L)$ on V and P_2 the operator $c_2(L)f_2(L)$, so

$$P_1 + P_2 = I. \tag{10.8}$$

Now $P_1P_2 = 0$, since it contains as a factor $f_1(L)f_2(L) = f(L)$ which vanishes on V. Finally multiplying (10.8) by P_1 and then P_2 gives $P_1^2 = P_1$ and $P_2^2 = P_2$. So by Corollary 4.6.3 (the converse statement) if $U_i = Im(P_i)$, then $V = U_1 \oplus U_2$. Now U_1 is in the nullspace N_2 of $f_2(L)$, since any vector written as $c_1(L)f_1(L)(\mathbf{v})$ goes to 0 when $f_2(L)$ is applied, since a factor of $f(L)$ appears. The same argument shows $U_2 \subset N_1$. Next we show $N_2 \subset U_1$. Let $\mathbf{v}$ be in N_2, so $c_2f_2(L)\mathbf{v} = 0$. Multiply (10.7) on the right by the column vector $\mathbf{v}$. Then $\mathbf{v} = c_1(x)f_1(L)\mathbf{v}$, so $\mathbf{v} \in U_1$. In the same way we show $N_1 \subset U_2$, so $N_2 = U_1$ and $N_1 = U_2$. This gives the direct sum decomposition.

Now we assume V is finite dimensional, so it have a minimal polynomial $f(x)$ for the operator L. The polynomial $f_1(L)$ vanishes when considered as an operator on N_1. If it is not the minimal polynomial of L restricted to N_1, then there is a polynomial $g(x)$ of smaller degree than that of $f_1(x)$ such that $g(L)$ vanishes on N_1. But then the polynomial $g(x)f_2(x)$ is such that $g(L)f_2(L)$ vanishes: $g(L$ vanishes on

$N_1 \oplus 0$ and $f_2(L)$ vanishes on $0 \oplus N_2$. This contradicts the assertion that $f_1(x)f_2(x)$ is the minimal polynomial for A. □

Now we want to generalize this result to the finest possible decomposition of $f(x)$ that still allows the techniques of the proof to go through. Write $f(x)$ according to its decomposition in Theorem C.5.3. Let $f_i(x) = p_i(x)^{m_i}$. Then the $f_1(x)$ are relatively prime, so the theorem will go through. Here is the statement:

Theorem 10.5.2 (Primary Decomposition Theorem). *Let L be a linear operator acting on the vector space V, and let $f(x)$ a polynomial that vanishes on L: $f(L) = 0$. Assume that $f(x)$ factors as a product of primary polynomials $f_i(x) = p_i(x)^{m_i}$, where the irreducible p_i are distinct, $1 \leq i \leq k$. Let N_i be the nullspace of $f_i(A)$. Then $V = N_1 \oplus N_2 \oplus \cdots \oplus N_k$, where each N_i is invariant under L. Thus L restricts to an operator L_i on N_i*

If V is finite dimensional, let A be a matrix representing L in a basis for V consisting first of a basis of N_1 followed by a basis for N_2, ... followed by a basis of N_k. Then A can be written in block diagonal form

$$A = \begin{pmatrix} A_1 & 0 & \dots & 0 \\ 0 & A_2 & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & A_k \end{pmatrix}$$

where A_i is the restriction of A to N_i.

Furthermore if $f(x)$ is the minimal polynomial of L on V, then $f_1(x)$ is the minimal polynomial of L restricted to $N_2 \oplus \cdots \oplus N_k$ and so on.

Proof. We prove this by induction on the number of factors k by writing $f(x)$ as $f_1(x)g(x)$, where $g(x) = \prod_{i=2}^{k} f_i(x)$. The key point is that $f_1(x)$ and $g(x)$ are relatively prime. By induction we may assume that the theorem is true for $g(x)$; then we must establish it for $f(x)$. Still needs finishing □

The value of this theorem is that it shows that to understand an operator A acting on a vector space V, it suffices to decompose its minimal polynomial $f(x)$ into its primary components. Then the operator can be decomposed into a sum of operators each (called A_i) acting on a subspace N_i that is part of a direct sum decomposition of V. Thus we have reduced the original problem into a collection of smaller problems - because each smaller problem concerns an operator whose minimal polynomial is primary, acting on a vector space of smaller dimension.

For the following corollary we use the terminology of Theorem 10.5.1.

Corollary 10.5.3. *In N_i there is a vector $\mathbf{v}_i$ whose A_i-annihilator is $p_i(x)^{m_i}$. Furthermore the A annihilator of $\mathbf{v}_1 + \cdots + \mathbf{v}_k$ is the minimal polynomial $f(x)$ of A.*

Proof. Recall that by Theorem 10.4.4, the A_i-annihilator of any vector in N_i divides the minimal polynomial $p_i(x)^{m_i}$. Thus if there is no vector with annihilator $p_i(x)^{m_i}$, then all vectors are annihilated by a smaller power of $p_i(x)$, which then must be the minimal polynomial. Thus there is a vector $\mathbf{v}_i$ whose A_i annihilator is $p_i(x)^{m_i}$.

Now consider the vector $\mathbf{v} = \sum_i \mathbf{v}_i$. It is obviously annihilated by $f(x) = \prod p_i(x)^{m_i}$, but not by any non-trivial divisor of $f(x)$. □

Polynomials $x - a_i$ of degree 1 are are all irreducible. They are relatively prime if and only if the coefficient a_i are distinct. (When the base field F is the complex numbers, then the irreducible polynomials are all of degree 1.)

Example 10.5.4. If we are the situation of Theorem 10.1.5, where $k = n$, with n eigenvectors $\mathbf{v}_i$ with distinct eigenvalues λ_i, then the nullspace N_i of each $L - \lambda_i I$ is non-trivial, so each N_i has dimension 1. The Primary Decomposition Theorem tells us that in the basis of the $\mathbf{v}_i$ the matrix of L is diagonal with the λ_i along the diagonal. Then the polynomial $\prod_{i=1}^{n}(x - \lambda_i)$ is the minimal polynomial, as you should check. So we have a converse in this case to the last statement in Theorem 10.5.2, and the degree of the minimal polynomial is the dimension of the space. This is another way of establishing Theorem 10.1.5.

In general we do not know how to construct the minimal polynomial, except in special cases such as the one above, when we can find a cyclic vector. The only constructive method we have is to solve the system of equations (10.5) in $n^2 + 1$ variables. We will remedy this problem (in part) by defining the characteristic polynomial $f(x)$ of A, and showing that $f(A)$ vanishes: this is the famous Cayley-Hamilton Theorem, proved in the next chapter. Since the characteristic polynomial has degree n, the homogeneous system of equations is much easier to solve.

Remark 10.5.5. The minimal polynomial seems to depend on the base field. Suppose that the matrix A has coefficients in $\mathbb{R}$ as often happens in examples. Then we can define the minimal polynomial for A over $\mathbb{A}$. One can prove that this polynomial is the same as the minimal polynomial of A over $\mathbb{C}$, even though the irreducible factors are different. We will settle this in §10.8.

10.6 The Jordan Canonical Form

In the remainder of this chapter we show how to write a linear operator L acting on a complex vector space V as a direct sum of operators whose matrix is a Jordan block, that we shall soon define. This representation is essentially unique as we shall see in §10.7. In the last section we show what happens over $\mathbb{R}$.

In this section we deal with any operator whose minimal polynomial only has linear factors, so it is the product

$$\prod_{i=1}^{k}(x - a_i)^{m_i} = (x - a_1)^{m_1}(x - a_2)^{m_2} \dots (x - a_k)^{m_k}, \tag{10.9}$$

where $a_i \neq a_j$ for all $i \neq j$. By the Fundamental Theorem of Algebra (see §C.6) this covers all operators over $\mathbb{C}$.

An important corollary of the main result we prove in this section is

Corollary 10.6.1. *Any linear operator whose minimal polynomial factors into linear factors is triangulable. In particular all complex linear operators are triangulable.*

Since the linear polynomials in (10.9) are relatively prime, by the Primary Decomposition Theorem we only need to consider the case of one factor. So we may assume the minimal polynomial of L is just $(x-a)^m$. Therefore $(x-a)$ is the irreducible polynomial $p(x)$ of the previous sections, so there is a $\mathbf{v}$ so that $(L-aI)^{m-1}\mathbf{v} \neq \mathbf{0}$, and yet of course $(L-aI)^m\mathbf{v} = \mathbf{0}$. So each non-zero $\mathbf{v}$ generates a cyclic $(L-aI)$ subspace as in Lemma 5.9.1.

Definition 10.6.2. A *Jordan block* of size r for a is the $r \times r$ matrix

$$J_r(a) = \begin{pmatrix} a & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & a & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & a & \ddots & 0 & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 & a & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 & a \end{pmatrix} \tag{10.10}$$

with a along the diagonal, 1 on the subdiagonal, and 0 everywhere else. When a is given, we write just J_r.

We prove:

Theorem 10.6.3. *Let $L\colon V \to V$ be a linear operator on the vector space V of dimension n, with minimal polynomial $(x-a)^m$. Then $V = \oplus V_i$, where V_i is an $L-aI_n$-invariant cyclic subspace of size r_i, so that if $L_i - aI_{r_i}$ is the restriction of $L-aI$ to the r_i-dimensional V_i, then there is a vector $\mathbf{v}_i \in V_i$ so that a basis for V_i is, generalizing* (5.18)*,*

$$\{\mathbf{v}_i, (L_i-aI)\mathbf{v}_i, (L_i-aI)^2\mathbf{v}_i, \dots, (L_i-aI)^{r_i-1}\mathbf{v}_i\}, \text{ for some } r_i \leq m, \tag{10.11}$$

and the $r_i \times r_i$ matrix A_i for L_i in this basis is the Jordan block J_{r_i}.

This explains the simpler matrix we get at the end: (10.10) compared to (5.16). Since the Jordan blocks are all triangular, this establishes Corollary 10.6.1.

Proof. Consider the operator $M = L - aI$. The key elementary observation is that M is nilpotent, since by hypothesis $M^m\mathbf{v} = 0$ for all $\mathbf{v} \in V$.

Therefore we may apply Theorem 5.9.5 to M. Thus in a suitable basis the matrix of M is block diagonal with standard nilpotent blocks N_r along the diagonal. Adding the matrix aI, we get the Jordan blocks $J_r(a)$ as required.

□

Corollary 10.6.4. *The minimal polynomial of L is $(x-a)^{r_i}$, where r_i is the maximum of the periods over the direct sum. So it is the size of the largest Jordan block.*

Write down the general Jordan matrix for 1 eigenvalue, then several.

Theorem 10.6.5. *Let L be an operator on a space V of dimension n whose minimal polynomial is $\prod_{i=1}^{k}(x-a_i)^{m_i}$. Then the degree of the minimal polynomial is strictly less than n unless there is only one Jordan block for each factor.*

Because all polynomiais factor as a product of polynomials of degree 1 over $\mathbb{C}$, we have achieved one of the major goals of linear algebra: we have a complete classification of all linear operators over $\mathbb{C}$.

Remark 10.6.6. In some books the Jordan block is defined as the transpose of (10.10). This is simply because the basis has been taken in the opposite order:

$$\mathbf{v}_{r-1}, \mathbf{v}_{r-2}, \ldots, \mathbf{v}_1, \mathbf{v}.$$

The $r \times r$ permutation matrix

$$P_r = \begin{pmatrix} 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & 1 & 0 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 1 & \ldots & 0 & 0 \\ 1 & 0 & \ldots & 0 & 0 \end{pmatrix} \tag{10.12}$$

is called the *reversal matrix*, since it simply reverses the order of the variables. It is symmetric and orthogonal, so $P^t = P^{-1}$.

Then

$$J_r^t = P_r J_r P_r^{-1}. \tag{10.13}$$

The difference between the two definitions of the Jordan block is therefore unimportant since J_r and J_r^t are similar.

10.7 Uniqueness of the Jordan Form

To finish the classification over $\mathbb{C}$, we show that the decomposition into Jordan blocks that we have found is unique, in the sense that for a given irreducible factor $(x-a)$, the number n_i of Jordan blocks of size r_i is uniquely determined for all i.

We can do this one irreducible factor at a time, so assume that we have only one factor $(x-a)$ in the minimal polynomial. Let m be the degree of the minimal polynomial. The information that we have (the givens of the problem) are the dimension n of the space, the degree m of the minimal polynomial, and the dimension K_i of the nullspace of $(L-aI)^i$. Obviously we only need consider the K_i for $i \leq m$. Since the nullspace of $(L-aI)^i$ is contained in the nullspace of $(L-aI)^{i+1}$, we have $K_i \leq K_{i+1}$. Obviously $K_0 = 0$.

Since there is a contribution of 1 to the nullspace of $(L-aI)$ from each cyclic subspace, we have

$$K_1 = \sum_{i=1}^{m} n_i.$$

For $(L-aI)^2$, we already have K_1 from the previous nullspace. Only the cyclic subspaces of period $r_i \geq 2$ contribute again, so

$$K_2 = K_1 + \sum_{i=2}^{m} n_i.$$

More generally,

$$K_{j+1} = K_j + \sum_{i=j}^{m} n_i$$

and of course $K_m = n$. So we have a system of m linear equations in m variables n_1, n_2, $\ldots n_m$. To make it simpler to solve, write $k_1 = K_1$, and $k_j = K_j - K_{j-1}$. So our system of equations is (remember that the variables are the n_i) $A\mathbf{n} = \mathbf{k}$, where A is the upper-triangular matrix

$$\begin{pmatrix} 1 & 1 & \ldots & 1 & 1 \\ 0 & 1 & \ldots & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \ldots & 1 & 1 \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix}$$

This matrix is invertible: in fact its inverse was seen in Exercise 2.3.6 to be

$$\begin{pmatrix} 1 & -1 & \ldots & 0 & 0 \\ 0 & 1 & -1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & 1 & -1 \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix}$$

so we get the unique solution $\mathbf{k} = A^{-1}\mathbf{n}$, which shows that the sizes of the Jordan blocks are uniquely determined by the operator.

Example 10.7.1. Assume that L is diagonalizable. This means that all its Jordan blocks have size 1. So $n_1 = n$, and $n_i = 0$ for $i > 1$. So $A\mathbf{n}$ applied to the vector $(1,0,\ldots,0)$ gives $k_1 = n_1$, and $k_i = 0$ for $i > 1$.

At the other extreme, assume that there is only one Jordan block of size r. So $n_r = 1$, and all the others are 0. Apply A to this vector to get $k_1 + k_2 + \cdots + k_r = 1$. The $n = m = r$

From this we get the important theorem:

Theorem 10.7.2. *Suppose given two linear operators whose minimal polynomials factor into the same linear terms, and whose Jordan decompositions on each factor are the same. Then they are similar.*

The proof is easy: construct an isomorphism on each Jordan block, and since the sum is direct, this isomorphism extends to the entire space.

10.8 The Jordan Form over the Real Numbers

So far we have only determined all the Jordan forms over the complex numbers. To get the Jordan form of a real $n \times n$ matrix A over the reals, we complexify, as in §5.8. Therefore we view A as acting on $\mathbb{C}^n$. We get a complex Jordan decomposition for A, in which some of the eigenvalues λ are real, while the others appear in complex conjugate pairs. The Jordan blocks $J_k(a)$, $a \in \mathbb{R}$ of A over $\mathbb{R}$ are the same as those over $\mathbb{C}$. A may have blocks $J_k(a)$, where a is not in $\mathbb{R}$. Then

Theorem 10.8.1. *Let λ be a non-real eigenvalue for A. Then the number and size of the Jordan blocks for λ are the same as those for the complex conjugate $\overline{\lambda}$.*

Proof. In the factorization of a real polynomial over $\mathbb{C}$, there are as many factors $(x-\lambda)$ as there are of factors $(x-\overline{\lambda})$. □

Theorem 10.8.2. *A Jordan block for corresponding to the complex conjugate pair of eigenvalues $\lambda = a+\mathrm{i}b$ and $\overline{\lambda} = a-\mathrm{i}b$ is, when written in 2×2 blocks:*

$$\begin{pmatrix} \Lambda & 0 & 0 & \dots & 0 \\ I_2 & \Lambda & 0 & \dots & 0 \\ 0 & I_2 & \Lambda & 0 & \dots \\ \vdots & \ddots & I_2 & \Lambda & 0 \\ 0 & 0 & \dots & I_2 & \Lambda \end{pmatrix} \tag{10.14}$$

where

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and

$$\Lambda = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}.$$

Therefore it is not triangular.

These results can be proven without too much difficulty but with a great deal of book keeping. Proofs can be found in Shafarevich– Remizov and Horn–Johnson [13]. It would take us too far afield to prove them.

10.9 An Application of the Jordan Canonical Form

Here is an important and surprising application.

Theorem 10.9.1. *Let A be an arbitrary $n \times n$ matrix. Then A is similar to A^t.*

Proof. Since for an invertible matrix $(C^t)^{-1} = (C^{-1})^t$ we write this just C^{-t} for simplicity of notation. We start by assume that A has only one Jordan block J, therefore of size n. By construction A is similar to J, so there is an invertible matrix C such that $A = CJC^{-1}$, so

$$A^t = C^{-t}J^tC^t = C^{-t}PJPC^t$$

using (10.12) and (10.13). Now $J = C^{-1}AC$ so substituting that in the right hand side, we get

$$A^t = C^{-t}PC^{-1}ACPC^t. \tag{10.15}$$

Since the inverse of $D = CPC^t$ is $C^{-t}P^{-1}C^{-1} = C^{-t}PC^{-1}$, then (10.15) becomes $A^t = D^{-1}AD$, so A and A^t are indeed similar. To do the general case just replace P by the block diagonal matrix with blocks P_r of appropriate size along the diagonal. This matrix is clearly still orthogonal and symmetric, and transforms the Jordan form of A into its transpose form. Then the proof goes through as before. □

A second major application is to the solutions of systems of linear differential equations with constant coefficients. That is covered in §16.4.

Chapter 11
The Determinant

Abstract If you compare elimination from linear equations in Chapter 1 to Gaussian elimination in Chapter 2, you notice that one of the main differences is that in the first it was not necessary to keep careful track of the order of the equations or of the order of the variables, since it is clear that the zero set of a set of linear equations depends on neither. The situation is different in Gaussian elimination. Because the matrix of coefficients has numbered rows and columns: we have to define a row operation that interchanges the rows of a matrix. In the same way we can define a column operation that interchanges the columns of a matrix. So in this chapter we first look at the mathematics behind row interchanges or column interchanges of a matrix: this is the mathematics of permutations. Why do we need this? In order to define the determinant of a square matrix. We already wrote it down in Chapter 1 for square matrices of size 2 in (1.11) and size 3 in (1.15). We now want to write down the general formula, and understand the properties of the determinant. The most important square matrices are those that are invertible. We know that this have maximum rank, meaning that their rank is equal to their size. We want to get a new criterion for invertibility of square matrices: their determinant is non–zero.

11.1 Permutations

If you need background on the language of sets, see §B.1.

Definition 11.1.1. A permutation of a finite set S is a one-to-one map from S to itself.

Example 11.1.2. If the set S has three elements, that we label 1, 2 and 3 then the map σ from S to S such that $\sigma(1) = 2$, $\sigma(2) = 3$, $\sigma(3) = 1$ is a permutation. So is the map τ such that $\tau(1) = 2$, $\tau(2) = 1$, $\tau(3) = 3$

We denote permutations by lower case Greek letters, such as σ, τ, υ, γ. We reserve the Greek letter iota, written ι, for the trivial permutation, namely the permutation that sends every element k of S to itself: $\iota(k) = k$. If the set S has n elements,

we write it as S_n. We usually use the integers $\{1,2,\ldots,n\}$ to label the elements of S_n, but on occasion we are forced to use a different set of labels, for example $\{0,1,\ldots,n-1\}$ or $\{1,2,\ldots,k-1,k+1,\ldots,n+1\}$ for any integer k between 1 and $n+1$ except for the integer k.

Example 11.1.3. The simplest permutations after the trivial one ι are the *transpositions*, which interchange two integers but do not move the others. For example, the permutation σ with values $\sigma(1)=2$, $\sigma(2)=1$, and $\sigma(i)=i$ for $i\neq 1,2$ is a transposition.

There are exactly two permutations on $\{1,2\}$, the trivial permutation and the transposition exchanging 1 and 2.

A standard way of writing a permutation σ on n elements consists in writing the integers 1 through n on a top row; then beneath each integer i write the value $\sigma(i)$. So, for example

$$\begin{vmatrix}1&2&3&4\\2&4&3&1\end{vmatrix} \tag{11.1}$$

denotes the permutation σ sending 1 to 2, 2 to 4, 3 to 3 and 4 to 1. In this notation, the fact that a permutation is one-to-one is expressed by the fact that each integer from 1 to n appears exactly once in the second row. Notice that if you interchange the rows you get a new permutation τ, where $\tau(1)=4$, $\tau(2)=1$, $\tau(3)=3$,$\tau(4)=1$,

Exercise 11.1.4. Enumerate all the permutations on $\{1,2,3\}$, listing the trivial permutation, then all the transpositions, and then the remaining ones.

Recall that $n!=n(n-1)(n-2)\cdots(2)(1)$. Therefore $2!=2$, $3!=6$ and $4!=24$.

Proposition 11.1.5. *There are $n!$ distinct permutations of a set with n elements. We write the set of permutations on n elements as $\mathscr{S}_n$, not to be confused with S_n.*

Proof. We prove this by induction on n. The starting case $n=1$ is trivially verified, and we have already noticed that it is true for $n=2$. Suppose that we have proved that there are $n!$ permutations on n elements. We use this to prove the case $n+1$. The new element $n+1$ can be mapped to any integer k between 1 and $n+1$. For each choice of k the remaining integers $(1,2,,\ldots,n)$ can be mapped bijectively to the n integers $(1,\ldots,k-1,k+1,\ldots,n+1)$. By induction there are $n!$ of doing this for each of the $n+1$ choice of k. So in total there are $n!\cdot(n+1)=(n+1)!$ permutations. □

Definition 11.1.6. We can follow a permutation σ in $\mathscr{S}_n$ by another permutation τ in $\mathscr{S}_n$, yielding a third permutation $\tau\circ\sigma$ called the composition of the two permutations or the product permutation. It sends the element k to $\tau\circ\sigma(k)$). We often drop the circle in the representation of the composition of permutations, writing just $\tau\sigma$ for $\tau\circ\sigma$.

So using the σ and τ from Example 11.1.2 $(\tau\circ\sigma)(1)=1$, $(\tau\circ\sigma)(2)=3$ and $(\tau\circ\sigma)(3)=2$. We can also form $\sigma\circ\tau$. In this example we get $(\sigma\circ\tau)(1)=3$,

$(\sigma\circ\tau)(2)=2$ and $(\sigma\circ\tau)(3)=1$. In particular it is not always true that $\tau\circ\sigma=\sigma\circ\tau$. We say that composition of permutations is not commutative.

Any permutation has an inverse, namely the permutation σ^{-1} that undoes the effect of σ: for all k, $\sigma^{-1}(\sigma(k))=k$. So, for example the inverse of the permutation in (11.1) is

$$\begin{vmatrix}1&2&3&4\\4&1&3&2\end{vmatrix}$$

obtained just by interchanging the order of the rows

By Theorem B.1.5, composition of permutations is associative, so

$$\upsilon\circ(\tau\circ\sigma)=(\upsilon\circ\tau)\circ\sigma,$$

so we can omit the parentheses. We sometimes also omit the $\circ$ and just write $\upsilon\tau\sigma$.

Exercise 11.1.7. Write the inverse of the permutation $\sigma\in\mathscr{S}_3$ sending $1\to 2$, $2\to 3$, $3\to 1$. Write σ^2 and σ^3. Here, we write σ^2 for $\sigma\sigma$, and σ^3 for $\sigma\sigma\sigma$.

In our permutation notation (11.1)

$$\sigma^{-1}=\begin{vmatrix}1&2&3\\3&1&2\end{vmatrix}.$$

Similarly you will find that $\sigma^2=\sigma^{-1}$, from which it follows that σ^3 is the identity.

Let's write the transposition that interchanges i and j as τ_{ij}. By definition it fixes all the other integers.

Proposition 11.1.8. *Any permutation in $\mathscr{S}_n$ can be written as the composition of at most $n-1$ transpositions.*

Proof. We prove this by induction on n. The result is obvious for $n=1$. Assume we have established it for the integer n. We need to prove it for $n+1$. Let σ is a permutation of the $n+1$ integers $[1,\dots,n+1]$. Then σ sends $n+1$ to some element k. Compose σ with the permutation $\tau_{k,n+1}$: $\tau_{k,n+1}\circ\sigma$. If $k=n+1$, the $\tau_{n+1,n+1}$ means the identity permutation ι, otherwise it means the transposition $j\leftrightarrow n+1$. Then the composition $\tau_{k,n}\circ\sigma$ fixes $n+1$, so it is a permutation of $[1,\dots,n]$. By induction this is a composition of $n-1$ transpositions, so σ is a composition of n transpositions, as required. □

Exercise 11.1.9. Write the permutation

$$\begin{vmatrix}1&2&3&4\\2&3&4&1\end{vmatrix},$$

as a composition of transpositions.

Definition 11.1.10. Assume $(i_1,i_2,\dots,i_m)$ are m distinct elements in $[1,\dots,n]$, Let γ be the permutation that sends i_1 to i_2, i_2 to i_3, …, i_{m-1} to i_m and finally i_m to i_1. Thus

$$\begin{vmatrix} i_1 & \dots & i_{m-1} & i_m \\ i_2 & \dots & i_m & i_1 \end{vmatrix},$$

while all the remaining integers are fixed by γ and are omitted. Then γ is called a *cycle* of order m. When $m = 1$, γ is the identity, when $m = 2$ it is the transposition $\tau_{i_1 i_2}$. When $m > 2$ it can be written as the composition of the $m-1$ transpositions:

$$\tau_{i_1 i_2} \tau_{i_2 i_3} \dots \tau_{i_{m-1} i_m},$$

as you should check.

Example 11.1.11. Consider the permutation

$$\upsilon = \begin{vmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 2 & 1 \end{vmatrix},$$

Then $\upsilon = \gamma_{135} \circ \tau_{24}$. We say υ is the composition of disjoint cycles. Note υ is also equal to $\tau_{24} \circ \gamma_{135}$.

Exercise 11.1.12. Show any permutation can be written as the composition of disjoint cycles, and devise an algorithm for computing the cycles. Also show that disjoint cycles commute.

11.2 Permutation Matrices

Given a permutation σ on $[1, \dots, n]$, there are two matrices that can be naturally associated to σ.

Definition 11.2.1. First the matrix $P^\sigma = (p^\sigma_{ij})$ with entries:

$$p^\sigma_{ij} = \begin{cases} 1, & \text{if } j = \sigma(i); \\ 0, & \text{otherwise.} \end{cases}$$

and second the matrix $Q^\sigma = (q^\sigma_{ij})$ with entries:

$$q^\sigma_{ij} = \begin{cases} 1, & \text{if } i = \sigma(j); \\ 0, & \text{otherwise.} \end{cases}$$

So P^σ has ones is positions $(i, \sigma(i))$ while Q^σ has ones is positions $(\sigma(j), j)$.

Example 11.2.2. For any n, the matrices P^ι and Q^ι associated to the identity permutation are both the identity matrix. When $n = 2$, the matrices associated to the unique non-trivial element of $\mathscr{S}_2$ are both

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Example 11.2.3. Consider the permutation σ defined by

$$\begin{vmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{vmatrix}.$$

So $\sigma(1) = 2$, $\sigma(2) = 3$, and $\sigma(3) = 1$, It is a cycle of order 3. Then

$$P^\sigma = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \text{ and } Q^\sigma = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

The following property is often used to define a permutation matrix.

Theorem 11.2.4. *A square matrix is a permutation matrix if and only if it has exactly one* 1 *in each row and each column, and all other entries equal to* 0

Proof. To each matrix with exactly one 1 in each row and column, and 0 everywhere else, there corresponds a permutation. Indeed if the matrix has a 1 in position (i, j), then the associated permutation σ is defined by $\sigma(j) = i$, for all j, and the matrix is Q^σ. This is well defined because there is only one non-zero element ion column j. On the other hand, taking the permutation τ such that $\tau(i) = j$, then the same matrix is P^τ. □

Theorem 11.2.5. *Q^σ is the transpose of P^σ and $P^\sigma Q^\sigma = I$, so Q^σ is the inverse of P^σ.*

Proof. The $\sigma(i)$-th row of Q^σ is the $\sigma(i)$-th column of P^σ, and this gives both results. □

Example 11.2.6. Now consider the transposition σ_{12} in $\mathscr{S}_3$. Then

$$P^\sigma = Q^\sigma = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Next we look at products of permutation matrices. First an exercise:

Exercise 11.2.7. Compute the matrix product $P^\sigma P^\tau$ for the matrices above. Also compute $P^\tau P^\sigma$. Note they are not equal, but that each one is a permutation matrix.

Next we work out the general case:

Theorem 11.2.8. *If σ and τ are two permutations on $[1, \ldots, n]$, with permutation matrices defined as above, then:*

$$P^\sigma P^\tau = P^{\tau\sigma} \quad \textit{and} \quad Q^\sigma Q^\tau = Q^{\sigma\tau}.$$

Thus in the second case, the matrix $Q^{\sigma\tau}$ of the composition of permutations is the matrix product $Q^\sigma Q^\tau$ of the matrices of permutations. In the first case the order is reversed.

Proof. We first prove the result for the P matrices. Where does $P^\sigma P^\tau$ have a 1 in the i-th row? Work out the multiplication: take the i-row of P^σ: it has a 1 in the $\sigma(i)$-th position. So we are looking for the column of P^τ with a 1 in the $\sigma(i)$-th position, i.e. the $\sigma(i)$-th row of P^τ. That is column $\tau(\sigma(i))$. This shows that the product matrix is $P^{\tau(\sigma}$, proving the first statement.

We get the second statement by taking transposes of the first:

$$(P^\sigma P^\tau)^t = (P^{\tau\sigma})^t \text{ so } (P^\tau)^t (P^\sigma)^t = (P^{\tau\sigma})^t.$$

The last equality is the desired conclusion, since $(P^\tau)^t = Q^\tau$. □

Corollary 11.2.9. *The composition of two permutation matrices is a permutation matrix.*

Example 11.2.10. The permutation matrix of the transposition τ exchanging 1 and 2 has matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

and therefore is its own inverse.

Theorem 11.2.11. *There are exactly $n!$ permutation matrices of size n.*

This is obvious, since there is one Q matrix for each permutation.

The following theorem is important.:

Theorem 11.2.12. *If $\mathbf{v}$ is the column vector $[v_1, \ldots, v_n]$, then $P^\sigma \mathbf{v}$ is the column vector $[v_\sigma(1), \ldots, v_\sigma(n)]$. More generally if A is a $n \times m$ matrix, the matrix $P^\sigma A$ is the matrix whose rows are, in order, $\mathbf{a}^{\sigma(1)}$, $\mathbf{a}^{\sigma(2)}$, $\ldots$, $\mathbf{a}^{\sigma(n)}$. If A is a $r \times n$ matrix, then AQ^σ is the $r \times n$ matrix whose columns are, in order, $\mathbf{a}_{\sigma(1)}$, $\mathbf{a}_{\sigma(2)}$, $\ldots$, $\mathbf{a}_{\sigma(n)}$.*

Proof. Just work out the multiplication. □

Thus multiplying a matrix A on the left by a permutation matrix gives a row operation on A: this generalizes what we did in §2.8. Indeed the elementary matrix called T_{rs} in Definition 2.8.1 is simply the permutation matrix of the transposition τ_{ij}. By Proposition 11.1.8 we now see that by multiplying on the left repeatedly by different elementary matrices interchanging rows, we may achieve any permutation of the rows of A that we desire.

Example 11.2.13. Here is a computation:

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 4 & 5 & 6 \\ 7 & 8 & 9 \\ 1 & 2 & 3 \end{pmatrix}$$

as you should check.

Definition 11.2.14. Now fix a permutation σ and its permutation matrix P^σ. Its non-zero entries are by definition $(p^\sigma_{i,\sigma(i)})$. Consider two distinct entries $p_{i,\sigma(i)}$ and $p_{j,\sigma(j)}$. They are *reversed* if $i < j$ and $\sigma(i) > \sigma(j)$. Thus the row index of $p_{i,\sigma(i)}$ is less than that of $p_{j,\sigma(j)}$, while the opposite is true for the column indices. Thus if we draw an arrow from the $(i,\sigma(i))$ entry to the $(j,\sigma(j))$ in the standard representation of the matrix, it goes north-west if they are reversed.

Example 11.2.15. The permutation matrix in Example 11.2.13 has two reversals: first, the one between the first and second columns, that we write $1 \leftrightarrow 2$, and then $1 \leftrightarrow 3$.

Consider the two permutation matrices

$$\begin{pmatrix} 1&0&0&0\\ 0&0&1&0\\ 0&1&0&0\\ 0&0&0&1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0&0&1&0\\ 0&1&0&0\\ 1&0&0&0\\ 0&0&0&1 \end{pmatrix}$$

The one on the left has one reversal $2 \leftrightarrow 3$, and the one on the right three reversals $1 \leftrightarrow 2$, $1 \leftrightarrow 3$, and $2 \leftrightarrow 3$.

The number of reversals for the matrix P^σ, or, which is the same, for the permutation σ is written $\mathrm{rev}(\sigma)$.

A key step is the following:

Exercise 11.2.16. For the transposition of adjacent elements $\tau_{i,i+1}$, $\mathrm{rev}(\tau_{i,i+1}) = 1$. For the transposition $\tau_{i,i+k}$, $\mathrm{rev}(\tau_{i,i+k}) = 2k-1$.

Hint: To get started consider the transposition $\tau_{2,3}$ in $\mathscr{S}_3$. There is one reversal $2 \leftrightarrow 3$. Now consider the transposition $\tau_{1,3}$. Its matrix is

$$\begin{pmatrix} 0&0&1\\ 0&1&0\\ 1&0&0 \end{pmatrix}$$

There are three reversals $1 \leftrightarrow 2$, $1 \leftrightarrow 3$, and $2 \leftrightarrow 3$ as noted above. Just continue in the same way.

Definition 11.2.17. We define the sign $\varepsilon(\sigma)$ of σ to be $(-1)^{\mathrm{rev}(\sigma)}$. So ε can be either 1 or -1. If $\mathrm{rev}(\sigma)$ is even, then $\varepsilon(\sigma) = 1$, in which case we say that σ is an even permutation; while if $\mathrm{rev}(\sigma)$ is odd, $\varepsilon(\sigma) = -1$: we say σ is an odd permutation.

Thus by Exercise 11.2.16, the sign of any transposition is -1. Here is the theorem that explains the importance of the sign.

Theorem 11.2.18.

$$\varepsilon(\tau\sigma) = \varepsilon(\tau)\varepsilon(\sigma).$$

Thus the composition of two even permutations is even, the composition of two odd permutations is even, and the composition of an odd and an even permutation is odd.

Proof. Since every permutation can be written as a composition of transpositions, it is enough to prove this when τ is a transposition. So assume that τ is the transposition τ_{ij}, $j = i+k$. From the argument of Exercise 11.2.16, we see that τ_{ij} interchanges the order in $2k-1$ pairs. □

Theorem 11.2.19. *In $\mathscr{S}_n$, $n \geq 2$, there are as many even permutations as odd permutations.*

Proof. The proof is simple and beautiful. Let E be the set of even permutations on $n \geq 2$ elements, and O the set of odd permutations. Let τ be any transposition. Then composition by τ gives a map from the even permutations to the odd permutations by Theorem 11.2.18. This map is

- injective: if σ_1 and if σ_2 get mapped to the same element when multiplied by τ, we get $\tau\sigma_1 = \tau\sigma_2$. Now apply by τ^{-1} to see that $\sigma_1 = \sigma_2$.
- It is also surjective. We must show that any odd permutation σ comes from an even one. Indeed, if comes from $\tau^{-1}\sigma$.

Thus E and O have the same number of elements. □

Since the total number of permutations is $n!$ by Proposition 11.1.5, the number of even and odd permutations is $n!/2$ except when $n = 1$..

Exercise 11.2.20. Write down the even and odd permutations for $n = 2,3,4$.

Exercise 11.2.21. By Example 11.1.10, if γ is a cycle of order m, then $\varepsilon(\gamma) = (-1)^{m-1}$.

11.3 Permutations and the Determinant

In (1.10), we gave the formula for a quantity that must be non-zero for there to be a unique solution to an linear system of 2 equations in 2 variables:

$$a_{11}a_{22} - a_{12}a_{21}.$$

We have only two permutations in two variable: the identity permutation ι and the transposition τ that exchanges 1 and 2, and therefore has sign $\varepsilon(\tau) = -1$. Then we notice that our formula can be written

$$a_{1,\iota(1)}a_{2,\iota(2)} + \varepsilon(\tau)a_{1,\tau(1)}a_{2,\tau(2)}.$$

In (1.14) we did the same thing for three equations in three variables, getting the expression

$$a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}. \quad (11.2)$$

Here we have six terms, three with a positive sign and three with a negative sign. In each term, say $a_{13}a_{21}a_{32}$ the first subscript of the coefficients are arranged $\{1,2,3\}$,

while the second coefficients describe all six permutations on $\{1,2,3\}$. Furthermore the three terms corresponding to transpositions have coefficient -1. This implies that (11.2) can be written

$$\sum_{\sigma\in\mathscr{S}_3} \varepsilon(\sigma) a_{1,\sigma(1)} a_{2,\sigma(2)} a_{3,\sigma(3)}.$$

Written this way, this formula generalizes nicely to n variables, and leads us to our definition.

Definition 11.3.1 (The Leibniz Formula). Let $A = (a_{ij})$ be a square matrix of size n. The its determinant $\det(A)$ is

$$\det(A) = \sum_{\sigma\in\mathscr{S}_n} \varepsilon(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)} \tag{11.3}$$

We also occasionally use the traditional notation

$$\det A = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}$$

when we want to show the coefficients. So the determinant is a function that associates to any square matrix a number.

Exercise 11.3.2. Show that (11.3) generalizes the formulas in the cases $n = 2$ and $n = 3$ above.

Exercise 11.3.3. Show that in each product $\prod_{i=1}^{n} a_{i,\sigma(i)}$ there is an exactly one entry from each row and each column of A.

Theorem 11.3.4. *Assume A is upper-triangular, so that $a_{ij} = 0$ if $i > j$. Then* $\det(A)$ *is the product of the diagonal elements: $a_{11}a_{22}\dots a_{nn}$. Similarly when A is lower-triangular.*

Proof. All we have to do is show that for any non-trivial permutation σ, there is an i such that $i < \sigma(i)$: then all the other terms in the sum (11.3) vanish. For σ to be non-trivial means that there is an i such that $i \neq \sigma(i)$. We may therefore assume that $i > \sigma(i)$, since otherwise we get 0. Let $j = \sigma(i)$, and consider the integers $S_j = \{1,2,\dots,j\}$. Where are they mapped by the bijection σ? Since j is already the image of $i > j$, as least one of them, say $k \leq j$ has to be sent to an element outside S_j in other words $\sigma(k) > k$, so we are done. □

If our matrix A is written in blocks, and is block triangular according to Definition 2.9.7, then the previous theorem generalizes:

Theorem 11.3.5. *Assume A is block triangular, with diagonal blocks A^{11}, …, A^{ss}. Then*

$$\det(A) = \prod_{i=1}^{s} \det(A^{ii}).$$

Proof. Just do the same analysis as in the previous theorem, but do it block by block. Show that the only permutations σ that contribute a non-zero term to the determinant of A are the ones that permute the first n_1 elements among themselves, then the next n_2 elements and so on. The result then follows. □

A criterion under which a determinant vanishes is given in §11.11. It will not be used later.

We have already shown that $\det(A)$ has the following desirable property when $n = 2$ and $n = 3$: $\det(A) \neq 0$ is equivalent to the fact that any square system of linear equations where A is the matrix of coefficients of the variables, and any right hand side $\mathbf{b}$ can be solved uniquely. Another way of saying this is that the $n \times n$ matrix A has rank n if and only if $\det(A) \neq 0$. A third way is A is invertible if and only if $\det(A) \neq 0$. We still have to do the same in the the general case. First an easy result, a corollary of Theorem 11.3.4.

Theorem 11.3.6. *For the identity matrix I,*

$$\det(I) = 1.$$

Proof. There is only one non-zero term in the sum (11.3), and it is the product of n 1s. □

Definition 11.3.7. For a square matrix A whose rows are $\mathbf{a}^1$, …, $\mathbf{a}^n$, the function $d(\mathbf{a}^1, \ldots, \mathbf{a}^n) = \det(A)$.

Theorem 11.3.8. *Hold fixed all the rows of A except the i-th one, say, so writing the i-th row as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Then*

$$d(\mathbf{a}^1, \ldots, \mathbf{a}^{i-1}, \mathbf{x}, \mathbf{a}^{i+1}, \ldots, \mathbf{a}^n)$$

is a linear function of $\mathbf{x}$. So for any scalar c,

$$d(\mathbf{a}^1, \ldots, \mathbf{a}^{i-1}, c\mathbf{x}, \ldots, \mathbf{a}^n) = cd(\mathbf{a}^1, \ldots, \mathbf{a}^{i-1}, \mathbf{x}, \mathbf{a}^{i+1}, \ldots, \mathbf{a}^n)$$

and

$$d(\ldots, \mathbf{a}^{i-1}, \mathbf{x}+\mathbf{y}, \mathbf{a}^{i+1}, \ldots) = d(\ldots, \mathbf{a}^{i-1}, \mathbf{x}, \mathbf{a}^{i+1}, \ldots) + d(\ldots, \mathbf{a}^{i-1}, \mathbf{y}, \mathbf{a}^{i+1}, \ldots).$$

Finally $\det(A) = 0$ *if a row of A consists entirely of* 0*s.*

Proof. This is obvious once you notice that in any term of (11.3) there is exactly one term belonging to the i-th row. □

Here is the important point to establish. It is the reason for introducing the sign ε of a permutation.

First

Definition 11.3.9. If τ is a permutation on $[1,\dots,n]$ and A is a matrix with n rows, let $A^\tau = (a^\tau_{ij})$ be the matrix with the rows permuted by τ, meaning that the i-th row of A becomes the $\tau(i)$-th row of A^τ, or, what is the same, k-th row of A^τ is the $\tau^{-1}(k)$-th row of A. Thus if we write $A^\tau = (a^\tau_{ij})$, we have $a_{ij} = a^\tau_{\tau(i),j}$, or $a^\tau_{k,j} = a_{\tau^{-1}(k),j}$.

Using the definition of P^τ of the previous section,

$$A^\tau = P^\tau A.$$

.

Theorem 11.3.10. *Then*

$$\det(A^\tau) = \varepsilon(\tau)\det(A).$$

Proof. Write $A^\tau = (a^\tau_{ij})$, so that by definition $a^\tau_{ij} = a_{\tau^{-1}(i),j}$. The determinant we want to compute is

$$\det(A^\tau) = \sum_{\sigma\in\mathscr{S}_n} \varepsilon(\sigma)\prod_{i=1}^n a^\tau_{i,\sigma(i)}.$$

By definition $a^\tau_{i,\sigma(i)} = a_{\tau^{-1}(i),\sigma(i)}$, so

$$\varepsilon(\sigma)\prod_{i=1}^n a^\tau_{i,\sigma(i)} = \varepsilon(\sigma)\prod_{i=1}^n a_{\tau^{-1}(i),\sigma(i)}. \tag{11.4}$$

Switching to the summation variable $j = \tau^{-1}(i)$, we can write (11.4) as

$$\varepsilon(\sigma)\prod_{j=1}^n a_{j,\sigma\tau(j)}.$$

Next we change the summation variable σ that enumerates all the permutations to $\upsilon = \sigma\tau$ which also enumerates all the permutations as σ varies, remembering that τ is fixed. Since $\sigma = \upsilon\tau^{-1}$,

$$\det(A^\tau) = \sum_{\upsilon\in\mathscr{S}_n} \varepsilon(\upsilon\tau^{-1})\prod_{j=1}^n a_{j,\upsilon(j)} = \varepsilon(\tau^{-1}) \sum_{\upsilon\in\mathscr{S}_n} \varepsilon(\upsilon)\prod_{j=1}^n a_{j,\upsilon(j)} = \varepsilon(\tau)\det(A).$$

We used the fact that $\varepsilon(\upsilon\tau^{-1}) = \varepsilon(\upsilon)\varepsilon(\tau^{-1})$ by Theorem 11.2.18, to pull the constant factor $\varepsilon(\tau^{-1})$ out of each term in the sum. We also used $\varepsilon(\tau^{-1}) = \varepsilon(\tau)$, which is true for any permutation τ. □

From these theorems we get the corollaries we are interested in.

Corollary 11.3.11. *If the square matrix A has two equal rows, then* $\det(A) = 0$.

Proof. Let τ be the transposition that permutes the two equal rows. On one hand $\varepsilon(\tau) = -1$, so $\det(A^\tau) = -\det(A)$ by Theorem 11.3.10. On the other hand $A^\tau = A$. so obviously $\det(A^\tau) = \det(A)$. The only way this can be reconciled is if $\det(A) = 0$. □

Corollary 11.3.12. *Here is how the determinant of a square matrix A is transformed under the elementary row operations of Definition 2.5.2.*

1. *If a row of A is multiplied by the non-zero scalar c, then the determinant of the new matrix is* $c\det(A)$
2. *If two rows of A are interchanged, then the determinant is multiplied by* -1.
3. *If you add to a row of A a multiple of a different row, then the determinant does not change.*

Proof. (1) follows from Theorem 11.3.8.
(2) is an immediate corollary of Theorem 11.3.10.
(3) uses linearity followed by Corollary 11.3.11:

$$\begin{aligned} d(\dots,\mathbf{a}^i + c\mathbf{a}^j,\dots,\mathbf{a}^j,\dots) &= d(\dots,\mathbf{a}^i,\dots,\mathbf{a}^j,\dots) + d(\dots,c\mathbf{a}^j,\dots,\mathbf{a}^j,\dots) \\ &= d(\dots,\mathbf{a}^i,\dots,\mathbf{a}^j,\dots) + cd(\dots,\mathbf{a}^j,\dots,\mathbf{a}^j,\dots) \\ &= d(\dots,\mathbf{a}^i,\dots,\mathbf{a}^j,\dots). \end{aligned}$$

□

This is important because it shows that if the determinant of a matrix is non-zero, then the determinant of any row equivalent matrix is also non-zero.

Corollary 11.3.13. *A square matrix A of size n has rank less than n if and only if its determinant is* 0.

Proof. By using row operations, which only multiply the determinant by a non-zero constant, by Theorem 2.8.11 we can modify A to

- a matrix with a row of 0s;
- or to the identity matrix I.

In the first case by Theorem 11.3.8, $\det(A) = 0$; in the second case, because we know $\det(I) = 1$, the determinant is non-zero. □

11.4 Properties of the Determinant

Now determinant results with a different flavor.

Theorem 11.4.1. *The determinant of a square matrix is equal to that of its transpose:* $\det(A^t) = \det(A)$. *Similarly* $\det(A^*) = \det(A)$.

Proof. The beautiful proof is a simple change of variables, just like the proof of Theorem 11.3.10. Write the entries of the transpose $A^t as (a^t_{ij})$. Then by definition

$$\det(A^t) = \sum_{\sigma \in \mathscr{S}_n} \varepsilon(\sigma) \prod_{i=1}^{n} a^t_{i,\sigma(i)}.$$

Because A^t is the transpose of A, we can write this in terms of the entries of A.

$$\det(A^t) = \sum_{\sigma \in \mathscr{S}_n} \varepsilon(\sigma) \prod_{i=1}^{n} a_{\sigma(i),i}. \tag{11.5}$$

Make the change of summation variable $\upsilon = \sigma^{-1}$. As we have already noted $\varepsilon(\upsilon) = \varepsilon(\sigma)$. As σ runs over all permutations, so does its inverse υ. Finally, instead of using i as the variable, use $j = \upsilon(i)$. Then (11.5) is rewritten

$$\det(A^t) = \sum_{\upsilon \in \mathscr{S}_n} \varepsilon(\upsilon) \prod_{i=j}^{n} a_{j,\upsilon(j)} = \det(A).$$

The second statement is left to you. □

The most important result concerning determinants is

Theorem 11.4.2. *If A and B are two square matrices of size n, then* $\det(AB) = \det(A)\det(B)$. *Therefore* $\det(AB) = \det(BA)$.

