Basic Concepts of Linear Algebra

by

Jim Carrell
Department of Mathematics
University of British Columbia

Chapter 1

Introductory Comments to the Student

This textbook is meant to be an introduction to abstract linear algebra for first, second or third year university students who are specializing in mathematics or a closely related discipline. We hope that parts of this text will be relevant to students of computer science and the physical sciences. While the text is written in an informal style with many elementary examples, the propositions and theorems are carefully proved, so that the student will get experince with the theorem-proof style. We have tried very hard to emphasize the interplay between geometry and algebra, and the exercises are intended to be more challenging than routine calculations. The hope is that the student will be forced to think about the material.

The text covers the geometry of Euclidean spaces, linear systems, matrices, fields ($\mathbb{Q}, \mathbb{R}$, $\mathbb{C}$ and the finite fields $\mathbb{F}_p$ of integers modulo a prime p), vector spaces over an arbitrary field, bases and dimension, linear transformations, linear coding theory, determinants, eigen-theory, projections and pseudo-inverses, the Principal Axis Theorem for unitary matrices and applications, and the diagonalization theorems for complex matrices such as the Jordan decomposition. The final chapter gives some applications of symmetric matrices positive definiteness. We also introduce the notion of a graph and study its adjacency matrix. Finally, we prove the convergence of the QR algorithm. The proof is based on the fact that the unitary group is compact.

Although, most of the topics listed above are found in a standard course on linear algebra, some of the topics such as fields and linear coding theory are seldom treated in such a course. Our feeling is, however, that because

coding theory is such an important component of the gadgets we use everyday, such as personal computers, CD players, modems etc., and because linear coding theory gives such a nice illustration of how the basic results of linear algebra apply, including it in a basic course is clearly appropriate. Since the vector spaces in coding theory are defined over the prime fields, the students get to see explicit situations where vector space structures which don't involve the real numbers are important.

This text also improves on the standard treatment of the determinant, where either its existence in the $n \times n$ case for $n > 3$ is simply assumed or it is defined inductively by the Laplace expansion, an d the student is forced to believe that all Laplace expansions agree. We use the classical definition as a sum over all permutations. This allows one to give a quite easy proof of the Laplace expansion, for example. Much of this material here can be covered in a 13-15 week semester.

Throughout the text, we have attempted to make the explanations clear, so that students who want to read further will be able to do so.

Contents

Chapter 2

Euclidean Spaces and Their Geometry

By Euclidean n-space, we mean the space $\mathbb{R}^n$ of all (ordered) n-tuples of real numbers. This is the domain where much, if not most, of the mathematics taught in university courses such as linear algebra, vector analysis, differential equations etc. takes place. And although the main topic of this book is algebra, the fact is that algebra and geometry can hardly be seperated: we need a strong foundation in both. The purpose of this chapter is thus to provide a succinct introduction to Euclidean space, with the emphasis on its geometry.

2.1 $\mathbb{R}^n$ and the Inner Product.

2.1.1 Vectors and n-tuples

Throughout this text, $\mathbb{R}$ will stand for the real numbers. Euclidean n-space, $\mathbb{R}^n$, is by definition the set of all (ordered) n-tuples of real numbers. An n-tuple is just a sequence consisting of n real numbers written in a column like

$$\mathbf{r} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}.$$

Sometimes the term sequence is replaced by the term or string or word. The entries $r_1, \ldots, r_n$ are called the *components* of the n-tuple, r_i being the ith *component*. It's important to note that the order of the components matters:

e.g.

$$\begin{pmatrix}1\\2\\3\end{pmatrix} \neq \begin{pmatrix}2\\3\\1\end{pmatrix}.$$

Definition 2.1. The elements of $\mathbb{R}^n$ will be called *vectors*, and $\mathbb{R}^n$ itself will be called a *Euclidean n-space*.

Vectors will be denoted by a bold faced lower case letters $\mathbf{a}, \mathbf{b}, \mathbf{c} \ldots$ and so forth. To simplify our notation a bit, we will often express a vector as a row expression by putting

$$\begin{pmatrix}r_1\\r_2\\\vdots\\r_n\end{pmatrix} = (r_1, r_2, \ldots, r_n)^T.$$

A word of explanation about the term vector is in order. In physics books and in some calculus books, a vector refers any directed line segment in $\mathbb{R}^2$ or $\mathbb{R}^3$. Of course, a vector $\mathbf{r}$ in $\mathbb{R}^n$ is a directed line segment starting at the origin $\mathbf{0} = (0, 0, \ldots, 0)^T$ of $\mathbb{R}^n$. This line segment is simply the set of points of the form $t\mathbf{r} = (tr_1, tr_2, \ldots tr_n)^T$, where $0 \leq t \leq 1$. More generally, the term vector may refer to the set of all directed line segments parallel to a given segment with the same length. But in linear algebra, the term vector is used to denote an element of a vector space. The vector space we are dealing with here, as will presently be explained, is $\mathbb{R}^n$, and its vectors are therefore real n-tuples.

2.1.2 Coordinates

The Euclidean spaces $\mathbb{R}^1$, $\mathbb{R}^2$ and $\mathbb{R}^3$ are especially relevant since they physically represent a line, plane and a three space respectively. It's a familiar assumption that the points on a line L can be put into a one to one correspondence with the real numbers $\mathbb{R}$. If $a \in \mathbb{R}$ (that is, if a is an element of $\mathbb{R}$), then the point on L corresponding to a has distance $|a|$ from the origin, which is defined as the point corresponding to 0. Such a one to one correspondence puts a coordinate system on L.

Next, we put a coordinate system called xy-coordinates on a plane by selecting two (usually orthogonal) lines called an x-axis and a y-axis, each having a coordinate system, and identifying a point P in the plane with the element $(p_1, p_2)^T$ of $\mathbb{R}^2$, where p_1 is the projection of P parallel to the y-axis

onto the x-axis, and p_2 is the projection of P parallel to the x-axis onto the y-axis. This is diagrammed in the following figure.

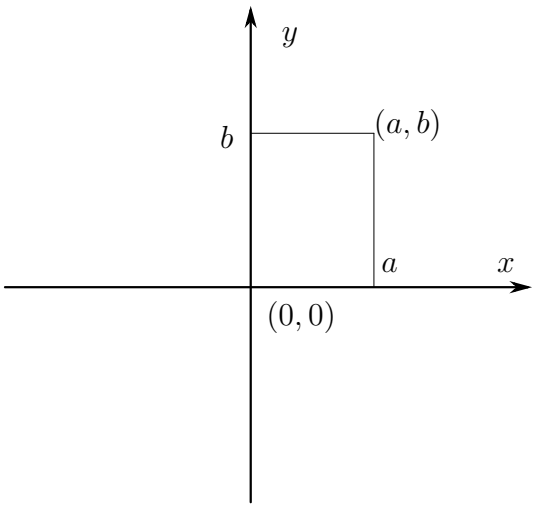


FIGURE (Euclidean PLANE)

In the same manner, the points of Euclidean 3-space are parameterized by the ordered 3-tuples of real numbers, i.e. $\mathbb{R}^3$; that is, that is, every point is uniquely identified by assigning it xyz-coordinates. Thus we can also put a coordinate system on $\mathbb{R}^3$.

FIGURE (Euclidean 3-space)

But just as almost everyone eventually needs more storage space, we may also need more coordinates to store important data. For example, if we are considering a linear equation such as

$$3x + 4y + 5z + 6w = 0,$$

where the solutions are 4-tuples, we need $\mathbb{R}^4$ to store them. While extra coordinates give more degrees of freedom, our geometric intuition doesn't work very well in dimensions bigger than three. This is where the algebra comes in.

2.1.3 The Vector Space $\mathbb{R}^n$

Vector addition in $\mathbb{R}^2$ or $\mathbb{R}^3$ is probably already very familiar to you. Two vectors are added by a rule called the Parallelogram Law, which we will review below. Since n may well be bigger than 3, we define vector addition in a different, in fact much simpler, way by putting

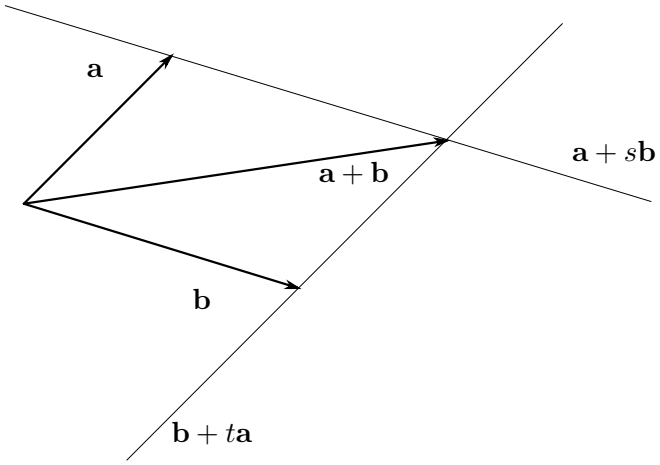
$$\mathbf{a}+\mathbf{b}=\begin{pmatrix}a_1\\a_2\\\vdots\\a_n\end{pmatrix}+\begin{pmatrix}b_1\\b_2\\\vdots\\b_n\end{pmatrix}=\begin{pmatrix}a_1+b_1\\a_2+b_2\\\vdots\\a_n+b_n\end{pmatrix}. \tag{2.1}$$

Thus addition consists of adding the corresponding components of the two vectors being added.

There is a second operation called *scalar multiplication*, where a vector $\mathbf{a}$ is dilated by a real number r. This is defined (in a rather obvious way) by

$$r\mathbf{a}=r\begin{pmatrix}a_1\\a_2\\\vdots\\a_n\end{pmatrix}=\begin{pmatrix}ra_1\\ra_2\\\vdots\\ra_n\end{pmatrix}. \tag{2.2}$$

These two operations satisfy the axioms which define a vector space. They will be stated explicitly in Chapter 4. Scalar multiplication has an obvious geometric interpretation. Multiplying $\mathbf{a}$ by r stretches or shrinks $\mathbf{a}$ along itself by the factor $|r|$, changing its direction if $r<0$. The geometric interpretation of addition is the Parallelogram Law.



PARALLELOGRAM LAW

Parallelogram Law: *The sum* $\mathbf{a}+\mathbf{b}$ *is the vector along the diagonal of the parallelogram with vertices at* $\mathbf{0}$, $\mathbf{a}$ *and* $\mathbf{b}$.

Thus vector addition (2.1) agrees with the classical way of defining addition. The Parallelogram Law in $\mathbb{R}^2$ by showing that the line through $(a,b)^T$ parallel to $(c,d)^T$ meets the line through $(c,d)^T$ parallel to (a,b) at $(a+c,b+d)^T$. Note that lines in $\mathbb{R}^2$ can be written in the form $rx+sy=t$, where $r,s,t\in\mathbb{R}$, so this is an exercise in writing the equation of a line and computing where two lines meet. (See Exercise 2.29.)

Checking the Parallelogram Law in $\mathbb{R}^3$ requires that we first discuss how to represent a line in $\mathbb{R}^3$. The Parallelogram Law in $\mathbb{R}^n$, will follow in exactly the same way. We will treat this matter below.

2.1.4 The dot product

We now take up *measurements* in $\mathbb{R}^n$. The way we measure things such as length and angles is to use an important operation called either the *inner product* or the *dot product*.

Definition 2.2. The inner product of two vectors $\mathbf{a}=(a_1,a_2,\ldots,a_n)^T$ and $\mathbf{b}=(b_1,b_2,\ldots,b_n)^T$ in $\mathbb{R}^n$ is defined to be

$$\mathbf{a}\cdot\mathbf{b}:=\sum_{i=1}^{n}a_ib_i. \tag{2.3}$$

Note that if $n=1$, $\mathbf{a}\cdot\mathbf{b}$ is the usual product. The inner product has several important properties. Let $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ be arbitrary vectors and r any scalar (i.e., $r\in\mathbb{R}$). Then

(1) $\mathbf{a}\cdot\mathbf{b}=\mathbf{b}\cdot\mathbf{a}$,

(2) $(\mathbf{a}+\mathbf{b})\cdot\mathbf{c}=\mathbf{a}\cdot\mathbf{c}+\mathbf{b}\cdot\mathbf{c}$,

(3) $(r\mathbf{a})\cdot\mathbf{b}=\mathbf{a}\cdot(r\mathbf{b})=r(\mathbf{a}\cdot\mathbf{b})$, and

(4) $\mathbf{a}\cdot\mathbf{a}>0$ unless $\mathbf{a}=\mathbf{0}$, in which case $\mathbf{a}\cdot\mathbf{a}=0$.

These properties are all easy to prove, so we will leave them as exercises.

The *length* $|\mathbf{a}|$ of $\mathbf{a}\in\mathbb{R}^n$ is defined in terms of the dot product by putting

$$\begin{aligned}|\mathbf{a}| &= \sqrt{\mathbf{a}\cdot\mathbf{a}}\\ &= (\sum_{i=1}^{n}a_i^2)^{1/2}.\end{aligned}$$

This definition generalizes the usual square root of the sum of squares definition of length for vectors in $\mathbb{R}^2$ and $\mathbb{R}^3$. Notice that

$$|r\mathbf{a}| = |r||\mathbf{a}|\,.$$

The *distance* between two vectors $\mathbf{a}$ and $\mathbf{b}$ is defined as the length of their difference $\mathbf{a} - \mathbf{b}$. Denoting this distance by $d(\mathbf{a}, \mathbf{b})$, we see that

$$\begin{aligned} d(\mathbf{a}, \mathbf{b}) &= |\mathbf{a} - \mathbf{b}| \\ &= \big((\mathbf{a} - \mathbf{b}) \cdot (\mathbf{a} - \mathbf{b})\big)^{1/2} \\ &= \big(\sum_{i=1}^{n} (a_i - b_i)^2\big)^{1/2}. \end{aligned}$$

2.1.5 Orthogonality and projections

Next we come to an important notion which involves both measurement and geometry. Two vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^n$ are said to be *orthogonal* (a fancy word for perpendicular) if $\mathbf{a} \cdot \mathbf{b} = 0$. Note that the zero vector $\mathbf{0}$ is orthogonal to every vector, and by property (4) of the dot product, $\mathbf{0}$ is the only vector orthogonal to itself. Two vectors in $\mathbb{R}^2$, say $\mathbf{a} = (a_1, a_2)^T$ and $\mathbf{b} = (b_1, b_2)^T$, are orthogonal if and only if and only if $a_1 b_1 + a_2 b_2 = 0$. Thus if $a_1, b_2 \neq 0$, then $\mathbf{a}$ and $\mathbf{b}$ are orthogonal if and only if $a_2/a_1 = -b_1/b_2$. Thus, the slopes of orthogonal vectors in $\mathbb{R}^2$ are negative reciprocals.

For vectors in $\mathbb{R}^n$, the meaning of orthogonality follows from the following property.

Proposition 2.1. *Two vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *in* $\mathbb{R}^n$ *are orthogonal if and only if* $|\mathbf{a} + \mathbf{b}| = |\mathbf{a} - \mathbf{b}|$.

Let's prove this geometrically, at least for $\mathbb{R}^2$. Consider the triangle with vertices at $\mathbf{0}, \mathbf{a}, \mathbf{b}$. The hypotenuse of this triangle is a segment of length $|\mathbf{a}-\mathbf{b}|$, which follows from the Parallelogram Law. Next consider the triangle with vertices at $\mathbf{0}, \mathbf{a}, -\mathbf{b}$. The hypotenuse of this triangle is a segment of length $|\mathbf{a}+\mathbf{b}|$, which also follows from the Parallelogram Law. Now suppose $|\mathbf{a} + \mathbf{b}| = |\mathbf{a} - \mathbf{b}|$. Then by the side side side criterion for congruence, which says that two triangles are congruent if and only if they have corresponding sides of equal length, the two triangles are congruent. It follows that $\mathbf{a}$ and $\mathbf{b}$ are orthogonal. For the converse direction, suppose $\mathbf{a}$ and $\mathbf{b}$ are orthogonal. Then the side angle side criterion for congruence applies, so our triangles are congruent. Thus $|\mathbf{a} + \mathbf{b}| = |\mathbf{a} - \mathbf{b}|$.

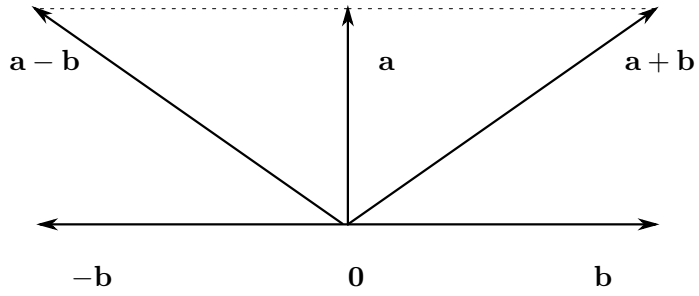


DIAGRAM FOR PROOF

In fact, it is much easier to use algebra (namely the dot product). The point is that $\mathbf{a} \cdot \mathbf{b} = 0$ if and only if $|\mathbf{a} + \mathbf{b}| = |\mathbf{a} - \mathbf{b}|$. The details are left as an exercise.

One of the most fundamental applications of the dot product is the *orthogonal decomposition* of a vector into two or more mutually orthogonal components.

Proposition 2.2. *Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ be given, and suppose that $\mathbf{b} \neq \mathbf{0}$. Then there exists a unique scalar r such that $\mathbf{a} = r\mathbf{b} + \mathbf{c}$ where $\mathbf{b}$ and $\mathbf{c}$ are orthogonal. In fact,*

$$r = (\frac{\mathbf{a} \cdot \mathbf{b}}{\mathbf{b} \cdot \mathbf{b}}),$$

and

$$\mathbf{c} = \mathbf{a} - (\frac{\mathbf{a} \cdot \mathbf{b}}{\mathbf{b} \cdot \mathbf{b}})\mathbf{b}.$$

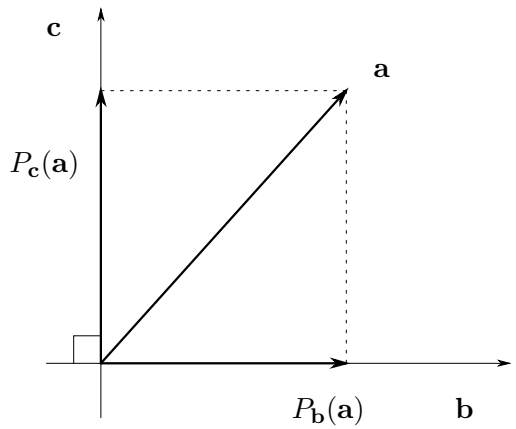
Proof. We see this as follows: since we want $r\mathbf{b} = \mathbf{a} - \mathbf{c}$, where $\mathbf{c}$ has the property that $\mathbf{b} \cdot \mathbf{c} = 0$, then

$$r\mathbf{b} \cdot \mathbf{b} = (\mathbf{a} - \mathbf{c}) \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{b} - \mathbf{c} \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{b}.$$

As $\mathbf{b} \cdot \mathbf{b} \neq 0$, it follows that $r = \mathbf{a} \cdot \mathbf{b} / \mathbf{b} \cdot \mathbf{b}$. The reader should check that $\mathbf{c} = \mathbf{a} - (\frac{\mathbf{a} \cdot \mathbf{b}}{\mathbf{b} \cdot \mathbf{b}})\mathbf{b}$ is orthogonal to $\mathbf{b}$. Thus we get the desired orthogonal decomposition

$$\mathbf{a} = (\frac{\mathbf{a} \cdot \mathbf{b}}{\mathbf{b} \cdot \mathbf{b}})\mathbf{b} + \mathbf{c}.$$

□



ORTHOGONAL DECOMPOSITION

Definition 2.3. The vector

$$P_{\mathbf{b}}(\mathbf{a}) = (\frac{\mathbf{a} \cdot \mathbf{b}}{\mathbf{b} \cdot \mathbf{b}})\mathbf{b}$$

will be called the *orthogonal projection* of $\mathbf{a}$ on $\mathbf{b}$.

By the previous Proposition, another way to express the orthogonal decomposition of $\mathbf{a}$ into the sum of a component parallel to $\mathbf{b}$ and a component orthogonal to $\mathbf{b}$ is

$$\mathbf{a} = P_{\mathbf{b}}(\mathbf{a}) + (\mathbf{a} - P_{\mathbf{b}}(\mathbf{a})). \tag{2.4}$$

Now suppose $\mathbf{b}$ and $\mathbf{c}$ are any two nonzero orthogonal vectors in $\mathbb{R}^2$, so that $\mathbf{b} \cdot \mathbf{c} = 0$. I claim that any vector $\mathbf{a}$ orthogonal to $\mathbf{b}$ is a multiple of $\mathbf{c}$. Reason: if $\mathbf{b} = (b_1, b_2)^T$ and $\mathbf{a} = (a_1, a_2)^T$, then $a_1 b_1 + a_2 b_2 = 0$. Assuming, for example, that $b_1 \neq 0$, then

$$a_1 = -\frac{b_2}{b_1} a_2 = \frac{c_1}{c_2} a_2,$$

and the claim follows.

It follows that for any $\mathbf{a} \in \mathbb{R}^2$, there are scalars r and s so that $\mathbf{a} = r\mathbf{b} + s\mathbf{c}$. We can solve for r and s by using the dot product as before. For example, $\mathbf{a} \cdot \mathbf{b} = r\mathbf{b} \cdot \mathbf{b}$. Hence we can conclude that if $\mathbf{b} \neq \mathbf{0}$, then

$$r\mathbf{b} = P_{\mathbf{b}}(\mathbf{a}),$$

and similarly, if $\mathbf{c} \neq \mathbf{0}$, then

$$s\mathbf{c} = P_{\mathbf{c}}(\mathbf{a}).$$

Therefore, we have now proved a fundamental fact which we call the **projection formula** for $\mathbb{R}^2$.

Proposition 2.3. *If* $\mathbf{b}$ *and* $\mathbf{c}$ *are two non zero mutually orthogonal vectors in* $\mathbb{R}^2$*, then any vector* $\mathbf{a}$ *in* $\mathbb{R}^2$ *can be uniquely expressed as the sum of its projections. In other words,*

$$\mathbf{a} = P_{\mathbf{b}}(\mathbf{a}) + P_{\mathbf{c}}(\mathbf{a}) = (\frac{\mathbf{a}\cdot\mathbf{b}}{\mathbf{b}\cdot\mathbf{b}})\mathbf{b} + (\frac{\mathbf{a}\cdot\mathbf{c}}{\mathbf{c}\cdot\mathbf{c}})\mathbf{c}. \tag{2.5}$$

Projections can be written much more simply if we bring in the notion of a unit vector. When $\mathbf{b} \neq \mathbf{0}$, the *unit vector along* $\mathbf{b}$ is defined to be the vector of length one given by the formula

$$\widehat{\mathbf{b}} = \frac{\mathbf{b}}{(\mathbf{b}\cdot\mathbf{b})^{1/2}} = \frac{\mathbf{b}}{|\mathbf{b}|}.$$

(Check that $\widehat{\mathbf{b}}$ is indeed of length one,) Unit vectors are also called *directions.* Keep in mind that the direction $\widehat{\mathbf{a}}$ exists only when $\mathbf{a} \neq \mathbf{0}$. It is obviously impossible to assigne a direction to the zero vector. If $\widehat{\mathbf{b}}$ and $\widehat{\mathbf{c}}$ are unit vectors, then the projection formula (2.5) takes the simpler form

$$\mathbf{a} = (\mathbf{a}\cdot\widehat{\mathbf{b}})\widehat{\mathbf{b}} + (\mathbf{a}\cdot\widehat{\mathbf{c}})\widehat{\mathbf{c}}. \tag{2.6}$$

Example 2.1. Let $\mathbf{b} = (3,4)^T$ and $\mathbf{c} = (4,-3)^T$. Then $\widehat{\mathbf{b}} = \frac{1}{5}(3,4)^T$ and $\widehat{\mathbf{c}} = \frac{1}{5}(4,-3)^T$. Let $\mathbf{a} = (1,1)$. Thus, for example, $P_{\mathbf{b}}(\mathbf{a}) = \frac{7}{5}(3,4)^T$, and $\mathbf{a} = \frac{7}{5}(3,4)^T + \frac{1}{5}(4,-3)^T$.

2.1.6 The Cauchy-Schwartz Inequality and Cosines

If $\mathbf{a} = \mathbf{b} + \mathbf{c}$ is an orthogonal decomposition in $\mathbb{R}^n$ (which just means that $\mathbf{b}\cdot\mathbf{c} = 0$), then

$$|\mathbf{a}|^2 = |\mathbf{b}|^2 + |\mathbf{c}|^2.$$

This is known as Pythagoras's Theorem (see Exercise 4).

If we apply Pythagoras' Theorem to (2.4), for example, we get

$$|\mathbf{a}|^2 = |P_{\mathbf{b}}(\mathbf{a})|^2 + |\mathbf{a} - P_{\mathbf{b}}(\mathbf{a})|^2.$$

Hence,

$$|\mathbf{a}|^2 \geq |P_{\mathbf{b}}(\mathbf{a})|^2 = (\frac{\mathbf{a}\cdot\mathbf{b}}{\mathbf{b}\cdot\mathbf{b}})^2|\mathbf{b}|^2 = \frac{(\mathbf{a}\cdot\mathbf{b})^2}{|\mathbf{b}|^2}.$$

Cross multiplying and taking square roots, we get a famous fact known as the Cauchy-Schwartz inequality.

Proposition 2.4. *For any* $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^n$, *we have*

$$|\mathbf{a} \cdot \mathbf{b}| \leq |\mathbf{a}||\mathbf{b}|.$$

Moreover, if $\mathbf{b} \neq \mathbf{0}$, *then equality holds if and only if* $\mathbf{a}$ *and* $\mathbf{b}$ *are collinear.*

Note that two vectors $\mathbf{a}$ and $\mathbf{b}$ are said to be *collinear* whenever one of them is a scalar multiple of the other. If either $\mathbf{a}$ and $\mathbf{b}$ is zero, then automatically they are collinear. If $\mathbf{b} \neq \mathbf{0}$ and the Cauchy-Schwartz inequality is an equality, then working backwards, one sees that $|\mathbf{a} - P_{\mathbf{b}}(\mathbf{a})|^2 = 0$, hence the validity of the second claim. You are asked to supply the complete proof in Exercise 6.

Cauchy-Schwartz says that for any two unit vectors $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$, we have the inequality

$$-1 \leq \widehat{\mathbf{a}} \cdot \widehat{\mathbf{b}} \leq 1.$$

We can therefore define the angle θ between any two non zero vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^n$ by putting

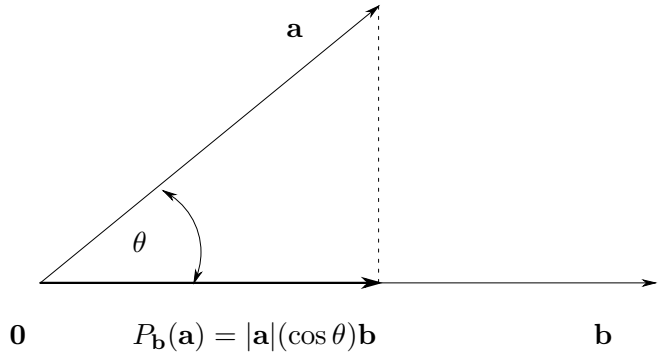
$$\theta := \cos^{-1}(\hat{\mathbf{a}} \cdot \hat{\mathbf{b}}).$$

Note that we don't try to define the angle when either $\mathbf{a}$ or $\mathbf{b}$ is $\mathbf{0}$. (Recall that if $-1 \leq x \leq 1$, then $\cos^{-1} x$ is the unique angle θ such that $0 \leq \theta \leq \pi$ with $\cos\theta = x$.) With this definition, we have

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta \tag{2.7}$$

provided $\mathbf{a}$ and $\mathbf{b}$ are any two non-zero vectors in $\mathbb{R}^n$. Hence if $|\mathbf{a}| = |\mathbf{b}| = 1$, then the projection of $\mathbf{a}$ on $\mathbf{b}$ is

$$P_{\mathbf{b}}(\mathbf{a}) = (\cos\theta)\mathbf{b}.$$



PROJECTION

Thus another way of expressing the projection formula is

$$\widehat{\mathbf{a}} = (\cos\beta)\widehat{\mathbf{b}} + (\cos\gamma)\widehat{\mathbf{c}}.$$

Here β and γ are the angles between $\mathbf{a}$ and $\mathbf{b}$ and $\mathbf{c}$ respectively, and $\cos\beta$ and $\cos\gamma$ are the corresponding direction cosines.

In the case of $\mathbb{R}^2$, there is already a notion of the angle between two vectors, defined in terms of arclength on a unit circle. Hence the expression $\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta$ is often (especially in physics) taken as definition for the dot product, rather than as definition of angle, as we did here. However, defining $\mathbf{a}\cdot\mathbf{b}$ in this way has the disadvantage that it is not at all obvious that elementary properties such as the identity $(\mathbf{a}+\mathbf{b})\cdot\mathbf{c} = \mathbf{a}\cdot\mathbf{c}+\mathbf{b}\cdot\mathbf{c}$ hold. Moreover, using this as a definition in $\mathbb{R}^n$ has the problem that the angle between two vectors must also be defined. The way to solve this is to use arclength, but this requires bringing in an unnecessary amount of machinery. On the other hand, the algebraic definition is easy to state and remember, and it works for any dimension. The Cauchy-Schwartz inequality, which is valid in $\mathbb{R}^n$, tells us that it possible two define the angle θ between $\mathbf{a}$ and $\mathbf{b}$ via (2.7) to be $\theta := \cos^{-1}(\hat{\mathbf{a}}\cdot\hat{\mathbf{b}})$.

2.1.7 Examples

Let us now consider a couple of typical applications of the ideas we just discussed.

Example 2.2. A film crew wants to shoot a car moving along a straight road with constant speed x km/hr. The camera will be moving along a straight track at y km/hr. The desired effect is that the car should appear to have exactly half the speed of the camera. At what angle to the road should the track be built?

Solution: Let θ be the angle between the road and the track. We need to find θ so that the projection of the velocity vector $\mathbf{v}_R$ of the car on the track is exactly half of the velocity vector $\mathbf{v}_T$ of the camera. Thus

$$\left(\frac{\mathbf{v}_R\cdot\mathbf{v}_T}{\mathbf{v}_T\cdot\mathbf{v}_T}\right)\mathbf{v}_T = \frac{1}{2}\mathbf{v}_T$$

and $\mathbf{v}_R\cdot\mathbf{v}_T = |\mathbf{v}_R||\mathbf{v}_T|\cos\theta$. Now $|\mathbf{v}_R| = x$ and $|\mathbf{v}_T| = y$ since speed is by definition the magnitude of velocity. Thus

$$\frac{xy}{y^2}\cos\theta = \frac{1}{2}$$

Consequently, $\cos\theta = y/2x$. In particular the camera's speed cannot exceed twice the car's speed.

Example 2.3. What we have seen so far can be applied to finding a formula for the distance from a point $\mathbf{v} = (v_1, v_2)^T$ in $\mathbb{R}^2$ to a line $ax + by = c$. Of course this problem can be solved algebraically by considering the line through $(v_1, v_2)^T$ orthogonal to our line. A more illuminating way to proceed, however, is to use projections since they will give a method which can be used in any $\mathbb{R}^n$, whereas it isn't immediately clear how to extend the first method. The way to proceed, then, is to begin by converting the line into a more convenient form. The way we will do this is to choose two points $(x_0, y_0)^T = \mathbf{a}$ and $(x_1, y_1)^T = \mathbf{b}$ on the line. Then the line can also be represented as the set of all points of the form $\mathbf{a}+t\mathbf{c}$, where $\mathbf{c} = \mathbf{b}-\mathbf{a}$. Since distance is invariant under translation, we can replace our original problem with the problem of finding the distance d from $\mathbf{w} = \mathbf{v} - \mathbf{a}$ to the line $t\mathbf{c}$. Since this distance is the length of the component of $\mathbf{w}$ orthogonal to $\mathbf{c}$, we get the formula

$$\begin{aligned} d &= |\mathbf{w} - P_{\mathbf{c}}(\mathbf{w})| \\ &= \left|\mathbf{w} - \left(\frac{\mathbf{w}\cdot\mathbf{c}}{\mathbf{c}\cdot\mathbf{c}}\right)\mathbf{c}\right| \end{aligned}$$

We will give another way to express this distance below.

Example 2.4. Suppose ℓ is the line through $(1,2)^T$ and $(4,-1)^T$. Let us find the distance d from $(0,6)^T$ to ℓ. Since $(4,-1)^T - (1,2)^T = (3,-3)^T$, we may as well take

$$\mathbf{c} = 1/\sqrt{2}(1,-1)^T.$$

We can also take $\mathbf{w} = (0,6)^T - (1,2)^T$, although we could also use $(0,6)^T - (4,-1)^T$. The formula then gives

$$\begin{aligned} d &= \left|(-1,4)^T - \left(\frac{(-1,4)^T\cdot(1,-1)^T}{\sqrt{2}}\right)\frac{(1,-1)^T}{\sqrt{2}}\right| \\ &= \left|(-1,4)^T - \left(\frac{-5}{2}\right)(1,-1)^T\right| \\ &= \left|\left(\frac{3}{2},\frac{3}{2}\right)^T\right| \\ &= \frac{3\sqrt{2}}{2}. \end{aligned}$$

Exercises

Exercise 2.1. Verify the four properties of the dot product on $\mathbb{R}^n$.

Exercise 2.2. Verify the assertion that $\mathbf{b}\cdot\mathbf{c} = 0$ in the proof of Theorem 2.2.

Exercise 2.3. Prove the second statement in the Cauchy-Schwartz inequality that $\mathbf{a}$ and $\mathbf{b}$ are collinear if and only if $|\mathbf{a}\cdot\mathbf{b}| \leq |\mathbf{a}||\mathbf{b}|$.

Exercise 2.4. A nice application of Cauchy-Schwartz is that if $\mathbf{a}$ and $\mathbf{b}$ are unit vectors in $\mathbb{R}^n$ such that $\mathbf{a}\cdot\mathbf{b} = 1$, then $\mathbf{a} = \mathbf{b}$. Prove this.

Exercise 2.5. Show that $P_{\mathbf{b}}(r\mathbf{x} + s\mathbf{y}) = rP_{\mathbf{b}}(\mathbf{x}) + sP_{\mathbf{b}}(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $r, s \in \mathbb{R}$. Also show that $P_{\mathbf{b}}(\mathbf{x})\cdot\mathbf{y} = \mathbf{x}\cdot P_{\mathbf{b}}(\mathbf{y})$.

Exercise 2.6. Prove the vector version of Pythagoras's Theorem. If $\mathbf{c} = \mathbf{a} + \mathbf{b}$ and $\mathbf{a}\cdot\mathbf{b} = 0$, then $|\mathbf{c}|^2 = |\mathbf{a}|^2 + |\mathbf{b}|^2$.

Exercise 2.7. Show that for any $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^n$,

$$|\mathbf{a} + \mathbf{b}|^2 - |\mathbf{a} - \mathbf{b}|^2 = 4\mathbf{a}\cdot\mathbf{b}.$$

Exercise 2.8. Use the formula of the previous problem to prove Proposition 2.1, that is to show that $|\mathbf{a} + \mathbf{b}| = |\mathbf{a} - \mathbf{b}|$ if and only if $\mathbf{a}\cdot\mathbf{b} = 0$.

Exercise 2.9. Prove the law of cosines: If a triangle has sides with lengths $a,\ b,\ c$ and θ is the angle between the sides of lengths a and b, then $c^2 = a^2 + b^2 - 2ab\cos\theta$. (Hint: Consider $\mathbf{c} = \mathbf{b} - \mathbf{a}$.)

Exercise 2.10. Another way to motivate the definition of the projection $P_{\mathbf{b}}(\mathbf{a})$ is to find the minimum of $|\mathbf{a} - t\mathbf{b}|^2$. Find the minimum using calculus and interpret the result.

Exercise 2.11. Orthogonally decompose the vector $(1, 2, 2)$ in $\mathbb{R}^3$ as $\mathbf{p} + \mathbf{q}$ where $\mathbf{p}$ is required to be a multiple of $(3, 1, 2)$.

Exercise 2.12. Use orthogonal projection to find the vector on the line $3x + y = 0$ which is nearest to $(1, 2)$. Also, find the nearest point.

Exercise 2.13. How can you modify the method of orthogonal projection to find the vector on the line $3x + y = 2$ which is nearest to $(1, -2)$?

2.2 Lines and planes.

2.2.1 Lines in $\mathbb{R}^n$

Let's consider the question of representing a line in $\mathbb{R}^n$. First of all, a line in $\mathbb{R}^2$ is cut out by a singlle linear equation $ax + by = c$. But a single equation $ax + by + cz = d$ cuts out a plane in $\mathbb{R}^3$, so a line in $\mathbb{R}^3$ requires at least two equations, since it is the intersection of two planes. For the general case, $\mathbb{R}^n$, we need a better approach. The point is that every line is determined two points. Suppose we want to express the line through $\mathbf{a}$ and $\mathbf{b}$ $\mathbb{R}^n$. Notice that the space curve

$$\mathbf{x}(t) = \mathbf{a} + t(\mathbf{b} - \mathbf{a}) = (1 - t)\mathbf{a} + t\mathbf{b}, \tag{2.8}$$

where t varies through $\mathbb{R}$, has the property that $\mathbf{x}(0) = \mathbf{a}$, and $\mathbf{x}(1) = \mathbf{b}$. As you can see from the Parallelogram Law, this curve traces out the line through $\mathbf{a}$ parallel to $\mathbf{b} - \mathbf{a}$ as in the diagram below.

Equation (2.8) lets us define a line in any dimension. Hence suppose $\mathbf{a}$ and $\mathbf{c}$ are any two vectors in $\mathbb{R}^n$ such that $\mathbf{c} \neq \mathbf{0}$.

Definition 2.4. The *line through* $\mathbf{a}$ *parallel to* $\mathbf{c}$ is defined to be the path traced out by the curve $\mathbf{x}(t) = \mathbf{a} + t\mathbf{c}$ as t takes on all real values. We will refer to $\mathbf{x}(t) = \mathbf{a} + t\mathbf{c}$ as the *vector form* of the line.