Proof. This is an exercise in matrix multiplication, followed by a use of the properties of determinants given in Theorems 11.3.8 and 11.3.10. Write C for the product AB. Then as noted in (2.8), the rows of C can be written in terms of the rows of B as

$$\mathbf{c}^i = a_{i1}\mathbf{b}^1 + \cdots + a_{in}\mathbf{b}^n = \sum_{j=1}^{n} a_{ij}\mathbf{b}^j. \tag{11.6}$$

We need to compute $\det(C) = d(\mathbf{c}^1, \ldots, \mathbf{c}^n)$. We replace each $\mathbf{c}^i$ by its expression (11.6) in terms of the entries of A and B and use the linearity of d in each row. To start, using the linearity, for each i we get:

$$d(\mathbf{c}^1, \ldots, \mathbf{c}^{i-1}, \sum_{j=1}^{n} a_{ij}\mathbf{b}^j, \mathbf{c}^{i+1}, \ldots, \mathbf{c}^n) = \sum_{j=1}^{n} a_{ij} d(\mathbf{c}^1, \ldots, \mathbf{c}^{i-1}, \mathbf{b}^j, \mathbf{c}^{i+1}, \ldots, \mathbf{c}^n).$$

Key remark: when you get $\mathbf{b}^j$ in the slot for the i-th row entry of the determinant, a coefficient a_{ij} gets pulled out as shown above.

When we do this for all the rows of C, we get a sum of terms with certain coefficients: the product of n of the a_{ij} multiplied by determinants of matrices where each entry consists of a *different* row of B. We know that we only have to consider the case of distinct rows, because whenever a determinant of rows has two equal rows, it is 0. So for each term in the sum, the order of the $\mathbf{b}$ is described by a permutation of $\{1, \ldots, n\}$, call it σ. So we write

$$d(\mathbf{b}^{\sigma(1)}, \mathbf{b}^{\sigma(2)}, \ldots, \mathbf{b}^{\sigma(n)}).$$

By Theorem 11.3.10 , this is $\varepsilon(\sigma)\det(B)$. What is the coefficient of this term? By the key remark it is $a_{1,\sigma(1)}a_{2,\sigma(2)}\ldots a_{n,\sigma(n)}$. So we have shown that

$$\det(C) = \det(B)\big(\sum_{\sigma\in\mathscr{S}_n} \varepsilon(\sigma)\prod_{i=1}^{n} a_{i,\sigma(i)}\big).$$

By (11.3) this is $\det(B)\det(A)$, which is the same thing as $\det(A)\det(B)$, since these are just scalars. □

By Corollary 11.3.13, a square matrix A has an inverse if and only if $\det(A) \neq 0$.

Corollary 11.4.3. *If the square matrix A has an inverse, written A^{-1} as usual, then*

$$\det(A^{-1}) = \frac{1}{\det(A)}.$$

Proof. This is immediate, since $AA^{-1} = I$ and $\det(I) = 1$. □

The following corollary is very important because it says that the determinant of a square matrix A is an invariant of the similarity class of A, so that it is an invariant of the linear operator represented by A.

Corollary 11.4.4. *For any invertible matrix C,* $\det(CAC^{-1}) = \det(A)$

Proof.

$$\begin{aligned}\det(CAC^{-1}) = \det(C)\det(AC^{-1}) &= \det(AC^{-1})\det(C)\\ &= \det(A)\det(C^{-1})\det(C) = \det(A)\det(I) = \det(A)\end{aligned}$$

□

We have already noted how to compute the determinant of a block triangular matrix in Theorem 11.3.5. This becomes easy to prove using Theorem 11.4.2. We only do a simple case.

Assume the matrix A is written in block-diagonal form (see §2.9)

$$A = \begin{pmatrix} B & 0_{rs} \\ 0_{sr} & C \end{pmatrix}$$

where B and C are square matrix of size r and s respectively, and the other two are matrices of the marked size.

Note that

$$A = \begin{pmatrix} B & 0_{rs} \\ 0_{sr} & I_s \end{pmatrix}\begin{pmatrix} I_r & 0_{rs} \\ 0_{sr} & C \end{pmatrix}$$

The determinants of the matrices of the right are easily seen to be $\det(B)$ and $\det(C)$, so that $\det(A) = \det(B)\det(C)$ by Theorem 11.4.2.

If the matrix A is written in block-diagonal form as

$$A = \begin{pmatrix} B & D \\ 0_{sr} & C \end{pmatrix}$$

where B and C are square matrix of size r and s respectively, and D is an arbitrary matrix of size $r \times s$. Assume $s \geq r$ and proceed as follows. If C has maximum rank, we can reduce to the previous corollary by doing row operations involving the last s rows on the first r ones. Note that these row operations do not modify b, so we are done. If B does not have maximum rank, then $\det(B) = 0$ and $\det(A) = 0$ so we get 0 on both sides. If $s < r$ do column operations instead.

Corollary 11.4.5. *In Example 2.5.1, we looked at an invertible matrix A that could be factored as a lower triangular matrix times an upper triangular matrix: $A = LU$. This implies that both L and U are invertible, so that their diagonal elements are all non-zero.*

11.5 The Laplace Expansion

Our next result is the Laplace expansion for the matrix. We need some preliminary definitions. In §2.4, we defined and gave notation for a submatrix of a matrix A. Here we need a special kind of submatrix for a square matrix A of size n, for which we use a simpler notation.

Definition 11.5.1. Write A_{ij} for the square submatrix of size $n-1$ from which the i-th row and the j-column of A have been removed. This submatrix has a determinant, called the *minor* m_{ij}. It is more common to consider the *cofactor* $\hat{m}_{ij}$, which is just $(-1)^{i+j} m_{ij}$.

Then

Theorem 11.5.2 (Laplace Expansion). *For the square matrix A of size n, and for any i,*

$$\det(A) = \sum_{i=1}^{n} a_{ij} \hat{m}_{ij}.$$

Similarly for any j,

$$\det(A) = \sum_{j=1}^{n} a_{ij} \hat{m}_{ij}.$$

Notice that only the index in the summation changes. These two expansions are called the expansions along the j-column and the expansion along the i-th row, respectively.

Proof. If is enough to prove this in one of the two cases. We will prove the second. Each term in the formula for the determinant contains exactly one term from the i-tu row, i.e. a term a_{ij}. So the determinant of A can be written

$$\Big(\sum_{\sigma\in\mathscr{S}_{i1}}\varepsilon(\sigma)a_{2\sigma(2)}\dots a_{2\sigma(n)}\Big)a_{i1}+\dots \\ +\Big(\sum_{\sigma\in\mathscr{S}_{ij}}\varepsilon(\sigma)a_{2\sigma(2)}\dots a_{2\sigma(n)}\Big)a_{ij}+\dots \\ +\Big(\sum_{\sigma\in\mathscr{S}_{in}}\varepsilon(\sigma)a_{2\sigma(2)}\dots a_{2\sigma(n)}\Big)a_{in} \quad (11.7)$$

where $\mathscr{S}_{ik}$ is the subset of permutations sending i to k. Notice that the n sets $\mathscr{S}_{ik}$, $1\le k\le n$, are disjoint and each has $(n-1)!$ elements, so we get all permutations in $\mathscr{S}$ in this way.

Each $\sigma\in\mathscr{S}_{ij}$ yields a permutation τ on $\{1,2,\dots,n-1\}$ excluding the integer i since it always is mapped to j. Here is the rule:

$$\tau(k)=\begin{cases}\sigma(k), & \text{if } k<i \text{ and } \sigma(k)<j;\\ \sigma(k)-1, & \text{if } k<i \text{ and } \sigma(k)>j;\\ \sigma(k+1), & \text{if } k\ge i \text{ and } \sigma(k+1)<j;\\ \sigma(k+1)-1, & \text{if } k\ge i \text{ and } \sigma(k+1)>j.\end{cases} \quad (11.8)$$

To understand what this means look at the following matrix of size 4 and suppose we are looking at a permutation taking 2 to 3. Then to find the 'residual' permutation on $\{1,2,3\}$ remove the second row and the third column from A:

$$\begin{pmatrix} a_{11} & a_{12} & \mathbf{a}_{13} & a_{14}\\ \mathbf{a}_{21} & \mathbf{a}_{22} & \mathbf{a}_{23} & \mathbf{a}_{24}\\ a_{31} & a_{32} & \mathbf{a}_{33} & a_{34}\\ a_{41} & a_{42} & \mathbf{a}_{43} & a_{44}\end{pmatrix} \quad (11.9)$$

The row and column in bold split the remaining part of the matrix in four regions corresponding to the four cases of (11.8), which explains how to renumber the entries to get a permutation on $\{1,2,\dots,n-1\}$. Suppose that $\sigma\in\mathscr{S}_{23}$ is

$$\begin{vmatrix}1 & \mathbf{2} & 3 & 4\\ 2 & \mathbf{3} & 1 & 4\end{vmatrix}$$

Since $i=2$ and $j=3$, the rule of (11.8) tells us that the corresponding τ is

$$\begin{vmatrix}1 & 2 & 3\\ 2 & 1 & 3\end{vmatrix}$$

We now need to compare $\varepsilon(\sigma)$ to $\varepsilon(\tau)$. The argument that follows is completely elementary, and is best understood by referring to a geometric diagram. Consider for example the diagram based on the matrix in (11.9) and a specific permutation σ such that $\sigma(2)=3$, for example the permutation

$$\begin{vmatrix}1 & 2 & 3 & 4\\ 2 & 3 & 4 & 1\end{vmatrix}$$

Then this permutation (a 4-cycle, by the way) gives the term $\pm a_{12}a_{23}a_{34}a_{41}$ in the determinant, so mark them in bold in the matrix:

$$\begin{pmatrix} * & a_{12} & \bullet & * \\ \bullet & \bullet & \mathbf{a}_{23} & \bullet \\ * & * & \bullet & a_{34} \\ a_{41} & * & \bullet & * \end{pmatrix} \tag{11.10}$$

How do you detect a reversal in a permutation from this diagram? When there are two entries so that the left one is below the right one. So, in the diagram above there are three reversals:

$$(a_{41},a_{12}),(a_{41},a_{23}),(a_{41},a_{34})$$

If you remove the i-th row and the j-the column all the reversals containing a_{ij} disappear. In our example we are left with

$$(a_{41},a_{12}),(a_{41},a_{34})$$

These reversals are the same as those of the permutation τ associated to σ. In our example

$$\begin{pmatrix} * & a_{12} & * \\ * & * & a_{23} \\ a_{31} & * & * \end{pmatrix}$$

Notice the renumbering, which derives from the rule (11.8). All the reversals that do not involve the removed element a_{23} survive: the formal proof is left to you.

So to understand the relationship between $\varepsilon(\sigma)$ and $\varepsilon(\tau)$ we need only consider the reversals that involve a_{ij}. If we number the quadrants clockwise in the matrix created by the i-th row and the j-th column by

$$\begin{cases} Q_1 = (k,l) & \text{such that } k < i \text{ and } l < j; \\ Q_2 = (k,l) & \text{such that } k < i \text{ and } l > j; \\ Q_3 = (k,l) & \text{such that } k > i \text{ and } l > j; \\ Q_4 = (k,l) & \text{such that } k > i \text{ and } l < j. \end{cases} \tag{11.11}$$

For a given permutation σ let n_i be the number of $a_{k,\sigma(k)}$ in Q_i. Note that a_{ij} is in none of them, so that the total number is $n-1$. Because there is one of these in each row and each column (other than the i-th row and the j-th column), we have the four linear equations:

$$\begin{aligned} n_1 + n_2 &= i-1; \\ n_3 + n_4 &= n-i; \\ n_1 + n_4 &= j-1; \\ n_2 + n_3 &= n-j. \end{aligned}$$

Exercise 11.5.3. Show that the matrix associated to this linear system has rank 3.

Exercise 11.5.4. Determine the n_i for the example above.

Now the number of reversals involving $a_{i,\sigma(i)}$ is the number of elements in quadrants 2 and 4, so it is n_2+n_4. By the linear equations above , we see that $n_2-n_4 = i-j$, so n_2+n_4 has the same parity as $i+j$. This shows that $\varepsilon(\sigma) = (-1)^{i+j}\varepsilon(\tau)$. This shows that

$$\det(A) = \sum_{i=1}^{n}(-1)^{i+j}a_{ij}m_{ij}$$

and completes the proof. □

This gives the familiar checkerboard pattern for the signs used in each minor:

$$\begin{pmatrix} + & - & + \\ - & + & - \\ + & - & + \end{pmatrix}$$

Many examples here.

Example 11.5.5. Consider the $n \times n$ matrix, called the Vandermonde matrix:

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{n-1} & x_2^{n-1} & \dots & x_n^{n-1} \end{pmatrix} \tag{11.12}$$

We wish to compute its determinant. In §6.2, we considered the transpose of this matrix, which we also called the Vandermonde matrix, and showed that it is invertible when the scalars x_i are distinct. Therefore we already know its determinant is non-zero in that case. In fact we can compute the determinant explicitly.

So each column of the matrix consists in the first n powers of the scalar x_i, starting with the 0-th power. We will compute the determinant of this polynomial by induction on n. When $n = 2$ it is $(x_2 - x_1)$. When $n = 3$, we do some row operations. We start with

$$\begin{pmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ x_1^2 & x_2^2 & x_3^2 \end{pmatrix}$$

First subtract the second row multiplied by x_1 from the third to get

$$\begin{pmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ 0 & x_2^2 - x_1x_2 & x_3^2 - x_1x_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ 0 & x_2(x_2-x_1) & x_3(x_3-x_1) \end{pmatrix}$$

First subtract the first row multiplied by x_1 from the second to get

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & x_2 - x_1 & x_3 - x_1 \\ 0 & x_2(x_2 - x_1) & x_3(x_3 - x_1) \end{pmatrix}$$

So by the Laplace expansion along the first column, the determinant is that of

$$\begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ x_2(x_2 - x_1) & x_3(x_3 - x_1) \end{pmatrix}$$

We can factor $(x_2 - x_1)$ from the first column and $(x_3 - x_1)$ from the second column. So the determinant is

$$(x_2 - x_1)(x_3 - x_1)\begin{vmatrix} 1 & 1 \\ x_2 & x_3 \end{vmatrix} = (x_2 - x_1)(x_3 - x_1)(x_3 - x_1)$$

This suggests that the general answer in the $n \times n$ case is

$$\prod_{i>j}(x_i - x_j) \tag{11.13}$$

so that there are $\frac{(n(n-1)}{2}$ factors. Assume this is true in the case $(n-1)$. Then just do the same row operations we did in the 3×3 case to get the Vandermonde determinant for $n-1$: subtract x_1 times the next-to-last row from the last row, and work backwards. Use the Laplace expansion as before to reduce to the $n-1$ Vandermonde matrix. We are done.

Notice that the determinant of the Vandermonde matrix vanishes when $x_i = x_j$, since the matrix then has two equal columns. This is confirmed by our computation. We could compute the determinant by using this remark: it must be a product of factors $(x_i - x_j)$, for $i \neq j$. So there are $\frac{(n(n-1)}{2}$ factors needed, just as in (11.13). So up to a constraint factor, we have the right answer.

Exercise 11.5.6. Multiply (11.12) by its transpose, and compute the determinant in terms of sums of powers of the x_i.

11.6 Cramer's Rule

We now prove one of the most famous theorems in linear algebra, proved very early in its history by Cramer.

Theorem 11.6.1 (Cramer's Rule). *Consider the system of linear equations*

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

where A is a square matrix of size n with $\det(A) \neq 0$*. Then writing as usual* $\mathbf{a}_1, \ldots,$ $\mathbf{a}_n$ *for the columns of A, we get for the entries of the unique solution* $\mathbf{x}$ *of this system*

$$x_j = \frac{d(\mathbf{a}_1, \ldots, \mathbf{a}_{j-1}, \mathbf{b}, \mathbf{a}_{j+1}, \ldots, \mathbf{a}_n)}{\det(A)}.$$

Proof. We can give a simple conceptual proof of this. Rewrite the equation as

$$\mathbf{b} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_b\mathbf{a}_n \tag{11.14}$$

Consider the determinant

$$d(\mathbf{a}_1, \ldots, \mathbf{a}_{j-1}, \mathbf{b}, \mathbf{a}_{j+1}, \ldots, \mathbf{a}_n)$$

where we have simply replaced the j-th column of A by $\mathbf{b}$. Now expand $\mathbf{b}$ by (11.14) and use the linearity of the determinant in its columns and the fact that the determinant is 0 when two of its columns are multiples of each other to get

$$d(\mathbf{a}_1, \ldots, \mathbf{a}_{j-1}, \mathbf{b}, \mathbf{a}_{j+1}, \ldots, \mathbf{a}_n) = x_j d(\mathbf{a}_1, \ldots, \mathbf{a}_{j-1}, \mathbf{a}_j, \mathbf{a}_{j+1}, \ldots, \mathbf{a}_n).$$

The right hand side is just $x_j \det(A)$, so just divide by $\det(A)$ to get the desired result. □

Remarks about how this is not useful for computation: always solve using Gaussian elimination. So from a modern point of view this result is a curiosity. Still, it is nice that there is such a beautiful and simple formula.

A 2×2 and 3×3 example here.

11.7 The Adjugate Matrix

In this optional section, given a square matrix A of size n, consider the square matrix

$$\hat{A} = (\hat{m}_{ij}),$$

where the entries of $\hat{A}$ are the cofactors (see Definition 11.5.1) of the matrix A. So $\hat{A}$ also has size n. Then the *adjugate matrix* of A is the transpose of $\hat{A}$. It is written $\operatorname{adj} A$.

This definition allows us to rewrite all the Laplace expansions as one matrix multiplication.

Corollary 11.7.1.

$$A \operatorname{adj} A = (\operatorname{adj} A)A = (\det A)I.$$

Here I is the identity matrix of size n.

Proof. For the terms on the diagonal, notice that this result just expresses all the Laplace expansions as one. So the only issue is to show that the off-diagonal terms are zero. □

11.8 The Cauchy-Binet Theorem

Now assume $C = AB$, where C is an $m \times m$ matrix, A a $m \times n$ matrix and B of course a $n \times m$ matrix. Then by (2.7)

$$\mathbf{c}_k = b_{1k}\mathbf{a}_1 + \cdots + b_{nk}\mathbf{a}_n = \sum_{j=1}^{n} b_{jk}\mathbf{a}_j. \tag{11.15}$$

Then exactly as in the proof of Theorem 11.4.2, we get

$$\det(C) = \sum_{k_1,\ldots,k_m=1}^{n} b_{k_{11}} b_{k_{22}} \ldots b_{k_{mm}} d(\mathbf{a}_{k_1}, \mathbf{a}_{k_2}, \ldots, \mathbf{a}_{k_m}). \tag{11.16}$$

Here the summation means that you sum over each of the k_i.

Let's make sure the right hand side makes sense. Each column of A is a m-vector, so in the determinant we take m columns of A, since we need a square matrix.

All the terms in the sum are 0 unless A has at least m columns, since otherwise we have to repeat a column, in which case the determinant is 0. So we get a first interesting result:

Theorem 11.8.1. *If the square matrix C of size m can be written as the product of a $m \times n$ matrix A by a $n \times m$ matrix B with $n < m$, then* $\det(C) = 0$.

Theorem 11.4.2 is the case $n = m$. What happens if $n > m$?

Consider the terms in the sum on the right hand side of (11.16). As always, if the k_i are not distinct, the corresponding determinant vanishes.

For any collection K of m integers k_i, $1 \le k_1 < k_2 < \cdots < k_m \le n$, let $\mathscr{S}_K$ be the collection of all one-to-one mappings from $[1,\ldots,m]$ to $(k_1,\ldots,k_m)$ Then we can rewrite (11.16) as

$$\det(C) = \sum_{K} \sum_{\sigma \in \mathscr{S}_K} b_{k_{11}} b_{k_{22}} \ldots b_{k_{mm}} \varepsilon(\sigma)(d(\mathbf{a}_{k_1}, \mathbf{a}_{k_2}, \ldots, \mathbf{a}_{k_m}). \tag{11.17}$$

Now denote by A_K the square submatrix of size m of A using the columns $\mathbf{a}_k$ for $k \in K$, and B^K the square submatrix of size m of B using the rows $\mathbf{b}^k$, for $k \in K$.

Theorem 11.8.2 (Cauchy-Binet Theorem).

$$\det C = \sum_{K} \det(A_K) \det(B^K),$$

where the sum of over all the subsets of $[1,\ldots,n]$ consisting of m elements.

This follows immediately from (11.17).

Example 11.8.3. Let A the 2×3 matrix (a_{ij}) and B the 3×2 matrix (b_{jk}). Then their product C is the 2×2 matrix (c_{ik}),

$$C = \begin{pmatrix} a_{11}b_{11}+a_{12}b_{21}+a_{13}b_{31} & a_{11}b_{12}+a_{12}b_{22}+a_{13}b_{32} \\ a_{21}b_{11}+a_{22}b_{21}+a_{23}b_{31} & a_{21}b_{12}+a_{22}b_{22}+a_{23}b_{32} \end{pmatrix}$$

There are three sets K: $\{1,2\}$, $\{1,3\}$ and $\{2,3\}$. So for example

$$A_{\{1,3\}} = \begin{pmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{pmatrix} \text{ and } B^{\{1,3\}} = \begin{pmatrix} b_{11} & b_{12} \\ b_{31} & b_{32} \end{pmatrix}$$

so

$$\det(A_{\{1,3\}}) = a_{11}a_{23} - a_{13}a_{21} \quad \text{and } \det(B^{\{1,3\}}) = b_{11}b_{32} - b_{12}b_{31}.$$

So one of the three terms in the sum giving $\det(C)$ is

$$\det(A_{\{1,3\}})\det(B^{\{1,3\}}) = (a_{11}a_{23} - a_{13}a_{21})(b_{11}b_{32} - b_{12}b_{31})$$

Thus you can explicitly compute the expressions on both sides of the Cauchy-Binet formula, and check that they match.

We now generalize the Cachy-Binet Theorem. We use the notation of §2.4 to describe the square submatrices of a square matrix C that is the product AB of two square matrices A and B: see (2.13).

Theorem 11.8.4. *Assume $C = AB$, where all three matrices are square of size n. Let I and J are fixed subsets of $[1,\dots,n]$ consisting of k elements each. Then*

$$\det C_J^I = \sum_K \det(A_K^I)\det(B_J^K),$$

where the sum is over all subsets K of $[1,\dots,n]$ consisting of k elements.

Proof. We simply reduce to the previous theorem. The key remark follows the very definition 2.2.3 of matrix multiplication: see (2.7) and (2.8). Thus every entry in C_J^I can be computed just from rows of A with index in J and the columns in B with index in I. In other words

$$C_J^I = A^J B_I.$$

You should check this one entry c_{ij} of C at a time. So now just apply Theorem 11.8.2 to this triple of matrices to get

$$\det(C_J^I) = \sum_K \det(A_K^J)\det(B_I^K).$$

□

We will use this result in Chapter 12.

11.9 Gaussian Elimination via Determinants

We now want to look back at Gaussian elimination using determinants, especially the minors of a matrix. We show how this works for a square system. We follow the presentation and largely the notation of Gantmacher [8], chapter II, §1.

So assume we have a system of n equations in n variables:

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

Assume A has rank r. This means there is a $r \times r$ submatrix of A which has rank r. Then we can multiply A on the left by a suitable permutation matrix P^σ, and on the right by a suitable permutation matrix Q^τ so that the leaning principal submatrix of size r of $P^\sigma A Q^\tau$ has rank r. This follows from Theorem 11.2.12. This just means that we reorder the equations (using P^σ) and the variables (using Q^τ) to move the submatrix into the right position. In fact, by a sufficiently careful reordering (which, unfortunately we can only determine during Gaussian elimination) we may assume that all the leading principal submatrices of size k of $P^\sigma A Q^\tau$ have maximal rank. The system of equations is now $P^\sigma A Q^\tau = P^\sigma \mathbf{b} Q^\tau$. This system is equivalent to the previous system, in the sense of Chapter 1. To simplify notation we continue to write it as $A\mathbf{x} = \mathbf{b}$.

Remark 11.9.1. What we do next is to show that with these hypotheses $A = LU$, where L is lower triangular, and D is upper triangular. In other words we only use the elementary transformation of Definition 2.8.1 of type 1 and of type 3, but where the matrix $E_{rs}(c)$ is lower triangular, so $r > s$, so the product of all the elementary matrices used is itself lower triangular. As we already noted in Example 2.5.1, it is easy to solve the equation $LU\mathbf{x} = \mathbf{b}$.

By assumption the leading principal minors D_k of A, $1 \leq k \leq r$, are all non-zero, while $D_k = 0$ when $k > r$. The last inequalities are implied by the rank of A being r. In particular $D_1 = a_{11} \neq 0$. Then:

Definition 11.9.2. a_{11} is the first *pivot* d_1 of A.

Let E_1 be the elementary $n \times n$ matrix:

$$E_1 = \begin{pmatrix} 1 & 0 \dots 0 \\ -\frac{a_{21}}{a_{11}} & 1 \dots 0 \\ \vdots & \vdots \ddots \vdots \\ -\frac{a_{n1}}{a_{11}} & 0 \dots 1 \end{pmatrix}$$

So E_1 is invertible with determinant equal to one, since it is lower triangular with ones on the diagonal. We write $A^{(1)}$ for the product matrix $E_1 A$. By construction:

$$A^{(1)} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & \dots & a_{3n}^{(1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & a_{n3}^{(1)} & \dots & a_{nn}^{(1)} \end{pmatrix}$$

where the matrix entries without suffix are the original entries of A, while those with an upper (1) are by definition the entries of $A^{(1)}$.

Proposition 11.9.3. *We compute the second diagonal element of the matrix $A^{(1)}$, and show it is non-zero, under the assumption that the rank of A is at least 2, so it will serve as our second pivot d_2:*

$$a_{22}^{(1)} = a_{22} - \frac{a_{12}a_{21}}{a_{11}} = \frac{D_2}{D_1}. \tag{11.18}$$

Proof. Because $A^{(1)}$ was obtained from A by adding the first row of A multiplied by a constant, the minors that contain that row (in particular the leading principal minors) do not change when one passes from A to $A^{(1)}$, by Corollary 11.3.11. On the other hand, the second leading principal minor of $A^{(1)}$ is simply $a_{11}a_{22}^{(1)}$, because that principal matrix is triangular. So $a_{11}a_{22}^{(1)} = D_2$, and since $D_1 = a_{11}$, this is what we found by direct computation. This computation establishes the result, since by hypothesis, the leading principal minor D_2 is non-zero. □

This simple but important argument will generalize as we create more zeroes by Gaussian elimination.

Exercise 11.9.4. Write down the definition of E_2 using that of E_1 as a model.

We write $A^{(2)}$ for the matrix $E_2A^{(1)}$. By construction:

$$A^{(2)} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \dots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \dots & a_{nn}^{(2)} \end{pmatrix}$$

We claim, as before, that if $3 \leq r$, where r is the rank of A, then $a_{33}^{(2)} \neq 0$, because

$$a_{33}^{(2)} = \frac{D_3}{D_2}.$$

by the same argument as in (11.18). So this gives us the third pivot d_3.

So if $2 < r$ we can continue the elimination process until we reach the rank r of the matrix. For simplicity, first consider the case of maximum rank $r = n$. At each step we get a new non-zero pivot

$$d_k = a_{kk}^{(k-1)} = \frac{D_k}{D_{k-1}}.$$

Thus in the end we get the upper triangular matrix:

$$A^{(n-1)} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1,n-1} & a_{1n} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2,n-1}^{(1)} & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \dots & a_{3,n-1}^{(2)} & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{n-1,n-1}^{(n-2)} & a_{n-1,n}^{(n-2)} \\ 0 & 0 & 0 & \dots & 0 & a_{nn}^{(n-1)} \end{pmatrix}$$

with an accompanying lower triangular matrix $E = E_{n-1}E_{n-2}\cdots E_2E_1$. By construction $A^{(n-1)} = EA$.

Now let us consider the more general case where A only has rank r, that can be less than n. Then by left multiplication by invertible matrices we get, after $r-1$ steps:

$$A^{(r-1)} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1r} & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2r}^{(1)} & \dots & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \dots & a_{3r}^{(2)} & \dots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{r,r}^{(r-1)} & \dots & a_{r,n}^{(r-1)} \\ 0 & 0 & 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 & \dots & 0 \end{pmatrix}$$

together with an invertible matrix E so that $EA = A^{(r-1)}$. The non-zero pivots are $\frac{D_k}{D_{k-1}}$, $1 \le k \le r$, so their product is the determinant of the leading principal matrix of size r. This is the classic way of producing an upper-triangular matrix that is row equivalent to A. We finish solving by doing back substitution on the upper triangular $A^{(r-1)}$.

Symmetric Case

Now we assume that A is symmetric. This will allow us to do the back-substitution to make A diagonal in a conceptually simple way. . Since the matrix E used to make A upper-triangular is lower triangular, E^t is upper triangular. So $(EA)E^t$, the product of two upper triangular matrices, is upper triangular. But EAE^t is symmetric: just compute its transpose. The only symmetric upper triangular matrices are diagonal, so EAE^T is diagonal and we have achieved the goal of Gaussian elimination without any further computation. We record this special case as a theorem.

Theorem 11.9.5. *Assume A is a symmetric matrix of size n such that all its leading principal minors are non zero. Then Gaussian elimination can be accomplished by*

left multiplication by an invertible lower triangular matrix E of determinant 1. *The k-th diagonal element of the diagonal matrix* EAE^t *is* $d_k = \frac{D_k}{D_{k-1}}$, *where the* D_k, $1 \leq k \leq n$ *are the leading principal minors of A, and* $D_0 = 1$ *by convention.*

We now generalize this to matrices of smaller rank It can also be used to compute the signature of a quadratic form in many cases, as explained in [8], volume 1, p.302.

We make the simple but important remark: by definition, Gaussian elimination applied to symmetric matrices as above yields a matrix is the same congruence class as the original matrix.

Theorem 11.9.6. *A is an* $n \times n$ *symmetric matrix of rank r with non-zero leading principal minors* D_k, $1 \leq k \leq r$. *Then Gaussian elimination can be performed to produce zeroes below and to the right of the first r diagonal elements of the matrix. Denoting the pivots of A by* $d_k, 1 \leq k \leq n$, *we have*

$$d_k = \frac{D_k}{D_{k-1}} \text{ for } 1 \leq k \leq r,$$

where $D_0 = 1$ *by definition.*

Proof. After the first $k-1$ columns of A have been cleared by forward elimination, the k-th leading submatrix A_k is upper triangular with the first k pivots on the diagonal. So $D_k = \det(A_k) = \prod_{i=1}^{k} d_i$. Further Gaussian elimination does not modify A_k. Thus, if all leading principal minors of A are non-zero, then so are all the pivots, which means that Gaussian elimination can occur without row exchanges. □

11.10 Determinants and Volumes

We restrict the scalars to $\mathbb{R}$ and assume we are working in $\mathbb{R}^n$ equipped with the standard inner product.

First we compute a determinant by expansion by minors.

Example 11.10.1. If A is the matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

then the submatrices and their minors from the first row are:

$$A_{11} = \begin{pmatrix} 5 & 6 \\ 8 & 9 \end{pmatrix} \quad \text{and } m_{11} = -3;$$

$$A_{12} = \begin{pmatrix} 4 & 6 \\ 7 & 9 \end{pmatrix} \quad \text{and } m_{12} = -6;$$

and

$$A_{13} = \begin{pmatrix} 4 & 5 \\ 7 & 8 \end{pmatrix} \quad \text{and } m_{13} = -3.$$

Example 11.10.2. The determinant of the matrix A from Example 11.10.1, following our formula of expansion by minors, is given by:

$$1(5\cdot 9 - 6\cdot 8) - 2(4\cdot 9 - 6\cdot 7) + 3(4\cdot 8 - 5\cdot 7) = -3 + 12 - 9 = 0.$$

Exercise 11.10.3. Compute the determinant of

$$M = \begin{pmatrix} 1 & -2 & 0 & 0 \\ -3 & 2 & 0 & 0 \\ 0 & 0 & -1 & 3 \\ 0 & 7 & 2 & 1 \end{pmatrix} \text{ and } N = \begin{pmatrix} 1 & -2 & 0 & 0 \\ -3 & 2 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 1 & 1 & 2 & 1 \end{pmatrix}.$$

Definition 11.10.4. In $\mathbb{R}^n$ start with n vectors $\mathbf{v}^1, \ldots, \mathbf{v}^n$. Then the n-dimensional *parallelepiped* spanned by the origin and these n vectors is the set of linear combinations $\sum_{i=1}^n x_i \mathbf{v}_i$, $0 \le x_i \le 1$.

Example 11.10.5. When $n = 2$, you have the parallelogram with four vertices $\mathbf{0}$, $\mathbf{v}^1$, $\mathbf{v}^2$, and $\mathbf{v}^1 + \mathbf{v}^2$. When $n = 3$, you have the parallelepiped with eight vertices. The first four as when $n = 2$: $\mathbf{0}$, $\mathbf{v}^1$, $\mathbf{v}^2$, $\mathbf{v}^1 + \mathbf{v}^2$, to which one adds the sum of each with $\mathbf{v}^3$.

Corollary 11.10.6. *If V is the n-dimensional parallelepiped in $\mathbb{R}^n$ spanned by the origin and n vectors $\mathbf{v}^1, \ldots, \mathbf{v}^n$, then the n-volume of V is $|\det A|$, where A is the $n \times n$ matrix with columns the $\mathbf{v}^i$.*

When you enter the columns of A in a different order, the determinant either stays the same or is multiplied by -1, so the volume does not change. The sign of the determinant tells us whether the linear transformation T_A associated to multiplication by the matrix A is orientation-preserving $(+)$ or orientation-reversing $(-)$.

Example 11.10.7. Because the determinant of the matrix A from Example 11.10.1 is 0, as we saw in Example 11.10.2, A is not invertible. Indeed, its nullspace is one-dimensional, generated by the vector $(1, -2, 1)$. Take the cube in $\mathbb{R}^3$ with vertices the origin and the three unit vectors. Under the linear transformation T_A, what happens to this cube? In other words, what are the points $A\mathbf{x}$, where $\mathbf{x}$ is the column vector with all coordinates between 0 and 1.

Also see Minsky p. 416 and Franklin p. 153. Also Serre p. 120.

11.11 The Birkhoff-Koenig Theorem

Definition 11.11.1. Given a $n \times n$ matrix A, for each permutation σ of $[1, \ldots, n]$, we call the list of entries $a_{i,\sigma(i)}$ a *generalized diagonal* of A. Thus there are as many generalized diagonals as there are permutations, namely $n!$. We say that a generalized diagonal vanishes if $\prod_{i=1}^n a_{i,\sigma(i)} = 0$.

Theorem 11.11.2 (Birkhoff-Koenig Theorem). *All the generalized diagonals of a square matrix A of size n vanish if and only if there is a $r \times s$ submatrix of A with $r+s = n+1$ that is $0_{r,s}$.*

First note that the vanishing of all the generalized diagonals implies by definition that the determinant of A is 0.

Proof. First we assume that A has a $r \times s$ submatrix B of 0s. Let the rows of B be $i_1, \ldots, i_r$ and the columns of B be $j_1, \ldots, j_s$. Assume that the result is false, so that there is a generalized diagonal D that does not vanish. The in the rows $i_1, \ldots, i_r$, the entry of D must be outside $j_1, \ldots, j_s$. We have r elements to place, but only $n-s = n-(n+1-r) = r-1$ places to put them. Contradiction.

The other implication is harder. We assume that every generalized diagonal of A vanishes, and prove by induction that there is a $r \times s$ submatrix of A what is the zero matrix. There is nothing to do to start the induction at $n = 1$. So assume that the result is true for all square matrices of size less than n, and establish it for n. If A is the zero matrix, again there is nothing to prove. So we may assume that some entry $a_{ij} \neq 0$, so the result is true by induction. We call the submatrix with that row and column removed A_{ij} as in Definition 11.5.1. By induction A_{ij} has a $r \times s$ submatrix 0, with $r+s = n$. By permuting rows and columns of A to get a matrix A', we may assume that the $r \times s$ submatrix in the upper right hand corner of A' is the zero matrix. Thus in block notation

$$A' = \begin{pmatrix} B & 0 \\ C & D \end{pmatrix}$$

where B is a square matrix of size r and D a square matrix of size s. Assume that there is a generalized diagonal of D that does not vanish. Then extending this diagonal in all possible way to A', we see that all the diagonals of B vanish, so that by induction applied to B there is a $t \times u$ submatrix of B, with $t+u = r+1$ that is 0. By permuting the first $n-r$ rows of A' we get a matrix A'' whose first $n-r$ rows can be written in block notation as

$$E = \begin{pmatrix} B' & 0 & 0 \\ B'' & B''' & D' \end{pmatrix}$$

This submatrix of A visibly contains a $u \times (t+s)$ submatrix of 0s. Since $u+t+s = r+1+s = n+1$, the matrix A'' satisfies the conclusion. Now just undoing the permutations made to get A'', the result is true for A. Finally we need to consider the case where there is a generalized submatrix of D that does not vanish. That case is handled exactly the same way, and is left to you. □

Of course a determinant can be 0 even if not all the generalized diagonals are 0.

Chapter 12
The Characteristic Polynomial

Abstract We continue our work on understanding all linear operators L from a vector space V to itself. The fundamental approach is to find invariant subspace for L. An invariant subspace is a proper subspace $W \subset V$ such that $L(W) \subset W$. The primary decomposition theorem of Chapter C allowed use to do this in terms of the decomposition of the minimal polynomials into relatively prime factors. It is especially useful because it allows the decomposition of the space into a direct sum of invariant subspaces. We also fixed a vector $\mathbf{v}$ and looked at the subspace generated by applications of L to $\mathbf{v}$. This generates a subspace W that is invariant under L, and on which we get a minimal polynomial with degree equal to the dimension of W. This led us naturally to the notion of eigenvector: a non-zero vector such that the generated subspace has dimension 1. How to find the eigenvectors? We construct the characteristic polynomial of L, and show that it is always divisible by the minimal polynomial of L: this is the famous Cayley-Hamilton Theorem. We compute the characteristic polynomial in several important cases.

12.1 The Characteristic Polynomial

First some motivation for the introduction of the characteristic polynomial

Given a linear operator L on V, we know that an invariant subspace of dimension 1, spanned by a single vector, call it $\mathbf{v}$, satisfies $L(\mathbf{v}) = \lambda\mathbf{v}$, for some scalar λ. By Definition 10.1.3 the non-zero vector $\mathbf{v}$ is called an eigenvector of the linear transformation L and the scalar λ in the base field, is called the eigenvalue associated to $\mathbf{v}$.

We can also make this definition in terms of any matrix A representing L in some basis, and the coordinate vector of $\mathbf{v}$ in that basis, which we still write $\mathbf{v}$: $A\mathbf{v} = \lambda\mathbf{v}$.

The key remark is that $\mathbf{v}$ is an eigenvector with eigenvalue λ if $\mathbf{v}$ is in the nullspace of the operator $\lambda I_n - L$. As usual I_n is the identity operator, or, when needed, the $n \times n$ identity matrix. Passing to matrix notation, if A is the matrix of L

for a certain basis of V, then the matrix $\lambda I_n - A$ is not invertible, so that its determinant is 0.

We do not know what λ is, other than a scalar. So replacing λ by the variable x, to find the eigenvalues we need to solve the polynomial equation $\det(xI_n - A) = 0$. An eigenvalue of A is a root of this polynomial. Conversely any root λ of this polynomial is an eigenvalue. It is associated to any non-zero vector in the nullspace of $\lambda I_n - A$. Notice that the eigenvalues depend of the base field, since the roots of a polynomial depend on the field considered. For example, the polynomial $x^2 + 1$ has no roots in $\mathbb{R}$, but factors as $(x+\mathrm{i})(x-\mathrm{i})$ in $\mathbb{C}$.

Thus the polynomial $\det(xI_n - A)$ is quite important.

Definition 12.1.1. The *characteristic polynomial*[1] $p(x)$ of the $n \times n$ matrix A is

$$\det(xI_n - A)$$

where, as usual, I_n is the $n \times n$ identity matrix, and x is a variable.

As already noted the roots of the characteristic polynomial are the eigenvalues. This may depend of the field of coefficients: for example quadratic polynomials that are irreducible over $\mathbb{R}$ factor over $\mathbb{C}$.

Theorem 12.1.2. *The characteristic polynomial $p(x$ is a monic polynomial of degree n in x, which we write with alternating signs as*

$$p(x) = x^n - p_1 x^{n-1} + p_2 x^{n-2} + \cdots + (-1)^k p_k x^{n-k} + \cdots + (-1)^n p_n, \qquad (12.1)$$

where the p_i are in the base field F.

Proof. We use Definition 11.3.1 of the determinant. So $p(x)$ is a sum of a product of terms, one from each row and column of A, corresponding to a permutation of n elements. Note that the variable x only appears in the terms along the diagonal, and to the first power with coefficient 1. This implies that the degree of $p(x)$ is at most n. Furthermore the only term where n diagonal elements occur corresponds to the identity permutation. We get leading coefficient 1 since the product is

$$\prod_{i=1}^{n} (x - a_{ii}). \qquad (12.2)$$

Since (12.1) is just a monic polynomial of degree n, we are done. □

Theorem 12.1.3. *The characteristic polynomial of a upper or lower triangular matrix A of size n is $\prod_{i=1}^{n}(x - a_{ii})$. In particular, if A has r distinct scalars $\lambda_1, \ldots, \lambda_r$ along the diagonal, then A has at least r eigenvectors $\mathbf{v}_i$ with eigenvalues λ_i, $1 \le i \le r$. These eigenvectors are linearly independent.*

[1] Note that some authors (for example [28]) define the characteristic polynomial as $\det(A - xI_n)$, while others (in particular [8], [16], and [24] use the one given here. It is a simple exercise using properties of the determinant to show that the two possible definitions differ by a factor of $(-1)^n$.

Proof. Since the characteristic polynomial is a determinant, we simply apply Theorem 11.3.4. The last statement is just Theorem 10.1.5. □

An important property of the characteristic polynomial is that it is invariant under similarity, just like the determinant. Thus we may talk about the characteristic polynomial of a linear operator. In other words

Theorem 12.1.4. *The characteristic polynomial of A is the same as that of CAC^{-1}, for any invertible matrix C of size n.*

Proof. We need to compute $\det(xI - CAC^{-1})$. Since

$$xI - CAC^{-1} = C(xI - A)C^{-1}$$

this follows immediately from Corollary 11.4.4. □

This implies that all the coefficients p_i in (12.1) are invariant for similarity.

Theorem 12.1.5. *For any two square matrices A and B of size n, the characteristic polynomial of AB is equal to that of BA.*

Proof. If either one is invertible, let's say A, then $A(BA)A^{-1} = AB$, so the matrices AB and BA are similar: now just use Theorem 12.1.4.

If A is not invertible, replace A by $A - tI$ for $|t| > |\lambda_i|$, for any eigenvalue λ_i of A. Then $A - tI$ is invertible, so by the previous case the characteristic polynomial of $(A - tI)B$ is equal to that of $B(A - tI)$, so

$$\det(xI - (A - tI)B = \det(xI - B(A - tI).$$

Now fix x, so that we have an equality of polynomials of degree n in t that agree for an infinite number of values of t. Theorem 6.2.3 says that they are equal for all values of t including the value $t = 0$: in fact all we need is that they agree at $n + 1$ points, by using the computation of the determinant of the Vandermonde matrix in Example 11.5.5. So setting $t = 0$ we get $\det(xI - AB) = \det(xI - B(A)$ for all values of x, which is what we wanted to prove. □

In particular

Corollary 12.1.6. *The eigenvalues of AB are the same as those of BA, counted each with their multiplicity as roots of the corresponding characteristic polynomial.*

You may well ask why we chose to write 12.1 with alternating coefficients. It is because the coefficients then have an interesting interpretation. For example

Theorem 12.1.7. *The coefficient p_1 in* (12.1) *is the trace of A, namely:*

$$p_1 = \operatorname{tr} A = \sum_{i=1}^{n} a_{ii}.$$

The coefficient p_n in (12.1) *is the determinant of A.*

Proof. This is easy to prove using the expansion (11.3) of the determinant of $xI_n - A$ in terms of permutations. First we look at the coefficient of the term of degree 0. We get it by setting $x = 0$, so the constant term is $\det(-A) = (-1)^n \det(A) = (-1)^n p_n$. Finally we look at the term of degree $n-1$. It must come from a term with at least $n-1$ diagonal elements, to get a high enough degree. Since each term in the determinant expansion involves a term from each row and column, this forces the remaining term to be a diagonal term too. So in the product (12.2) we choose the term x from $n-1$ factors, and the term $-a_{ii}$ from the last one. This gives the result. □

Exercise 12.1.8. Compute the characteristic polynomials of the matrices A and B:

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix} \text{ and } B = \frac{1}{2}\begin{pmatrix} 7 & -3 \\ -1 & 1 \end{pmatrix}$$

and show that the determinant and the trace agree with the computation in (12.1).

Simple examples of the characteristic polynomial here.

Example 12.1.9. We compute the characteristic polynomial of the permutation matrix A from Example 10.4.5 by Laplace expansion. You will see that you get the minimal polynomial for A, which follows from the Cayley Hamilton Theorem we will prove soon. We will also do a more general computation in Theorem 12.6.1.

12.2 The Multiplicity of Eigenvalues

Associated to any eigenvalue λ of a linear operator L w can associate two numbers. First the multiplicity of λ as a root of the characteristic polynomial of L, which we call the algebraic multiplicity, and second the dimension of the nullspace of $L - \lambda I$ which we call the geometric multiplicity.

Then:

Theorem 12.2.1. *The algebraic multiplicity of any eigenvalue of L is greater than or equal to the geometric multiplicity.*

Proof. Assume that the geometric multiplicity of the eigenvector λ is g. So there are g linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_g$ that are in the nullspace of $L - \lambda I$, so they form a basis for the nullspace of $L - \lambda I$. Complete this basis to a basis of the vector space V on which L is acting. In this basis, L has a matrix A. The eigenvalues of L are those of A, by Theorem 12.1.4, so we need only work with A. By construction, in this basis the first g columns of A are 0 except for the entry λ is position (i,i). In particular the upper right hand corner of A of size g is λI. This means that A is block upper triangular. By Theorem 11.3.5, the determinant of $xI - A$ is the product of the determinants of the blocks, which implies that the characteristic polynomial has the root λ with multiplicity at least g, which is what we needed to prove. □

Now deal with generalized eigenvectors.

12.3 The Trace and the Determinant

In this section we consider the $n \times n$ matrix A of a linear operator. We may assume that we have chosen the most convenient basis , so if A is complex, we may assume A is triangular. Consider the characteristic polynomial of A. Over $\mathbb{C}$ it has n roots, its eigenvalues counted with multiplicity: $\lambda_1, \ldots, \lambda_n$. Since we may assume A is triangular, by Theorem we know that these values are the diagonal entries of A. Thus

Theorem 12.3.1. *The trace of A is the sum of the eigenvalues counted with multiplicities.*

If A has only real entries, and we want to consider it as the matrix of a real linear operator, when A does not have enough eigenvalues. We can however consider it as a complex matrix, in which case Theorem 12.3.1. As we know, the eigenvalues then are either real or occur in complex conjugate pairs. In either case we see that the sum of the eigenvalues is real. While we cannot diagonalize A over $\mathbb{R}$ we can block diagonalize it, with blocks of the form

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}, \quad b \neq 0,$$

along the diagonal. Over $\mathbb{C}$ this has eigenvalues $a \pm \mathrm{i}b$, so their sum is $2a$ as expected.

Now we turn to the determinant. Over $\mathbb{C}$, we may choose as before a triangular A, in which case we know that the determinant is the product of the diagonal entries, so

Theorem 12.3.2. *The determinant of A is the product of the eigenvalues counted with multiplicities.*

Again, if A is real, so that its eigenvalues occur in complex conjugate pairs, we see that its determinant is real, as it must be.

Now we go back to the polynomial techniques of Chapter 10. Let $f(x)$ be any polynomial. We want to compute the eigenvalues of $f(A)$.

Theorem 12.3.3 (Spectral Mapping Theorem). *Assume $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the $n \times n$ matrix, considered over $\mathbb{C}$. Then for any polynomial $f(x)$, the eigenvalues of $f(A)$ are $f(\lambda_1), \ldots, f(\lambda_n)$.*

Proof. The key point is that A is similar to an upper triangular matrix, i.e. $A = CTC^{-1}$. Then, since $A^k = CTC^{-1} \ldots CTC^{-1} = CT^kC_{-1}$, $f(A) = Cf(T)C^{-1}$. $f(T)$ is upper triangular, and its diagonal elements are $f(t_{ii}$, and therefore its eigenvalues. Since they are also the eigenvalues of $f(A)$, we are done. □

We can apply this result to the characteristic polynomial $p(x)$ of A.

Corollary 12.3.4. *All the eigenvalues of $p(A)$ are 0.*

Proof. Indeed, the theorem tells us that all the eigenvalues of $p(A)$ are of the form $p(\lambda)$, where λ is an eigenvalue of A. Since the eigenvalues of A are the roots of $p(x)$ we are done. □

In the next section we will improve this result.

12.4 The Cayley-Hamilton Theorem

We did something unusual when we formed the characteristic polynomial. We worked with a matrix whose coefficients are not in a field, but in the set of polynomials $F[x]$. It has all the properties of addition and multiplication of a field, but we cannot divide in it. Fortunately when we form a determinant, only addition and multiplication are needed, and, as we saw in §C.1, these behave in the expected way.

Here is the statement of the Cayley-Hamilton Theorem, using the terminology of §10.3.

Theorem 12.4.1 (Cayley-Hamilton). *If A is a square matrix of size n, and $p(x)$ its characteristic polynomial, then $p(x)$ vanishes on A, so $p(A) = 0$.*

Thus for any matrix of size n, the minimal polynomial has degree at most n, and divides the characteristic polynomial Recall that from the elementary linear algebra techniques used in Theorem 10.3.1, the best bound we were able to produce for the degree of the minimal polynomial is n^2, so the bound provided by the Cayley-Hamilton Theorem is much better. On the other hand, in Theorem 10.4.3 we showed that in some cases the minimal polynomial has degree n, so it must be the characteristic polynomial. Furthermore if we take for A a diagonal matrix where the diagonal entries are all distinct, it is easy to see that the minimal polynomial is the characteristic polynomial. In fact this is a good exercise. Thus the Cayley Hamilton Theorem is not completely unexpected.

Proof. Let $P_x = xI - A$ be the characteristic matrix of A. Its entries are polynomials of degree at most 1 in x. As in §11.7, we can form the adjugate matrix R_x of P_x. We can do this since that just requires forming minors of P_x, which are determinants, exactly as in the case of the characteristic polynomial. Each minor considered has degree at most $n-1$ in x, since it is the product of $n-1$ terms of P_x. So the adjugate R_x of P_x is a square matrix of size n, whose entries are polynomials in x of degree at most $n-1$. We can therefore write it as

$$R_x = R_0 x^{n-1} + R_1 x^{n-2} + \cdots + R_i x^{n-i-1} + \cdots + R_{n-2} x + R_{n-1}$$

where the R_i in the right hand side are $n \times n$ matrices of scalars. Clearly $xI - A$ is the parallel representation of P_x.

By Corollary 11.7.1 , we have

$$\det(P_x) I_n = P_x R_x. \tag{12.3}$$

Multiply out $P_x R_x$ using the representations above:

$$\begin{aligned} P_x R_x = {} & R_0 x^n + R_1 x^{n-1} + \cdots + R_i x^{n-i} + \cdots + R_{n-2} x^2 + R_{n-1} x \\ & - \left(A R_0 x^{n-1} + A R_1 x^{n-2} + \cdots + A R_i x^{n-i-1} + \cdots + A R_{n-2} x + A R_{n-1} \right) \end{aligned} \quad (12.4)$$

Now we enter this into (12.3), writing the left hand side as the scalar

$$\det(P_x) I_n = p(x) I_n = x^n I_n + p_1 x^{n-1} I_n + \cdots + p_i x^{n-i} I_n + \cdots + p_n I_n$$

without the alternation of sign of the coefficients, as before, since it serves no purpose here. So (12.3) becomes, after equating the constant matrices of polynomials appearing as coefficients of x^i on each side:

$$\begin{aligned} I &= R_0 && \text{coefficient of } x^n \\ p_1 I &= R_1 - A R_0 && \text{coefficient of } x^{n-1} \\ \ldots &= \ldots \\ p_i I &= R_i - A R_{i-1} && \text{coefficient of } x^{n-i} \\ \ldots &= \ldots \\ p_{n-1} I &= R_{n-1} - A R_{n-2} && \text{coefficient of } x \\ p_n I &= -A R_{n-1} && \text{constant coefficient} \end{aligned}$$

as you should check. Notice that each equation is a matrix equation that involves equating all the corresponding coefficients of $n \times n$ matrices. Now a miracle occurs: multiply the equation giving the coefficient of x^i on the left by A^i, and sum the equations. The right hand side telescopes to the zero matrix, and the left hand side is

$$A^n + p_1 A^{n-1} + \cdots + p_{n-1} A + p_n$$

which is just the characteristic polynomial evaluated at the matrix A. We have shown that the characteristic polynomial vanishes at A, so we are done. □

Examples, especially the 2×2 case.
Explain the miracle.

12.5 The Schur Unitary Triangularization Theorem

After the boring generalization to **C** of results that are true over **R**, we prove an interesting factorization theorem.

Theorem 12.5.1 (Schur Unitary Triangularization Theorem). *Let A be an $n \times n$ complex matrix, and let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be its eigenvalues, listed with the multiplicity with which they appear in the characteristic polynomial. They are written in any*

order. Then there is a unitary matrix U such that $U^*AU = T$, where T is an upper triangular matrix with $t_{ii} = \lambda_i$.

How does this compare to what we already know? Because $U^* = U^{-1}$, the theorem says that A is similar to T, but the similarity is achieved using a smaller collection of matrices: unitary rather than more general invertible matrices. We already know (Corollary 10.6.1 that A is similar to an upper triangular matrix, but our proof requires establishing the Jordan canonical form: the proof here is much more direct. Theorem 12.1.3 tells us that the characteristic polynomial of the upper triangular T is $\prod_i^n (x - t_{ii})$ so that the t_{ii} are the eigenvalues. One point that is new is that we can put the eigenvalues of A on the diagonal of T in any order.

Proof. Pick one of the eigenvalues of A and a unit length eigenvector $\mathbf{u}_1$ for this eigenvalue, which we write λ_1.

The first thing we will do is construct an orthogonal matrix U that will act as a change of basis matrix so that the new matrix $B = U^*AU$, which is similar to A since $U^* = U^{-1}$ and therefore has the same eigenvalues as A, has the unit coordinate vector $\mathbf{e}_1$ as eigenvector corresponding to λ_1.

Here is how we do that: Gram-Schmidt orthogonalization says we can complete $\mathbf{u}_1$ to a orthogonal basis $\mathbf{u}_i$, $2 \leq i \leq n$. By simply dividing each vector by its length we get an orthonormal basis. The matrix U whose columns are the unit length $\mathbf{u}_i$ is unitary. Consider the matrix $B = U^*AU$, so $UB = UU^*AU = AU$. Then by elementary matrix multiplication $U\mathbf{e}_1 = \mathbf{u}_1$. Then

$$UB\mathbf{e}_1 = AU\mathbf{e}_1 = A\mathbf{u}_1 = \lambda_1\mathbf{u}_1 = \lambda_1 U\mathbf{e}_1.$$

Multiplying both sides by U^*, we get $B\mathbf{e}_1 = \lambda_1\mathbf{e}_1$, as claimed. By obvious matrix multiplication. this means

$$B = \begin{pmatrix} \lambda_1 & \star \\ 0_{n-1,1} & B_1 \end{pmatrix} \tag{12.5}$$

so that B is block triangular, with diagonal blocks 1×1 (λ_1) and a $n-1 \times n-1$ block B_1 on the diagonal. Because B was obtained from A by similarity, the eigenvalues of B_1 are the remaining eigenvalues $\lambda_2, \ldots, \lambda_n$ of A.

We can repeat this operation on B_1 using a unitary matrix U_1, and so on, until we get a upper triangular matrix T whose diagonal elements are the λ_i in the order we chose. □

Geometrically here is what we are doing. We first pick any eigenvalue λ_1, with an associated eigenvector $\mathbf{v}_1$. We make a unitary change of coordinates to move $\mathbf{v}_1$ to $\mathbf{e}_1$. Then we restrict to the orthogonal complement of $\mathbf{v}_1$, a space of dimension $n-1$, and restrict the operator to this space. Its eigenvalues are the unused ones of the original operator. Find a new eigenvalue eigenvector pair, do another unitary coordinate change to move it to $\mathbf{e}_2$, and repeat.

We can use the Schur Theorem to improve the Spectral Mapping Theorem 12.3.3.

Theorem 12.5.2. *If the eigenvalues of the $n \times n$ matrix A are $\lambda_1, \lambda_2, \ldots, \lambda_n$, listed with the multiplicity with which they appear in the characteristic polynomial. Then the eigenvalues of $f(A)$, for any polynomial $f(x)$, are $f(\lambda_1), \ldots, f(\lambda_n)$.*

Proof. By the Schur Theorem (or by Jordan Canonical Form) we know that A is similar to a triangular matrix T, so $A = CTC^{-1}$. We do not need to assume that the change of basis matrix C is unitary. Then $f(A) = f(CTC^{-1}) = Cf(T)C^{-1}$ as we saw in Theorem 10.2.4. The diagonal entries of $f(T)$ are the $f(\lambda_i)$. Since A is similar to T, these are the eigenvalues of $f(A)$. □

12.6 The Characteristic Polynomial of the Companion Matrix

We compute the characteristic polynomial of the companion matrix (5.16). We already know the answer: indeed in Theorem 10.4.3 we showed that (12.6) is the minimal polynomial of the matrix: since it has degree n it is also the characteristic polynomial. We do the computation here by a different method.

Theorem 12.6.1. *The characteristic polynomial of the companion matrix is the polynomial*

$$f(x) = x^n - a_{n-1}x^{n-1} - a_{n-2}x^{n-2} - \cdots - a_0 \tag{12.6}$$

Proof. The characteristic polynomial of A is the determinant

$$\begin{vmatrix} x & 0 & 0 & \ldots & 0 & -a_0 \\ -1 & x & \ldots\ldots & & 0 & -a_1 \\ 0 & -1 & x & \ldots & 0 & -a_2 \\ 0 & 0 & -1 & \ldots & 0 & -a_3 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & x - a_{n-1} \end{vmatrix} \tag{12.7}$$

We compute the Laplace expansion along the first row, using induction on n. First we do the case $n = 2$. The determinant we need is

$$\begin{vmatrix} x & -a_0 \\ -1 & x - a_1 \end{vmatrix} = x(x - a_1) - a_0 = x^2 - a_1 x - a_0,$$

as required.

Now we do the case n. By Laplace expansion of the determinant along the first row we get two terms:

$$x \begin{vmatrix} x & \ldots\ldots & 0 & -a_1 \\ -1 & x & \ldots & -a_2 \\ 0 & -1 & \ldots & -a_3 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & x - a_{n-1} \end{vmatrix} + (-1)^n a_0 \begin{vmatrix} -1 & x & \ldots\ldots & 0 \\ 0 & -1 & x & \ldots & 0 \\ 0 & 0 & -1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & x \\ 0 & 0 & 0 & \ldots & -1 \end{vmatrix}$$

By induction, since the first term is x times the characteristic polynomial in the case $n-1$, we get

$$x(x^{n-1} - a_{n-1}x^{n-2} - a_{n-2}x^{n-3} - \cdots - a_2x - a_1)$$

while the second term gives $a_0(-1)^n(-1)^{n-1} = -a_0$, since the matrix is triangular with -1 along the diagonal.

Thus we do get the polynomial (12.6) as characteristic polynomial of (5.16). □

Remark 12.6.2. Starting with any monic polynomial $f(x)$ of degree n, we can ask for all similarity classes of $n \times n$ matrices whose characteristic polynomial is $f(x)$. We know that there is at least one, namely (5.16).

If $F = \mathbb{C}$, then $f(x)$ factors linearly with n complex roots $b_1, \ldots, b_n$, not necessarily distinct. Then we can construct an operator M whose matrix is a lower triangular $n \times n$ matrix B over $\mathbb{C}$, whose diagonal entries are the b_i, taken with their multiplicity as zeroes of $f(x)$. Then the characteristic polynomial of B is $f(x)$ by Theorem 12.1.3. Can we guarantee that these two operators are distinct? If they are similar, then they must have the same minimal polynomial as well as the same characteristic polynomial. Now Corollary 10.6.4 shows that the minimal polynomial for each i is the size of the largest Jordan block for that b_i. So if there is more than one Jordan block for at least one of the b_i, the minimal polynomial has degree less than n, so the operators are not similar.

When $F = \mathbb{R}$ the situation is more complicated, since the irreducible factorization of $f(x)$ over $\mathbb{R}$ contains polynomials of degree 1 and degree 2. Let us look at the 2×2 case: the real matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{12.8}$$

with characteristic polynomial

$$g(x) = x^2 - (a+d)x + (ad - bc).$$

The trace of this matrix is $a+d$ and its determinant $ad-bc$. Notice how they appear in the characteristic polynomial. For this to be irreducible over $\mathbb{R}$, by the quadratic formula we must have

$$(a+d)^2 - 4(ad-bc) = (a-d)^2 + 4bc < 0$$

or $(a-d)^2 < 4bc$.

The companion matrix of this polynomial is

$$\begin{pmatrix} 0 & -(ad-bc) \\ 1 & a+d \end{pmatrix}$$

The full value of the companion matrix only reveals itself when one takes smaller subfields of $\mathbb{C}$, for example the field of rational numbers $\mathbb{Q}$. Over such a field there

are irreducible polynomials of arbitrary high degree: for example the *cyclotomic polynomial*

$$\Phi_p(x) = x^{p-1} + x^{p-2} + \cdots + x + 1$$

for p a prime number. Since $(x-1)\Phi_p(x) = x^p - 1$, the roots of $\Phi(t)$ are complex numbers on the circle of radius 1, thus certainly not rational. It is a bit harder to show that $\Phi_p(x)$ is irreducible, but it only requires the elementary theory of polynomials in one variable. A good reference is Steven H. Weintraub's paper Several Proofs of the Irreducibility of the Cyclotomic Polynomial

12.7 The Minors of a Square Matrix

Now assume that A and B are two square matrices of size n. We know that if they are similar, meaning that there is an invertible matrix C so that $B = CAC^{-1}$, they must have the same minimal polynomial and the same characteristic polynomial. Is that enough to insure they are similar? We will now see it is not, using the Cauchy-Binet Theorem 11.8.4.

Definition 12.7.1. For the matrix A of size n let $D_A(k)$ be the greatest common divisor of all the minors of $xI - A$ of size k. Thus $D_A(k)$ is a polynomial in x, that we normalize so that its leading coefficient is 1: it is monic.

It is immediate that $D_A(n)$ is the characteristic polynomial of A. For convenience we set $D_A(0) = 1$. Then $D_k(A)$ divides $D_{k+1}(A)$ for $0 \le k \le k-1$.

Write $\bar{A}$ for $xI - A$ and $\bar{B}$ for $xI - B$. We know that if A and B are similar, then $\bar{B} = C\bar{A}C^{-1}$. For an arbitrary minor $\bar{B}^I_J$ of size k of B, by two uses of the Cauchy-Binet Theorem 11.8.4, with the notation there, we get

$$\det(\bar{B}^I_J) = \sum_K \Big(\sum_L \det(C^I_L) \det(\bar{A}^L_K) \Big) \det((C^{-1})^K_J)$$

where the sums over K and L are over all subsets of $[1, \ldots, n]$ containing k elements.

Remembering that the matrices C and C^{-1} involve only constants this implies that $\det(\bar{B}^I_J)$ divides all the $\det(\bar{A}^L_K)$. In particular the greatest common divisor $D_B(k)$ of the $\det(\bar{B}^I_J)$ divides the greatest common $D_A(k)$ of the $\det(\bar{A}^I_J)$. Running the same argument by moving the C and C^{-1} to the other side, we see that $D_A(k)$ divides $D_B(k)$, so they are the same monic polynomial. So we have proved:

Theorem 12.7.2. *If A and B are similar matrices, then $D_A(k) = D_B(k)$ for all k.*

Example 12.7.3. If A is the $n \times n$ companion matrix (5.16), let's compute $D_A(n-1)$. Remove the first row and the last column from the characteristic polynomial A, given in (12.7). The determinant of that minor of size $n-1$ is clearly ± 1, since it is upper triangular with -1 along the diagonal. Thus $D_A(n-1) = 1$, so all the $D_A(k) = 1$, for $1 \le k \le n-1$, since the others must divide $D_A(n-1)$.

Example 12.7.4. To be sure that this is a new invariant, beyond the minimal polynomial and the characteristic polynomial, we should exhibit two polynomial that have the same minimal and characteristic polynomials, and yet are not similar, by showing that $D_A(k) \neq D_B(k)$ for some k.

For example take

$$A = \begin{pmatrix} 0&0&0&0 \\ 1&0&0&0 \\ 0&0&0&0 \\ 0&0&1&0 \end{pmatrix} \quad \text{so} \quad xI - A = \begin{pmatrix} x&0&0&0 \\ -1&x&0&0 \\ 0&0&x&0 \\ 0&0&-1&x \end{pmatrix}$$

and

$$B = \begin{pmatrix} 0&0&0&0 \\ 1&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \end{pmatrix} \quad \text{so} \quad xI - B = \begin{pmatrix} x&0&0&0 \\ -1&x&0&0 \\ 0&0&x&0 \\ 0&0&0&x \end{pmatrix}$$

Then the minimal polynomial for both is x^2 and the characteristic polynomial x^4, as you should check. However $D_A(2) = 1$ while $D_B(2) = x$, and $D_A(3) = x$ while $D_B(3) = x^2$ so these matrices are not similar.

12.8 Computation of Eigenvectors

Put the material about iteration of A and importance of eigenvector computations? Probability matrices? Markov chains?

12.9 The Big Picture

A summary of what happens first general results, then over $\mathbb{C}$, then over $\mathbb{R}$.

Talk about iterations A^k of linear operators: what happens as k gets large and what do the eigenvalues tells you.

12.10 The Coefficients of the Characteristic Polynomial

We have already described the two most important coefficients in the characteristic polynomial: the trace and the determinant. In this optional section we explain the others. This is a more complicated combinatorial problem.

Before stating the result, we need notation concerning submatrices and their determinants, called minors, of A. We first considered submatrices and established

notation in §2.4; then we looked at their determinants in §11.5, but only for submatrices of size $n-1$ of a matrix of size n. Here we need some more general notation.

Definition 12.10.1. Let $J = \{i_1, \dots, i_k\}$ be a set of k distinct integers in the interval $[1,n]$, listed in increasing order. If A is a square matrix, let A_J or $A(i_1, \dots, i_k)$ (which is the notation used in §2.4) be the principal submatrix of A formed by the rows and the columns of A with indices J. Let $m_J = m(i_1, \dots, i_k)$ be its determinant, or minor. Then A_J is a principal submatrix of A, and its determinant m_J or $m(i_1, \dots, i_k)$ a principal minor of A.

For each k between 1 and n, the submatrices $A(1,2,\dots,k)$ are the leading principal submatrices, and their determinants the leading principal minors of A.

Example 12.10.2. So if A is the matrix

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

then

$$A(1,2) = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}, A(2,3) = \begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix}, \text{ and } A(1,2,3) = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 3 \\ 0 & 3 & 4 \end{pmatrix}$$

Obviously, if A is symmetric as in this example, A_J is symmetric. For each k, there is only one leading principal matrix, but $\binom{n}{k}$ principal minors of order k.

For the matrix A above, we have already computed $A(1,2)$ and $A(1,2,3)$; $A(1)$ is the 1×1 matrix (2) and $A(1,2,3,4)$ is A. So the determinant $m(1) = 2$, $m(1,2) = 8$, and $m(1,2,3) = 2(16-9) = 14$.

Then we compute the coefficients of the characteristic polynomial in the following theorem. The proof is complicated: you may want to look at Example 12.10.5 first, since it deals with the simplest case: a diagonal matrix. In any case, we only use this result much later in the book.

Theorem 12.10.3. *For each index j, $1 \le j \le n$ we have*

$$p_j = \sum_J \det A_J = \sum_J m_J$$

where the sum is over all choices of j elements $J = \{i_1, \dots, i_j\}$ from $[1,n]$, and A_J is the corresponding principal submatrix of A, and m_J its determinant. Thus the sum has $\binom{n}{j}$ terms.

Example 12.10.4. We first check this in the two cases we already know from Theorem 12.1.7.

- If $j = n$, then there is only one choice for J: all the integers between 1 and n. Theorem 12.10.3 then says that $p_n = \det A$.

- If $j = 1$, the sets J have just one element, so Theorem 12.10.3 says

$$p_1 = a_{11} + a_{22} + \cdots + a_{nn}$$

which is indeed the trace.

Proof. To do the general case, we use the definition of the determinant (11.3) in terms of permutations.

To pass from the determinant to the characteristic equation, we must make the following substitutions: we replace each term a_{ij} by $\delta_{ij}x - a_{ij}$, where x is a variable and δ_{ij} is the *Kronecker delta*

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j; \\ 0, & \text{otherwise.} \end{cases}$$

With this notation the characteristic polynomial is the sum over all permutations σ of

$$\varepsilon(\sigma)(\delta_{1,\sigma(1)}x - a_{1,\sigma(1)})(\delta_{2,\sigma(2)}x - a_{2,\sigma(2)})\ldots(\delta_{n,\sigma(n)} - a_{n,\sigma(n)}). \tag{12.9}$$

We fix an integer k in order to study the coefficient p_k of x^{n-k} in the characteristic polynomial.

Then we consider a subset $I = \{i_1, i_2, \ldots, i_k\}$, $1 \le i_1 < i_2 < \cdots < i_k \le n$, of the first n positive integers, and we ask how a term such as (12.9) can produce a factor

$$\pm a_{i_1,\sigma(i_1)}a_{i_2,\sigma(i_2)}\ldots a_{i_k,\sigma(i_k)}$$

The permutation σ must fix at least the $n-k$ integers not in I, meaning that $\sigma(i) = i$ for $i \notin I$.

The **key point** is that σ then restricts to a permutation $\tau\colon I \to I$. Conversely a permutation $\tau\colon I \to I$ extends uniquely to a permutation σ of the first n integers by defining it to be constant on the integers not in I. Note that $\varepsilon(\sigma) = \varepsilon(\tau)$.

Thus the term of degree $n-k$ in the characteristic polynomial can be written

$$\sum_I \sum_\sigma \varepsilon(\sigma) \prod_{J_I} (\delta_{j,\sigma(j)} \prod_I (-a_{i,\sigma(i)}) \tag{12.10}$$

where the first sum is over all the subsets I of k elements as defined above; the second sum is over all the permutations of $[1,\ldots,n]$. J_I is the complementary set to I of $n-k$ elements in $[1,\ldots,n]$, and the product

$$\prod_{J_I} (\delta_{j,\sigma(j)}$$

is over all the $j \in J$. In other words in the product of binomials of (12.10) we pick $n-k$ times the Kronecker delta term, and k times the $-a_{i,\sigma(i)}$ term, in all possible ways.

Unless the permutation σ fixes the elements in J, the corresponding term in (12.10) vanishes, so by the key point we can rewrite (12.10) as

$$\sum_I \Big(\sum_\tau \varepsilon(\tau) \prod_{i \in I} (-a_{i,\tau(i)}) \Big) \tag{12.11}$$

where τ varies over all the permutations of I. Then the interior sum is just $(-1)^k$ times the determinant of the principal submatrix A_I of A, which proves the result. □

Example 12.10.5. If the matrix A is diagonal, then the characteristic polynomial of A is just $\prod_{i=1}^{n}(x-a_{ii})$, Therefore by Definition 12.11, the coefficients of the characteristic polynomial are just the elementary symmetric polynomials in the a_{ii}, which are of course the eigenvalues of A. It is easy to see in this case that the elementary symmetric polynomials are the sum of the determinants of the principal sub minors of the appropriate size of A, as Theorem 12.10.3 says they are.

More examples here.

Chapter 13
The Spectral Theorem

Abstract The main result of the lecture, one of the most important results in linear algebra, indeed, all of mathematics, is the Spectral Theorem 13.3.1. It tells us that the eigenvalues and eigenvectors of a real symmetric matrix (and of a complex Hermitian matrix) are real. We prove it first by using an argument over $\mathbb{C}$, and a second time (in the real symmetric case) , without introducing complex number, using the Rayleigh Quotient instead. This argument requires a famous theorem in real analysis: the maximum theorem for a continuous function on a compact set, that we do not prove. An immediate corollary of the Spectral Theorem is Theorem 13.3.4, which shows that we can diagonalize real symmetric matrices using orthogonal matrices, defined in Definition 8.3.6 precisely for this appearance. This gives another verification of Algorithm 7.5.3. Then we list various ways of characterizing positive definite and semidefinite forms: Theorem 13.5.4.

13.1 Triangulation of Complex Operators

In this section we prove a variant of Corollary 10.6.1, which was derived indirectly from the Jordan canonical form.

First recall some definitions: a vector space V is an inner-product space if it is either a Euclidean space (therefore real) or a Hermitian space (therefore complex). In both cases we can associate to an operator $L\colon V \to V$ its adjoint, that we will write here as L^*. See Sections 9.1 and 9.2.

Theorem 13.1.1. *Assume V is an inner-product space, L an operator on V. Let W be a subspace of V invariant under L. Then $W^\perp$ is invariant under L^*.*

Proof. By hypothesis, if $\mathbf{w} \in W$, then $L\mathbf{w} \in W$. Let $\mathbf{v}$ be in $W^\perp$. Then

$$0 = \langle L\mathbf{w}, \mathbf{v}\rangle = \langle \mathbf{w}, L^*\mathbf{v}\rangle.$$

Since $\mathbf{w}$ is an arbitrary element of W, this shows that every $L^*\mathbf{v}$ is orthogonal to W. □

Now we apply this in the complex case.

Theorem 13.1.2. *Let V be a Hermitian space, and L any operator on V. Then there is an orthonormal basis for V in which the matrix A of L is upper-triangular. thus L is triangulable.*

Proof. We prove this by induction on the dimension n of V. If $n = 1$ there is nothing to prove. Because V is complex, the adjoint L^* has an eigenvalue λ with associated eigenvector $\mathbf{v}$ of length 1. Let W be the orthogonal complement of the one-dimension space spanned by $\mathbf{v}$. So W is invariant under L. We can restrict L to . an operator L_W on W. Since W has dimension $n-1$, by induction we know that L_W is triangulable in an suitable orthonormal basis $\mathbf{v}_2, \ldots, \mathbf{v}_n$ of W. I claim that the matrix of L is upper-triangular in the orthonormal basis $\mathbf{v}, \mathbf{v}_2, \ldots, \mathbf{v}_n$ of W. We do not know how L acts on $\mathbf{v}$, since we only know that $\mathbf{v}$ is an eigenvector for the adjoint L^*. But that does not matter, since it is the first vector. □

This is the desired result.

13.2 The Rayleigh Quotient

Now V is a Euclidean space of dimension n. Let A be the matrix of the symmetric operator L in an orthonormal basis. Thus A is a symmetric matrix. Our goal is to prove that all the eigenvalues of A are real. We will do this in two different ways. The first proof does not involve the complex numbers, but requires a well-known theorem in multivariable calculus you may not have seen. We start on that proof here. First recall that since A is symmetric: $\langle \mathbf{x}, A\mathbf{x}\rangle = \mathbf{x}^t A\mathbf{x} = \langle \mathbf{A}x, \mathbf{x}\rangle$, which just expresses the fact that L is self-adjoint.

Starting from the quadratic form $\mathbf{x}^t A\mathbf{x}$, we define a function that is the key to the proof.

Definition 13.2.1. The *Rayleigh quotient* is the real-valued function, defined for all non-zero vectors $\mathbf{x}$ as:

$$R(\mathbf{x}) = \frac{\langle \mathbf{x}, A\mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle}.$$

R is clearly continuous everywhere it is defined, namely everywhere except at the origin. In fact it is infinitely differentiable there, since it is just a quotient of polynomials in the variables. Moreover it has the following useful property:

Lemma 13.2.2. *For any non-zero $t \in \mathbb{R}$, $R(t\mathbf{x}) = R(\mathbf{x})$*

Proof. Just substitute $t\mathbf{x}$ for $\mathbf{x}$ in the definition of $R(\mathbf{x})$. □

Next we define a simple geometric concept.

Definition 13.2.3. A *ray* emanating from the origin through a non-zero point $\mathbf{e}$ is the set of $t\mathbf{e}$, for $t > 0$. Thus a ray is a half-line starting at the origin and passing through $\mathbf{e}$.

The reason we formed the Rayleigh quotient is

Lemma 13.2.4. *All the values of $R(\mathbf{x})$ are attained on the unit sphere $U = \{\mathbf{x} | \|\mathbf{x}\| = 1\}$.*

Proof. Lemma 13.2.2 says that the Rayleigh quotient is constant along rays emanating from the origin. Since every point in $\mathbb{R}^n \setminus \mathbf{0}$ is on a (unique) ray emanating from the origin, and since each such ray meets the unit sphere (why?), all values of R are attained on the unit sphere U. □

Theorem 13.2.5. *$R(\mathbf{x})$ attains its maximum M and its minimum m on the unit sphere.*

Proof. We do not prove this result. A famous theorem in multivariable calculus called the Weierstrass maximum theorem says $R(\mathbf{x})$, when restricted to U attains both its minimum and its maximum values, because $R(\mathbf{x})$ is continuous on U and U is closed and bounded. We will not define what it means to be closed, but roughly it means that every limit of a converging sequence of points in U is in U. Bounded just means that U is contained in a ball of finite radius, which is obviously true for U. By Lemma 13.2.4, the maximum and minimum of $R(\mathbf{x})$ on all of $\mathbb{R}^n \setminus \mathbf{0}$ are attained on U, so the values M and m we have found are not only the maximum and the minimum for R when restricted to U, but for R on all of $\mathbb{R}^n \setminus \mathbf{0}$. □

Recall that the gradient $R(\mathbf{x})$ is the vector of partial derivatives $\left(\frac{\partial R}{\partial x_i}\right)$, for $1 \leq i \leq n$. We now use without proof a well-known theorem of multivariable calculus: any point where $R(\mathbf{x})$ attains its maximum or minimum is a critical point for R, namely a point $\mathbf{e}$ where $\nabla R(\mathbf{e}) = \mathbf{0}$.

Here is the key computation.

Proposition 13.2.6. *Let $\mathbf{e} \in U$ be a critical point for R. Then $\mathbf{e}$ is an eigenvector of A with eigenvalue $\lambda = R(\mathbf{e})$.*

Proof. Let $\mathbf{f}$ be an arbitrary but fixed non-zero vector in $\mathbb{R}^n$, and let t be a real variable. We evaluate the Rayleigh quotient at $\mathbf{e} + t\mathbf{f}$, and write the composite function as

$$g(t) = R(\mathbf{e} + t\mathbf{f}).$$

The numerator of $g(t)$ is

$$p(t) = \langle \mathbf{e} + t\mathbf{f}, A(\mathbf{e} + t\mathbf{f}) \rangle = \langle \mathbf{e}, A\mathbf{e} \rangle + 2t \langle \mathbf{e}, A\mathbf{f} \rangle + t^2 \langle \mathbf{f}, A\mathbf{f} \rangle \tag{13.1}$$

and its denominator is

$$r(t) = \langle \mathbf{e} + t\mathbf{f}, \mathbf{e} + t\mathbf{f} \rangle = \langle \mathbf{e}, \mathbf{e} \rangle + 2t \langle \mathbf{e}, \mathbf{f} \rangle + t^2 \langle \mathbf{f}, \mathbf{f} \rangle \tag{13.2}$$

so they are both quadratic polynomials in the variable t. Now the derivative $g'(t) = \langle \nabla R((\mathbf{e} + t\mathbf{f}), \mathbf{f} \rangle$ by the chain rule. We evaluate $g'(t)$ at $t = 0$. Since the gradient $\nabla R(\mathbf{e}) = 0$ by the hypothesis that we are at a critical point, we get $g'(0) = 0$.

On the other hand, since $g(t) = p(t)/r(t)$, by the quotient rule we get

$$g'(0) = \frac{p'(0)r(0) - p(0)r'(0)}{r^2(0)} = 0.$$

Now $r^2(0) = \langle \mathbf{e}, \mathbf{e} \rangle = 1$, since $\mathbf{e}$ is on U. Furthermore $p(0) = R(\mathbf{e})$, which we denote λ. So we get:

$$g'(0) = p'(0) - \lambda r'(0) = 0. \tag{13.3}$$

Next we compute the derivatives of $p(t)$ and $r(t)$ at 0, using (13.1) and (13.2) respectively.

$$p'(0) = 2\langle \mathbf{f}, A\mathbf{e} \rangle$$
$$r'(0) = 2\langle \mathbf{f}, \mathbf{e} \rangle$$

Equation (13.3) reads, after substituting in these values:

$$2\langle \mathbf{f}, A\mathbf{e} \rangle - 2\lambda \langle \mathbf{f}, \mathbf{e} \rangle = 0 \text{ , or } \quad \langle \mathbf{f}, A\mathbf{e} - \lambda \mathbf{e} \rangle = 0.$$

Since $\mathbf{f}$ is an arbitrary vector in $\mathbb{R}^n$, this means that $A\mathbf{e} - \lambda \mathbf{e}$ is perpendicular to every vector, which can only happen if it is the zero vector: $A\mathbf{e} - \lambda \mathbf{e} = \mathbf{0}$. Thus $\mathbf{e}$ is an eigenvector of A with eigenvalue $\lambda = R(\mathbf{e})$, which concludes the proof of the proposition. □

Theorem 13.2.7. *Let m be the minimum, and M the maximum value of $R(\mathbf{x})$. Then m and M are eigenvalues of A, so A has a real eigenvalue.*

Proof. Theorem 13.2.5 says $R(\mathbf{x})$ has a maximum and a minimum, therefore it has a least one critical point. In fact, unless $R(\mathbf{x})$ is constant, it has two critical points. By Proposition 13.2.6 each critical point $\mathbf{e}$ is an eigenvector with eigenvalue $R(\mathbf{e})$. Since $R(\mathbf{x})$ is a real-valued function each such eigenvalue is real. □

The minimum m is therefore the smallest value of $R(\mathbf{x})$ at a critical point, and the maximum the largest.

13.3 The Spectral Theorem for a Real Symmetric Matrix

Using the results on the Rayleigh quotient, we give a first proof of the spectral theorem for real self-adjoint operators on a Euclidean space V. We work with its matrix A in an orthonormal basis of V, which is therefore symmetric.

Theorem 13.3.1 (The Spectral Theorem). *If A is a real symmetric $n \times n$ matrix, then its eigenvalues are real and its eigenvectors can be selected to form an orthonormal basis of the vector space V.*

The spectrum of a matrix is the set of its eigenvalues. This theorem is called the spectral theorem because it describes the eigenvalues of a real symmetric matrix: they are real. The first paragraph of Steen [25] discusses the early history of the

spectral theorem, at the time it was called the principal axis theorem. We have already seen the contribution of Sylvester in his law of inertia 7.7.6. We have used the method of Lagrange (§7.5) to diagonalize quadratic forms.

Example 13.3.2. Before starting the proof, let's work out the familiar 2×2 case. Let A be an arbitrary 2×2 matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

To compute the eigenvalues of A, we need the roots of the characteristic polynomial of A, namely the determinant

$$\begin{vmatrix} t-a & -b \\ -c & t-d \end{vmatrix} = t^2 - (a+d)t + ad - bc.$$

The quadratic formula tells us that this polynomial has real roots if and only if the discriminant is non-negative. The discriminant is

$$(a+d)^2 - 4(ad-bc) = a^2 + 2ad + d^2 - 4ad + 4bc = (a-d)^2 + 4bc.$$

When the matrix is symmetric, $b = c$, so we get $(a-d)^2 + 4b^2$, a sum of squares, which is always non-negative. So the eigenvalues:

$$\lambda_i = \frac{a+d \pm \sqrt{(a-d)^2 + 4b^2}}{2}$$

are real.

What about the eigenvectors? We could compute them, but we only need to show they are orthogonal. First assume the matrix has a double eigenvalue. This corresponds to the discriminant being 0, which means that $b = 0$ and $a = d$. Because the matrix is diagonal, any non-zero vector in the plane is an eigenvector. There is therefore no difficulty in finding two eigenvectors that are orthogonal. Now assume that the eigenvalues are distinct. Let $\mathbf{v}_1$ be the unit eigenvector corresponding to the first (real) eigenvalue λ_1. Let $\mathbf{v}_2$ be a unit vector generating the orthogonal complement of $\mathbf{v}_1$. So

$$\langle A\mathbf{v}_1, \mathbf{v}_2 \rangle = \lambda_1 \langle \mathbf{v}_1, \mathbf{v}_2 \rangle = 0.$$

On the other hand, since A is symmetric

$$\langle A\mathbf{v}_1, \mathbf{v}_2 \rangle = \langle \mathbf{v}_1, A\mathbf{v}_2 \rangle.$$

The two equations together say that $A\mathbf{v}_2$ is orthogonal to $\mathbf{v}_1$, so it is a multiple of $\mathbf{v}_2$. Thus $\mathbf{v}_2$ is the second eigenvector. This settles the theorem in dimension 2.

Proof (Proof of the Spectral Theorem). We now do the case of general n, generalizing the ideas of the two-dimensional case. By Theorem 13.2.7 we have one real eigenvalue λ_1 with its eigenvector $\mathbf{v}_1$, which we can normalize to length 1 .

Apply Theorem 13.1.1 to L, using for W the one-dimension space spanned by $\mathbf{v}_1$. Then the orthogonal $W^\perp$ is invariant under L^* , namely under L, since L is self-

adjoint. So choose an orthonormal basis for V that starts with $\mathbf{v}_1$. Then the remaining elements of the basis are a basis of $W^\perp$. In this basis, the matrix A for L has only zeroes in the first row and column except at $a_{11} = \lambda_1$. Because L is self-adjoint, A is symmetric. This shows that L restricts to a self-adjoint operator L_1 on $W^\perp$. Its matrix in the basis above is just the principal submatrix of A consisting of rows and columns $(2,3,\ldots,n)$. Repeating the argument using the Rayleigh quotient applied to L_1, we find a second real eigenvalue λ_2 with unit length eigenvector $\mathbf{v}_2$.

We just repeat this until we get n mutually orthonormal eigenvectors $\mathbf{v}_i$ with real eigenvalues λ_i. □

If at each step we take the eigenvalue corresponding to the minimum of the Rayleigh quotient, we get $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.

Definition 13.3.3. Let A be a symmetric $n \times n$ matrix. Let $\mathbf{v}_1$, $\mathbf{v}_2$, …, $\mathbf{v}_n$ be the collection of orthonormal eigenvectors found in the Spectral Theorem, and λ_i the corresponding eigenvalues. Let Q be the matrix whose i-th column is the eigenvector $\mathbf{v}_i$. Then Q is called the *matrix of eigenvectors* of A, and $\lambda = (\lambda_1,\ldots,\lambda_n)$ the *vector of eigenvalues*. The basis $\{\mathbf{v}_1,\ldots,\mathbf{v}_n\}$ of V is called a *spectral basis* for A.

We write D for $D(\lambda_1,\lambda_2,\ldots,\lambda_n)$, the diagonal matrix with diagonal entries the eigenvalues.

Theorem 13.3.4. *Let A be a real symmetric $n \times n$ matrix, Q its matrix of unit length eigenvectors, and λ its vector of eigenvalues. Then Q is an orthogonal matrix. and*

$$Q^{-1}AQ = D \quad or \quad A = QDQ^t \tag{13.4}$$

Proof. That the matrix Q is orthogonal follows immediately from the fact that its columns, the eigenvectors, have length 1 and are orthonormal. We can write all the eigenvector-eigenvalue equations in one matrix equation:

$$AQ = QD, \tag{13.5}$$

as a moment's thought will confirm. Multiply on the left by Q^{-1}, to get $Q^{-1}AQ = Q^{-1}QD = D$. Because Q is an orthogonal matrix, $Q^{-1} = Q^t$. □

Exercise 13.3.5. Show that (13.5) encodes all the eigenvector-eigenvalues, as claimed.

Review Definition 7.7.3 for the meaning of p, k, and m in the next result. If you have not read that chapter, just define them using the following result.

Corollary 13.3.6. *Start with a symmetric matrix A. Its rank is the number of non-zero eigenvalues. p is the number of positive eigenvalues, k is the number of zero eigenvalues, and m is the number of negative eigenvalues.*

Proof. The matrix $D = D(\lambda_1,\lambda_2,\ldots,\lambda_n)$ is congruent to A because Q is an orthogonal matrix, so $Q^{-1} = Q^t$. Now p, k, and m are invariants of the congruence class. They are easy to compute for the matrix D. □

13.4 The Spectral Theorem for Self-Adjoint Operators

We now prove the Spectral Theorem for self-adjoint operators over $\mathbb{R}$ or $\mathbb{C}$. In the real case, the proof we give here could replace the part of the proof involving the Rayleigh quotient in §13.2, at the cost of using techniques involving complex variables. Since we already have a proof for the real case, we focus on the complex case, but the proof works in both cases.

First we prove a general result that holds only over $\mathbb{C}$. It gives a simple proof of a result we proved very indirectly using the minimal polynomial, primary decomposition and Jordan normal form: see Corollary 10.6.1.

Theorem 13.4.1. *Let V be a Hermitian space and L any operator on V. Then there is an orthonormal basis for V in which the matrix A representing L is upper-triangular.*

Proof. We prove this by induction of the dimension n of V. Because V is a complex vector space, the adjoint L^* has a unit eigenvector $\mathbf{v}$, therefore a one-dimensional invariant subspace W spanned by $\mathbf{v}$. By Theorem 9.2.4, the orthogonal complement $W^\perp$ is invariant under $L = (L^*)^*$. $W^\perp$ has dimension $n-1$, so by induction the theorem applies to the restriction of L to $W^\perp$. To choose a orthonormal basis for V that starts with $\mathbf{v}$ and then continues with an orthonormal basis of $W^\perp$. The vector $\mathbf{v}$ is mapped under L to some linear combination of all the basis vectors. The matrix A representing L in this basis is upper-triangular, so we are done. □

Now we turn to the spectral theorem. Notice how the next theorem improves the result we just proved.

Theorem 13.4.2. *If L is a self-adjoint operator on an inner-product space then any eigenvalue of L is real, and the eigenvectors corresponding to distinct eigenvalues are orthogonal.*

Proof. Assume $L\mathbf{v} = \lambda\mathbf{v}$. Then

$$\langle L\mathbf{v}, \mathbf{v}\rangle = \langle \lambda\mathbf{v}, \mathbf{v}\rangle = \lambda\langle \mathbf{v}, \mathbf{v}\rangle$$

because of the linearly of the inner product in the first variable. On the other hand,

$$\langle \mathbf{v}, L\mathbf{v}\rangle = \langle \mathbf{v}, \lambda\mathbf{v}\rangle = \overline{\lambda}\langle \mathbf{v}, \mathbf{v}\rangle$$

since the inner product is conjugate linear in the second variable. These two expressions are equal. Now $\lambda\langle \mathbf{v}, \mathbf{v}\rangle \neq 0$, because $\mathbf{v}$, being an eigenvector is non-zero and the inner product is positive definite. Thus we must have $\lambda = \overline{\lambda}$, which just says that λ is real.

In the real case, we must show that the eigenvector we used can be chosen to be real. Work with the matrix A of L in any orthogonal basis. Then for the vector of coefficients $\mathbf{z}$ of $\mathbf{v}$ we have $A\mathbf{z} = \lambda\mathbf{z}$. The matrix A is real, and, as we just proved A is real. Write $\mathbf{z} = \mathbf{x} + i\mathbf{y}$, where both $\mathbf{x}$ and $\mathbf{y}$ are real. Then both $\mathbf{x}$ and $\mathbf{y}$ are real eigenvectors associated to λ.

Now we turn to the second statement. Assume that we have two distinct eigenvalues λ and μ with associated eigenvectors $\mathbf{v}$ and $\mathbf{w}$. Then

$$\lambda\langle \mathbf{v}, \mathbf{w}\rangle = \langle L\mathbf{v}, \mathbf{w}\rangle = \langle \mathbf{v}, L\mathbf{w}\rangle = \overline{\mu}\langle \mathbf{v}, \mathbf{w}\rangle.$$

Since μ is an eigenvalue, therefore real, $\overline{\mu} = \mu$. Since $\lambda \neq \mu$ we must have $\langle \mathbf{v}, \mathbf{w}\rangle = 0$ as required. □

This theorem does not immediately imply that in the real case L has an eigenvalue. However let the real symmetric matrix A act on the complexification $V_{\mathbb{C}}$ of V. Then A is Hermitian, and viewed as a complex matrix has the same characteristic polynomial as when viewed as a real matrix. Since as a complex matrix it has an eigenvalue and eigenvector, the theorem above shows that the eigenvalue is real and the eigenvector can be chosen to be real, too.

In both the real and complex cases we can use the spectral basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of eigenvectors as an orthonormal basis for V. Then we have generalizations of Theorem 13.3.4 and Corollary 13.3.6, where you just replace the word symmetric by Hermitian and orthogonal by unitary. Because the eigenvalues are real by Theorem 13.4.2, the statement of Corollary 13.3.6 still makes sense in the complex case.

13.5 Positive Definite Matrices

We started out study of positive definite operators and matrices in §9.3, where the key definitions are given. We continue treating the real and complex cases in parallel, but focussing on the real case. In Theorem 9.3.1 we showed that a positive definite matrix A is invertible, and that its inverse is positive definite. We can say more now.

Corollary 13.5.1. *If A is positive definite, the eigenvalues of A^{-1} are $1/\lambda_i$ with the same unit eigenvectors $\mathbf{v}_i$, and therefore the same eigenvector matrix Q, so A^{-1} is also positive definite. Then the eigenvalue-eigenvector decomposition of A^{-1} can be written:*

$$Q^{-1}A^{-1}Q = D(1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_n)$$

Proof. All the matrices in (13.4) are invertible, so just compute the inverse using the fact that the inverse of the orthogonal matrix Q is its transpose, that the inverse of the diagonal matrix $D(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is $D(1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_n)$ and that computing the inverse of a product of invertible matrices reverses the factors of the product. □

So Q is a change of basis matrix that diagonalizes the quadratic form, as in Theorem 7.4.1. It is an orthogonal matrix in the real case, and a unitary matrix in the complex case. It is a "better" change of basis because it preserves distance and angle - that is what being orthogonal means. Note finally, that the diagonal matrix obtained by this method is uniquely defined (up to order), since it consists in the eigenvalues of A.

Why not always diagonalize by this method? The answer is that it is harder (and more expensive computationally) to compute the eigenvalues than to do Gaussian elimination.

Example 13.5.2 (Example 7.7.14 once again). In §7.5 we computed a diagonal matrix (associated to the quadratic form given by (7.12) by change of basis), and obtain $D(1,3/4,2/3)$. In (7.7.14) we computed the eigenvalues of the same form q, and obtained $D(1/2,1/2,2)$. From the preceding remark see that $D(1/2,1/2,2)$ can also be viewed as being obtained from a change of basis. Thus, as we claimed in the remark before Definition 7.7.3, the matrix D itself is not unique. However, in accordance with the Law of Inertia 7.7.6, the numbers p_+, p_0 and p_- are the same: indeed, for both, we get $(3,0,0)$. The form q is positive definite.

Example 13.5.3. The permutation matrices A and A^σ (recall Definition 11.3.9 have the same type: if one is positive definite, the other is; if one is positive semidefinite, the other is, and so on.

Indeed, they have the same characteristic polynomial and therefore the same eigenvalues. Therefore by Corollary 13.3.6 they have the same signature.

Our goal is to develop tests for positive definiteness and positive semidefiniteness for symmetric matrices in the real case, and conjugate symmetric matrices in the complex case . Here is the first result, expressed using operators

Theorem 13.5.4. *The following conditions on the self-adjoint operator L are equivalent.*

1. *L is positive (semi)definite.*
2. *The eigenvalues of L are all positive (non negative).*
3. *There exists self-adjoint operator S on V such that $L = S^2$, which is invertible if and only if L is positive definite. S^* is as usual the adjoint of S.*
4. *The index of negativity of L is* 0 *in both cases.*

Proof. By the Spectral Theorem we can find an orthonormal basis of eigenvectors for any self-adjoint operator. We only consider the positive definite case, and leave the semidefinite case as an exercise. Let A be the matrix of L is the spectral basis of unit eigenvectors of L, so that A is a diagonal matrix with the eigenvalues along the diagonal.

For $(1) \Rightarrow (2)$, assume $\mathbf{v}$ is an eigenvector with eigenvalue λ: $L\mathbf{v} = \lambda\mathbf{v}$. Then

$$\langle L\mathbf{v}, \mathbf{v}\rangle = \langle \lambda\mathbf{v}, \mathbf{v}\rangle = \lambda\langle \mathbf{v}, \mathbf{v}\rangle.$$

This must be positive. Since $\langle \mathbf{v}, \mathbf{v}\rangle > 0$, we must have $\lambda > 0$. Thus all the eigenvalues are positive. For $(2) \Rightarrow (1)$, assume all the eigenvalues are positive. Write an arbitrary element $\mathbf{v} \in V$ as a linear combination of the orthonormal eigenvectors $\mathbf{v}_i$, which form a basis of V:

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$$

for complex numbers c_i. Then

$$\begin{aligned}\langle L\mathbf{v},\mathbf{v}\rangle &= c_1\langle L\mathbf{v}_1,\mathbf{v}\rangle + c_2\langle L\mathbf{v}_2,\mathbf{v}\rangle + \cdots + c_n\langle L\mathbf{v}_n,\mathbf{v}\rangle \\ &= |c_1|^2\langle L\mathbf{v}_1,\mathbf{v}_1\rangle + |c_2|^2\langle L\mathbf{v}_2,\mathbf{v}_2\rangle + \cdots + |c_n|^2\langle L\mathbf{v}_n,\mathbf{v}_n\rangle \\ &= \lambda_1|c_1|^2 + \lambda_2|c_2|^2 + \cdots + \lambda_n|c_n|^2\end{aligned}$$

which is positive unless all the coefficients c_i are 0.