In this form, we are defining $\mathbf{x}(t)$ as a vector-valued function of t. The vector form $\mathbf{x}(t) = \mathbf{a} + t\mathbf{c}$ leads directly to *parametric form* of the line. In the parametric form, the components $x_1, \ldots, x_n$ of $\mathbf{x}$ are expressed as linear functions of t as follows:

$$x_1 = a_1 + tc_1,\ x_2 = a_2 + tc_2, \ldots, x_n = a_n + tc_n. \tag{2.9}$$

Letting $\mathbf{a}$ vary while $\mathbf{b}$ is kept fixed, we get the family of all lines of the form $\mathbf{x} = \mathbf{a} + t\mathbf{c}$. Every point of $\mathbb{R}^n$ is on one of these lines, and two lines either coincide or don't meet at all. (The proof of this is an exercise.) We will say that two lines $\mathbf{a} + t\mathbf{c}$ and $\mathbf{a}' + t\mathbf{c}'$ are *parallel* if $\mathbf{c}$ and $\mathbf{c}'$ are collinear. We will also say that the line $\mathbf{a} + t\mathbf{c}$ is parallel to $\mathbf{c}$.

Example 2.5. Let's find an expression for the line in $\mathbb{R}^4$ passing through $(3, 4, -1, 0)$ and $(1, 0, 9, 2)$. We apply the trick in (2.8). Consider

$$\mathbf{x} = (1 - t)(3, 4, -1, 0) + t(1, 0, 9, 2).$$

Clearly, when $t = 0$, $\mathbf{x} = (3, 4, -1, 0)$, and when $t = 1$, then $\mathbf{x} = (1, 0, 9, 2)$. We can also express $\mathbf{x}$ in the vector form $\mathbf{x} = \mathbf{a} + t(\mathbf{b} - \mathbf{a})$ where $\mathbf{a} = (3, 4, -1, 0)$ and $\mathbf{b} = (1, 0, 9, 2)$. The parametric form is

$$x_1 = -2t + 1, \quad x_2 = -4t + 4, \quad x_3 = 10t + 1, \quad x_4 = 2t.$$

Example 2.6. The intersection of two planes in $\mathbb{R}^3$ is a line. Let's show this in a specific example, say the planes are $x+y+z = 1$ and $2x-y+2z = 2$. By inspection, $(1, 0, 0)$ and $(0, 0, 1)$ lie on both planes, hence on the intersection. The line through these two points is $(1 - t)(1, 0, 0) + t(0, 0, 1) = (1 - t, 0, t)$. Setting $x = 1 - t$, $y = 0$ and $z = t$ and substituting this into both plane equations, we see that this line does indeed lie on both planes, hence is in the intersection. But by staring at the equations of the planes (actually by subtracting twice the first equation from the second), we see that every point (x, y, z) on the intersection has $y = 0$. Thus all points on the intersection satisfy $y = 0$ and $x + z = 1$. But any point of this form is on our line, so we have shown that the intersection of the two planes is the line.

Before passing to planes, let us make a remark about the Parallelogram Law for $\mathbb{R}^n$, namely that $\mathbf{a} + \mathbf{b}$ is the vector along the diagonal of the parallelogram with vertices at $\mathbf{0}$, $\mathbf{a}$ and $\mathbf{b}$. This is valid in any $\mathbb{R}^n$, and can be seen by observing (just as we noted for $n = 2$) that the line through $\mathbf{a}$ parallel to $\mathbf{b}$ meets the line through $\mathbf{b}$ parallel to $\mathbf{a}$ at $\mathbf{a} + \mathbf{b}$. We leave this as an exercise.

2.2.2 Planes in $\mathbb{R}^3$

The solution set of a linear equation

$$ax + by + cz = d \tag{2.10}$$

in three variables x, y and z is called a *plane in* $\mathbb{R}^3$. The linear equation (2.10) expresses that the dot product of the vector $\mathbf{a} = (a, b, c)^T$ and the variable vector $\mathbf{x} = (x, y, z)^T$ is the constant d:

$$\mathbf{a} \cdot \mathbf{x} = d.$$

If $d = 0$, the plane passes through the origin, and its equation is said to be *homogeneous*. In this case it is easy to see how to interpret the plane equation. The plane $ax + by + cz = 0$ consists of all $(r, s, t)^T$ orthogonal to $\mathbf{a} = (a, b, c)^T$. For this reason, we call $(a, b, c)^T$ a *normal* to the plane. (On a good day, we are normal to the plane of the floor.)

Example 2.7. Find the plane through $(1,2,3)^T$ with nomal $(2,3,5)^T$. Now $\mathbf{a} = (2,3,5)^T$, so in the equation (2.10) we have $d = (2,3,5)^T \cdot (1,2,3)^T = 23$. Hence the plane is $2x + 3y + 5z = 23$.

Holding $\mathbf{a} \neq \mathbf{0}$ constant and varying d gives a family of planes filling up $\mathbb{R}^3$ such that no two distinct planes have any points in common. Hence the family of planes $ax+by+cz = d$ (a, b, c fixed and d arbitrary) are all *parallel*. By drawing a picture, one can see from the Parallelogram Law that every vector $(r,s,t)^T$ on $ax + by + cz = d$ is the sum of a fixed vector $(x_0, y_0, z_0)^T$ on $ax + by + cz = d$ and an arbitrary vector $(x,y,z)^T$ on the parallel plane $ax + by + cz = 0$ through the origin.

FIGURE

2.2.3 The distance from a point to a plane

A nice application of our projection techniques is to be able to write down a simple formula for the distance from a point to a plane P in $\mathbb{R}^3$. The problem becomes quite simple if we break it up into two cases. First, consider the case of a plane P through the origin, say with equation $ax + by + cz = 0$. Suppose $\mathbf{v}$ is an arbitrary vector in $\mathbb{R}^3$ whose distance to P is what we seek. Now we can decompose $\mathbf{v}$ into orthogonal components where one of the components is along the normal $\mathbf{n} = (a,b,c)^T$, say

$$\mathbf{v} = P_{\mathbf{n}}(\mathbf{v}) + (\mathbf{v} - P_{\mathbf{n}}(\mathbf{v})), \tag{2.11}$$

where

$$P_{\mathbf{n}}(\mathbf{v}) = \big(\frac{\mathbf{v} \cdot \mathbf{n}}{\mathbf{n} \cdot \mathbf{n}}\big)\mathbf{n}.$$

It's intuitively clear that the distance we're looking for is

$$d = |P_{\mathbf{n}}(\mathbf{v})| = |\mathbf{v} \cdot \mathbf{n}|/\sqrt{\mathbf{n} \cdot \mathbf{n}},$$

but we need to check this carefully. First of all, we need to say that the distance from $\mathbf{v}$ to P means the minimum value of $|\mathbf{v} - \mathbf{r}|$, where $\mathbf{r}$ is on P. To simplify notation, put $\mathbf{p} = P_{\mathbf{n}}(\mathbf{v})$ and $\mathbf{q} = \mathbf{v} - \mathbf{p}$. Since $\mathbf{v} = \mathbf{p} + \mathbf{q}$,

$$\mathbf{v} - \mathbf{r} = \mathbf{p} + \mathbf{q} - \mathbf{r}.$$

Since P contains the origin, $\mathbf{q} - \mathbf{r}$ lies on P since both $\mathbf{q}$ and $\mathbf{r}$ do, so by Pythagoras,

$$|\mathbf{v} - \mathbf{r}|^2 = |\mathbf{p}|^2 + |\mathbf{q} - \mathbf{r}|^2.$$

But $\mathbf{p}$ is fixed, so $|\mathbf{v}-\mathbf{r}|^2$ is minimized by taking $|\mathbf{q}-\mathbf{r}|^2 = 0$. Thus $|\mathbf{v}-\mathbf{r}|^2 = |\mathbf{p}|^2$, and the distance $D(\mathbf{v}, P)$ from $\mathbf{v}$ to P is indeed

$$D(\mathbf{v}, P) = |\mathbf{p}| = \frac{|\mathbf{v} \cdot \mathbf{n}|}{(\mathbf{n} \cdot \mathbf{n})^{\frac{1}{2}}} = |\mathbf{v} \cdot \widehat{\mathbf{n}}|.$$

Also, the point on P nearest $\mathbf{v}$ is $\mathbf{q}$. If $\mathbf{v} = (r, s, t)^T$, the distance is

$$D(\mathbf{v}, P) = \frac{|ar + bs + ct|}{\sqrt{a^2 + b^2 + c^2}}.$$

We now attack the general problem by reducing it to the first case. We want to find the distance $D(\mathbf{v}, Q)$ from $\mathbf{v}$ to an arbitrary plane Q in $\mathbb{R}^3$. Suppose the equation of Q is $ax + by + cz = d$, and let $\mathbf{c}$ be a vector on Q. I claim that the distance from $\mathbf{v}$ to Q is the same as the distance from $\mathbf{v}-\mathbf{c}$ to the plane P parallel to Q through the origin, i.e. the plane $ax + by + cz = 0$. Indeed, we already showed that every vector on Q has the form $\mathbf{w}+\mathbf{c}$ where $\mathbf{w}$ is on P. Thus let $\mathbf{r}$ be the vector on Q nearest $\mathbf{v}$. Since $d(\mathbf{v}, \mathbf{r}) = |\mathbf{v}-\mathbf{r}|$, it follows easily from $\mathbf{r} = \mathbf{w} + \mathbf{c}$ that $d(\mathbf{v}, \mathbf{r}) = d(\mathbf{v} - \mathbf{c}, \mathbf{w})$. Hence the problem amounts to minimizing $d(\mathbf{v} - \mathbf{c}, \mathbf{w})$ for $\mathbf{w} \in P$, which we already solved. Thus

$$D(\mathbf{v}, Q) = |(\mathbf{v} - \mathbf{c}) \cdot \widehat{\mathbf{n}}|,$$

which reduces to the formula

$$D(\mathbf{v}, Q) = \frac{|ar + bs + ct - d|}{\sqrt{a^2 + b^2 + c^2}},$$

since

$$\mathbf{c} \cdot \widehat{\mathbf{n}} = \frac{\mathbf{c} \cdot \mathbf{n}}{(\mathbf{n} \cdot \mathbf{n})^{\frac{1}{2}}} = \frac{d}{\sqrt{a^2 + b^2 + c^2}}.$$

In summary, we have

Proposition 2.5. *Let Q be the plane in $\mathbb{R}^3$ defined by $ax+by+cz = d$, and let $\mathbf{v}$ be any vector in $\mathbb{R}^3$, possibly lying on Q. Let $D(\mathbf{v}, Q)$ be the distance from $\mathbf{v}$ to Q. Then*

$$D(\mathbf{v}, Q) = \frac{|ar + bs + ct - d|}{\sqrt{a^2 + b^2 + c^2}}\,.$$

In fact, the problem we just solved has a far more general version known as the least squares problem. We will come back to this topic in a later chapter.

Ii is as an exercises to find a formula for the distance from a point to a line. A more challenging exercise is to find the distance between two lines. If one of the lines is parallel to $\mathbf{a}$ and the other is parallel to $\mathbf{b}$, then it turns out that what is needed is a vector orthogonal to both $\mathbf{a}$ and $\mathbf{b}$. This is the same problem encountered if one wants to find the plane through three non collinear points. What is needed is a vector orthogonal to $\mathbf{q} - \mathbf{p}$ and $\mathbf{r} - \mathbf{p}$. Both of these problems are solved by using the cross product, which we take up in the next section.

Exercises

Exercise 2.14. Express the line $ax + by = c$ in $\mathbb{R}^2$ in parametric form.

Exercise 2.15. Express the line with vector form $(x, y)^T = (1, -1)^T + t(2, 3)^T$ in the form $ax + by = c$.

Exercise 2.16. Find the line through the points $\mathbf{a}$ and $\mathbf{b}$ in the following cases:

(i) $\mathbf{a} = (1, 1, -3)^T$ and $\mathbf{b} = (6, 0, 2)^T$, and

(ii) $\mathbf{a} = (1, 1, -3, 4)^T$ and $\mathbf{b} = (6, 0, 2, -3)^T$.

Exercise 2.17. Prove the Parallelogram Law in $\mathbb{R}^n$ for any n.

Exercise 2.18. Find the line of intersection of the planes $3x - y + z = 0$ and $x - y - z = 1$ in parametric form.

Exercise 2.19. Do the following:

(a) Find the equation in vector form of the line through $(1, -2, 0)^T$ parallel to $(3, 1, 9)^T$.

(b) Find the plane perpendicular to the line of part (a) passing through $(0, 0, 0)^T$.

(c) At what point does the line of part (a) meet the plane of part (b)?

Exercise 2.20. Determine whether or not the lines $(x, y, z)^T = (1, 2, 1)^T + t(1, 0, 2)^T$ and $(x, y, z)^T = (2, 2, -1)^T + t(1, 1, 0)^T$ intersect.

Exercise 2.21. Consider any two lines in $\mathbb{R}^3$. Suppose I offer to bet you they don't intersect. Do you take the bet or refuse it? What would you do if you knew the lines were in a plane?

Exercise 2.22. Use the method of § 2.2.2 to find an equation for the plane in $\mathbb{R}^3$ through the points $(6, 1, 0)^T$, $(1, 0, 1)^T$ and $(3, 1, 1)^T$

Exercise 2.23. Compute the intersection of the line through $(3, -1, 1)^T$ and $(1, 0, 2)^T$ with the plane $ax + by + cz = d$ when

(i) $a = b = c = 1$, $d = 2$,

(ii) $a = b = c = 1$ and $d = 3$.

Exercise 2.24. Find the distance from the point $(1, 1, 1)^T$ to

(i) the plane $x + y + z = 1$, and

(ii) the plane $x - 2y + z = 0$.

Exercise 2.25. Find the orthogonal decomposition $(1,1,1)^T = \mathbf{a}+\mathbf{b}$, where $\mathbf{a}$ lies on the plane P with equation $2x + y + 2z = 0$ and $\mathbf{a} \cdot \mathbf{b} = 0$. What is the orthogonal projection of $(1,1,1)^T$ on P?

Exercise 2.26. Here's another bet. Suppose you have two planes in $\mathbb{R}^3$ and I have one. Furthermore, your planes meet in a line. I'll bet that all three of our planes meet. Do you take this bet or refuse it. How would you bet if the planes were all in $\mathbb{R}^4$ instead of $\mathbb{R}^3$?

Exercise 2.27. Show that two lines in $\mathbb{R}^n$ (any n) which meet in two points coincide.

Exercise 2.28. Verify that the union of the lines $\mathbf{x} = \mathbf{a} + t\mathbf{b}$, where $\mathbf{b}$ is fixed but $\mathbf{a}$ is arbitrary is $\mathbb{R}^n$. Also show that two of these lines are the same or have no points in common.

Exercise 2.29. Verify the Parallelogram Law (in $\mathbb{R}^n$) by computing where the line through $\mathbf{a}$ parallel to $\mathbf{b}$ meets the line through $\mathbf{b}$ parallel to $\mathbf{a}$.

2.3 The Cross Product

2.3.1 The Basic Definition

The cross product of two non parallel vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^3$ is a vector in $\mathbb{R}^3$ orthogonal to both $\mathbf{a}$ and $\mathbf{b}$ defined geometrically as follows. Let P denote the unique plane through the origin containing both $\mathbf{a}$ and $\mathbf{b}$, and let $\mathbf{n}$ be the choice of unit vector normal to P so that the thumb, index finger and middle finger of your right hand can be lined up with the three vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{n}$ without breaking any bones. In this case we call $(\mathbf{a}, \mathbf{b}, \mathbf{n})$ a *right handed triple*. (Otherwise, it's a left handed triple.) Let θ be the angle between $\mathbf{a}$ and $\mathbf{b}$, so $0 < \theta < \pi$. Then we put

$$\mathbf{a} \times \mathbf{b} = |\mathbf{a}||\mathbf{b}| \sin\theta \mathbf{n}. \tag{2.12}$$

If $\mathbf{a}$ and $\mathbf{b}$ are collinear, then we set $\mathbf{a} \times \mathbf{b} = \mathbf{0}$. While this definition is very pretty, and is useful because it reveals the geometric properties of the cross product, the problem is that, as presented, it isn't computable unless $\mathbf{a} \cdot \mathbf{b} = 0$ (since $\sin\theta = 0$). For example, one sees immediately that $|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}||\mathbf{b}| \sin\theta$.

To see a couple of examples, note that $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ and $(\mathbf{i}, -\mathbf{j}, -\mathbf{k})$ both are right handed triples, but $(\mathbf{i}, -\mathbf{j}, \mathbf{k})$ and $(\mathbf{j}, \mathbf{i}, \mathbf{k})$ are left handed. Thus $\mathbf{i} \times \mathbf{j} = \mathbf{k}$, while $\mathbf{j} \times \mathbf{i} = -\mathbf{k}$. Similarly, $\mathbf{j} \times \mathbf{k} = \mathbf{i}$ and $\mathbf{k} \times \mathbf{j} = -\mathbf{i}$. In fact, these examples point out two of the general properties of the cross product:

$$\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a},$$

and

$$(-\mathbf{a}) \times \mathbf{b} = -(\mathbf{a} \times \mathbf{b}).$$

The question is whether or not the cross product is computable. In fact, the answer to this is yes. First, let us make a temporary definition. If $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, put

$$\mathbf{a} \wedge \mathbf{b} = (a_2 b_3 - a_3 b_2, a_3 b_1 - a_1 b_3, a_1 b_2 - a_2 b_1).$$

We call $\mathbf{a} \wedge \mathbf{b}$ the *wedge product* of $\mathbf{a}$ and $\mathbf{b}$. Notice that $\mathbf{a} \wedge \mathbf{b}$ is defined without any restrictions on $\mathbf{a}$ and $\mathbf{b}$. It is not hard to verify by direct computation that $\mathbf{a} \wedge \mathbf{b}$ is orthogonal to both $\mathbf{a}$ and $\mathbf{b}$, so $\mathbf{a} \wedge \mathbf{b} = r(\mathbf{a} \times \mathbf{b})$ for some $r \in \mathbb{R}$.

The key fact is the following

Proposition 2.6. *For all* $\mathbf{a}$ *and* $\mathbf{b}$ *in* $\mathbb{R}^3$,

$$\mathbf{a} \times \mathbf{b} = \mathbf{a} \wedge \mathbf{b}.$$

This takes care of the computability problem since $\mathbf{a} \wedge \mathbf{b}$ is easily computed. An outline the proof goes as follows. The wedge product and the dot product are related by the following identity:

$$|\mathbf{a} \wedge \mathbf{b}|^2 + (\mathbf{a} \cdot \mathbf{b})^2 = (|\mathbf{a}||\mathbf{b}|)^2. \tag{2.13}$$

The proof is just a calculation, and we will omit it. Since $\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta$, and since $\sin\theta \geq 0$, we deduce that

$$|\mathbf{a} \wedge \mathbf{b}| = |\mathbf{a}||\mathbf{b}| \sin\theta. \tag{2.14}$$

It follows that $\mathbf{a} \wedge \mathbf{b} = \pm|\mathbf{a}||\mathbf{b}| \sin\theta \mathbf{n}$. The fact that the sign is $+$ proven by showing that

$$(\mathbf{a} \wedge \mathbf{b}) \cdot \mathbf{n} > 0.$$

The proof of this step is a little tedious so we will omit it.

2.3.2 Some further properties

Before giving applications, we let us give some of the algebraic properties of the cross product.

Proposition 2.7. *Suppose* $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3$. *Then :*

(*i*) $\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a}$,

(*ii*) $(\mathbf{a} + \mathbf{b}) \times \mathbf{c} = \mathbf{a} \times \mathbf{c} + \mathbf{b} \times \mathbf{c}$, *and*

(*iii*) *for any* $r \in \mathbb{R}$,

$$(r\mathbf{a}) \times \mathbf{b} = \mathbf{a} \times (r\mathbf{b}) = r(\mathbf{a} \times \mathbf{b}).$$

Proof. The first and third identities are obvious from the original definition. The second identity, which says that the cross product is distributive, is not at all obvious from the definition. On the other hand, it is easy to check directly that

$$(\mathbf{a} + \mathbf{b}) \wedge \mathbf{c} = \mathbf{a} \wedge \mathbf{c} + \mathbf{b} \wedge \mathbf{c},$$

so (*ii*) has to hold also since $\otimes = \wedge$. □

Recalling that $\mathbb{R}^2$ can be viewed as the complex numbers, it follows that vectors in $\mathbb{R}^1 = \mathbb{R}$ and $\mathbb{R}^2$ can be multiplied, where the multiplication is both associative and commutative. Proposition 2.7 says that the cross product gives a multiplication on $\mathbb{R}^3$ which is distributive, but not commutative. It is in fact anti-commutative. Also, the cross product isn't associative: $(\mathbf{a} \times \mathbf{b}) \times \mathbf{c}$ and $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$ are not in general equal. Instead of the usual associative

law for multiplication, the cross product satisfies a famous identity known as the Jacobi identity:

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) + \mathbf{b} \times (\mathbf{c} \times \mathbf{a}) + \mathbf{c} \times (\mathbf{a} \times \mathbf{b}) = \mathbf{0}.$$

The Jacobi Identity and the anti-commutativity $\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a}$ are the basic axioms for what is called a Lie algebra, which is an important structure in abstract algebra with many applications in mathematics and physics. The next step in this progression of algebras (that is, a product on $\mathbb{R}^4$ is given by the the quaternions, which are fundamental, but won't be considered here.

2.3.3 Examples and applications

The first application is to use the cross product to find a normal $\mathbf{n}$ to the plane P through $\mathbf{p}, \mathbf{q}, \mathbf{r}$, assuming they don't all lie on a line. Once we have $\mathbf{n}$, it is easy to find the equation of P. We begin by considering the plane Q through the origin parallel to P. First put $\mathbf{a} = \mathbf{q} - \mathbf{p}$ and $\mathbf{b} = \mathbf{r} - \mathbf{p}$. Then $\mathbf{a}, \mathbf{b} \in Q$, so we can put $\mathbf{n} = \mathbf{a} \times \mathbf{b}$. Suppose $\mathbf{n} = (a, b, c)^T$ and $\mathbf{p} = (p_1, p_2, p_3)^T$. Then the equation of Q is $ax + by + cz = 0$, and the equation of P is obtained by noting that

$$\mathbf{n} \cdot ((x, y, z)^T - (p_1, p_2, p_3)^T) = 0,$$

or, equivalently,

$$\mathbf{n} \cdot (x, y, z)^T = \mathbf{n} \cdot (p_1, p_2, p_3)^T.$$

Thus the equation of P is

$$ax + by + cz = ap_1 + bp_2 + cp_3.$$

Example 2.8. Let's find an equation for the plane in $\mathbb{R}^3$ through $(1, 2, 1)^T$, $(0, 3, -1)^T$ and $(2, 0, 0)^T$. Using the cross product, we find that a normal is $(-1, 2, 1)^T \times (-2, 3, -1)^T = (-5, -3, 1)^T$. Thus the plane has equation $-5x - 3y + z = (-5, -3, 1)^T \cdot (1, 2, 1)^T = -10$. One could also have used $(0, 3, -1)^T$ or $(2, 0, 0)^T$ on the right hand side with the same result, of course.

The next application is the area formula for a parallelogram.

Proposition 2.8. *Let* $\mathbf{a}$ *and* $\mathbf{b}$ *be two noncollinear vectors in* $\mathbb{R}^3$*. Then the area of the parallelogram spanned by* $\mathbf{a}$ *and* $\mathbf{b}$ *is* $|\mathbf{a} \times \mathbf{b}|$.

We can extend the area formula to 3-dimensional (i.e. solid) parallelograms. Any three noncoplanar vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ in $\mathbb{R}^3$ determine a solid

parallelogram called a parallelepiped. This parallelepiped $\mathcal{P}$ can be explicitly defined as

$$\mathcal{P} = \{r\mathbf{a} + s\mathbf{b} + t\mathbf{c} \mid 0 \leq r, s, t \leq 1\}.$$

For example, the parallelepiped spanned by $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ is the unit cube in $\mathbb{R}^3$ with vertices at $\mathbf{0}$, $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$, $\mathbf{i}+\mathbf{j}$, $\mathbf{i}+\mathbf{k}$, $\mathbf{j}+\mathbf{k}$ and $\mathbf{i}+\mathbf{j}+\mathbf{k}$. A parallelepiped has 8 vertices and 6 sides which are pairwise parallel.

To get the volume formula, we introduce the *triple product* $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$ of $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$.

Proposition 2.9. *Let* $\mathbf{a}$, $\mathbf{b}$ *and* $\mathbf{c}$ *be three noncoplanar vectors in* $\mathbb{R}^3$. *Then the volume of the parallelepiped they span is* $|\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})|$.

Proof. We leave this as a worthwhile exercise. □

By the definition of the triple product,

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = a_1(b_2c_3 - b_3c_2) - a_2(b_3c_1 - b_1c_3) + a_3(b_1c_2 - b_2c_1).$$

The right hand side of this equation is a 3×3 determinant which is written

$$\det \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix}.$$

We'll see somewhat later how the volume of n-dimensional parallelepiped is expressed as the absolute value of a certain $n \times n$ determinant, which is a natural generalization of the triple product.

Example 2.9. We next find the formula for the distance between two lines. Consider two lines ℓ_1 and ℓ_2 in $\mathbb{R}^3$ parameterized as $\mathbf{a}_1 + t\mathbf{b}_1$ and $\mathbf{a}_2 + t\mathbf{b}_2$ respectively. We want to show that the distance between ℓ_1 and ℓ_2 is

$$d = |(\mathbf{a}_1 - \mathbf{a}_2) \cdot (\mathbf{b}_1 \times \mathbf{b}_2)|/|\mathbf{b}_1 \times \mathbf{b}_2|.$$

This formula is somewhat surprising because it says that one can choose any two initial points $\mathbf{a}_1$ and $\mathbf{a}_2$ to compute d. First, let's see why $\mathbf{b}_1 \times \mathbf{b}_2$ is involved. This is in fact intuitively clear, since $\mathbf{b}_1 \times \mathbf{b}_2$ is orthogonal to the directions of both lines. But one way to see this concretely is to take a tube of radius r centred along ℓ_1 and expand r until the tube touches ℓ_2. The point $\mathbf{v}_2$ of tangency on ℓ_2 and the center $\mathbf{v}_1$ on ℓ_1 of the disc (orthogonal to ℓ_1) touching ℓ_2 give the two points so that $d = d(\mathbf{v}_1, \mathbf{v}_2)$, and, by construction, $\mathbf{v}_1 - \mathbf{v}_2$ is parallel to $\mathbf{b}_1 \times \mathbf{b}_2$. Now let $\mathbf{v}_i = \mathbf{a}_i + t_i\mathbf{b}_i$

for $i = 1, 2$, and denote the unit vector in the direction of $\mathbf{b}_1 \times \mathbf{b}_2$ by $\widehat{\mathbf{u}}$. Then

$$\begin{aligned} d &= |\mathbf{v}_1 - \mathbf{v}_2| \\ &= (\mathbf{v}_1 - \mathbf{v}_2) \cdot \frac{(\mathbf{v}_1 - \mathbf{v}_2)}{|\mathbf{v}_1 - \mathbf{v}_2|} \\ &= |(\mathbf{v}_1 - \mathbf{v}_2) \cdot \widehat{\mathbf{u}}| \\ &= |(\mathbf{a}_1 - \mathbf{a}_2 + t_1\mathbf{b}_1 - t_2\mathbf{b}_2) \cdot \widehat{\mathbf{u}}| \\ &= |(\mathbf{a}_1 - \mathbf{a}_2) \cdot \widehat{\mathbf{u}}|. \end{aligned}$$

The last equality is due to the fact that $\mathbf{b}_1 \times \mathbf{b}_2$ is orthogonal to $t_1\mathbf{b}_1 - t_2\mathbf{b}_2$ plus the fact that the dot product is distributive. This is the formula we sought.

For other applications of the cross product, consult *Vector Calculus* by Marsden and Tromba.

Exercises

Exercise 2.30. Using the cross product, find the plane through the origin that contains the line through $(1, -2, 0)^T$ parallel to $(3, 1, 9)^T$.

Exercise 2.31. Using the cross product, find

(a) the line of intersection of the planes $3x+2y-z = 0$ and $4x+5y+z = 0$, and

(b) the line of intersection of the planes $3x+2y-z = 2$ and $4x+5y+z = 1$.

Exercise 2.32. Is $\mathbf{x} \times \mathbf{y}$ orthogonal to $2\mathbf{x} - 3\mathbf{y}$? Generalize this property.

Exercise 2.33. Find the distance from $(1, 2, 1)^T$ to the plane containing $1, 3, 4)^T$, $(2, -2, -2)^T$, and $(7, 0, 1)^T$. Be sure to use the cross product.

Exercise 2.34. Formulate a definition for the angle between two planes in $\mathbb{R}^3$. (Suggestion: consider their normals.)

Exercise 2.35. Find the distance from the line $\mathbf{x} = (1, 2, 3)^T + t(2, 3, -1)^T$ to the origin in two ways:

(i) using projections, and

(ii) using calculus, by setting up a minimization problem.

Exercise 2.36. Find the distance from the point $(1, 1, 1)^T$ to the line $x = 2 + t, y = 1 - t, z = 3 + 2t$,

Exercise 2.37. Show that in $\mathbb{R}^3$, the distance from a point $\mathbf{p}$ to a line $\mathbf{x} = \mathbf{a} + t\mathbf{b}$ can be expressed in the form

$$d = \frac{|(\mathbf{p} - \mathbf{a}) \times \mathbf{b}|}{|\mathbf{b}|}.$$

Exercise 2.38. Prove the identity

$$|\mathbf{a} \times \mathbf{b}|^2 + (\mathbf{a} \cdot \mathbf{b})^2 = (|\mathbf{a}||\mathbf{b}|)^2.$$

Deduce that if $\mathbf{a}$ and $\mathbf{b}$ are unit vectors, then

$$|\mathbf{a} \times \mathbf{b}|^2 + (\mathbf{a} \cdot \mathbf{b})^2 = 1.$$

Exercise 2.39. Show that

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c}.$$

Deduce from this $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$ is not necessarily equal to $(\mathbf{a} \times \mathbf{b}) \times \mathbf{c}$. In fact, can you say when they are equal?

Chapter 3

Linear Equations and Matrices

3.1 Linear equations: the beginning of algebra

The subject of algebra arose from the study of equations. The simplest kind of equations are linear equations, which are equations of the form

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = c,$$

where $a_1, a_2, \ldots a_n$ are a set of numbers called the coefficients, $x_1, x_2, \ldots x_n$ are the variables and c is the constant term. In most familiar situations, the coefficients are real numbers, but in some of the other settings we will encounter later, such as coding theory, the coefficients might be elements of some a finite field. Such considerations will be taken up in later chapters.

The simplest linear equation one can imagine is an equation with only one variable, such as $ax = b$. For example, consider $3x = 4$. This equation is easy to solve since we can express the solution as $x = 3/4$. In general, if $a \neq 0$, then $x = \dfrac{b}{a}$, and this is the only solution. But if $a = 0$ and $b \neq 0$, there is no solution, since the equation is $0 = b$. And in the case where a and b are both 0, every real number x is a solution. This points outs a general property of linear equations. Either there is a unique solution (i.e. exactly one), no solution or infinitely many solutions.

Let's take another example. Suppose you are planning to make a cake using 10 ingredients and you want to limit the cake to 2000 calories. Let a_i be the number of calories per gram of the ith ingredient. Presumably, each a_i is nonnegative, although this problem may eventually be dealt with.

Next, let x_i be the number of grams of the ith ingredient. Then $a_1x_1+a_2x_2+\cdots+a_{10}x_{10}$ is the total number of calories in the recipe. Since you want the total number of calories in your cake to be at most 2000, you could consider the equation $a_1x_1 + a_2x_2 + \cdots + a_{10}x_{10} = 2000$. The totality of possible solutions $x_1, x_2, \ldots, x_{10}$ to this equation is the set of all possible recipes you can concoct with exactly 2000 calories. Decreasing the amount of any ingredient will then clearly decrease the total number of calories. Of course, any solution where some x_i is negative don't have a physical meaning.

A less simple example is the question of finding all common solutions of the equations $z = x^2 + xy^5$ and $z^2 = x + y^4$. Since the equations represent two surfaces in $\mathbb{R}^3$, we would expect the set of common solutions to be a curve. It's impossible to express the solutions in closed form, but we can study them locally. For example, both surfaces meet at $(1, 1, 1)^T$, so we can find the tangent line to the curve of intersection at $(1, 1, 1)^T$ by finding the intersection of the tangent planes of the surfaces at this point. This will at least give us a linear approximation to the curve.

General, nonlinear systems are usually very difficult to solve; their theory involves highly sophisticated mathematics. On the other hand, it turns out that systems of linear equations can be handled much more simply. There are elementary methods for solving them, and modern computers make it possible to handle gigantic linear systems with great speed. A general linear system having of m equations in n unknowns $x_1, \ldots, x_n$ can be expressed in the following form:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{23}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m. \end{aligned} \tag{3.1}$$

Here, all the coefficients a_{ij} and all constants b_i are assumed for now to be real numbers. When all the constants $b_i = 0$, we will call the system *homogeneous*. The main problem, of course, is to find a procedure or algorithm for describing the *solution set* of a linear system as a subset of $\mathbb{R}^n$.

For those who skipped Chapter 1, let us insert a word about notation. A solution of (3.1) is an n-tuple of real numbers, i. e. an element of $\mathbb{R}^n$. By

convention, n-tuples are always written as column vectors. To save space, we will use the notation

$$(u_1, u_2, \ldots, u_n)^T = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

The meaning of the superscript ^{T} will be clarified below. We would also like to point out that a brief summary of the highlights of this chapter may be found in the last section.

3.1.1 The Coefficient Matrix

To simplify notation, we will introduce the coefficient matrix.

Definition 3.1. The *coefficient matrix* of the above linear system is the $m \times n$ array

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{23} & \ldots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}. \tag{3.2}$$

The *augmented coefficient matrix* is the $m \times (n+1)$ array

$$(A|\mathbf{b}) = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{23} & \ldots & a_{2n} & b_2 \\ \vdots & \vdots & \cdots & \vdots & \\ a_{m1} & a_{m2} & \ldots & a_{mn} & b_m \end{pmatrix}. \tag{3.3}$$

In general, an $m \times n$ *matrix* is simply a rectangular $m \times n$ array as in (3.2). When $m = n$, we will say that A is a square matrix of *degree* n.

Now let's look at the strategy for finding solving the system. First of all, we will call the set of solutions the *solution set*. The strategy for finding the solution set is to replace the original system with a sequence of new systems so that each new system has the same solution set as the previous one, hence as the original system.

Definition 3.2. Two linear systems are said to be *equivalent* if they have the same solution sets.

Two equivalent systems have the same number of variables, but don't need to have the same number of equations.

3.1.2 Gaussian reduction

The procedure for solving an arbitrary system is called *Gaussian reduction*. Gaussian reduction is an algorithm for solving an arbitrary system by performing a sequence of explicit operations, called *elementary row operations*, to bring the augmented coefficient matrix $(A|\mathbf{b})$ in (3.3) to a form called *reduced form*, or *reduced row echelon form*. First of all, we define reduced row echelon form.

Definition 3.3. A matrix A is said to be in *reduced row echelon form*, or simply, to be *reduced*, if it has three properties.

(i) The first non zero entry in each row of A is 1.

(ii) The first non zero entry in every row is to the right of the first non zero entry in all the rows above it.

(iii) Every entry above a first non zero entry is zero.

We will call a first non zero entry in a row its *corner entry*. A first non zero entry in a row which has not been made into 1 by a dilation is called the *pivot* of the row. Pivots aren't required to be 1.

For reasons that will be explained later, an $n \times n$ matrix in reduced row echelon form is called the $n \times n$ *identity matrix*. For example,

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Here are some more examples of reduced matrices:

$$\begin{pmatrix} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 5 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 & 0 & 9 \\ 0 & 0 & 0 & 1 & 4 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 3 & 0 & 9 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Notice that the last matrix in this example would be the coefficient matrix of a system in which the variable x_1 does not actually appear. The only variables that occur are $x_2, \ldots, x_5$. Note also that the $n \times n$ identity matrix I_n and every matrix of zeros are also reduced.

3.1.3 Elementary row operations

The strategy in Gaussian reduction is to use a sequence of steps called *elementary row operations* on the rows of the coefficient matrix A to bring A into reduced form. There are three types of elementary row operations defined as follows:

- (Type I) Interchange two rows of A.
- (Type II) Multiply a row of A by a non zero scalar.
- (Type III) Replace a row of A by itself plus a multiple of a different row.

We will call Type I operations *row swaps* and Type II operations *row dilations*. Type III operations are called *transvections*. We will boycott this term. The main result is that an arbitrary matrix A can always be put into reduced form by a sequence of row operations. Before proving this, we will work an example.

Example 3.1. Consider the counting matrix

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

We can row reduce C as follows:

$$C \overset{R_2-4R_1}{\to} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 7 & 8 & 9 \end{pmatrix} \overset{R_3-7R_1}{\to} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & -6 & -12 \end{pmatrix}$$

$$\overset{R_3-2R_2}{\to} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & 0 & 0 \end{pmatrix} \overset{(-1/3)R_2}{\to} \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix} \overset{R_1-2R_2}{\to} \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Notice that we have indicated the row operations.

Proposition 3.1. *Every matrix A can be put into reduced form by some (not unique) sequence of elementary row operations.*

Proof. If $a_{11} \neq 0$, we can make it 1 by the dilation which divides the first row by a_{11}. We can then use row operations to make all other entries in the first column zero. If $a_{11} = 0$, but the first column has a non zero entry somewhere, suppose the first non zero entry is in the ith row. Then swapping the first and ith rows puts a non zero entry in the $(1,1)$ position. Now proceed as above, dividing the new first row by the inverse of the new $(1,1)$ entry, getting a corner entry in the $(1,1)$ position. Now that there we have a corner entry in $(1,1)$, we can use operations of type III to make all elements in the first column below the $(1,1)$ entry 0. If the first column consists entirely of zeros, proceed directly to the second column and repeat the procedure just described on the second column. The only additional step is that if the $(2,2)$ entry is a corner, then the $(1,2)$ entry may be made into 0 by another Type III operation. Continuing in this manner, we will eventually obtain a reduced matrix. □

REMARK: Of course, the steps leading to a reduced form are not unique. Nevertheless, the reduced form of A itself is unique. We now make an important definition.