For (1) $\Rightarrow$ (3), again let $\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n\}$ be an orthonormal basis of eigenvectors for L. By the first equivalence L is positive definite if and only if the λ_i are all positive. Then take for S the operator that maps $\mathbf{v}_i \mapsto \sqrt{\lambda_i}\mathbf{v}_i$. S is invertible if and only if all the λ_i are positive. It is obvious that S is self-adjoint, since in the orthonormal basis $\{\mathbf{v}_1,\ldots,\mathbf{v}_n\}$ its matrix is diagonal with real numbers on the diagonal and that $S^2 = L$. S is called the square root of L. Note that a positive semidefinite operator also has a square root, but it is invertible only if L is positive definite.

(3) $\Rightarrow$ (1) is a special case of Theorem 9.3.3. We reprove it in the language of operators. We take any operator S on V. S^*S is self-adjoint because

$$\langle S^*S\mathbf{v},\mathbf{w}\rangle = \langle S\mathbf{v},S\mathbf{w}\rangle = \langle v,S^*S\mathbf{w}\rangle.$$

It is positive semidefinite because, by the same computation,

$$\langle S^*S\mathbf{v},\mathbf{v}\rangle = \langle S\mathbf{v},S\mathbf{v}\rangle \geq 0$$

for all $\mathbf{v} \in V$. If S is invertible, then $S\mathbf{v} \neq \mathbf{0}$ when $\mathbf{v} \neq \mathbf{0}$. Then $\langle S\mathbf{v},S\mathbf{v}\rangle > 0$ because the scalar product itself is positive definite.

(1) $\Leftrightarrow$ (4) follows easily from the definition of the index of positivity. More generally, here is the definition of the more commonly used *inertia* of A: it is the triple of non-negative integers (n_+,n_-,n_0), where n_+ is the index of positivity, n_0 the index of nullity and $n_- = n - n_1 - n_3$ what could be called by analogy the index of negativity: thus if $\{\mathbf{v}_1,\ldots,\mathbf{v}_n\}$ is any orthogonal basis of V, then

1. n_+ is the number of basis elements $\mathbf{v}_i$ such that $\mathbf{v}_i^t A\mathbf{v}_i > 0$,
2. n_- is the number of basis elements $\mathbf{v}_i$ such that $\mathbf{v}_i^t A\mathbf{v}_i < 0$,
3. n_0 is the number of basis elements $\mathbf{v}_i$ such that $\mathbf{v}_i^t A\mathbf{v}_i = 0$,

The fact that these numbers do not depend on the choice of orthogonal basis is the content of Sylvester's Theorem. Thus, essentially by definition:

- A is positive definite if and only if its inertia is $(n,0,0)$.
- A is positive semidefinite if and only if its inertia is $(n_+,0,n_0)$.

This concludes the proof. □

There are two other useful criteria for positive definiteness.

First we recall a result for all square matrices proved in §12.10. We will use it in the next tests for positive (semi)definiteness. As there write the characteristic polynomial of the $n \times n$ matrix A over a field F as

$$P(t) = t^n - p_1 t^{n-1} + p_2 t^{n-2} - \cdots + (-1)^n p_n,$$

where $p_i \in F$ for all i, $1 \leq i \leq n$. Then by Theorem 12.10.3,

$$p_j = \sum_J \det A_J = \sum_J m_J,$$

where the sum is over all choices of j elements $J = \{i_1, \ldots, i_j\}$ from $[1,n]$, and A_J is the corresponding principal matrix of A, and m_J its determinant, called a principal minor. Thus the sum has $\binom{n}{j}$ terms.

We now turn to the last two tests for positive definiteness. The first result is

Theorem 13.5.5. *A symmetric matrix A is positive definite if and only if all its leading principal minors are positive. It is positive semidefinite if and only if all its principal minors are non-negative.*

Notice the subtle difference between the two cases: to establish that A is positive semidefinite, you need to check all the principal minors, not just the leading ones.

Example 13.5.6. Consider the matrix

$$\begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$$

The leading principal minors of this matrix are both 0, and yet it obviously not positive semidefinite, since its eigenvalues (0 and -1) are not both non-negative.

To prove Theorem 13.5.5 we use and extend the notation of Definition 12.10.1

Definition 13.5.7. Given $J = \{i_1, \ldots, i_k\}$ as in Theorem 12.10.3, for any n-vector $\mathbf{x} = (x_1, \ldots, x_n)$ let $\mathbf{x}_J$ be the k-vector $(x_{i_1}, x_{i_2} \ldots, x_{i_k})$. Let $\tilde{\mathbf{x}}_J$ be the n-vector whose i-th entry $\tilde{x}_i$ is given by

$$\tilde{x}_i = \begin{cases} x_i, & \text{if } i \in J; \\ 0, & \text{otherwise.} \end{cases}$$

Then it is clear that

$$\boxed{\tilde{\mathbf{x}}_J^t A \tilde{\mathbf{x}}_J = \mathbf{x}_J^t A_J \mathbf{x}_J.} \tag{13.6}$$

Example 13.5.8. If $n = 4$, $k = 2$ and $J = \{2,4\}$, then

$$\mathbf{x}_J = \begin{pmatrix} x_2 \\ x_4 \end{pmatrix}, \text{ and } \tilde{\mathbf{x}}_J = \begin{pmatrix} 0 \\ x_2 \\ 0 \\ x_4 \end{pmatrix}.$$

Since

$$A_J = \begin{pmatrix} a_{22} & a_{24} \\ a_{42} & a_{44} \end{pmatrix}$$

you can easily verify (13.6) in this case.

This allows us to prove:

Proposition 13.5.9. *If A is positive definite, the symmetric matrix A_J is positive definite. If A is positive samidefinite, then A_J is positive semidefinite.*

Proof. If A is positive definite, the left hand side of (13.6) is positive if $\tilde{\mathbf{x}}_J \neq \mathbf{0}$. So the right hand side is positive when $\mathbf{x}_J \neq \mathbf{0}$, since $\mathbf{x}_J$ is just $\tilde{\mathbf{x}}_J$ with $n-k$ zero entries removed. That is the definition of positive definiteness for A_J, so we are done. The positive semidefinite case is even easier. □

If A_J is positive definite, its determinant, which is the principal minor m_J of A, is positive: indeed the determinant is the product of the eigenvalues, which are all positive. This shows that all the principal minors are positive, and finishes the easy implication in the proof of Theorem 13.5.5. A similar argument handles the positive semidefinite case.

Before proving the more difficult implication of Theorem 13.5.5, we look at some examples.

Example 13.5.10. When the set J has just one element, so $k = 1$, we are looking at 1×1 principal minors. So we get: $m(i) = a_{ii} > 0$. However there are symmetric matrices with positive diagonal entries that are not positive definite. The matrix

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

is not positive definite: test the vector $(1,1)$:

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

When J has two elements, so $J = \{i, j\}$, we get:

Corollary 13.5.11. *Let A be a positive definite matrix. Then for $i \neq j$,*

$$\left|a_{ij}\right| \leq \sqrt{a_{ii}a_{jj}}.$$

In Example 13.5.10, $|a_{12}| = 2$, while $a_{11}a_{22} = 1$, so the matrix is not positive definite.

Example 13.5.12. The matrix

$$A = \begin{pmatrix} 2 & 0 & 0 & 2 \\ 0 & 4 & 3 & 0 \\ 0 & 3 & 4 & 0 \\ 2 & 0 & 0 & 2 \end{pmatrix}$$

is not positive definite, by applying the lemma to $i = 1$, $j = 4$. It is positive semidefinite, however.

A weaker result implied by this corollary is useful when just scanning the matrix.

Corollary 13.5.13. *If A is positive definite, the term of largest absolute value must be on the diagonal.*

Now we return to the proof of Theorem 13.5.5. In the positive definite case it remains to show that if all the leading principal minors of the matrix A are positive, then A is positive definite. Here is the strategy. From Sylvester's Theorem, if U be any invertible $n \times n$ matrix, then the symmetric matrix A is positive definite if and only if U^tAU is positive definite, since this is just a computation of the signature of the symmetric matrix. We have the following obvious facts for diagonal matrices , which guide us.

Proposition 13.5.14. *If the matrix A is diagonal, then it is positive definite if and only if all its diagonal entries are positive.*

This can be rewritten: A is positive definite if and only if all its its leading principal minors are positive.

Proof. If the diagonal entries are $\{d_1, d_2, \ldots, d_n\}$, the leading principal minors are $D_1 = d_1$, $D_2 = d_1d_2$, $\ldots$, $D_n = d_1d_2\cdots d_n$. So the positivity of the d_i is equivalent to that of the D_i. □

The key step in the proof of Theorem 13.5.5 is the following result:

Proposition 13.5.15. *If A has positive leading minors, it can be diagonalized by an invertible lower triangular matrix U with $u_{ii} = 1$ for all i. In other words UAU^t is diagonal.*

Proof. We rewrite the diagonalization algorithm for symmetric matrices 7.5.3 in terms of determinants: see §11.9. Then we examine why we can avoid using elementary matrices other than lower triangular matrices on the left side to diagonalize. This follows from Theorem 11.9.5 which expresses the pivots in terms of quotients of principal minors. □

More generally, the proof implies a result that is interesting in its own right.

Proposition 13.5.16. *Let A be any symmetric matrix that can be diagonalized by a product U of lower triangular matrices as in Proposition 13.5.15. Then the leading principal minors of $A' = UAU^t$ are equal to those of A.*

Proof. As before, this immediately follows from the fact that if you add to a row (or column) of a square matrix a multiple of another row (or another column), then the determinant of the matrix does not change. Just apply this to the leading principal minors. □

Exercise 13.5.17. State the result concerning negative definite matrices that is analogous to the main theorem, noting that Proposition 13.5.16 applies. Do the same for Theorem 13.5.4.

We now finish the proof of the main theorem in the positive–definite case.

Assume that A is a symmetric matrix whose leading principal minors D_k are all positive. Proposition 13.5.15 tells us that A can be diagonalized to a matrix $A^{(n-1)} = UAU^t$ by a lower diagonal matrix U with 1's on the diagonal The diagonal matrix $A^{(n-1)}$ obtained has all its diagonal entries $a_{11}, a_{22}^{(1)}, \ldots, a_{nn}^{(n-1)}$ positive, so it is positive definite by the easy Proposition 13.5.14. By Proposition 13.5.16 A is positive definite, so we are done in the positive–definite case.

We now prove Theorem 13.5.5 in the positive semidefinite case. Proposition 13.5.9 establishes one of the implications. For the other implication we prove:

Proposition 13.5.18. *If A is positive semidefinite, then $A + \varepsilon I$ is positive definite for any $\varepsilon > 0$.*

More generally we could prove that for any symmetric matrix A, there is a positive number c such that $A + cI$ is positive definite.

We also need Theorem 12.10.3. Write the characteristic polynomial of A as in (12.1):

$$P(t) = t^n - p_1 t^{n-1} + p_2 t^{n-2} - \cdots + p_n.$$

Since all the principal minors of A are non-negative, Theorem 12.10.3 says that all the p_i are non-negative. We have the elementary proposition:

Proposition 13.5.19. *Assume the characteristic polynomial of A is written as in* (12.1)*. Then if all the p_i are non-negative, A is positive semidefinite. If all the p_i are positive, then A is positive definite.*

Proof. We first note that all the roots of $P(t)$ are non negative, only using the non-negativity of the p_i. Assume we have a negative root λ. Then all the terms of $P(\lambda)$ have the same sign, meaning that if n is even, all the terms are non-negative, while if n is odd, all the terms are non-positive. Since the leading term λ^n is non-zero, this is a contradiction. Thus all the roots are non-negative, and A is therefore positive semidefinite by Theorem 13.5.4, (1). If the p_i are all positive (in fact if a single one of them is positive) then the polynomial cannot have 0 as a root, so by the same criterion A is positive definite. □

This concludes the proof of Theorem 13.5.5.

The last test is the characteristic polynomial test. We write the characteristic polynomial of A as in (12.1). With this notation, we get a new test for positive definiteness.

Theorem 13.5.20. *A is positive definite if and only if all the p_i, $1 \leq i \leq n$, are positive. A is positive semidefinite if and only if all the p_i, $1 \leq i \leq n$, are nonnegative.*

Proof. One implication follows immediately from Theorem 13.5.19.

For the reverse implication, we must show that if A is positive definite, then all the constants p_i are positive. This follows immediately from Theorem 12.10.3 and Proposition 13.5.9: all the principal minors are positive (non-negative) and the p_i are sums of them. □

13.6 The Spectral Theorem for Unitary Operators

In this section and the next one we prove a spectral theorem for isometries, meaning operators that are orthogonal over $\mathbb{R}$ or unitary over $\mathbb{C}$. Unusually we start with the complex case, which is easier and has a simpler result.

So let V be a Hermitian space of dimension n and L a unitary operator on V, as studied in §9.5. By definition $\langle L\mathbf{v}, L\mathbf{w}\rangle = \langle \mathbf{v}, \mathbf{w}\rangle$ for all $\mathbf{v}$ and $\mathbf{w}$ in V, and as we proved in Theorem 9.5.2 this is verified if and only if $\|L\mathbf{v}\| = \|\mathbf{v}\|$ for all $\mathbf{v} \in V$. We say L preserves length, or is an isometry. In particular L is invertible and its inverse L^{-1} is its adjoint by Theorem

Since we are over $\mathbb{C}$, L has an eigenvector $\mathbf{v}$: $L\mathbf{v} = \lambda\mathbf{v}$, for some complex number λ. Because L is an isometry, $|\lambda| = 1$, so that $\lambda = e^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta$.

The one-dimensional space spanned by the eigenvector $\mathbf{v}$ is invariant under L. We will apply the following theorem to it:

Theorem 13.6.1. *If L is unitary, and the subspace W of V is invariant under L, then its orthogonal complement $W^{\perp}$ is invariant under L.*

Proof. We must prove that if $\mathbf{u} \in W^{\perp}$, then $L\mathbf{u} \in W^{\perp}$, If $\mathbf{u} \in W^{\perp}$, then $\langle \mathbf{u}, \mathbf{w}\rangle = 0$ for all $\mathbf{w} \in W$. Because L is invertible $L^{-1}\mathbf{w} \in W$. Then

$$\langle L\mathbf{u}, \mathbf{w}\rangle = \langle \mathbf{u}, L^{*}\mathbf{w}\rangle = \langle \mathbf{u}, L^{-1}\mathbf{w}\rangle = 0$$

by hypothesis. □

We can apply this theorem to the space W generated by an eigenvector, so L restricts to an operator on $W^{\perp}$, which has dimension $n-1$. The restriction is unitary, so by induction on the dimension as in the self-adjoint case, we can construct a spectral basis of V consisting of unit eigenvectors for L: see Theorem 13.3.4. Now let A be the matrix of L in any orthonormal basis of V, so that A is a unitary matrix. Then

Theorem 13.6.2. *Let A be a unitary $n \times n$ matrix, U its matrix of unit eigenvectors, and λ its vector of eigenvalues. Then U is a unitary matrix, and $U^{*}AU$ is the diagonal matrix $D(\lambda_1, \ldots, \lambda_n)$.*

Proof. As before, the system of equations $AU = UD$ is simply a convenient way of writing all the eigenvector-eigenvalue equations. Since $U^{-1} = U^{*}$, we get $U^{*}AU = D$. □

13.7 The Spectral Theorem for Orthogonal Operators

Finally we treat the case of an orthogonal operator L on a Euclidean space V. This is slightly harder than the cases we have already considered, because we cannot

guarantee that L has an eigenvalue: consider the case of a rotation in $\mathbb{R}^2$. Fortunately, this turns out to be the only problem.

So we perform the same trick as at the end of §13.4. First we pick an orthonormal basis for V, and consider the matrix A of L in this basis. Since the real matrix A acts on the $\mathbb{R}^n$ of vector coordinates, it also acts on $\mathbb{C}^n$ in the same way. The orthogonal matrix A is obviously a unitary matrix acting on $\mathbb{C}^n$, so the results of the previous section apply. Let $\mathbf{v}$ be an eigenvector of A on $\mathbb{C}^n$ with eigenvalue λ. We know that λ is a complex number of modulus 1. If λ is real, it is ± 1. Otherwise we still have an eigenvalue-eigenvector equation $A\mathbf{v} = \lambda\mathbf{v}$, where both λ and $\mathbf{v}$ are complex. Since $|\lambda| = 1$ we can write $\lambda = \cos\theta + \mathrm{i}\sin\theta$. We also write $\mathbf{v} = \mathbf{x} + \mathrm{i}\mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are real vectors. Then take the complex conjugate of the eigenvalue-eigenvector equation. Because A is real we get: $A\overline{\mathbf{v}} = \overline{\lambda}\,\overline{\mathbf{v}}$. This says that $\overline{\lambda}$ is an eigenvalue of A with eigenvector $\overline{\mathbf{v}}$. Write everything out in terms of real and imaginary parts:

$$\begin{aligned} A\mathbf{x} + \mathrm{i}A\mathbf{y} &= (\cos\theta + \mathrm{i}\sin\theta)(\mathbf{x} + \mathrm{i}\mathbf{y}) \\ &= \cos\theta\mathbf{x} - \sin\theta\mathbf{y} + \mathrm{i}(\cos\theta\mathbf{y} + \mathrm{i}\sin\theta\mathbf{x}). \end{aligned}$$

Take the real and imaginary parts of this equation:

$$\begin{aligned} A\mathbf{x} &= \cos\theta\mathbf{x} - \sin\theta\mathbf{y}; \\ A\mathbf{y} &= \cos\theta\mathbf{y} + \sin\theta\mathbf{x}. \end{aligned}$$

Because λ is not real, the vectors $\mathbf{x}$ and $\mathbf{y}$ are linearly independent. This equations say that A leaves the plane spanned by $\mathbf{x}$ and $\mathbf{y}$ invariant. The matrix of this restriction using the basis $\{\mathbf{x}, \mathbf{y}\}$ is

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{13.7}$$

which is the matrix of the rotation by θ radians.

The analog of Theorem 13.6.1 holds for orthogonal matrices, so by repeating this construction we can construct an orthonormal basis on V so that in this basis the orthogonal operator is block diagonal, with one-dimensional blocks corresponding to the real eigenvalues, and two-dimensional blocks (13.7) of rotations.

Note that the orthogonal transformations of determinant one play a special role, since the composition of two of them is also an orthogonal transformation of determinant. These are sometimes called proper orthogonal transformations or rotations.

Example 13.7.1. Consider orthogonal operators L on $\mathbb{R}^2$. Then L has

- either two real eigenvalues, which must be either 1 or -1. If they are both 1, then the transformation is the identity. If they are both -1, then the linear transformation is rotation by π. The most interesting case occurs when one eigenvalue is 1 and the other is -1, in which case the operator is called a reflection along the eigenvector $\mathbf{e}_1$ with eigenvalue 1. Let $\mathbf{v}$ be any vector in $\mathbb{R}^2$ and let $c\mathbf{e}$ be the component of v in the one-dimensional eigenspace spanned by $\mathbf{e}$, so that $\langle \mathbf{v} - c\mathbf{e}_1, \mathbf{e}_1 \rangle = 0$. Then the image of $\mathbf{v}$ under L is $\mathbf{v} - 2c\mathbf{e}_1$ as you should check. The determinant of L is -1.

- or no real eigenvalues, in which case it is a rotation by an angle θ for which both $\cos\theta$ and $\sin\theta$ are non-zero. The determinant is 1.

Example 13.7.2. Now consider orthogonal operators L on $\mathbb{R}^3$.

- If the eigenvectors of L are 1, 1 and -1, describe the motion as a reflection: in which linear subspace?
- The most interesting case is that where there is one real eigenvector $\mathbf{e}$ with eigenvalue ± 1, and one block of type (13.7). The orthogonal complement of $\mathbf{e}$ is a two-dimensional subspace V, so we have reduced to Example 13.7.1 in V. If the eigenvalue of $\mathbf{e}$ is 1, then we have a rotation in V The construction shows that to specify a rotation in $\mathbb{R}^3$ you need an axis of rotation (spanned by the eigenvector $\mathbf{e}$) and the angle θ given by the complex eigenvectors in the plane perpendicular to $\mathbf{e}$.

Exercise 13.7.3. Prove all assertions in Examples 13.7.1 and 13.7.2. In the last example, if the eigenvector of $\mathbf{e}$ is -1, describe the motion of $\mathbb{R}^3$ obtained in terms of rotations and reflections.

13.8 The Spectral Theorem for Normal Operators

We defined normal operators and normal matrices in §9.6: the operator L is normal if it commutes with its adjoint: $LL^* = L^*L$, and the square matrix A is normal if it commutes with its conjugate transpose: $A\overline{A^*} = \overline{A^*}A$. We proved some preliminary results on normal operators in §9.6, which you should review now.

The first technical result is:

Theorem 13.8.1. *If V is an inner product space, and L a normal operator on V, then $\mathbf{v}$ is an eigenvector for L with eigenvalue λ if and only if $\mathbf{v}$ is an eigenvector for L^* with eigenvalue $\overline{\lambda}$.*

Proof. The easy but key remark is that $\mathbf{v}$ is an eigenvector of L with eigenvalue λ if and only if

$$\|(L-\lambda I)\mathbf{v}\| = 0.$$

The operator $L-\lambda I$ is normal for any $\lambda \in \mathbb{C}$, since its conjugate is $L^* - \overline{\lambda} I$:

$$\langle (L-\lambda I)\mathbf{v}, \mathbf{w}\rangle = \langle L\mathbf{v}, \mathbf{w}\rangle - \langle \lambda\mathbf{v}, \mathbf{w}\rangle = \langle \mathbf{v}, L^*\mathbf{w}\rangle - \langle \mathbf{v}, \overline{\lambda}\mathbf{w}\rangle = \langle \mathbf{v}, (L^* - \overline{\lambda} I)\mathbf{w}\rangle.$$

By Theorem 9.6.5, we have

$$\|(L-\lambda I)\mathbf{v}\| = \|(L^* - \overline{\lambda} I)\mathbf{v}\|$$

for any $\mathbf{v} \in V$. We are done. □

We will be mostly interested in complex normal operators. Recall (Theorem 13.4.1) that for any operator L on a Hermitian space V there is a unitary basis $\mathbf{v}_1$,

$\ldots$, $\mathbf{v}_n$ in which the matrix A representing L is upper triangular. Then in that basis the matrix of a normal operator is diagonal. This is a natural generalization of the statement that a symmetric or Hermitian matrix that is upper triangular is diagonal.

Theorem 13.8.2. *If L is a normal operator on the Hermitian space V of dimension n, then in any basis such that the matrix A representing L is upper triangular, it is in fact diagonal.*

Proof. Because A is upper triangular, the column vector $(0,\ldots,0,1)$ is an eigenvector for A with eigenvalue a_{nn}. By Theorem 13.8.1 it is also an eigenvector for the conjugate transpose $\overline{A^t}$ of A. Thus the last row of $\overline{A^t}$ has zeroes in all positions except the n-th where it has $\overline{a_{nn}}$. Therefore A is block diagonal, where the top, left square block A^{11} of size $n-1$ is upper-triangular, the two off-diagonal blocks A^{21} and A^{12} are both zero, the the bottom right square block A^{22} of size 1 is just (a_{nn}). By induction we may assume that the result is true for A^{11}, so that it is diagonal. Therefore A is diagonal. □

This gives us the spectral theorem for complex normal operators. Conversely we see that normal operators are the only operators for which such a theorem is true.

Corollary 13.8.3. *If L is a normal operator on the Hermitian space V of dimension n, then V has a spectral basis for L, meaning a unitary basis of eigenvectors for L. Therefore if A is the matrix for L in a spectral basis, U is the spectral basis, and D is the diagonal matrix of eigenvalues, then $U^{-1}AU = D$.*

Conversely if L is a complex operator on a Hermitian space such that there is a unitary basis in which the matrix representing L is diagonal, then L is normal.

Proof. We only need to prove the easy converse: a diagonal matrix D is obviously normal, since D^* is also diagonal, so the operator represented by D in the basis is normal. □

Thus we know exactly which operators admit a spectral theorem. We could easily formulate such a result in the real case too, but it could have to be more complicated, since one cannot diagonalize orthogonal operators.

We may ask how to characterize the various subclasses of normal matrices using their eigenvalues.

Theorem 13.8.4. *Let L be a normal operator on a Hermitian space V. Then*

1. *L is Hermitian if and only if its eigenvalues are real;*
2. *L is unitary if and only if its eigenvalues have modulus* 1.
3. *L is positive definite if and only if its eigenvalues are positive;*
4. *L is positive semidefinite if and only if its eigenvalues are non-negative;*

Proof. In each case one direction has already been established. So we establish the other implication. By assumption L is normal, so it is diagonalizable in all four cases. So choose a spectral basis for V, and let D be the diagonal matrix for A in this basis. The diagonal entries of D are the eigenvalues of L. It is now almost trivial to

establish the result. For example, assume that the entries of D are all real. Then the matrix D is self-adjoint, so the operator L it represents is self adjoint. Now assume all the entries of D are real and positive. To show that L is positive definite, we must show that $\langle L\mathbf{v}, \mathbf{v}\rangle > 0$ for all non-zero $\mathbf{v}$. Just expand $\mathbf{v}$ in terms of the basis $\{\mathbf{v}_i\}$ of eigenvectors to get

$$\langle L\mathbf{v}, \mathbf{v}\rangle = \lambda_1 \|\mathbf{v}_1\|^2 + \cdots + \lambda_n \|\mathbf{v}_n\|^2.$$

For this to be positive for any non-zero $\mathbf{v}$, all the λ_i must be positive. The same argument works in the positive-semidefinite case. If all the eigenvectors have modulus 1, then using the expansion of any vector $\mathbf{v}$ in terms of eigenvectors, we see that the operator L preserves length. So by Theorem 9.5.2 we are done. □

What can we say about normal operators on a Euclidean space? To be written....

Finally what can we say about the minimal polynomial of a normal operator on an inner product space? Over $\mathbb{C}$, Corollary 13.8.3 says that the operator is diagonalizable. As we have seen, the minimal polynomial of a diagonal matrix is simply the product of $(x-\lambda_i)$, where the product is over all the distinct eigenvalues of L.

13.9 The Polar Decomposition

In this section we do the real and complex case simultaneously. Here V is an inner product space, L an arbitrary operator on V, and L^* its adjoint. First we assume L is invertible.

Theorem 13.9.1 (Polar Decomposition). *There is a unique positive definite transformation S and a unique isometry Q such that $L = SQ$.*

Proof. By Theorem 9.3.2 the self adjoint operator L^*L is positive definite. Thus by Theorem 13.5.4 there is a unique self-adjoint operator S such that $S^2 = L^*L$, and S is invertible. Then consider the operator $Q = LS^{-1}$. By Theorem 9.1.4 or 9.2.3 $Q^* = S^{-1}L*$, since S is self-adjoint. Then

$$Q^*Q = S^{-1}L*LS^{-1} = S^{-1}S^2S^{-1} = I$$

so by Theorem 9.4.5 or 9.5.3 Q is an isometry and we are done. □

A slightly more involved argument shows that if L is not invertible it can be written $L = SQ$ where S is again the uniqueness determined square root of S, which is only positive semidefinite, and Q is an isometry.

This theorem is called the polar theorem because it generalizes the decomposition of a complex number (the arbitrary transformation L) into a positive real number (the positive definite S) and a complex number of length 1 (the isometry Q).

13.10 The Singular Value Decomposition

For the final construction of this chapter we will work over $\mathbb{R}$ for simplicity and clarity only. Our goal is to generalize past constructions to matrices that are not square, so do not correspond to operators, contrary to the theme of the rest of the chapter.

In §5.4 we studied linear maps L of a vector space of dimension n to a vector space of dimension m and showed how to choose bases of the two spaces so that the matrix of L in these bases is diagonal. Now we want to do something similar, but for inner product spaces, where we only allow orthogonal bases. It is not obvious that diagonalization can be achieved, but we now show it can.

So let $L\colon V \to W$ be a linear map from a Euclidean space V of dimension n and a Euclidean space W of dimension m. After picking a orthonormal basis for each space, we can view them as $\mathbb{R}^n$ and $\mathbb{R}^m$ with the standard inner product. Let A be the matrix of L in these bases.

So A be an arbitrary real $m \times n$ matrix. Consider the square matrix $S = A^t A$ of size n. It is symmetric, so we can apply the spectral theorem: it has real eigenvalues $\lambda_1, \ldots, \lambda_n$ corresponding to eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ that are mutually orthogonal and of length 1. So

$$\mathbf{v}_i^t S \mathbf{v}_i = \lambda_i \mathbf{v}_i^t \mathbf{v}_i = \lambda_i \quad \text{and when } i \neq j, \quad \mathbf{v}_i^t S \mathbf{v}_j = \lambda_i \mathbf{v}_i^t \mathbf{v}_j = 0, \tag{13.8}$$

since the $\mathbf{v}_i$ form an orthonormal basis of $\mathbb{R}^n$. On the other hand we also have

$$\mathbf{v}_i^t S \mathbf{v}_i = \mathbf{v}_i^t A^t A \mathbf{v}_i = (A\mathbf{v}_i)^t (A\mathbf{v}_i), \tag{13.9}$$

the length squared of a vector. Since $\mathbf{v}_i^t \mathbf{v}_i$ is positive, in fact $= 1$, this shows that $\lambda_i \geq 0$. Finally when $i \neq j$,

$$(A\mathbf{v}_i)^t (A\mathbf{v}_j) = \mathbf{v}_i^t S \mathbf{v}_j = \lambda_i \mathbf{v}_i^t \mathbf{v}_j = 0. \tag{13.10}$$

So $A^t A$ is a positive semidefinite matrix of size n. Let r be the number of positive eigenvalues. We reorder the basis so they come first: they correspond to $\mathbf{v}_1, \ldots, \mathbf{v}_r$. By Gram-Schmidt we can complete this with vectors $\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n$ to an orthonormal basis of V. The matrix U whose columns are the $\mathbf{v}_j$ is orthogonal by construction.

Now we construct an orthonormal basis for W. Let $\sigma_i = \sqrt{\lambda_i}$, $1 \leq i \leq r$, and consider the elements

$$\mathbf{q}_i = \frac{A\mathbf{v}_i}{\sigma_i}, \quad 1 \leq i \leq r.$$

By (13.9) and (13.10) these vectors in W have length 1. By (13.10) these r vectors are mutually orthogonal:

$$\langle \mathbf{q}_i, \mathbf{q}_j \rangle = \mathbf{q}_i^t \mathbf{q}_j = \frac{\mathbf{v}_i^t A^t A \mathbf{v}_j}{\sigma_i \sigma_j} = 0.$$

By Gram-Schmidt again, we can complete $\mathbf{q}_1, \ldots, \mathbf{q}_r$ with vectors $\mathbf{q}_{r+1}, \ldots, \mathbf{q}_m$ to get an orthonormal basis of W. Let Q be the square matrix of size m with columns the $\mathbf{q}_i$. Then Q is an orthogonal matrix. On the other hand $A\mathbf{v}_j = 0$, for $r+1, \ldots n$. This makes it obvious that in the basis of the $\mathbf{v}_j$ for V and the basis $\mathbf{q}_i$ for W the matrix of L is diagonal with the elements σ_i, $1 \leq i \leq r$ as the only non-zero entries in the first r diagonal positions.

Finally let Σ be the $m \times n$ matrix whose only non-zero terms are the first r diagonal elements, which are σ_i. The change of basis computation from §5.3 tells us that is our original basis the matrix of L is

$$A = Q\Sigma U^{-1} = Q\Sigma U^t \tag{13.11}$$

since U is orthogonal.

STATE the theorem here.

By construction the basis of V used for A is a spectral basis for A^tA. Similarly the basis used for W is a spectral basis for AA^t. Indeed by (13.9)

$$AA^t = Q\Sigma U^t U\Sigma Q^t = Q\Sigma^2 Q^t$$

and by taking transposes in the opposite order

$$A^tA = U\Sigma Q^t Q\Sigma U^t = U\Sigma^2 U^t.$$

Finally this shows that the eigenvalues of AA^t and A^tA are the same: the diagonal entries of Σ^2, which are the λ_i. The rank of A is of course r.

The σ_i are invariants of the transformation L. Therefore linear transformations on Euclidean spaces can be partitioned into equivalence classes corresponding first to it rank, and second to the positive numbers $\sigma_1, \ldots, \sigma_r$, which are called the singular values of L.

The most important application of the SVD is the definition of the pseudoinverse of an arbitrary matrix.

Definition 13.10.1. If the $n \times n$ matrix A has singular value decomposition $A = Q^t\Sigma U$, then its pseudoinverse is the $n \times m$ matrix

$$A^+ = U^t\Sigma^+ Q$$

where Σ^+ is $n \times m$ quasi diagonal matrix with entries $1/\sigma_i$ on the diagonal, where the σ_i are the diagonal entries of Σ.

So

$$A^+A = U^t\Sigma^+ QQ^t\Sigma U = U^t\Sigma^+\Sigma U = U^t I_r U$$

and

$$AA^+ = Q^t\Sigma^+ UU^t\Sigma Q = Q^t\Sigma^+\Sigma Q = Q^t I_r Q$$

This shows that if A is square and invertible then $A^+A = AA^+ = I$.

Obviously $A^{++} = A$.

Chapter 14
The Method of Least Squares

Abstract In Chapter 8 we showed how to determine the 'best' approximation for the solution of an inconsistent system of linear equations. Here we show how this is equivalent to one of the most important minimization techniques in science: the method of least squares. This method is useful in statistics, as it leads to the regression line. In §14.3 we show how to interpret what we have done into the language of data measurement. Then we study a related but more symmetric minimization problem that goes under the name orthogonal least squares. Finally we report briefly on the methods for computing the least squares solution.

14.1 The Method of Least Squares

Here we give an application of the methods of §8.5. First some background.

Assume you run an experiment that has n inputs $q_1, \ldots, q_n$ and one output r. You have reason to believe, or are trying to confirm, that when measured in the appropriate units, the output of the experiment satisfies an affine equation:

$$r = a_0 + a_1 q_1 + \cdots + a_n q_n.$$

You do not know the coefficients $a_0, a_1, \ldots, a_n$ of the affine equation, which determine the relationship between the output r and the inputs $q_1, \ldots, q_n$. You wish to determine them by experiment. In each experiment there is measurement error, so you run the experiment a large number of times, say $m > n+1$ times. Let r_i and a_{i1}, $\ldots a_{in}$ be the results obtained in the i-th experiment. In each experiment we want to use the same coefficients a_i. For these coefficients, in the i-th equation, instead of getting r_i when using $q_{i1}, \ldots q_{in}$, you get the linear combination

$$s_i = a_0 + a_1 q_{i1} + a_2 q_{i2} + \cdots + a_n q_{in}$$

of the inputs. In matrix notation if Q is the $m \times n$ matrix

$$Q = \begin{pmatrix} 1 & q_{11} & \dots & q_{1n} \\ 1 & q_{21} & \dots & q_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & q_{m1} & \dots & q_{mn} \end{pmatrix}$$

and $\mathbf{s}$ the m-vector with entries s_i, and $\mathbf{a}$ the $(n+1)$-vector with entries a_j, the equation can be written $\mathbf{s} = Q\mathbf{a}$.

The question is: how to choose the coefficients a_i to minimize the sum of the errors $s_i - r_i$, in some sense. We want to make all the errors non-negative, so there is no cancelation. So you could minimize the sum of the absolute value of the errors. Instead, here we minimize the sum of the square $(s_i - r_i)^2$, hence the name 'method of least squares'. So we minimize the expression

$$\sum_{i=1}^{m} (r_i - s_i)^2 = (\mathbf{r} - Q\mathbf{a})^t (\mathbf{r} - Q\mathbf{a}) \tag{14.1}$$

the dot product of the vector $(\mathbf{r} - Q\mathbf{a})$ with itself. It can be written

$$\mathbf{a}^t Q^t Q\mathbf{a} - 2\mathbf{a}^t Q\mathbf{r} + \mathbf{r}^t \mathbf{r}$$

The variables are denoted $a_0, \dots, a_n$ while the constants are denoted r_i and q_{ij}. This is the same equation as Remark 8.5.6. Equation (14.1) is quadratic in each one of the $n+1$ variables a_j, so we can take the partial derivatives without difficulty, and a minimum can only occur when the partials with respect to a_j are 0, getting

$$Q^t Q\mathbf{a} - Q^t \mathbf{r} = 0.$$

These are the same normal equations as in the previous section.

We assume that m is greater than $n+1$. To apply the method of §8.5, we need to know Q has rank $n+1$. Then $Q^t Q$ is invertible matrix of size $n+1$ and the unique minimum is

$$\mathbf{a} = (Q^t Q)^{-1} Q\mathbf{r}. \tag{14.2}$$

This gives us the coefficients of the affine space that is the best fit to the data according to the method of least squares.

14.2 Fitting to a Line

The most important case of the method of least squares occurs when the data consists of only one input variable denoted q, and only output variable r. So this is a special case of the one considered in §14.1

We wish to fit our data to a line $r = a_0 + a_1 q$. By choosing to write our line in this way, we exclude the possibility that the line be vertical, but that is of little importance. Let's repeat the derivation of §14.1 in this special case. So we use the

method of least squares to fit a line $r = a_0 + a_1 q$ to m data points (q_i, r_i), which we can think of as m points in the plane

Give names to the sums of the coordinates of the data points:

$$s_q = \sum_{i=1}^{m} q_i \quad \text{and} \quad s_r = \sum_{i=1}^{m} r_i. \tag{14.3}$$

and to the squares:

$$S_{qq} = \sum_{i=1}^{m} q_i^2, \quad S_{qr} = \sum_{i=1}^{m} q_i r_i, \text{ and } S_{rr} = \sum_{i=1}^{m} r_i^2. \tag{14.4}$$

Definition 14.2.1. The average or mean $(\overline{q}, \overline{r})$ of the data is given by

$$\overline{q} = \frac{\sum_i q_i}{m} \quad \text{and} \quad \overline{r} = \frac{\sum_i r_i}{m}.$$

The point $(\overline{q}, \overline{r})$ is called the centroid of the data. Note that $\overline{q}$ and $\overline{r}$ are scalars, not vectors.

The matrix Q in our special case is

$$Q = \begin{pmatrix} 1 & q_1 \\ 1 & q_2 \\ \vdots & \vdots \\ 1 & q_m \end{pmatrix}$$

so

$$Q^t Q = \begin{pmatrix} m & s_q \\ s_q & S_{qq} \end{pmatrix} \tag{14.5}$$

which has determinant $mS_{qq} - s_q^2$. Since we assume Q has rank 2, we know that $Q^t Q$ is positive definite, so that $mS_{qq} - s_q^2 > 0$.

How does this computation compare with what we did in the general case? Equation (14.2) is written

$$\begin{pmatrix} m & s_q \\ s_q & S_{qq} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} s_r \\ S_{qr} \end{pmatrix}$$

The inverse of $Q^t Q$ can be computed explicitly as

$$\frac{1}{mS_{qq} - s_q^2} \begin{pmatrix} S_{qq} & -s_q \\ -s_q & m \end{pmatrix}.$$

The main computation of this section gives:

$$\begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \frac{1}{mS_{qq} - s_q^2} \begin{pmatrix} S_{qq} & -s_q \\ -s_q & m \end{pmatrix} \begin{pmatrix} s_r \\ S_{qr} \end{pmatrix} = \frac{1}{mS_{qq} - s_q^2} \begin{pmatrix} S_{qq} s_r - S_{qr} s_q \\ -s_q s_r + m S_{qr} \end{pmatrix}$$

or

$$a_0 = \frac{S_{qq}s_r - S_{qr}s_q}{mS_{qq} - s_q^2} \quad \text{and} \quad a_1 = \frac{mS_{qr} - s_q s_r}{mS_{qq} - s_q^2}. \tag{14.6}$$

We are done, having found simple explicit formulae for a_0 and a_1.

Remark 14.2.2. The minimizing line passes through the centroid of the points. Indeed just change coordinates so the centroid is the origin: then s_q and s_r are 0. Then (14.6) says that $a_0 = 0$, so the minimizing line goes through the origin. In §14.4 we will proceed differently: we will first show that the minimizing line must go through the centroid, and then minimize over all lines going through the centroid. We could have done that here, too. Consult Proposition 14.4.2.

Corollary 14.2.3. *Assume you have performed the experiment m times, and have computed s_q, s_r , S_{qq} and S_{qr} for those m experiments. Now perform the experiment one more time, getting values q_{m_1} and r_{m_1}. You can update* (14.6) *easily since*

$$\begin{aligned} s_q &\mapsto s_q + q_{m+1} \\ s_r &\mapsto s_r + r_{m+1} \\ S_{qq} &\mapsto S_{qq} + q_{m+1}^2 \\ S_{qr} &\mapsto S_{qr} + q_{m+1}r_{m+1} \end{aligned}$$

Exercise 14.2.4. Write done an explicit example with $m = 3$ and corresponding values. Then add a fourth row to Q and do the update.

Exercise 14.2.5. Show that if you change units by the same factor in both q and r, then the slope does not change, and b scales by the same factor.

Remark 14.2.6. Interchange the coordinates $q_i \leftrightarrow r_i$ for all m points. Then the slope of the minimizing line for the new configuration of points is

$$a = \frac{mS_{qr} - s_q s_r}{nS_{rr} - s_r^2}.$$

This is the slope if you use the method of least squares with the role of the variables q and r interchanged. Note that it is different from the original formula unless

$$mS_{qq} - s_q^2 = mS_{rr} - s_r^2.$$

Let's work out completely one example where we have run the experiment 3 times. The measurements for the input variable q are $(1,2,3)$ and the output variable r are $(1,1,3)$.

Solve and draw a graph and show what the errors are on the graph.

Compare the two answers for the special case above.

14.3 Connection to Statistics

This section is for readers who want to see the connection of this material to elementary statistics: measurement error, correlation and regression. More details, useful for novices in statistics, are given in [7], especially Part III. Most of Part One of [26] gives a detailed history of the method of least squares as expounded first by Legendre in 1805. Chapter IX of [32] gives a different, more technical historical coverage of Least Squares. Finally for an advanced mathematical treatment see [30] Chapter VII. The main point is to connect what we do to the cognate concepts in statistics.

This is mainly a question of giving new names to concepts we have already defined in §14.2. So for example we have a random variables **x** that we sample m times, to get measurements x_i.

For simplicity, we only cover the case of a *scatter diagram* in the plane, meaning a collection of points (x_i, y_i), $1 \leq i \leq n$. These points represent the data that has been collected. For example each x_i represents the height of a mother, and y_i the height of her daughter, or x_i is time and y_i the measurement of the quantity of some good produced at time x_i. We write **x** for the vector $(x_1, x_2, \ldots, x_n)$, and similarly for **y**. We consider **x** as the independent variable and **y** as the dependent variable.

We can take the mean of the samples for **x**, which is by definition

$$\overline{x} = \frac{\sum_{i=1}^{m} x_i}{m}.$$

This is s_x/m,writing as in §14.2 $s_x = \sum_{i=1}^{m}(x_i$. Then the *variance* of this sample of the random variable is by definition

$$\sigma_{xx} = \frac{\sum_{i=1}^{m}(x_i - \overline{x})^2}{m},$$

the mean of the squares of the values centered at their mean. The standard deviation $\sigma_{\mathbf{x}}$ of **x** is the square root of the variance: its advantage is that it has the same dimension as **x**. Writing as in §14.2 $S_{xx} = \sum_{i=1}^{m}(x_i^2$, the variance is

$$\frac{\sum_{i=1}^{m}(x_i^2 - 2\overline{x}x_i + \overline{x}^2)}{m} = \frac{S_{xx} - \overline{x}s_x}{m} = \frac{S_{xx} - s_x^2/m}{m} = \frac{mS_{xx} - s_x^2}{m^2}.$$

as a little computation shows. We can of course define the same quantities for **y**.

The *covariance* of **x** and **y** is

$$\sigma_{xy} = \frac{\sum_{i=1}^{m}(x_i - \overline{x})(\sum_{i=1}^{m}(y_i - \overline{y})}{m},$$

so that the variance is the special case $\mathbf{x} = \mathbf{y}$ of the covariance. In the notation of §14.2 the covariance is

$$\sigma_{xy} = \frac{mS_{xy} - s_x s_y}{m^2}.$$

Our main goal is to write the coefficients a_0 and a_1 of the line found by the method of least squares in terms of the variables we have just introduced. They are given by (14.6), so the slope of the line is

$$a_1 = \frac{mS_{xy} - s_x s_y}{mS_{xx} - s_x^2} = \frac{\sigma_{xy}}{\sigma_{xx}}.$$

Finally since

$$\sigma_{xx}\overline{x} - \sigma_{xy}\overline{x} = \left(\frac{mS_{xx} - s_x^2}{m^2}\right)\frac{s_y}{m} - \left(\frac{mS_{xy} - s_x s_y}{m^2}\right)\frac{s_x}{m} = \frac{S_{xx}s_y}{m^2} - \frac{S_{xy}s_x}{m^2}$$

we get for the constant term:

$$a_0 = \frac{S_{xx}s_y - S_{xy}s_x}{mS_{xx} - s_x^2} = \frac{\sigma_{xx}\overline{y} - \sigma_{xy}\overline{x}}{\sigma_{xx}} = \overline{y} - \frac{\sigma_{xy}\overline{x}}{\sigma_{xx}}$$

This line is called the *recession line*.

We defined the centroid whose coordinates are the means $(\overline{x}, \overline{y})$ in Definition 14.2.1.

It is easy to see what happens to the centroid if you change the coordinate system in two important ways.

1. Suppose we move the origin of the coordinate system to the point $(-a, -b)$. In this new coordinate system, for all i, x_i is replaced by $x_i + a$, which we can write in vector notation as $\mathbf{x}' = \mathbf{x} + \mathbf{a}$, for the vector $\mathbf{a}$ with the constant a in all entries. Then the mean $\overline{x+a} = \overline{x} + a$, and $\overline{y+\mathbf{b}} = \overline{y} + b$.
2. Suppose the scales on the axes are changed independently by factors of c and d, so that any point with coordinates (x, y) now has coordinates (cx, dy). Then for all i, the coordinates of (x_i, y_i) become (cx_i, dy_i), and $(\overline{cx}, \overline{dy}) = (c\overline{x}, d\overline{y})$.

We defined the standard deviation, also called the *root mean square* of the data, as the square root of the variance, so:

Definition 14.3.1.

$$\sigma_{\mathbf{x}} = \sqrt{\frac{\sum_i (x_i - \overline{x})^2}{n}} \quad \text{and} \quad \sigma_{\mathbf{y}} = \sqrt{\frac{\sum_i (y_i - \overline{y})^2}{n}}.$$

Using the same notation as for the average, and letting $\mathbf{a}$ be the constant variable of value a we get

1. $\sigma_{\mathbf{x}+\mathbf{a}} = \sigma_{\mathbf{x}}$;
2. $\sigma_{a\mathbf{x}} = |a|\sigma_{\mathbf{x}}$.

Next we convert to the standard deviation coordinate system, meaning that we translate the origin so that it is at $(\overline{x}, \overline{y})$ and rescale the axes so that the unit of measurement on the x axis is $\sigma_{\mathbf{x}}$, and on the y axis is $\sigma_{\mathbf{y}}$. Thus a point $\mathbf{a} = (a_1, a_2)$ in the original coordinate system has coordinates

$$\left(\frac{a_1-\bar{x}}{\sigma_{\mathbf{x}}}, \quad \frac{a_2-\bar{y}}{\sigma_{\mathbf{y}}}\right) \tag{14.7}$$

in the standard deviation system associated to the data points (x_i, y_i).

In particular if the data (x_i, y_i) is in standard deviation coordinates for itself, then by definition

$$\sqrt{\frac{\sum_i x_i^2}{n}} = 1 \quad \text{and} \quad \sqrt{\frac{\sum_i y_i^2}{n}} = 1$$

or $\sum_i x_i^2 = n$ and $\sum_i y_i^2 = n$.

Let the *standard deviation line* be the line in the standard deviation coordinate system that goes through the origin and has slope 1. Then in the original coordinate system this line goes through the centroid $(\bar{x}, \bar{y})$ and has slope $\frac{\sigma_{\mathbf{y}}}{\sigma_{\mathbf{x}}}$.

Exercise 14.3.2. Generalize (14.7) to a change to an arbitrary coordinate system, by which I mean one where the origin is chosen to be any point, and the scale on the axes are changed independently.

Next we define the *correlation* of $\mathbf{x}$ and $\mathbf{y}$ to be the covariance divided by the standard deviations:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}}$$

Theorem 14.3.3. *The correlation ρ_{xy} satisfies three properties:*

1. *it is invariant under adding the same number to each of the x_i or each of the y_i;*
2. *it is invariant under multiplication of all of the x_i or all of the y_i by the same positive number.*
3. *it is invariant under the interchange $\mathbf{x} \leftrightarrow \mathbf{y}$;*

The first two properties say that r is independent of the coordinate system used. Therefore we drop the subscripts. The last one says that it is symmetric.