Definition 3.4. The reduced form of an $m \times n$ matrix A is denoted by A_{red}. The *row rank*, or simply, the *rank* of an $m \times n$ matrix A is the number of non-zero rows in A_{red}.

3.2 Solving Linear Systems

Let A be an $m \times n$ matrix and consider the linear system whose augmented coefficient matrix is $(A|\mathbf{b})$. The first thing is to point out the role of row operations.

3.2.1 Equivalent Systems

Recall that two linear systems are said to be equivalent if they have the same solution sets. The point about Gaussian reduction is that performing a row operation on the augmented coefficient matrix of a linear system gives a new system which is equivalent to the original system.

For example, a row swap, which corresponds to interchanging two equations, clearly leaves the solution set unchanged. Similarly, multiplying the ith equation by a non-zero constant a does likewise, since the original system can be recaptured by multiplying the ith equation by a^{-1}. The only question is whether a row operation of Type III changes the solutions. Suppose the ith equation is replaced by itself plus a multiple k of the jth equation,

where $i \neq j$. Then any solution of the original system is still a solution of the new system. But any solution of the new system is also a solution of the original system since subtracting k times the jth equation from the ith equation of the new system gives us back the original system. Therefore the systems are equivalent.

To summarize this, we state

Proposition 3.2. *Performing a sequence of row operations on the augmented coefficient matrix of a linear system gives a new system which is equivalent to the original system.*

3.2.2 The Homogeneous Case

We still have to find a way to write down the solution set. The first step will be to consider the homogeneous linear system with coefficient matrix $(A|\mathbf{0})$. This is the same as dealing with the coefficient matrix A all by itself.

Definition 3.5. The solution set of a homogeneous linear system with coefficient matrix A is denoted by $\mathcal{N}(A)$ and called the *null space* of A.

The method is illustrated by the following example.

Example 3.2. Consider the homogeneous linear system

$$\begin{aligned} 0x_1 + x_2 + 2x_3 + 0x_4 + 3x + 5 - x_6 &= 0 \\ 0x_1 + 0x_2 + 0x_3 + x_4 + 2x_5 + 0x_6 &= 0. \end{aligned}$$

The coefficient matrix$A$ is already reduced. Indeed,

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

Our procedure will be to solve for the variables in the corners, which we will call the *corner variables*. We will express these variables in terms of the remaining variables, which we will call the *free variables*. In A above, the corner columns are the second and fourth, so x_2 and x_4 are the corner variables and the variables x_1, x_3, x_5 and x_6 are the free variables. Solving gives

$$\begin{aligned} x_2 &= -2x_3 - 3x_5 + x_6 \\ x_4 &= -2x_5 \end{aligned}$$

In this expression, the corner variables are dependent variables since they are functions of the free variables. Now let $(x_1, x_2, x_3, x_4, x_5, x_6)$ denoted

an arbitrary vector in $\mathbb{R}^6$ which is a solution to the system, and let us call this 6-tuple *the general solution vector*. Replacing the corner variables by their expressions in terms of the free variables gives a new expression for the general solution vector involving just the free variables. Namely

$$\mathbf{x} = (x_1, -2x_3 - 3x_5 + x_6,\ x_3,\ -2x_5,\ x_5,\ x_6)^T.$$

The general solution vector now depends only on the free variables, and there is a solution for any choice of these variables.

Using a little algebra, we can compute the vector coefficients of each one of the free variables in $\mathbf{x}$. These vectors are called the *fundamental solutions*. In this example, the general solution vector $\mathbf{x}$ gives the following set of fundamental solutions:

$$\mathbf{f}_1 = (1,0,0,0,0,0)^T,\ \mathbf{f}_2 = (0,0,-2,\ 1,\ 0,\ 0)^T,\ \mathbf{f}_3 = (0,-3,\ 0,\ -2,\ 1,\ 0)^T,$$

and

$$\mathbf{f}_4 = (0,-1,\ 0,\ 0,\ 0,1)^T.$$

Hence the general solution vector has the form

$$\mathbf{x} = x_1\mathbf{f}_1 + x_3\mathbf{f}_2 + x_4\mathbf{f}_3 + x_5\mathbf{f}_4.$$

In other words, the fundamental solutions span the solution space, i.e. every solution is a linear combination of the fundamental solutions.

This example suggest the following

Proposition 3.3. *In an arbitrary homogeneous linear system with coefficient matrix A, any solution is a linear combination of the fundamental solutions, and the number of fundamental solutions is the number of free variables. Moreover,*

$$\#\text{corner variables} + \#\text{free variables} = \#\text{variables}. \tag{3.4}$$

Proof. The proof that every solution is a linear combination of the fundamental solutions goes exactly like the above example, so we will omit it. Equation (3.4) is an obvious consequence of the fact that every variable is either a free variable or a corner variable, but not both. □

There is something strange in Example 3.2. The variable x_1 never actually appears in the system, but it does give a free variable and a corresponding fundamental solution $(1,0,0,0,0,0)^T$. Suppose instead of A the coefficient matrix is

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

Now $(1,0,0,0,0,0)^T$ is no longer a fundamental solution. In fact the solution set is now a subset of $\mathbb{R}^5$. The corner variables are x_1 and x_3, and there are now only three fundamental solutions corresponding to the free variables x_2, x_4, and x_5.

Even though (3.4) is completely obvious, it gives some very useful information. Here is a typical application.

Example 3.3. Consider a system involving 25 variables and assume there are 10 free variables. Then there are 15 corner variables, so the system has to have at least 15 equations, that is, there must be at least 15 linear constraints on the 25 variables.

We can also use (3.4) to say when a homogeneous system with coefficient matrix A has a unique solution (that is, exactly one solution). Now $\mathbf{0}$ is always a solution. This is the so called *trivial solution*. Hence if the solution is to unique, then the only possibility is that $\mathcal{N}(A) = \{\mathbf{0}\}$. But this happens exactly when there are no free variables, since if there is a free variable there will be non trivial solutions. Thus a homogeneous system has a unique solution if and only if every variable is a corner variable, which is the case exactly when the number of corner variables is the number of columns of A. It follows that if a homogeneous system has more variables than equations, there have to be non trivial solutions, since there has to be at least one free variable.

3.2.3 The Non-homogeneous Case

Next, consider the system with augmented coefficient matrix $(A|\mathbf{b})$. If $\mathbf{b} \neq \mathbf{0}$, the system is called *non-homogeneous*. To resolve the non-homogeneous case, we need to observe a result sometimes called the *Super-Position Principle*.

Proposition 3.4. *If a system with augmented coefficient matrix $(A|\mathbf{b})$ has a particular solution $\mathbf{p}$, then any other solution has the form $\mathbf{p} + \mathbf{x}$, where $\mathbf{x}$ varies over all solutions of the associated homogeneous equation. That is, $\mathbf{x}$ varies over $\mathcal{N}(A)$.*

Proof. Let us sketch the proof. (It is quite easy.) Suppose $\mathbf{p} = (p_1, \ldots, p_n)$ and let $\mathbf{x} = (x_1, \ldots, x_n)$ be an element of $\mathcal{N}(A)$. Then substituting $p_i + x_i$ into the system also gives a solution. Conversely, if $\mathbf{q}$ is another particular solution, then $\mathbf{p} - \mathbf{q}$ is a solution to the homogeneous system, i.e. an element of $\mathcal{N}(A)$. Therefore $\mathbf{q} = \mathbf{p} + \mathbf{x}$, where $\mathbf{x} = \mathbf{q} - \mathbf{p} \in \mathcal{N}(A)$. This completes the proof. □

Example 3.4. Consider the system involving the counting matrix C of Example 3.1:

$$\begin{aligned} 1x_1 + 2x_2 + 3x_3 &= a \\ 4x_1 + 5x_2 + 6x_3 &= b \\ 7x_1 + 8x_2 + 9x_3 &= c, \end{aligned}$$

where a, b and c are fixed arbitrary constants. This system has augmented coefficient matrix

$$(C|\mathbf{b}) = \begin{pmatrix} 1 & 2 & 3 & a \\ 4 & 5 & 6 & b \\ 7 & 8 & 9 & c \end{pmatrix}.$$

We can use the same sequence of row operations as in Example 3.1 to put $(C|\mathbf{b})$ into reduced form $(C_{red}|\mathbf{c})$ but to minimize the arithmetic with denominators, we will actually use a different sequence.

$$(C|\mathbf{b}) \overset{R_2-R_1}{\rightarrow} \begin{pmatrix} 1 & 2 & 3 & a \\ 3 & 3 & 3 & b-a \\ 7 & 8 & 9 & c \end{pmatrix} \overset{R_3-2R_2}{\rightarrow} \begin{pmatrix} 1 & 2 & 3 & a \\ 3 & 3 & 3 & b-a \\ 1 & 2 & 3 & c-2b+2a \end{pmatrix} \overset{R_3-R_1}{\rightarrow}$$

$$\begin{pmatrix} 1 & 2 & 3 & a \\ 3 & 3 & 3 & b-a \\ 0 & 0 & 0 & c-2b+a \end{pmatrix} \overset{(-1/3)R_3}{\rightarrow} \begin{pmatrix} 1 & 2 & 3 & a \\ -1 & -1 & -1 & (1/3)a-(1/3)b \\ 0 & 0 & 0 & c-2b+a \end{pmatrix} \overset{R_2+R_1}{\rightarrow}$$

$$\begin{pmatrix} 1 & 2 & 3 & a \\ 0 & 1 & 2 & (4/3)a-(1/3)b \\ 0 & 0 & 0 & c-2b+a \end{pmatrix} \overset{R_1-2R_2}{\rightarrow} \begin{pmatrix} 1 & 0 & -1 & (-5/3)a+(2/3)b \\ 0 & 1 & 2 & (4/3)a-(1/3)b \\ 0 & 0 & 0 & c-2b+a \end{pmatrix}.$$

The reduced system turns out to be the same one we obtained by using the sequence in Example 11.2. We get

$$\begin{aligned} 1x_1 + 0x_2 - 1x_3 &= (-5/3)a + (2/3)b \\ 0x_1 + 1x_2 + 2x_3 &= (4/3)a - (1/3)b \\ 0x_1 + 0x_2 + 0x_3 &= a - 2b + c \end{aligned}$$

Clearly the above system may in fact have no solutions. In fact, from the last equation, we see that whenever $a - 2b + c \neq 0$, there cannot be a solution. Such a system is called *inconsistent*. For a simpler, example, think of three lines in $\mathbb{R}^2$ which don't pass through a common point. This is an example where the system has three equations but only two variables.

Example 3.5. Let's solve the system of Example 3.4 for $a = 1$, $b = 1$ and $c = 1$. In that case, the original system is equivalent to

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

It follows that $x_1 = -1 + x_3$ and $x_2 = 1 - 2x_3$. This represents a line in $\mathbb{R}^3$.

The line if the previous example is parallel to the line of intersection of the three planes

$$\begin{aligned} 1x_1 + 2x_2 + 3x_3 &= 0 \\ 4x_1 + 5x_2 + 6x_3 &= 0 \\ 7x_1 + 8x_2 + 9x_3 &= 0, \end{aligned}$$

These planes meet in a line since their normals are contained in a plane through the origin. On the other hand, when $a - 2b + c \neq 0$, what happens is that the line of intersection of any two of the planes is parallel to the third plane (and not contained in it).

that slightly perturbing the lines will

3.2.4 Criteria for Consistency and Uniqueness

To finish our treatment of systems (for now), we derive two criteria, one for consistency and the other for uniqueness. Consider the $m \times n$ linear system (3.1) with coefficient matrix A and augmented coefficient matrix $(A|\mathbf{b})$.

Proposition 3.5. *Suppose the coefficient matrix A has rank k, that is A_{red} has k corners. Then $\mathcal{N}(A) = \{\mathbf{0}\}$ if and only if $k = n$. The (possibly non-homogeneous) linear system $(A|\mathbf{b})$ is consistent if and only if the rank of A and of $(A|\mathbf{b})$ coincide. If $(A|\mathbf{b})$ is consistent and $k = n$, then the solution is unique. Finally, if A is $n \times n$, the system (3.1) is consistent for all $\mathbf{b}$ if and only if the rank of A equals n.*

Proof. The first statement is a repetition of a result we already proved. The second follows as in the previous example, because if the rank of $(A|\mathbf{b})$ is greater than k, then the last equation amounts to saying $0 = 1$. If A is $n \times n$ of rank n, then it is clear that $(A|\mathbf{b})$ and A have the same rank, namely n. It remains to show that if A and $(A|\mathbf{b})$ have the same rank for all $\mathbf{b}$, then A has rank n. But if the rank of A is less than n, one can (exactly as in

Example 3.4) produce a $\mathbf{b}$ for which $(A|\mathbf{b})$ has rank greater than the rank of A. □

Systems where $m = n$ are an important special case as they are neither under determined (fewer equations than unknowns) nor over determined (more equations than unknowns). When A is $n \times n$ of rank n, the system (3.1) is called *nonsingular*. Thus the nonsingular systems are the square systems which are always consistent and always have unique solutions. We also say that an $n \times n$ matrix *nonsingular* if it has maximal rank n. If the rank of A is less than n, we say that A is *singular*.

Example 3.6. An amusing geometric criterion for a 3×3 matrix

$$A = \begin{pmatrix} \mathbf{a_1} \\ \mathbf{a_2} \\ \mathbf{a_3} \end{pmatrix}$$

to be nonsingular is that

$$\mathbf{a_1} \cdot (\mathbf{a_2} \times \mathbf{a_3}) \neq 0.$$

Indeed, we know that $\mathbf{a_1}$, $\mathbf{a_2}$, and $\mathbf{a_3}$ are not in a plane through the origin if and only if $\mathbf{a_1} \cdot (\mathbf{a_2} \times \mathbf{a_3}) \neq 0$. But the above Proposition also says that the rank of A is three precisely when there is no non-zero vector orthogonal to each of $\mathbf{a_1}$, $\mathbf{a_2}$, and $\mathbf{a_3}$.

The expression $\mathbf{a_1} \cdot (\mathbf{a_2} \times \mathbf{a_3})$ is called the determinant of A and abbreviated $\det(A)$. In algebraic terms, we have

$$\det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \tag{3.5}$$

$$a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}.$$

Determinants will be taken up in a later chapter.

Exercises

Exercise 3.1. Consider the linear system

$$\begin{aligned} x_1 + 2x_2 + 4x_3 + 4x_4 &= 7 \\ x_2 + x_3 + 2x_4 &= 3 \\ x_1 + 0x_2 + 2x_3 + 0x_4 &= 1 \end{aligned}$$

(a) Let A be the coefficient matrix of the associated homogeneous system. Find the reduced form of A.

(b) Determine whether the system is consistent and, if so, find the general solution.

(c) Find the fundamental solutions of $A\mathbf{x} = \mathbf{0}$ and show that the fundamental solutions span $\mathcal{N}(A)$.

(d) Is the system $A\mathbf{x} = \mathbf{b}$ consistent for all $\mathbf{b} \in \mathbb{R}^3$? If not, find an equation which the components of $\mathbf{b}$ must satisfy.

Exercise 3.2. If A is 9×27, explain why the system $A\mathbf{x} = \mathbf{0}$ must have at least 18 fundamental solutions.

Exercise 3.3. Consider the system $A\mathbf{x} = \mathbf{0}$ where $A = \left(\begin{smallmatrix} 1 & -1 & 2 & -1 & 1 \\ -2 & 2 & 1 & -2 & 0 \end{smallmatrix}\right)$. Find the fundamental solutions and show they span $\mathcal{N}(A)$.

Exercise 3.4. Let A be the 2×5 matrix of Problem 3.3. Solve the compounded linear system

$$\left(A \middle| \begin{smallmatrix} 1 & -1 \\ -2 & 0 \end{smallmatrix}\right).$$

Exercise 3.5. Set up a linear system to determine whether $(1, 0, -1, 1)$ is a linear combination of $(-1, 1, 2, 0)$, $(2, 1, 0, 1)$ and $(0, 1, 0, -1)$ with real coefficients. What about when the coefficients are in $\mathbb{Z}_3$? Note that in $\mathbb{Z}_3$, $-1 = 2$.

Exercise 3.6. A baseball team has won 3 mores games at home than on the road, and lost 5 more at home than on the road. If the team has played a total of 42 games, and if the number of home wins plus the number of road losses is 20, determine the number of home wins, road wins, home losses and road losses.

Exercise 3.7. For what real values of a and b does the system

$$\begin{aligned} x + ay + a^2z &= 1 \\ x + ay + abz &= a \\ bx + a^2y + a^2bz &= a^2b \end{aligned}$$

have a unique solution?

Exercise 3.8. True or False: If the normals of three planes in $\mathbb{R}^3$ through the origin lie in a plane through the origin, then then the planes meet in a line.

Exercise 3.9. Suppose A is a 12×15 matrix of rank 12. How many fundamental solutions are there in $\mathcal{N}(A)$?

Exercise 3.10. . How many 2×2 matrices of rank 2 are there if we impose the condition that thhe entries are either 0 or 1? What about 3×3 matrices of rank 3 with the same condition?

Exercise 3.11. Find the ranks of each of the following matrices:

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 4 & 9 \\ 1 & 8 & 27 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 2 \\ 1 & 4 & 4 \\ 1 & 8 & 8 \end{pmatrix}.$$

Can you formulate a general result from your results?

Exercise 3.12. If a chicken and a half lay and egg and a half in a day and a half, how many eggs does a single chicken lay in one day? Bonus marks for relating this to linear equations.

3.3 Matrix Algebra

Our goal in this section is to introduce matrix algebra and to show how it it is closely related to the theory of linear systems.

3.3.1 Matrix Addition and Multiplication

Let $\mathbb{R}^{m \times n}$ denote the set of all $m \times n$ matrices with real entries. There are three basic algebraic operations on matrices. These are addition, scalar multiplication and matrix multiplication. There are conditions which govern when two matrices can be added and when they can be multiplied. In particular, one cannot add or multiply any pair of matrices. First of all,

suppose $A = (a_{ij}) \in \mathbb{R}^{m\times n}$ and r is a scalar. Then the scalar multiple rA of A is the matrix $rA = (ra_{ij}) \in \mathbb{R}^{m\times n}$ in which every element of A has been multiplied by r. For example, if A is 2×3, then

$$3A = \begin{pmatrix} 3a_{11} & 3a_{12} & 3a_{13} \\ 3a_{21} & 3a_{22} & 3a_{23} \end{pmatrix}.$$

Matrix addition can only be carried out on matrices of the same dimension. When A and B have the same dimension, say $m \times n$, we take their sum in the obvious manner. If $A = (a_{ij})$ and $B = (b_{ij})$, then $A + B$ is defined to be the $m \times n$ matrix $A + B := (a_{ij} + b_{ij})$. In other words, the (i, j) entry of $A + B$ is $a_{ij} + b_{ij}$. For example,

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{pmatrix}.$$

Addition and scalar multiplication can be combined in the usual way to give linear combinations of matrices (of the same dimension). Here is an example.

Example 3.7. Let

$$A = \begin{pmatrix} 1 & 1 & 0 & 2 \\ 2 & -4 & 0 & 1 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 2 & 1 & 2 \end{pmatrix}.$$

Then

$$3A = \begin{pmatrix} 3 & 3 & 0 & 6 \\ 6 & -12 & 0 & 3 \end{pmatrix} \qquad \text{and} \qquad A + B = \begin{pmatrix} 2 & 2 & 1 & 2 \\ 2 & -2 & 1 & 3 \end{pmatrix}.$$

The $m \times n$ matrix such that every entry is 0 is called the $m \times n$ *zero matrix*. Clearly, the $m \times n$ zero matrix is an additive identity for addition of $m \times n$ matrices. Now that the additive identity is defined, we can also note that any $m \times n$ matrix A has as an additive inverse $-A$, since $A + (-A) = O$.

3.3.2 Matrices Over $\mathbb{F}_2$: Lorenz Codes and Scanners

So far we have only considered matrices over the real numbers. After we define fields, in the next Chapter, we will be able to compute with matrices over other fields, such as the complex numbers $\mathbb{C}$. Briefly, a field is a set with addition and multiplication which satisfies the basic algebraic properties of the integers, but where we can also divide.

The smallest field is $\mathbb{F}_2$, the integers mod 2, which consists of 0 and 1 with the usual rules of addition and multiplication, except that $1 + 1$ is

defined to be 0: $1 + 1 = 0$. The integers mod 2 are most used in computer science since. (Just look at the on-off switch on a PC.) Adding 1 represents a change of state while adding 0 represents status quo.

Matrices over $\mathbb{F}^2$ are themselves quite interesting. For example, since $\mathbb{F}_2$ has only two elements, there are precisely 2^{mn} such matrices. Addition on such matrices also has some interesting properties, as the following example shows.

Example 3.8. For example,

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

and

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

In the first sum, the parity of every element in the first matrix is reversed. In the second, we see every matrix over $\mathbb{F}_2$ is its own additive inverse.

Example 3.9. Random Key Crypts. Suppose Rocky wants to send a message to Bullwinkle, and he wants to make sure that Boris and Natasha won't be able to learn what it says. Here is what the ever resourceful flying squirrel can do. First he encodes the message as a sequence of zeros and ones. For example, he can use the binary expansions of 1 through 26, thinking of 1 as a, 2 as b etc. Note that $1 = 1, 2 = 10, 3 = 11, 4 = 100 \ldots, 26 = 11010$. Now he represents each letter as a five digit string: $a = 00001$, $b = 00010$, $c = 00011$, and so on, and encodes the message. Rocky now has a long string zeros and ones, which is usually called the *plaintext*. Finally, to make things more compact, he arranges the plaintext into a 01 matrix by adding line breaks at appropriate places. Let's denote this matrix by P, and suppose P is $m \times n$. Now the fun starts. Rocky and Bullwinkle have a list of $m \times n$ matrices of zeros and ones that only they know. The flying squirrel selects one of these matrices, say number 47, and tells Bullwinkle. Let E be matrix number 47. Cryptographers call E the *key*. Now he sends the *ciphertext* $enc_E(P) = P + E$ to Bullwinkle. If only Rocky and Bullwinkle know E, then the matrix P containing the plaintext is secure. Even if Boris and Natasha succeed in overhearing the ciphertext $P + E$, they will still have to know E to find out what P is. The trick is that the key E has to be sufficiently random so that neither Boris nor Natasha can guess it. For example, if E is the all ones matrix, then P isn't very secure since Boris

and Natasha will surely try it. Notice that once Bullwinkle receives the ciphertext, all he has to do is add E and he gets P. For

$$enc_E(P) + E = (P + E) + E = P + (E + E) = P + O = P.$$

This is something even a mathematically challenged moose can do. Our hero's encryption scheme is extremely secure if the key E is sufficiently random and it is only used once. (Such a crypt is called a *one time pad*.) However, if he uses E to encrypt another plaintext message Q, and Boris and Natasha pick up both $enc_E(P) = P + E$ and $enc_E(Q) = Q + E$, then they can likely find out what both P and Q say. The reason for this is that

$$(P + E) + (Q + E) = (P + Q) + (E + E) = P + Q + O = P + Q.$$

The point is that knowing $P + Q$ may be enough for a cryptographer to deduce both P and Q. However, as a one time pad, the random key is very secure (in fact, apparently secure enough for communications on the hot line between Washington and Moscow).

Example 3.10. (**Scanners**) We can also interpret matrices over $\mathbb{F}_2$ in another natural way. Consider a black and white photograph as being a rectangular array consisting of many black and white dots. By giving the white dots the value 0 and the black dots the value 1, our black and white photo is therefore transformed into a matrix over $\mathbb{F}_2$. Now suppose we want to compare two black and white photographs whose matrices A and B have the same dimensions, that is, both are $m \times n$. It turns out to be inefficient for a computer to scan the two matrices to see in how many positions they agree. However, suppose we consider the sum $A + B$. When $A + B$ has a 1 in the (i, j)-component, then $a_{ij} \neq b_{ij}$, and when it has 0, then $a_{ij} = b_{ij}$. Hence the sum two identical photographs will be the zero matrix, and the sum of two complementary photographs will sum to the all ones matrix. An obvious and convenient measure of how similar the two matrices A and B are is the number of non zero entries of $A+B$. This number. which is easily tabulated, is known as the *Hamming distance* between A and B.

3.3.3 Matrix Multiplication

The third algebraic operation, *matrix multiplication*, is the most important and the least obvious to define. For one thing, the product of two matrices of the same dimension is only defined if the matrices are square. The product AB of two matrices A and B exists only when the number of columns of A equals the number of rows of B.

Definition 3.6. Let A be $m \times n$ and B $n \times p$. Then the *product* AB of A and B is the $m \times p$ matrix C whose entry in the ith row and kth column is

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

Thus

$$AB = \big(\sum_{j=1}^{n} a_{ij} b_{jk}\big).$$

Put another way, if the columns of A are $\mathbf{a}_1, \ldots, \mathbf{a}_n$, then the rth column of AB is

$$b_{1r}\mathbf{a}_1 + b_{2r}\mathbf{a}_2 + \ldots b_{nr}\mathbf{a}_n.$$

Hence the rth column of AB is the linear combination of the columns of A using the entries of the rth column of B as the scalars. One can also express AB as a linear comination of the rows of B. This turns out to be connected with row operations. The reader is invited to work this out explicitly.

Example 3.11. Here are two examples.

$$\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 6 & 0 \\ -2 & 7 \end{pmatrix} = \begin{pmatrix} 1 \cdot 6 + 3 \cdot (-2) & 1 \cdot 0 + 3 \cdot 7 \\ 2 \cdot 6 + 4 \cdot (-2) & 2 \cdot 0 + 4 \cdot 7 \end{pmatrix} = \begin{pmatrix} 0 & 21 \\ 4 & 28 \end{pmatrix}.$$

Note how the columns of the product are linear combinations. Computing the product in the opposite order gives a different result:

$$\begin{pmatrix} 6 & 0 \\ -2 & 7 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 6 \cdot 1 + 0 \cdot 2 & 6 \cdot 3 + 0 \cdot 4 \\ -2 \cdot 1 + 7 \cdot 2 & -2 \cdot 3 + 7 \cdot 4 \end{pmatrix} = \begin{pmatrix} 6 & 18 \\ 12 & 22 \end{pmatrix}.$$

From this example, we have a pair of 2×2 matrices A and B such that $AB \neq BA$. More generally, multiplication of $n \times n$ matrices is not commutative, although there is a notable exception: if A and B are 1×1, then $AB = BA$.

3.3.4 The Transpose of a Matrix

Another operation on matrices is *transposition*, or taking the transpose. If A is $m \times n$, the *transpose* A^T of A is the $n \times m$ matrix $A^T := (c_{rs})$, where $c_{rs} = a_{sr}$. This is easy to remember: the ith row of A^T is just the ith column of A. Here are two obvious facts. First,

$$(A^T)^T = A.$$

Second, a matrix and its transpose have the same diagonal. A matrix A which is equal to its transpose (that is, $A = A^T$) is called *symmetric*. Clearly, every symmetric matrix is square. The symmetric matrices over $\mathbb{R}$ turn out to be especially important, as we will see later.

Example 3.12. If

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

then

$$A^T = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}.$$

An example of a 2×2 symmetric matrix is

$$\begin{pmatrix} 1 & 3 \\ 3 & 5 \end{pmatrix}.$$

Notice that the dot product $\mathbf{v} \cdot \mathbf{w}$ of any two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ can be expressed as a matrix product, provided we use the transpose. In fact,

$$\mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T \mathbf{w} = \sum v_i w_i.$$

The transpose of a product has an amusing property:

$$\big(AB\big)^T = B^T A^T.$$

This transpose identity can be seen as follows. The (i, j) entry of $B^T A^T$ is the dot product of the ith row of B^T and the jth column of A^T. Since this is the same thing as the dot product of the jth row of A and the ith column of B, which is the (j, i) entry of AB, and hence the (i, j) entry of $(AB)^T$, we see that $(AB)^T = B^T A^T$. Suggestion: try this out on an example.

3.3.5 The Algebraic Laws

Except for the commutativity of multiplication, the expected algebraic properties of addition and multiplication all hold for matrices. Assuming all the sums and products below are defined, matrix algebra obeys following laws:

(1) **Associative Law**: Matrix addition and multiplication are associative:

$$\big(A + B\big) + C = A + \big(B + C\big) \qquad \text{and} \qquad \big(AB\big)C = A\big(BC\big).$$

(2) **Distributive Law**: Matrix addition and multiplication are distributive:

$$A\big(B+C\big) = AB + AC \qquad \text{and} \qquad \big(A+B\big)C = AC + BC.$$

(3) **Scalar Multiplication Law**: For any scalar r,

$$\big(rA\big)B = A\big(rB\big) = r\big(AB\big).$$

(4) **Commutative Law for Addition**: Matrix addition is commutative: $A + B = B + A$.

Verifying these properties is a routine exercise. I suggest working a couple of examples to convince yourself, if necessary. Though seemingly uninteresting, the associative law for multiplication will often turn to be a very important property.

Recall that the $n \times n$ *identity matrix* I_n is the matrix having one in each diagonal entry and zero in each entry off the diagonal. For example,

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Note that it makes sense to refer to the identity matrix over $\mathbb{F}_2$, since ! is the multiplicative identity of $\mathbb{F}^2$.

The identity matrix I_n is a multiplicative identity for matrix multiplication. More precisely, we have

Proposition 3.6. *If A is an $m \times n$ matrix, then $AI_n = A$ and $I_m A = A$.*

Proof. This is an exercise. □

Exercises

Exercise 3.13. Make up three matrices A, B, C so that AB and BC are defined. Then compute AB and $(AB)C$. Next compute BC and $A(BC)$. Compare your results.

Exercise 3.14. Suppose A and B are symmetric $n \times n$ matrices. (You can even assume $n = 2$.)

(a) Decide whether or not AB is always symmetric. That is, whether $(AB)^T = AB$ for all symmetric A and B?

(b) If the answer to (a) is no, what condition ensures AB is symmetric?

Exercise 3.15. Suppose B has a column of zeros. How does this affect any product of the form AB? What if A has a row or a column of zeros?

Exercise 3.16. Let A be the 2×2 matrix over $\mathbb{F}_2$ such that $a_{ij} = 1$ for each i, j. Compute A^m for any integer $m > 0$. Does this question make sense if $m < 0$? (Note A^j is the product $AA \cdots A$ of A with itself j times.)

Exercise 3.17. Generalize this question to 2×2 matrices over $\mathbb{F}_2 p$.

Exercise 3.18. Let A be the $n \times n$ matrix over $\mathbb{R}$ such that $a_{ij} = 2$ for all i, j. Find a formula for A^j for any positive integer j.

Exercise 3.19. Verify Proposition 3.6 for all $m \times n$ matrices A over $\mathbb{R}$.

Exercise 3.20. Give an example of a 2×2 matrix A such that every entry of A is either 0 or 1 and $A^2 = I_2$ as a matrix over $\mathbb{F}_2$, but $A^2 \neq I_2$ as a matrix over the reals.

3.4 Elementary Matrices and Row Operations

The purpose of this section is make an unexpected connection between matrix multiplication and row operations. We will see that in fact row operations can be done by matrix multiplication. For example, in the $2 \times n$ case, we use the following three types of 2×2 matrices:

$$E_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad E_2 = \begin{pmatrix} r & 0 \\ 0 & 1 \end{pmatrix} \text{ or } \begin{pmatrix} 1 & 0 \\ 0 & r \end{pmatrix}, \qquad E_3 = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \text{ or } \begin{pmatrix} 1 & 0 \\ s & 1 \end{pmatrix}.$$

These matrices enable us to do row operations of types I, II and III respectively via left or pre-multiplication, so they are called *elementary matrices*. For example,

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} c & d \\ a & b \end{pmatrix},$$

$$\begin{pmatrix} r & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} ra & rb \\ c & d \end{pmatrix},$$

and

$$\begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a + sc & b + sd \\ c & d \end{pmatrix}.$$

Suitably modified, the same procedure works in general. An $m \times m$ matrix obtained from I_m by performing a single row operation is called an *elementary $m \times m$ matrix*. Here is the main point.

Proposition 3.7. *Let A be an $m \times m$ matrix, and assume E is an elementary $m \times m$ matrix. Then EA is the matrix obtained by performing the row operation corresponding to E on A.*

Proof. This follows from the fact that

$$EA = (EI_m)A,$$

and EI_m is the result of applying the row operation corresponding to E to I_m. Thus left multiplication by E will do the same thing to A that it does to I_m. □

Since any matrix can be put into reduced form by a sequence of row operations, and since row operations can be performed by left multiplication by elementary matrices, we have

Proposition 3.8. *An arbitrary $m \times n$ matrix A can be put into reduced form by a performing sequence of left multiplications on A using only $m \times m$ elementary matrices.*

Proof. This follows from the above comments. □

This procedure can be expressed as follows: starting with A and replacing it by $A_1 = E_1A$, $A_2 = E_2(E_1A))$ and so forth, we get the sequence

$$A \to A_1 = E_1A \to A_2 = E_2(E_1A) \to \cdots \to E_k(E_{k-1}(\cdots(E_1A)\cdots)),$$

the last matrix being A_{red}. What we obtain by this process is a matrix

$$B = (E_k(E_{k-1}\cdots(E_1A)\cdots))$$

with the property that $BA = A_{red}$. We want to emphasize that although B is expressed as a certain product of elementary matrices, the way we have chosen these matrices is never unique. However, it will turn out that B is unique in certain cases, one of which is the case where A is a nonsingular $n \times n$ matrix.

Note that we could have expressed B without parentheses writing it simply as $B = E_kE_{k-1}\cdots E_1$, due to the fact that, by the associative law, the parens can be rearranged at will.

When we are reducing a matrix A with entries in $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{C}$ or even $\mathbb{F}_2$, then the elementary matrices we need to use also have entries in $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{C}$ or $\mathbb{F}_2$, hence the matrix B which brings A into reduced form also has entries in the corresponding place. Hence we may state

Proposition 3.9. *For any $m \times n$ matrix A (with entries in $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{C}$ or $\mathbb{F}_2$), there is an $m \times m$ matrix B (with entries in $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{C}$ or $\mathbb{F}_2$), which is a product of elementary matrices, such that $BA = A_{red}$.*

Example 3.13. Let's compute the matrix B produced by the sequence of row operations in Example 3.1 which puts the counting matrix C in reduced form. Examining the sequence of row operations, we see that B is the product

$$\begin{pmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1/3 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -7 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -4 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus

$$B = \begin{pmatrix} -5/3 & 2/3 & 0 \\ 4/3 & -1/3 & 0 \\ 1 & -2 & 1 \end{pmatrix}.$$

Be careful to express the product in the correct order. The first row operation is the made by the matrix on the right and the last by the matrix on

the left. Thus

$$BC = \begin{pmatrix} -5/3 & 2/3 & 0 \\ 4/3 & -1/3 & 0 \\ 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

That is, $BC = C_{red}$.

In the above computation, you should not be explicitly multiplying the elementary matrices out. Start at the right and apply the sequence of row operations working to the left. A convenient way of doing this is to begin with the 3×6 matrix $(A|I_3)$ and carry out the sequence of row operations. The final result will be $(A_{red}|B)$. Thus if we start with

$$(A|I_3) = \begin{pmatrix} 1 & 2 & 3 & 1 & 0 & 0 \\ 4 & 5 & 6 & 0 & 1 & 0 \\ 7 & 8 & 9 & 0 & 0 & 1 \end{pmatrix},$$

we end with

$$(A_{red}|B) = \begin{pmatrix} 1 & 0 & -1 & -5/3 & 2/3 & 0 \\ 0 & 1 & 2 & 4/3 & -1/3 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{pmatrix}.$$

3.4.1 Application to Linear Systems

How does this method apply to solving linear systems? Note that the linear system (3.1 can be expressed in the compact matrix form

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

where A is the coefficient matrix, $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ is the column of variables, and $\mathbf{b} = (b_1, b_2, \ldots, b_m)^T$ is the column of constants.

Starting with a system $A\mathbf{x} = \mathbf{b}$, where A is $m \times n$, multiplying this equation by any elementary matrix E gives a new linear system $EA\mathbf{x} = E\mathbf{b}$, which we know is equivalent to the original system. Therefore, applying Proposition 3.9, we obtain

Proposition 3.10. *Given the linear system $A\mathbf{x} = \mathbf{b}$, there exists a square matrix B which is a product of elementary matrices, such that the original system is equivalent to $A_{red}\mathbf{x} = B\mathbf{b}$.*

What's useful is that given E, there exists an elementary matrix F such that $FE = I_m$. It follows (after a little thought) that there exists a square

matrix C such that $CB = I_m$. We will expand on this in the following section.

The advantage of knowing the matrix B which brings A into reduced form is that at least symbolically one can handle an arbitrary number of systems as easily as one. In other words, one can just as easily solve a matrix linear equation $A\mathbf{X} = \mathbf{D}$, where $\mathbf{X} = (x_{ij})$ is a matrix of variables and $\mathbf{D} = (D_{jk})$ is a matrix of constants. If A is $m \times n$ and D has p columns, then X is $n \times p$ and D is $m \times p$. This matrix equation is equivalent to $A_{red}X = BD$.

Exercises

Exercise 3.21. Find the reduced row echelon form for each of the following matrices, which are assumed to be over $\mathbb{R}$:

$$A_1 = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 3 & 1 \\ 1 & 2 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 2 & -1 & 1 \\ 2 & 3 & 1 & 0 \\ 0 & 1 & 2 & 1 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

Exercise 3.22. Repeat Exercise 3.21 for the following matrices, except assume that each matrix is defined over $\mathbb{Z}_2$:

$$C_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \quad C_2 = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \quad C_3 = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

Exercise 3.23. Find matrices B_1, B_2 and B_3 which are products of elementary matrices such that B_iA_i is reduced, where A_1, A_2, A_3 are the matrices of Exercise 1.

Exercise 3.24. Find matrices D_1, D_2 and D_3 defined over $\mathbb{Z}_2$ which are products of elementary matrices such that D_iC_i is reduced, where C_1, C_2, C_3 are the matrices of Exercise 2.