Proposition 14.3.4. *Furthermore $-1 \le \rho_{xy} \le 1$.*

Proof. We may as well work in standard deviation coordinates, so we assume our data (x_i, y_i), $1 \le i \le n$, satisfies $\bar{x} = 0$, $\bar{y} = 0$, $\sum_i x_i^2 = n$ and $\sum_i y_i^2 = n$. We need to show

$$-n \le \sum_i x_i y_i \le n.$$

The key idea is to expand the expression

$$\sum_{i=1}^{n} (x_i \pm y_i)^2 = \sum_{i=1}^{n} x_i^2 \pm 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2 = 2n \pm 2\sum_{i=1}^{n} x_i y_i,$$

remembering that we are in standard deviation coordinates. Since this is a sum of squares, it is non-negative. So $n - \sum_{i=1}^{n} x_i y_i \ge 0$, showing $r \le 1$ and $n + \sum_{i=1}^{n} x_i y_i \ge 0$, showing $r \ge -1$. □

Exercise 14.3.5. Show that this proposition is just the Cauchy-Schwarz inequality in disguise.

Example 14.3.6. What does it mean for ρ_{xy} to be 1? The previous proof shows that it implies $x_i = y_i$ for all i in standard deviation coordinates. Similarly for the other extreme case: if $r = -1$, then $x_i = -y_i$ for all i in standard deviation coordinates.

We now can compare the slope $\frac{\sigma_{xy}}{\sigma_{xx}}$ of the recession line with the slope $\frac{\sigma_{\mathbf{y}}}{\sigma_{\mathbf{x}}}$ of the standard deviation line. A little computation shows that

$$\frac{\sigma_{xy}}{\sigma_{xx}} = \rho_{xy}\frac{\sigma_{\mathbf{y}}}{\sigma_{\mathbf{x}}}.$$

Since $|\rho_{xy}| \leq 1$ the recession line has smaller slope than the standard deviation line. Thus if the correlation ρ_{xy} between the $\{x_i\}$ and the $\{y_i\}$ is 0, then the least squares line is horizontal.

Now there is a second regression line, where we interchange the roles of x and y. It goes through the same point $(\bar{x},\bar{y})$, but its slope is $\frac{1}{\rho_{xy}}\frac{\sigma_{\mathbf{y}}}{\sigma_{\mathbf{x}}}$, therefore greater than the slope of the standard deviation line.

Now we consider the case of n variables for which we have m observations. Then we get a $m \times n$ matrix X whose columns $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are the m measurements for each variable. Assume we are in centroid coordinates.

Theorem 14.3.7. *Then the symmetric matrix*

$$\frac{X^tX}{m^m} = \begin{pmatrix} \sigma_{x_1x_1} & \sigma_{x_1x_2} & \cdots & \sigma_{x_1x_n} \\ \sigma_{x_2x_1} & \sigma_{x_2x_2} & \cdots & \sigma_{x_2x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_mx_1} & \sigma_{x_mx_2} & \cdots & \sigma_{x_mx_n} \end{pmatrix}$$

is called the variance-covariance matrix of the n samples of data.

If the input data are not *collinear*, then by definition the matrix X^tX is invertible. Therefore it is positive definite. Consider the least squares problem considered in §14.1 with input data given by the $\mathbf{x}_j$, and a single output of data $\mathbf{y}$. The only difference with what we did early is that we leave out the column of constants. Then an easy computation shows that the least squares hyperplane in $\mathbb{R}^n$ that gives the best fit to the data has coefficients

$$\mathbf{a} = (X^tX)^{-1}X\mathbf{y}.$$

This is the analog of (14.2). So the variance-covariance matrix gives the solution.

14.4 Orthogonal Least Squares

In the method of least squares, explained in §14.1, we noticed that the output is treated differently from the inputs. One could ask for a more symmetric treatment. Here we give one in the case that we have a collection of n points (x_i, y_i) in the plane, and we ask for lines written in the form $x\cos\theta + y\sin\theta + c = 0$ that are the best fit to the data in the following sense. We want the sum of the squares of the distance of the points from the line, in the sense of perpendicular projection, to be minimum. This problem is known as the orthogonal least squares problem.

As you know from calculus, the perpendicular distance of the point (x_i, y_i) from the line $x\cos\theta + y\sin\theta y + c = 0$ is

$$x_i\cos\theta + y_i\sin\theta + c$$

So we wish to minimize the function of two variables

$$f(\theta, c) = \sum_{i=1}^{n} (x_i\cos\theta + y_i\sin\theta + c)^2$$

Lemma 14.4.1. *Let $x_1, \ldots, x_n$ be points on a line. Consider the real number t_0 that minimizes the function*

$$g(t) = \sum_{i=1}^{n} (x_i - t)^2.$$

Then

$$t_0 = \sum_{i=1}^{n} \frac{x_i}{n}.$$

Proof. Just take the derivative of $g(t)$ and set it to 0. □

Proposition 14.4.2. *The line minimizing distance passes through the centroid $(\bar{x}, \bar{y})$ of the data set.*

Proof. Consider a line minimizing distance. If it does not pass through the centroid, take the parallel line that does. A little computation using the lemma says that the new line has smaller $f(\theta, c)$ than the old line. □

In Remark 14.2.2, we proved in passing that the minimizing line for least squares goes through the centroid. We could have proceeded as we do here.

So we change the coordinate system so that the centroid of the data is the origin. Then the lines to consider are of the form $x\cos\theta + y\sin\theta = 0$, $-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$ and the function to minimize is

$$f(\theta) = \sum_{i=1}^{n} (x_i\cos\theta + y_i\sin\theta)^2$$

a function of a single variable θ. So we compute the derivative:

$$f'(\theta) = 2\sum_{i=1}^{n}(x_i\cos\theta + y_i\sin\theta)(-x_i\sin\theta + y_i\cos\theta)$$
$$= 2\sum_{i=1}^{n}\big(-x_i^2\sin\theta\cos\theta + x_iy_i(\cos^2\theta - \sin^2\theta) + y_i^2\sin\theta\cos\theta\big)$$

We wish to divide by $\cos\theta$ to express everything in terms of $\tan\theta$. We first need to consider what happens if $\cos\theta = 0$. Then for the derivative to be 0 we must have $\sum_{i=1}^{n} x_iy_i = 0$. This shows that $\sin\theta = 0$ is a second root of the equation.

So now we assume that $\cos\theta \neq 0$, so dividing by $\cos\theta$ we get for the vanishing of the derivative the quadratic equation in $\tan\theta$:

$$-S_{xy}(\tan\theta)^2 + (S_{xx} - S_{yy})\tan\theta + S_{xy} = 0$$

using the notation for sums of (14.4). We can divide by S_{xy} since we have already considered the case where it vanishes, so we are left with

$$(\tan\theta)^2 + \frac{S_{yy} - S_{xx}}{S_{xy}}\tan\theta - 1 = 0$$

The discriminant of this quadratic is obviously positive, so it has two distinct roots giving the slopes of two lines:

$$-\frac{S_{yy} - S_{xx}}{2S_{xy}} \pm \sqrt{\Big(\frac{S_{yy} - S_{xx}}{2S_{xy}}\Big)^2 + 1}.$$

Since their product is -1, the lines corresponding to the two roots are perpendicular. One is the minimum that we are looking for, and the other is the maximum over all the lines going through the centroid.

Let's compare these slopes with the slope m of the minimizing line for the ordinary least squares problem found in (14.6). So compare, shift the coordinates there so that the line goes through the centroid of the data. Then the slope is

$$m_{LS} = \frac{S_{xy}}{S_{xx}}.$$

Example 14.4.3. Assume we have three points (so $r = 3$ with coordinates $(0,1)$, $(1,1)$ and $(2,3)$. These points are obvious not aligned. We get $s_x = 3$, $s_y = 5$, so the centroid is $(1,5/3)$. Shift the points so the centroid is the origin. The three points are now $(-1,-2/3)$, $(0,-2/3)$ and $(1,4/3)$. In that coordinate system we have

$S_{xx} = 2$, $S_{xy} = 2$ and $S_{yy} = 20/9$. The slope of the ordinary least squares line is $m = \frac{S_{xy}}{S_{xx}} = 1$.

The two possible slopes of the orthogonal least squares lines are

$$-\frac{2}{36} \pm \sqrt{\Big(\frac{2}{36}\Big)^2 + 1}$$

so the minimum clearly occurs for the + sign. The slope of the orthogonal least squares line is smaller than that of the least squares line as a little computation shows.

Here is a graph:

14.5 Computational Techniques in Least Squares

The key reference for this material is [9], Chapter 6. We use the notation of §8.5. As there, we assume for simplicity that the $m \times n$ matrix A of coefficients has rank n, so that there is a unique minimizer for the problem. The error, usually called the *residual* for the problem is the vector $\mathbf{b} - \mathbf{p}$. As noted in (8.13) it is orthogonal to the columns of A.

By the singular value decomposition (SVD) of A, we can write

$$A = Q_1 D Q_2^t,$$

where Q_1 is an orthogonal matrix of size m whose columns are the eigenvectors of AA^t, Q_2 is the orthogonal matrix whose columns are the eigenvalues of A^tA, and the $m \times n$ matrix D only has non-zero entries along the diagonal, and they are the square roots of the non-zero eigenvalues of AA^t and A^tA.

Notice that $A^t = Q_2 D^t Q_1^t$, so

$$AA^t = (Q_1 D Q_2^t)(Q_2 D^t Q_1^t) = Q_1 D D^t Q_1^t \quad \text{and} \quad A^t = Q_2 D^t D Q_2^t.$$

So we have the eigenvector-eigenvalue decomposition for the matrices AA^t and A^tA, which are both symmetric so that the spectral theorem applies. Furthermore A^tA is positive definite, so its eigenvalues are positive so we can take their square roots, which are the n non-zero diagonal elements of D and D^t.

Given the matrix A we define its *pseudoinverse* A^+ using its SVD of A by the four Moore-Penrose conditions. A^+ is the unique $n \times m$ matrix X so that

1. $AXA = A$;
2. $XAX = X$;
3. $(AX)^t = AX$;
4. $(XA)^t = XA$.

These conditions imply that AA^+ is the orthogonal projection to the range of A, and A^+A the orthogonal projection to the range of A^t. In other words in our setup

$$A^+ = (A^tA)^{-1}A^t.$$

Then the first method of solving least squares is to use the *Cholesky factorization* of the positive definite $C = A^tA$. This means we write $C = GG^t$. Because C is symmetric and positive definite, its LDM decomposition, where L is lower triangular, D is diagonal and M is upper triangular, has $M = L^t$, and all the diagonal entries of

D are positive. So $D = D_1^2$, where D_1 is again diagonal. So let $G = LD_1$. Thus G is lower triangular with positive elements on the diagonal Then to solve the least squares problem written in normal equations as $A^tA\mathbf{x} = A^t\mathbf{b}$, write $A^tA = GG^t$ and $\mathbf{d} = A^t\mathbf{b}$.

Algorithm

First solve $G\mathbf{y} = \mathbf{d}$ and then $G^t\mathbf{x} = \mathbf{y}$.

Thus this is a variant of the solution of a system of linear equations by writing the coefficient matrix as a product of lower triangular and upper triangular, which can be done quickly.

The second method is to use QR factorization. Write $A = QR$, where A is an orthogonal $m \times m$ matrix and R is an upper triangular $m \times n$ matrix, so that its last $m - n$ rows are $\mathbf{0}$. Write R_1 for the square matrix of size n consists of its top n rows. Then write $Q^t\mathbf{b} = \mathbf{c}$. Write the first n entries of $\mathbf{c}$ as the n vector $\mathbf{c}_1$, and the last $m - m$ entries as the vector $\mathbf{c}_2$. Then since Q is orthogonal, and therefore preserves distance, the Pythagorian theorem gives us:

$$\|A\mathbf{x} - \mathbf{b}\|^2 = \|Q^tA\mathbf{x} - Q^t\mathbf{b}\|^2 = \|R_1\mathbf{x} - \mathbf{c}_1\|^2 + \|\mathbf{c}_2\|^2.$$

So the least squares solution is given by the easy triangular system $R_1\mathbf{x} = \mathbf{c}_1$ and the error is $\|\mathbf{c}_2\|^2$.

Chapter 15
Linear Inequalities and Polyhedra

Abstract In this chapter we restrict to real coefficients, so that we can compare constants and write down inequalities. First in a section that introduces the tools of the chapter, we study *affine geometry*, which is essentially linear geometry, but where we do not have an origin. Then we show how systems of linear inequalities can be studied and solved by elimination much in the same way as systems of equalities. The set of solutions of families of linear inequalities is called a polyhedron: they represent a special case of convex sets, about which we study next. Using projections we can say a great deal about polyhedra, and prove that bounded polyhedra are polytopes.

15.1 Affine Geometry

In the linear algebra we have studied so far we have only given a name to subspaces of $\mathbb{R}^n$ given by the vanishing of a collection of homogeneous linear equations, which we can write in matrix form as the $\mathbf{x}$ that satisfy $A\mathbf{x} = \mathbf{0}$. These subspaces are of course called linear subspaces. Now we generalize.

Definition 15.1.1. An *affine subspace* or a *flat*[1] of $\mathbb{R}^n$ is a subset of $\mathbf{x} \in \mathbb{R}^n$ given by the vanishing of a family of equations $A\mathbf{x} = \mathbf{b}$, where A is as usual a $m \times n$ matrix, so $\mathbf{b}$ is an m-vector.

The goal of this section is to extend linear algebra to include flats as the basic objects of study. This could be called affine algebra, but it is always referred to as affine geometry.

A flat $T = \{\mathbf{t} \in \mathbb{R}^n | A\mathbf{t} = \mathbf{b}\}$ can be empty, precisely when $\mathbf{b}$ is not in the column space of A. Assume T is non-empty. Let W be the linear subspace of $\mathbb{R}^n$ given by $W = \{\mathbf{w} \in \mathbb{R}^n | A\mathbf{w} = \mathbf{0}\}$. We call W the linear subspace associated to the flat T, following Definition 1.1.6. It is obviously uniquely determined by T. If the matrix

[1] A term used in many books on affine geometry: see for example Lay [17], p. 12.

A has rank n, then T is a single point. Conversely any point $\mathbf{p}$ is the flat associated to the family of equations $I\mathbf{x} = \mathbf{p}$, where I is the $n \times n$ identity matrix.

Lemma 15.1.2. *Fix an element $\mathbf{t}_0 \in T$. Any element $\mathbf{t}$ in T can be written uniquely as $\mathbf{t}_0 + \mathbf{w}$, for an element $\mathbf{w} \in W$.*

Proof. This is easy. First note that any element of the form $\mathbf{t}_0 + \mathbf{w}$ is in T. Then assume that $\mathbf{t} \in T$ has two different representations $\mathbf{t}_0 + \mathbf{w}_1$ and $\mathbf{t}_0 + \mathbf{w}_2$, and show that $\mathbf{w}_1 = \mathbf{w}_2$. This is closely related to Theorem 1.1.7. □

Definition 15.1.3. The dimension of a non-empty flat T is the dimension of its associated linear subspace W.

Thus if $T = \{\mathbf{t} \in \mathbb{R}^n | A\mathbf{t} = \mathbf{b}\}$, the dimension of T is $n - r$, where r is the column rank of A.

Remark 15.1.4. The union of a flat T of dimension m that does not go through the origin and its associated linear space W is contained in a linear subspace U of dimension $m+1$. Indeed a basis for U is given by a basis of W plus any $\mathbf{t} \in T$.

What we have done so far defines flats by the equations that vanish on it. Now we want to argue directly on the points of the flat: what makes it affine?

Its associated linear space W is determined by $n - r$ linearly independent points on it (a basis). To determine T we also need the additional information given by the m-vector $\mathbf{b}$, whose coordinates tells us how far the equations are from being homogeneous.

We want a more intrinsic description, so we make a new definition:

Definition 15.1.5. A set S in $\mathbb{R}^n$ is *affine* if for every pair of points $\mathbf{p}$ and $\mathbf{q}$ in S and every real number λ, the point $\lambda\mathbf{p} + (1-\lambda)\mathbf{q}$ is in S.

In other words, if two distinct points are in S, then any point on the line joining the two points is in S. As we now see all affine sets are flats.

Theorem 15.1.6. *A non-empty subset S of $\mathbb{R}^n$ is affine if and only if it is a flat of $\mathbb{R}^n$.*

Proof. It is easy to show that every flat is an affine subset. Assume the flat T is the set of solutions $\mathbf{x}$ of $A\mathbf{x} = \mathbf{b}$. Then T is an affine subset. Indeed take two points $\mathbf{p}$ and $\mathbf{q}$ in T, so that $A\mathbf{p} = \mathbf{b}$ and $A\mathbf{q} = \mathbf{b}$. Then by linearity

$$A(\lambda\mathbf{p} + (1-\lambda)\mathbf{q}) = \lambda A\mathbf{p} + (1-\lambda)A\mathbf{q} = \lambda\mathbf{b} + (1-\lambda)\mathbf{b} = \mathbf{b}$$

so we are done,

The other direction is harder. Assume that S is an affine subset. Pick any $\mathbf{s}_0 \in S$. Consider the subset $V = \{-\mathbf{s}_0 + \mathbf{s}|$ for all $\mathbf{s} \in S\}$. It contains the origin: take $\mathbf{s} = s_0$. To show V is a subspace, we must additionally show that it is closed under scalar multiplication and vector addition.

First we consider scalar multiplication. The following graph explains the argument:

If S consists in just one point $\mathbf{s}_0$, there is nothing to do, since V is just the origin. Otherwise take any other point $\mathbf{s} \in S$. Then $\mathbf{s} - \mathbf{s}_0$ is in V. We need to show that for any $\lambda \in \mathbb{R}$, $\lambda(\mathbf{s} - \mathbf{s}_0) \in V$. Since S is affine $\lambda\mathbf{s}_0 + (1-\lambda)\mathbf{s} \in S$, so

$$\lambda\mathbf{s}_0 + (1-\lambda)\mathbf{s} - \mathbf{s}_0 = (\lambda - 1)\mathbf{s}_0 + (1-\lambda)\mathbf{s} = (1-\lambda)(\mathbf{s} - \mathbf{s}_0) \text{ is in } V.$$

Finally we do vector addition. Take any $\mathbf{s}_1$ and $\mathbf{s}_2$ in S. So $-\mathbf{s}_0 + \mathbf{s}_1$ and $-\mathbf{s} + \mathbf{s}_2$ are in V. We must show that for any λ_1 and λ_2 in $\mathbb{R}$,

$$\lambda_1(-\mathbf{s}_0 + \mathbf{s}_1) + \lambda_2(-\mathbf{s}_0 + \mathbf{s}_2) = -(\lambda_1 + \lambda_2)\mathbf{s}_0 + \lambda_1\mathbf{s}_1 + \lambda_2\mathbf{s}_2 \text{ is in } V.$$

This element is in V if and only if when $\mathbf{s}_0$ is added to it, the new element is in S. This new element is written

$$(1 - \lambda_1 - \lambda_2)\mathbf{s}_0 + \lambda_1\mathbf{s}_1 + \lambda_2\mathbf{s}_2 = \text{ is in } V.$$

This is true when $\lambda_1 + \lambda_2 = 1$, and then by multiplying $\mathbf{s}_2$ by the appropriate scalar using the first part of the proof, we get it for all λ_1 and λ_2.[2] □

Theorem 15.1.7. *The intersection of any collection of affine sets in $\mathbb{R}^n$ is affine, possibly empty.*

Proof. The easy method is just to notice that the intersection of flats is a flat and use Theorem 15.1.6. We can also argue directly. Let S and T be affine. Consider the intersection $S \cap T$. Pick any two points $\mathbf{p}$ and $\mathbf{q}$ in it. The for any λ, $\lambda\mathbf{p} + (1-\lambda)\mathbf{q}$ is in S and in T, so it is in the intersection.

Definition 15.1.8. The *affine hull* of an arbitrary set $S \in \mathbb{R}^n$ is the intersection of all affine sets containing S.

Corollary 15.1.9. *The affine hull of a set S is affine.*

Definition 15.1.10. Let $\mathbf{x}_1, \ldots, \mathbf{x}_r$ be a collection of r points in $\mathbb{R}^n$, where r is any positive integer. Then $\mathbf{x}$ is an *affine combination* of the points $\mathbf{x}_i$ if there exists real numbers λ_i, $\sum_{i=1}^r \lambda_i = 1$, such that

$$\mathbf{x} = \sum_{i=1}^{r} \lambda_i \mathbf{x}_i$$

Theorem 15.1.11 (The Affine Combinations Theorem). *A set S is affine if and only if all finite affine combinations of points of S are in S.*

Proof. One direction is obvious: if all finite affine combinations of points in S are in S then S is affine, since affineness is just the case $r = 2$.

For the other direction assume that S is affine. We could argue on the flat, just as at the end of the proof of Theorem 15.1.6, but it is easier to do it directly: we prove

[2] This is a special case of Theorem 15.1.11.

by induction that if any affine combination of $r-1$ points is in S, then so is any affine combination of r points. We start the induction at $r=2$ where we just use the definition of affine. So take an affine linear combination of r points: $\mathbf{x}=\sum_{i=1}^{r}\lambda_i\mathbf{x}_i$, with $\sum_{i=1}^{r}\lambda_i=1$. If any of the λ_i is 0 then we have an affine combination of $r-1$ points, so $\mathbf{x}$ is in S. So we may assume that $\lambda_r\neq 0$. The following combination of the first $n-1$ points is affine:

$$\mathbf{y}=\frac{\lambda_1}{1-\lambda_r}\mathbf{x}_1+\frac{\lambda_2}{1-\lambda_r}\mathbf{x}_2+\cdots+\frac{\lambda_{r-1}}{1-\lambda_r}\mathbf{x}_{r-1}.$$

So by induction $\mathbf{y}\in S$. Then since S is affine, the affine combination $(1-\lambda_r)\mathbf{y}+\lambda_r\mathbf{x}_r$ is in S. □

Definition 15.1.12. For any set S, let $A(S)$ be the set of all finite affine combinations of points of S.

By taking the number of points r in the affine combination to be 1, so that $\lambda_1=1$, we have

$$S\subset A(S) \tag{15.1}$$

Theorem 15.1.13. *For any set S, $A(S)$ is an affine set.*

Proof. Pick two points $\mathbf{p}$ and $\mathbf{q}$ in $A(S)$. Then

$$\begin{aligned}\mathbf{p}&=\sum_{i=1}^{r}\lambda_i\mathbf{x}_i \text{ with } \mathbf{x}_i\in S \text{ and } \sum_{i=1}^{r}\lambda_i=1.\\ \mathbf{q}&=\sum_{i=1}^{r}\mu_i\mathbf{y}_i \text{ with } \mathbf{y}_i\in S \text{ and } \sum_{i=1}^{r}\mu_i=1.\end{aligned} \tag{15.2}$$

We must show that for any number ν, $\nu\mathbf{p}+(1-\nu)\mathbf{q}$ is in $A(S)$. In this expression replace $\mathbf{p}$ and $\mathbf{q}$ by the sums in (15.2):

$$\begin{aligned}\nu\mathbf{p}+(1-\nu)\mathbf{q}&=\nu(\sum_{i=1}^{r}\lambda_i\mathbf{x}_i)+(1-\nu)(\sum_{i=1}^{r}\mu_i\mathbf{y}_i\\ &=\sum_{i=1}^{r}\nu\lambda_i\mathbf{x}_i+\sum_{i=1}^{r}(1-\nu)\mu_i\mathbf{y}_i.\end{aligned} \tag{15.3}$$

Since $\sum_{i=1}^{r}\nu\lambda_i+\sum_{i=1}^{r}(1-\nu)\mu_i=1$ this is an affine combination of the points $\{\mathbf{x}_1,\ldots,\mathbf{x}_r,\mathbf{y}_1,\ldots,\mathbf{y}_s\}$. □

Theorem 15.1.14. *For any set S, the affine hull is equal to the set of affine combinations.*

Proof. Temporarily denote by H the affine hull of A. Since H is the intersection of all affine sets containing S and $A(S)$ is affine and contains S, we have $H\subset A(S)$. We need to prove the other inclusion: it is an immediate consequence of the Affine Combinations Theorem. □

Definition 15.1.15. The points $\mathbf{x}_0, \ldots, \mathbf{x}_k$ are *affinely dependent* if there are real numbers a_i, with $\sum_{i=0}^{k} a_i = 0$ and not all $a_i = 0$ such that

$$\sum_{i=0}^{k} a_i \mathbf{x}_i = \mathbf{0}. \tag{15.4}$$

Otherwise we say they are *affinely independent*.

Linearly independent sets of points are affinity independent, but as the next example shows, there are affinely independent sets of points that are not linearly independent.

Proposition 15.1.16. *Any subset of a collection of affinely independent points is affinely independent.*

Proof. This is trivial. Suppose that the points $\mathbf{x}_0, \ldots, \mathbf{x}_k$ are affinely independent, but that some subset, say $\mathbf{x}_0, \ldots, \mathbf{x}_l$, is affinely dependent. Then there is a collection a_i, $0 \leq i \leq l$, non all zero, with $\sum_{i=0}^{l} a_i = 0$. Then define $a_{l+1} = \cdots = a_k = 0$. This gives an equation of linear dependence for the original set, a contradiction.. □

Example 15.1.17. The points $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$ in $\mathbb{R}^3$ are affinely independent. Indeed, if you add the origin $(0,0,0)$ to this set of points, it is still affinely independent.

Exercise 15.1.18. Show that if there is repetition in the list of points $\mathbf{x}_0, \ldots, \mathbf{x}_k$, so for example if $\mathbf{x}_0 = \mathbf{x}_1$, the points are affinely dependent.

Proposition 15.1.19. *The points $\mathbf{x}_0, \ldots, \mathbf{x}_k$ are affinely dependent if and only if the vectors $\mathbf{x}_i - \mathbf{x}_0$, $1 \leq i \leq k$, are linearly dependent.*

Proof. Assume that $\mathbf{x}_0, \ldots, \mathbf{x}_k$ are affinely dependent, so there are real numbers a_i satisfying (15.4). Then

$$a_0 = -\sum_{i=1}^{k} a_i. \tag{15.5}$$

If $a_0 \neq 0$, substitute a_0 into the equation of affine dependence, getting

$$\sum_{i=1}^{k} a_i(\mathbf{x}_i - \mathbf{x}_0) = 0. \tag{15.6}$$

Not all the coefficients in this equation are zero by (15.4), so this is the required equation of linear dependence between the $\mathbf{x}_i - \mathbf{x}_0$.

To get the other implication, start from the equation of linear dependence (15.6) and define a_0 by (15.5). This gives (15.4), the required equation of affine dependence. □

Exercise 15.1.20. Prove that if $\mathbf{x}$ is an affine combination of $\mathbf{x}_0, \ldots, \mathbf{x}_k$, and if the $\mathbf{x}_i$, $0 \leq i \leq k$, are affinely dependent, then $\mathbf{x}$ is an affine combination of a smaller number of the $\mathbf{x}_i$.

The importance of affine independence comes from the following theorem, which is analogous to Theorem 3.4.12 in linear algebra.

Theorem 15.1.21. *Let* $\mathbf{x}_0, \ldots, \mathbf{x}_m$ *be a collection of affinely independent points in a flat of dimension m. Then any point* $\mathbf{x}$ *is the flat can be written uniquely as an affine combination of the* $\mathbf{x}_i$, $0 \leq i \leq m$.

Proof. First we prove the uniqueness. Suppose

$$\begin{aligned} \mathbf{x} &= \sum_{i=0}^{m} \lambda_i \mathbf{x}_i && \text{where } \sum_{i=0}^{m} \lambda_i = 0; \\ &= \sum_{i=0}^{m} \mu_i \mathbf{x}_i && \text{where } \sum_{i=0}^{m} \mu_i = 0; \end{aligned}$$

and where not all the λ_i are zero and not all the μ_i are zero. Then subtract the two representations. You get an equation of affine dependence between the $\mathbf{x}_i$ unless all the coefficients are 0, meaning that $\lambda_i = \mu_i$ for all i.

The existence follows from Lemma 15.1.2: this is where you use the hypothesis on the dimension. □

Thus affinely independent sets of $m+1$ elements in an affine subset of dimension m form the analog of a basis in a vector space of dimension m.

Definition 15.1.22. Assume S is a flat of dimension $n-1$ in $\mathbb{R}^n$, and W its associated linear space. Then S is called a *hyperplane*, the zeroes of a unique equation

$$a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = b$$

which we can write in terms of matrix multiplication as

$$\mathbf{a}^t \mathbf{x} = b$$

or in terms of the standard inner product on $\mathbb{R}^n$ as

$$\langle \mathbf{a}, \mathbf{x} \rangle = b. \tag{15.7}$$

We will always assume that the vector $\mathbf{a}$ is not the zero vector. So $b = \langle \mathbf{a}, \mathbf{s} \rangle$, for any $\mathbf{s} \in S$. Since V is a hyperplane through the origin we get $\langle \mathbf{a}, \mathbf{x} \rangle = 0$, so that $\mathbf{a}$ is orthogonal to any $\mathbf{v} \in V$ Thus $\mathbf{a}$ is called a *normal* (meaning perpendicular) vector to V, and by extension also to S where for any two vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ in S, we have $\langle \mathbf{a}, \mathbf{s}_1 - \mathbf{s}_2 \rangle = 0$. The normal vector $\mathbf{a}$ is only defined up to multiplication by a non-zero scalar. We write $S = H_{\mathbf{a},b}$ and $V = H_{\mathbf{a},0}$. Note that $H_{\mathbf{a},b} = H_{c\mathbf{a},cb}$ for any non-zero scalar c.

Lemma 15.1.23. *Take n linearly independent points* $\mathbf{b}_i$, $1 \leq i \leq n$, *in* $\mathbb{R}^n$. *Then there is a unique hyperplane H passing through these points.*

Proof. Write B for the $n \times n$ matrix whose i-th row are the coordinates of $\mathbf{b}_i$. Write the equation of H as

$$c_1x_1 + c_2x_2 + \dots c_nx_n = d,$$

so the vector $\mathbf{c}$ and the unknown number d satisfy the system of n equation $B\mathbf{c} = \mathbf{d}$, where $\mathbf{d} = (d, d, \dots, d)$. Linear independence of the points $\mathbf{b}^i$ is equivalent to saying that the matrix B is invertible, so there is a unique solution $\mathbf{c} = B^{-1}\mathbf{d}$ to the system of equations, up to scaling by a non-zero constant. Thus the hyperplane H is unique. □

Remark 15.1.24. The distance between a hyperplane S and its associated linear space V is

$$\frac{\sqrt{\langle \mathbf{a}, \mathbf{s} - \mathbf{v} \rangle}}{\|\mathbf{a}\|}$$

for any $\mathbf{s} \in S$ and $\mathbf{v} \in V$, as you should check.

Example 15.1.25. Let's write the equations of the affine line in $\mathbb{R}^3$ through the points $(3,0,1)$ and $(1,2,0)$. That means we need to find all the solutions in (a,b,c,d) of the equations

$$ax + by + cz = d$$

that verify $3a + c = d$ and $a + 2b = d$. We have 2 equations in 4 unknowns, and we can easily solve in terms of a and b:

$$c = -2a + 2b$$
$$d = a + 2b$$

Therefore there are two affine equations vanishing on the points, and therefore on the line. On the other hand there is only one *linear* equation through the points: indeed, when $d = 0$, we have , up to a scalar, $-2x + y + 6z = 0$. The two affine equations, are, for example, $x - 2z = 1$ and $y + 2z = 2$.

15.2 Systems of Linear Inequalities and Polyhedra

For simplicity we only consider weak inequalities ($\leq$ and $\geq$) even though we could also use strong inequalities ($<$, $>$): they just make the bookkeeping more complicated. We can arrange that all the inequalities be oriented in the same way: this is no restriction, since it can always be achieved by multiplying the inequality by -1.

We keep the same notation as in the equality case described in §1.3, so the system S we study has m inequalities in n variables:

$$a_{i1}x_1 + \cdots + a_{in}x_n \leq b_i \quad (1 \leq i \leq m) \tag{15.8}$$

As usual we write

$$f_i(x_1,\ldots,x_n) = a_{i1}x_1 + \cdots + a_{in}x_n - b_i$$

for the function associated to the i-th row. We also write $A\mathbf{x} \leq \mathbf{b}$, where A is the $m \times n$ matrix of coefficients of the x_j and $\mathbf{b}$ in n vector of b_j. The first question is: when is the set of solutions $Z(S)$, namely the set of $(\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_n)$ satisfying the m inequalities $\mathbf{a}^j\mathbf{x} \leq b_j$ in the system S nonempty? Thus what we are doing is completely parallel to the case of equalities.

We give a name to the set of solutions:

Definition 15.2.1. A *polyhedron* P in $\mathbb{R}^n$ is the set of solutions $\mathbf{x}$ of (15.8).

We write $P(A,\mathbf{b})$ when we need to indicate the dependence on A and $\mathbf{b}$.

The polyhedron $P(A,\mathbf{b})$ could very well be empty, as could be set of solutions of a system of inhomogeneous equations. We say that the system of inequalities (15.8) is consistent if the polyhedron is not empty. Otherwise it is inconsistent.

We define a *consequence* of (15.8) as any linear inequality

$$c_1x_1 + c_2x_2 + \ldots c_nx_n \leq c$$

that is true for any $\mathbf{x} \in P(A,\mathbf{b})$. As in the case of equalities, we only use this concept if the system (15.8) is consistent, since otherwise any linear inequality is a consequence since we only have to verify it on the empty set.

We can produce consequences of S as follows: for any set of non-negative constants $c_1,c_2,\ldots,c_m$ form the inequality

$$c_1f_1 + \cdots + c_mf_m \leq 0$$

which can be written out as

$$(c_1a_{11} + \cdots + c_ma_{m1})x_1 + \cdots + (c_1a_{1n} + \cdots + c_ma_{mn})x_n \leq c_1a_1 + \cdots + c_ma_m$$

Because we have restricted the coefficients c_i to non-negative numbers, we always get in this way consequences of S: the direction of the inequality is preserved, so the inequality is preserved.

We now show that elimination proceeds exactly as in the case of equalities. Before stating the result, let's do some simple examples. Even the simplest possible example is instructive.

Example 15.2.2. Assume there is only one variable, so the inequalities can be written: $a_ix \leq b_i$, $1 \leq i \leq m$. If $a_i = 0$ in some equation, we get a contradiction if $b_i < 0$. This is the situation we always reduce to when trying to get a contraction. If $b_i \geq 0$, we get an identity which can be ignored. Assume that we are in that case.

Now consider the equations where $a_i \neq 0$. Let P denote the indices p where $a_p > 0$ and Q those where $a_q < 0$. Then for any $p \in P$ and any $q \in Q$ the linear combination $-a_qa_px + a_pa_qx \leq -a_qb_p + a_pb_q$ follows from the original inequalities. Thus $a_qb_p \leq a_pb_q$, so if this is not verified we get a contradiction.

Finally assume that all the inequalities obtained are consistent. Solving for x in each equation we have

$$x \leq \frac{b_p}{a_p} \qquad \text{for all } p \in P, \tag{15.9}$$
$$\frac{b_q}{a_q} \leq x \qquad \text{for all } q \in Q.$$

Then we have the following simple lemma, which tells us that the solutions form an interval, which could be unbound negatively, positively or both.

Lemma 15.2.3. *The equations 15.9 are satisfied if and only if*

$$\max_{q \in Q} \frac{b_q}{a_q} \leq \min_{p \in P} \frac{b_p}{a_p}. \tag{15.10}$$

The easy proof is left to you.

Thus in this case the polyhedron associated to this family of inequalities is non-empty if and only if (15.10) is verified, and for equations where $a_i = 0$, $b_i \geq 0$.

Now an example in $\mathbb{R}^2$.

Example 15.2.4. We work in the plane with unknowns x and y to simplify the notation. Assume we have the two inequalities:

$$\begin{aligned} -4x - y &\leq -12 \\ -x + 2y &\leq 8 \end{aligned} \tag{15.11}$$

First think about this geometrically. The two lines $y = -4x + 12$ and $y = x/2 + 4$ meet at the point $(16/9, 44/9)$. The region that satisfies both inequalities is the region above the line with negative slope and below the line with positive slope. Therefore it is the

GRAPH HERE

Because the sign of the coefficients of x is the two equations is the same, we cannot eliminate x and produce a consequence using an equation involving y alone. But we can eliminate y by adding twice the first equation to the second. So we get

$$-9x \leq -16 \quad \text{or} \quad x \geq \frac{16}{9}. \tag{15.12}$$

For later purposes it is more convenient to put y by itself on one side of each equation in (15.11), getting

$$\begin{aligned} 12 - 4x &\leq y \\ y &\leq x/2 + 4 \end{aligned}$$

so combining the two, by dropping the middle inequality involving y, we get

$$12 - 4x \leq x/2 + 4.$$

This is just another way of producing (15.12). For any $\bar{x}$ satisfying (15.12), any $\bar{y}$ in the interval $[12 - 4\bar{x}, \quad \bar{x}/2 + 4]$ gives a solution $(\bar{x}, \bar{y})$ to the original system. The

interval is non-empty because of the restriction (15.12) on $\bar{x}$, Draw a graph to show what is happening. Thus the polyhedron of (15.12) is the same as the polyhedron of

$$\begin{aligned} -x &\leq \frac{16}{9} \\ -4x-y &\leq -12 \\ -x/2+y &\leq 4 \end{aligned}$$

The advantage of this representation is that the new first equation only involves x.

Example 15.2.5. We make the previous example a little more complicated. Add two inequalities, so that the full set is:

$$\begin{aligned} -4x-y &\leq -12 \\ 2x+y &\leq 12 \\ x-y &\leq 0 \\ -x+2y &\leq 8 \end{aligned} \tag{15.13}$$

Are there any solutions? For variety we first eliminate the variable x, and then determining the constraints on y. Then we finish, just as in the equality case, by showing that for each solution $\bar{y}$ of the new system $S^{(1)}$ in y alone , there is a solution $(\bar{x},\bar{y})$ for the original system S.

Since we are in the plane, we can first understand the configuration of the four lines obtained by replacing each inequality in (15.13) by an equality. This is useful, because the region where one inequality is satisfied is one of the two half planes separated by the line where there is equality. The lines are:

$$\begin{aligned} L_1 &= \{(x,y)|4x+y=12\}; \\ L_2 &= \{(x,y)|2x+y=12\}; \\ L_3 &= \{(x,y)|x-y=0\}; \\ L_4 &= \{(x,y)|x-2y=-8\}. \end{aligned}$$

We can easily find the point of intersection of each pair of lines by Gaussian elimination. We get 6 points, distinct in this case, that we label $p_{ij} = L_i \cap L_j$, for $i < j$. So $p_{12} = (0,12)$, $p_{13} = (2.4,2.4)$, $p_{14} = (16/9,44/9)$, $p_{23} = (4,4)$, $p_{24} = (16/5,28/5)$, $p_{34} = (8,8)$.

Here is a graph of the region that satisfied the inequalities.

How would be determine this if we were not able to draw the graph? By elimination.

How do we eliminate x? Note that in the second and third equations, the coefficient of x is positive while in the first and fourth it is negative. We consider two inequalities where the coefficients of x have opposite signs, and we form the linear combinations with positive coefficients that makes the coefficient of x vanish. This new inequality is a consequence of the original two, and that does not contain x. There are four ways of doing this

So

- $f_1 + 2f_2$ gives $y \leq 12$.
- $f_3 + f_4$ gives $y \leq 8$.
- $f_1 + 4f_3$ gives $5y \geq 12$.
- $f_2 + 2f_4$ gives $5y \leq 28$.

So we have 4 inequalities involving y: one limiting it from below, and three from above. Using Lemma 15.2.3, take the most stringent inequality on either side, so

$$\frac{12}{5} \leq y \leq \frac{28}{5}. \tag{15.14}$$

This interval is not empty so we can find such a y.

Now we prove that for any $\bar{y}$ satisfying (15.14), there is a point $(\bar{x}, \bar{y})$ satisfying the original inequalities (15.13) in x and y. You should work this out geometrically on the graph.

The polyhedron of (15.13) is therefore the polyhedron of

$$\begin{aligned} y &\leq \frac{28}{5} \\ -y &\leq -\frac{12}{5} \\ -4x - y &\leq -12 \\ 2x + y &\leq 12 \\ x - y &\leq 0 \\ -x + 2y &\leq 8 \end{aligned} \tag{15.15}$$

which is easier to work with because of the added first two inequalities.

We continue by working out a three dimensional example, where you can still visualize what is going on.

Example 15.2.6. We work in ordinary space with unknowns x, y and z to simplify the notation. Assume we have the six inequalities:

$$\begin{aligned} x + y + z &\geq -1 \\ -x - y - z &\geq -1 \\ x - y + z &\geq -1 \\ -x + y - z &\geq -1 \\ x - y - z &\geq -1 \\ -x + y + z &\geq -1 \end{aligned} \tag{15.16}$$

Notice how the inequalities are grouped in consecutive pairs. if you multiply the second inequality in each pair by -1, you should be able to see what is going on geometrically. We eliminate x. For each equation where x has a positive coefficient

(always 1 in our example) we take a positive linear combination with an equation where x has a negative coefficient (always -1 in the example) so that x disappears from the new equation. Since we have 3 equations with positive x coefficient, and 3 with negative coefficient, we will end up with 9 inequalities from which x has been eliminated. They are

$$\begin{aligned} 0 &\geq -2 \quad \text{Equations 1 and 2} \\ y &\geq -1 \quad \text{Equations 1 and 4} \\ y+z &\geq -1 \quad \text{Equations 1 and 6} \\ -y &\geq -1 \quad \text{Equations 3 and 2} \\ 0 &\geq -2 \quad \text{Equations 3 and 4} \\ z &\geq -1 \quad \text{Equations 3 and 6} \\ -y-z &\geq -1 \quad \text{Equations 5 and 2} \\ -z &\geq -1 \quad \text{Equations 5 and 4} \\ 0 &\geq -2 \quad \text{Equations 5 and 6} \end{aligned} \tag{15.17}$$

Three of the inequalities are identically true, so we are left with six inequalities in y and z. y is already eliminated from two of them, so we are left with 4 inequalities on which we have to perform elimination of y:

$$\begin{aligned} y &\geq -1 \\ -y &\geq -1 \\ y+z &\geq -1 \\ -y-z &\geq -1 \end{aligned} \tag{15.18}$$

After elimination of y we get, listing first the two equations from the first step on which y was already eliminated:

$$\begin{aligned} z &\geq -1 \\ -z &\geq -1 \\ 0 &\geq -2 \end{aligned} \tag{15.19}$$

where repeats have been eliminated. This implies that the values of z that produce solutions after two steps are $-1 \leq z \leq 1$. The inequalities from (15.18) say that for any $\bar{z}$ in the interval, y must satisfy

$$-1-\bar{z} \leq y \leq 1+\bar{z}$$

This always has a solution when $-1 \leq z \leq 1$, since then $-1-\bar{z} \leq 1+\bar{z}$. Finally, for any solution $\bar{y}, \bar{z}$, the theorem tells us there is always a solution in x. You should work this out, and graph the solutions for any $\bar{y}, \bar{z}$.

A simple way of solving this problem is noting that the three vectors $\mathbf{u} = \mathbf{x}+\mathbf{y}+\mathbf{z}$, $\mathbf{v} = \mathbf{x}-\mathbf{y}+\mathbf{z}$ and $\mathbf{w} = \mathbf{u} = -\mathbf{x}+\mathbf{y}+\mathbf{z}$ are linearly independent. Using these as a

basis, the inequalities become easy to understand, and the solution set is just a cube bounded by -1 and 1 in all three coordinates.

Now we state and prove the general case, reverting to the notation in (15.8). Let $P^{(0)}$ be the polyhedron associated to this system of inequalities. We pick one variable, which after renaming the variables we may assume is x_1, and group the inequalities into three groups.

- The indices p where a_{p1} is positive; the set of such p we call P.
- The indices q where a_{q1} is negative; the set of such q we call Q.
- The indices r where a_{r1} is zero; the set of such r we call R.

Our goal is to eliminate x_1. For the inequalities with index in R there is nothing to do. So if $\#R$ denotes the cardinality of R, then we obviously get $\#R$ inequalities from which x_i has been eliminated. Now consider an inequality with index p in P. Dividing by $a_{p1} > 0$ and isolating x_1 on the left hand side, we get

$$x_1 \leq -\frac{a_{p2}}{a_{p1}}x_2 - \cdots - \frac{a_{pn}}{a_{p1}}x_n + \frac{b_p}{a_{p1}}.$$

On the other hand, for each q in Q we get

$$-\frac{a_{q2}}{a_{q1}}x_2 - \cdots - \frac{a_{qn}}{a_{q1}}x_n + \frac{b_q}{a_{q1}} \leq x_1.$$

Let $\#P$ denote the number of inequalities in P and $\#Q$ denote the number of inequalities in Q. Then from any pair of inequalities, one in Q and the other in P we get an inequality from which x_i has been eliminated:

$$-\frac{a_{q2}}{a_{q1}}x_2 - \cdots - \frac{a_{qn}}{a_{q1}}x_n + \frac{b_q}{a_{q1}} \leq -\frac{a_{p2}}{a_{p1}}x_2 - \cdots - \frac{a_{pn}}{a_{p1}}x_n + \frac{b_p}{a_{p1}}$$

or

$$\left(\frac{a_{p2}}{a_{p1}} - \frac{a_{q2}}{a_{q1}}\right)x_2 + \cdots + \left(\frac{a_{pn}}{a_{p1}} - \frac{a_{qn}}{a_{q1}}\right)x_n \leq \frac{b_p}{a_{p1}} - \frac{b_q}{a_{q1}}. \tag{15.20}$$

Therefore we get a new polyhedron $P^{(1)}$ in $\mathbb{R}^{n-1}$ given by the inequalities in R and the inequalities 15.20 for each $p \in P$ and $q \in Q$.

We have established

Theorem 15.2.7. *The polyhedron $P^{(1)}$ is the projection of $P^{(0)}$ into $\mathbb{R}^{n-1}$. The polyhedron $P^{(0)}$ is non-empty if and only if $P^{(1)}$ is non-empty.*

By repeating this elimination process, we can determine when the polyhedron is non-empty. If it is non-empty we would like to study its properties, much as we did for affine sets. The most important question is: is the polyhedron bounded? This just means it can be put inside a ball of finite radius in $\mathbb{R}^n$.

We can determine this somewhat crudely by our projection technique. Keep projecting the polyhedron until you get its projection onto the line with coordinate x_i, for some i. Then there are four possibilities:

1. There are no constrains on the projection: in other words the projection is the entire line.
2. There is one constraint above: $x_i \leq M$: so the projection is a half-line;
3. There is one constraint below: $m \leq x_i$: again the projection is a half-line;
4. There are constraints on both sides: $m \leq x_i \leq M$: the projection is a finite interval.

Clearly the only way the original polyhedron can be bounded is if we are in the last case. This is a necessary but not sufficient condition, as you should convince yourself by taking the polyhedron in $\mathbb{R}^2$ given by $x_2 \leq M$ and $m \leq x_2$, where $m < M$. Then the projection to the x_2 line gives an interval but the projection to the x_1 line puts us in case 1. However we get the following easy theorem:

Theorem 15.2.8. *A polyhedron is bounded if and only if its projection to all the coordinate axes is an interval.*

The easy proof is left to you. Note that the polyhedron in Example 15.2.6 is bounded.

Application 15.2.9 (Linear Optimization) *Using this elimination technique we can determine when a linear function*

$$f(x_1, x_2, \ldots, x_n) = c_1x_1 + c_2x_2 + \cdots + c_nx_n, \quad c_i \in \mathbb{R}$$

has a maximum or a minimum on a polyhedron P given by $A\mathbf{x} \leq \mathbf{b}$.

Define a new variable $x_{n+1} = c_1x_1 + \cdots + c_nx_n$, and use this equation to eliminate one of the x_i, $1 \leq i \leq n$ with $c_i \neq 0$ from the inequalities defining P. Then we have a new polyhedron in the n variables: x_{n+1} and the original variables with x_i removed. This is in fact the same polyhedron but described in a new coordinate system. Then eliminate the original variables by the technique above until we get down to the line with variable x_{n+1}.

Assuming again that the original polyhedron was non-empty, we have the four possibilities enumerated above. Here is what they say about the maximum or minimum of the function $f(\mathbf{x})$ on the polyhedron.

1. If the projection is the entire line, $f(\mathbf{x})$ takes all values on the polyhedron;
2. If $x_{n+1} \leq M$, then M is the maximum of $f(\mathbf{x})$ on P;
3. If $m \leq x_{n+1}$, then m is the minimum of $f(\mathbf{x})$ on P.;
4. If $m \leq x_{n+1} \leq M$, then $f(\mathbf{x})$ has both a maximum and a minimum on P

To say more the tool we need is convexity.