Exercise 3.25. Prove carefully the if E is an elementary matrix and F is the elementary matrix that performs the inverse operation, then $FE = EF = I_n$.

Exercise 3.26. Write down all the 3×3 elementary matrices E over $\mathbb{Z}_2$. For each E, find the matrix F defined in the previous exercise such that $FE = EF = I_3$.

Exercise 3.27. Repeat Exercise 3.26 for the elementary matrices over $\mathbb{Z}_3$.

Exercise 3.28. List all the row reduced 2×3 matrices over $\mathbb{Z}_2$.

3.5 Matrix Inverses

Given an elementary matrix E, we noted in the last section that there exists another elementary matrix F such that $FE = I_m$. A little thought will convince you that not only is $FE = I_m$, but $EF = I_m$ as well. Doing a row operation then undoing it produces the same result as first undoing it and then doing it. Either way you are back to where you started. The essential property is pointed out in the next

Definition 3.7. Suppose two $m \times m$ matrices A and B have the property that $AB = BA = I_m$. Then we say A is an *inverse* of B (and B is an inverse of A).

We will use A^{-1} to denote an inverse of A. In the 2×2 examples above,

$$E_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \Rightarrow E_1^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

$$E_2 = \begin{pmatrix} r & 0 \\ 0 & 1 \end{pmatrix} \Rightarrow E_2^{-1} = \begin{pmatrix} r^{-1} & 0 \\ 0 & 1 \end{pmatrix},$$

and

$$E_3 = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \Rightarrow E_3^{-1} = \begin{pmatrix} 1 & -s \\ 0 & 1 \end{pmatrix}.$$

3.5.1 A Necessary and Sufficient for Existence

Recall that an $n \times n$ matrix A is called invertible if A has rank n. We therefore have the

Proposition 3.11. *If an $n \times n$ matrix A is invertible, there exists an $n \times n$ matrix B such that $BA = I_n$.*

Proof. This follows from Proposition 3.9 since $A_{red} = I_n$. □

This relates the notion of invertibilty say for real numbers with that of invertibilty for matrices. Thus we would like to know when A has a two sided inverse, not just an inverse on the left.

Theorem 3.12. *An $n \times n$ matrix A has an inverse B if and only if A has rank n. Moreover, there can only be one inverse. Finally, if an $n \times n$ matrix A has some left inverse, then A is invertible and the left inverse is the unique inverse A^{-1}.*

Proof. Suppose first that A has two inverses B and C. Then

$$B = BI_n = B(AC) = (BA)C = I_nC = C.$$

Thus t$B = C$, so the inverse is unique. Next, suppose A has a left inverse B. We will show that the rank of A is n. For this, we have to show that if $A\mathbf{x} = \mathbf{0}$, then $\mathbf{x} = \mathbf{0}$. But if $A\mathbf{x} = \mathbf{0}$, then

$$B(A\mathbf{x}) = (BA)\mathbf{x} = I_n\mathbf{x} = \mathbf{0}. \tag{3.6}$$

Thus indeed, A does have rank n. Now suppose A has rank n. Then we know the system $A\mathbf{x} = \mathbf{b}$ has a solution for all $\mathbf{b}$. It follows that there exists an $n \times n$ matrix X so that $AX = I_n$. This follows from knowing the system $A\mathbf{x} = \mathbf{e}_i$ has a solution for each i, where $\mathbf{e}_i$ is the ith column of I_n. Thus there exist $n \times n$ matrices B and X so that $BA = AX = I_n$. We now show that $B = X$. Repeating the above argument, we have

$$B = BI_n = B(AX) = (BA)X = I_nX = X.$$

Thus A has an inverse if and only if it has rank n. To finish the proof, suppose A has a left inverse B: that is B is $n \times n$ and $BA = I_n$. But we just showed in (3.6) that A has rank n, so (as we concluded above), A^{-1} exists and equals B. □

This theorem explains why we call square matrices of maximal rank invertible.

Corollary 3.13. *If A invertible, then the system*

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

has the unique solution $\mathbf{x} = A^{-1}\mathbf{b}$.

Proof. We leave this as an exercise. □

The product of any two invertible $n \times n$ matrices A and B is also invertible. Indeed, $(AB)^{-1} = B^{-1}A^{-1}$. For

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

This is used in the proof of the following useful Proposition.

Proposition 3.14. *Every invertible matrix is a product of elementary matrices.*

The proof is left as an exercise. Of course, by the previous Proposition, the converse is also true: any product of elementary matrices is invertible.

3.5.2 Methods for Finding Inverses

We have two ways of finding the matrix B so that $BA = A_{red}$. The first is simply to multiply out the sequence of elementary matrices which reduces A. This is not as bad as it sounds since multiplying elementary matrices is very easy. The second method is to form the augmented matrix $(A|I_n)$ and row reduce. The final result will be in the form $(I_n|A^{-1})$. This is the method used in most textbooks. Let's begin with an example.

Example 3.14. Suppose we want to find an inverse for

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

Since we only need to solve the matrix equation $XA = I_3$, we can use our previous strategy of row reducing $(A|I_3)$.

$$(A|I_3) = \begin{pmatrix} 1 & 2 & 0 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & -1 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 & 1 \end{pmatrix} \rightarrow$$

$$\begin{pmatrix} 1 & 2 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & 1 & -1 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & -2 & 2 & -1 \\ 0 & 0 & 1 & 1 & -1 & 1 \end{pmatrix} \rightarrow$$

$$\begin{pmatrix} 1 & 0 & 0 & 5 & -4 & 2 \\ 0 & 1 & 0 & -2 & 2 & -1 \\ 0 & 0 & 1 & 1 & -1 & 1 \end{pmatrix}.$$

Hence

$$A^{-1} = B = \begin{pmatrix} 5 & -4 & 2 \\ -2 & 2 & -1 \\ 1 & -1 & 1 \end{pmatrix},$$

since, by construction, $BA = I_3$.

Example 3.15. To take a slightly more interesting example, let

$$A = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

The catch is that we will assume that the entries of A are elements of $\mathbb{F}_2 = \{0, 1\}$. Imitating the above procedure, we obtain that

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

Note that the correctness of this result should be checked by computing directly that

$$I_4 = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

There is somewhat less obvious third technique which is sometimes also useful. If we form the *augmented coefficient matrix* $(A \mid \mathbf{b})$, where $\mathbf{b}$ represents the column vector with components $b_1, b_2, \ldots b_m$ and perform the row reduction of this augmented matrix, the result will be in the form $(I_n \mid \mathbf{c})$, where the components of $\mathbf{c}$ are certain linear combinations of the components of $\mathbf{b}$. The coefficients in these linear combinations give us the entries of A^{-1}. Here is an example.

Example 3.16. Again let

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

Now form

$$\begin{pmatrix} 1 & 2 & 0 & a \\ 1 & 3 & 1 & b \\ 0 & 1 & 2 & c \end{pmatrix}$$

and row reduce. The result is

$$\begin{pmatrix} 1 & 0 & 0 & 5a - 4b + 2c \\ 0 & 1 & 0 & -2a + 2b - c \\ 0 & 0 & 1 & a - b + c \end{pmatrix}.$$

Thus we see the inverse is

$$\begin{pmatrix} 5 & -4 & 2 \\ -2 & 2 & -1 \\ 1 & -1 & 1 \end{pmatrix}.$$

3.5.3 Matrix Groups

In this section, we will give some examples of what are called matrix groups. The basic example of a matrix group is the set $GL(n, \mathbb{R})$ of all invertible elements of $\mathbb{R}^{n\times n}$. Thus,

$$GL(n, \mathbb{R}) = \{A \in \mathbb{R}^{n\times n} \mid A^{-1} \text{ exists}\}. \tag{3.7}$$

Notice that, by definition, every element in $GL(n, \mathbb{R})$ has an inverse. Moreover, I_n is an element of $GL(n, \mathbb{R})$, and if A and B are elements of $GL(n, \mathbb{R})$, then so is their product AB. These three properties define what we mean by a matrix group.

Definition 3.8. A subset G of $\mathbb{R}^{n\times n}$ is called a *matrix group* if the following three conditions hold:

(i) if $A, B \in G$, then $AB \in G$,

(ii) $I_n \in G$, and

(iii) if $A \in G$, then $A^{-1} \in G$.

It turns out that these three axioms are broad enough to give each matrix group an extremely rich structure. Of course, as already noted above, $GL(n, \mathbb{R})$ is a matrix group. In fact, if $G \subset \mathbb{R}^{n\times n}$ is any matrix group, then $G \subset GL(n, \mathbb{R})$ (why?). A subset of $GL(n, \mathbb{R})$ which is also a matrix group is called a *subgroup* of $GL(n, \mathbb{R})$. Thus we want to consider subgroups of $GL(n, \mathbb{R})$. The simplest example of a subgroup of $GL(n, \mathbb{R})$ is $\{I_n\}$: this is the so called *trivial subgroup*.

To get some simple yet interesting examples, let us consider permutation matrices.

Example 3.17 (Permutation Matrices). A matrix P obtained from I_n by a finite sequence of row swaps is called a *permutation matrix*. In other words, a permutation matrix is a matrix $P \in \mathbb{R}^{n\times n}$ such that there are row swap matrices $S_1, \ldots, S_k \in \mathbb{R}^{n\times n}$ for which $P = S_1 \cdots S_k$. (Recall that a row swap matrix is by definition an elementary matrix obtained by interchanging two rows of I_n.) Clearly, I_n is a permutation matrix (why?), and any product of permutation matrices is also a permutation matrix. Thus we only need to see that the inverse of a permutation matrix is also a permutation matrix. Let $P = S_1 \cdots S_k$ be a permutation matrix. Then $S^{-1} = S_k^{-1} \cdots S_1^{-1}$, so P^{-1} is indeed a permutation matrix since $S_i^{-1} = S_i$ for each index i.

Let $P(n)$ denote the set of $n \times n$ permutation matrices. One can also describe $P(n)$ as the set of all matrices obtained form I_n by permuting the rows of I_n. Thus $P(n)$ is the set of all $n \times n$ matrices whose only entries are 0 or 1 such that every row and every column has exactly one non-zero entry. It follows from elementary combinatorics that $P(n)$ has exactly n elements.

The inverse of a permutation matrix has a beautiful expression.

Proposition 3.15. *If P is a permutation matrix, then $P^{-1} = P^T$.*

Proof. This follows from the above discussion. We leave the details for the exercises. □

To give an explicit example, let us compute $P(3)$.

Example 3.18. $P(3)$ consists of the following six 3×3 permutation matrices; namely I_3 and

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Definition 3.9 (The orthogonal group). Let $Q \in \mathbb{R}^{n \times n}$. Then we say that Q is *orthogonal* if and only if $QQ^T = I_n$. The set of all $n \times n$ orthogonal matrices is denoted by $O(n, \mathbb{R})$. We call $O(n, \mathbb{R})$ the *orthogonal group*.

Proposition 3.16. *$O(n, \mathbb{R})$ is a subgroup of $GL(n, \mathbb{R})$.*

Proof. It follows immediately from the definition and Theorem 3.12 that if Q is orthogonal, then $Q^T = Q^{-1}$. Consequently, since $QQ^T = I_n$ implies $Q^TQ = I_n$, whenever Q is orthogonal, so is Q^{-1}. The identity I_n is clearly orthogonal, so it remains to show that the product of two orthogonal matrices is orthogonal. Let Q and R be orthogonal. Then

$$QR(QR)^T = QR(R^TQ^T) = Q(RR^T)Q^T = QQ^T = I_n.$$

□

By Proposition 3.15, we have $P(n) \subset O(n, \mathbb{R})$. That is, every permutation matrix is orthogonal. Hence $P(n)$ is a subgroup of $O(n, \mathbb{R})$.

The condition $Q^TQ = I_n$ which defines an orthogonal matrix Q is equivalent to the property that as a transformation of $\mathbb{R}^n$ to itself, Q preserves inner products. That is, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$Q\mathbf{x} \cdot Q\mathbf{y} = \mathbf{x} \cdot \mathbf{y}. \tag{3.8}$$

Indeed, since $Q^TQ = I_n$,

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T\mathbf{y} = \mathbf{x}^TQ^TQ\mathbf{y} = (Q\mathbf{x})^TQ\mathbf{y} = Q\mathbf{x} \cdot Q\mathbf{y}. \tag{3.9}$$

Conversely, if $Q\mathbf{x} \cdot Q\mathbf{y} = \mathbf{x} \cdot \mathbf{y}$, then $Q^TQ = I_n$. (Just let $\mathbf{x} = \mathbf{e}_i$ and $\mathbf{y} = \mathbf{e}_j$.) In particular, we can now conclude

Proposition 3.17. *Every element of the orthogonal group $O(n, \mathbb{R})$ preserves lengths of vectors and also distances and angles between vectors.*

Proof. This follows from the identity $\mathbf{x} \cdot \mathbf{y} = |\mathbf{x}||\mathbf{y}| \cos\theta$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$. □

The preceding Proposition tells us that the orthogonal group $O(n, \mathbb{R})$ is intimately related to the geometry of $\mathbb{R}^n$. If Q is orthogonal, then the columns of Q are mutually orthogonal unit vectors, which is a fact we will frequently use.

The orthogonal group for $O(2, \mathbb{R})$ is especially interesting. It has an important subgroup $SO(2)$ called the *rotation group* which consists of the rotation matrices

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

Note that $R_0 = I_2$. The fact that $SO(2)$ is a subgroup of $O(2, \mathbb{R})$ follows from trigonometry. For example, the sum formulas for $\cos(\theta + \mu)$ and $\sin(\theta + \mu)$, are equivalent to the geometrically obvious formulas

$$R_\theta R_\mu = R_\mu R_\theta = R_{\theta+\mu} \tag{3.10}$$

for all θ and μ. We will discuss this later that in more detail.

Example 3.19. There are three subgroups of $GL(n, \mathbb{R})$ which we encountered in Chapter 4: the group $\mathcal{D}_n$ of all invertible diagonal matrices (those diagonal matrices with no zeros on the diagonal), the group $\mathcal{L}_n$ of all lower triangular matrices with only ones on the diagonal, and the group $\mathcal{U}_n$ of all upper triangular matrices with ones on the diagonal. In the proof of Theorem 3.19, we actually used the fact that $\mathcal{L}_n$ and $\mathcal{U}_n$ are matrix groups.

Exercises

Exercise 3.29. Find the inverse of each of the following real matrices or show that the inverse does not exist.

(a) $\begin{pmatrix} 1 & 2 \\ 4 & 1 \end{pmatrix}$ (b) $\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & 1 & 0 \end{pmatrix}$ (c) $\begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}$ (d) $\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$.

Exercise 3.30. If the field is $\mathbb{Z}_2$, which of the matrices in Exercise 1 are invertible?

Exercise 3.31. Suppose $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, and assume that $\Delta = ad - bc \neq 0$. Show that $A^{-1} = \frac{1}{\Delta}\left(\begin{smallmatrix} d & -b \\ -c & a \end{smallmatrix}\right)$. What does the condition $\Delta \neq 0$ mean in terms of the rows of A?

Exercise 3.32. Suppose A has an inverse. Find a formula for the inverse of A^T?

Exercise 3.33. Prove Proposition 3.14.

Exercise 3.34. Suppose A is $n \times n$ and there exists a right inverse B, i.e. $AB = I_n$. Show A invertible.

Exercise 3.35. Let $C = \left(\begin{smallmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{smallmatrix}\right)$. Find a general formula for C^{-1}.

Exercise 3.36. Show that if A and B are $n \times n$ and have inverses, then $(AB)^{-1} = B^{-1}A^{-1}$. What is $(ABCD)^{-1}$ if all four matrices are invertible?

Exercise 3.37. Suppose A is invertible $m \times m$ and B is $m \times n$. Solve the equation $AX = B$.

Exercise 3.38. Suppose A and B are both $n \times n$ and AB is invertible. Show that both A and B are invertible. (See what happens if $B\mathbf{x} = \mathbf{0}$.)

Exercise 3.39. Let A and B be two $n \times n$ matrices over $\mathbb{R}$. Suppose $A^3 = B^3$, and $A^2B = B^2A$. Show that if $A^2 + B^2$ is invertible, then $A = B$. (Hint: Consider $(A^2 + B^2)A$.)

Exercise 3.40. Without computing, try to guess the inverse of the matrix

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

(Hint: are the columns orthogonal?)

Exercise 3.41. Is it TRUE or FALSE that if an $n \times n$ matrix with integer entries has an inverse, then the inverse also has integer entries?

Exercise 3.42. Consider the matrix

$$B = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 2 & 1 \end{pmatrix}.$$

Find B^{-1}, if it exists, when B is considered to be a real matrix. Is B invertible when it is considered as a matrix over $\mathbb{Z}_2$?

Exercise 3.43. A real $n \times n$ matrix Q such that $Q^T Q = I_n$ is called *orthogonal*. Find a formula for the inverse of an arbitrary orthogonal matrix Q. Also show that the inverse of an orthogonal matrix is also orthogonal.

Exercise 3.44. Show that the product of two orthogonal matrices is orthogonal.

Exercise 3.45. A matrix which can be expressed as a product of row swap matrices is called a *permutation matrix*. These are the matrices obtained by rearranging the rows of I_n. Show that every permutation matrix is orthogonal. Deduce that if P is a permutation matrix, then $P^{-1} = P^T$.

Exercise 3.46. Show that the following two matrices are permutation matrices and find their inverses:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Exercise 3.47. You are a code-breaker (more accurately, a cryptographer) assigned to crack a secret cipher constructed as follows. The sequence 01 represents A, 02 represents B and so forth up to 26, which represents Z. A space between words is indicated by inserting 00. A text can thus be encoded as a sequence. For example, 190804000209070004150 7 stands for "the big dog". We can think of this as a vector in $\mathbb{R}^{22}$. Suppose a certain text has been encoded as a sequence of length 14,212=44×323, and the sequence has been broken into 323 consecutive intervals of length 44. Next, suppose each sub-interval is multiplied by a single 44×44 matrix C. The new sequence obtained by laying the products end to end is called the cipher text, because

it has now been enciphered, and it is your job to decipher it. Discuss the following questions.

(i) How does one produce an invertible 44×44 matrix in an efficient way, and how does one find its inverse?

(ii) How many of the sub-intervals will you need to decipher to break the whole cipher by deducing the matrix C?

Exercise 3.48. Prove Proposition3.15.

Exercise 3.49. Show that if Q is orthogonal, then the columns of Q are mutually orthogonal unit vectors. Prove that this is also true for the rows of Q.

Exercise 3.50. * Show that every element H of $O(2, \mathbb{R})$ that isn't a rotation satisfies $H^T = H$ and $H^2 = I_2$. (Note I_2 is the only rotation that satisfies these conditions.)

3.6 The $LPDU$ Decomposition

In this section, we will show that every $n \times n$ invertible matrix A has an interesting factorization $A = LPDU$, where each of the matrices L, P, D, U has a particular form. This factorization is frequently used in solving large systems of linear equations by a process known as back substitution. We will not go into back substitution here, but the reader can consult a text on applied linear algebra, e.g. *Linear Algebra and its Applications* by G. Strang. In addition to its usefulness in applied linear algebra, the $LPDU$ decomposition is also of theoretical interest, since each P gives a class of matrices called a Schubert cell, which has many interesting properties.

In order to describe the necessary ingredients L, D, P, and U, we need column operations, pivots and permutation matrices. Of these, the permutation matrices are especially interesting, as we will encounter them in a number of other contexts later.

3.6.1 The Basic Ingredients: L, P, D and U

In the $LPDU$ decomposition, L is lower triangular and has only 1's on its diagonal, P is a permutation matrix, D is a diagonal matrix without any zeros on its diagonal, and U is upper triangular and has only 1's on its diagonal. The notable feature of these types of matrices is that each one can be constructed from just one kind of elementary matrix.

Let's introduce the cast of characters starting with lower triangular matrices. An $n \times n$ matrix $L = (l_{ij})$ is called *lower triangular* if all entries of L strictly above its diagonal are zero. In other words, $l_{ij} = 0$ if $i < j$. Similarly, a matrix is called *upper triangular* if all entries strictly below its diagonal are zero. Clearly, the transpose of a lower triangular matrix is upper triangular and vice versa. We will only be dealing with the upper or lower triangular matrices all of whose diagonal entries are 1's. These matrices are called, respectively, *upper triangular unipotent* and *lower triangular unipotent* matrices.

Example 3.20. Every lower triangular 3×3 unipotent matrix has the form

$$L = \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & c & 1 \end{pmatrix}.$$

The transpose $U = L^T$ is

$$U = \begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix}.$$

We've already used lower triangular unipotent matrices for row reduction, namely, the elementary matrices of Type III which are also lower triangular. For these matrices, exactly one of a, b, c is different from zero. Left multiplication on A by such a matrix replaces a row of A by itself plus a certain multiple of one the rows above it. You can check without any difficulty the basic property that the product of two or more lower triangular elementary matrices of Type III is again lower triangular. Moreover, all the diagonal entries of the product will be ones (why?). More generally, the product of two or more lower triangular matrices is again lower triangular, and the product of two or more lower triangular unipotent matrices is again lower triangular unipotent.

Notice also that if L is lower triangular unipotent, then we can find lower triangular elementary matrices of Type III, say $E_1, E_2, \ldots, E_k$ so that $E_k \cdots E_2 E_1 L = I_n$. Since the inverse of each E_i is another lower triangular elementary matrix of type III, we therefore see that $L = E_1^{-1} E_2^{-1} \cdots E_k^{-1}$. Thus both L and L^{-1} can be expressed as a product of lower triangular elementary matrices of Type III. In particular, the inverse of a lower triangular unipotent matrix is also lower triangular unipotent.

We summarize this discussion in the following proposition:

Proposition 3.18. *The product of two lower triangular unipotent matrices is also lower triangular unipotent, and the inverse of a lower triangular unipotent matrix is a lower triangular unipotent matrix. The corresponding statements in the upper triangular unipotent case also hold.*

What this Proposition says is that the lower (resp. upper) triangular unipotent matrices is closed under the operations of taking products and inverses. Hence they form a pair of matrix groups.

Recall that an $n \times n$ matrix which can be expressed as a product of elementary matrices of Type II (row swaps for short) is called a *permutation matrix*. The $n \times n$ permutation matrices are exactly those matrices which can be obtained by rearranging the rows of I_n. We have already seen that they form a subgroup of $O(n, \mathbb{R})$. In particular, inverse of a permutation matrix P is P^T.

3.6.2 The Main Result

We now come to the main theorem.

Theorem 3.19. *Let $A = (a_{ij})$ be an invertible matrix over $\mathbb{R}$. Then A can be expressed in the form $A = LPDU$, where L is lower triangular unipotent, U is upper triangular unipotent, P is a permutation matrix, and D is a diagonal matrix with all its diagonal entries non zero. Furthermore, the matrices P and D are unique.*

This result gives the invertible matrices an interesting structure. Each of L, P, D, U is constructed from just one kind of elementary matrix. Note that we need to add that every invertible diagonal matrix is a product of Type I elementary matrices. Because of our assertion that the diagonal matrix D is unique, its diagonal entries have quite a bit of significance. The ith diagonal entry d_{ii} of D is usually called the ith *pivot* of A.

Example 3.21. Let $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ be invertible. If $a \neq 0$, then the $LPDU$ decomposition of A is

$$A = \begin{pmatrix} 1 & 0 \\ -c/a & 1 \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & (ad-bc)/a \end{pmatrix} \begin{pmatrix} 1 & -b/a \\ 0 & 1 \end{pmatrix}.$$

However, if $a = 0$, then $bc \neq 0$ and A can be expressed either as

$$LPD = \begin{pmatrix} 1 & 0 \\ d/b & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} c & 0 \\ 0 & b \end{pmatrix}$$

or

$$PDU = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} c & 0 \\ 0 & b \end{pmatrix} \begin{pmatrix} 1 & d/c \\ 0 & 1 \end{pmatrix}.$$

This example tells us that L and U are not necessarily unique.

Proof of Theorem 3.19. This proof has the desirable feature that it is algorithmic. That is, it gives a clear procedure for finding the $LPDU$ factorization. The first step in the proof is to scan down the first column of A until we find the first non zero entry. Such an entry exists since A is invertible. Let $\sigma(1)$ denote the row this entry is in, and let $d_{\sigma(1)} = a_{\sigma(1),1}$ denote the entry itself. Now perform a sequence of row operations to make the entries below $a_{\sigma(1),1}$ equal to zero. This transforms the first column of A into

$$(0, \ldots, 0, d_{\sigma(1)}, 0, \ldots, 0)^T. \tag{3.11}$$

This reduction is performed by pre-multiplying A by a sequence of lower triangular elementary matrices of the third kind. We therefore obtain a lower

triangular unipotent matrix L_1 so that the first column of L_1A has the form (3.11). The next step is to use the non zero entry $d_{\sigma(1)}$ in the first column to annihilate all the entries in the $\sigma(1)$-st row to the right of $d_{\sigma(1)}$. Since post multiplying by elementary matrices performs column operations, we can multiply L_1A on the right by a sequence of upper triangular elementary matrices of the third kind to produce a matrix of the form $(L_1A)U_1$ whose first column has the form (3.11) and whose $\sigma(1)$-st row is

$$(d_{\sigma(1)}, 0, \ldots, 0). \tag{3.12}$$

Moreover, Proposition 3.18 guarantees that U_1 will be upper triangular unipotent. We now have the first column and $\sigma(1)$-st row of A in the desired form and from now on, they will be unchanged.

To continue, we repeat this procedure on the second column of L_1AU_1. Suppose the first non zero entry of the second column sits in the k-th row. Since we already cleared out all non zero entries of the $\sigma(1)$st row, $k \neq \sigma(1)$. Now set $\sigma(2) = k$ and repeat the same procedure we carried out for the first column and $\sigma(1)$st row. Continuing, we eventually obtain lower triangular unipotent matrices L_i and upper triangular unipotent matrices U_i and a rearrangement $\sigma(1), \sigma(2), \ldots, \sigma(n)$ of $1, 2, \ldots, n$ so that

$$(L_nL_{n-1}\cdots L_1)A(U_1U_2\cdots U_n)^{-1} = Q,$$

where Q is the matrix whose ith column has $d_{\sigma(i)}$ as its $\sigma(i)$th entry and zeros in every other entry, where $d_{\sigma(i)}$ is the first non zero entry in the ith column of $(L_{i-1}\cdots L_1)A$. We can clearly factor Q as PD where D is the diagonal matrix whose ith diagonal entry is $d_{\sigma(i)}$ and P is the permutation matrix with ones exactly where Q had non zero entries. This gives us the expression

$$A = (L_nL_{n-1}\cdots L_1)^{-1}PD(U_1U_2\cdots U_n)^{-1}.$$

But this is the desired factorization $A = LPDU$. Indeed, $L' = L_nL_{n-1}\cdots L_1$ is lower triangular unipotent since it is a product of lower triangular elementary matrices of type III, and hence its inverse L is also lower triangular unipotent. The same remarks hold when we put $U = (U_1U_2\cdots U_n)^{-1}$, so we have established the existence of the $LPDU$ factorization. We will leave the proof of the uniqueness of P and D as an exercise. □

Example 3.22. To illustrate the proof, let

$$A = \begin{pmatrix} 0 & 2 & -2 \\ 0 & 4 & -5 \\ -1 & -2 & -1 \end{pmatrix}.$$

Since the first non zero entry in the first column of A is $a_{13} = -1$, the first two steps are to subtract the first column twice from the second and to subtract it once from the third. The result is

$$AU_1U_2 = \begin{pmatrix} 0 & 2 & -2 \\ 0 & 4 & -5 \\ -1 & 0 & 0 \end{pmatrix}.$$

Next we subtract twice the first row from the second, which gives

$$L_1AU_1U_2 = \begin{pmatrix} 0 & 2 & -2 \\ 0 & 0 & -1 \\ -1 & 0 & 0 \end{pmatrix}.$$

Finally, we add the second column to the third and then factor, getting

$$Q = L_1AU_1U_2U_3 = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & -1 \\ -1 & 0 & 0 \end{pmatrix}.$$

Now write

$$Q = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} -1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

Thus we obtain the $LPDU$ factorization

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

3.6.3 The Case $P = I_n$

In the case where $P = I_n$, A can be row reduced without using row interchanges. In fact, in a sense that can be made precise, general invertible matrices do not require a row interchange. This is because a row interchange is only necessary when a zero shows up on a diagonal position during row reduction.

Proposition 3.20. *If an invertible matrix A admits an LDU decomposition, then the matrices L, D and U are all unique.*

Proof. If A has two LDU decompositions, say $A = L_1D_1U_1 = L_2D_2U_2$, then we can write $L_1^{-1}L_2D_2U_2 = D_1U_1$. Hence $L_1^{-1}L_2D_2 = D_1U_1U_2^{-1}$. But in this equation, the matrix on the left hand side is lower triangular and

the matrix on the right hand side is upper triangular. This tells us that immediately that $D_1 = D_2$, and also that $L_1^{-1}L_2 = U_1U_2^{-1} = I_n$ (why?). Hence $L_1 = L_2$ and $U_1 = U_2$, so the proof is completed. □

Going back to the 2×2 case considered in Example 3.21, the LDU decomposition for A is therefore unique when $a \neq 0$. We also pointed out in the same example that if $a = 0$, then L and U are not unique.

If one were only interested in solving a square system $A\mathbf{x} = \mathbf{b}$, then finding the $LPDU$ factorization of A isn't necessary. In fact, it turns out that one can post multiply A by a permutation matrix Q which is concocted to move zero pivots out of the way. That is, if Q is chosen carefully, there exists a factorization $AQ = LDU$. The only affect on the system is to renumber the variables, replacing $\mathbf{x}$ by $Q^{-1}\mathbf{x}$. The L, D, U and Q are no longer unique.

3.6.4 The symmetric LDU decomposition

Suppose A is an invertible symmetric matrix which has an LDU decomposition. Then it turns out that L and U are not only unique, but they are related. In fact, $U = L^T$. This makes finding the LDU decomposition very simple. The reasoning for this goes as follows. If $A = A^T$ and $A = LDU$, then

$$LDU = (LDU)^T = U^T D^T L^T = U^T D L^T$$

since $D = D^T$. Therefore the uniqueness of L, D and U implies that $U = L^T$. The upshot is that to factor $A = LDU$, all one needs is to do row operations of Type III on A such that higher rows act on lower rows to bring A into upper triangular form B. This means all we need to do is to find a lower triangular unipotent matrix L' so that $L'A$ is upper triangular, i.e. $L'A = B$. Then the matrices D and U can be found by inspection. In fact, D is diagonal so $d_{ii} = b_{ii}$ for all i, and since all the b_{ii} are all nonzero, we can write $B = DU$, where U is an upper triangular unipotent matrix. This means $L'A = DU$, so we have found both U and D. Hence L is also known since, by the uniqueness result just proved,

$L = U^T$. Of course, it's also the case that $L = (L')^{-1}$, but this is the hard way to go.

Example 3.23. Consider the symmetric matrix

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

The strategy is to apply Type III row operations, only allowing higher rows to operate on lower rows, to bring A into upper triangular form, which is our DU. Doing so, we find that A reduces to

$$DU = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & -2 \\ 0 & 0 & 1 \end{pmatrix}.$$

Hence

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and

$$U = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus $A = LDU$ where U is as above, $L = U^T$ and $D = \mathrm{diag}(1, 2, 1)$.

Summarizing, we state

Proposition 3.21. *If A is an (invertible) $n \times n$ symmetric matrix without zero pivots, then the $LPDU$ decomposition of A has the form $A = LDL^T$.*

When A is invertible and symmetric but has zero pivots, then a permutation matrix $P \neq I_n$ is needed. Expressing $A = LPDU$, it turns out that we may construct L and U so that $U = L^T$ still holds. This says that PD is also symmetric (why?). Since $P^T = P^{-1}$, we see that

$$PD = (PD)^T = D^T P^T = DP^{-1},$$

so $PDP = D$. I claim two conditions must be fulfilled in order to have this. The first is that since P is a permutation matrix and hence PD is D with its rows permuted, PDP cannot be a diagonal matrix unless $P = P^{-1}$. Since P is a permutation matrix, $P = P^{-1}$ if and only if $P = P^T$. We can therefore conclude that P is a symmetric permutation matrix. Moreover, this tells us that $PD = DP$, so P and D commute. Now look at PDP^{-1}. It can be shown that PDP^{-1} is always a diagonal matrix with the same diagonal entries as D, except in a different order. In fact, let the ith diagonal entry of PDP^{-1} be $d_{\sigma(i)}$. Then σ is the permutation which is determined by P. This gives the second condition; since $PDP^{-1} = D$, the diagonal of D must be left unchanged by the permutation σ. Thus D and P cannot be arbitrary when A is symmetric. We therefore have

Proposition 3.22. *Let A be a symmetric invertible matrix. Then there exists an expression $A = LPDU$ with L, P, D, U as usual and furthermore :*

(i) $U = L^T$,

(ii) $P = P^T = P^{-1}$, *and*

(iii) $PD = DP$.

Conversely, if L, P, D, U satisfy the above three conditions, then $LPDU$ is symmetric.

The next example shows how the disussion preceeding the last Propositon works.

Example 3.24. Let

$$A = \begin{pmatrix} 0 & 2 & 4 & -4 \\ 2 & 4 & 2 & -2 \\ 4 & 2 & -8 & 7 \\ -4 & -2 & 7 & -8 \end{pmatrix}.$$

The first step is to make the (3,1) an (4,1) entries of A zero while keeping A symmetric. This is done by using symmetric row and column operations. That is, we replace A by $A_1 = E_1 A E_1^T$, $A_2 = E_2^T A_1 E_2^T$ etc. Begin with E_1 which differs from I_4 only in the $(3,2)$-entry, which is -2 instead of 0. Computing $E_1 A E_1^T$ gives

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 2 & 4 & -4 \\ 2 & 4 & 2 & -2 \\ 4 & 2 & -8 & 7 \\ -4 & -2 & 7 & -8 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} =$$

$$\begin{pmatrix} 0 & 2 & 0 & -4 \\ 2 & 4 & -6 & -2 \\ 0 & -6 & 0 & 11 \\ -4 & -2 & 7 & -8 \end{pmatrix}.$$

Notice A_1 is symmetric (why?). Next let $A_2 = E_2 A_1 E_2^T$, where E_2 is obtained from I_4 by adding twice the second row to the fourth row. The result is

$$A_2 = \begin{pmatrix} 0 & 2 & 0 & 0 \\ 2 & 4 & -6 & 6 \\ 0 & -6 & 0 & -1 \\ 0 & 6 & -1 & 0 \end{pmatrix}.$$

The next step is to remove the 4 in the $(2,2)$-position of A_2. This is done by symmetric elimination. We subtract the first row from the second row and the first column from the second column. This gives

$$A_3 = \begin{pmatrix} 0 & 2 & 0 & 0 \\ 2 & 0 & -6 & 6 \\ 0 & -6 & 0 & -1 \\ 0 & 6 & -1 & 0 \end{pmatrix}.$$

It is easy to see how to finish. After two more eliminations, we end up with

$$PD = \begin{pmatrix} 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{pmatrix}.$$

Hence we get

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

and

$$D = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

We leave it as an exercise to find L and hence U.

Exercises

Exercise 3.51. Find the $LPDU$ decompositions of the following matrices:

$$\begin{pmatrix} 0 & 1 & 1 \\ 2 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 3 \\ 0 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & -1 \\ 1 & -1 & 0 \end{pmatrix}.$$

Exercise 3.52. Find the inverse of

$$\begin{pmatrix} 1 & a & b & c \\ 0 & 1 & d & e \\ 0 & 0 & 1 & f \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Exercise 3.53. Show directly that an invertible upper triangular matrix B can be expressed $B = DU$, where D is a diagonal matrix with non zero diagonal entries and U is upper an triangular matrix all of whose diagonal entries are ones. Is this still true if B is singular?

Exercise 3.54. Show that the product of any two lower triangular matrices is lower triangular. Also show that the inverse of a lower triangular invertible matrix is lower triangular. What are the diagonal entries of the inverse?

Exercise 3.55. Let A be a 3×3 lower triangular unipotent matrix. Find a formula expressing A as a product of lower triangular elementary matrices of type III.

Exercise 3.56. Find the $LPDU$ decomposition of

$$\begin{pmatrix} 0 & 1 & 2 & 1 \\ 1 & 1 & 0 & 2 \\ 2 & 0 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix}.$$

Exercise 3.57. Find the LDU decomposition of

$$\begin{pmatrix} 1 & 1 & 2 & 1 \\ 1 & -1 & 0 & 2 \\ 2 & 0 & 0 & 1 \\ 1 & 2 & 1 & -1 \end{pmatrix}.$$

Exercise 3.58. Prove that in any $LPDU$ decomposition, P and D are unique. (Let $LPDU = L'P'D'U'$ and consider $L^{-1}L'P'D' = PDUU'^{-1}$.)

Exercise 3.59. Find a 3×3 matrix A such that the matrix L in the $A = LPDU$ decomposition isn't unique.

Exercise 3.60. Let A be the matrix of Example 3.24. Find the matrices L and U. Also, show that $PD = DP$.

3.7 Summary

This chapter was an introduction to linear systems and matrices. We began by introducing the general linear system of m equations in n unknowns with real coefficients. There are two types of systems called homogeneous and non-homogeneous according to whether the constants on the right hand sides of the equations are all zeros or not. The solutions make up the solution set. If the system is homogeneous, the solution set is a subspace of $\mathbb{R}^n$. In order to write the system in a compact form, we introduced the coefficient matrix for homogeneous systems and the augmented coefficient matrix for non-homogeneous systems. We then wrote down the three row operations of Gaussian reduction. The row operations give a specific set of rules for bringing the coefficient matrix and augmented coefficient matrix into a normal form known as reducued form. The point is that performing a row operation on the coefficient matrix (or augmented coefficient matrix) gives a new coefficient matrix (or augmented coefficient matrix) whose associated linear system has exactly the same solution space (or set in the non-homogeneous case).

After a matrix is put into reducued form, we can read off its rank (the number of non-zero rows). We then obtained criteria which are necessary and sufficient for the existence and uniqueness of solutions. A non-homogeneous system has a solution if and only if its augmented coefficient matrix and coefficient matrix have the same rank. A unique solution exists if and only if the augmented coefficient matrix and coefficient matrix have the same rank and the rank is the number of unknowns.

We next introduced matrix algebra, addition and multiplication. Matrices of the same size can always be added but to form AB, the number of rows of B must be the same as the number of columns of A. We saw how elementary matrices perform row opertions, so that matrices can be row reduced by multiplication. This lead to the notion of the inverse of an $n \times n$ matrix A, a matrix B such that $AB = BA = I_n$. We saw $BA = I_n$ is enough to guarantee $AB = I_n$, and also, the invertible $n \times n$ matrices are exactly those of rank n. A key fact is that a square linear system $A\mathbf{x} = \mathbf{b}$ with A invertible has unique solution $\mathbf{x} = A^{-1}\mathbf{b}$.

We then introduced matrix groups and gave several examples. After that, we discussed a way of factoring an invertible matrix as $LPDU$. This is an often used method both in applied mathematics in solving large systems and in pure mathematics in the study of matrix groups.

Chapter 4

Fields and vector spaces

4.1 Elementary Properties of Fields

4.1.1 The Definition of a Field

In the previous chapter, we noted unecessarily that one of the main concerns of algebra is the business of solving equations. Beginning with the simplest, most trivial equation, the equation $ax = b$, we see that there is a subtle point. We are used to considering equations which involve integers. To divide three apples among 4 persons, we have to consider the equation $4x = 3$. This doesn't have an integer solution. More generally, we obviously cannot function without all quotients p/q, where p, q are integers and $q \neq 0$. The set of all such quotients is called the the set of *rational numbers*. We will denote them by $\mathbb{Q}$. Recall that addition and multiplication in $\mathbb{Q}$ is defined by:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}, \tag{4.1}$$

and

$$\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}. \tag{4.2}$$

Clearly the sum and product of two rational numbers is another rational number.

Next suppose we have been given the less trivial job of solving two linear equations in two unknowns, say x an y. These equations might be written out as

$$\begin{aligned} ax + by &= m \\ cx + dy &= n. \end{aligned}$$

Assuming $ad - bc \neq 0$, there is a unique solution which is expressed as

$$\begin{aligned} x &= \frac{dm - bn}{ad - bc} \\ y &= \frac{-cm + an}{ad - bc}. \end{aligned}$$

(In fact we derived this in the previous chapter.) The main point of this is that to express the solutions of such a linear system, one needs all the available algebraic operations: addition, subtraction, multiplication and division. A set, such as the rationals or reals, where all these operations exist is called a field.

Before defining the notion of a field, we need to define the notion of a *binary operation* on a set. Addition and multiplication on the set of integers, $\mathbb{Z}$, are two basic examples of binary operations. Let S be any set, finite or infinite. Recall that the Cartesian product of S with itself is the set $S \times S$ of all ordered pairs (x, y) of elements $x, y \in S$. Note, we call (x, y) an ordered pair since $(x, y) \neq (y, x)$ unless $x = y$. Thus,

$$S \times S = \{(x, y) \mid x, y \in S\}.$$

Definition 4.1. A *binary operation* on S is a function $F : S \times S \to S$, that is, a function F whose domain is $S \times S$ which takes its values $F(x, y)$ in S.

Note: when A and B are sets, we will write $F : A \to B$ to indicate that F is a function with domain A and values in B. Also, we often express a binary operation by writing something like $x \cdot y$ or $x * y$ for $F(x, y)$. So, for example, the operation of addition on $\mathbb{Z}$ may be thought of as being a binary operation $+$ on $\mathbb{Z}$ such that $+(x, y) = x + y$.

We now define the notion of a field.

Definition 4.2. Assume given a set $\mathbb{F}$ with two binary operations called addition and multiplication. The sum and product of two elements $a, b \in \mathbb{F}$ will be denoted by $a + b$ and ab respectively. Suppose addition and multiplication satisfy the following properties:

(i) $a + b = b + a$ (addition is commutative);

(ii) $(a + b) + c = a + (b + c)$ (addition is associative);

(iii) $ab = ba$ (multiplication is commutative);

(iv) $a(bc) = (ab)c$ (multiplication is associative);

(v) $a(b+c) = ab + ac$ (multiplication is distributive);

(vi) $\mathbb{F}$ contains an additive identity 0 and a multiplicative identity 1 distinct from 0; the additive and multiplicative identities have the property that $a + 0 = a$ and $1a = a$ for every $a \in \mathbb{F}$;

(vii) for every $a \in \mathbb{F}$, there is an element $-a$ called the *additive inverse* of a such that $a + (-a) = 0$; and

(viii) for every $a \neq 0$ in $\mathbb{F}$, there is an element a^{-1}, called the *multiplicative inverse* of a such that $aa^{-1} = 1$.

Then $\mathbb{F}$ is called a *field*.

Note that we will often express $a+(-b)$ as $a-b$. In particular, $a-a = 0$. In any field $\mathbb{F}$, $a0 = 0$ for all a. For

$$a0 = a(0+0) = a0 + a0,$$

so adding $-a0$ to both sides and using the associativity of addition, we get

$$0 = a0 - a0 = (a0 + a0) - a0 = a0 + (a0 - a0) = a0 + 0 = a0.$$

Hence $a0 = 0$ for all $a \in \mathbb{F}$.

Using this fact, we next show

Proposition 4.1. *In any field $\mathbb{F}$, whenever $ab = 0$, either a or b is zero. Put another way, if neither a nor b is zero, then $ab \neq 0$.*

Proof. Suppose $a \neq 0$ and $b \neq 0$. If $ab = 0$, it follows that

$$0 = a^{-1}0 = a^{-1}(ab) = (a^{-1}a)b = 1b = b.$$

This is a contradiction, so $ab \neq 0$. □

The conclusion that $ab = 0$ implies either a or b is zero is one of the field properties that is used repeatedly. We also have

Proposition 4.2. *In any field $\mathbb{F}$, the additive and multiplicative identities are unique. Moreover, the additive and multiplicative inverses are also unique.*

Proof. We will show 0 is unique. The proof that 1 is unique is similar. Let 0 and $0'$ be two additive identities. Then

$$0' = 0' + 0 = 0$$

so 0 is indeed unique. We next show additive inverses are unique. Let $a \in \mathbb{F}$ have two additive inverses b and c. Using associativity, we see that

$$b = b + 0 = b + (a + c) = (b + a) + c = 0 + c = c.$$

Thus $b = c$. The rest of the proof is similar. □

4.1.2 Examples

We now give some examples.

First of all, it's easy to see that the rational numbers satisfy all the field axioms, so $\mathbb{Q}$ is a field. In fact, verifying the field axioms for $\mathbb{Q}$ simply boils down to the basic arithmetic properties of the integers: associativity, commutativity and distributivity and the existence of 0 and 1. Indeed, all one needs to do is to use (4.1) and (4.2) to prove the field axioms for $\mathbb{Q}$ from these properties of the integers.

The integers $\mathbb{Z}$ are not a field, since field axiom (viii) isn't satisfied by $\mathbb{Z}$. Indeed, the only integers which have multiplicative inverses are ± 1.

The second example of a field is the set of real numbers $\mathbb{R}$. The construction of the real numbers is actually somewhat technical, so we will skip it. For most purposes, it suffices to think of $\mathbb{R}$ as being the set of all decimal expansions

$$a_1 a_2 \cdots a_r . b_1 b_2 \cdots ,$$

where all a_i and b_j are integers between 0 and 9. Note that there can be infinitely many b_j to the right of the decimal point. We also have to make appropriate identifications for repeating decimals such as $1 = .999999\ldots$. A very useful fact is that $\mathbb{R}$ is ordered; that is, any real number x is either positive , negative or 0, and the product of two numbers with the same sign is positive. This makes it possible to solve systems of linear inequalities such as $a_1 x_1 + a_2 x_2 + \cdots + a_n x_n > c$. In addition, the reals have what is called the *Archimedian property*: if $a, b > 0$, then there exists an $x > 0$ so that $ax > b$.

The third basic field is $\mathbb{C}$, the field of complex numbers. This is a very important field. We will discuss it in the next section.

4.1.3 An Algebraic Number Field

Many examples of fields arise by extending an already given field. We will now give an example of a field called an algebraic number field which is obtained by adjoining the square root of an integer to the rationals $\mathbb{Q}$. Let us first recall the

Theorem 4.3 (Fundamental Theorem of Arithmetic). *Let m be an integer greater than 1. Then m can be factored $m = p_1 p_2 \cdots p_k$, where $p_1, p_2, \ldots, p_k$ are primes. Moreover, this factorization is unique up to the order of the factors.*

Recall that a positive integer p is called *prime* if $p > 1$ and its only positive factors are 1 and itself. For a proof of the Fundamental Theorem of Arithmetic, the reader is referred to a text on elementary number theory. We say that a positive integer m is *square free* if its prime factorization has no repeated factors. For example, $10 = 2 \cdot 5$ is square free while $12 = 4 \cdot 3$ isn't.

Let $m \in \mathbb{Z}$ be positive and square free, and let $\mathbb{Q}(\sqrt{m})$ denote the set of all real numbers of the form $a + b\sqrt{m}$, where a and b are arbitrary rational numbers. It is easy to see that sums and products of elements of $\mathbb{Q}(\sqrt{m})$ give elements of $\mathbb{Q}(\sqrt{m})$. Clearly 0 and 1 are elements of $\mathbb{Q}(\sqrt{m})$. Hence, assuming the field axioms for $\mathbb{R}$ allows us to conclude without any effort that all but one of the field axioms are satisfied in $\mathbb{Q}(\sqrt{m})$. We still have to prove that any non zero element of $\mathbb{Q}(\sqrt{m})$ has a multiplicative inverse.

So assume $a + b\sqrt{m} \neq 0$. Thus at least one of a or b is non zero. By clearing away the denominators, we can assume the a and b are integers (why?). Furthermore, we can assume they don't have any common prime factors; that is, a and b are *relatively prime*. (This will also mean that both a and b are non zero.) The trick is to notice that

$$(a + b\sqrt{m})(a - b\sqrt{m}) = a^2 - mb^2.$$

Hence

$$\frac{1}{a + b\sqrt{m}} = \frac{a - b\sqrt{m}}{a^2 - mb^2}.$$

Thus, if $a^2 - mb^2 \neq 0$, then $(a + b\sqrt{m})^{-1}$ exists in $\mathbb{R}$ and is an element of $\mathbb{Q}(\sqrt{m})$.

To see that indeed $a^2 - mb^2 \neq 0$, suppose to the contrary. Then

$$a^2 = mb^2.$$

But this implies that m divides a^2, hence any prime factor p_i of m has to divide a itself. In other words, m divides a. Given that, we may cancel m on both sides and get an equation

$$cm = b^2,$$

where c is an integer. Repeating the argument, any prime factor of m has to divide b^2, hence b. The upshot of this is that the original assumption that a and b had no common factor has been violated, so the equation $a^2 = mb^2$ is impossible. Therefore we have proven

Proposition 4.4. *If m is a square free positive integer, then $\mathbb{Q}(\sqrt{m})$ is a field.*

The field $\mathbb{Q}(\sqrt{m})$ is in fact the smallest field containing both the rationals $\mathbb{Q}$ and $\sqrt{m}$.

4.1.4 The Integers Modulo p

The integers modulo p form a class of fields which you may find surprising. Start with any prime number $p > 1$, and let

$$\mathbb{F}_p = \{0, 1, 2, \ldots p-1\}.$$

Using modular arithmetic, we will make $\mathbb{F}_p$ into a field in a natural way. Modular arithmetic can succinctly be described as adding and multiplying using only remainders.

Before that, however, let's look at he most important and simple example, the case $p = 2$. Recall that we already considered this field in the previous chapter. The field $\mathbb{F}_2 = \{0, 1\}$ consists simply of the additive and multiplicative identities. Addition and multiplication for $\mathbb{F}_2$ are determined for us by the field axioms. I claim we have to have the following:

$$0 + 0 = 0, \quad 0 + 1 = 1 + 0 = 1, \quad 1 + 1 = 0, \quad 0 \cdot 1 = 1 \cdot 0 = 0, \quad 1 \cdot 1 = 1.$$

Indeed, since $0 \neq 1$, $1 + 1 \neq 1$, so $1 + 1 = 0$.

Proposition 4.5. *$\mathbb{F}_2$ is a field.*

Proof. One would have to check by going through all possible cases, that the 8 field axioms are true. But we can also obtain $\mathbb{F}_2$ via arithmetic modulo 2, as we will see below. □

The reason $\mathbb{F}_2$ is so useful is that 0 and 1 represent the states on and off and adding 1 causes a change of the state. It also plays an important role

in coding theory and information theory, two disciplines that have arisen because of the electronic revolution. We will study linear coding theory in some detail later.

Now suppose p is any prime number greater than one. Let us see how to make $\mathbb{F}_p = \{0, 1, 2, \ldots, p-1\}$ into a field. We have to define addition and multiplication so that all eight field axioms are satisfied. To add two elements a and b in $\mathbb{F}_p$, first take their sum in the usual way to get the integer $a + b$. If $a + b < p$, then we define their sum in $\mathbb{F}_p$ to be $a + b$. However, if $a + b \geq p$, we need to use *division with remainder*. This is the principle which says that if a and p are non-negative integers with $p \neq 0$, then one can uniquely express a as $a = qp + r$, where q is a non-negative integer and $0 \leq r < p$. (For division with remainder, we don't need that p is prime.)

Thus, if $a + b \geq p$, write

$$a + b = qp + r,$$

where q is a nonnegative integer and r is an integer such that $0 \leq r < p$. Then the sum of a and b in $\mathbb{F}_p$ is defined to be r. This operation is called *modular addition*. To multiply a and b in $\mathbb{F}_p$, we use, in an entirely similar fashion, the remainder upon dividing ab by p. You should check that the construction of $\mathbb{F}_2$ forced upon us above agrees with our definition here.

The next example we look at is $\mathbb{F}_3$, which has three elements.

Example 4.1. Let us construct $\mathbb{F}_3$. Of course, 0 and 1 are always the identities. Now $\mathbb{F}_3 = \{0, 1, 2\}$, so to completely determine the addition, we only have to define $1 + 1$, $1 + 2$ and $2 + 2$. First of all, $1 + 1 = 2$. To find $2 + 2$, first take the ordinary sum 4, then divide 4 by 3. Since the remainder is 1, $2 + 2 = 1$ in $\mathbb{F}_3$. Similarly, $1 + 2 = 0$ in $\mathbb{F}_3$. Thus $-2 = 1$ and $-1 = 2$. To find all products, it is actually sufficient to just find $2 \cdot 2$ (why?). But $2 \cdot 2 = 4$ in usual arithmetic, so $2 \cdot 2 = 1$ in $\mathbb{F}_3$. Thus $2^{-1} = 2$. A good way to describe addition and multiplication in $\mathbb{F}_3$ (or more generally any $\mathbb{F}_p$) is to construct addition and multiplication tables. The addition table for $\mathbb{F}_3$ will be

$+$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

We will skip the proofs that addition and multiplication defined on $\mathbb{F}_p$ using modular arithmetic satisfy the field axioms (i) through (v).

For arbitrary primes, the existence of additive inverses is easy to see (the inverse of a is $p - a$), but it is not so obvious that multiplicative inverses always exist. To prove that they do, let us first prove that $\mathbb{F}_p$ satisfies the property of Proposition 4.1:

Proposition 4.6. *Let p be a prime number. If $ab = 0$ in $\mathbb{F}_p$, then either $a = 0$ or $b = 0$ (or both).*

Proof. Since $ab = 0$ in $\mathbb{F}_p$ is the same thing as saying that p divides the usual product ab, the Propositionfollows from the following fact about prime numbers: if the prime number p divides ab, then it divides a or it divides b. (This latter fact follows easily from the Fundamental Theorem of Arithmetic.) □

Put another way, we can say that a fixed non-zero element $a \in \mathbb{F}_p$ induces a *one-to-one* map

$$\begin{aligned} \phi_a : \mathbb{F}_p \setminus \{0\} &\longrightarrow \mathbb{F}_p \setminus \{0\} \\ x &\longmapsto ax \end{aligned}$$

by multiplication. Here $\mathbb{F}_p \setminus \{0\}$ is the set $\mathbb{F}_p$ without 0. To see that ϕ_a is one-to-one, note that $\phi_a(x) = \phi_a(y)$ implies $ax = ay$, which implies $a(x-y) = 0$, which implies $x - y = 0$ (by Proposition 4.6) and hence $x = y$. Since $F_p \setminus \{0\}$ is a *finite* set, this one-to-one map has to be *onto*. In particular, there exists an $x \in \mathbb{F}_p \setminus \{0\}$, such that $ax = 1$, which is the required inverse of a.

Putting the above facts together, we get

Proposition 4.7. *If p is a prime, then $\mathbb{F}_p$, as defined above, is a field.*

If the requirement of having multiplicative inverses is taken out of the definition of a field, the resulting system is called a *ring*. For example, $\mathbb{Z}_4$ is a ring, but not a field since $2 \cdot 2 = 0$. In fact, if q is a composite number, then $\mathbb{Z}_q$ (defined exactly as above) is a ring but not a field. Note that the integers $\mathbb{Z}$ also form a ring.

We now make some definitions from elementary number theory. For any integers a and b which are not both 0, let $d > 0$ be the largest integer which divides both a and b. We call d the *greatest common divisor* of a and b. The greatest common divisor, or simply, gcd of a and b is traditionally denoted (a, b). For example, $(4, 10) = 2$.

Definition 4.3. Let a, b, c be integers. Then we say a is *congruent to b modulo c* if $a - b$ is divisible by c. If a is congruent to b modulo c, we write $a \equiv b \bmod c$.

Proposition 4.8. *Let a, b, q be positive integers. Then the congruence equation $ax \equiv 1$ mod q has a solution if and only if $(a, q) = 1$.*

This proposition again implies that non-zero elements of $\mathbb{F}_p$ have multiplicative inverses. It can also me used to prove that $\mathbb{F}_p$ is not a field, unless p is prime.

The following amusing result of Fermat gives a formula for finding the inverse of any element in $\mathbb{F}_p$.

Fermat's Little Theorem: *Suppose p is a prime greater than 1. Then for any integer $a \not\equiv 0 \bmod p$, $a^{(p-1)} \equiv 1 \bmod p$.*

We will give a group theoretic proof of Fermat's Little Theorem later. From this, we get

Proposition 4.9. *If p is a prime and $a \neq 0$ in $\mathbb{F}_p$, then the reduction modulo p of $a^{(p-2)}$ is the inverse of a in $\mathbb{F}_p$.*

For example, suppose we want to compute the inverse of 5 in $\mathbb{F}_{23}$. Since $5^{21} = 476837158203125$, we simply reduce 476837158203125 modulo 23, which gives 14. If you weren't able to do this calculation in your head, it is useful to have a math package such as Maple or Mathematica. Of course, $5 \cdot 14 = 70 = 3 \cdot 23 + 1$, which is easier to see than the above value of 5^{21}.

Note that Fermat's Little Theorem is not Fermat's Last Theorem that Fermat is famous for having stated (without proof): namely, there are no integer solutions $m > 2$ of $a^m + b^m = c^m$ where $a, b, c \in \mathbb{Z}$ are all non zero.

Amusingly, Fermat's Last Theorem is false in $\mathbb{F}_p$. Indeed, the Binomial Theorem implies that the following identity holds for all $a, b \in \mathbb{F}_p$:

$$(a + b)^p = a^p + b^p. \tag{4.3}$$

Hence the sum of two pth powers is a pth power.

4.1.5 The characteristic of a field

If $\mathbb{F}$ is a finite field, then some multiple r of the identity $1 \in \mathbb{F}$ has to be 0. The reason for this is that since $\mathbb{F}$ is finite, the multiples $r1$ of 1 can't all be different. Hence there have to be $m > n$ such that $m1 = n1$ in $\mathbb{F}$. But this implies $(m - n)1 = 0$. Now I claim that the least positive integer r such that $r1 = 0$ is a prime. For if r can be expressed as a product $r = st$, where s, t are positive integers, then, $r1 = (st)1 = (s1)(t1) = 0$. But, by the minimality of r, $s1 \neq 0$ and $t1 \neq 0$, so a factorization $r = st$ is impossible unless either s or t is 1. Therefore r is prime. One calls this prime p the *characteristic of* $\mathbb{F}$. In general, we make the following definition:

Definition 4.4. Let $\mathbb{F}$ be an arbitrary field. If some multiple $q1$ of 1 equals 0, we say that $\mathbb{F}$ has *positive characteristic*, and, in that case, the *characteristic*

of $\mathbb{F}$ is defined to be the least positive integer q such that $q1 = 0$. If all multiples $q1$ are nonzero, we say $\mathbb{F}$ has *characteristic* 0.

Proposition 4.10. *If a field* $\mathbb{F}$ *has positive characteristic, then its characteristic is a prime. Otherwise, its characteristic is 0.*

Clearly the characteristics of $\mathbb{Q}$ and $\mathbb{R}$ are both 0.

4.1.6 Polynomials

Let $\mathbb{F}$ be a field and suppose x denotes a variable. We will assume it makes sense to talk about the powers x^i, where i is any positive integer. Define $\mathbb{F}[x]$ to be the set of all polynomials

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

with coefficients $a_i \in \mathbb{F}$ for each i, where n is an arbitrary non-negative integer. If $a_n \neq 0$, we say that f has degree n. Of course, if $a_i = 0$, we interpret $a_i x^i$ as being zero also. Addition of polynomials is defined by adding the coefficients of each x^i. We may also multiply two polynomials in the natural way using $x^i x^j = x^{i+j}$ and the distributive law.

Note that by definition, two polynomials $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ and $q(x) = b_k x^k + b_{k-1}x^{k-1} + \cdots + b_1 x + b_0$ are equal if and only if $a_i = b_i$ for each index i.

Exercises

Exercise 4.1. Prove that in any field $(-1)a = -a$.

Exercise 4.2. Prove Fermat's Little Theorem. (Note: it is alright to consult an elementary number theory book.

Exercise 4.3. Verify the identity (4.3), and use it to conclude that if $a^p = b^p$ in $\mathbb{F}_p$, then $a = b$.

Exercise 4.4. Prove that the characteristic of the field $\mathbb{F}_p$.

Exercise 4.5. Show that $\mathbb{F}_p$ is *perfect*. That is, every element in $\mathbb{F}_p$ is a pth power.

Exercise 4.6. Show directly that $\mathbb{F} = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$ is a field under the usual operations of addition and multiplication in $\mathbb{R}$. Also, find $(1 - \sqrt{2})^{-1}$ and $(3 - 4\sqrt{2})^{-1}$.

Exercise 4.7. Describe addition and multiplication for the field $\mathbb{F}_p$ having p elements for $p = 5$. That is, construct addition and multiplication tables for $\mathbb{F}_5$ as in Example 1.1. Check that every element $a \neq 0$ has a multiplicative inverse.

Exercise 4.8. Use Fermat's Theorem to find 9^{-1} in $\mathbb{F}_{13}$. Use this to solve the equation $9x \equiv 15 \bmod 13$.

Exercise 4.9. Find at least one primitive element β for $\mathbb{F}_{13}$? (Calculators should be used here.) Also, express 9^{-1} using this primitive element instead of Fermat's Theorem.

Exercise 4.10. Let $\mathbb{Z}$ denote the integers. Consider the set $\mathcal{Q}$ of all pairs (a, b) where $a, b \in \mathbb{Z}$ and $b \neq 0$. Consider two pairs (a, b) and (c, d) to be the same if $ad = bc$. Now define operations of addition and multiplication on $\mathcal{Q}$ as follows:

$$(a, b) + (c, d) = (ad + bc, bd) \quad \text{and} \quad (a, b)(c, d) = (ac, bd).$$

Show that $\mathcal{Q}$ is a field. Can you identify $\mathcal{Q}$?.

Exercise 4.11. Write out the addition and multiplication tables for $\mathbb{F}_6$. Is $\mathbb{F}_6$ is a field? If not, why not?

Exercise 4.12. Find both $-(6 + 6)$ and $(6 + 6)^{-1}$ in $\mathbb{F}_7$.

Exercise 4.13. Let $\mathbb{F}$ be a field and suppose that $\mathbb{F}' \subset \mathbb{F}$ is a subfield, that is, $\mathbb{F}'$ is a field for the operations of $\mathbb{F}$. Show that $\mathbb{F}$ and $\mathbb{F}'$ have the same characteristic.

4.2 The Field of Complex Numbers

We will now introduce the field $\mathbb{C}$ of complex numbers. The complex numbers are incredibly rich. Without them, mathematics would be a far less interesting discipline. From our standpoint, the most notable fact about the complex numbers is that they form an *algebraically closed field*. That is, $\mathbb{C}$ contains all roots of any polynomial

$$x^n + a_1 x^{n-1} + \dots a_{n-1}x + a_n = 0$$

with complex coefficients. This statement, which is due to C. F. Gauss, is called the Fundamental Theorem of Algebra.

4.2.1 The Definition

The starting point for considering complex numbers is the problem that if a is a positive real number, then $x^2 + a = 0$ apparently doesn't have any roots. In order to give it roots, we have to make sense of an expression such as $\sqrt{-a}$. The solution turns turns out to be extremely natural. The real xy-plane $\mathbb{R}^2$ with its usual component-wise addition also has a multiplication such that certain points (namely points on the y-axis), when squared, give points on the negative x-axis. If we interpret the points on the x-axis as real numbers, this solves our problem. It also turns out that under this multiplication on $\mathbb{R}^2$, every nonzero pair $(a, b)^T$ has a multiplicative inverse. The upshot is that we obtain the field $\mathbb{C}$ of complex numbers. The marvelous and deep consequence of this definition is that $\mathbb{C}$ contains not only numbers such as $\sqrt{-a}$, it contains the roots of all polynomial equations with real coefficients.

Let us now give the details. The definition of multiplication on $\mathbb{R}^2$ is easy to state and has a natural geometric meaning discussed below. First of all, we will call the x-axis the *real axis*, and identify a point of the form $(a, 0)^T$ with the real number a. That is, $(a, 0)^T = a$. Hence multiplication on $\mathbb{R}$ can be reformulated as $ab = (a, 0)^T \cdot (b, 0)^T = (ab, 0)^T$. We extend this multiplication to all of $\mathbb{R}^2$ by putting

$$(a, b)^T \cdot (c, d)^T = (ac - bd, ad + bc)^T. \tag{4.4}$$

(Note: do not confuse this with the dot product on $\mathbb{R}^2$.)

We now make the following definition.

Definition 4.5. Define $\mathbb{C}$ to be $\mathbb{R}^2$ with the usual component-wise addition (vector addition) and with the multiplication defined by (4.4).

Addition and multiplication are clearly binary operations. Notice that $(0,a)^T \cdot (0,a)^T = (-a^2,0)^T$, so that $(0,a)^T$ is a square root of $-a^2$. It is customary to denote $(0,1)^T$ by i so

$$i = \sqrt{-1}.$$

Since any point of $\mathbb{R}^2$ can be uniquely represented

$$(a,b)^T = a(1,0)^T + b(0,1)^T, \tag{4.5}$$

we can therefore write

$$(a,b)^T = a + ib.$$

In other words, by identifying the real number a with the vector $a(1,0)^T$ on the real axis, we can express any element of $\mathbb{C}$ as a sum of a real number, its *real part*, and a multiple of i, its *imaginary part*. Thus multiplication takes the form

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

Of course, $\mathbb{R}$ is explicitly given as a subset of $\mathbb{C}$, namely the real axis.

The Fundamental Theorem of Algebra is formally stated as follows:

Theorem 4.11. *A polynomial equation*

$$p(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0 = 0$$

with complex (but possibly real) coefficients has n complex roots.

There are many proofs of this theorem, but unfortunately, none of them are elementary enough to repeat here. Every known proof draws on some deep result from another field, such as complex analysis or topology.

An easy consequence is that given any polynomial $p(z)$ with complex coefficients, there exist $r_1, \ldots, r_n \in \mathbb{C}$ which are not necessarily all distinct such that

$$p(z) = (z-r_1)(z-r_2)\ldots(z-r_n).$$

We now prove

Theorem 4.12. *$\mathbb{C}$ is a field containing $\mathbb{R}$ as a subfield.*

Proof. By saying $\mathbb{R}$ is a subfield, we mean that addition and multiplication in $\mathbb{C}$ extend the addition and multiplication in $\mathbb{R}$ (after identifying $\mathbb{R}$ and the real axis). The verification of this theorem is simply a computation. The real number 1 is the identity for multiplication in $\mathbb{C}$, and $0 = (0,0)^T$

is the identity for addition. If $a + ib \neq 0$, then $a + ib$ has a multiplicative inverse, namely

$$(a + ib)^{-1} = \frac{a - ib}{a^2 + b^2}. \tag{4.6}$$

The other properties of a field follow easily from the fact that $\mathbb{R}$ is a field. □

4.2.2 The Geometry of $\mathbb{C}$

We now make some more definitions which lead to some beautiful geometric properties of $\mathbb{C}$. First of all, the *conjugate* $\overline{z}$ of $z = a + ib$ is defined by $\overline{z} = a - ib$. It is easy to check the following identities:

$$\overline{w + z} = \overline{w} + \overline{z} \quad \text{and} \tag{4.7}$$

$$\overline{wz} = \overline{w}\,\overline{z}. \tag{4.8}$$

The real numbers are obviously the numbers which are equal to their conjugates. Complex conjugation is the transformation from $\mathbb{R}^2$ to itself which sends a point to its reflection through the real axis.

Formula (4.6) for $(a + ib)^{-1}$ above can now be expressed in a new way. Let $z = a + ib \neq 0$. Since $z\overline{z} = a^2 + b^2$, we get

$$z^{-1} = \frac{\overline{z}}{a^2 + b^2}.$$

Notice that the denominator of the above formula is the square of the length of z. The length of a complex number $z = a + ib$ is called its *modulus* and is denoted by $|z|$. Thus

$$|z| = (z\overline{z})^{1/2} = (a^2 + b^2)^{1/2}.$$

Since $\overline{wz} = \overline{w}\,\overline{z}$, we obtain the nice formula for the modulus of a product, namely

$$|wz| = |w||z|. \tag{4.9}$$

In particular, the product of two unit length complex numbers also has length one. Now the complex numbers of unit length are just those on the unit circle C=$\{x^2 + y^2 = 1\}$. Every point of C can be represented in the form $(\cos\theta, \sin\theta)$ for a unique angle θ such that $0 \leq \theta < 2\pi$. It is convenient to use a complex valued function of $\theta \in \mathbb{R}$ to express this. We define the *complex exponential* to be the function

$$e^{i\theta} := \cos\theta + i\sin\theta. \tag{4.10}$$

The following proposition is geometrically clear.

Proposition 4.13. *Any $z \in \mathbb{C}$ can be represented as $z = |z|e^{i\theta}$ for some $\theta \in \mathbb{R}$. θ is unique up to a multiple of 2π.*

The value of θ in $[0, 2\pi)$ such that $z = |z|e^{i\theta}$ is called the *argument* of z. The key property of the complex exponential is the identity

$$e^{i(\theta+\mu)} = e^{i\theta}e^{i\mu}, \tag{4.11}$$

which follows from the standard trigonometric formulas for the sine and cosine of the sum of two angles. (We will give a simple proof of this when we study rotations in the plane.) This gives complex multiplication a geometric interpretation. Writing $w = |w|e^{i\mu}$, we see that

$$wz = (|w|e^{i\mu})(|z|e^{i\theta}) = (|w||z|)(e^{i\mu}e^{i\theta}) = |wz|e^{i(\mu+\theta)}.$$

In other words, the product wz is obtained by multiplying the lengths of w and z and adding their arguments.

Exercises

Exercise 4.14. Find all solutions of the equation $z^3 + 1 = 0$ and interpret them as complex numbers. Do the same for $z^4 - 1 = 0$.

Exercise 4.15. Find all solutions of the linear system

$$\begin{aligned} ix_1 + 2x_2 + (1-i)x_3 &= 0 \\ -x_1 + ix_2 - (2+i)x_3 &= 0 \end{aligned}$$

Exercise 4.16. Suppose $p(x) \in \mathbb{R}[x]$. Show that the roots of $p(x) = 0$ occur in conjugate pairs, that is $\lambda, \mu \in \mathbb{C}$ where $\overline{\lambda} = \mu$.

4.3 Vector spaces

4.3.1 The notion of a vector space

In mathematics, there many situations in which one deals with sets of objects which can be added and multiplied by scalars, so that these two operations behave like vector addition and scalar multiplication in $\mathbb{R}^n$. A fundamental example of this is the set of all real valued functions whose domain is a closed interval $[a, b]$ in $\mathbb{R}$, which one frequently denotes as $\mathbb{R}^{[a,b]}$. Addition and scalar multiplication of functions is defined pointwise, as in calculus. That is, if f and g are functions on $[a, b]$, then $f + g$ is the function whose value at $x \in [a, b]$ is

$$(f + g)(x) = f(x) + g(x),$$

and if r is any real number, then rf is the function whose value at $x \in [a, b]$ is

$$(rf)(x) = rf(x).$$

The key point is that we have defined sums and scalar multiples so that the sum of $f, g \in \mathbb{R}^{[a,b]}$ and all scalar multiples of a single $f \in \mathbb{R}^{[a,b]}$ are also elements of $\mathbb{R}^{[a,b]}$. When a set S admits an addition (resp. scalar multiplication) with this property, we will say that S is *closed* under addition (resp. scalar multiplication).

A more familiar example is the set $C(a, b)$ of all continuous real valued functions on $[a, b]$. Since $C(a, b) \subset \mathbb{R}^{[a,b]}$, we will of course use the definitions of addition and scalar multiplication already given for $\mathbb{R}^{[a,b]}$. In order to know that $C(a, b)$ is closed under addition and scalar multiplication, we need to know that sums and scalar multiples of continuous functions are continuous. But this is guaranteed by a basic theorem usually discussed in calculus: the sum of two continuous functions is continuous and any scalar multiple of a continuous function is continuous. Hence

$f + g$ and rf belong to $C(a, b)$ for all f and g in $C(a, b)$ and any real scalar r.

We now give the definition of a vector space over a field $\mathbb{F}$. It will be clear that, under the definitions of addition and scalar multiplication given above, $\mathbb{R}^{[a,b]}$ is a vector space over $\mathbb{R}$.

Definition 4.6. Let $\mathbb{F}$ be a field and V a set. Assume that there is an operation on V called addition which assigns to each pair of elements elements $\mathbf{a}$ and $\mathbf{b}$ of V a unique sum $\mathbf{a} + \mathbf{b} \in V$. Assume also that there is a second operation, called scalar multiplication, which assigns to any $r \in \mathbb{F}$ and any

$\mathbf{a} \in V$ a unique scalar multiple $r\mathbf{a} \in V$. Suppose that addition and scalar multiplication satisfy the following axioms.

(1) Vector addition is commutative. That is, $\mathbf{a}+\mathbf{b} = \mathbf{b}+\mathbf{a}$ for all $\mathbf{a}, \mathbf{b} \in V$.

(2) Vector addition is also associative. That is, $(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c})$ for all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in V$.

(3) There is an additive identity $\mathbf{0} \in V$ so that $\mathbf{0} + \mathbf{a} = \mathbf{a}$ for all $\mathbf{a} \in V$.

(4) Every element of V has an additive inverse. That is, given $\mathbf{a} \in V$, there is an element denoted $-\mathbf{a} \in V$ so that $\mathbf{a} + (-\mathbf{a}) = \mathbf{0}$.

(5) $1\mathbf{a} = \mathbf{a}$, for all $\mathbf{a} \in V$.

(6) Scalar multiplication is associative. If $r, s \in \mathbb{F}$ and $\mathbf{a} \in V$, then $(rs)\mathbf{a} = r(s(\mathbf{a}))$.

(7) Scalar multiplication is distributive. If $r, s \in \mathbb{F}$ and $\mathbf{a}, \mathbf{b} \in V$, then $r(\mathbf{a} + \mathbf{b}) = r\mathbf{a} + r\mathbf{b}$, and $(r + s)\mathbf{a} = r\mathbf{a} + s\mathbf{a}$.

Then V is called a *vector space over* $\mathbb{F}$.

You will eventually come to realize that all of the above conditions are needed. Just as for fields, the additive identity $\mathbf{0}$ and additive inverses unique: each vector has exactly one negative. We usually call $\mathbf{0}$ the *zero vector*.

Let's look at some more examples.

Example 4.2. The first example is the obvious one: if $n \geq 1$, then $\mathbb{R}^n$ with the usual component-wise addition and scalar multiplication is a real vector space, that is a vector space over $\mathbb{R}$. We usually call $\mathbb{R}^n$ real n-space.

Example 4.3. More generally, for any field $\mathbb{F}$ and $4 \geq 1$, the set $\mathbb{F}^n$ of all n-tuples $(a_1, a_2, \ldots, a_n)^T$ of elements of $\mathbb{F}$ can be made into a vector space over $\mathbb{F}$ in exactly the same way. That is,

$$(a_1, a_2, \ldots, a_n)^T + (b_1, b_2, \ldots, b_n)^T = (a_1 + b_1, a_2 + b_2, \ldots, a_n + b_n)^T$$

and, for all $r \in \mathbb{F}$,

$$r(a_1, a_2, \ldots, a_n)^T = (ra_1, ra_2, \ldots, ra_n)^T.$$

Example 4.4. When $\mathbb{F} = \mathbb{F}_2$, the elements of $\mathbb{F}^n$ are called *n-bit strings*. For example, if $n = 4$, we have 4-bit strings such as 0000, 1000, 0100, 1100 and so forth. Since there are 4 places to put either a 0 or a 1, there are $2^4 = 16$ 4-bit strings. Binary strings have the nice property that each string is its own additive inverse. Also, the string 1111 changes the parity of each component. That is, $0101 + 1111 = 1010$. The space of n-bit strings are the fundamental objects of coding theory.

Example 4.5. Similarly, if $\mathbb{F} = Z_p$, we can consider the space of all p-ary strings $a_1a_2 \dots a_n$ of elements of $\mathbb{F}_p$. Note that when we consider strings, we often, for simplicity, drop the commas. However, you have to remember that a string $a_1a_2 \dots a_n$ can also be confused with a product in $\mathbb{F}$. The space $(\mathbb{F}_p)^n$ is frequently denoted as $V(n, p)$.