15.3 Convex Sets

We start with some geometric considerations to motivate the definition of convexity. Take two distinct points $\mathbf{p}$ and $\mathbf{q}$ in $\mathbb{R}^n$. There is a unique straight line L passing through both of them. By extension of the notation in $\mathbb{R}$ we denote $[\mathbf{p}, \mathbf{q}]$ and $(\mathbf{p}, \mathbf{q})$

the closed and open segments of points on L bounded by $\mathbf{p}$ and $\mathbf{q}$. The points $\mathbf{r}$ of $(\mathbf{p},\mathbf{q})$ can be parametrized by

$$\mathbf{r} = \lambda\mathbf{p} + (1-\lambda)\mathbf{q}, \text{ for } \lambda \in \mathbb{R}, 0 < \lambda < 1. \tag{15.21}$$

If we think of λ as time and $\mathbf{r}(\lambda)$ being the position of a particle at time λ, then at time 0 it is at $\mathbf{q}$ and then it moves at constant speed to $\mathbf{p}$ which it reaches at time 1.

Definition 15.3.1. A point $\mathbf{r}$ is *between* $\mathbf{p}$ and $\mathbf{q}$ if it satisfies (15.21), so that it is in the open segment $(\mathbf{p},\mathbf{q})$.

Definition 15.3.2. A set S in $\mathbb{R}^n$ is *convex* if for every pair of points $\mathbf{p}$ and $\mathbf{q}$ in S, any point of the open segment joining $\mathbf{p}$ and $\mathbf{q}$ is in S. In other words, every point between $\mathbf{p}$ and $\mathbf{q}$ is in S.

Before giving some examples of convex sets, we make some definitions and set up some notation that will be used throughout this chapter.

Definition 15.3.3. Start with the hyperplane $H_{\mathbf{a},b}$ of (15.7). Then the two *closed half-spaces* associated to this hyperplane are:

$$H^+_{\mathbf{a},b} = \{\mathbf{x} \mid \langle \mathbf{a},\mathbf{x}\rangle \geq b\}, \text{ and } H^-_{\mathbf{a},b} = \{\mathbf{x} \mid \langle \mathbf{a},\mathbf{x}\rangle \leq b\}.$$

Note that $H^+_{\mathbf{a},c}$ is the half-space the normal vector $\mathbf{a}$ points into. The hyperplane $H_{\mathbf{a},b}$ is called the *face* of both half-spaces.

Here graph in plane.

Theorem 15.3.4. *The intersection of any collection of convex sets in $\mathbb{R}^n$ is convex.*

Proof. Let C_α, $\alpha \in I$, be such a collection, where the index set I may be infinite. If the intersection is empty, we are done; if there is just one point in the intersection, likewise. So take any two points $\mathbf{p}$ and $\mathbf{q}$ in the intersection. For every $\alpha \in I$, the segment $[p,q]$ is in C_α, so it is in the intersection, which is therefore convex. □

Example 15.3.5. Using this theorem, we can show the following sets are convex:
The empty set;[3]
A point;
A line or a segment on a line;
Any affine hyperplane $H_{\mathbf{a},b}$; More generally, any linear space;
A half-space $H^+_{\mathbf{a},b}$ or $H^-_{\mathbf{a},b}$. More generally, any polyhedron (see Definition 15.2.1).

We are primarily interested in polyhedra. However it is worth knowing about the following example.

[3] In a few texts, the empty set is not taken to be convex. The majority of references say that the empty set is convex: [3], [4], [5], [14], [17], [21], [23], [27]. This is simply a matter of convention.

Example 15.3.6. Consider the closed ball $\overline{N}_r(\mathbf{c})$ in $\mathbb{R}^n$ of points at distance less than or equal to r from the center $\mathbf{c}$. It is convex. Change coordinates so that its center is at the origin. Then the closed ball $\overline{N}_r(\mathbf{0})$ is just the set of points $\mathbf{x} \in \mathbb{R}^n$ such that $\|\mathbf{x}\| \leq r$. Given two points $\mathbf{p}$ and $\mathbf{q}$ such that $\|\mathbf{p}\| \leq r$ and $\|\mathbf{q}\| \leq r$, we must show that $\|\lambda\mathbf{p} + (1-\lambda)\mathbf{q}\| \leq r$ for all λ, $0 < \lambda < 1$. By the triangle inequality

$$\|\lambda\mathbf{p} + (1-\lambda)\mathbf{q}\| \leq \lambda\|\mathbf{p}\| + (1-\lambda)\|\mathbf{q}\| \leq \lambda r + (1-\lambda)r = r,$$

so we are done.

Definition 15.3.7. A point $\mathbf{r}$ of a convex set S is an *extreme point* of S if it is not between two points of S.

In other words, one cannot find distinct points $\mathbf{p}$ and $\mathbf{q}$ in S so that (15.21) is satisfied.

Example 15.3.8. Let T be the polyhedron in $\mathbb{R}^2$ that is the set of solutions of the three inequalities $A\mathbf{x} \leq \mathbf{b}$, where A is a 3×2 matrix and x_1, x_2 are the coordinates on $\mathbb{R}^2$. Assume that any two rows of A are linearly independent. Then there is a unique point of intersection $\mathbf{p}_k$ of the line $L_i = \langle \mathbf{a}^i, \mathbf{x} \rangle = b_i$ and $L_j = \langle \mathbf{a}^j, \mathbf{x} \rangle = b_j$, where $\{i, j, k\}$ are all distinct. Also assume that the three points $\mathbf{p}_1$, $\mathbf{p}_2$ and $\mathbf{p}_3$ are distinct. Finally sssume the normal $\mathbf{a}^i$ of L_i, the i-th row of A, points to the opposite half-plane that the point $\mathbf{p}_i$ lies in. Then T is the triangle whose vertices are the $\mathbf{p}_i$. Convince yourself T is convex. The extreme points of T are the vertices of the triangle.

Example 15.3.9. The extreme points of the closed ball $\overline{N}_r(\mathbf{c})$ in $\mathbb{R}^n$ are the $(n-1)$-sphere $S_r(\mathbf{p})$ of points at distance exactly r from the center $\mathbf{c}$.

Exercise 15.3.10. If you remove an extreme point from a convex set, what remains is convex. Conversely, if you remove a point from a convex set, and the remainder is convex, the removed point was an extreme point. Combining Example 15.3.9 and this exercise, we see that balls ball $N_r(\mathbf{c})$ in $\mathbb{R}^n$ of points at distance less than r from the center $\mathbf{c}$ are convex.

Definition 15.3.11. The *convex hull* of a set $S \in \mathbb{R}^n$ is the intersection of all convex sets containing S. It is denoted $C(S)$.

Corollary 15.3.12. *The convex hull of any set S is convex.*

Definition 15.3.13. A n-simplex is the convex hull of $n+1$ affinely independent points in $\mathbb{R}^n$. If the points are $\mathbf{a}_0, \ldots, \mathbf{a}_n$, then each $\mathbf{a}_i$ is an extreme point of the simplex, also called a vertex of the simplex, and the segments $[\mathbf{a}_i, \mathbf{a}_j]$ are the *edges*. More generally the convex hull of any collection of $\mathbf{a}_i$ is called a face of the simplex.

We write the simplex as:

$$H(\mathbf{a}_0, \ldots, \mathbf{a}_n) = \{\mathbf{x} = \sum_{i=0}^{n} \lambda_i \mathbf{a}_i \mid \lambda_i \geq 0\,,\ \sum_{i=0}^{m} \lambda_i = 1\} \tag{15.22}$$

Definition 15.3.14. Let $\mathbf{x}_1, \ldots, \mathbf{x}_r$, be a collection of r points in $\mathbb{R}^n$, where r is any positive integer. Then $\mathbf{x}$ is a *convex combination* of the points $\mathbf{x}_i$ if there exist non-negative real numbers λ_i, $\sum_{i=1}^r \lambda_i = 1$ such that

$$\mathbf{x} = \sum_{i=1}^{r} \lambda_i \mathbf{x}^i \tag{15.23}$$

This is very similar to the definition of affine combination: the only difference is that the λ_i are required to be non-negative.

Exercise 15.3.15. The low dimensional simplices.

- if $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$ are three distinct, non-aligned points in the plane, then the set of convex combinations of $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$ is the triangle and the inside of the triangle formed by the three points.
- if $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are four distinct points in $\mathbb{R}^3$, such that any three span a plane, and the four points do not lie in a plane, then the set of convex combinations of $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ is a tetrahedron[4] and its interior.

Theorem 15.3.16 (The Convex Combinations Theorem). *A set S is convex if and only if all finite convex combinations of points of S are in S.*

Proof. By definition, S is convex if convex combinations of two points of S are in S. So half of the theorem is clear, and we only need to show that a convex combination of r points of a convex set S is in S, for any $r \geq 2$. We do this by induction on r. We start the induction at $r = 2$: this is the definition of convexity, so there is nothing to do.

Next we assume that the result is known for $r \geq 2$, namely that any convex combination of r points is in S, and we prove it for $r+1$. Let $\mathbf{x}^1, \ldots, \mathbf{x}^{r+1}$ be $r+1$ arbitrary points of S, and let

$$\mathbf{x} = \sum_{i=1}^{r+1} \lambda_i \mathbf{x}^i \text{ , where all } \lambda_i \geq 0 \text{ and } \sum_{i=1}^{r+1} \lambda_i = 1.$$

We need to show $\mathbf{x} \in S$. We may assume that $\lambda_i > 0$ for all i, since otherwise there is nothing to prove since there are only r terms. Let $\gamma = \sum_{i=1}^r \lambda_i$, so by the last remark $0 < \gamma < 1$. Then let

$$\gamma_i = \lambda_i / \gamma \text{ , } 1 \leq i \leq r,$$

so that the point $\mathbf{y} = \sum_{i=1}^r \gamma_i \mathbf{x}^i$ is a convex combination of r points of S, and is therefore in S by induction. Then $\mathbf{x} = \gamma \mathbf{y} + \lambda_{r+1} \mathbf{x}^{r+1}$, and $\gamma + \lambda_{r+1} = 1$, so $\mathbf{x}$ is a convex combination of two points of S and is therefore in S, since S is convex. □

Definition 15.3.17. For any set S, let $K(S)$ be the set of all finite convex combinations of points of S.

[4] If you do not remember what a tetrahedron is, you can use this as a definition.

By taking just one point in the convex combination, so $r = 1$ and $\lambda_1 = 1$, we see that $S \subset K(S)$. When S is empty, $K(S)$ is empty.

Theorem 15.3.18. *For any set S, $K(S)$ is a convex set.*

Proof. To show that $K(S)$ is convex, we need to show that if $\mathbf{k}_1$ and $\mathbf{k}_2$ are points of $K(S)$ then for any λ, $0 \leq \lambda \leq 1$, $\lambda\mathbf{k}_1 + (1-\lambda)\mathbf{k}_2$ is in $K(S)$. Since $\mathbf{k}_1$ is a convex combination of points of S, we have

$$\mathbf{k}_1 = \sum_{i=1}^{n} \mu_i \mathbf{x}_i, \text{ for } \mu_i \geq 0 \ , \ \sum_{i=1}^{n} \mu_i = 1,$$

and similarly for $\mathbf{k}_2$:

$$\mathbf{k}_2 = \sum_{j=1}^{m} \nu_j \mathbf{y}_j, \text{ for } \nu_j \geq 0 \ , \ \sum_{j=1}^{m} \nu_j = 1,$$

where the $\mathbf{x}_i$ and $\mathbf{y}_j$ are all in S. Then

$$\lambda\mathbf{k}_1 + (1-\lambda)\mathbf{k}_2 = \sum_{i=1}^{n} \lambda\mu_i \mathbf{x}_i + \sum_{j=1}^{m} (1-\lambda)\nu_j \mathbf{y}_j. \tag{15.24}$$

To show that the right-hand side is a convex combination of the $n+m$ points $\{x_i\}$ and $\{y_j\}$ we need to show that all the coefficients in (15.24) are non-negative, which is easy, and that they sum to 1, which we check:

$$\sum_{i=1}^{n} \lambda\mu_i + \sum_{j=1}^{m} (1-\lambda)\nu_j = \lambda \sum_{i=1}^{n} \mu_i + (1-\lambda) \sum_{j=1}^{m} \nu_j = \lambda + 1 - \lambda = 1,$$

so this is in $K(S)$. □

Theorem 15.3.19. *For any set S, the hull is equal to the set of convex combinations: $C(S) = K(S)$.*

Proof. By Theorem 15.3.18 $K(S)$ is convex, and it contains S. Since $C(S)$ is the intersection of all convex sets containing S, we have:

$$C(S) \subset K(S)$$

To get the opposite inclusion, take a convex combination $\sum_{i=1}^{r} \lambda_i \mathbf{x}_i$ of elements $\mathbf{x}_i$ of S, and an arbitrary convex set T containing S. All we need to do is show that this convex combination is in T. Since the $\mathbf{x}_i$ are in S, they are in T, and Theorem 15.3.16 shows that all convex combinations of points of T are in T, so we are done. □

An immediate corollary of this theorem is that any point in the convex hull of a set S can be written as a finite convex combination of points in S.

Since we no longer need to make the distinction between the convex hull and the set of all convex combinations, in both cases we write $K(S)$ and refer to it as the convex hull.

Theorem 15.3.20. *Let $T: V \to W$ be a linear transformation between two vector spaces V and W. Let S be a convex set in V. Then its image $T(S)$ under T is convex in W.*

Proof. Take any two points $\mathbf{p}$ and $\mathbf{q}$ in $T(S)$. We must show that for any λ, $0 < \lambda < 1$, $\lambda\mathbf{p} + (1-\lambda)\mathbf{q}$ is in $T(S)$. By definition of $T(S)$, there is a $\mathbf{a} \in S$ such that $T(\mathbf{a}) = \mathbf{p}$ and a $\mathbf{b} \in S$ such that $T(\mathbf{b}) = \mathbf{q}$. Since S is convex, for our choice of λ, $\lambda\mathbf{a} + (1-\lambda)\mathbf{b}$ is in S. By linearity of T,

$$T(\lambda\mathbf{a} + (1-\lambda)\mathbf{b}) = \lambda T(\mathbf{a}) + (1-\lambda)T(\mathbf{b}) = \lambda\mathbf{p} + (1-\lambda)\mathbf{q},$$

which is therefore in $T(S)$, as required. □

Definition 15.3.21. If S and T are non-empty subsets of $\mathbb{R}^n$, and a and b are fixed real numbers, then the *Minkowski sum* of S and T with coefficients a and b, written $aS + bT$, is

$$aS + bT := \{a\mathbf{s} + b\mathbf{t} \mid \forall \mathbf{s} \in S, \forall \mathbf{t} \in T\}.$$

If T is empty, then $aS + bT := aS$. Similarly, if S is empty, $aS + bT := bT$.

Proposition 15.3.22. *If S and T are convex, then so is the Minkowski sum $aS + bT$, for any choice of a and b.*

Proof. Pick two points $a\mathbf{s}_1 + b\mathbf{t}_1$ and $a\mathbf{s}_2 + b\mathbf{t}_2$ in $aS + bT$. We must show that for any λ, $0 < \lambda < 1$,

$$\lambda(a\mathbf{s}_1 + b\mathbf{t}_1) + (1-\lambda)(a\mathbf{s}_2 + b\mathbf{t}_2)$$

is in $aS + bT$. This can be written

$$a(\lambda\mathbf{s}_1 + (1-\lambda)\mathbf{s}_2) + b(\lambda\mathbf{t}_1 + (1-\lambda)\mathbf{t}_2)$$

and since S and T are both convex, this is in $aS + bT$. □

Exercise 15.3.23. Let S be a convex set in the plane with coordinates x and y. Assume S contains an entire line L. For simplicity, and without loss of generality, let L be the line with equation $y = 0$, namely the x-axis. What are all the possibilities for S?

Hint: S could be just the line L, or the entire plane, or the upper half-plane $y \geq 0$, or the lower half-plane $y \leq 0$. In order to analyze the remaining cases, assume that S only contains points in the upper half-plane. Assume that it contains a point $\mathbf{p}$ with second coordinate $y = a$, for some $a > 0$. Then show, by connecting $\mathbf{p}$ to points on the lines with very large and very small x coordinates, that S contains the entire strip of points (x, y) with $0 \leq y < a$. Finally let b be the greatest lower bound of y-coordinates of points in the upper half-plane that are not in S. Note that b is greater than or equal to any a found previously. Then show that S is contained in the strip of points (x, y) with $0 \leq y \leq b$. Then what can you say?

Exercise 15.3.24. If S is the closed ball of radius r_1 centered at $\mathbf{c}_1$, and T the closed ball of radius r_2 centered at $\mathbf{c}_2$, then $S + T$ is the closed ball B of radius $r_1 + r_2$ centered at $\mathbf{c}_1 + \mathbf{c}_2$.

Hint: First show that $S+T \subset B$, because every point in $S+T$ is at most at distance r_1+r_2 from $\mathbf{c}_1+\mathbf{c}_2$. Then show the opposite inclusion, by writing every point of the boundary of B as the sum of points from S and T. Make a picture in $\mathbb{R}^2$.

Definition 15.3.25. The *dimension* of a convex set C is the dimension of the affine hull of C, namely the flat of smallest dimension containing C.

Just as we defined linear independence and affine independence for a set of points, we can do the same for convex independence.

Definition 15.3.26. A set S of two or more points is *convexly independent* if no point s_0 in S is in the convex hull of the remaining points. A single point is convexly independent.

Exercise 15.3.27. Show that if a (necessarily finite) set of points is linearly independent, then it is affinely independent. If a set is affinely independent, then it is convexly independent. Given an example of

1. An infinite set of points that is convexly independent. Because it is infinite, it cannot be affinely independent;
2. A finite set of points that is convexly independent, and not affinely independent.
3. An affinely independent set of points that is not linearly independent.

The following lemma will be used in the proof of Theorem 15.3.29. Its proof is a simple exercise, and is left to you.

Lemma 15.3.28. *Assume a set S is not convexly independent, so that there is a point $\mathbf{s}_0 \in S$ that is a convex combination of other points of S. Then $\mathbf{s}_0$ is not extreme for the convex hull of S.*

Theorem 15.3.29. *If S is a finite set of points, then the extreme points E of the convex hull of S form the unique convexly independent subset of S with convex hull equal to the convex hull $K(S)$ of S.*

Proof. If the set S is not convexly independent, then an $\mathbf{s}_0 \in S$ can be written as a convex combination of the remaining points of S. Then remove $\mathbf{s}_0$ from S: the remaining points have the same convex hull as S. Continue doing this one point at a time until you are left with a convexly independent subset S^0 with the same convex hull as S. None of the removed points is extreme by Lemma 15.3.28, and conversely it is easy to see that the extreme points are all contained in S^0. Write the points of S^0 as $\mathbf{a}_i$, $0 \le i \le m$. To conclude the proof we must show that all the $\mathbf{a}_i$ are extreme. We prove this by contradiction. Assume, without loss of generality, that $\mathbf{a}_m$ is not extreme. Then it can be written as a combination $\mathbf{a}_m = \lambda\mathbf{p}+(1-\lambda)\mathbf{q}$, with $0<\lambda<1$ and $\mathbf{p}$ and $\mathbf{q}$ in $K(S)=K(S^0)$, with $\mathbf{p} \neq \mathbf{a}_m \neq \mathbf{q}$. Since $\mathbf{p}$ and $\mathbf{q}$ are in the convex hull of the S^0, they can be written

$$\mathbf{p} = \sum_{i=0}^{m} \mu_i \mathbf{a}_i, \quad \sum_{i=0}^{m} \mu_i = 1, \mu_i \geq 0;$$
$$\mathbf{q} = \sum_{i=0}^{m} \nu_i \mathbf{a}_i, \quad \sum_{i=0}^{m} \nu_i = 1, \nu_i \geq 0;$$

so that

$$\mathbf{a}_m = \lambda \mathbf{p} + (1-\lambda)\mathbf{q} = \sum_{i=0}^{m} \big(\lambda \mu_i + (1-\lambda)\nu_i\big)\mathbf{a}_i.$$

For all i, $0 \leq i \leq m$, define

$$\pi_i = \lambda \mu_i + (1-\lambda)\nu_i. \tag{15.25}$$

Then $\pi_i \geq 0$, as you should check, and

$$\sum_{i=0}^{m} \pi_i = \lambda \sum_{i=0}^{m} \mu_i + (1-\lambda) \sum_{i=0}^{m} \nu_i = \lambda + (1-\lambda) = 1. \tag{15.26}$$

Moving the term in $\mathbf{a}_m$ to the left-hand side, we get:

$$(1-\pi_m)\mathbf{a}_m = \sum_{i=0}^{m-1} \pi_i \mathbf{a}_i$$

If $1-\pi_m > 0$, divide by it to get

$$\mathbf{a}_m = \sum_{i=0}^{m-1} \frac{\pi_i}{1-\pi_m} \mathbf{a}_i.$$

Since all the coefficients in this sum are non-negative, and

$$\sum_{i=0}^{m-1} \frac{\pi_i}{1-\pi_m} = 1,$$

this expresses $\mathbf{a}_m$ as a convex combination of the remaining $\mathbf{a}_i$: a contradiction to the assumption of convex independence.

If $1-\pi_m = 0$, the only other possibility, then all the other π_i are 0, since they are non-negative and (15.26) holds. By (15.25), since λ and $1-\lambda$ are both positive, this forces $\mu_i = \nu_i = 0$, for $0 \leq i \leq m-1$. This in turn says that $\mathbf{p} = \mathbf{q} = \mathbf{a}_m$, so that $\mathbf{a}_m$ is extreme. So all the points in S^0 are extreme. □

15.4 Polyhedra and Polytopes

We use the notation of (15.8) for our polyhedron $P = P(A, \mathbf{b})$ in $\mathbb{R}^n$.

Definition 15.4.1. Given a vector $\mathbf{c} \in \mathbb{R}^n$, assume that there is a point $\mathbf{p} \in P$ such that for all $\mathbf{q} \in P$, $\langle \mathbf{c}, \mathbf{p} \rangle \leq \langle \mathbf{c}, \mathbf{q} \rangle$. Let $d = \langle \mathbf{c}, \mathbf{p} \rangle$. Geometrically this means that the affine hyperplane with equation $H_{\mathbf{c},d}$ meets P in at least one point - the point $\mathbf{p}$, and that the whole polyhedron lies in the half space $H^-_{\mathbf{c},d}$. Then the *face* of P defined by $\mathbf{c}$ is the intersection of P with the hyperplane $H_{\mathbf{c},d}$.

Theorem 15.4.2. *A simplex is a polyhedron.*

Proof. Suppose the simplex S is given by $n+1$ affinely independent vectors $\mathbf{a}_0$, $\mathbf{a}_1$, $\ldots$, $\mathbf{a}_n$ in $\mathbb{R}^n$. To establish the result, we will write S as the intersection of $n+1$ half-spaces.

For any j, $0 \leq j \leq n$, let H'_j be the flat that goes through $\mathbf{b}_i = \mathbf{a}_i - \mathbf{a}_j$, for all i except j. Since the $n+1$ points $\mathbf{a}_i$ are affinely independent, the n points $\mathbf{b}_i$ are linearly independent by Proposition 15.1.19, so H'_j is uniquely determined, as we showed in Lemma 15.1.23. Write the equation of H'_j as $c^j_1 x_1 + \cdots + c^j_n x_n = d_j$. The equation for the hyperplane H_j passing through the $\mathbf{a}_i$, $i \neq j$, is $c_j 1 x_1 + \cdots + c_{jn} x_n = e_j$, where $e_j = d_j + f(\mathbf{a}_j)$, so only the right-hand side of the equation changes. If you substitute for the x_k the coordinates a_{ik} of the i-th point $\mathbf{a}_i$, then the c_{jk} and e_j must satisfy these n equations. Now let H^+_j be the half-space bounded by H_j that contains the last generator $\mathbf{a}_j$ of the simplex S. Clearly H^+_j contains S and is convex. So the intersection $C := \cap^n_{j=0} H^+_j$ contains S and is convex. Any point $\mathbf{p}$ in $\mathbb{R}^n$ is an affine combination of the $\mathbf{a}_i$, so it can be written

$$\mathbf{p} = \sum_{i=1}^{k} \lambda_i \mathbf{a}_i$$

with $\sum^k_{i=1} \lambda_i = 1$. Those that are convex combinations of the $\mathbf{a}_i$ also have all $\lambda_i \geq 0$. Suppose that there is a point $\mathbf{p} \in C$ that is not a convex combination of the $\mathbf{a}^i$. Then there is an i such that $\lambda_i < 0$. We evaluate H_i on $\mathbf{p}$. By linearity we see that its value is λ_i times the value at $\mathbf{a}_i$, since it vanishes at all the other $\mathbf{a}_j$. Since λ_i is negative, the point $\lambda_i \mathbf{a}^i$ is inside the half-space H^-_j, so it is not in C, and we have our contradiction. □

Definition 15.4.3. The λ_i in (15.22) are the *barycentric coordinates* of the point $\mathbf{x}$ in the n-simplex $H(\mathbf{a}_0, \ldots, \mathbf{a}_n)$. The *barycenter* or *centroid* of the n-simplex is the point

$$\mathbf{c} := \frac{1}{n+1}(\mathbf{a}_0 + \mathbf{a}_1 + \cdots + \mathbf{a}_n) \tag{15.27}$$

of the simplex.

The barycentric[5] coordinates are uniquely determined for every point in the affine hull of the vertices of the simplex, as we saw in Theorem 15.1.21.

Definition 15.4.4. A *polytope* is the convex hull of a finite number of points.

[5] Introduced by Möbius in 1827: see [14], p. 134

The simplest polytope is the simplex.

Example 15.4.5. The unit cube in $\mathbb{R}^n$ is the polytope with vertices all the points whose coordinates are either 0 or 1. Thus it has 2^n vertices. Its vertices are the 2^n points whose coordinates are either 0 or 1. So in $\mathbb{R}^2$, the vertices are $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$.

The *crosspolytope* is the polytope in $\mathbb{R}^n$ with vertices the $2n$ points whose coordinates are all 0 except in one position, where the coordinate is either 1 or -1. The crosspolytope in $\mathbb{R}^2$ has vertices $(-1,0)$, $(1,0)$, $(0,-1)$, $(0,1)$ and thus is just a rotated square. In $\mathbb{R}^3$ its vertices are the six points $(-1,0,0)$, $(1,0,0)$, $(0,-1,0)$, $(0,1,0)$, $(0,0,-1)$, $(0,0,1)$, so it is not a rotated cube: it does not have enough vertices.

Remark 15.4.6. By Theorem 15.3.29, it suffices to consider polytopes on convexly independent sets of points, which are then called the *vertices* of the polytope. This definition agrees with the more general definition of vertex since we are just talking about the extreme points of the convex set.

Example 15.4.7. A simplex is *regular* if all its edges have the same length. Regular simplices exist in all dimensions. To construct a regular simplex of dimension n, take the $n+1$ unit coordinate vectors $\mathbf{e}_0$, $\mathbf{e}_1$, …, $\mathbf{e}_n$ in $\mathbb{R}^{n+1}$. These points all lie on the affine hypersurface with equation $x_0 + x_1 + \cdots + x_n = 1$, an n dimensional affine space. The $\mathbf{e}_i$ are affinely independent. Therefore any point in the polytope generated by the $\mathbf{e}_i$ can be written uniquely as

$$\sum_{i=0}^{n} \lambda_i \mathbf{e}_i$$

where, as usual, all the λ_i are nonnegative and $\sum_{i=0}^{n} \lambda_i = 1$. These are the barycentric coordinates of the simplex. The barycenter is the point in $\mathbb{R}^{n+1}$ with all coordinates equal to $1/(n+1)$, so by construction the distance of each vertex $\mathbf{e}_1$ from the barycenter is the length of the vector

$$\frac{1}{n+1}\big(n,-1,\ldots,-1\big).$$

Show this vector has length $r_n = \sqrt{n/(n+1)}$. Thus this regular simplex is inscribed in a sphere in $\mathbb{R}^n$ of radius r_n with center the barycenter, and inscribes the sphere of radius x centered at the barycenter. As n increases, r_n increases.

If the vertices of a polytope are $\mathbf{a}_0, \ldots, \mathbf{a}_m$, and we write A for the $(m+1) \times n$ matrix with rows $\mathbf{a}_i$, then we denote the polytope on these points by

$$P_A = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} = \sum_{i=0}^{m} \lambda_i \mathbf{a}_i, \text{ for } 0 \le \lambda_i \le 1, 0 \le i \le m, \sum_{i=0}^{m} \lambda_i = 1\}. \tag{15.28}$$

We will prove later that any bounded polyhedron is a polytope, Theorem 15.6.7, and then that any polytope is a polyhedron, the famous Weyl-Minkowski Theorem

15.7.11). Thus there must be a way of passing from the representation $P(A,\mathbf{b})$ for the bounded polyhedron to the representation P_A for the same set considered as a polytope: the matrix A will be different, of course. We have already shown how to do this for a simplex.

15.5 Carathéodory's Theorem

Next a beautiful and important theorem that tells us how many terms r we need in the convex combination of the convex hull of any set in $\mathbb{R}^n$.

Theorem 15.5.1 (Carathéodory's Theorem). *If S is a set in $\mathbb{R}^n$ and $\mathbf{x}$ a point in the convex hull of S, then $\mathbf{x}$ can be written as a convex combination of at most $n+1$ points in S.*

Proof. Theorem 15.3.19 says any $\mathbf{x} \in K(S)$ can be written as a convex combination of points in S, but it does not give us a bound. We find a bound by arguing by contradiction. Assume there is a point $\mathbf{x} \in K(S)$ for which the shortest representation as a convex combination of points of S required N points, $N > n+1$, so

$$\mathbf{x} = \sum_{i=1}^{N} \lambda_i \mathbf{x}_i \text{ where all } \mathbf{x}_i \in S \text{ , and } \lambda_i > 0 \text{ , and } \sum_{i=1}^{N} \lambda_i = 1$$

Consider the $N-1$ points $(\mathbf{x}_i - \mathbf{x}_N)$, $1 \le i \le N-1$. Since $N-1 > n$, these points are linearly dependent in $\mathbb{R}^n$, so we write an equation of linear dependence (so not all the coefficients are 0)

$$\sum_{i=1}^{N-1} \gamma_i (\mathbf{x}_i - \mathbf{x}_N) = 0$$

or

$$\sum_{i=1}^{N} \gamma_i \mathbf{x}_i = 0$$

where we have set $\gamma_N = -\sum_i^{N-1} \gamma_i$.

The following argument will be used many times in this chaptrt so it is worth remembering. Let t be a real variable. For every $t \in \mathbb{R}$ we can write

$$\mathbf{x} = \sum_{i=1}^{N} (\lambda_i - t\gamma_i)\mathbf{x}_i$$

Setting $\eta_i(t) = \lambda_i - t\gamma_i$, and recalling that the λ_i are all positive, our goal is to find a value of t so that all the η_i are non-negative, and at least one is 0. For such value of t the $\eta_i(t)$ give a representation of $\mathbf{x}$ as a convex combination of at most $N-1$ points of S, the desired contradiction. So look at the set I_+ of indices i where γ_i is positive. Since the sum of all the γ is 0, this set is non-empty. Consider the set of ratios λ_i/γ_i, $i \in I_+$. Pick an index i_0 for which this ratio is minimal, and let $t_0 = \lambda_{i_0}/\gamma_{i_0}$, so that

$\eta_{i_0}(t_0) = 0$ and all the other η are non-negative. Then $\mathbf{x}$ is a convex combination of fewer than N of the $\mathbf{x}_i$, the desired contradiction. □

Corollary 15.5.2. *If the dimension of the convex hull of S is $d < n$ then the estimate in Carathéodory's Theorem improves to $d+1$.*

Exercise 15.5.3. According to Definition 15.3.13, the n-simplex in $\mathbb{R}^n$ is the convex combination of its $n+1$ vertices S spanning $\mathbb{R}^n$. Show that there are points in the simplex that are not a convex combination of fewer than $n+1$ points, showing that Carathéodory's Theorem gives the best general bound for the number of points needed.

Definition 15.5.4. Let the polytope P (see Definition 15.4.4) in $\mathbb{R}^n$ be the convex hull of its m extreme points $\mathbf{a}^1, \ldots, \mathbf{a}^m$. Without loss of generality we can assume that the dimension of P is n, so $m \geq n+1$. To each set of $n+1$ affinely independent subsets of the m points $\mathbf{a}^i$ we can associate a simplex S_j with that set of points as vertices. These simplices are called the *subsimplices* of P.

Note that a simplex has only one subsimplex: itself.

Corollary 15.5.5. *A polytope is the union of its subsimplices.*

Proof. Just use the main argument in the proof of the theorem. □

Example 15.5.6. Find the subsimplices of the cube and the crosspolytope (see 15.4.5) in $\mathbb{R}^3$.

15.6 Minkowski's Theorem

The main theorem of this section is sometimes also called the Krein Milman Theorem. We will only prove the result for polytopes and bounded polyhedra, since those are the only cases that interest us.

Theorem 15.6.1 (Minkowski's Theorem). *Let C be a compact convex set, and let E be the set of extreme points of C. Then E is non-empty and C is the convex hull of E.*

Proof. We prove this by induction on the dimension of the convex set C. The result is clear if C has dimension 1 - and therefore is a closed interval: every point in C is in the convex hull of the two end points of the interval. Assume the result true for dimension $n-1$. Let C be a compact convex set of dimension n.

First let $\mathbf{x}$ be a boundary point of C, and H a supporting hyperplane of C through $\mathbf{x}$. $C \cap H$ is a compact convex set of dimension at most $n-1$. Thus by induction, $\mathbf{x}$ can be written as a convex combination of extreme points of $C \cap H$. But an extreme point of $C \cap H$ is an extreme point of C: $\mathbf{x}$ is not an interior point of a segment $[\mathbf{a}, \mathbf{b}] \in H$, because $\mathbf{x}$ is extreme in $C \cap H$. On the other hand $\mathbf{x}$ is not an interior

point of a segment $[\mathbf{a},\mathbf{b}]$ transverse to H, thus meeting H just in the point $\mathbf{x}$, since H is a supporting hyperplane of C, so that C is contained in one of the closed half-planes delimited by H.

Now assume $\mathbf{x}$ is not a boundary point of C: take any line ℓ through $\mathbf{x}$. Because C is compact, ℓ intersects C in two points $\mathbf{x}_1$ and $\mathbf{x}_2$ in the boundary of C, with $\mathbf{x} \in [\mathbf{x}_1,\mathbf{x}_2]$. By the previous argument, $\mathbf{x}_1$ and $\mathbf{x}_2$ are in the convex hull of extreme points, so is $\mathbf{x}$. □

As Example 15.3.9 shows, the number of extreme points of a compact convex set need not be finite. By Definition 15.4.4 a polytope has only a finite number of extreme points, and Corollary 15.6.4 shows the same is true for polyhedra.

Definition 15.6.2. If the compact convex set C has a finite number of extreme points, each extreme point of C is called a *vertex*.

We will use the word vertex and extreme point interchangeably.

Theorem 15.6.3. *Let* $\mathbf{p}$ *be a boundary point of the polyhedron* $P = P(A,\mathbf{b})$. *Then* $\mathbf{p}$ *is an extreme point of the polyhedron if and only if the normal vectors of the constraints that are active at* $\mathbf{p}$ *span* $\mathbb{R}^n$. *In particular there must be at least n active constraints at* $\mathbf{p}$ *for it to be an extreme point.*

Proof. First assume that the active normal vectors do not span. Then the intersection of the hyperplanes $H_{\mathbf{a}^i,b_i}$ is a positive dimensional linear space containing $\mathbf{p}$. So we can find a line segment $\mathbf{p}+t\mathbf{u}$, $-\varepsilon \le t \le \varepsilon$, $\varepsilon > 0$ in P so $\mathbf{p}$ is not extreme.

Next assume that the active $\mathbf{a}^i$ at $\mathbf{p}$ span. Assume that $\mathbf{p}$ is not extreme, and derive a contradiction. If $\mathbf{p}$ is not extreme, we can find $\mathbf{q}$ and $\mathbf{r}$ different from $\mathbf{p}$ in P with

$$\mathbf{p} = \frac{\mathbf{q}}{2} + \frac{\mathbf{r}}{2}. \tag{15.29}$$

For each active i, we have

$$\begin{aligned}
\langle \mathbf{a}^i, \mathbf{q}\rangle &\le b_i, \quad \text{because } \mathbf{q} \in P;\\
\langle \mathbf{a}^i, \mathbf{r}\rangle &\le b_i, \quad \text{because } \mathbf{r} \in P;\\
\langle \mathbf{a}^i, \mathbf{p}\rangle &= b_i, \quad \text{because } i \text{ is active at } \mathbf{p};\\
\frac{1}{2}\langle \mathbf{a}^i, \mathbf{q}\rangle + \frac{1}{2}\langle \mathbf{a}^i, \mathbf{r}\rangle &= b_i, \quad \text{by (15.29).}
\end{aligned}$$

Thus, for each active constraint, we have

$$\langle \mathbf{a}^i, \mathbf{p}\rangle = \langle \mathbf{a}^i, \mathbf{q}\rangle = \langle \mathbf{a}^i, \mathbf{r}\rangle = \mathbf{b}_i.$$

Since the active $\mathbf{a}^i$ span, the system of all $\langle \mathbf{a}^i, \mathbf{x}\rangle = \mathbf{b}_i$ has only one solution, so $\mathbf{q}$ and $\mathbf{r}$ are not distinct from $\mathbf{p}$. Thus $\mathbf{p}$ is extreme. □

Corollary 15.6.4. *A polyhedron has either no extreme points, or a finite number of extreme points.*

Proof. Theorem 15.6.3 tells us that the extreme points are the points where any set of at least n linear equations with linearly independent left-hand sides meet. For each set of n such equations there is at most one solution, so all in all there are only a finite number of solutions and therefore only a finite number of vertices. Indeed, the number of solutions is at most $\binom{m}{n}$. In particular if $m < n$ there are no extreme points. □

The corollary does not exclude the possibility that a polyhedron has no extreme points. Indeed, any polyhedron defined by fewer than n half-spaces has no extreme points.

Example 15.6.5. Consider the polyhedron P in $\mathbb{R}^3$ given by the inequalities $x \geq 0$, $y \geq 0$, $z \geq 0$ and $x+y \leq 3$, $-1 \leq z-x \leq 2$, and $y+z \leq 4$. We want to find the vertices of P, by considering the points in P where three inequalities with linearly independent directions vanish. Clearly the origin $\mathbf{0}$ is a vertex: it satisfies all the constraints and the three positivity constraints are active there. The next easiest vertices to find are those that are the intersection of two positivity constraints and one other equation. An easy computation gives the vertices $(1,0,0)$, $(0,3,0)$ and $(0,0,2)$. Next we find those where only one coordinate vanishes. Checking cases, we get $(1,2,0)$, $(2,0,4)$, $(3,0,4)$, $(3,0,2)$, $(0,3,1)$, $(0,2,2)$. There are no vertices where all three coordinates are non-zero: this is because the directions of the constraints (other than the positivity constraints) only span a 2-dimensional vector space. We end up with a compact polyhedron with 10 vertices: so it is the convex hull of these vertices.

The following corollary is interesting when attempting to make a polytope from a polyhedron..

Corollary 15.6.6. *Let P be a non-empty polyhedron in $\mathbb{R}^n$ given as the intersection of m half-spaces $H^-_{\mathbf{a}^i,b_i}$. Assume that the normal vectors $\mathbf{a}^i$ span $\mathbb{R}^n$, so that $m \geq n$. Then P has at least one extreme point.*

Proof. Pick a collection of n normal vectors that form a basis of $\mathbb{R}^n$. By reordering the half-spaces, we can assume they are $\mathbf{a}^i$, $1 \leq i \leq n$. The polyhedron P_0 which is the intersection of these n half-spaces clearly has a unique extreme point: the intersection $\mathbf{p}$ of the n linearly independent hyperplanes $H_{\mathbf{a}^i,b_i}$, $1 \leq i \leq n$. Next define the polyhedron P_1 to be the intersection of P_0 with $H_{\mathbf{a}^{n+1},b_{n+1}}$. If $\mathbf{p}$ is in $H_{\mathbf{a}^{n+1},b_{n+1}}$, it is an extreme point of P_1. Otherwise linear algebra tells us that we can find a subset of $n-1$ of the n half-spaces defining P_0, such that their normal vectors and $\mathbf{a}^{n+1}$ form a basis of $\mathbb{R}^n$. The intersection point $\mathbf{p}_1$ of the corresponding hyperplanes is an extreme point of P_1. Continuing in this way, adding one half-space at a time, gives the result. □

Compare this to Exercise 15.6.9.

Theorem 15.6.7. *A bounded polyhedron P is a polytope.*

Proof. Our goal is to apply Minkowski's Theorem 15.6.1. Since P is the intersection of m half-spaces given by $\mathbf{a}^i \cdot \mathbf{x} \leq b_i$, P is closed. Since it is bounded, it is compact.

Since it is a polyhedron, it is convex. Minkowski's Theorem tells us that P is the convex hull of its extreme points, which are finite in number by Corollary 15.6.4. Thus P is a polytope. □

We will prove the converse later: Corollary 15.7.11. We already proved the result for simplices in Example 15.4.2.

Exercise 15.6.8. Prove the following result. If C is a compact convex set, then a point $\mathbf{p} \in C$ at maximum distance from the origin is an extreme point of C. There is nothing special about the origin in this statement: any point will do.

Hint: If $\mathbf{p}$ is not extreme, then it is *between* two points $\mathbf{q}$ and $\mathbf{r}$ in C. A little geometry in the plane spanned by the three points $\mathbf{q}$, $\mathbf{r}$ and $\mathbf{0}$ gives the result.

Example 15.6.9. Prove that a closed convex set C has an extreme point if and only if it does not contain a line.

Hint: First assume C contains a line L. A point on the line clearly is not an extreme point. Pick a point $\mathbf{p}$ off the line that is extreme. Then Exercise 15.3.23 shows that in the plane spanned by L and $\mathbf{p}$, C contains a strip bounded by L on one side, and by the line L' parallel to L through $\mathbf{p}$ on the other. Because C is closed, L' is in C, and $\mathbf{p}$ is not extreme. This contradiction proves the result.

Now assume that C does not contain a line. Pick a point $\mathbf{q}$ in C. We now construct a function whose domain is the set of lines ℓ through $\mathbf{q}$. This set of lines is compact, by an argument similar to the one used in the proof of Theorem 13.2.7. Consider the function $d(\ell)$ that associates to ℓ the distance from $\mathbf{q}$ to the closest point where ℓ intersects the boundary of C. Since C contains no lines, this distance is finite. Show $d(\ell)$ is continuous, so it has a maximum. Conclude using Exercise 15.6.8.

15.7 Polarity for Convex Sets

It is not difficult to show that a closed set S is convex if and only if it is the intersection of all the half-spaces containing it. How can we describe these half-spaces? Here is an approach.

In Example 15.3.5 we wrote $H^-_{\mathbf{a},c} = \{\mathbf{x} \mid \langle \mathbf{a}, \mathbf{x} \rangle \leq c\}$. As long as $c \neq 0$, we get the same half-space by dividing the equation by c, so we look at: $H^-_{\mathbf{a},1} = \{\mathbf{x} \mid \langle \mathbf{a}, \mathbf{x} \rangle \leq 1\}$. This suggests that to the set S we associate all the vectors $\mathbf{a}$ so that S is contained in $H^-_{\mathbf{a},1}$. We make this into a definition:

Definition 15.7.1. Let S by a non-empty set in $\mathbb{R}^n$. Then the *polar set* S^* of S is given by

$$S^* = \{\mathbf{y} \in \mathbb{R}^n \mid \langle \mathbf{y}, \mathbf{x} \rangle \leq 1 \text{ for all } \mathbf{x} \in S\}. \tag{15.30}$$

Thus S lies in the intersection of the half-spaces $H^-_{\mathbf{y},1}$, for all $\mathbf{y} \in S^*$. Dually, S^* is the intersection of the half-spaces:

$$S^* = \bigcap_{\mathbf{x} \in S} H^-_{\mathbf{x},1}.$$

Example 15.7.2. If the set S contains a single point $\mathbf{a}$ other than the origin, then S^* is the closed half-space bounded by the hyperplane $H_{\mathbf{a},1}$ with equation $\langle \mathbf{a}, \mathbf{x} \rangle = 1$, that contains the origin.

If S only contains the origin, then S^* is all of $\mathbb{R}^n$.

Proposition 15.7.3. *If $S = \overline{N}_r(\mathbf{0})$, the closed ball of radius r centered at the origin, then $S^* = \overline{N}_{1/r}(\mathbf{0})$.*

Proof. This follows from the Cauchy-Schwarz inequality 8.2.7. To test if a non-zero element $\mathbf{y}$ is in S^*, dot it with the unique element $\mathbf{x}$ on the same ray through the origin and on the boundary of S. Then $\|x\| = r$ and

$$\langle \mathbf{y}, \mathbf{x} \rangle = \|\mathbf{x}\| \|\mathbf{y}\| = r\|\mathbf{y}\| \leq 1$$

so $\|\mathbf{y}\| \leq 1/r$. If this is true, then the Cauchy-Schwarz inequality shows us that for any $\mathbf{x} \in \overline{N}_r(\mathbf{0})$,

$$\langle \mathbf{y}, \mathbf{x} \rangle \leq \|\mathbf{x}\| \|\mathbf{y}\| \leq 1,$$

as required. □

We have the elementary

Theorem 15.7.4. *If $\{S_\alpha\}$ is an arbitrary collection of sets indexed by α, then the polar of the union of the $\{S_\alpha\}$ is the intersection of the polars of the S_α.*

From this we deduce the useful:

Theorem 15.7.5. *The polar of an arbitrary set S is a closed and convex set containing the origin.*

Proof. Write S as the union of its points, and notice from Example 15.7.2 that the polar of a point is convex, closed and contains the origin. By Theorem 15.3.4, any intersection of convex sets is convex, and any intersection of closed sets is closed (see Theorem 13.2.5, so we are done. □

Another elementary consequence of Theorem 15.7.4 is

Theorem 15.7.6. *If $S \subset T$, then $T^* \subset S^*$.*

Proof. Because $S \subset T$, S^* is the intersection of a smaller number of half-spaces than T^*, so certainly $T^* \subset S^*$. □

Theorem 15.7.7. *Assume that the polytope P has the points $\mathbf{a}^0, \ldots, \mathbf{a}^m$ as vertices. Then*

$$P^* = \{\mathbf{y} \mid \langle \mathbf{a}^i, \mathbf{y} \rangle \leq 1 \text{ for all } i = 0, \ldots, m\}. \tag{15.31}$$

Proof. This is easy. The right-hand side of (15.31) contains the left-hand side by the definition of the polar, so all we need is the opposite inclusion. So take any $\mathbf{y}$ satisfying the right-hand side inequalities. An arbitrary point in the polytope is given by (15.28). Dot this expression with $\mathbf{y}$ to get

$$\sum_{i=0}^{m} \lambda_i \langle \mathbf{a}^i, \mathbf{y} \rangle \leq \sum_{i=0}^{m} \lambda_i = 1,$$

since the λ_i are non-negative, and $\sum_{i=0}^{m} \lambda_i = 1$. Thus $\mathbf{y}$ is in P^*. □

Thus the polar of a polytope P_A is the polyhedron $P(A, \mathbf{1})$, using the notation of (15.28) and Definition 15.2.1.

The first important result of the section is

Theorem 15.7.8. *Let S be a compact and convex set of dimension n in $\mathbb{R}^n$ that contains the origin in its interior. Then S^* is a compact convex set of dimension n containing the origin in its interior.*

Proof. Theorem 15.7.5 tells us that S^* is closed, convex and contains the origin. Thus we need only prove that S^* is bounded and that the origin is an interior point. Because the origin is an interior point of S, for some small radius r, the ball $\overline{N}_r(\mathbf{0})$ is contained in S. But then S^* is contained in the ball $\overline{N}_{1/r}(\mathbf{0})$ by Proposition 15.7.3 and Theorem 15.7.6, which shows that S^* is bounded. Because S is compact it is bounded, so is a subset of $\overline{N}_R(\mathbf{0})$ for a large enough R. Proceeding as before, this shows that the ball $\overline{N}_{1/R}(\mathbf{0})$ is contained in S^*, showing that the origin is an interior point. □

The next step is to apply polarity twice.

Definition 15.7.9. The *bipolar* S^{**} of a set S as the polar of the polar of S, $S^{**} = (S^*)^*$.

Then in complete generality we have $S \subset S^{**}$. Indeed, rewrite (15.30) for S^*:

$$S^{**} = \{\mathbf{x} \in \mathbb{R}^n \mid \langle \mathbf{y}, \mathbf{x} \rangle \leq 1 \text{ for all } \mathbf{y} \in S^*\}. \tag{15.32}$$

Comparing this to (15.30) shows that if $\mathbf{x}$ is in S, then it is in S^{**}, so $S \subset S^{**}$.

Now the main result of this section. It uses the Separation Theorem for closed convex sets, that we have not proved.