Example 4.6. This example generalizes $\mathbb{R}^{[a,b]}$. Let S be any set and define $\mathbb{R}^S$ to be the set of all real valued functions whose domain is S. We define addition and scalar multiplication pointwise, exactly as for $\mathbb{R}^{[a,b]}$. Then R^S is a vector space over $\mathbb{R}$. Notice that $\mathbb{R}^n$ is nothing but $\mathbb{R}^S$, where $S = \{1, 2, \dots, n\}$. This is because specifying the n-tuple $\mathbf{a} = (a_1, a_2, \dots a_n)^T \in \mathbb{R}^n$ is the same as defining a function $f_{\mathbf{a}} : S \to \mathbb{R}$ by setting $f_{\mathbf{a}}(i) = a_i$.

Example 4.7. The set $\mathcal{P}_n$ of all polynomials

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

with real coefficients having degree at most n is a real vector space under pointwise addition and scalar multiplication defined as above. Pointwise addition of two polynomials amounts to adding the coefficients of x^i in each polynomial, for every i. Scalar multiplication by r is multiplying each term $a_i x^i$ by r. Notice the similarity between these operations on polynomials and component-wise addition and scalar multiplication on $\mathbb{R}^{n+1}$.

$$(a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0) + (b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0) =$$

$$(a_n + b_n)x^n + (a_{n-1} + b_{n-1})x^{n-1} + \cdots + (a_1 + b_1)x + (a_0 + b_0),$$

while

$$\begin{aligned}\mathbf{a} + \mathbf{b} &= (a_0, a_1, a_2, \dots a_n)^T + (b_0, b_1, b_2, \dots b_n)^T \\ &= (a_0 + b_0, a_1 + b_1, a_2 + b_2, \dots a_n + b_n)^T.\end{aligned}$$

In this sense, $\mathcal{P}_n$ and $\mathbb{R}^{n+1}$ are indistinguishable as vector spaces.

Example 4.8. Consider the differential equation

$$y'' + ay' + by = 0, \tag{4.12}$$

where a and b are real constants. This is an example of a homogeneous linear second order differential equation with constant coefficients. The set of twice differentiable functions on $\mathbb{R}$ which satisfy (4.12) is a real vector space.

4.3.2 Inner product spaces

The set $C(a, b)$ of continuous real valued functions on the interval $[a, b]$ defined in the previous subsection is one of the most basic vector spaces in mathematics. Although $C(a, b)$ is much more complicated than $\mathbb{R}^n$, it has an important structure in common with $\mathbb{R}^n$ which lets us partially extend our intuition about $\mathbb{R}^n$ to $C(a, b)$. Namely, we can define an inner product (f, g) of $f, g \in C(a, b)$ by

$$(f, g) = \int_a^b f(t)g(t)\mathrm{d}t.$$

The first three axioms for the Euclidean inner product (dot product) on $\mathbb{R}^n$ are verified by applying standard facts about integration proved (or at least stated) in any calculus book. Recall that the last axiom requires that $(f, f) \geq 0$ and $(f, f) = 0$ only if $f = 0$. The verification of this requires some argument, and we leave it as an exercise in elementary real analysis.

If a real vector space admits an inner product, then the notions of length and distance can be introduced by just copying the definitions used for $\mathbb{R}^n$ in Chapter 1. The length $||f||$ of any $f \in C(a, b)$ is defined to be

$$||f|| := (f, f)^{1/2} = \big(\int_a^b f(t)^2 \mathrm{d}t \big)^{1/2},$$

and the distance between $f, g \in C(a, b)$ is defined to be

$$d(f, g) = ||f - g|| = \big(\int_a^b (f(t) - g(t))^2 \mathrm{d}t \big)^{1/2}.$$

Just as for the Euclidean inner product on $\mathbb{R}^n$, we can say two functions $f, g \in C(a, b)$ are *orthogonal* if $(f, g) = \int_a^b f(t)g(t)\mathrm{d}t = 0$. Then the tools we developed from the Euclidean inner product on $\mathbb{R}^n$ such as projections and orthogonal decompositions extend word by word to $C(a, b)$. For example, $\cos t$ and $\sin t$ are orthogonal on $[0, 2\pi]$ because $\int_0^{2\pi} \cos t \sin t \mathrm{d}t = 0$.

Although the notion of orthogonality for $C(a,b)$ doesn't have any obvious geometric meaning, it nevertheless enables us to extend our intuitive concept of orthogonality into a new situation. In fact, this extension turns out to be extremely important since it leads to the idea of expanding a function in terms of possibly infinitely many mutually orthogonal functions. These infinite series expansions are called Fourier series. For example, the functions $\cos mx$, $m = 0, 1, 2, \ldots$ are orthogonal on $[0, 2\pi]$, and the Fourier cosine series for $f \in C(0, 2\pi)$ has the form

$$f(x) = \sum_{m=0}^{\infty} a_m \cos mx,$$

where

$$a_m = \int_0^{2\pi} f(t) \cos mt \mathrm{d}t / \int_0^{\pi} \cos^2 mt \mathrm{d}t.$$

We call a_m the *Fourier coefficient* of f with respect to $\cos mt$. Notice that $a_m \cos mx$ is the projection of f on $\cos mx$. This series is an infinite version of the formula in Proposition 2.3.

If we only take finitely many terms of the above Fourier series, we obtain a *least squares approximation* to f.

Example 4.9. Suppose $[a,b] = [-1,1]$. Then the functions 1 and x are orthogonal. In fact, x^k and x^m are orthogonal if k is even and m is odd, or vice versa. Indeed,

$$(x^k, x^m) = \int_{-1}^{1} x^k \cdot x^m \mathrm{d}x = \int_{-1}^{1} x^{k+m} \mathrm{d}x = 0,$$

since $k+m$ is odd. On the other hand, the projection of x^2 on the constant function 1 is $r1$, where $r = \frac{1}{2}\int_{-1}^{1} 1 \cdot x^2 \mathrm{d}x = \frac{1}{3}$. Thus, $x^2 - 1/3$ is orthogonal to the constant function 1 on $[-1,1]$, and $x^2 = (x^2 - 1/3) + 1/3$ is an orthogonal decomposition of x^2 on $[-1,1]$.

Similarly, by arguing exactly as in §2, we immediately obtain a Cauchy-Schwartz inequality on $C(a,b)$.

Cauchy-Schwartz Inequality for $C(a,b)$. *For any $f, g \in C(a,b)$, the inequality*

$$|\int_a^b f(t)g(t)\mathrm{d}t| \leq \big(\int_a^b f(t)^2 \mathrm{d}t\big)^{1/2} \big(\int_a^b g(t)^2 \mathrm{d}t\big)^{1/2}$$

holds. Equality holds if and only if one of the functions is a constant multiple of the other.

4.3.3 Subspaces and Spanning Sets

We next consider the extremely important notion of a subspace.

Definition 4.7. Let V be vector space over a field $\mathbb{F}$. A non-empty subset W of V is called a *linear subspace of* V, or simply a *subspace*, provided $\mathbf{a} + \mathbf{b} \in W$ and $r\mathbf{a}$ are in W whenever $\mathbf{a}, \mathbf{b} \in W$ and $r \in \mathbb{F}$.

The following Proposition is immediate.

Proposition 4.14. *Every subspace W of V is a vector space over $\mathbb{F}$ in its own right.*

Proof. This is left as an exercise. □

Notice that every subspace of a vector space contains the zero vector $\mathbf{0}$ (why?). In fact, $\{\mathbf{0}\}$ is itself a subspace, called the *trivial subspace*. Hence, if the constant term of a homogeneous linear equation $ax + by + cz = d$ above is nonzero, then the solution set cannot be a subspace.

Here is a fundamental example of a subspace of $\mathbb{R}^3$.

Example 4.10. The solutions $(x, y, z)^T \in \mathbb{R}^3$ of a homogeneous linear equation $ax + by + cz = 0$, with $a, b, c \in \mathbb{R}$ make up the plane consisting of all vectors orthogonal to $(a, b, c)^T$. By the properties of the dot product, the sum of any two solutions is another solution, and any scalar multiple of a solution is a solution. Hence the solution set of a homogeneous linear equation in three variables is a subspace of $\mathbb{R}^3$. More generally, the solution set of a homogeneous linear equation in n variables with real coefficients is a subspace of $\mathbb{R}^n$. If the coefficients are in the field $\mathbb{F}$, then the solutions in $\mathbb{F}^n$ make up a subspace of $\mathbb{F}^n$.

The subspaces of $\mathbb{R}^2$ and $\mathbb{R}^3$ can be easily described. For $\mathbb{R}^2$, they are $\{\mathbf{0}\}$, any line through $\mathbf{0}$ and $\mathbb{R}^2$ itself. We will consider the subspaces of $\mathbb{R}^3$ below. Try to guess what they are before reading further.

A basic method for constructing subspaces of a given vector space is to take linear combinations.

Definition 4.8. Let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be vectors in V, and let $r_1, \ldots, r_k$ be any elements of $\mathbb{F}$. Then the vector

$$\mathbf{w} = \sum_{i=1}^{k} r_i \mathbf{v}_i$$

is called a *linear combination* of $\mathbf{v}_1, \ldots, \mathbf{v}_k$. A subspace W which consists of all linear combinations of a collection of vectors in V, say $\mathbf{v}_1, \ldots, \mathbf{v}_k$, is said to be spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

Proposition 4.14 says that subspaces are closed under taking linear combinations. It also asserts the converse. The set of all linear combinations of a collection of vectors in V is a subspace of V. We will denote the subspace spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$ by $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$.

As previously noted, lines L and planes P in $\mathbb{R}^3$ containing $\mathbf{0}$ are subspaces of $\mathbb{R}^3$. Every line L is by definition $\text{span}\{\mathbf{a}\}$ for some (in fact, any) nonzero $\mathbf{a} \in L$. Is every plane P the span of a set of vectors? Well, if $\mathbf{a}$ and $\mathbf{b}$ are two non-collinear vectors in P, then $W := \text{span}\{\mathbf{a}, \mathbf{b}\}$ is contained in P. The question remains as to whether $W = P$. To see why the answer is yes, as expected, you can argue as follows. Let $\mathbf{n}$ denote any non-zero normal to P, and take any $\mathbf{c} \in P$. The line through $\mathbf{a}$ and $\mathbf{0}$ and the line through $\mathbf{b}$ and $\mathbf{0}$ both lie on P. Now any vector of the form $\mathbf{c} + t\mathbf{b}$ is orthogonal to $\mathbf{n}$, so the line through $\mathbf{c}$ parallel to $\mathbf{b}$ also lies on P. This line meets the line through $\mathbf{a}$ and $\mathbf{0}$ at some $r\mathbf{a}$ (why?). Next construct $s\mathbf{b}$ in the same way by interchanging the roles of $\mathbf{a}$ and $\mathbf{b}$. Then clearly, $\mathbf{c} = r\mathbf{a} + s\mathbf{b}$, because $\mathbf{c}$ is the intersection of the line through $r\mathbf{a}$ parallel to $\mathbf{b}$ and the line through $s\mathbf{b}$ parallel to $\mathbf{a}$. Hence $\mathbf{c} \in \text{span}\{\mathbf{a}, \mathbf{b}\}$, so $P = \text{span}\{\mathbf{a}, \mathbf{b}\}$.

On the other hand, if $\mathbf{a}$ and $\mathbf{b}$ are two non-collinear vectors in $\mathbb{R}^3$, then $\mathbf{n} = \mathbf{a} \times \mathbf{b}$ is orthogonal to any linear combination of $\mathbf{a}$ and $\mathbf{b}$. Thus we obtain a homogeneous equation satisfied by exactly those vectors in $P = \text{span}\{\mathbf{a}, \mathbf{b}\}$. (We just showed above that every vector orthogonal to $\mathbf{n}$ is on P.) If $\mathbf{n} = (r, s, t)^T$, then an equation is $rx + sy + tz = 0$.

Example 4.11. Let P be the plane spanned by $(1, 1, 2)^T$ and $(-1, 0, 1)^T$. Then $(1, 1, 2)^T \times (-1, 0, 1)^T = (1, -3, 1)^T$ is a normal to P, so an equation for P is $x - 3y + z = 0$.

4.3.4 Linear Systems and Matrices Over an Arbitrary Field

Although we developed the theory of linear systems over the reals, the only reason we didn't use an arbitrary field is that the definition hadn't yet been made. In fact, the material covered in Chapter 3 pertaining to linear systems and matrices goes through word for word when we use an arbitrary field $\mathbb{F}$. Thus we have $m \times n$ matrices over $\mathbb{F}$, which will be denoted by $\mathbb{F}^{m \times n}$, and linear systems $A\mathbf{x} = \mathbf{b}$, where $A \in \mathbb{F}^{m \times n}$, $\mathbf{x} \in \mathbb{F}^n$ and $\mathbf{b} \in \mathbb{F}^m$. Row reduction, matrix inversion etc. all go through as for $\mathbb{R}$. We will not bother to restate all the results, but we will use them when needed. The matrix group $GL(n, \mathbb{R})$ is replaced by its counterpart $GL(n, \mathbb{F})$, which is also a matrix group. One thing to be careful of, however, is that in $\mathbb{R}^n$, if $\mathbf{x}^T\mathbf{x} = 0$, then $\mathbf{x} = \mathbf{0}$. This is false for most other fields such as $\mathbb{F}_p$. It is

even false for $\mathbb{C}$ since $(1 \;\; i)\begin{pmatrix}1\\ i\end{pmatrix} = 1 + i^2 = 0$.

Exercises

Exercise 4.17. Let V be a vector space. Show that $0\mathbf{a} = \mathbf{0}$.

Exercise 4.18. Let V be a vector space. Show that for any $\mathbf{a} \in V$, the vector $(-1)\mathbf{a}$ is an additive inverse of V. In other words, prove the formula $(-1)\mathbf{a} = -\mathbf{a}$.

Exercise 4.19. Describe all subspaces of $\mathbb{R}^3$.

Exercise 4.20. Which of the following subsets of $\mathbb{R}^2$ is not a subspace?

(a) The line $x = y$;

(b) The unit circle;

(c) The line $2x + y = 1$;

s(d) The first octant $x, y \geq 0$.

Exercise 4.21. Prove that every line through the origin and plane through the origin in $\mathbb{R}^3$ are subspaces.

Exercise 4.22. Find all the subspaces of the vector space $V(n, p) = (\mathbb{F}_p)^n$ in the following cases:

(i) $n = p = 2$;

(ii) $n = 2$, $p = 3$; and

(iii) $n = 3$, $p = 2$.

Exercise 4.23. How many points lie on a line in $V(n, p)$? On a plane?

Exercise 4.24. Let $\mathbb{F} = \mathbb{F}_2$. Find all solutions in $\mathbb{F}^4$ of the equation $w+x+y+z = 0$. Compare the number of solutions with the number of elements $\mathbb{F}^4$ itself has?

Exercise 4.25. Consider the real vector space $V = C(0, 2\pi)$ with the inner product defined in §4.3.2.

(a) Find the length of $\sin^2 t$ in V.

(b) Compute the inner product $(\cos t, \sin^2 t)$.

(c) Find the projection of $\sin^2 t$ on each of the functions $1, \cos t$, and $\sin t$ in V.

(d) Are 1, $\cos t$ and $\sin t$ mutually orthogonal as elements of V?

(e) How would you define the orthogonal projection of $\sin^2 t$ onto the subspace W of V spanned by $1, \cos t$, and $\sin t$?

(f) Describe the subspace W of part (e).

Exercise 4.26. Assume $f \in C(a, b)$. Recall that the average value of f over $[a, b]$ is defined to be

$$\frac{1}{b-a}\int_a^b f(t)dt.$$

Show that the average value of f over $[a, b]$ is the projection of f on 1. Does this suggest an interpretation of the average value?

Exercise 4.27. Let $f, g \in C(a, b)$. Give a formula for the scalar t which minimizes

$$||f - tg||^2 = \int_a^b (f(x) - tg(x))^2 dx.$$

Exercise 4.28. Find a spanning set for the plane $3x - y + 2z = 0$ in $\mathbb{R}^3$.

Exercise 4.29. Find an equation for the plane in $\mathbb{R}^3$ through the origin containing both $(1, 2, -1)^T$ and $(3, 0, 1)^T$.

Exercise 4.30. Let L be the line obtained by intersecting the two planes in the previous two exercises. Express L as span$\{\mathbf{a}\}$ for some $\mathbf{a}$.

Exercise 4.31. Describe all subspaces of $\mathbb{R}^4$ and $\mathbb{R}^5$.

4.4 Summary

The purpose of this chapter was to introduce two fundamental notions: fields and vector spaces. Fields are the number systems where we can add, subtract, multiply and divide in the usual sense. The basic examples were the rationals $\mathbb{Q}$, which form the smallest field containing the integers, the reals (which are hard to define, so we didn't), the prime fields $\mathbb{F}_p$, which are the systems which support modular arithmetic, and the queen of all fields, the complex numbers $\mathbb{C}$. The basic property of $\mathbb{C}$ is that it contains $\mathbb{R}$ and is algebraically closed. A vector space is what happens when a field is cloned. That is, we get the space $\mathbb{F}^n$ of n-tuples of elements in $\mathbb{F}$. In a vector space, we can add elements and operate on them by scalars. General vector spaces do not have a multiplication, although some specific examples do. Vector spaces V have subspaces, the most common example of a subspace being the set of all linear combinations of a subcollection of the vectors in V. We mentioned a special class of vector spaces over $\mathbb{R}$, namely inner product spaces. These spaces are just like $\mathbb{R}^n$ except for the fact that they are frequently not spanned by finite sets as $\mathbb{R}^n$ is. However, some of the properties we developed for $\mathbb{R}^n$, such as orthogonal projection and the Cauchy-Schwartz Inequalty, go through in the general case just as they did in $\mathbb{R}^n$.

For example, $C(a, b)$ is an inner product space that doesn't have this property. We also pointed out that the theory of linear systems and matrix theory, two themes that were carried out over $\mathbb{R}$ in Chapter 3, have identical versions over an arbitrary field.

Chapter 5

The Theory of Finite Dimensional Vector Spaces

5.1 Some Basic concepts

Vector spaces which are spanned by a finite number of vectors are said to be *finite dimensional*. The purpose of this chapter is explain the elementary theory of such vector spaces, including linear independence and notion of the dimension. Indeed, the development of a workable definition for this notion was one of the first important achievements in basic algebra. We will also explain the construction of a number basic vector spaces such as direct sums, duals and quotients.

5.1.1 The Intuitive Notion of Dimension

Roughly speaking, the dimension of a vector space should describe the number of degrees of freedom an inhabitant of the space has. It is clear what this means for for subsets of $\mathbb{R}^n$ provided $n = 1, 2$ or 3. For example, the path traced out by a point moving smoothly through $\mathbb{R}^3$ is intuitively one dimensional. A smooth surface without any thickness is a two dimensional object. (On the other hand, the notion of the dimension of non-smooth paths and surfaces can be very hard to formulate. In fact, such dimensions may turn out to real, that is non-integral.) The objects we will be treating here, however, are linear, and we will see that their dimensions are defined in a natural way.

In particular, we will see that any subspace of $\mathbb{F}^n$ is finite dimensional. Since our intuition tells us that $\mathbb{R}^1$, $\mathbb{R}^2$ and $\mathbb{R}^3$ should have dimensions one,

two and three respectively, we should expect that our final definition will have the property that the dimension of $\mathbb{R}^n$ is n. Thus, the dimension of $\mathbb{F}^n$ should also be n.

5.1.2 Linear Independence

Let V denote a vector space over a field $\mathbb{F}$. Before defining the notion of the dimension of V, we need to discuss the concept of linear independence. One way of putting the definition is to say that a set of vectors is linearly independent if no one of them can be expressed as a linear combination of the others. This means that if you have two vectors, they are linearly independent when they don't lie on the same line through the origin (i.e. they aren't collinear), and three vectors are linearly independent when they don't all lie on a plane through the origin. (Of course, any three vectors lie on a plane, but the plane will not necessarily contain the origin.) Thus the situation of two, three or any finite number of vectors failing to be linearly independent will involve a constraint. Let us now formulate a definition.

Definition 5.1. Let $\mathbf{w}_1, \ldots, \mathbf{w}_k$ in V. Then we say that $\mathbf{w}_1, \ldots, \mathbf{w}_k$ are *linearly independent* (or, simply, *independent*) if and only if the vector equation

$$x_1\mathbf{w}_1 + x_2\mathbf{w}_2 + \cdots + x_k\mathbf{w}_k = \mathbf{0} \tag{5.1}$$

has only the trivial solution $x_1 = x_2 = \cdots = x_k = 0$. If a non trivial solution exists, we will call the vectors *linearly dependent* (or, simply, *dependent*).

One of the first things to notice is any set of vectors in V that includes $\mathbf{0}$ is dependent (why?). We begin with a reformulation of the concept of indepedence.

Proposition 5.1. *A set of vectors is linearly dependent if and only if one of them can be expressed as a linear combination of the others.*

Proof. Suppose first that one of the vectors, say $\mathbf{w}_1$, is a linear combination of the others. That is

$$\mathbf{w}_1 = a_2\mathbf{w}_2 + \cdots + a_k\mathbf{w}_k.$$

Thus

$$\mathbf{w}_1 - a_2\mathbf{w}_2 - \cdots - a_k\mathbf{w}_k = \mathbf{0},$$

so (5.1) has a solution with $x_1 = 1$, thus a non trivial solution. Therefore $\mathbf{w}_1, \ldots, \mathbf{w}_k$ are dependent. Conversely, suppose $\mathbf{w}_1, \ldots, \mathbf{w}_k$ are dependent.

This means that there is a solution $x_1, x_2, \ldots, x_k$ of (5.1), where some $x_i \neq 0$. We can assume (just by reordering the vectors) that the nonzero coefficient is x_1. Then we can write

$$\mathbf{w}_1 = a_2\mathbf{w}_2 + \ldots a_k\mathbf{w}_k,$$

where $a_i = -x_i/x_1$, so the proof is done. □

FIGURE
(LINEARLY DEPENDENT, INDEPENDENT)

The following fact gives one of the important properties of linearly independent sets.

Proposition 5.2. *Assume that $\mathbf{w}_1, \ldots, \mathbf{w}_k$ are linearly independent vectors in V and suppose $\mathbf{v}$ is in their span. Then there is exactly one linear combination of $\mathbf{w}_1, \ldots, \mathbf{w}_k$ which gives $\mathbf{v}$.*

Proof. Suppose $\mathbf{v}$ can be expressed in two ways, say

$$\mathbf{v} = r_1\mathbf{w}_1 + r_2\mathbf{w}_2 + \cdots + r_k\mathbf{w}_k$$

and

$$\mathbf{v} = s_1\mathbf{w}_1 + s_2\mathbf{w}_2 + \cdots + s_k\mathbf{w}_k$$

where the r_i and s_i are all elements of $\mathbb{F}$. By subtracting and doing a bit of algebraic manipulation, we get that

$$\mathbf{0} = \mathbf{v} - \mathbf{v} = (r_1 - s_1)\mathbf{w}_1 + (r_2 - s_2)\mathbf{w}_2 + \cdots + (r_k - s_k)\mathbf{w}_k.$$

Since the $\mathbf{w}_i$ are independent, every coefficient $r_i - s_i = 0$, which proves the Proposition. □

When $V = \mathbb{F}^n$, the definition of linear independence involves considering a linear system. Recalling that vectors in $\mathbb{F}^n$ are viewed as column vectors, consider the $n \times k$ matrix $A = (\mathbf{w}_1 \ \ldots \ \mathbf{w}_k)$. By the theory of linear systems (Chapter 2), we have

Proposition 5.3. *The vectors $\mathbf{w}_1, \ldots, \mathbf{w}_k$ in $\mathbb{F}^n$ are linearly independent exactly when the system $A\mathbf{x} = \mathbf{0}$ has no non trivial solution which is the case exactly when the rank of A is k. In particular, more than n vectors in $\mathbb{F}^n$ are linearly dependent.*

5.1.3 The Definition of a Basis

As usual let V be a vector space over a field $\mathbb{F}$.

Definition 5.2. A collection of vectors in V which is both linearly independent and spans V is called a *basis* of V.

Notice that we have not required that a basis be a finite set. Usually, however, we will deal with vector spaces that have a finite basis. One of the questions we will investigate is whether a finite dimensional vector space has a basis. Of course, $\mathbb{F}^n$ has a basis, namely the standard basis vectors, or, in other words, the columns of the identity matrix I_n over $\mathbb{F}$. A non zero vector in $\mathbb{R}^n$ spans a line, and clearly a single non zero vector is linearly independent. Hence a line has a basis consisting of a single element. A plane P through the origin is spanned by any two non collinear vectors on P, and two any two non collinear vectors on P are linearly independent. Thus P has a basis consisting of two vectors. It should noted that the trivial vector space $\{\mathbf{0}\}$ does not have a basis, since in order to contain a linearly independent subset it has to contain a nonzero vector.

Proposition 5.2 allow us to deduce an elementary property of bases.

Proposition 5.4. *The vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ in V form a basis of W if and only if every vector $\mathbf{v}$ in V admits a unique expression*

$$\mathbf{v} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_r\mathbf{v}_r,$$

where $a_1, a_2, \ldots a_r$ are elements of $\mathbb{F}$.

Proof. We leave this as an exercise. □

Example 5.1. Suppose A is an $m \times n$ matrix over $\mathbb{F}$. The column space $\mathrm{col}(A)$ of A is a subspace of $\mathbb{F}^m$ which we have already considered. Using our new terminology, if A has rank n, then its columns are independent and hence form a basis of the column space. This gives a useful criterion for determining whether or not a given set of vectors in $\mathbb{F}^m$ is a basis of the subspace they span. If the rank of A is less than n, the columns are dependent, so there is still the problem of finding a basis. More generally, this is the problem of extracting a basis from a spanning set that may be dependent. We will solve this for $\mathbb{F}^m$ below.

Example 5.2. Let A be an $m \times n$ matrix over $\mathbb{F}$. As pointed out in Chapter 3, the theory of linear systems, which was developed in Chapter 3 for the case $\mathbb{F} = \mathbb{R}$, extends over word for word to a linear equation (or system) $A\mathbf{x} = \mathbf{b}$ over any field $\mathbb{F}$ and any $A \in \mathbb{F}^{m\times n}$. For example, the fundamental

solutions of $A\mathbf{x} = \mathbf{0}$ are a basis of the null space $\mathcal{N}(A)$, which is a subspace of $\mathbb{F}^n$, and we still have the identity

$$\dim \mathcal{N}(A) = n - \operatorname{rank}(A),$$

which was originally stated in (3.4).

Exercises

Exercise 5.1. Are the vectors $(0,2,1,0)^T$, $(1,0,0,1)^T$ and $(1,0,1,1)^T$ in $\mathbb{R}^4$ are independent? Can they form a basis of $\mathbb{R}^4$?

Exercise 5.2. Are $(0,0,1,0)^T$, $(1,0,0,1)^T$ and $(1,0,1,1)^T$ independent in $\mathbb{F}_2^4$?

Exercise 5.3. Show that any subset of a linearly independent set is linearly independent.

Exercise 5.4. Suppose $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$ are mutually orthogonal unit vectors in $\mathbb{R}^m$. Show $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$ are independent.

Exercise 5.5. Show that m independent vectors in $\mathbb{F}^m$ are a basis.

Exercise 5.6. Find a basis for the space $\mathbb{R}[x]$ of all polynomials with real coefficients.

Exercise 5.7. True or False: Four vectors in $\mathbb{R}^3$ are dependent. (Supply reasoning.)

Exercise 5.8. Prove the assertions made in Example 5.2 that the fundamental solutions are a basis of $\mathcal{N}(A)$ and $\dim \mathcal{N}(A) = n - \operatorname{rank}(A)$.

Exercise 5.9. Use the theory of linear systems to show the following:

(i) More than m vectors in $\mathbb{F}^m$ are dependent.

(ii) Fewer than m vectors in $\mathbb{F}^m$ cannot span $\mathbb{F}^m$.

Exercise 5.10. Let $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ be a basis of $\mathbb{R}^3$.

(a) Determine whether or not $3\mathbf{u} + 2\mathbf{v} + \mathbf{w}$, $\mathbf{u} + \mathbf{v} + 0\mathbf{w}$, and $-\mathbf{u} + 2\mathbf{v} - 3\mathbf{w}$ are independent.

(b) Find a general necessary and sufficient condition for the vectors $a_1\mathbf{u} + a_2\mathbf{v} + a_3\mathbf{w}$, $b_1\mathbf{u} + b_2\mathbf{v} + b_3\mathbf{w}$ and $c_1\mathbf{u} + c_2\mathbf{v} + c_3\mathbf{w}$ to be independent, where $a_1, a_2, \ldots, c_3$ are arbitrary scalars.

Exercise 5.11. Find a basis for the set of invertible 3×3 real matrices. (Be careful.)

5.2 Bases and Dimension

We will now (finally) define the notion of dimension and prove the basic results about bases. As we already noted above (see Exercise 5.9) $\mathbb{F}^n$ can't contain more than n independent vectors. Our definition of dimension will in fact amount to saying that the dimension of an $\mathbb{F}$-vector space V is the maximal number of independent vectors. This definition gives the right answer for the dimension of a line (one), a plane (two) and more generally $\mathbb{F}^n$ (n).

5.2.1 The Definition of Dimension

We start with the following definition.

Definition 5.3. Let V be a vector space over an arbitrary field $\mathbb{F}$. Then we say that V is *finite dimensional* if it is spanned by a finite set of vectors.

For the remainder of this section, we will only consider finite dimensional vector spaces.

Definition 5.4. The *dimension* of a finite dimensional vector space V is the number of elements in a basis of V. For convenience, we will define the dimension of the trivial vector space $\{\mathbf{0}\}$ to be 0, even though $\{\mathbf{0}\}$ doesn't have a basis. The dimension of V will be denoted by $\dim V$ or by $\dim_{\mathbb{F}} V$ in case there is a chance of confusion about which field is being considered.

This definition obviously assumes that a finite dimensional vector space (different from $\{\mathbf{0}\}$) has a basis. Less obviously, it also assumes that any two bases have the same number of elements. Hence, we have to prove these two facts before we can use the definition. Thses assertions will be part of the Dimension Theorem, which will be proved below.

In order to get some feeling for the definition, let's consider $\mathbb{F}^n$ as a special case. I claim that if $n > 0$, any basis of $\mathbb{F}^n$ has n elements. In fact, this is just the result of Exercise 5.9, since it says that a basis, being independent, cannot have more than n elements and, being a spanning set, has to have at least n elements. In fact, we can even say more.

Proposition 5.5. *Every basis of $\mathbb{F}^n$ contains exactly n vectors. Moreover, n linearly independent vectors in $\mathbb{F}^n$ span $\mathbb{F}^n$ and hence are a basis. Similarly n vectors in $\mathbb{F}^n$ that span $\mathbb{F}^n$ are also linearly independent and hence are also a basis.*

Proof. We already verified the first statement. Now if $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are independent and $A = (\mathbf{w}_1 \ \ldots \ \mathbf{w}_n)$, then $\mathcal{N}(A) = \{\mathbf{0}\}$, so A has rank n. Since A is $n \times n$, we know A has an inverse, so the system $A\mathbf{x} = \mathbf{b}$ is solvable for any $\mathbf{b} \in \mathbb{F}^n$. Thus $\mathbf{w}_1, \ldots, \mathbf{w}_n$ span $\mathbb{F}^n$. Similarly, if $\mathbf{w}_1, \ldots, \mathbf{w}_n$ span $\mathbb{F}^n$, they have to be independent for the same reason: A is $n \times n$ of rank n. □

There is a slight sublety, however, which is illustrated by what happens when $\mathbb{F} = \mathbb{C}$. Since $\mathbb{C} = \mathbb{R}^2$, $\mathbb{C}^n$ is in some sense the same as $\mathbb{R}^{2n}$. Thus, if we ask what is the dimension of $\mathbb{C}^n$, we see that the answer could be either n or $2n$ and still be consistent with having the dimension of $\mathbb{F}^n$ be n. Hence when we speak of the dimension of $\mathbb{C}^n = \mathbb{R}^{2n}$, we need to differentiate between whether we are speaking of the real dimension (which is $2n$) or the complex dimension (which is n). In other words, $\dim_{\mathbb{R}} \mathbb{C}^n = 2n$ while $\dim_{\mathbb{C}} \mathbb{C}^n = n$.

5.2.2 Some Examples

We now consider some examples.

Example 5.3. Let $\mathbf{e}_i$ denote the ith column of I_n. As mentioned above, the vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ are the so called standard basis of $\mathbb{R}^n$. In fact, $\mathbf{e}_1, \ldots, \mathbf{e}_n$ make sense for any field $\mathbb{F}$ and, by the same reasoning, are a basis of $\mathbb{F}^n$.

Example 5.4. The dimension of a line is 1 and that of a plane is 2. The dimension of the hyperplane $a_1x_1 + \cdots + a_nx_n = 0$ in $\mathbb{R}^n$ is $n-1$, provided some $a_i \neq 0$. Note that the $n-1$ fundamental solutions form a basis of the hyperplane.

Example 5.5. Let $A = (\mathbf{w}_1 \ \mathbf{w}_2 \ \ldots \ \mathbf{w}_n)$ be $n \times n$ over $\mathbb{F}$, and suppose A has rank n. Then the columns of A are a basis of $\mathbb{F}^n$. Indeed, the columns span $\mathbb{F}^n$ since we can express an arbitrary $\mathbf{b} \in \mathbb{F}^n$ as a linear combinations of the columns due to the fact that the system $A\mathbf{x} = \mathbf{b}$ is consistent for all $\mathbf{b}$. We are also guaranteed that $\mathbf{0}$ is the unique solution of the system $A\mathbf{x} = \mathbf{0}$. Hence the columns of A are independent. Thus, the columns of an $n \times n$ matrix over $\mathbb{F}^n$ of rank n are a basis of $\mathbb{F}^n$. (Note that we have essentially just repeated part of the proof of Proposition5.5.)

Example 5.6. For any positive integer n, let $\mathcal{P}_n$ denote the space of polynomials with real coefficients of degree at most n (cf. Example 4.7). Let's determine a basis of $\mathcal{P}_3$. Consider the polynomials $1, x, x^2, x^3$. I claim they are linearly independent. To see this, we have to show that if

$$y = \sum_{i=0}^{3} a_i x^i = 0$$

for every x, then each $a_i = 0$. Now if $y = 0$, then

$$y(0) = a_0 = 0,\ y'(0) = a_1 = 0,\ y''(0) = a_2 = 0,\ y'''(0) = a_3 = 0.$$

Hence we have the asserted linear independence. It is obvious that $1, x, x^2, x^3$ span $\mathcal{P}_3$, so our job is done.

Example 5.7. Let $a_1, \ldots, a_m$ be real constants. Then the solution space of the homogeneous linear differential equation

$$y^{(m)} + a_1 y^{(m-1)} + \cdots + a_{m-1} y' + a_m y = 0$$

is a vector space over $\mathbb{R}$. It turns out, by a theorem on differential equations, that the dimension of this space is m. For example, when $m = 4$ and $a_i = 0$ for $1 \leq i \leq 4$, then we are dealing with the vector space $\mathcal{P}_3$ of the last example. The solution space of the equation $y'' + y = 0$ consists of all linear combinations of the functions $\sin x$ and $\cos x$.

5.2.3 The Dimension Theorem

We will next establish the basic result needed to show that the definition of dimension makes sense.

Theorem 5.6 (The Dimension Theorem). *Assume V is a finite dimensional vector space over a field $\mathbb{F}$ containing a non zero vector. Then V has a basis. In fact, any spanning set for V contains a basis, and any linearly independent subset of V is contained in a basis. Moreover, any two bases of V have the same number of elements.*

Proof. We first show every spanning set contains a basis. Let $\mathbf{w}_1, \ldots, \mathbf{w}_k$ span V. Of course, we may certainly assume that every $\mathbf{w}_i \neq \mathbf{0}$. Now consider the set of all subsets of $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ which also span V, and let $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ be any such subset where r is minimal. There is no problem showing this subset exists, since $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ has only 2^k subsets.

I claim that $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are independent. For, if

$$a_1\mathbf{v}_1 + \cdots + a_r\mathbf{v}_r = \mathbf{0},$$

and some $a_i \neq 0$, then

$$\mathbf{v}_i = \frac{-1}{a_i} \sum_{j \neq i} a_j \mathbf{v}_j,$$

so if $\mathbf{v}_i$ is deleted from $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$, we still have a spanning set, which contradicts the minimality of r. Thus $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are independent, so every

spanning set contains a basis. In particular, since V has a finite spanning set, it has a basis.

We next show that any linearly independent set in V can be extended to a basis. Let $\mathbf{w}_1, \ldots, \mathbf{w}_m$ be independent, and put $W = \text{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$. I claim that if $\mathbf{v} \notin W$, then $\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{v}$ are independent. To see this, suppose

$$a_1\mathbf{w}_1 + \cdots + a_m\mathbf{w}_m + b\mathbf{v} = \mathbf{0}.$$

If $b \neq 0$, it follows (as in the last argument) that $\mathbf{v} \in W$, contrary to the choice of $\mathbf{v}$. Thus $b = 0$. But then each $a_k = 0$ also since the $\mathbf{w}_k$ are independent. This proves the claim.

Now suppose $W \neq V$. We will use the basis $\mathbf{v}_1, \ldots, \mathbf{v}_r$ obtained above. If each $\mathbf{v}_i \in W$, then $W = V$ and we are done. Otherwise, let i be the first index such that $\mathbf{v}_i \notin W$. By the previous claim, $\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{v}_i$ are independent. Hence they form a basis for $W_1 = \text{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{v}_i\}$. Clearly we may continue, at each step adding one of the $\mathbf{v}_j$, if necessary, always maintaining an independent subset of V. Eventually we have to obtain a subspace containing $\mathbf{v}_1, \ldots, \mathbf{v}_r$, so our original independent vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$ are contained in a basis.