Theorem 15.7.10 (The Bipolar Theorem). *Let S be a closed convex set containing the origin. Then the bipolar S^{**} of S is equal to S.*

Proof. We have just established the inclusion $S \subset S^{**}$. To get the opposite inclusion, pick a point $\mathbf{b}$ not in S. We must show it is not in S^{**}. Since S is convex and closed, by the Separation Theorem, we can find a hyperplane $H = \{\mathbf{x} \mid \langle \mathbf{a}, \mathbf{x} \rangle = 1\}$ strictly separating S and $\mathbf{b}$. Because $\mathbf{0} \in S$, we have $\langle \mathbf{a}, \mathbf{x} \rangle < 1$ for all $\mathbf{x} \in S$, and $\langle \mathbf{a}, \mathbf{b} \rangle > 1$. The first inequality says that $\mathbf{a}$ is in S^*, from which the second inequality say that $\mathbf{b}$ is not in S^{**}, and we are done. □

By this result and Theorem 15.7.5 we see that $S = S^{**}$ if and only if S is a closed convex set containing the origin.

Corollary 15.7.11. *A polytope is a polyhedron, and a bounded polyhedron is a polytope.*

Proof. The last statement is Theorem 15.6.7, so we need only prove the first one. By restricting to the affine hull of the polytope P, we can assume it has maximum dimension, so that it has a non-empty interior. Then by translating it, we can make the origin an interior point. Then by Theorem 15.7.8, P^* is compact, and by Theorem 15.7.7 it is a polyhedron, therefore a bounded polyhedron. So by Theorem 15.6.7, P^* is a polytope, so its polar $(P^*)^*$ is a polyhedron. By the Bipolar Theorem, $(P^*)^* = P$, so P is a polyhedron as claimed. □

We now see that bounded polyhedra and polytopes are the same. This result is known as the Weyl-Minkowski Theorem: see [31].

We could pursue this line of inquiry by determining the polar of a given polytope. Example 15.4.2 shows that the polar polytope of a simplex is again a simplex. This is investigated in [17], chapter 9, which is a good reference for the material in this section. A more advanced reference is [3], chapter IV.

Example 15.7.12. Show that the polar polytope of the cube is the crosspolytope. from which is follows that the polar polytope of the crosspolytope is the cube. First work this out in $\mathbb{R}^2$ and $\mathbb{R}^3$.

Chapter 16
Linear Differential Equations

Abstract Linear algebra is a very useful tool for studying differential equations. We illustrate this with the simplest differential equations: linear differential equations with constant coefficients. Our approach is to solve the equations without the existence and uniqueness theorems for linear equations: instead we use only elementary results of linear algebra. We first study a single linear differential equation with constant coefficients of high order, and determine all its solutions using the primary decomposition theorem of Chapter §10. Then we study a system of linear equations with constant coefficients of order one. We study these, which generalize what we did in the first part of the chapter, which leads us to the matrix exponential and an application of the Jordan form in the same chapter.

16.1 Differential Calculus Review

Our vector space will be the space $\mathscr{C}^\infty$ of infinitely differentiable functions $f(t)$ defined on the entire line $\mathbb{R}$. We call the variable t because we are thinking of it as time. $\mathscr{C}^\infty$ is a vector space because the sum of two infinitely differentiable functions is infinitely differentiable, as is a scalar multiple of such a function. Unlike most of the vector space we have studied it is infinite dimensional. Indeed, inside $\mathscr{C}^\infty$ lies the subspace of polynomials of all degrees. It is already infinite dimensional with basis $1, t, t^2, \ldots, t^n, \ldots$. Consider the map from $\mathscr{C}^\infty$ to $\mathscr{C}^\infty$ given by differentiation d/dt. It is a linear operator since if $f(t)$ and $g(t)$ are in $\mathscr{C}^\infty$ and c is a real number, we have:

$$\frac{d}{dt}\big(f(t)+g(t)\big) = \frac{d}{dt}f(t)+\frac{d}{dt}g(t)$$
$$\frac{d}{dt}\big(cf(t)\big) = c\frac{d}{dt}f(t)$$

and the range is in $\mathscr{C}^\infty$ since $\frac{d}{dt}f(t)$ is infinitely differentiable. For simplicity of notation we write d/dt as D. As is the case for any linear operator we can compose D

with itself, getting the higher order derivatives. Consider polynomials in the operator D wisth real coefficients:

$$D^n + a_{n-1}D^{n-1} + \cdots + a_1 D + a_0 I, \tag{16.1}$$

where the a_i are real numbers. They are linear operators as studied in Chapter 10. They all commute.

Calculus tells us that the nullspace $\mathscr{N}(D)$ of D itself consists of the constant functions, so it has dimension 1. Similarly the nullspace of D^n has dimension n, as you see by iterating D: it consists of the polynomials of degree at most $n-1$. We also need to know that the exponential e^t and the trigonometric functions $\sin t$ and $\cos t$ are in $\mathscr{C}^\infty$, and that their derivatives are, respectively, e^t, $\cos t$ and $-\sin t$. Finally recall that e^t does not vanish on $\mathbb{R}$.

We first consider differential operators in the form (16.1). To solve the differential equation associated to this operator means determining its nullspace, that is the functions $f(t)$ satisfying

$$D^n f + a_{n-1}D^{n-1} f + \cdots + a_1 Df + a_0 f = 0, \tag{16.2}$$

Thus we can associate the polynomial

$$x^n + a_{n-1}x^{n-1} f + \cdots + a_1 x + a_0 = 0 \tag{16.3}$$

to the differential equation, simply by replacing the linear operator D by the polynomial variable x.

16.2 Examples

We start with the simplest possible differential equation after $d^n f = 0$, where we have already determined the solutions. Then we have:

Example 16.2.1. The differential equation $Df(t) - af(t)I = 0$, where $f(t) \in \mathscr{C}^\infty$. We now show that its only solutions, that is, the functions is the nullspace of the operator $D - aI$, are the ce^{at}, for an arbitrary $c \in \mathbb{R}$. Thus the nullspace has dimension 1. The constant c is the initial value $f(0)$, so there is a unique solution with given initial value.

Proof. An easy differentiation shows that ce^{at} is a solution. As noted e^{at} is nowhere 0 on $\mathbb{R}$. So consider any other solution $g(t)$ of the differential equation. Compute the derivative of

$$\frac{g(t)}{e^{at}}.$$

The quotient rule and the chain rule say that the derivative is

$$\frac{g'(t) - ag(t)}{e^{at}}.$$

Since $g(t)$ is a solution of the differential equation, this vanishes, so the derivative of $\frac{g(t)}{e^{at}}$ is 0. Thus $g(t) = ce^{at}$ for some constant c and $g(0) = c$. □

Example 16.2.2. Now we compute the solutions of the differential equation

$$f'' + c^2 f = 0 \text{ for some real number } c \neq 0.$$

Then the solutions are linear combinations of $\cos(ct)$ and $\sin(ct)$. For initial values $f(0) = c$ and $f'(0) = d$ there is a unique solution $f(t) = c\cos(ct) + d\sin(ct)$.

Proof. Differentiation shows that $\cos(ct)$ and $\sin(ct)$ are solutions. These two solutions are linearly independent over $\mathbb{R}$: assume by contradiction that there is an equation of linear dependence, and evaluate at $t = 0$ and $t/(2c)$ to get a contradiction. Finally we show that any solution $f(t)$ is a linear combination of these two.

Suppose that $g(t)$ is either $\sin(ct)$ or $\cos(ct)$.

Consider the expression $f(t)g'(t) - f'(t)g'$. Its derivative is

$$fg'' + f'g' - f'g' + f''g = fg'' + f''g.$$

If $g(t) = \sin(ct)$ this derivative is

$$-f(t)c^2\sin(t) - f''\sin(t) = -\sin(t)(c^2 f(t) + f(t) = 0$$

since f is assumed to be a solution. Thus

$$-f(t)c\cos(ct) - f'(t)\sin(t) = a, \tag{16.4}$$

a constant.

Similarly setting $g(t) = \cos(t)$, we get

$$-f(t)c\sin(ct) + f'(t)\cos(ct) = b, \tag{16.5}$$

another constant. We do Gaussian elimination on functions by multiplying (16.4) by $\cos(ct)$ and (16.5) by $\sin(ct)$, and adding. We get

$$cf(t) = a\cos(ct) + b\sin(ct)$$

so $f(t)$ is a linear combination of the two previous solutions. Thus the space of solutions has dimension 2. At $t = 0$ we have $cf(0) = a$ and $cf'(0) = bc$ so there is a unique solution satisfying these initial conditions. □

Example 16.2.3. Next we compute the solutions of the differential equation

$$f'' - c^2 f = 0 \text{ for some real number } c \neq 0.$$

Differentiation shows that e^{ct} is a solution as is e^{-ct}. These functions are linearly independent, so we have a set of solutions of dimension at least two. As before for initial values $f(0) = d$ and $f'(0) = e$ there is a unique solution $f(t) = ce^{ct} + de^{ct}$.

Proof. By the same method as in the previous example, we show that linear combinations of these are the only solutions. Take any solution $f(t)$. Then as before form the functions

$$f(t)ce^{ct} - f'e^{ct}$$
$$-f(t)ce^{-ct} - f'e^{-ct}$$

and compute their derivatives as in the previous example. In both case you get a factor $f'' - c^2 f$, so since f is a solution of the differential equations both derivatives are 0. Thus the functions are constants, so for example

$$cf(t)e^{ct} - f'e^{ct} = a$$
$$-cf(t)e^{-ct} - f'e^{-ct} = b$$

Multiply the first equation by e^{-ct} and the second by e^{ct} and subtract. The derivative terms vanish, so since $e^{ct}e^{-ct} = 1$, we get $cf(t) = ae^{-ct} + be^{ct}$ as required. So we get a two-dimensional space of solutions. Obviously $cf(0) = a + b$ and $cf'(0) = -ca + bc$ showing that for given initial conditions there is a unique pair (a,b) as required. □

Example 16.2.4. In Example 16.2.2 the polynomial $t^2 + c^2$ has roots $\pm$i. We will also need to analyze what happens when the roots, more generally, are any two complex conjugate numbers $a \pm ib$, where a and $b \neq 0$ are real. These complex numbers are the two roots of the real polynomial $t^2 - 2at + a^2 + b^2$. So consider the differential operator $D^2 - 2aD + (a^2 + b^2)I$. From Example 16.2.1 we see that the functions $e^{(a\pm ib)t}$ are complex solutions of the differential equation. By taking the right complex linear combinations $c_1 e^{(a+ib)t} + c_2 e^{(a-ib)t}$, we get for real solutions the functions $e^{at}\cos bt$ and $e^{at}\sin bt$, using the fact that $e^{(a+ib)t} = e^{at}(\cos(bt) + i\sin(bt))$, as we point out in the complex variables review in Appendix B.5. So we get a two-dimensional space of solutions. Instead of proving that these are the only solutions using a trick similar to the one above, we will rely on the general theorem 16.3.1. For later on, note that the equation $D^2 f - 2aD + (a^2 + b^2)I$ can be replaced by the 2×2 system in the two unknown functions f and its derivative f':

$$\begin{pmatrix} f' \\ f'' \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -(a^2+b^2) & 2a \end{pmatrix} \begin{pmatrix} f \\ f' \end{pmatrix} \tag{16.6}$$

Example 16.2.5. Our fifth example is the differential equation $(D - aI)^n f = 0$, Writing this out with derivatives, and expanding using the binomial theorem we have

$$f^{(n)} - naf^{(n-1)} + \binom{n}{2} a^2 f^{(n-2)} - \cdots + (-1)^{n-1} na^{n-1} f' + (-1)^n a^n = 0.$$

e^{at} is a solution because it is already a solution of $(D - aI)f = 0$. We use the same trick as before. Let $f(t)$ be any solution, and take the function $\frac{f(t)}{e^{at}}$. Compute its derivatives:

$$D\Big(\frac{f(t)}{e^{at}}\Big) = \frac{f' - af}{e^{at}}$$
$$D^2\Big(\frac{f(t)}{e^{at}}\Big) = \frac{f'' - 2af' + a^2 f}{e^{at}}$$
$$\vdots = \ddots$$
$$D^n\Big(\frac{f(t)}{e^{at}}\Big) = \frac{f^{(n)} - naf^{(n-1)} + \binom{n}{2}a^2 f^{(n-2)} - \cdots + (-1)^{n-1}na^{n-1}f' + (-1)^n a^n}{e^{at}}$$

Notice that the numerator on the rhs is just the lhs of the differential equation, so for the solution f it is 0. Thus the n-th derivative of $\frac{f(t)}{e^{at}}$ is a constant, so $\frac{f(t)}{e^{at}}$ is a polynomial of degree $\leq n-1$. Thus

$$f(x) = a_0 e^{at} + a_1 t e^{at} + \cdots + a_{n-1} t^{n-1} e^{at},$$

for any choice of constants a_i. As we have already noticed the monomials 1, t, $\ldots$, t^{n-1} are linearly independent , so multiplying by e^{at}, linear independence is preserved. So the vector space of solutions has dimension n.

16.3 The General Case

After these examples we consider the general case, the differential equation

$$\frac{d^n}{dt^n}f(t) + a_{n-1}\frac{d^{n-1}}{dt^{n-1}}f(t) + \cdots + a_1\frac{d}{dt}f(t) + a_0 = 0. \tag{16.7}$$

To solve it, factor the polynomial

$$P(t) = t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0$$

into distinct factors, so we have

$$P(t) = (t - r_1)^{m_1}(t - r_2)^{m_2}\ldots(t - r_k)^{m_k}$$

Since the polynomial $P(t)$ has real coefficients, in order to get linear factors we must allow the r_i to be complex. They come in complex conjugate pairs, as we have seen.

Theorem 16.3.1. *The vector space $\mathscr{N}$ of solutions of* (16.7) *is the direct sum of the solutions of*

$$\Big(\frac{d}{dt} - r_i I\Big)^{m_i} f(t) = 0. \tag{16.8}$$

Furthermore $\mathscr{N}$ has dimension n.

Proof. This is just a restatement of the Primary Decomposition Theorem 10.5.2. Indeed the linear operator D vanishes on the space $\mathscr{N}(D)$ by definition. The dimension statement was proved in (16.8) in Example 16.2.5. □

Finally we should consider the case where an r is complex. This is the case we analyzed, with multiplicity one, in Example 16.2.2, where the complex conjugate roots are $\pm$i. We showed that we can produce two linearly independent real solutions.

Next we handle the case $(D^2+bD+cI)f=0$ where the discriminant $\sqrt{b^2-4c}$ of the quadratic is negative. The roots can be written $a(b\cos\theta+\ddot{\imath}c\sin\theta)$, where $b^2+c^2=1$. Then take a power of this operator. Must show we get the right dimension for the real solutions.

16.4 Systems of First Order Differential Equations

Take any $n\times n$ matrix A over $\mathbb{R}$, and let $\mathbf{f}(t)$ be a vector of n $\mathbb{R}$ valued functions $f_j(t)$, $1\le j\le n$. Then form the system of linear differential equations:

$$A\mathbf{f}(t)=\mathbf{f}'(t) \tag{16.9}$$

with the initial value condition $\mathbf{f}(0)=\mathbf{k}$, where $\mathbf{k}$ is a given vector of constants.

Example 16.4.1. We can always put the high order differential equation (16.2) into this form, so what we do now generalizes what we did in the first part of this chapter. Indeed let $f_1(t)=f'(t)$, $f_2(t)=f''(t),\dots,f_{n-1}(t)=f^{(n-2}(t)$, so that (16.2) become

$$Df_{n-1}+a_{n-1}f_{n-1}+\cdots+a_1f_1+a_0f=0$$

Renaming f to f_0 for convenience, our single differential equation is replaced by the $n\times n$ system

$$\mathbf{f}'+A\mathbf{f}=0,$$

where $\mathbf{f}=(f_0,f_1,\dots,f_{n-1})$, $\mathbf{f}'=(f_0',f_1',\dots,f_{n-1'})$ and

$$A=\begin{pmatrix} 0 & -1 & 0 & \dots & 0 \\ 0 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \dots & 0 & -1 \\ -a_0 & -a_1 & \dots & \dots & -a_{n-1} \end{pmatrix}$$

Unsurprisingly this is the companion matrix (12.7) of the polynomial (16.3), up to transpose and a sign. Therefore we know that the minimal polynomial of the matrix A is its characteristic polynomial. If the characteristic polynomial has distinct roots over $\mathbb{C}$, then A is diagonalizable over $\mathbb{C}$. However if the characteristic polynomial has a repeated root λ, then for that λ A has a single Jordan block of size equal to the multiplicity of the root.

The most favorable case for the system (16.9) occurs when A is diagonal, so that the functions are "decoupled": the equations are $f_j'(t) = a_{jj} f_j(t) = 0$, $1 \leq j \leq n$, which we know how to solve by Theorem 16.2.1.

Now take an arbitrary A. By definition A is similar to its Jordan form J, so there exists an invertible matrix C so that $A = CJC^{-1}$. Then define $\mathbf{g}(t) = C^{-1}\mathbf{f}(t)$ and $\mathbf{l} = C^{-1}\mathbf{k}$. Since $A\mathbf{f}(t) = \mathbf{f}'(t)$, we get $AC\mathbf{g}(t) = C\mathbf{g}'(t)$ or $C^{-1}AC\mathbf{g}(t) = \mathbf{g}'(t)$ with initial condition $\mathbf{l} = C^{-1}\mathbf{f}_0$. So clearly it is equivalent to solve the initial condition problem in $\mathbf{f}$ or the new one in $\mathbf{g}$. So we need only consider the case where A is already in Jordan form.

Thus the only case that remains is that of a Jordan block (10.10) of size $r \geq 2$, since there still is interaction between the variables in the block. If a is the term on the diagonal of J, then the equations are

$$\begin{aligned} f_1'(t) &= af_1(t) \\ f_2'(t) &= f_1(t) + af_2(t) \\ \dots &= \dots \\ f_n'(t) &= f_{n-1}(t) + af_n(t) \end{aligned}$$

So, using the initial conditions, $f_1(t) = l_1 e^{at}$, so the second equation is $f_2'(t) = l_1 e^{at} + af_2(t)$ which has a solution $f_2(t) = e^{at}(l_1 t + l_2)$, which gives the desired initial condition.

Continuing in this way, we can solve the system for any set of initial conditions:

$$f_k(t) = e^{at}\left(l_k + l_{k-1}t + l_{k-2}\frac{t^2}{2} + \cdots + l_{k-i}\frac{t^i}{(k-i)!} + \cdots + l_1 t^{k-1}\right)$$

Why is this the only solution. These differential equations are not homogenous. As in systems of linear equations, the difference of any two solutions of the inhomogeneous equation is a solution of the homogenous equation, and we have found all the solutions of the homogenous equation in the previous section.

16.5 Eigenvector Computations for Linear ODE

16.6 Difference Equations

Appendix A
Notation

A.1 Generalities

The Kronecker δ_{ij} is the function of i and j that takes the value 1 when $i = j$ and the value 0 when $i \neq j$. Here i and j are integers, usually the indices in double sums as occur often in linear algebra.

A.2 Real and Complex Vector Spaces

The term scalar denotes either a real or a complex number. F denotes either the field of real numbers or of complex numbers.

The n-th cartesian product of scalars F is written F^n. Lower-case bold letters such as $\mathbf{x}$ and $\mathbf{a}$ denote vectors in F^n, each with coordinates represented by non-bold letters $(x_1, \ldots, x_n)$ and $(a_1, \ldots, a_n)$, respectively. We typically use $\mathbf{x}$ (and $\mathbf{y}$, $\mathbf{z}$, etc.) for unknowns and $\mathbf{a}$ (and $\mathbf{b}$, $\mathbf{c}$, etc.) for constants.

Vectors are also called points, depending on the context. When the direction is being emphasized, it is called a vector.
With the exception of gradients, vectors are always column matrices.
In the body of the text, an expression such as $[a_1, a_2, \ldots, a_n]$ denotes a column vector while $(a_1, a_2, \ldots, a_n)$ denotes a row vector.
The length of a vector $\mathbf{v}$ is written $\|\mathbf{v}\|$. If $\mathbf{v}$ is real, this is $\sqrt{v_1^2 + \cdots + v_n^2}$ while if $\mathbf{v}$ is complex this is $\sqrt{v_1\overline{v_1} + \cdots + v_n\overline{v_n}}$. The inner product of $\mathbf{v}$ and $\mathbf{w}$ is $\langle \mathbf{v}, \mathbf{w} \rangle$, or, more rarely, $\mathbf{v} \cdot \mathbf{w}$. The context tells you whether it is a scalar product or a hermitian product.
The linear span of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_r$ in a vector space is written $\mathrm{lin}(\mathbf{x}_1, \ldots, \mathbf{x}_r)$.
The partial order in $\mathbb{R}^n$ leads to the following notation:

$\mathbf{x} \prec \mathbf{y}$ means that $x_i < y_i$ for all $1 \leq i \leq n$
$\mathbf{x} \preceq \mathbf{y}$ means that $x_i \leq y_i$ for all $1 \leq i \leq n$
$\mathbf{x} \preccurlyeq \mathbf{y}$ means that $x_i \leq y_i$ for all $1 \leq i \leq n$ and $x_j < y_j$ for some j

and therefore

$$\mathbb{R}^n_{\succ} = \{\mathbf{x} \mid \mathbf{x} \succ \mathbf{0}\}$$
$$\mathbb{R}^n_{\succeq} = \{\mathbf{x} \mid \mathbf{x} \succeq \mathbf{0}\}$$
$$\mathbb{R}^n_{\succcurlyeq} = \{\mathbf{x} \mid \mathbf{x} \succcurlyeq \mathbf{0}\}$$

The open ball of radius r centered at the point $\mathbf{p} \in \mathbb{R}^n$ is written

$$N_r(\mathbf{p}) = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{p}\| < r\}$$

and the closed ball

$$\overline{N}_r(\mathbf{p}) = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{p}\| \leq r\}$$

A.3 Matrices

Matrices are written with round brackets as in

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

for the $m \times n$ matrix A. Matrices are denoted by capital roman letters such as A, and have as entries the corresponding lower case letter. So $A = (a_{ij})$. A is an $m \times n$ matrix if it has m rows and n columns, so $1 \leq i \leq m$ and $1 \leq j \leq n$. We write the columns of A as $\mathbf{a}_j$ and the rows as $\mathbf{a}^i$.

If $\mathbf{b}_1, \dots, \mathbf{b}_n$ is any collection of m-vectors, then

$$\begin{pmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \dots & \mathbf{b}_n \end{pmatrix}$$

is the $m \times n$ matrix whose j-th column is $\mathbf{b}_j$.

This implies that if, as always, the $\mathbf{a}^i$ are the rows of the $m \times n$ matrix A, then the $n \times m$ matrix $\begin{pmatrix} \mathbf{a}^1 & \mathbf{a}^2 & \dots & \mathbf{a}^m \end{pmatrix}$ is A^t, the transpose of the matrix A. This is because the vectors $\mathbf{a}^i$ are considered as column vectors.
$D(d_1, d_2, \dots, d_n)$ is the $n \times n$ diagonal matrix

$$\begin{pmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ 0 & 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_n \end{pmatrix}$$

where the only non-zero terms are along the principal diagonal where the row and column indices are equal. I_n or just I stands for the $n \times n$ identity matrix $D(1,1,\dots,1)$.

A.4 Linear Transformations

If A is an $m \times n$ matrix, L_A is the linear transformation from F^m to F^n given by $L_A(\mathbf{x}) = A\mathbf{x}$, the matrix product of A by the n-column vector $\mathbf{x}$. The nullspace of T_A is written $\mathscr{N}(A)$, and its range $\mathscr{R}(A)$.
A linear operator is a linear transformation from a vector space to itself.

Appendix B
Math Review

Abstract We review some basic mathematics results that will be used throughout the book: sets, mappings and equivalence relations. There is then a longer section on complex numbers.

B.1 Sets and Maps

Before getting starting we need to set up some terminology concerning sets and maps. If S and T are two sets, then $S \cup T$ is their union, and $S \cap T$ their intersection. If the intersection is empty, we write $S \cap T = \emptyset$. If S is a subset of T, we write $S \subset T$. The elements of the set T that are not in S is denoted $T \setminus S$.

Some important sets for us are $\mathbb{N}$, the natural numbers, namely the positive integers; $\mathbb{Z}$ the integers; $\mathbb{Q}$ the rational numbers; $\mathbb{R}$ the real numbers; $\mathbb{C}$ the complex numbers. If $\alpha \in \mathbb{C}$, so it can be written $a + \mathrm{i}b$, for $a \in \mathbb{R}$ and $b \in \mathbb{R}$, then $\overline{\alpha}$ is the complex conjugate of α, namely $a - \mathrm{i}b$. Here, as usual i is the imaginary number such that $\mathrm{i}^2 = -1$.

We can build a new set $S \times T$ from two sets S and T. It is the cartesian product of S and T: the set of all pairs (s,t), for $s \in S$ and $t \in T$. If S and T are the same set, then we write S^2 for the cartesian product. Note that if $s_1 \neq s_2$, then the element (s_1, s_2) if different from the element (s_2, s_1), which is why the product is often called the ordered product. You are certainly familiar with this notation for the sets $\mathbb{R}^2$, and $\mathbb{C}^2$.

Exercise B.1.1. Explain how to build sets $\mathbb{R}^n$, and $\mathbb{C}^n$, for any $n \in \mathbb{N}$.

Let $f\colon S \to T$ be a map between two sets S and T. This simply means that to each $s \in S$, the map f assigns a well-defined element $f(s) \in T$. The map f is injective if whenever $f(s_1) = f(s_2)$, then $s_1 = s_2$. The map f is surjective (or onto) if for all $t \in T$ there is a $s \in S$ such that $f(s) = t$. Finally f is an isomorphism if it is both injective and surjective.

Here is another way of saying this. The inverse image of an element $t \in T$ is the set S_t of $s \in S$ such that $f(s) = t$. For a general f, S_t could be empty. Then

- f is injective if and only if for all $t \in T$, S_t is either empty or a single element;
- f is surjective is for all $t \in T$, S_t is not empty.

If f is a map from a set S to itself, then it is called the identity map if for all $s \in S$, $f(s) = s$.

If f is a map from S to T, and g a map from T to U, then we may define the composite map $g \circ f$ from S to U by setting

$$(g \circ f)(s) = g(f(s).$$

Note that the right hand side makes sense because $f(s) \in T$.

We say that the map $g\colon T \to S$ is the inverse of f if the two composite maps:

$$g \circ f\colon S \to S \text{ and } f \circ g\colon T \to T$$

are both the identity map.

Exercise B.1.2. If $f\colon S \to T$ is both injective and surjective, it has an inverse.

The following exercises show how this depends on the definition of the domain and the range of f.

Exercise B.1.3. Let S be the non-negative real numbers, and let f be the map from S to itself taking s to s^2. Show that f is invertible.

Exercise B.1.4. Let S be the real numbers and T the positive numbers. Let f be the exponential map, associating to any real number s its exponential e^s. Then f is an isomorphism from S to T. What is its inverse?

Finally we record a simple fact that is known as the associativity of maps. Suppose we have a third map $h\colon U \to W$. The we can form the composite: $h \circ g \circ f\colon S \to W$. This could potentially depend on the order in which the maps are evaluated: $h \circ (g \circ f)$ versus $(h \circ g) \circ f$.

Theorem B.1.5. *Composition of maps is associatve, so $h \circ (g \circ f)$ is the same map as $(h \circ g) \circ f$.*

Proof. We need to show that for all $s \in S$, the value is the same, regardless of the order of evaluation. Just one line.

$$h \circ (g \circ f)(s) = h((g \circ f)(s)) = h(g(f(s))) = (h \circ g)(f(s)) = (h \circ g) \circ (f)(s)).$$

□

Definition B.1.6 (Kronecker Delta). One simple but useful function from $\mathbb{N} \times \mathbb{N}$ which takes the values 0 or 1 is the Kroncker delta, always written δ_{ij}, for $i, j \in \mathbb{N}$ where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. The Kronecker delta is very useful when dealing with two indices i and j, and since double indices are rampant in linear algebra, this function with be very useful for us.

B.2 Equivalence Relations

In Chapter 1 we establish an equivalence relation between systems of linear equations in the same variables. We establish relationships between matrices later in this course by introducing equivalence relations. So let us review equivalence relations here.

If S is a set, a binary relation compares two elements of S. This means that we compare any two elements s and t of S, and associate a binary outcome to the comparison. By convention we say that the outcome is either true or false (we could also have said 1 or 0). it is traditional to express that the outcome is true by writing $s \sim t$.

We are only interested in a specific kind of binary relation, called an equivalence relation. It has three properties: it is reflexive, symmetric and transitive:

Definition B.2.1. Let $\sim$ be a binary relation on a set S. Then

1. $\sim$ is *reflexive* when $s \sim s$ for all $s \in S$.
2. $\sim$ is *symmetric* when $s \sim t$ implies $t \sim s$.
3. $\sim$ is *transitive* when $s \sim t$ and $t \sim u$ imply $s \sim u$.

Read "$s \sim t$" as "s is equivalent to t". The most familiar example of an equivalence relation is probably congruence:

Example B.2.2. Congruence modulo a positive integer k is an equivalence relation on the set of integers $\mathbb{Z}$, defined as follows: Two integers a and b are congruent modulo k, if they have the same remainder under division by k. Each equivalence class contains all the integers whose remainder modulo division by k is a fixed integer. Thus there are k equivalence classes, often denoted $\tilde{0}$, $\tilde{1}$, $\ldots$, $\widetilde{k-1}$. Thus the equivalence class $\tilde{0}$ contains all the multiples of k:

$$\ldots, -2k, -k, 0, k, 2k, \ldots$$

A key fact about an equivalence relation on a set S is that it partitions S into non-overlapping *equivalence classes*.

Definition B.2.3. A *partition* of a set S is a collection of non-overlapping subsets S_i, called *equivalence classes*, whose union is S. Thus for any two i and j in I, the intersection $S_i \cap S_j$ is empty, and the union $\cup_{i \in I} S_i = S$.

Proposition B.2.4. *A partition $\{S_i, i \in I\}$ defines an equivalence relation P on $S \times S$, whose domain and range is all of S: sPt if s and t are in the same subset S_i. Conversely any equivalence relation R defines a partition of S, where each equivalence class S_s consists of the elements $t \in S$ that are equivalent to a given element s.*

Proof. It is easy to show that P satisfies the three properties of an equivalence relation. For the converse, just show that the sets S_s are either the same, or disjoint. Their union is obviously S. □

B.3 Algorithms and Methods of Proof

Algorithms.

Proof by contradiction.

Proof by induction.

B.4 Dual Maps

Suppose we have sets S and T, and a fixed map $L\colon S \to T$ between them. Now take another set R, and consider the set $Mor(S,R)$ of all maps $f\colon S \to R$, and the set $Mor(T,R)$ of all maps $g\colon T \to R$. For simplicity of notation, since R is fixed, we write S^* for $Mor(S,R)$, and T^* for $Mor(T,R)$.

Then we get a map $L^*\colon T^* \to S^*$ which associates to a map $g\colon T \to R$ the composite map $g \circ L\colon S \to R$.

Need a diagram here.

$$\begin{array}{ccccc} S & \xrightarrow{L} & T & \xrightarrow{g} & R \\ S^* & \xleftarrow{L^*} & T^* & & \\ g\circ L \in S^* & \xleftarrow{L^*} & g \in T^* & & \end{array}$$

Furthermore if we have a second fixed map $M\colon T \to U$, and we also consider the set $U^* = Mor(U,R)$. So we can form $M^*\colon U^* \to T^*$ as before. Next we consider the composite $M \circ L\colon S \to U$, and we claim:

$$(M \circ L)^* = L^* \circ M^*.$$

Another diagram here.

$$\begin{array}{ccccccc} S & \xrightarrow{L} & T & \xrightarrow{M} & U & \xrightarrow{h} & R \\ S^* & \xleftarrow{L^*} & T^* & \xleftarrow{M^*} & U^* & & \\ h\circ M\circ L \in S^* & \xleftarrow{L^*} & h\circ M \in T^* & \xleftarrow{M^*} & h \in U^* & & \end{array}$$

We will use this material in §6.6, in the linear algebra situation where the sets are vector spaces and the maps linear maps. It is useful to see that the results are true in the most general situation possible.

B.5 Review of Complex Numbers

We now review some properties of complex numbers. Recall that every complex number α can be written uniquely as $a+b\mathrm{i}$, where a and b are real numbers and i is the imaginary number whose square is -1. If $b=0$ then α is real, and if $a=0$ α is called imaginary. You know how to add and multiply complex numbers: if $\beta=c+d\mathrm{i}$ is a second complex number, then

$$\alpha+\beta=(a+c)+(b+d)\mathrm{i} \text{ and } \alpha\beta=(ac-bd)+(ad+bc)\mathrm{i}. \tag{B.1}$$

Remark B.5.1. If α and β are both real, then their sum and product as complex numbers are the same as their sum and product as real numbers. Check this.

Every complex number $a+b\mathrm{i}$ has the additive inverse $-a-b\mathrm{i}$, meaning that their sum gives 0. Every complex number $a+b\mathrm{i}$ other than 0 has a multiplicative inverse, meaning a number $c+d\mathrm{i}$ such that

$$(a+b\mathrm{i})(c+d\mathrm{i})=1.$$

Finding the multiplicative inverse is a good exercise in solving two linear equations in two variables. Indeed, if the inverse of $a+b\mathrm{i}$ is $x+y\mathrm{i}$, then we have the two equations (for the real and imaginary parts):

$$\begin{aligned} ax-by&=1\\ bx+ay&=0 \end{aligned}$$

Now a and b are not both 0, so we can eliminate x by multiplying the first equation by b, the second by a, and subtracting, We get $y=-b/(a^2+b^2)$, so by substituting this value into the other equation we get $x=a/(a^2+b^2)$. We write the inverse of α as α^{-1}.

It is very useful to plot complex numbers in the plane. As you know, it is traditional to plot the real part along the horizontal axis and the imaginary part along the vertical axis, with the numbers 1 and i as units.

The number $r=\sqrt{a^2+b^2}$ is called the modulus of the complex number $a+b\mathrm{i}$. Notice that in the coordinate system we just set up, it is the distance of the complex number from the origin.

Exercise B.5.2. Show that the modulus of the multiplicative inverse of a non-zero complex number α is $1/r$, if r is the modulus of α.

We now define the complex conjugate of a complex number $\alpha=a+b\mathrm{i}$: it is $\overline{\alpha}=a-b\mathrm{i}$. So the complex conjugate of a real number is itself. Note that the product of a complex number with its conjugate is

$$\alpha\overline{\alpha}=a^2+b^2. \tag{B.2}$$

In particular this product is real, which motivates many of the definitions made in linear algebra. The multiplicative inverse of the non-zero α can be written

$$\frac{\overline{\alpha}}{a^2+b^2}$$

confirming what we found above.

Exercise B.5.3. Writing the complex numbers α and β in terms of their real and imaginary parts, show by direct computation that:

1. The complex conjugate of a sum is the sum of the complex conjugates:

$$\overline{\alpha+\beta} = \overline{\alpha}+\overline{\beta}.$$

2. The complex conjugate of a product is the product of the complex conjugates:

$$\overline{\alpha\beta} = \overline{\alpha}\,\overline{\beta}. \tag{B.3}$$

We can write a non-zero complex number $\alpha = a+b\mathrm{i}$ using the complex exponential function e^z as $\alpha = r\mathrm{e}^{\mathrm{i}\theta}$. Here $\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta$, where θ is the angle whose tangent is b/a. The angle θ is called its argument. The advantage of this representation is that the multiplicative inverse of $r\mathrm{e}^{\mathrm{i}\theta}$, when $r \neq 0$, is

$$\frac{1}{r}\mathrm{e}^{-\mathrm{i}\theta}.$$

Exercise B.5.4. Verify all these statements. For the last one you need the trigonometric identities called the addition formulas:

$$\begin{aligned}\cos(\theta_1+\theta_2) &= \cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2\\ \sin(\theta_1+\theta_2) &= \sin\theta_1\cos\theta_2 + \cos\theta_1\sin\theta_2\end{aligned}$$

Every polynomial over $\mathbb{C}$ factors as a product of linear polynomials. We state this result, known as the Fundamental Theorem of Algebra, in §C.6.

Appendix C
Polynomials

Abstract This appendix present the basic results on polynomials in one variable over a field F necessary for finding the invariants factors of a linear operator. Many readers will have already seen these elementary results.

C.1 Polynomials: Definitions

As always, F is a field, either $\mathbb{R}$ or $\mathbb{C}$. A polynomial $f(x)$ over F is an expression of the form

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0, \tag{C.1}$$

where the a_i, called the coefficients of the polynomial are in the field F, and x is a variable, also called an unknown. The x^i are powers of x. For each i the expression $a_i x^i$ is called a term of the polynomial, or a monomial.

So $f(x) = 2x^2 - 1/5$ is a polynomial over $\mathbb{R}$ and even over the rational numbers $\mathbb{Q}$.

The largest m such that $a_m \neq 0$ is called the degree of the polynomial, $a_m x^m$ is called the leading term, and a_m the leading coefficient. If the leading coefficient is 1, the polynomial is *monic*. There is one exception to this definition of the degree: if all the coefficients of $f(x)$ are 0, then we say its degree is $-\infty$. There is only one such polynomial, the polynomial 0. Polynomials of degree 0 are of the form c, where c is a non-zero element of F.

We can multiply a polynomial $f(x)$ by a constant $c \in F$:

$$cf(x) = ca_n x^n + ca_{n-1}x^{n-1} + \cdots + ca_1 x + ca_0.$$

Unless $c = 0$, $cf(x)$ has the same degree as $f(x)$.

We can add two polynomials: the polynomial $f(x)$ and the polynomial

$$g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x^1 + b_0 \tag{C.2}$$

by just adding the coefficients. If a term does not appear in one of the polynomials, we just take its coefficients of that term to be 0. Then, if $n \geq m$,

$$f(x)+g(x) = (a_n+b_n)x^n + \cdots +$$

What is the degree of $f(x)+g(x)$? If the degree of $f(x)$ and $g(x)$ are different, then the degree of the sum is just the maximum of the degrees.

However, if the degrees are the same, then the degree of $f(x)+g(x)$ is can be smaller than the common degree n of $f(x)$ and $g(x)$: namely when if $a_n+b_n = 0$.

Note that these definitions imply that the collection of all polynomials over F, which we write $F[x]$, is a vector space over F. However it is not finite-dimensional. It is clear that the polynomials x^i, $0 \leq i$, form a basis for $F[x]$.

On the other hand, for any integer n, the collection of polynomials of degree at most n forms a vector space of dimension $n+1$.

The most interesting fact about $K[x]$ is that it also has a multiplication. The product of the polynomials f and g given in (C.1) and (C.2), of degrees n and m respectively, is a polynomial

$$h(x) = c_p x^p + c_{p-1}x^{p-1} + \cdots + c_1 x + c_0,$$

where the coefficients are given by the following formula:

$$c_k = a_0 b_k + a_1 b_{k-1} + a_2 b_{k-2} + \cdots + a_k b_0. \tag{C.3}$$

This is sometimes called the convolution product of the vector $a_n\ a_{n-1}\ \ldots\ a_0$ by the vector $b_m\ b_{m-1}\ \ldots\ b_0$ From this formula it is easy to see that h has degree $p = n+m$ and that the coefficient of the leading term is $a_n b_m$

Example C.1.1. Numerical example of the product of degree 2 by degree 3.

Exercise C.1.2. Show that $F[t]$ satisfies all the axioms of a field, except that it is not true that all elements other than 0 have a multiplicative inverse.

Exercise C.1.3. Show that the only polynomials $f(t)$ as in (C.1) that have a multiplicative inverse have $a_0 \neq 0$. The next exercise shows that this is a necessary, but not a sufficient condition.

Exercise C.1.4. Show that the polynomial $t+1$ does not have a multiplicative inverse in $F[t]$ by assuming it has an inverse and then solving for a_i for increasing values of i.

Exercise C.1.5. Instead of taking polynomials, we could instead take for the scalars the integers, which we denote $\mathbb{N}$. Show that they fail being a field in exactly the same way as polynomials: elements other than 0 fail to have multiplicative inverses in the integers. We can enlarge the integers to the rational numbers $\mathbb{Q}$ to remedy this problem, so $\mathbb{Q}$ is a field, contained in the real numbers, with is itself contained in the complex numbers.

C.2 The Euclidean Algorithm

The Euclidean algorithm is one of the most famous algorithms of mathematics. We need it for polynomials; you have certainly seen it not only for polynomials but also for integers. Its goal is to find the greatest common divisor of two polynomials $f(x)$ and $g(x)$. This is a polynomial of largest degree dividing both $f(x)$ and $g(x)$. We do this by first explaining how to do long division of a polynomial $f(x)$ by a polynomial $g(x)$ of degree at most that of $f(x)$. The condition on the degree of $g(x)$ in the statement only excludes the polynomial $g(x)$ that is identically 0, to avoid division by 0.

Theorem C.2.1 (Euclidean Algorithm). *Given a polynomial $f(x)$ and a polynomial $g(x)$ of degree at least 0, there are polynomials $q(x)$ and $r(x)$ such that*

$$f(x) = q(x)g(x) + r(x) \text{ with } \deg r(x) < \deg g(x). \tag{C.4}$$

The polynomial $q(x)$ is the quotient and $r(x)$ is the remainder.

Proof. We do this by finding the coefficients of one power of x for both of both $q(x)$ and $r(x)$ at a time. Start with the polynomials $f(x)$ and $g(x)$ of (C.1) and (C.2), with $n \geq m$. Then we form the polynomial $f_1(x)$ given by

$$f_1(x) = f(x) - \frac{a_n}{b_m} x^{n-m} g(x).$$

By construction the degree n_1 of f_1 is less than the degree of f. If n_1 is still greater than or equal to m, we form

$$f_2(x) = f_1 - \frac{a^1_{n_1}}{b_m} x^{n_1-m} g(x),$$

where $a^1_{n_1}$ is the leading coefficient of f_1. We repeat this process until the degree n_k of $f_k(x)$ is less than m. So we have, for suitable scalars c_1

$$\begin{aligned} f_1 &= f - c_0 x^{n-m} g, \\ f_2 &= f_1 - c_1 x^{n_1-m} g \\ f_3 &= f_2 - c_2 x^{n_2-m} g \\ \dots &= \dots \\ f_k &= f_{k-1} - c_{k-1} x^{n_{k-1}-m} g \end{aligned}$$

Then backsubstituting into the last equation, we get

$$f_k(x) = f - \left(c_0 x^{n-m} + c_1 x^{n_1-m} + c_2 x^{n_2-m} + \cdots + c_{k-1} x^{n_{k-1}-m}\right) g(x).$$

Write $q(x)$ for polynomial multiplying g, and $r(x)$ for $f_k(x)$, we have proved that

$$f(x) = q(x)g(x) + r(x)$$

for a polynomial r of degree less than that of g. □

This process is algorithmic, meaning that what we do at each step is uniquely determined. Still, the proof that the conclusion of the theorem is uniquely defined is interesting.

Theorem C.2.2. *With the hypotheses of the Euclidean algorithm, the polynomials $q(x)$ and $r(x)$ are uniquely defined.*

Proof. Suppose we have a second representation $f(x) = q_1(x)g(x) + r_1(x)$, where the degree of $r_1(x)$ is still less than the degree of g. Taking the difference of the two representations, we get

$$\big(r_1(x) - r(x)\big) = \big(q(x) - q_1(x)\big)g(x).$$

The polynomial on the right hand side has degree at least m unless $q(x) = q_1(x)$, in which case we get the 0 polynomial. The left hand side has degree at most $m-1$ since it is the difference of two remainders. So the only possibility is $q(x) = q_1(x)$. Then, since the right hand side is 0, $r_1(x) = r(x)$. □

Definition C.2.3. If $r(x)$ is the zero polynomial, we say $g(x)$ divides $f(x)$.

Remark C.2.4. If you do not care about how to compute $q(x)$ and $r(x)$, it is easy to produce an existence proof of the representation (C.4). We know that we can always write $f(x) = q(x)g(x) + r(x)$ for some polynomials $q(x)$ and $r(x)$: for example take $q(x) = 0$ and $r(x) = f(x)$. We need to show that we can find such a representation with $\deg r(x) < \deg g(x)$. So take the representation $f(x) = q(x)g(x) + r(x)$ with $r(x)$ of smallest degree, and assume that $\deg r(x) \geq \deg g(x)$. Let the leading term of $r(x)$ be $r_{m+e}x^{m+e}$. Then adding to $q(x)$ the term $\frac{r_{m+e}}{b_m}x^e$ decreases the degree of $r(x)$, a contradiction. Notice that this is just one step of the Euclidean algorithm.

Exercise C.2.5. Write down long division of integers, represented in base 10, in the high school way.

Exercise C.2.6. Replacing the powers of 10 by powers of x, make the same representation of the Euclidean algorithm for polynomials, and show that it is exactly what we did above.

C.3 Roots of Polynomials

Definition C.3.1. Let $f(x)$ be the polynomial of (C.1). Substitute $a \in K$ in for the variable x: if

$$f(a) = a_n a^n + a_{n-1}a^{n-1} + \cdots + a_1 a + a_0 = 0$$

then a is a *root* of $f(x)$.

Now apply the Euclidean algorithm to f and the polynomial $x-a$, a polynomial of degree 1. Then

$$f(x) = q(x)(x-a)+r,$$

where r is a constant. If a is a root of f, evaluating the expression at $x=a$ shows that $r=0$, so $x-a$ divides $f(x)$.

So we have proved

Theorem C.3.2. *If a is a root of the polynomial $f(x)$ of degree n, then there is a polynomial $q(x)$ of degree $n-1$ such that $f(x) = q(x)(x-a)$. Thus $x-a$ divides $f(x)$.*

Corollary C.3.3. *A polynomial of degree n has at most n distinct roots.*

Proof. Otherwise f would have a number of factors greater than its degree, which is impossible. □

The number of roots depends on the field K. For example the polynomial x^2+1 has no roots over $\mathbb{R}$, while it has roots i and $-$i over $\mathbb{C}$.

Explain that every polynomial in $\mathbb{C}$ factors as a product of linear factors. Put proof in appendix?

Explain that every polynomial in $\mathbb{R}$ factors as a product of linear and quadratic factors. Which are the quadratics that cannot be factored over $\mathbb{R}$? Quadratic formula

Suppose we have a polynomial of degree exactly n, written as in (C.1). Then we can divide by the leading coefficient a_n to make the polynomial monic. We call it $f(x)$, which is therefore written

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0.$$

Also assume that $f(x)$ factors as a product of linear factors, so it has n roots

$$f(x) = (x-u_1)(x-u_2)\ldots(x-u_n).$$

As we just noted, this can always be done over $\mathbb{C}$. Then we can ask: what is the relationship between the a_i and the u_i, $1 \le i \le n$?

To express the answer we need to define the elementary symmetric functions of the a_i.

Definition C.3.4. The n elementary symmetric functions s_i of the n quantities u_i, $1 \le i \le n$ are

$$\begin{aligned}
s_1 &= u_1 + u_2 + \cdots + u_n;\\
s_2 &= u_1u_2 + u_1u_3 + \cdots + u_1u_n + u_2u_3 + \cdots + u_2u_n + \ldots u_{n-1}u_n;\\
\ldots &= \ldots;\\
s_k &= \sum_{1\le i_1<i_2<\cdots<i_k\le n} u_{i_1}u_{1_2}\ldots u_{i_k};\\
\ldots &= \ldots;\\
s_n &= u_1u_2\ldots u_n.
\end{aligned}$$

These functions are called symmetric functions, because given any permutation σ of $[1,\dots,n]$ acting by permuting the indices, so that for example

$$\sigma(s_1) = u_{\sigma(1)} + u_{\sigma(2)} + \cdots + u_{\sigma(n)}.$$

then $\sigma(s_i) = s_i$.

The classic result about polynomials is

Theorem C.3.5. *If the polynomial $f(x) = (x-u_1)(x-u_2)\dots(x-u_n)$, then it can be written*

$$f(x) = x^n - s_1 x^{n-1} + s_2 x^{n-2} - \cdots \pm s_n$$

where the s_i are the elementary symmetric functions of the roots u_i.

Proof. This is easily proved by induction on n. For $n = 1$ it is obvious. Assume it is true for $n-1$ and write the symmetric functions of the $n-1$ roots of the polynomial $g(x)$ as $s_1, \dots, s_{n-1}$, so that we may assume

$$g(x) = x^{n-1} - s_1 x^{n-2} + s_2 x^{n-3} - \cdots \pm s_{n-1}$$

Now multiply $g(x)$ by $(x-u_n)$ and write the product $f(x)$. Let t_i be the elementary symmetric functions of the n roots $u_1, \dots, u_n$. So $g(x)(x-u_n)$ can be written

$$\begin{aligned} f(x) = g(x)(x-u_n) = x^n - s_1 x^{n-1} + s_2 x^{n-2} - \cdots \pm s_{n-1} x \\ - u_n x^{n-1} + u_n s_1 x^{n-2} - u_n s_2 x^{n-3} - \cdots \mp u_n s_{n-1} \end{aligned} \quad \text{(C.5)}$$

So the coefficients of $f(x)$ are, up to alternation of sign, $t_1 = s_1 + u_n$, $t_2 = s_2 + u_n s_1$, $\dots$, $t_{n-1} = s_{n-1} + u_n s_n$, $t_n = u_n s_n - 1$. These are easily seen to be the symmetric functions in the n roots, so we are done. □

We will use this result when we study the characteristic polynomial in Chapter 12.