It remains to show that two bases of V have the same number of elements. This is proved by the so called the replacement principle. Suppose $\mathbf{u}_1, \ldots, \mathbf{u}_m$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are two bases of V with $m \neq n$. Without any loss of generality, suppose $m \leq n$. We can then write

$$\mathbf{v}_1 = r_1\mathbf{u}_1 + r_2\mathbf{u}_2 \cdots + r_m\mathbf{u}_m.$$

Since $\mathbf{v}_1 \neq \mathbf{0}$, some $r_i \neq 0$, so we may suppose, by renumbering indices if necessary, that $r_1 \neq 0$. I claim that this implies that $\mathbf{v}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ is also a basis of V. To see this, we must show $\mathbf{v}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ are independent and span. Suppose that

$$x_1\mathbf{v}_1 + x_2\mathbf{u}_2 \cdots + x_m\mathbf{u}_m = \mathbf{0}.$$

If $x_1 \neq 0$, then

$$\mathbf{v}_1 = y_2\mathbf{u}_2 + \cdots + y_j\mathbf{u}_m,$$

where $y_i = -x_i/x_1$. Since $r_1 \neq 0$, this gives two distinct ways of expanding $\mathbf{v}_1$ in terms of the first basis, which contradicts Proposition 5.4. Hence $x_1 = 0$. It follows immediately that all $x_i = 0$ (why?), so $\mathbf{v}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ are independent. I leave the proof that $\mathbf{v}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ span V as an exercise, hence we have produced a new basis of V where $\mathbf{v}_1$ replaces $\mathbf{u}_1$. I claim that

$\mathbf{u}_2, \ldots, \mathbf{u}_m$ can be renumbered so that $\mathbf{u}_2$ can be replaced by $\mathbf{v}_2$, giving a new basis $\mathbf{v}_1, \mathbf{v}_2, \mathbf{u}_3 \ldots, \mathbf{u}_m$ of V. To be explicit, we can write

$$\mathbf{v}_2 = r_1\mathbf{v}_1 + r_2\mathbf{u}_2 \cdots + r_m\mathbf{u}_m.$$

Then there exists an $i > 1$ such that $r_i \neq 0$ (why?). Renumbering so that $i = 2$ and applying the same reasoning as in the previous argument, we get the claim. Continuing this process, we will eventually replace all the $\mathbf{u}_i$'s, which implies that $\mathbf{v}_1, \ldots, \mathbf{v}_m$ must be a basis of V. But if $m < n$, it then follows that $\mathbf{v}_{m+1}$ is a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_m$, which contradicts the linear independence of $\mathbf{v}_1, \ldots, \mathbf{v}_n$. This is a contradiction, so we conclude $m = n$, and the Dimension Theorem is proven. □

5.2.4 Some Applications and Further Properties

Let's begin with an application. Let p be a prime and consider a finite dimensional vector space V over $\mathbb{F} = \mathbb{F}_p$. Then the dimension of V determines the number of elements of V.

Proposition 5.7. *The number of elements of V is exactly $p^{\dim_{\mathbb{F}_p} V}$.*

The proof goes as follows. Let $k = \dim V$ and choose a basis $\mathbf{w}_1, \ldots, \mathbf{w}_k$ of V, which we know is possible. Then every $\mathbf{v} \in W$ has a unique expression

$$\mathbf{v} = a_1\mathbf{w}_1 + a_2\mathbf{w}_2 + \cdots + a_k\mathbf{w}_k$$

where $a_1, a_2, \ldots, a_k$ are scalars, that is, elements of $\mathbb{F}_p$. Now it is simply a matter of counting such expressions. In fact, since $\mathbb{F}_p$ has p elements, there are p choices for each a_i, and, since different choices of the a_i give different elements of V (Proposition5.2), it follows that V contains exactly $p \cdot p \cdots p = p^k$ elements. □

Thus, for example, a line in $\mathbb{F}^n$ has p elements, a plane has p^2 and so forth.

Example 5.8. Consider for example a matrix over $\mathbb{F}_2$, for example

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Let V denote the row space of A, that is the subspace of $\mathbb{F}^4$ spanned by A's rows. Using row operations, A row reduces to the 4×4 identity matrix I_4.

We will see below that row operations leave the row space of A unchanged. Hence we conclude that the row space of A is $\mathbb{F}^4$. The original rows are a basis as are the rows of $A_{red} = I_4$.

We next establish some more properties of finite dimensional vector spaces. First of all, we prove a fact that is obvious for $\mathbb{F}^k$.

Proposition 5.8. *Let V be a finite dimensional vector space, say $\dim V = n$. Then any subset of V containing more that n elements is dependent.*

Proof. It suffices to show that any subset of $n+1$ elements of V is dependent. Let $\mathbf{v}_1, \ldots, \mathbf{v}_{n+1}$ be independent and suppose $\mathbf{u}_1, \ldots, \mathbf{u}_n$ give a basis of V. Applying the replacement principle as in the proof of the Dimension Theorem (Theorem 5.6), we get that $\mathbf{v}_1, \ldots, \mathbf{v}_n$ give a basis, so $\mathbf{v}_1, \ldots, \mathbf{v}_{n+1}$ can't be independent. □

We also need to show the not surprising fact that every subspace of a finite dimensional vector space is also finite dimensional.

Proposition 5.9. *Every subspace W of a finite dimensional vector space V is finite dimensional. In particular, for any subspace W of V, $\dim W$ is defined and $\dim W \leq \dim V$.*

Proof. We have to show that W is finite dimensional. Consider any set of independent vectors in W, say $\mathbf{w}_1, \ldots, \mathbf{w}_m$. If these vectors don't span W, then $\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{w}_{m+1}$ are independent for any choice of $\mathbf{w}_{m+1} \in W$ not in the span of $\mathbf{w}_1, \ldots, \mathbf{w}_m$. If $\dim V = n$, then by Proposition 5.8, more than n elements of V are dependent, so it follows that W has to have a finite spanning set with at most n elements. The assertion that $\dim W \leq \dim V$ also follows immediately from Proposition 5.8. □

5.2.5 Extracting a Basis Constructively

Theorem 5.6 guarantees that any spanning set of a finite dimensional vector space contains a basis. In fact, the subsets which give bases are exactly the minimal spanning subsets. Frequently, however, we need an explicit method for actually extracting one of these subsets. There is an explicit method for subspaces of $\mathbb{F}^n$ which is based on row reduction. Suppose $\mathbf{w}_1, \ldots, \mathbf{w}_k \in \mathbb{F}^n$, and let W be the subspace they span. Let us construct a subset of these vectors which spans W. Consider the $n \times k$ matrix $A = (\mathbf{w}_1 \ \ldots \ \mathbf{w}_k)$. We must find columns of A which are a basis of the column space $W = \mathrm{col}(A)$.

Proposition 5.10. *The columns of A that correspond to a corner entry in A_{red} are a basis of the column space* $\mathrm{col}(A)$ *of A. Therefore, the dimension of* $\mathrm{col}(A)$ *of A is the rank of A.*

Proof. The key observation is that $A\mathbf{x} = \mathbf{0}$ if and only if $A_{red}\mathbf{x} = \mathbf{0}$ (why?). This says any expression of linear dependence among the columns of A_{red} is also an expression of linear dependence among the columns of A. The converse statement is also true. For example, if column five of A_{red} is the sum of the first four columns of A_{red}, this also holds for the first five columns of A. But it is obvious that the columns of A_{red} containing a corner entry are a basis of the column space of A_{red} (of course, this says nothing about the column space of A). Hence the corner columns are also linearly independent in W. But we just saw that every non corner column in A_{red} is a linear combination of the corner columns of A_{red}, so the same is true for A from what we said above. Therefore, the corner columns in A span W, and the proof is complete. □

This result may seem a little surprising since it involves row reducing A which of course changes $\mathrm{col}(A)$.

Example 5.9. To consider a simple example, let

$$A = \begin{pmatrix} 1 & 2 & 2 \\ 4 & 5 & 8 \\ 7 & 8 & 14 \end{pmatrix}.$$

Then

$$A_{red} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Proposition 5.10 implies the first two columns are a basis of $\mathrm{col}(A)$. Notice that the first and third columns are dependent in both A and A_{red} as the Proposition guarantees. The Proposition says that the first two columns are a basis of the column space, but makes no assertion about the second and third columns, which in fact are also a basis.

5.2.6 The Row Space of A and the Rank of A^T

We now consider the row space of a matrix. The goal of this subsection is to relate the row space to row operations and then to derive a somewhat surprising result: namely that A and A^T have the same rank.

Definition 5.5. The *row space* of an $m \times n$ matrix A over $\mathbb{F}$ is the subspace $\mathrm{row}(A) \subset \mathbb{F}^n$ of A spanned by the rows of A.

The first fact about the row space of a matrix is about how row operations affect the row space (not!). Actually, we already let the cat out of the bag in Example 5.8.

Proposition 5.11. *Elementary row operations leave the row space of A unchanged. Consequently A and A_{red} always have the same row space. Moreover, the non-zero rows of A_{red} are a basis of* $\mathrm{row}(A)$*. Hence the dimension of the row space of A is the rank of A, that is*

$$\dim \mathrm{row}(A) = \mathrm{rank}(A).$$

Proof. The first assertion is equivalent to the statement that for any $m \times m$ elementary matrix E, $\mathrm{row}(EA) = \mathrm{row}(A)$. If E is a row swap or a row dilation, this is clear. So we only have to worry about what happens if E is an elementary row operation of the type III. Suppose E replaces the ith row $\mathbf{r}_i$ by $\mathbf{r'}_i = \mathbf{r}_i + k\mathbf{r}_j$, where $k \neq 0$ and $j \neq i$. Since the rows of EA and A are the same except that $\mathbf{r}_i$ is replaced by $\mathbf{r'}_i$, and since $\mathbf{r'}_i$ is itself a linear combination of two rows of A, every row of EA is a linear combination of some rows of A. Hence $\mathrm{row}(EA) \subset \mathrm{row}(A)$. But since E^{-1} is also of type III,

$$\mathrm{row}(A) = \mathrm{row}((E^{-1}E)A) = \mathrm{row}(E^{-1}(EA)) \subset \mathrm{row}(EA),$$

so $\mathrm{row}(EA) = \mathrm{row}(A)$. Therefore row operations do not change the row space, and the first claim of the proposition is proved.

It follows that the non zero rows of A_{red} span $\mathrm{row}(A)$. We will be done if the non zero rows of A_{red} are independent. But this holds for the same reason the rows of I_n are independent. Every non zero row of A_{red} has a 1 in the component corresponding to its corner entry, and in this column, all the other rows have a zero. Therefore the only linear combination of the non zero rows which can give the zero vector is the one where every coefficient is zero. Hence the non zero rows of A_{red} are also independent, so they form a basis of $\mathrm{row}(A)$. Thus $\dim \mathrm{row}(A)$ is the number of non zero rows of A_{red}, which is also the number of corners in A_{red}. Therefore, $\dim \mathrm{row}(A) = \mathrm{rank}(A)$, and this completes the proof. □

Here is a surprising corollary.

Corollary 5.12. *For any $m \times n$ matrix A over a field $\mathbb{F}$,*

$$\dim \mathrm{row}(A) = \dim \mathrm{col}(A).$$

Put another way, the ranks of A and A^T are the same.

Proof. We just saw that $\dim \mathrm{row}(A)$ equals $\mathrm{rank}(A)$. But in Proposition 5.10, we also saw that $\dim \mathrm{col}(A)$ also equals $\mathrm{rank}(A)$. Finally, $\mathrm{rank}(A^T) = \dim \mathrm{col}(A^T) = \dim \mathrm{row}(A)$, so we are done. □

This result is unexpected. There would seem to be no connection whatsoever between $\mathrm{row}(A)$ and $\mathrm{col}(A)$. But now we see they have the same dimensions. Let us cap off the discussion with some examples.

Example 5.10. The 3×3 counting matrix C of Example 3.1 has reduced form

$$C_{red} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

The first two rows are a basis of $\mathrm{row}(C)$ since they span $\mathrm{row}(C)$ and are clearly independent (why?).

Example 5.11. Suppose $\mathbb{F} = \mathbb{F}_2$ and

$$A = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

A is already reduced so its rows are a basis of $\mathrm{row}(A)$, which is thus a three dimensional subspace of $\mathbb{F}^6$. A little combinatorial reasoning will allow us to compute the number of elements in $\mathrm{row}(A)$. In fact, the answer was already given by Proposition 5.7. Repeating the argument, there are 3 basis vectors and each has 2 possible coefficients, 0 and 1. Thus there are $2 \cdot 2 \cdot 2 = 2^3$ vectors in all. The 7 non zero vectors are

$$(100111),\ (010101),\ (001011),\ (110010),\ (101100),\ (011110),\ (1111001).$$

Note that all combinations of 0's and 1's occur in the first three components, since the corners are in these columns. In fact, the first three components tell you which linear combination is involved. Examples of this type will come up again in linear coding theory.

Exercises

Exercise 5.12. Find a basis for the subspace of $\mathbb{R}^4$ spanned by

$$(1,0,-2,1),\ (2,-1,2,1),\ (1,1,1,1),\ (0,1,0,1),\ (0,1,1,0)$$

containing the first and fifth vectors.

Exercise 5.13. Consider the matrix $A = \begin{pmatrix} 1 & 2 & 0 & 1 & 2 \\ 2 & 0 & 1 & -1 & 2 \\ 1 & 1 & -1 & 1 & 0 \end{pmatrix}$ as an element of $\mathbb{R}^{3\times 5}$.

(i) Show that the fundamental solutions are a basis of $\mathcal{N}(A)$.

(ii) Find a basis of $\mathrm{col}(A)$.

(iii) Repeat (i) and (ii) when A is considered as a matrix over $\mathbb{F}_3$.

Exercise 5.14. Suppose V is a finite dimensional vector space over a field $\mathbb{F}$, and let W be a subspace of V.

(i) Show that if $\dim W = \dim V$, then $W = V$.

(ii) Show that if $w_1, w_2, \ldots, w_k$ is a basis of W and $v \in V$ but $v \notin W$, then $w_1, w_2, \ldots, w_k, v$ are independent.

Exercise 5.15. Let $\mathbb{F}$ be any field, and suppose V and W are subspaces of $\mathbb{F}^n$.

(i) Show that $V \cap W$ is a subspace of $\mathbb{F}^n$.

(ii) Let $V + W = \{u \in \mathbb{F}^n \mid u = v + w\ \exists v \in V, w \in W\}$. Show that $V + W$ is a subspace of $\mathbb{F}^n$.

Exercise 5.16. Consider the subspace W of $\mathbb{F}_2^4$ spanned by 1011, 0110, and 1001.

(i) Find a basis of W and compute $|W|$.

(ii) Extend your basis to a basis of $\mathbb{F}_2^4$.

Exercise 5.17. Find a basis of the vector space $\mathbb{R}^{n\times n}$ of real $n \times n$ matrices.

Exercise 5.18. A square matrix A over $\mathbb{R}$ is called symmetric if $A^T = A$ and called skew symmetric if $A^T = -A$.
(a) Show that the $n \times n$ symmetric matrices form a subspace of $\mathbb{R}^{n\times n}$, and compute its dimension.
(b) Show that the $n \times n$ skew symmetric matrices form a subspace of $\mathbb{R}^{n\times n}$ and compute its dimension.
(c) Find a basis of $\mathbb{R}^{3\times 3}$ using only symmetric and skew symmetric matrices.

Exercise 5.19. Show that the set of $n \times n$ upper triangular real matrices is a subspace of $\mathbb{R}^{n\times n}$. Find a basis and its dimension.

Exercise 5.20. If A and B are $n \times n$ matrices so that B is invertible (but not necessarily A), show that the ranks of A, AB and BA are all the same.

Exercise 5.21. True or False: $\operatorname{rank}(A) \geq \operatorname{rank}(A^2)$. Explain your answer.

Exercise 5.22. Let W and X be subspaces of a finite dimensional vector space V of dimension n. What are the minimum and maximum dimensions that $W \cap X$ can have? Discuss the case where W is a hyperplane (i.e. $\dim W = n - 1$) and X is a plane (i.e. $\dim X = 2$).

Exercise 5.23. Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ be mutually orthogonal unit vectors in $\mathbb{R}^n$. Are $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ a basis of $\mathbb{R}^n$?

Exercise 5.24. Given a subspace W of $\mathbb{R}^n$, define $W^\perp$ to be the set of vectors in $\mathbb{R}^n$ orthogonal to every vector in W. Show that $W^\perp$ is a subspace of $\mathbb{R}^n$ and describe a method for constructing a basis of $W^\perp$.

Exercise 5.25. Let W is a subspace of $\mathbb{R}^n$. Show that

$$\dim(W) + \dim(W^\perp) = n.$$

Exercise 5.26. Suppose W is a subspace of $\mathbb{F}^n$ of dimension k. Show the following:

(i) Any k linearly independent vectors in W span W, hence are a basis of W.

(ii) Any k vectors in W that span W are linearly independent, hence are a basis of W.

Exercise 5.27. Show that the functions

$$1, x, x^2, x^3, \ldots, x^n, \ldots$$

are linearly independent on any open interval (a, b).

Exercise 5.28. Is $\mathbb{R}$ a vector space over $\mathbb{Q}$? If so, is $\dim_{\mathbb{Q}} \mathbb{R}$ finite or infinite?

5.3 Some General Constructions of Vector Spaces

In this section, we consider some of the standard ways of producing new vector spaces: intersections, internal and external sums and quotients. We also will derive an interesting formula (the Hausdorff Intersection Formula) relating the dimensions of some of these spaces.

5.3.1 Intersections

Let V be a vector space over a field $\mathbb{F}$ with subspaces W and Y.

The most obvious way of building a new subspace is by taking the intersection $W \cap Y$.

Proposition 5.13. *The intersection $W \cap Y$ of the subspaces W and Y of V is also a subspace of V. More generally, the intersection of any number of subspaces of V is also a subspace.*

Proof. This is a simple exercise. □

Proposition 5.13 is simply a generalization of the fact that the solution space of a homogeneous linear system is a subspace of $\mathbb{F}^n$, the solution space is the intersection of a finite number of hyperplanes in $\mathbb{F}^n$, where each hyperplane is given by a homogeneous linear equation.

5.3.2 External and Internal Sums

First of all, let V and W be arbitrary vector spaces over the same field $\mathbb{F}$.

Definition 5.6. The *external direct sum* of V and W is the vector space denoted by $V \times W$ consisting of all pairs $(\mathbf{v}, \mathbf{w})$, where $\mathbf{v} \in V$ and $\mathbf{w} \in W$. Addition is defined by

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

and scalar multiplicatyion is defined by

$$r(\mathbf{v}, \mathbf{w}) = (r\mathbf{v}, r\mathbf{w}).$$

Of course, the wide awake reader will note that $\mathbb{F} \times \mathbb{F}$ is nothing else than $\mathbb{F}^2$. Thus the external direct sum is a generalization of the construction of $\mathbb{F}^n$. Hence, it can immediately be extended (inductively) to any number of vector spaces over $\mathbb{F}$. Once this is understood, $\mathbb{F}^n$ becomes the n-fold external direct sum of $\mathbb{F}$. It is also frequently called the n-fold Cartesian product of $\mathbb{F}$.

Note also that V and W can both be considered (in a natural way) as subspaces of $V \times W$ (why?).

Proposition 5.14. *If V and W are finite dimensional vector spaces over $\mathbb{F}$, then so is their external direct sum, and* $\dim(V \times W) = \dim V + \dim W$.

Proof. We leave this as an exercise. □

Now suppose W and Y are subspaces of the same vector space V. Then we can form the internal sum of W and Y.

Definition 5.7. The *internal sum* or simply *sum* of W and Y is the set

$$W + Y = \{\mathbf{w} + \mathbf{y} \mid \mathbf{w} \in W, \mathbf{y} \in Y\}.$$

More generally, we can in the same way form the sum of an arbitrary (finite) number of subspaces $V_1, V_2, \cdots, V_k$ of V. The sum $V_1 + \cdots + V_k$ is usually abbreviated as $\sum_{i=1}^{k} V_i$ or more simply as $\sum V_i$.

Proposition 5.15. *The sum $W = \sum_{i=1}^{k} V_i$ of the subspaces $V_1, V_2, \cdots, V_k$ of V is also a subspace of V. In fact, W is the smallest subspace of V containing each V_i.*

Proof. We leave the proof as an exercise. □

5.3.3 The Hausdorff Interesction Formula

We now ask a more interesting question: what are the dimensions of the sum $W + Y$ and the intersection $W \cap Y$ of two subspaces W and Y of V? It turns out that each depends on the other. The relation between them is called the Hausdoff Intersection Formula.

Theorem 5.16. *If W and Y are subspaces of a finite dimensional vector space V, then*

$$\dim(W + Y) = \dim W + \dim Y - \dim(W \cap Y). \tag{5.2}$$

Proof. We know $W \cap Y$ is a subspace of V, so, since V is finite dimensional, Proposition 5.8 and the Dimension Theorem tells us that $W \cap Y$ has a basis, say $\mathbf{x}_1, \ldots, \mathbf{x}_k$. We also know, by the Dimension Theorem again, that we can extend this basis to a basis of W, say $\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{w}_{k+1}, \ldots, \mathbf{w}_{k+r}$, and we can do likewise for Y, getting say $\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{y}_{k+1}, \ldots, \mathbf{y}_{k+s}$. I claim

$$\mathcal{B} = \{\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{w}_{k+1}, \ldots, \mathbf{w}_{k+r}, \mathbf{y}_{k+1}, \ldots, \mathbf{y}_{k+s}\}$$

is a basis of $W+Y$. It is not hard to see that $\mathcal{B}$ spans, so we leave this to the reader. To see $\mathcal{B}$ is independent, suppose

$$\sum_{i=1}^{k} \alpha_i \mathbf{x}_i + \sum_{j=k+1}^{k+r} \beta_j \mathbf{w}_j + \sum_{m=k+1}^{k+s} \gamma_m \mathbf{y}_m = \mathbf{0}. \tag{5.3}$$

Thus

$$\sum \gamma_m \mathbf{y}_m = -\big(\sum \alpha_i \mathbf{x}_i + \sum \beta_j \mathbf{w}_j\big) \in V.$$

In other words,

$$\sum \gamma_m \mathbf{y}_m \in Y \cap W.$$

Thus

$$\sum \gamma_m \mathbf{y}_m = \sum \delta_i \mathbf{x}_i$$

for some $\delta_i \in \mathbb{F}$. Substituting this into (5.3) gives the expression

$$\sum \alpha_i' \mathbf{x}_i + \sum \beta_j \mathbf{v}_j = \mathbf{0},$$

where $\alpha_i' = \alpha_i + \delta_i$. From this we infer that all the α_i' and β_j are 0 since the $\mathbf{x}_i$ and $\mathbf{w}_j$ are independent. Referring back to (5.3), and applying the same reasoning, we see that

$$\sum \alpha_i \mathbf{x}_i + \sum \gamma_m \mathbf{y}_m = \mathbf{0},$$

hence all α_i and γ_m are 0 too. This proves $\mathcal{B}$ is independent. It's clear that $\mathcal{B}$ spans $W+Y$, so it forms a basis of $W+Y$. Consequently $\dim(W+Y) = k+r+s$. To finish the proof, we need to count dimensions. Now

$$\dim(W+Y) = k+r+s = (k+r)+(k+s)-k,$$

which is exactly $\dim W + \dim Y - \dim(W \cap Y)$. □

This leads to a deeper understanding of how subspaces intersect.

Corollary 5.17. *If W and Y are subspaces of V, then*

$$\dim(W \cap Y) \geq \dim W + \dim Y - \dim V. \tag{5.4}$$

Proof. Since W and Y are both subspaces of V, $\dim(W+Y) \leq \dim V$. Now substitute this into the Hausdorff Formula (5.2). □

Example 5.12. Let us illustrate a typical application of (5.4). I claim that the intersection $P_1 \cap P_2$ of two planes in $\mathbb{R}^3$ has to contain a line. For $\dim(P_1+P_2) \geq 2+2-3 = 1$. More generally, the intersection $H_1 \cap H_2$ of two hyperplanes in $\mathbb{R}^n$ has dimension at least $2(n-1)-n = n-2$, hence it contains an $(n-2)$-dimensional subspace. On the other hand, the intersection of two planes in $\mathbb{R}^4$ does not have to contain a line since $2+2-4 = 0$.

5.3.4 Internal Direct Sums

The final concept in this section is the notion of an internal direct sum. As usual, let V be a vector space over $\mathbb{F}$ with subspaces W and Y.

Definition 5.8. We say that V is the *internal direct sum* (or simply the *direct sum*) of W and Y if $V = W + Y$ and for any $\mathbf{v} \in V$, the expression $\mathbf{v} = \mathbf{w} + \mathbf{y}$ with $\mathbf{w} \in W$ and $\mathbf{y} \in Y$ is unique. If V is the internal direct sum of W and Y, we write $V = W \oplus Y$. More generally, we say V is the direct sum of a collection of subspaces $V_1, \ldots, V_k$ if $V = \sum V_i$ and for any $\mathbf{v} \in V$, the expression $\mathbf{v} = \sum \mathbf{v}_i$, where each $\mathbf{v}_i \in V_i$, is unique. In this case, we write $V = \bigoplus_{i=1}^{k} V_i$.

Proposition 5.18. *Suppose V is finite dimensional. Then a necessary and sufficient condition that $V = W \oplus Y$ is that $V = W + Y$ and $W \cap Y = \{\mathbf{0}\}$. Equivalently, $V = W \oplus Y$ if and only if $\dim V = \dim W + \dim Y$ and $\dim(W \cap Y) = 0$.*

Proof. First, assume $V = W + Y$ and $W \cap Y = \{\mathbf{0}\}$. To see $V = W \oplus Y$, let $\mathbf{v}$ have two expressions $\mathbf{v} = \mathbf{w} + \mathbf{y} = \mathbf{w}' + \mathbf{y}'$. Then $\mathbf{w} - \mathbf{w}' = \mathbf{y}' - \mathbf{y}$ is an element of $W \cap Y = \{\mathbf{0}\}$, so $\mathbf{w} = \mathbf{w}'$ and $\mathbf{y}' = \mathbf{y}$. Hence $V = W \oplus Y$. On the other hand, if $V = W \oplus Y$ and $W \cap Y \neq \{\mathbf{0}\}$, then any non-zero $\mathbf{w} \in W \cap Y$ has two expressions $\mathbf{w} = \mathbf{w} + \mathbf{0} = \mathbf{0} + \mathbf{w}$. This violates the definition of a direct sum, so $W \cap Y = \{\mathbf{0}\}$.

Next, suppose $\dim V = \dim W + \dim Y$ and $\dim(W \cap Y) = 0$. Then, by the Hausdorff Intersection Formula, $\dim(W + Y) = \dim W + \dim Y$. Thus $W + Y$ is a subspace of V having the same dimension as V. Therefore $V = W + Y$. Since $\dim(W \cap Y) = 0$, we have $V = W \oplus Y$. The converse is proved by reversing this argument. □

We can extend Proposition5.18 to any number of subspaces as follows.

Proposition 5.19. *Suppose V is finite dimensional and $V_1, \ldots, V_k$ are subspaces. Then $V = \bigoplus_{i=1}^{k} V_i$ if and only if $V = \sum V_i$ and for every index i,*

$$V_i \cap \Big(\sum_{j \neq i} V_j\Big) = \{\mathbf{0}\}.$$

If $\sum V_i = V$ and $\sum_{i=1}^{k} \dim V_i = \dim V$, then $V = \bigoplus_{i=1}^{k} V_i$.

Proof. We leave this as an exercise. □

Example 5.13. In the last section, we defined the orthogonal complement $V^\perp$ of a subspace V of $\mathbb{R}^n$. Recall,

$$V^\perp = \{\mathbf{w} \in \mathbb{R}^n \mid \mathbf{w} \cdot \mathbf{v} = 0 \text{ for all } \mathbf{v} \in V\}.$$

Orthogonal complements in $\mathbb{R}^n$ provide examples of direct sums, since as we saw in Exercise 5.24, $\dim V + \dim V^\perp = n$ and $V \cap V^\perp = \{\mathbf{0}\}$ (why?). Thus, for any V,

$$\mathbb{R}^n = V \oplus V^\perp. \tag{5.5}$$

5.4 Vector Space Quotients

The final topic in this chapter is the construction of the quotient of a vector space V by a subspace W. This is a new vector space denoted as V/W. In a certain sense (despite the notation), V/W can be thought of as subtracting W from V. Not too much should be read into this claim. The meaning become clearer later.

5.4.1 Equivalence Relations

The notion of a quotient occurs in many different contexts in algebra. It is surely one of the most fundamental ideas in the area. The basis for this notion is the idea of an equivalence relation on a set. First, recall that if S and T are sets, the *product* $S \times T$ (as defined in the previous section) is the set of all pairs (s, t), where $s \in S$ and $t \in T$.

Definition 5.9. Let S be a non-empty set. A subset E of $S \times S$ is called a *relation on* S. If E is a relation on S, and a and b are elements of S, we will say a and b are *related* by E and write aEb if and only if $(a, b) \in E$. A relation E on S is called an *equivalence relation on* S when the following three conditions hold for all $a, b, c \in S$:

(i) (reflexivity) aEa,

(ii) (symmetry) if aEb, then bEa, and

(iii) (transitivity) if aEb and bEc, then aEc.

If E is an equivalence relation on S and $a \in S$, then the *equivalence class of* a is defined to be the set of all elements $b \in S$ such that bEa.

Proposition 5.20. *If E is an equivalence relation on S, every element $a \in S$ is in an equivalence class, and two equivalence classes are either disjoint or equal. Therefore S is the disjoint union of the equivalence classes of E.*

Proof. Every element is equivalent to itself, so S is the union of its equivalence classes. We have to show that if two equivalence classes C and C' contain a common element a, then $C = C'$. Let C and C' be two equivalence classes. If $a \in C \cap C'$, then for any $c \in C$ and $c' \in C'$, we have aEc and aEc'. By (ii) and (iii), it follows that cEc'. Hence every element equivalent to c is equivalent c', and conversely. Thus $C = C'$. □

5.4.2 Cosets

Now let V be a vector space over $\mathbb{F}$ and let W be a subspace. We are going to use W to define an equivalence relation on V. The elements of V/W will be the equivalence classes. The definition is given in the following Proposition.

Proposition 5.21. *Let V be a vector space over $\mathbb{F}$ and let W be a subspace. Given $\mathbf{v}$ and $\mathbf{y}$ in V, let us say that $\mathbf{v}E_W\mathbf{y}$ if and only if $\mathbf{v}-\mathbf{y}\in W$. Then E_W is an equivalence relation on V.*

Proof. Clearly $\mathbf{v}E_W\mathbf{v}$ since $\mathbf{v}-\mathbf{v}=\mathbf{0}\in W$. If $\mathbf{v}E_W\mathbf{y}$, then $\mathbf{y}E_W\mathbf{v}$ since W is closed under scalar multiplication. Finally, if $\mathbf{v}E_W\mathbf{y}$ and $\mathbf{y}E_W\mathbf{z}$, then $\mathbf{v}E_W\mathbf{z}$ since $\mathbf{v}-\mathbf{z}=(\mathbf{v}-\mathbf{y})+(\mathbf{y}-\mathbf{z})$ and W is closed under sums. Hence E_W is an equivalence relation on V. □

Definition 5.10. Let $\mathbf{v}\in V$ be fixed. Then the *coset* of W containing $\mathbf{v}$ is defined to be the set

$$\mathbf{v}+W=\{\mathbf{v}+\mathbf{w}\mid \mathbf{w}\in W\}. \tag{5.6}$$

The notion of a coset is nothing complicated. For example, if $V=\mathbb{R}^3$ and W is a plane through $\mathbf{0}$, then the coset $\mathbf{v}+W$ is simply the plane through $\mathbf{v}$ parallel to W.

Proposition 5.22. *The equivalence classes of the equivalence relation E_W on V are precisely the cosets of W. In particular, $\mathbf{v}+W=\mathbf{y}+W$ if and only if $\mathbf{v}-\mathbf{y}\in W$.*

Proof. Let C denote the equivalence class of $\mathbf{v}$ and consider the coset $\mathbf{v}+W$. If $\mathbf{y}E_W\mathbf{v}$, then $\mathbf{y}-\mathbf{v}=\mathbf{w}\in W$. Hence $\mathbf{y}=\mathbf{v}+\mathbf{w}$, so $\mathbf{y}\in\mathbf{v}+W$. Therefore $C\subset\mathbf{v}+W$. Arguing in reverse, we also conclude that $\mathbf{v}+W\subset C$. □

We now define the quotient space V/W to be the set of all cosets of W. We want to show that cosets can be added. Given two cosets $(\mathbf{v}+W)$ and $(\mathbf{y}+W)$, define their sum by

$$(\mathbf{v}+W)+(\mathbf{y}+W)=(\mathbf{v}+\mathbf{y})+W. \tag{5.7}$$

In order that this addition be a binary operation on V/W, we have to show that the rule (5.7) is independent of the way we write each coset. That is, suppose we have $\mathbf{v}+W=\mathbf{v}'+W$ and $\mathbf{y}+W=\mathbf{y}'+W$. Then we have to show that $(\mathbf{v}+\mathbf{y})+W=(\mathbf{v}'+\mathbf{y}')+W$. But this is so if and only if

$$(\mathbf{v}+\mathbf{y})-(\mathbf{v}'+\mathbf{y}')\in W,$$

which indeed holds since

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

Therefore, addition is well defined. Scalar multiplication on cosets is defined by

$$a(\mathbf{v}+W)=a\mathbf{v}+W. \tag{5.8}$$

A similar argument shows that this scalar multiplication is well defined.

We can now define the quotient vector space V/W and prove one of its main properties.

Theorem 5.23. *Let V be a vector space over a field $\mathbb{F}$ and suppose W is a subspce of V. Define V/W to be the set of cosets of W in V with addition and scalar multiplication defined as in (5.7) and (5.8). Then V/W is a vector space over $\mathbb{F}$. If V is finite dimensional, then*

$$\dim V/W=\dim V-\dim W.$$

Proof. The fact that V/W satisfies the vector space axioms is straightforward, so we will omit most of the details. The zero element is $\mathbf{0}+W$, and the additive inverse $-(\mathbf{v}+W)$ of $\mathbf{v}+W$ is $-\mathbf{v}+W$. Properties such as associativity and commutativity of addition follow from corresponding properties in V.

To check the dimension formula, first choose a basis $\mathbf{w}_1,\dots,\mathbf{w}_k$ of W, and extend this to a basis

$$\mathbf{w}_1,\dots,\mathbf{w}_k,\mathbf{v}_1,\dots,\mathbf{v}_{n-k}$$

of V. Then I claim the cosets $\mathbf{v}_1+W,\dots,\mathbf{v}_{n-k}+W$ are a basis of V/W. To see they are independent, put $\mathbf{v}_i+W=\alpha_i$ if $1\le i\le n-k$, and suppose there exist $a_1,\dots,a_{n-k}\in\mathbb{F}$ such that $\sum a_i\alpha_i=\mathbf{0}+W$. This means that $\sum_{i=1}^{n-k}a_i\mathbf{v}_i\in W$. Hence there exist $b_1,\dots,b_k\in\mathbb{F}$ such that

$$\sum_{i=1}^{n-k}a_i\mathbf{v}_i=\sum_{j=1}^{k}b_j\mathbf{w}_j.$$

But the fact that the $\mathbf{v}_i$ and $\mathbf{w}_j$ comprise a basis of V implies that all a_i and b_j are zero. Therefore we have the independence. We leave the fact that $\alpha_1,\dots,\alpha_{n-k}$ span V/W as an exercise. □

Exercises

Exercise 5.29. Prove that the cosets $\alpha_1, \ldots, \alpha_{n-k}$ defined in the proof of Theorem 5.23 span V/W.

5.5 Summary

In the previous chapter, we introduced the notion of a vector space V over an arbitrary field. The purpose of this chapter was to learn some of the basic theory of vector spaces. The main topics we considered were the twin concepts of bases and dimension. A basis of V is a subset $\mathcal{B}$ of V such that every vector in V can be uniquely expressed as a linear combination of elements of $\mathcal{B}$. That is, $\mathcal{B}$ spans and is linearly independent. The main fact is that if V is finite dimensional (it is spanned by a finite subset), then any two bases have the same number of vectors. Thus the dimension of a finite dimensional V can be defined as the number of elements in a basis of V. There are two other ways of thinking about a basis. It is a minimal spanning set and a maximal linearly independent subset of V.

After we covered dimension theory, we considered several examples such as the row and column spaces of a matrix. These turned out to have the same dimension, a very surprising fact. We also constructed some new vector spaces and computed their dimensions. For example, if U and W are subspaces of V, we defined the sum $U + W$ which is a new subspace of V and computed $\dim(U + W)$. The answer is given by the Hausdorff Intersection Formula. We also defined what it means to say V is the direct sum of subspaces U and W and gave examples.

If W is a subspace of V, we may also form the quotient space V/W whose elements are called the cosets of W. Its dimension is $\dim V - dim W$. The notion of a quotient vector space uses the important fundamental idea of an equivalence relation. The idea of constructing a quotient vector space is a fundamental one, which is under constant use. Finally, we still need to derive some simple but messy formulas for changing basis. This will be done with great care in the next chapter.

Chapter 6

Linear Coding Theory

6.1 Introduction

The purpose of this chapter is to give an introduction to linear coding theory. This is a topic that is not usually treated in linear algebra, but perhaps it should be. The point is that coding theory is based on elemnetary linear algebra, but it uses the finite fields $\mathbb{F}_p$ instead of the reals $\mathbb{R}$. Coding theory is an extremely important topic because without it, we wouldn't have PCs, modems, compact discs, DVDs and many other of the daily necessities.

Before starting, let's give a little history one of the contributions coding theory has made. In the 1960's and 70's, NASA launched several of the Mariner space probes in order to gather information about our planetary system. One of main the problems the NASA engineers faced was how to send the data which was gathered back to earth. Data which has to be sent electronically is encoded as binary strings, that is, strings of 0's and 1's. Since the space probes carried only a tiny, weak transmitter, there was a high probability that the transmitted data could be scrampled or entirely lost, due to the fact that there is a lot of radiation in space capable of disrupting commumications. Solar flares, for example, routinely make communications even here on earth an impossibility.