C.4 Great Common Divisors

As always, let $f(x)$ and $g(x)$ be two polynomials, where $n \geq m$. First we recall the definition of the greatest common divisor of $f(x)$ and $g(x)$.

Definition C.4.1. A polynomial $d(x)$ is the greatest common divisor (*gcd*) of $f(x)$ and $g(x)$ if it divides both of them, and if any other divisor of $f(x)$ and $g(x)$ divides $d(x)$.

It is not obvious that the *gcd* exists. We can make it unique by requiring that it also be a monic polynomial, so its leading coefficient is 1. We construct the gcd using the Euclidean algorithm. Do long division to obtain

$$f = q_0 g + r_0 \qquad (0)$$

where $\deg r_0 < \deg g$. Now divide g by r_0 to get

$$g = q_1 r_0 + r_1 \qquad (1)$$

where $\deg r_1 < \deg r_0$. Continue in this way, so that $\deg r_{i+1} < \deg r_i$:

$$\begin{aligned} r_0 &= q_2 r_1 + r_2 && (2)\\ r_1 &= q_3 r_2 + r_3 && (3)\\ \ldots &= \ldots \\ r_{k-3} &= q_{k-1} r_{k-2} + r_{k-1} && (k)\\ r_{k-2} &= q_k r_{k-1} + r_k && (k+1)\\ r_{k-1} &= q_{k+1} r_k + r_{k+1} && (k+2) \end{aligned}$$

until you get to a remainder (say r_{k+1}) that is the zero polynomial 0. We must always get to 0 since the degrees of the r_i are strictly decreasing. Thus equation (k+2) is actually: $r_{k-1} = q_{k+1} r_k$, so that r_k divides r_{k-1}. Using equation (k+1) we can express r_k as a linear combination of two terms with lower indices:

$$r_k = r_{k-2} - q_k r_{k-1}, \qquad (C.6)$$

so r_k divides r_{k-2}. Now that we have established that r_k divides two previous r_i with consecutive indices, our list of equation shows it divides all of them, and also $g(x)$ and $f(x)$. So r_k is a common divisor of the two polynomials we started with.

Next we establish it is the greatest common divisor. We do this by first establishing (C.9) below.

Equation (k), written as

$$r_{k-1} = r_{k-3} - q_{k-1} r_{k-2} \qquad (C.7)$$

shows that we can rewrite (C.6) as

$$r_k = r_{k-2} - q_k\big(r_{k-3} - q_{k-1} r_{k-2}\big) = (1 + q_k q_{k-1}) r_{k-2} - q_k r_{k-3}. \qquad (C.8)$$

Equation (k-1) allows the elimination of r_{k-2}, and at each step we can write r_k as a linear combination of two previous remainders with consecutive and lower indices. Continuing in this way, we see that r_k can be written as a linear combination of f and g.

$$r_k(x) = c_1(x) f(x) + c_2(x) g(x) \qquad (C.9)$$

for certain polynomials $c_1(x)$ and $c_2(x)$ that can be computed algorithmically from the q_i..

Theorem C.4.2. *The polynomial $r_k(x)$ is the greatest common divisor of $f(x)$ and $g(x)$. Let $c = \deg r_k(x)$. Then there is no common divisor of $f(x)$ and $g(x)$ of degree*

greater than c, and the only common divisors of degree c are scalar multiples of $r_k(x)$.

Proof. Let $h(x)$ be a common divisor of $f(x)$ and $g(x)$, so there are polynomials $d_1(x)$ and $d_2(x)$ such that $f(x) = d_1(x)h(x)$ and $g(x) = d_2(x)h(x)$. Substitute these two expression into (C.9), getting

$$r_k(x) = c_1(x)d_1(x)h(x) + c_2(x)d_2(x)h(x) = \big(c_1(x)d_1(x) + c_2(x)d_2(x)\big)h(x)$$

which shows that $h(x)$ divides $r_k(x)$. So $\deg h(x) \leq \deg r_k(x)$, with equality only if $h(x)$ is a scalar multiple of $r_k(x)$. □

Definition C.4.3. Two polynomials whose greatest common divisor is 1 are *relatively prime*.

Corollary C.4.4. *If two polynomials $f(x)$ and $g(x)$ are relatively prime, then there are polynomials $c_1(x)$ and $c_2(x)$ such that $1 = c_1(x)f(x) + c_2(x)g(x)$. The polynomials can be chosen so that $\deg c_1(x) < \deg g(x)$ and $\deg c_2(x) < \deg f(x)$, and then the representation is unique.*

Proof. The first statement is just (C.9). For the next statement, first note that $\deg(c_1(x)f(x)) = \deg(c_2(x)g(x))$, since the lead terms of both must agree to produce a polynomial of degree 0 on the left hand side. If $\deg c_1(x) > \deg g(x)$, just do long division: $c_1(x) = q_1(x)g(x) + r_1(x)$. By our first remark, then $\deg c_2(x) > \deg f(x)$, so also do long division : $c_2(x) = q_2(x)f(x) + r_2(x)$. Then we have

$$\begin{aligned} 1 &= \big(q_1(x)g(x) + r_1(x)\big)f(x) + \big(q_2(x)f(x) + r_2(x)\big)g(x) \\ &= \big(q_1(x) + q_2(x)\big)f(x)g(x) + r_1(x)\big)f(x) + r_2(x)\big)g(x) \end{aligned}$$

Then, since the last two terms have lower degree than the first term, the leading coefficient of $q_1(x)$ and $q_2(x)$ must cancel. We can continue the argument until we get to the situation where $\deg c_1(x) < \deg g(x)$ and $\deg c_2(x) < \deg f(x)$, as claimed. Now suppose there are two expressions with the same restriction on the degrees of $d_1(x)$ and $d_2(x)$:

$$\begin{aligned} 1 &= c_1(x)f(x) + c_2(x)g(x) \\ 1 &= d_1(x)f(x) + d_2(x)g(x) \end{aligned}$$

Subtract to get

$$0 = \big(c_1(x) - d_1(x)\big)f(x) + \big(c_2(x) - d_2(x)\big)g(x)$$

Because f and g are relatively prime, f must divide $c_2(x) - d_2(x)$ and g must divide $c_1(x) - d_1(x)$. But the degrees are too small, so $c_1(x) - d_1(x)$ and $c_2(x) = d_2(x)$. We are done. □

We can generalize this to

Theorem C.4.5. *Given k polynomials $f_1, \ldots, f_k$, they have a greatest common divisor $r(x)$, and $r(x)$ can be expressed as a linear combination of the f_i:*

$$r(x) = c_1(x)f_1(x) + \cdots + c_k(x)f_k(x)$$

Proof. We do this by first replacing f_1 and f_2 by their greatest common divisor d_1, and then working with the collection $d_1(x), f_3(x), \ldots, f_k(x)$, and using the fact that d_1 can be written as a linear combinations of f_1 and f_2. Then work with d_1 and f_3 to get their greatest common divisor $d_2(x)$, and continue in this way. □

Remark C.4.6. If you are familiar with the notion of ideal in a ring, that the results of this section can be obtained more quickly. Still the technique of proof given above is algorithmic, and allows the computation of the *gcd*.

C.5 Unique Factorization

Definition C.5.1. A polynomial over a field F is irreducible if its only divisors (polynomials in $K[x]$ are 1 and itself.

Irreducible polynomials are the analog of prime numbers for the factorization of integers.

It is important to understand that the notion of irreducibility depends on the field. For example there are different irreducible polynomials in $\mathbb{R}[x]$ than in $\mathbb{C}[x]$.

The key result that gives us unique factorization is:

Theorem C.5.2. *If an irreducible polynomial $p(x)$ divides the product of two polynomials $f(x)g(x)$, then it actually divides one of the two.*

Proof. Our hypothesis is that $f(x)g(x) = p(x)h(x)$ for some polynomial $h(x)$. Assume that $p(x)$ does not divide $f(x)$. Because $p(x)$ is irreducible, this forces the greatest common divisor of $p(x)$ and $f(x)$ to be 1. So by Corollary C.4.4 there are polynomials $c_1(x)$ and $c_2(x)$ that give

$$1 = c_1(x)f(x) + c_2(x)p(x).$$

Multiply this expression by $g(x)$:

$$\begin{aligned} g(x) &= c_1(x)f(x)g(x) + c_2(x)p(x)g(x) \\ &= c_1(x)p(x)h(x) + c_2(x)p(x)g(x) = p(x)\big(c_1(x)h(x) + c_2(x)g(x)\big) \end{aligned}$$

by our hypothesis. This says that $p(x)$ divides $g(x)$, so we are done. □

The main theorem is

Theorem C.5.3. *Any polynomial $f(x)$ in $K[x]$ can be factored as a product of irreducible polynomials $p_i \in K[x]$, each raised to a positive power m_i:*

$$f(x) = p_1(x)^{m_1} p_2(x)^{m_2} \dots p_k(x)^{m_k} = \prod_{i=1}^{k} p_i(x)^{m_i}$$

and the irreducible polynomials p_i are uniquely defined up to a constant factor and up to order.

Write the proof.

Definition C.5.4. A polynomial $f(x) = p(x)^m$, where $p(x)$ is irreducible, and m is an integer greater than or equal to 2, is called *primary*. When the irreducible polynomial $p(x)$ needs to be made explicit, $f(x)$ is called p-primary.

Example here about what uniqueness means.

C.6 The Fundamental Theorem of Algebra

Theorem C.6.1. *Any polynomial with coefficients in $\mathbb{C}$ factors over $\mathbb{C}$ as a product of polynomials of degree one (linear polynomials).*

We also have

Theorem C.6.2. *Any polynomial with coefficients in $\mathbb{R}$ factors over $\mathbb{R}$ as a product of irreducible polynomials of degree at most 2. The irreducible factors of degree 2 factor over $\mathbb{C}$ as $(x-\alpha)(x-\overline{\alpha})$, where α is a complex number that is not real, and $\overline{\alpha}$ is its complex conjugate.*

It would take us too far afield to prove the first theorem. Using it, it is not hard to prove the second one.

Proof. Indeed take a polynomial

$$f(x) = c_m x^m + c_{m-1} x^{m-1} + \dots + c_1 x^1 + c_0$$

with real coefficients c_i. By the Fundamental Theorem of Algebra, $f(x)$ factors in linear factors over $\mathbb{C}$. Assume that a complex number $\alpha = a + b\mathrm{i}$, $b \neq 0$ is a root. If there is no such root, then all the roots are real. We first show that the complex conjugate $\overline{\alpha} = a - b\mathrm{i}$ is also a root.

By (B.3), if $\alpha^k = a_k + b_k\mathrm{i}$, then $\overline{\alpha^k} = a_k - b_k\mathrm{i}$. Then, since the c_i are real and α is a root, the imaginary part of $f(\alpha)$ must vanish. So

$$c_m b_m + c_{m-1} b_{m-1} + \dots + c_1 b_1 + c_0 = 0.$$

Since this is also the imaginary part of $f(\overline{\alpha})$, $\overline{\alpha}$ is also a root. Therefore the product $(x-\alpha)(x-\overline{\alpha})$ divides $f(x)$. Now if $\alpha = a + b\mathrm{i}$,

$$(x-\alpha)(x-\overline{\alpha}) = x^2 - 2ax + a^2 + b^2,$$

a real polynomial that does not factor over $\mathbb{R}$ since its roots are not real. Thus it is irreducible. □

Exercise C.6.3. Use the quadratic formula to confirm that the roots of $x^2 - 2ax + a^2 + b^2$ are $a \pm \sqrt{-b^2} = a \pm b\mathrm{i}$.

Example of different factorizations over $\mathbb{R}$ and $\mathbb{C}$.

Appendix D
Matrices, Spreadsheets and Computer Systems

Abstract Avoiding hand computations. These notes first describe how matrix operations can be implemented in spreadsheets, especially Excel. In a second section they show how to do certain linear algebra computations in MatLab.

D.1 Matrices and Spreadsheets

We describe how matrix operations can be implemented in spreadsheets. I use Excel. These notes are probably too concise unless you already know the basics of Excel. I illustrate this using the Macintosh version of Excel. There may be changes required for other platforms.

In the world of spreadsheets the word matrix is never used. One only speaks of *arrays*, which means the same thing. The indexing of rows and columns is reversed: the columns, which are enumerated by letters, comes first, while the rows, enumerated by integers, come second. The entries of the array are called *cells*. So D3 represents the cell in the D-th column and the third row.

Example D.1.1. Here is the left-top corner of a spreadsheet, with a 3×4 array in it.

	A	*B*	*C*	*D*	*E*
1	1	2	4	−8	
2	4	5	6	−1	
3	7	8	9	2	
4					

So cell A1 contains the number 1, while cell C2 contains the number 6.

D.1.1 Row Operations

The easiest matrix operations to do in spreadsheets are elementary row operations, since no special Excel operations are required. They are less powerful and require more cutting and pasting than the Excel matrix functions of the next section.

Row swap On the example above, suppose you want to swap rows 2 and 3 of the matrix. First just copy row 2 of the array to somewhere else, for example in row 4 of the spreadsheet. Then paste row 3 where row 2 used to be. Finally copy the moved row 2 into row 3.

Subtract a multiple of a row from another row

Suppose we want to subtract 4 times row 1 from row 4 in the example. First make a copy of the matrix, and paste it, for example going from A5 to D7. Then in cell A6 write the expression

```
= A2 - $A$2 * A1
```

This formula means put into cell A6 the content of A2 minus the content of A2 times the content of A1. Then press Enter. A 0 should appear in cell A6. Now use the mouse to select all the array entries in row 6 of the spreadsheet. Go to the menu Edit and select the item Fill, at which point a submenu will appear: pick Right. All the entries of the matrix will be converted correctly: in cell B6 you will get

```
B2 - $A$2 * B1
```

The reason A2 did not get converted to B2 in this new formula, like the other two cells references, is that the $ signs make the cell reference into an absolute reference rather than a reference relative to the cell that is being considered.

Divide a row by a non-zero scalar After the row operation just performed we have the matrix

	A	*B*	*C*	*D*	*E*
7					
8	1	2	4	−8	
9	0	−3	−10	31	
10	7	8	9	2	
11					

We might want to divide the second row by -3 to make the next pivot 1. As before copy the array again, but this time do a Paste Special, choosing Values: you do not want the formulas in the cells, just the values. Imagine the array goes from A17 to D19, for example. Then in cell B18 type

```
=  B9 /$B$9.
```

After pressing Enter, a 1 should appear in cell B18. Then select the element of the array in row 18 from B18 to D18. Then do a Fill Right as before. You are done.

Continuing in this way, you can put a matrix in upper triangular or RREF form easily.

Example D.1.2. You should compute the RREF of the augmented matrix in the example above. Once you master the spreadsheet tools in the nest section, you can conclude that the the matrix of the first three columns is the identity matrix, because it is invertible, and the last column is the vector

$$\begin{bmatrix} 2 \\ 3 \\ -4 \end{bmatrix}$$

Why?

D.1.2 Matrix Algebra

Now we turn to the built-in matrix functions that even everyday users of Excel may not be familiar with.

Giving names

The first thing to do is to name an array in the spreadsheet. Type in numerical values into the cells of the array, then select the entire array. Go to the Insert menu, and select the item Name. A submenu appears: check Define. Then a dialog box called Define Name appears, and just type in a name, say mfm. The computer then tells you that you have assigned the name mfm to:

```
=Sheet1!$A$1:$C$3
```

Note how the array is specified by its upper left hand corner

```
$A$1
```

and its lower right hand corner

```
$C$3
```

separated by a colon indicating the range. The dollar signs indicate, as before, that this is an absolute reference. A common mistake is to use a name that could be a cell reference: letters followed by digits. So for example you would get an error message if you named your array mfm2.

Addition

To add two matrices of size 3×3 on the same sheet of the spreadsheet, now create a second matrix, call it msm, of the same size. Then select an empty 3×3 array in the spreadsheet, type in

```
=mfm+msm
```

and then hit command-shift-return simultaneously. The computer then returns the sum of the matrices in one operation. Common mistakes are to select a region of the wrong size or to forget to hit command-shift-return simultaneously

Scalar Multiplication Suppose you want to multiply mfm by 10, say. Choose an array of the size of mfm, type

```
=10*mfm
```

and then hit command-shift-enter simultaneously.

Matrix multiplication Follow the same steps as for addition, but this time let the first array, call it lefta, be a $n \times m$ array, and the second one, called righta be a $m \times p$ array. In other words the number of columns of lefta must be the same of the number of rows of righta, so that multiplication is possible. Next select an unused array in the spreadsheet of size $n \times p$, select it, and type

```
=MMULT(lefta, righta)
```

Then hit command-shift-return simultaneously.

Transposition After entering and naming a first matrix mfm of size $n \times m$, pick a region of the size $m \times n$, and type in

```
=TRANSPOSE(mfm)
```

and as always hit command-shift-return simultaneously. Here is another way of doing this, known as swapping rows and columns: 1) Select the array of size $n \times m$ called mfm in the spreadsheet and Copy. 2) Go to a free area of the spreadsheet of size $m \times n$, go to the Edit Menu, choose the paste special item and click on the box marked Transpose. The transposed matrix will appear.

Determinant Start with a square matrix, called mfm for instance. Pick a single cell and type

```
=MDETERM(mfm)
```

The determinant appears after you press command-shift-enter simultaneously. If it is 0, the matrix does not have an inverse, as we will learn.

Example D.1.3. Compute the determinant of the 3×3 coefficient matrix of Example D.1.1, meaning the array $A1 : C3$. You should get -3.

Matrix inverse Start with a square matrix, called mfm for instance. Then select a unused region of the same size as mfm in the spreadsheet and type in

```
=MINVERSE(mfm)
```

As always hit command-shift-enter simultaneously. The inverse will appear, unless the matrix has determinant 0 or close enough to 0 that Excel cannot do the computation.

Example D.1.4. Now compute the inverse of the 3×3 coefficient matrix of Example D.1.1. You should get

$$\begin{bmatrix} 1 & -4.666667 & 2.666667 \\ -2 & 6.333333 & -3.333333 \\ 1 & -2 & 1 \end{bmatrix}$$

The fractions are just the representation of numbers such at $2+2/3$ and the like. Why 3 in the denominator? As we will see later in the course, this is because the determinant is -3.

Solve a square system of equations If you have a $n \times n$ system of linear equations $A\mathbf{x} = \mathbf{b}$, you can find the unique solution to the system if the determinant is non-zero: first compute the inverse $B = A^{-1}$ of A using the MINVERSE command, and then compute the product $B\mathbf{b}$ using the MMULT command

```
=MMULT(B, b)
```

Notice that you only need to select an array of size $n \times 1$ to store the answer $\mathbf{b}$.

Example D.1.5. If you do this with the augmented matrix of Example D.1.1, using the inverse computed in the previous example, you get the column vector

$$\begin{bmatrix} 2 \\ 3 \\ -4 \end{bmatrix}$$

How to you check that this is correct? Just multiply the original coefficient matrix by this column vector using the MMULT command: you should get the right-most column of the original matrix, and you do.

LU form

Start out with a square matrix A of size n. We want to put it in LU form, meaning that we want to write $A = LU$, where L is a lower triangular matrix and U an upper triangular matrix. This is not always possible, because row operations may not give us the expected pivots. When it is, it is a way of replacing the row operations of §D.1.1 by matrix multiplications. Put the square matrix of size n you are interested in the cells in the upper-left hand corner of the matrix. For concreteness assume $n = 3$: Here is an example:

	A	*B*	*C*	*D*	*E*
1	2	1	1		
2	4	5	3		
3	6	9	9		
4					

Then enter the matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ -A2/A1 & 1 & 0 \\ -A2/A1 & 0 & 1 \end{bmatrix} \tag{D.1}$$

in cells A5 to C7, give it a name, and the multiply, using MMULT this matrix by the original matrix, putting the product in cells A8 to C10, for example. You get

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

Because the element in the second row and second column of this new matrix is not zero, it is the next pivot, ands multiply on the left by

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -B10/B9 & 1 \end{bmatrix} \tag{D.2}$$

You get the matrix

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

which is upper triangular, as required. The lower triangular matrix you want is the inverse of the product of the matrix (D.2) by the matrix (D.1). First we compute the product

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

and then take its inverse

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

Finally just check that this matrix times matrix (D.3) is the original matrix.

Diagonalization of symmetric matrices

Suppose the square matrix A you start with is symmetric: $A = A^T$. By what we just did we can find a matrix L such that LA is upper triangular. L is invertible and its transpose L^T is upper triangular. So the product LAL^T is symmetric. this matrix is both upper triangular and lower triangular, so it is symmetric. Thus we can just use the computation of §D.1.2 to diagonalize a symmetric matrix when all the pivots are non zero. An interesting question is: when does that occur?

D.2 Matrices in MatLab

Here is an involved example using MatLab. In these notes I work some examples of approximation of polynomials using orthogonal projections techniques.

First we establish a result called the interpolation of points by a polynomial. Take $n+1$ point in the plane, of the form (x_i, y_i), $0 \leq i \leq n$, where the x_i are distinct. Then these points lie on the graph of a unique polynomial $f(t)$ of degree at most n. We will discuss how the polynomials is determined.

Next, given the same $n+1$ points (x_i, y_i), we ask for the polynomial $g(t)$ of degree $n-1$ whose values $z_i = g(x_i)$ are as close as possible to values $y_i = f(x_i)$. By as close as possible we mean that

$$\sum_{i=0}^{n}(y_i - z_i)^2$$

is minimal over all all polynomials $g(t)$ of degree $n-1$. By the *Pythagorean* Theorem $g(t)$ is uniquely determined. Here we construct it. The key idea involves orthogonality. The (y_i) associated to polynomials of degree $\leq n$ form a vector space $\mathbb{R}^{n+1}$ with a natural basis, while the (z_i) associated to polynomials of degree $\leq n-1$ span a subspace H of dimension n. Using the standard inner product on $\mathbb{R}^{n+1}$, we can project the points to H. Since we have $n+1$ points in a vector space of dimension n, they are linearly dependent.

Then we make the same approximation, but using a different inner product, call the L^2 *inner product* on the vector space P_n of polynomials of degree n: For polynomials $f(x)$ and $g(x)$ it is defined by

$$\int_{-1}^{1} f(x)g(x)dx.$$

. In these notes we only integrate over the interval $[-1,1]$, but we could integrate over any finite interval of our choice.

An important purpose of these notices is to work an explicit example of both methods, explaining how to do the computations using MatLab.

A project for those who want to learn how to use MatLab is to work the exercise given in the last section.

D.2.1 Polynomials Passing Through Points

Suppose you are given $n+1$ points in the plane: (x_i, y_i), $0 \leq i \leq n$. We want to find a polynomial $f(t)$ of degree n passing through these points, where the x_i are distinct. If you write $f(t)$ as $c_n t^n + \ldots, c_1 t + c_0$, this means solving the system of linear equations of $n+1$ equations in $n+1$ variables:

$$A\mathbf{c} = \mathbf{y}, \tag{D.4}$$

where A is the $(n+1)\times(n+1)$ matrix with $a_{ij} = x_i^j$, $0 \leq i, j \leq n$. In other words the map that associates to a polynomial $f(x)$ of degree n the vector $[f(x_0, f(x_1), \ldots, f(x_n)] \in \mathbb{R}^{n+1}$ is a linear transformation T_A between two vector spaces of dimension $n+1$. The kernel of this transformation only contains the zero polynomial, since a non-trivial polynomial of degree $\leq n$ has at most n zeroes. Therefore T_A is an isomorphism. Notice that the first equation in (D.4) is

$$c_0 + c_1 x_0 + c_2 x_2^2 + \cdots + x_0^n = y_0.$$

The matrix A is known as the *Vandermonde matrix* at $x_0, x_1, \ldots, x_n$, so we have shown in a second way that it is invertible. We will show directly that is invertible if

and only if the values x_i are distinct, when we study determinants. When $n = 2$ the Vandermonde matrix is

$$\begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{pmatrix}$$

Since A is invertible by (D.4) we can solve for $\mathbf{c}$: $\mathbf{c} = A^{-1}\mathbf{y}$. In particular the coefficients c_i of the polynomial are uniquely determined by the points (x_i, y_i), showing that there is exactly one n-th degree polynomial going through $n+1$ points.

Example D.2.1. Here is an example that you would not want to do by hand. Take the four points $(-1,1)$, $(-1/2,-1)$, $(0,-3/2)$, $(1,0)$. The Vandermonde matrix for them is

$$A = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & -1/2 & 1/4 & -1/8 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

They are on the unique cubic

$$-\frac{2}{3}t^3 + 2t^2 + \frac{1}{6}t - \frac{3}{2}. \tag{D.5}$$

as I now prove using a MatLab computation.

As we have already noted, to find the coefficients of the polynomial just solve the linear system $A\mathbf{c} = \mathbf{y}$. Using MatLab, we compute the inverse of A:

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

Since $\mathbf{y}$ is the column vector $(1,-1,-3/2,-1)$ the coefficients $\mathbf{c}$ are

$$A^{-1}\mathbf{y} = (-3/2, 1/6, 2, -2/3).$$

Warning: MatLab has a `vander[x]` command that produces the Vandermonde matrix at the points $\mathbf{x}$, but it reverses the standard ordering of the columns: it goes from highest power to 0-th power. So in the example above, to get the Vandermonde matrix, first enter the vector of points at which you want to evaluate, and then use the function:

```
v = [-1 -1/2 0 1];
A = vander(v)
      -1.0000    1.0000   -1.0000    1.0000
      -0.1250    0.2500   -0.5000    1.0000
       0          0         0         1.0000
       1.0000    1.0000    1.0000    1.0000
```

To get the inverse of A:

```
Ainv = inv(A)
       -1.0000      2.6667     -2.0000      0.3333
        0.5000      0             -1.0000       0.5000
        0.5000     -2.6667       2.0000      0.1667
        0            0             1.0000       0
```

To get the column vector **y** just enter

```
y = [1;-1; -3/2; 0]
```

and then multiply to get

```
c= Ainv* y
 -0.6667
  2.0000
  0.1667
 -1.5000
```

remembering that the order of the coefficients is reversed. This confirms (D.5).

D.2.2 Orthogonal Projections

Next we want to find the polynomial of degree $n-1$ that passes as close as possible to the same $n+1$ points, in the least squares sense. How do you describe all polynomials of degree n_1, to be evaluated at $n+1$ points? Just take the first n columns of the matrix A. So D is a $n+1\times n$ matrix. Then the values of the polynomial of degree $n-1$ whose coefficients are $c_0,\ldots c_{n-1}$ at the points **x** are given by $D\mathbf{c}$. As the polynomial with coefficients **c** varies over all polynomials of degree $n-1$, its output varies in a subspace H of dimension n We want to project orthogonally the $n+1$ points $\mathbf{b}_i=(x_i,y_i)$ to points $\mathbf{p}_i$ in the image of multiplication by D: therefore the values of the n-first monomials c_0, c_1, …, c_{n-1} of the polynomial in t evaluated at the $n+1$ scalars x_i: $\mathbf{z}=D\mathbf{c}$. The points under projection are of the form $\mathbf{p}_i=(x_i,z_i)$ so that only the second coordinate of each point changes under projection. Because A is invertible, the columns of D are linearly independent. In that case we know that the projection is given by the $(n+1)\times(n+1)$ invertible matrix $P=D(D^tD)^{-1}D^t$. Then $\mathbf{c}=(D^tD)^{-1}D^t\mathbf{y}$. So we should compute the $(n+1)\times n$ matrix $(D^tD)^{-1}D^t$ first; then evaluate it at **y** to get the coefficients **c**, which we then multiply by D to get the z_i.

Example D.2.2. Here is the computation for Example D.2.1. D is just the first three columns of

$$D = \begin{pmatrix} 1 & -1 & 1 \\ 1 & -1/2 & 1/4 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

so

$$D^t D = \begin{pmatrix} 4 & -1 & 3/2 \\ -1 & 3/2 & -1 \\ 3/2 & -1 & 9/8 \end{pmatrix},$$

A symmetric matrix as expected. Next we need the inverse

$$(D^t D)^{-1} = \begin{pmatrix} 0.55 & -0.3 & -1 \\ -0.3 & 1.8 & 2 \\ -1 & 2 & 4 \end{pmatrix},$$

which is also symmetric. Next we compute

$$D_1 = (D^t D)^{-1} D^t = \begin{pmatrix} -0.15 & 0.45 & 0.55 & 0.15 \\ -0.1 & -0.7 & 0.3 & 1.1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

Here is the MatLab computation:

```
D = [  1.0000   -1.0000    1.0000;
       0.2500   -0.5000    1.0000;
       0         0         1.0000;
       1.0000    1.0000    1.0000]
Dtrans = transpose(D);
DtD = Dtrans*D;
DtDinv = inv(DtD)
DtDinv =
          1.2727   -0.0909   -0.7273
         -0.0909    0.4636    0.1091
         -0.7273    0.1091    0.6727

D1 = DtDinv*Dtrans
D1 =
    0.6364   -0.3636   -0.7273    0.4545
   -0.4455   -0.1455    0.1091    0.4818
   -0.1636    0.4364    0.6727    0.0545
y = [1;-1; -3/2; 0]
     y =
         1.0000
        -1.0000
        -1.5000
```

```
          0
c= D1*y
      c =
        2.0909
       -0.4636
       -1.6091
proj = D*c
        proj =
          0.9455
         -0.8545
         -1.6091
          0.0182
```

So the coefficients of the quadric that passes as close as possible to the original points are

$$\mathbf{c} = D_1\mathbf{y} = (-1.6091, \quad -0.4636, \quad 2.0909),$$

meaning that the quadric is $2.0909t^2 - 0.4636t - 1.6091$. Finally the second coordinate of the projected points are

$$D\mathbf{c} = (0.9455, \quad -0.8545, \quad -1.6091, \quad 0.0182).$$

Notice that they are close to the second coordinates of the original points: $\mathbf{y} = (1,-1,-1.5,0)$. This is expected, since it is the sum of the squares of the differences that we are minimizing.

Below is a graph of the cubic and the quadric of the example on the interval $[-1,1]$. The four points of approximation are between -1 and 1. First the code that produced it:

```
 x = linspace(-1, 1, 100);
y1 = -(2/3)*x.^3 +2*x.^2 +(1/6)*x -3/2;
y2 = 2.0909*x.^2 -0.4636*x -1.6091;
plot(x, y1,'--', x, y2, ':')
```

D.2.3 A Different Approximation

Once again start with a polynomial $f(t)$ of degree n. We want a way of measuring its distance in a difference sense, from a polynomial $g(t)$ of degree $n-1$ on the interval $[-1,1]$, for example. For that we need a new inner product on the vector space P_n of polynomials of degree n. Here is one such:

$$\langle f(t), h(t)\rangle = \int_{-1}^{1} f(t)h(t)dt.$$

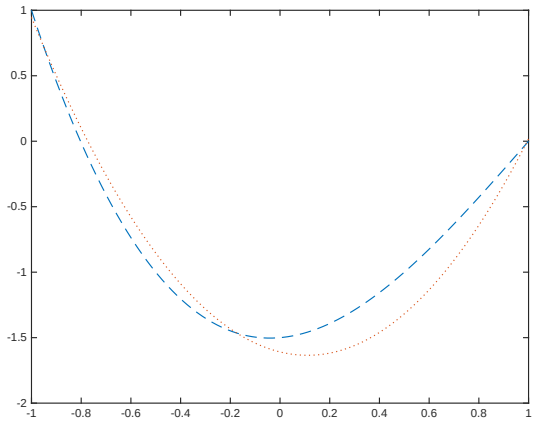

Fig. D.1 The Cubic and the Quadric

We already know that this is an inner product, because any polynomial is continuous. This inner product gives us a notion of distance between polynomials. It is, as always, $\sqrt{\langle f(t)-h(t), f(t)-h(t)\rangle}$ which can be written as an integral as given below.

Using it we want to project orthogonally any polynomial $f(t)$ of degree n to the subspace P_{n-1} of polynomials of degree $n-1$. P_{n-1} has dimension one less that the dimension of P_n. As we know by the Pythagorean Theorem, the orthogonal projection $g(t)$ is the point in P_{n-1} whose distance from $f(t)$ is smallest. This distance is

$$\int_{-1}^{1} (f(t)-g(t))^2 dt.$$

This is a generalization of what we did in the previous section, but where instead of using 4 points, we use all the points on the interval $[-1,1]$.

To make the orthogonal projection we need an orthogonal basis for the vector space of polynomials of degree n. We get it from the ordinary basis $\mathfrak{B} = \{1, t, t^2, \ldots, t^n\}$ of $P(n)$ by the Gram-Schmidt process. The polynomials in the orthogonal basis are called the Legendre polynomials.

Definition D.2.3. The formula for the Legendre polynomial of degree n is

$$f_n(t) = \frac{1}{2^n n!} \frac{d^n}{dt^n}(t^2 - 1)^n$$

We only use this amazing formula in the case $n = 3$, where the computations can be done by hand if you choose.

The first four Legendre polynomials are

$$\begin{aligned} le_0 &= 1; \\ le_1 &= t; \\ le_2 &= \frac{3t^2 - 1}{2}; \\ le_3 &= \frac{5t^3 - 3t}{2}. \end{aligned}$$

The square of their lengths, meaning

$$\langle f_i(t), f_i(t) \rangle = \int_{-1}^{1} f_i(t)^2 dt,$$

are 2, $2/3$, $2/5$, $2/7$. For example let's compute the square of the length of le_2 in MatLab. First we need to square the polynomial. This is a little tricky: MatLab stores the polynomial as the vector of its coefficients, so in our case as $[3/2 \quad 0 \quad -1/2]$. Note that the coefficients are given in decreasing order of the power of the variable. To find the coefficients of its square use the convolution function $conv(\cdot,\cdot)$. It is useful to see that convolution on the coefficients of two polynomials gives the coefficients of the product polynomial. Then transform the convolution back to a polynomial using the $poly2sym(q,t)$, where q is the vector of coefficients returned by the convolution function and t is the name of the variable.

```
p = [3/2 0 -1/2];
q = conv(p,p)
syms t;
p2 = poly2sym(q, t)
```

This returns as output

```
ans = (9*t^4)/4 - (3*t^2)/2 + 1/4
```

Now a kludge: add dots before each power sign by hand:

```
(9*t.^4)/4 - (3*t.^2)/2 + 1/4
```

Finally integrate numerically over the interval $[-1, 1]$:

```
syms t
integral(@(t) (9*t.^4)/4 - (3*t.^2)/2 + 1/4, -1, 1)
        ans = 0.4000
```

which is indeed $2/5$.

Thus these functions are not orthonormal: it would be easy enough to normalize them, but it is the tradition to write them as above, to get the beautiful Definition D.2.3. Let's check at least that two are orthogonal: for example le_2 and le_3. We just repeat the same code:

```
p = [3/2 0 -1/2];
q = [5/2 0 -3/2 0];
pq = conv(p,q);
syms t;
p2 = poly2sym(pq, t)
        (15*t^5)/4 - (7*t^3)/2 + (3*t)/4
integral(@(t) (15*t.^5)/4 - (7*t.^3)/2 + (3*t)/4, -1, 1)
        ans = 2.7756e-17
```

which is effectively 0. We of course knew that the answer is 0, since we are integrating an odd function over an interval symmetric around the origin.

By the usual projection formula, the projection of $f(t)$, a polynomial of degree n, is

$$g(t) = \sum_{i=0}^{n-1} \frac{\langle f, f_i \rangle}{\langle f_i, f_i \rangle} f_i.$$

Here $f_i = le_{i-1}$. By construction g is a polynomial of degree at most $n-1$. Therefore this involves computing a series of definite integrals of polynomials.

Example D.2.4. We start with the cubic $f(t) = -\frac{2}{3}t^3 + 2t^2 + \frac{1}{6}t - \frac{3}{2}$ of the earlier examples D.2.1 and D.2.2. We represent $f(t)$ in MatLab as:

```
ourp = [-2/3, 2, 1/6, -3/2];
syms t;
ourp2 = poly2sym(ourp, t)
        - (2*t^3)/3 + 2*t^2 + t/6 - 3/2
integral(@(t) - (2*t.^3)/3 + 2*t.^2 + t/6 - 3/2, -1, 1)
        ans = -1.6667
```

So the inner product with the constant term is just the definite integral of the cubic over the interval $[-1,1]$ divided by 2, giving $-5/6$.

The inner product with the linear term has coefficient

$$3/2 \int_{-1}^{1} -\frac{2}{3}t^4 + 2t^3 + \frac{1}{6}t^2 - \frac{3}{2}t = -0.2333$$

Use the MatLab code

```
syms t;
integral(@(t)3/2*(-2/3*t.^4 + 2*t.^3 + 1/6*t.^2-3/2*t), -1, 1)
```

to compute it.

The inner product with the quadratic term is computed by

```
integral(@(t) 5/2*(- t.^5 + 3*t.^4 + (7*t.^3)/12
     - (13*t.^2)/4 - t/12 + 3/4), -1, 1)
          ans = 1.3333
```

which is $4/3$. This tells us that the quadric that best approximates our cubic in this inner product is

$$-5/6 - 0.2333t + 1.3333(3/2t^2 - 1/2) = 2t^2 - 5/12t - 3/2$$

Here is how to make the comparison graph in MatLab:

```
t = linspace(-1, 1, 100);
y1 = -(2/3)*t.^3 +2*t.^2 +(1/6)*t -3/2;
y2 = 2*t.^2  -5/12 *t   -3/2;
plot(t, y1,'--', t, y2, ':')
```

The graph is given below. You should compare the two graphs.

D.2.4 Comparison

Finally we compare the two quadrics that are the two approximations of our cubic using different inner products.

Example D.2.5. Our original cubic is $C = -\frac{2}{3}t^3 + 2t^2 + \frac{1}{6}t - \frac{3}{2}$.

The first quadric computed in §D.2.2 is $Q_1 = 2.090t^2 - 0.4636t - 1.6091$

By definition the quadric $Q_2 = 2t^2 - 5/12t - 3/2$ computed in §D.2.3 is further away from C than Q_1 in the inner product used there. Let's check that. What are the points. 0n Q_2 for the 4 points used there: they are $(-1, 11/12)$, $(-1/2, -19/24)$, $(0, -3/2)$, $(1, 1/12)$. To compute the square of the distance of Q_1 from the cubic

```
q1= [2.0909 -0.4636 -1.6091];
val1 = [polyval(q1,-1) polyval(q1,-1/2) polyval(q1,0) polyval(q1,1)]
        val1 =0.9454   -0.8546   -1.6091    0.0182
 Cval = [1 -1 -3/2 0];
 dif1 = val1 - Cval
      dif1 = -0.0546    0.1454   -0.1091    0.0182
 dif1trans = transpose(dif1);
 norm1 = dif1*dit1trans
     norm1 =  0.0364
```

So this is the distance squared between the cubic and the first quadric.

Now we repeat for the quadric Q_2

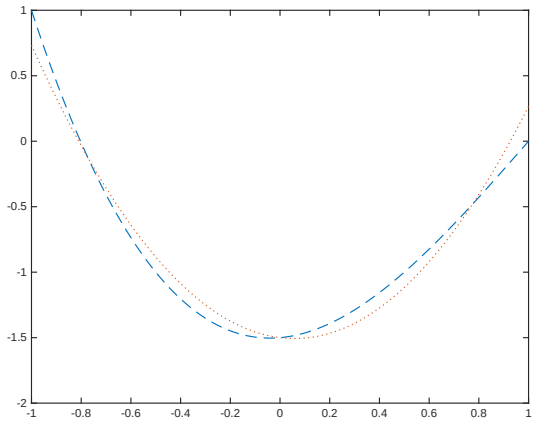

Fig. D.2 L2 norm approximation

```
q2 = [2 -5/12 -3/2]
val2 = [polyval(q2,-1) polyval(q2,-1/2) polyval(q2,0) polyval(q2,1)]
     val2 = 0.9167   -0.7917   -1.5000    0.0833
 dif2 = val1 - Cval
     dif2 = -0.9167   -0.2083         0    0.9167
 dif2trans = transpose(dif2);
 norm2 = dif2*dif2trans
      norm2 =  0.0573
```

So the distance of Q_2 from the original cubic in the first inner product is greater than that of Q_1, as it must be.

Now we compute in the second inner product:

```
q2 = [0 2 -5/12 -3/2];
C = [-2/3 2 1/6 -3/2];
diff2 = C - q2
     diff2 =-0.6667   0   0.5833   0
innerproddiff= conv(diff2, diff2)
     innerproddiff =0.4444    0   -0.7778    0   0.3403   0   0
 syms t;
 ourp2 = poly2sym(innerproddiff,t)
      ourp2 = 4*t^6)/9 - (7*t^4)/9 + (49*t^2)/144
 integral(@(t) (4*t.^6)/9 - (7*t.^4)/9 + (49*t.^2)/144, -1, 1)
     ans = 0.0427
```

That is the answer for Q_2. We repeat for Q_1:

```
q1= [2.0909 -0.4636 -1.6091];
C = [-2/3 2 1/6 -3/2];
diff1 = C - q1
     diff1 = -0.6667   -0.0909    0.6303    0.1091
innerproddiff1= conv(diff1, diff1)
     innerproddiff1 =
          0.4444    0.1212   -0.8321   -0.2600   0.3774   0.1375   0.0119
syms t;
ourp1 = poly2sym(innerproddiff1,t)
ourp1 =
   (4*t^6)/9 + (303*t^5)/2500 - (3747412578821693*t^4)/4503599627370496
   - (4684628960104061*t^3)/18014398509481984
   + (6798664461827979*t^2)/18014398509481984
  + (4954831006611443*t)/36028797018963968
  + 6861502807124709/576460752303423488

integral(@(t) (4*t.^6)/9 + (303*t.^5)/2500
     - (3747412578821693*t.^4)/4503599627370496
     - (4684628960104061*t.^3)/18014398509481984
     + (6798664461827979*t.^2)/18014398509481984
     + (4954831006611443*t)/36028797018963968
     + 6861502807124709/576460752303423488, -1, 1)
           ans = 0.0696
```

So indeed the distance of Q_1 from C is greater than that of Q_2 in this norm.

How should we compare the best approximation in the two norms: in the pictures, which of the two quadrics looks closer? Clearly Q_2 looks like a better approximation than Q_1.

D.2.5 Exercise

Take the four points $(-1,0)$, $(-1/3,-1/2)$, $(1/3,0)$, $(1,1.5)$, and find the cubic that goes through them, and then the two quadrics that give the best approximations in the two norms.

D.2.6 Computing the Interpolation Polynomial

Assume given $n+1$ points (x_i,y_i). where the x_i are distinct. in the plane. Let A be the Vandermonde matrix A of (D.2.1), evaluated at the x_i then finding the unique polynomial $p(x)$ of degree n, with coefficients c_0, c_1, ..., c_n just means solving the system $A\mathbf{c} = \mathbf{y}$. So the problem is solved by finding the inverse of A, for which there is a formula, albeit complicated. Inverting a large matrix is time consuming so this is not how the computation is actually done. Instead one finds a basis for the polynomials of degree Vn that is well suited for this problem. This is known as Newton interpolation. The basis is

$$\mathfrak{N} = \{\mathbf{n}_0 = 1, \mathbf{n}_1 = x - x_0, \mathbf{n}_2 = (x-x_0)(x-x_0), \ldots, \mathbf{n}_n = \prod_{i=0}^{n}(x-x_i)\}.$$

It is clear that this is a basis because the polynomials are of different degrees. The goal, therefore, is to find the expression of the Vandermonde equation in this basis, i.e. find the coefficients c_i such that

$$p(x) = c_0\mathbf{n}_0 + c_1\mathbf{n}_1 + \ldots c_n\mathbf{n}_n.$$

The advantage of this basis is that the interpolating matrix A is upper triangular in it, and that it is easy to increment the computation if a new point is added.

To do this we need to defined the divided difference of $\{x_0, x_1, \ldots, x_n\}$ and $\{y_0, y_1, \ldots, y_n\}$. We write these with square brackets. Then

$$\begin{aligned} f[x_i] &= y_i, \\ f[x_i, x_{i+1}] &= \frac{y_{i+1} - y_i}{x_{i+1} - x_i}, \\ \ldots &= \ldots \\ f[x_i, \ldots, x_{i+k}] &= \frac{f[x_{i+1}, \ldots, x_{i+k}] - f[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i} \end{aligned}$$

Thus any divided difference with k variables involves two divided differences with $k-1$ variables, so they can be computed recursively.

Theorem D.2.6. *The coefficients c_i of the interpolating polynomials in the Newton basis are*

$$c_j = f[x_0, \ldots, x_j]$$

Proof. □

D.2.7 The kernel of the rectangular Vandermonde determinant

Take the transpose of the Vandermonde matrix V, and remove its last row. This is the matrix we call D above. It is a $n \times (n+1)$ matrix of maximum rank: therefore its nullspace has dimension 1. It is of some interest to find a generator in the nullspace. it the case $n = 2$ it is easily seen to be

$$\left(\frac{1}{(x_0 - x_1)(x_0 - x_2)} \quad \frac{1}{(x_1 - x_0)(x_1 - x_2)} \quad \frac{1}{(x_2 - x_0)(x_2 - x_1)} \right)$$

In the general case:

Proposition D.2.7. *The i-th coordinate of a generator of the kernel can be written*

$$\frac{1}{(x_i - x_1) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)}$$

Find a direct proof.

The interesting fact about this generator is that all its coordinates are non-zero. I want to show that this implies that the unique relation between the projected points $p_0, p_1, \ldots, p_n$ in $\mathbb{R}^n$ has all its coefficients non-zero. That is clear because in the project one first applies D and later its transpose. The kernel and the linear combination of the projected points are given by the same vector.

References

1. Michael Artin. *Algebra*. Prentice Hall, New York, 1991.
2. Sheldon Axler. *Linear Algebra Done Right*. Undergraduate Texts in Mathematics. Springer, Upper Saddle River, NJ., 1997.
3. Alexander Barvinok. *A Course in Convexity*. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2002.
4. Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge U.P, Cambridge, 2004.
5. Monique Florenzano and Cuong Le Van. *Finite Dimensional Convexity and Op;timization*. Springer, New York, 2001.
6. Joel N. Franklin. *Matrix Theory*. Dover, New York, 1960.
7. David Freedman, Robert Pisani, and Roger Purves. *Statistics*. Norton, New York, 1998.
8. F. R. Gantmacher. *The Theory of Matrices*. Chelsea, New York, 1959. translation from the Russian original by K. A. Hirsch.
9. Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins U. P., Baltimore, 1996.
10. Paul Halmos. *Finite-Dimensional Vector Spaces*. Springer, New York, 1974.
11. Roger Hart. *The Chinese Roots of Linear Algebra*. Johns Hopkins U. P., Baltimore, 2011.
12. Kenneth Hoffman and Ray Kunze. *Linear Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1971.
13. Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1990.
14. Paul J. Kelly and Max L. Weiss. *Geometry and Convexity*. John Wiley, New York, 1979.
15. Serge Lang. *Linear Algebra*. Springer, New York, 1987.
16. Peter D. Lax. *Linear Algebra*. Wiley-Interscience, New York, 1997.
17. Steven R. Lay. *Convex Sets and Their Applications*. John Wiley, New York, 1982. reprint by Dover in 2007.
18. Thomas Muir. *The Theory of Determinants in the Historical Order of Development*. Dover, New York, 1960.
19. O. Neugebauer. *The Exact Sciences in Antiquity*. Princeton U. P., Princeton, NJ, 1952.
20. O. Neugebauer. *Vorlesungen über Geschichte der antiken mathematischen Wissenschaften, Erster Band*. Springer, New York, 1969.
21. A. Wayne Roberts and Dale E. Varberg. *Convex Functions*. Academic Press, New York, 1973.
22. Eleanor Robson. *Mathematics in Ancient Iraq, A Social History*. Princeton U. P., Princeton, NJ, 2008.
23. R. Tyrrell Rockafellar. *Convex Analysis*. Princeton U.P., Princeton, 1973.
24. Denis Serre. *Matrices*. Springer, New York, 2010.
25. L. A. Steen. Highlights in the history of spectral theory. *American Math Monthly*, 80(4):359–381, 1973.
26. Stephen M. Stigler. *The History of Statistics,The Measurement of Uncertainty before 1900*. Belknap Press of Harvard, Cambridge, Mass, 1986.
27. Josef Stoer and Christoph Witzgall. *Convexity and Optimization in Finite Dimensions I*. Number 163 in Die Grundlehren der mathematischen Wissenschlaften in Einzeldarstellungen. Springer-Verlag, New York, 1970.
28. Gilbert Strang. *Linear Algebra and its Applications*. Harcourt Brace Jovanovich, San Diego, CA, 1988.
29. J. J. Sylvester. A demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares. *Philosophical Magazine, IV*, 23:47–51, 1852.
30. B. L. van der Waerden. *Mathematical Statistics*. Springer-Verlag, Berlin-Heidelberg, 1969.
31. Herman Weyl. The elementary theory of convex polyhedra. (24), 1950.
32. Sir Edmund Whittaker and G. Robinson. *The Calculus of Observations*. Blackie & Son Limited, London and Glasgow, 1944.