To compensate for the possible damage to the transmitted binary strings, NASA used what are called error-correcting codes. In an error-correcting code, only certain of the strings 0's and 1's, called codewords, are used, so that for any received string which isn't a codeword, there may be a good choice as to which codeword should be substituted to restore the integrity of the transmission. For example, the Mariner probe to Venus used an error-correcting code consisting of 64 codewords. Each codeword was a string of

32 0's and 1's (thus the codewords are elements of $(\mathbb{F}_2)^{32}$). This code had the remarkably good property that it was able to correct an errant reception with to 7 errors. In other words, almost 25% of the digits could be off and the correct codeword would still be deducible.

For the reader who wishes to pursue coding theory more deeply, there are several elementary texts, such as *Introduction to Coding Theory* by R. Hill and *Introduction to the Theory of Error-Correcting Codes* by V. Pless. A more advanced book is *Applied Abstract Algebra* by R. Lidl and G. Pilz. Though more demanding, this book discusses many interesting applications of linear algebra besides coding theory. The web is also an excellent source of information. Just type your search topic into www.google.com.

6.2 Linear Codes

6.2.1 The Notion of a Code

The purpose of this section is to introduce the notion of a code. Recall that $V(n,p)$ denotes $\mathbb{F}^n$, where $\mathbb{F}$ is the prime field $\mathbb{F}_p$.

Definition 6.1. A *p-ary code of length n* is defined to be a subset of C of $V(n,p)$. The elements of C are called *codewords*. We will denote the number of elements of C by $|C|$.

Since $|V(n,p) = p^n$, every code $C \subset V(n,p)$ is finite.

Proposition 6.1. *The number of codes $C \subset V(n,p)$ is 2^{p^n}.*

Proof. The number of subsets of a set with k elements is 2^k, while $|V(np)| = p^n$. □

Definition 6.2. A linear subspace $C \subset V(n,p)$ is called a *linear code*. (or more precisely, a *p-ary linear code of length n*).

Thus a code $C \subset V(n,p)$ with the property that the sum of any two codewords is a codeword, which also contains the null word (i.e. zero vector) is a p-ary linear code. (Note that when the field is $\mathbb{F}_p$, a subset containing the null word which is closed under addition is a subspace.)

An important advantage of linear codes is that a linear code is determined by giving a set of codewords which span it.

Definition 6.3. If $C \subset V(n,p)$ is linear, then any set of codewords which gives a basis of C is called a set of *basic codewords*. If $\dim C = k$, we call C a *p-ary $[n,k]$-code*.

Proposition 6.2. *If C is a p-ary $[n,k]$-code, then $|C| = p^k$.*

Proof. This is a special case of Proposition 5.7, but let's repeat the proof anyway. Since $\dim C = k$, C has a basis consisting of k codewords, say $\mathbf{c}_1, \ldots \mathbf{c}_k$.

Now every codeword can be expressed in exactly one way as a linear combination

$$a_1\mathbf{c}_1 + a_2\mathbf{c}_2 + \cdots a_k\mathbf{c}_k,$$

where $a_1, a_2, \ldots, a_k$ vary over all elements of $\mathbb{F}_p$. Hence there are at most p^k possible linear combinations. But different linear combinations give different vectors, so in fact $|C| = p^k$. □

The most frequently used codes are *binary codes*, that is codes where $\mathbb{F} = \mathbb{F}_2$, so we will concentrate on these. The elements of $V(n,2)$ will be represented simply as strings of n 0's and 1's. We will frequently refer to these as *n-bit strings*. For example, the two-bit strings are 00, 01, 10, and 11.

Example 6.1. The equation $x_1+x_2+x_3+x_4 = 0$ defines a 4-bit linear code of dimension 3. Hence there are 8 codewords. Rewriting this equation as $x_1+x_2+x_3 = x_4$, we see that x_4 can be viewed as a check digit for x_1, x_2, x_3. In this code, the codewords are the 4 bit strings with an even number of 1's. A particular set of basic codewords is $\{1001, 0101, 0011\}$, although there are other possibly more natural choices.

Example 6.2. Let

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

and let C be the binary 6-bit linear code spanned by the rows of A. That is, $C = \mathrm{row}(A)$. Since A is in row reduced form, its rows are independent, hence form a set of basic codewords for C. Thus C is a three dimensional subspace of $V(6,2)$, so $|C| = 8$. The 7 non zero codewords are

$$(100111),\ (010101),\ (001011),\ (110010),\ (101100),\ (011110),\ (1111001).$$

Note that all possible combinations of 0's and 1's occur in the first three positions. These three letters tell you which linear combination of the basic codewords is involved. The last three letters are again check digits.

6.2.2 The International Standard Book Number

The International Standard Book Number (ISBN) is a reference number that is issued to books published by the mainstream publishing companies. Its purpose is to assist bookstores in making orders and to help librarians in cataloguing. The system has been in place since 1969. Each ISBN is a 10 digit string $a_1 \cdots a_9\ a_{10}$. The digits $a_1, \ldots, a_9$ are allowed to take any value between 0 and 9, but the last digit a_{10} can also take the value X, which is the Roman numeral denoting 10.

For example, the book *Fermat's Enigma* by Simon Singh, published in 1997 by Penguin Books, has ISBN 0-14-026869-3. The first digit 0 indicates that the book is in English, the digits between the first and second hyphens give the number assigned to the publisher, and the next set of digits indicates the title. The last digit is the check digit, which we will explain below. Major publishing companies like Penguin have small numbers (Penguin's is 14), while small publishers are given a larger number. Whitecap Books in Vancouver and Toronto has the 6 digit number 921061. Thus Penguin can publish 999,999 titles (in English), but Whitecap is restricted to 99.

ISBN's are based on a linear 11-ary [10,9] code, that is, a 9-dimensional linear subspace C of $V(10, 11)$. The code C is defined to be the solution space of the homogeneous linear equation in $a_1, \ldots, a_{10}$ given by

$$a_1 + 2a_2 + 3a_3 + \cdots + 9a_9 + 10a_{10} = 0.$$

Clearly, C can also be described as the null space of the rank one matrix $(1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10)$ over $\mathbb{F}_{11}$. Since $10+1 = 0$ in $\mathbb{F}_{11}$, the defining equation can also be expressed as

$$a_{10} = \sum_{i=1}^{9} i a_i.$$

The ISBN's are the codewords

$$a_1\ a_2\ a_3\ a_4 \ldots a_9\ a_{10}$$

described above (with hyphens inserted in appropriate places). Of course, not all 11^9 possible codewords can be used because of the restriction $a_i \neq 10$ except for a_{10}.

Example 6.3. For example, 0-15-551005-3 is an ISBN since $0 + 2 + 15 + 20 + 25 + 6 + 0 + 0 + 453 \equiv 3 \bmod(11)$, as is 0-14-026869-3 from the above example.

Example 6.4. Suppose that an ISBN is entered as 0-19-432323-1. With a minimum amount of technology, the machine in which the numbers are being entered will warn the librarian that 0-19-432323-1 is not an ISBN: that is, $(0, 1, 9, 4, 3, 2, 3, 2, 3)$ doesn't satisfy $\sum_{i=1}^{10} ia_i = 0$ in $\mathbb{F}_{11}$. Thus an error has been detected. But the type of error isn't. For example, there may be a single incorrect digit, or two digits might have been transposed. In fact, these two possibilities are the most common types of error. The next result says something about them.

Proposition 6.3. *A vector $\mathbf{a} = (a_1, \ldots, a_{10}) \in V(10, 11)$ that differs from an element of C in exactly one place cannot belong to C; in particular it cannot be an ISBN. Similarly, an element of $V(10, 11)$ obtained by transposing two unequal letters of an ISBN cannot be an ISBN.*

Proof. We will prove the first assertion but leave the second as an exercise. Suppose $\mathbf{c} = (c_1, \ldots, c_{10})$ is a codeword which differs from $\mathbf{a} \in V(10, 11)$ in one exactly component, say $c_i = a_i$ if $i \neq j$, but $c_j \neq a_j$. Then

$$\mathbf{v} := \mathbf{a} - \mathbf{c} = (0, \ldots, 0, a_j - c_j, 0, \ldots, 0).$$

If $\mathbf{a} \in C$, then $\mathbf{v} \in C$ too, hence $j(a_j - c_j) = 0$ in $\mathbb{F}_{11}$. But since neither j nor $a_j - c_j$ is zero in $\mathbb{Z}_{11}$, this contradicts the fact that $\mathbb{F}_{11}$ is a field. Hence $\mathbf{v} \notin C$, so $\mathbf{a} \notin C$ also. This proves the first assertion. □

Suppose you know all but the kth digit of an ISBN. Can you find the missing digit? Try this with an example, say 0-13-832x44-3. This is a sure way to astound your friends and relatives and maybe win a few bets. But don't bet with a librarian.

Exercises

Exercise 6.1. Determine all x such that 0-13-832x4-4 is an ISBN.

Exercise 6.2. Determine all x and y such that both 1-2-3832xy4-4 and 3-33-x2y377-6 are ISBNs.

Exercise 6.3. Prove the second assertion of Proposition 6.3.

6.3 Error detecting and correcting codes

6.3.1 Hamming Distance

In $\mathbb{R}^n$, the distance between two vectors is the square root of the sum of the squares of the differences of their components. This could never be used

to measure the distance between two elements of $V(n,p)$ since a sum of squares in $\mathbb{F}_p$ may well be 0. It turns out however that there is another way of measuring distances and lengths which works extremely well in the $V(n,p)$ setting.

Definition 6.4. Suppose $\mathbf{v} = (v_1, \ldots, v_n) \in V(n,p)$. Define the *weight* $\omega(\mathbf{v})$ of $\mathbf{v}$ to be the number of i such that $v_i \neq 0$. That is,

$$\omega(v_1, \ldots, v_n) = |\{i \mid v_i \neq 0\}|.$$

The *Hamming distance* (or simply the distance) $d(\mathbf{u}, \mathbf{v})$ between $\mathbf{u}$ and $\mathbf{v}$ in $V(n,p)$ is defined by

$$d(\mathbf{u}, \mathbf{v}) = \omega(\mathbf{u} - \mathbf{v}).$$

For example, $\omega(1010111) = 5$. Note that the only vector of weight zero is the zero vector. Therefore $\mathbf{u} = \mathbf{v}$ exactly when $\omega(\mathbf{u} - \mathbf{v}) = 0$. In fact what makes the Hamming distance function d so useful is that it satisfies the three properties which are used to characterize (or define) a distance function in general.

Proposition 6.4. *Suppose* $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V(n,p)$. *Then:*

(i) $d(\mathbf{u}, \mathbf{v}) \geq 0$, *and* $d(\mathbf{u}, \mathbf{v}) = 0$ *if and only if* $\mathbf{u} \neq \mathbf{v}$;

(ii) $d(\mathbf{u}, \mathbf{v}) = d(\mathbf{v}, \mathbf{u})$; *and*

(iii) $d(\mathbf{u}, \mathbf{w}) \leq d(\mathbf{u}, \mathbf{v}) + d(\mathbf{v}, \mathbf{w})$.

These properties clearly hold for the usual distance function on $\mathbb{R}^n$. Property (iii) is the *triangle inequality*, so named because in $\mathbb{R}^n$ it says that the length of any side of a triangle can't exceed the sum of the lengths of the other two sides. The first two properties of the Hamming distance are easy to see, but the triangle inequality requires proof.

Proof. First consider the case where $\mathbf{u}$ and $\mathbf{v}$ differ in every component. Thus $d(\mathbf{u}, \mathbf{v}) = n$. Let $\mathbf{w}$ be any vector in $V(n,p)$, and suppose $d(\mathbf{u}, \mathbf{w}) = k$. Then $\mathbf{u}$ and $\mathbf{w}$ agree in $n - k$ components, which tells us that $\mathbf{v}$ and $\mathbf{w}$ cannot agree in those $n - k$ components, so $d(\mathbf{v}, \mathbf{w}) \geq n - k$. Thus

$$d(\mathbf{u}, \mathbf{v}) = n = k + (n - k) \leq d(\mathbf{u}, \mathbf{w}) + d(\mathbf{v}, \mathbf{w}).$$

In the general case, let $\mathbf{u}, \mathbf{v}, \mathbf{w}$ be given. Now let $\mathbf{u}', \mathbf{v}'$ and $\mathbf{w}'$ denote the vectors obtained by dropping the components where $\mathbf{u}$ and $\mathbf{v}$ agree. Then we are in the previous case, so

$$d(\mathbf{u}, \mathbf{v}) = d(\mathbf{u}', \mathbf{v}') \leq d(\mathbf{u}', \mathbf{w}') + d(\mathbf{u}, \mathbf{w}').$$

But $d(\mathbf{u}', \mathbf{w}') \leq d(\mathbf{u}, \mathbf{w})$ and $d(\mathbf{v}', \mathbf{w}') \leq d(\mathbf{v}, \mathbf{w})$. Therefore,

$$d(\mathbf{u}, \mathbf{v}) \leq d(\mathbf{u}, \mathbf{w}) + d(\mathbf{v}, \mathbf{w}),$$

and the triangle inequality is established. □

One of the most desirable features for a (not necessarily linear) code C is that the minimum distance between two any codewords as large as possible. Let $d(C)$ denote this minimum distance. For a code which isn't linear, $d(C)$ has to be computed in the old fashioned way, that is the distance between every pair of codewords has to be taken into account. In general, if there are m codewords, then this means doing

$$\binom{m}{2} = \frac{m(m-1)}{2}$$

calculations (check this). But if C is linear, finding $d(C)$ takes much less.

Proposition 6.5. *If $C \subset V(n, 2)$ is linear, then $d(C)$ is the minimum of the weights of all the non zero codewords.*

We will leave the proof as an exercise.

6.3.2 The Main Result

An code $C \subset V(n, p)$ of length n such that $|C| = M$ and $d(C) = d$ is often called an (n, M, d)-code. If C is also linear, $M = p^k$ for some positive integer k. In that case, we say that C is a *p-ary $[n, k, d]$-code*. In general, one wants to maximize $d(C)$. The reason for this is given in the next result.

Proposition 6.6. *A (not necessarily linear) (n, M, d)-code C can detect up to $d - 1$ errors, i.e. if $d(\mathbf{v}, \mathbf{c}) \leq d - 1$ for some $\mathbf{c} \in C$, then $\mathbf{v} \notin C$. Moreoever, C corrects up to $e = (d - 1)/2$ errors. That is, if $d(\mathbf{v}, \mathbf{c}) \leq e$, for some codeword $\mathbf{c}$, then this $\mathbf{c}$ is the unique codeword with this property, and thus $\mathbf{c}$ corrects the errors in the non-codeword $\mathbf{v}$.*

The error-correcting assertion can be succinctly phrased by saying that any $\mathbf{v}$ within Hamming distance $e = (d - 1)/2$ of C is within e of a unique codeword. So if you know all but e digits of a codeword, you know them all.

Example 6.5. Suppose C is a 6-bit code with $d = 3$. Then $e = 1$. If $\mathbf{c} = 100110$ is a codeword, then $\mathbf{v} = 000110$ can't be one, but 100110 is the unique codeword within Hamming distance 1 of the non-codeword 000110.

We will leave the first assertion of Proposition6.6 as an exercise and prove the harder second assertion. Assume $d(\mathbf{v},\mathbf{c}) \leq (d-1)/2$, and suppose there exists an $\mathbf{c}' \in C$ such that $d(\mathbf{v},\mathbf{c}') \leq d(\mathbf{v},\mathbf{c})$. Thus,

$$d(\mathbf{v},\mathbf{c}') \leq d(\mathbf{c},\mathbf{v}) \leq (d-1)/2.$$

The idea is to use the triangle identity to estimate $d(\mathbf{c},\mathbf{c}')$, which we know is at least $d(C) = d$ if $\mathbf{c} \neq \mathbf{c}'$. But by the triangle inequality,

$$d(\mathbf{c},\mathbf{c}') \leq d(\mathbf{c},\mathbf{v}) + d(\mathbf{v},\mathbf{c}') \leq (d-1)/2 + (d-1)/2 = d-1,$$

so indeed we have, we have $\mathbf{c} = \mathbf{c}'$. □

For the binary [4,3]-code given by $x_1 + x_2 + x_3 + x_3 + x_4 = 0$, one sees easily that $d(C) = 2$. Thus C detects a single error, but can't correct an error because $(d-1)/2 = 1/2 < 1$. However, if some additional information is known, such as the component where the error occurs, it can be corrected using the linear equation defining the code.

6.3.3 Perfect Codes

We can also interpret Proposition 6.6 geometrically. If $r > 0$, define the *ball of radius r centred at $\mathbf{v} \in V(n,p)$* to be

$$B_e(\mathbf{v}) = \{\mathbf{w} \in V(n,p) \mid d(\mathbf{w},\mathbf{v}) \leq e\}. \tag{6.1}$$

Extending Proposition6.6, we show

Proposition 6.7. *Let $C \subset V(n,p)$ satisfy $d(C) = d$, and let $e = (d-1)/2$. For any $\mathbf{c} \in C$,*

$$B_e(\mathbf{c}) \cap C = \{\mathbf{c}\}.$$

Hence, an element $\mathbf{v} \in V(n,p)$ which lies in one of the balls $B_e(\mathbf{c})$ lies in exactly one of them.

Proof. This follows immediately from Proposition6.6. □

Of course, the Propositiondoesn't say much unless $d(C) > 2$. Thus the union of the balls $B_e(\mathbf{c})$ as $\mathbf{c}$ varies over C is the set of elements of $V(n,p)$ which are within e of a unique codeword. The nicest situation is that these balls cover $V(n,p)$, that is

$$V(n,p) = \bigcup_{\mathbf{c}\in C} B_e(\mathbf{c}). \tag{6.2}$$

Definition 6.5. A code $C \subset V(n,p)$ with $d(C) = d > 2$ is said to be *perfect* if (6.2) holds. That is, C is perfect if and only if every element of $V(n,p)$ is within e of a (unique) codeword.

We will consider perfect codes in more detail in §6.14. It turns out that perfect codes are not so abundant. There are infinitely many perfect binary linear codes C with $d(C) = 3$, hence $e = 1$. These single error correcting codes are known as *Hamming codes*. A result of Pless says that there are only two perfect linear codes with $e > 1$. One is a binary $[23, 12]$-code with $d = 7$ and the other is a ternary $[11, 6]$-code with $d = 5$.

6.3.4 A Basic Problem

One of the basic problems in coding theory is to design codes $C \subset V(n,p)$ such that both $|C|$ and $d(C)$ are large. More precisely, the problem is to maximize $|C|$ among all length codes C for which $d(C) \geq m$ for some given integer m. The maximum will then depend on n and m. If we also impose the condition that C is linear, then we are actually seeking to maximize $\dim(C)$, since $|C| = p^{\dim(C)}$. An example which has this property is the binary $(8, 16, 4)$ code C_8 defined in the next example.

Example 6.6. Consider the following matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

The row space C_8 of A is called the *extended Hamming code*. Notice that every row of A has weight 4, so the minimum distance of C_8 is at most 4. In fact, it can be shown that $d(C_8) = 4$. Hence C_8 is an $(8, 16, 4)$-linear code.

Proposition 6.8. *The code C_8 maximizes $|C|$ among all 8-bit binary linear codes with $d(C) \geq 4$.*

Proof. Since $\dim C_8 = 4$, we have to show that there are no 8-bit binary linear codes C with $d(C) \geq 4$ and $|C| > 16$. Suppose C is in fact one such code. Then by taking a spanning set for C as the rows of a $k \times 8$ matrix A, we can use row operations to put A into reduced row echelon form A_{red} without changing C. For simplicity, suppose that A_{red} has the form $(I_r|M)$. It follows that $|C| = 2^r$, so since $|C| > 16$, we see that $r \geq 5$. Hence M has at most three columns. Now the only way $d(C) \geq 4$ is if all entries of M

are 1. But then subtracting the second row of A_{red} from the first gives an element of C of weight 2, which contradicts $d(C) \geq 4$. Thus $r \leq 4$. □

In fact, by a similar argument, we can show the *singleton bound* for $d(C)$.

Proposition 6.9. *If C is a linear $[n,k]$-code, then*

$$d(C) \leq n - k + 1.$$

Put another way, a linear code C of length n satisfies

$$\dim C + d(C) \leq n + 1.$$

We leave the proof as an exercise. In the next section, we will consider a class of non-linear binary where both $|C|$ and $d(C)$ are large. Let us a make a final definition.

Definition 6.6. A linear $[n,k]$-code C with $d(C) = n - k + 1$ is said to be *maximal distance separating.*

6.3.5 Linear Codes Defined by Generating Matrices

The purpose of this subsection is to consider linear codes C given as the row space of a so called *generating matrix*. We already considered some examples of generating matrices in the last subsection.

Definition 6.7. A *generating matrix* for a linear $[n,k]$-code C is a $k \times n$ matrix over $\mathbb{F}_p$ of the form $M = (I_k \mid A)$ such that $C = \text{row}(M)$.

Example 6.7. Let C be the binary [4,2]-code with generating matrix

$$M = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Taking all the linear combinations of the rows, we find that

$$C = \{0000, 1011, 0101, 1110\}.$$

A check to make sure that we have found all of C is to note that since $\dim C = 2$ and $p = 2$, C has $4 = 2^2$ elements.

Example 6.8. Let

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

One checks easily that besides the rows of M, the elements of C are

$$(000000),\ (110010),\ (101100),\ (011110),\ (111001).$$

Clearly $d(C) = 3$, so C is one error-correcting .

The reader should also recall the code C_8 considered in Example 6.6. Every element of $C = \text{row}(M)$ can be expressed as a matrix product of the form $(x_1\ \dots\ x_k)M$. (To see this, transpose the fact that the column space of M^T consists of all vectors of the form $M^T(y_1\ \dots\ y_n)^T$.) Now, to any $\mathbf{x} = (x_1 \dots x_k) \in \mathbb{F}^k$, there is a unique codeword $\mathbf{c}(\mathbf{x}) = (x_1\ \dots\ x_k)M \in C$. For a generating matrix M as above,

$$\mathbf{c}(\mathbf{x}) = x_1 \dots x_k \sum_{i=1}^{k} a_{i1}x_i \cdots \sum_{i=1}^{k} a_{i(n-k)}x_i.$$

Since $x_1, \dots, x_k$ are completely arbitrary, the first k entries $x_1 \dots x_k$ are called the *message digits* and the last $n-k$ digits are called the *check digits*.

Exercises

Exercise 6.4. Prove the second assertion of Proposition 6.3.

Exercise 6.5. Prove the first two parts of Proposition6.5.

Exercise 6.6. Consider the binary code $C \subset V(6,2)$ which consists of 000000 and the following nonzero codewords:

$$(100111), (010101), (001011), (110010), (101100), (011110), (111001).$$

(i) Determine whether or not C is linear.

(ii) Compute $d(C)$.

(iii) How many elements of C are nearest to (011111)?

(iv) Determine whether or not 111111 is a codeword. If not, is there a codeword nearest 111111?

Exercise 6.7. Prove the first part of Proposition 6.6.

Exercise 6.8. Compute $d(C)$ for the code C of Example 6.2.

Exercise 6.9. Consider the binary code C_7 defined as the row space of the matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

in $V(7,2)$.

(i) Compute $d(C)$ and e.

(ii) Find the unique element of C that is nearest to 1010010. Do the same for 1110001.

Exercise 6.10. Let r be a positive integer and consider the ball $B_r(\mathbf{x}) \subset V(n,2)$ about $\mathbf{x} \in V(n,2)$. Show that

$$|B_r(\mathbf{x})| = \sum_{i=0}^{r} \binom{n}{i}.$$

Exercise 6.11. Generalize Exercise 6.10 from $V(n,2)$ to $V(n,p)$.

Exercise 6.12. * Show that if C is a linear $[2,k]$-code and C is e-error-correcting , then

$$\sum_{i=0}^{e} \binom{n}{i} \leq 2^{(n-k)}.$$

In particular, if C is 1-error-correcting , then $|C| \leq 2^{(n-k)}/(1+n)$.

Exercise 6.13. Show that if P is a permutation matrix, then P defines a transformation $T : V(n,p) \to V(n,p)$ which preserves the Hamming distance.

Exercise 6.14. Show that if C is a linear code, then

$$d(C) = \min\{\omega(\mathbf{x}) \mid \mathbf{x} \in C, \mathbf{x} \neq \mathbf{0}\}.$$

That is, $d(C)$ is the minimum weight among the non zero vectors in C. Use the result to find $d(C)$ for the code C used to define ISBN's? Is this code error-correcting ?

Exercise 6.15. Prove Proposition6.9. (Note,

Exercise 6.16. Taking $\mathbb{F} = \mathbb{F}_{11}$, compute the generating matrix for the ISBN code.

6.4 Hadamard matrices (optional)

We will next consider an interesting class of binary codes based on Hadamard matrices, which are named after the French mathematician J. Hadamard. As mentioned above, these so called *Hadamard codes* have the property that both $|C|$ and $d(C)$ have a large values (as opposed to what we saw in Proposition6.9 for linear codes). Hadamard matrices are themselves of interest, since their properties are not that well understood. As Hadamard codes are nonlinear, we consider this topic to be a sightseeing trip.

6.4.1 Hadamard Matrices

A *Hadamard matrix* is an $n \times n$ matrix H such that $h_{ij} = \pm 1$ for all $1 \leq i, j \leq n$ and

$$HH^T = nI_n.$$

Proposition 6.10. *If H is an $n \times n$ Hadamard matrix, then:*

(i) $H^T H = nI_n$,

(ii) any two distinct rows or any two distinct columns are orthogonal;

(iii) n is either 1, 2 or a multiple of 4; and

(iv) if $n > 1$, then any two rows of H agree in exactly $n/2$ places.

The only assertion that isn't clear is (iii), although it isn't hard to see that that n is even. It's still an open problem as to whether there is a $4k \times 4k$ Hadamard matrix for every $k > 0$. This is known for $k \leq 106$, but it doesn't seem to be known whether there is a 428×428 Hadamard matrix.

Example 6.9. Examples of $n \times n$ Hadamard matrices for $n = 2, 4, 8$ are

$$H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad H_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix},$$

and

$$H_8 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \end{pmatrix}.$$

After this point, it is no longer instructive to write them down. One can produce other Hadamard matrices from these by the transformation $H \mapsto PHQ$, where P and Q are permutation matrices.

6.4.2 Hadamard Codes

We will now define a Hadamard code. Let H be any $n \times n$ Hadamard matrix. Consider the $n \times 2n$ matrix $(H| - H)$, and let $\mathcal{H}$ be the binary matrix obtained by replacing all -1's by 0's.

Definition 6.8. The *Hadamard code* C associated to H is by definition the set of columns of $\mathcal{H}$. It is a binary n-bit code with $2n$-codewords.

Proposition 6.11. *Let C be an n-bit Hadamard code. Then $d(C) = n/2$. Thus C is a binary $(n, 2n, n/2)$-code.*

Proof. Recall that n is a multiple of 4, so $n/2$ is an even integer. The fact that the ith and jth columns of H are orthogonal if $i \neq j$ implies they must differ in exactly $n/2$ components since all the components are ± 1. But the ith and jth columns of H and $-H$ are also orthogonal if $i \neq j$, so they differ in $n/2$ places too. Moreover, the ith columns of H and $-H$ differ in n places. This proves $d(C) = n/2$, as asserted. □

For example, the Hadamard matrix H_2 gives the $(2, 4, 1)$-code

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

The code that was used in the transmission of data from the Mariner space probes to Venus in the 1970's was a binary $(32, 64, 16)$ Hadamard code. Since $(16 - 1)/2 = 7.5$, this code corrects 7 errors. An interesting footnote is that the Mariner space probes, now billions of miles from earth, are still transmitting data.

6.5 The Standard Decoding Table, Cosets and Syndromes

6.5.1 The Nearest Neighbour Decoding Scheme

Suppose that $C \subset V(n,p)$ is a p-ary linear n-bit code used in transmitting data from a satellite. Assume a codeword $\mathbf{c} = c_1 \dots c_n$ has been transmitted and received as $\mathbf{d} = d_1 \dots d_n$. Due to atmospheric interference and a number of other possible sources of noise, the received word $\mathbf{d}$ is not a codeword. The people manning the communication center therefore have to consider the *error* $\mathbf{E} := \mathbf{d} - \mathbf{c}$, which of course is unknown to them. The object is to make an intelligent guess as to what $\mathbf{c}$ is.

One popular scheme is the *nearest neighbour* decoding scheme, which uses a *standard decoding table* for C (SDT for short). The idea is to organize the elements of $V(n,p)$, into the cosets of C, which were defined in Definition 5.6. For the convenience of readers who skipped §5.4, we will recall the basic properties of cosets below. The cosets of C are disjoint subsets of $V(n,p)$ whose union is all of $V(n,p)$, such that any elements of the same coset have the same error. Cosets will turn out to be analogous to the set of all parallel planes in $\mathbb{R}^3$. This is illustrated in the following example.

Example 6.10. Let $C \subset V(4,2)$ be the linear code of Example 6.7. We will now construct an SDT for C. The SDT is a rectangular array listing all 2^4 elements of $V(4,2)$ as follows. The first row consists of the elements of C, putting $\mathbf{0} = 0000$ in the first column. To construct the second row, choose an element $\mathbf{E}_1$ not in C, and put $\mathbf{E}_1$ directly below 0000. Now add $\mathbf{E}_1$ to each element $\mathbf{c}$ in the first row to the right of 0000 and write the result $\mathbf{E}_1 + \mathbf{c}$ directly below $\mathbf{c}$. Thus the first row is the coset $\mathbf{0} + C$ of $\mathbf{0}$, and the second row is the coset $\mathbf{E}_1 + C$ of $\mathbf{E}_1$. Next, select a $\mathbf{E}_2 \in V(4,2)$ which isn't in either of the first two rows, assuming one exists, and repeat the previous step with $\mathbf{E}_2$. Continuing this construction will eventually exhaust $V(4,2)$, and the final result is the standard decoding table . This construction, however, may lead to many different standard decoding tables since there aren't any canonical choices for the error vectors that appear in the first column. Below is an example of a standard decoding table for C.

$$\begin{array}{cccc} 0000 & 1011 & 0101 & 1110 \\ 1000 & 0011 & 1101 & 0110 \\ 0100 & 1111 & 0001 & 1010 \\ 0010 & 1001 & 0111 & 1100 \end{array} .$$

Every row of an standard decoding table for a subspace $C \subset V(n,p)$ has the form $\mathbf{E} + C$ for some $\mathbf{E} \in V(n,p)$. The first row is $C = \mathbf{0} + C$, and the potential errors $\mathbf{E}_i$ we've selected vary through the first column. One obvious comment is that it makes sense to choose errors of minimal weight. If a codeword $\mathbf{c}$ has been transmitted but a non-codeword such as 0111 is received, then scan the standard decoding table until 0111 is located. In the example, 0111 occurs in the last row directly below the codeword 0101. The nearest neighbour decoding scheme assumes that the error is the leading element 0010 of the last row, so the correct codeword is $0101 = 0111 - 0010$.

Notice that it can happen that a row of a standard decoding table contains more than one element of minimal weight. This happens in the third row of the above table, where there are two elements of weight one. There is no reason to prefer decoding 1111 as 1011 rather than 1110. The non zero elements of least nonzero weight in $V(n,2)$ are standard basis vectors. If $\dim C = k$, then at most k of the standard basis vectors can lie in C. These vectors are therefore natural candidates for the leading column. In fact, it seems desirable to seek codes C so that there is a standard decoding table such that in each row, there is a unique vector of minimal length. We will see presently that this objective is achieved by perfect linear codes.

6.5.2 Cosets

From an inspection of the above standard decoding table, three properties are apparent:

(a) different rows don't share any common elements;

(b) any two rows have the same number of elements; and

(c) every element of $V(4,2)$ is in some row.

These properties follow from the fact that the rows of a standard decoding table are the *cosets* of C.

Definition 6.9. Let V be a vector space over $\mathbb{F}$, and suppose A and B are subsets of V. We define $A + B$ to be the subset consisting of all vectors of the form $a + b$, where $a \in A$ and $b \in B$. If C is a subspace of V, then a subset of V of the form $\{\mathbf{v}\} + C$ is called a *coset of* C.

To simplify the notation, we will denote $\{\mathbf{v}\} + C$ by $\mathbf{v} + C$. For example, each row of a standard decoding table is a coset of the linear code C, since it has the form $\mathbf{E} + C$. The properties (a), (b) and (c) stated above all follow from

Proposition 6.12. *Let V be a vector space over $\mathbb{F}_p$ of dimension n, and let C be a linear subspace of V of dimension k. Every element of V lies in a coset of C, and two cosets are either disjoint or equal. In fact, $\mathbf{v}+C = \mathbf{w}+C$ if and only if $\mathbf{w}-\mathbf{v} \in C$. Finally, there are $p^{(n-k)}$ cosets of C, every coset contains p^k elements.*

Proof. Certainly $\mathbf{v} \in \mathbf{v}+C$, so the first claim is true. If $\mathbf{w}+C$ and $\mathbf{v}+C$ contain an element $\mathbf{y}$, then $\mathbf{y} = \mathbf{w}+\mathbf{c} = \mathbf{v}+\mathbf{d}$ for some $\mathbf{c}, \mathbf{d} \in C$. Thus $\mathbf{w} = \mathbf{v}+\mathbf{c}-\mathbf{d}$. Since $\mathbf{c}-\mathbf{d} \in C$, it follows that $\mathbf{w}+C = \mathbf{v}+(\mathbf{c}-\mathbf{d})+C$. But since C is a subspace, $(\mathbf{c}-\mathbf{d})+C = C$, so $\mathbf{w}+C = \mathbf{v}+C$. This proves the second claim. If $\mathbf{v}+C = \mathbf{w}+C$, then $\mathbf{w}-\mathbf{v} \in C$, and conversely. To prove the last assertion, recall that $|C| = p^k$. Hence $|\mathbf{v}+C| = p^k$ too. For $\mathbf{v} \to \mathbf{v}+\mathbf{c}$ is a bijection from C to $\mathbf{v}+C$. It follows that there are $p^{(n-k)}$ cosets, which completes the proof. □

In coding theory, the error elements $\mathbf{E}$ in the first column of a particular standard decoding table are sometimes called *coset leaders*, although in other contexts, they are known as coset representatives..

6.5.3 Syndromes

One can modifying the construction of an standard decoding table so that it isn't necessary to scan the whole table to find a given entry. This is important since scanning is an inefficient process in terms of computer time. The amount of scanning can be greatly reduced by using syndromes. The simplification come about by introducing the notion of a parity check matrix.

Definition 6.10. Let $C \subset V(n,p)$ be a linear code given by a generating matrix $M = (I_k \mid A)$. A *parity check matrix* for C is an $(n-k)\times n$ matrix H such that C is the null space of H after identifying row and column vectors.

One of the advantages of using a generating matrix in the form $M = (I_k \mid A)$ is that the parity check matrix H is simple to write down.

Proposition 6.13. *Suppose $C \subset V(n,p)$ is the code obtained as the row space of a generating matrix $M = (I_k \mid A)$. Then $\mathbf{c} \in C$ if and only if $(-A^T \mid I_{n-k})\mathbf{c}^T = \mathbf{0}$. Thus, $H = (-A^T \mid I_{n-k})$ is a parity check matrix for C. Furthermore, two vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ in $V(n,p)$ are in the same coset of C if and only if $H\mathbf{d}_1^T = H\mathbf{d}_2^T$.*

We will leave the proof that H is a parity check matrix as an exercise. When C is a binary code, the parity check matrix $H = (A^T \mid I_{n-k})$, since $-A^T = A^T$.

We now give the proof of the second assertion. Note that $\mathbf{d}_1$ and $\mathbf{d}_2$ are in the same coset of C if and only if $\mathbf{d}_1 - \mathbf{d}_2 \in C$ if and only if $H(\mathbf{d}_1 - \mathbf{d}_2)^T = \mathbf{0}$ if and only if $H\mathbf{d}_1^T = H\mathbf{d}_2^T$. □

Definition 6.11. We call $\mathbf{d}H^T = (H\mathbf{d}^T)^T$ the *syndrome* of $\mathbf{d} \in V(n,p)$ with respect to C.

To incorporate syndromes in a standard decoding table , we insert an extra column consisting of the syndromes of the cosets. Thus each row consists of a coset and the syndrome of that coset. The different syndromes identify the different cosets, so instead of having to scan the whole decoding table to find $\mathbf{d}$, it suffices to first scan the column of syndromes to find the syndrome of $\mathbf{d}$ and then scan that row.

Example 6.11. Let M be the generating matrix of Example 6.7. Recall

$$M = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Thus

$$H^T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The syndromes are found by taking the matrix product

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Thus the standard decoding table with syndromes is

0000	1011	0101	1110	00
1000	0011	1101	0110	11
0100	1111	0001	1010	01
0010	10001	0111	1100	10

Exercises

Exercise 6.17. Prove Proposition 6.12.

Exercise 6.18. Prove Proposition **??**.

Exercise 6.19. Construct the standard decoding table with syndromes for the binary code C with generating matrix

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

(Note, this is rather complicated since C has 8 elements. Thus the standard decoding table with syndromes is a 8×9 table (counting the syndromes as one column) since $V(6,2)$ has 64 elements. Perhaps you can write a program.)

Exercise 6.20. Let C be the code of the previous problem.

(a) How many errors does C detect?

(b) How many errors does it correct?

(c) Use the standard decoding table with syndromes you constructed in Exercise 6.20 to decode 101111 and 010011.

Exercise 6.21. Show that, indeed, $C = \{\mathbf{c} \in V(n,2) \mid \mathbf{c}H^T = \mathbf{0}\}$. (Suggestion: begin by showing that $MH^T = \mathbf{O}$. Hence every row of M lies in the left null space of H. Now compute the dimension of the left null space of H, using the fact that H and H^T have the same rank.)

Exercise 6.22. Construct the standard decoding table for the binary code with generating matrix

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Exercise 6.23. Let C be a binary linear n-bit code with $n \geq 3$ with parity check matrix H. Show that if no two columns of H are dependent (i.e. equal), then $d(C) \geq 3$.

Exercise 6.24. Generalize Exercise 6.23 by showing that if $C \subset V(n,p)$ is a linear code with a parity check matrix H having the property that no m columns of H are linearly dependent, then $d(C) \geq m+1$.

Exercise 6.25. Show that any coset of the ISBN contains a unique standard basis vector. In particular, any 10 digit number differs form an ISBN in one digit. Can you determine in general when it is possible, for a given C, to have a standard decoding table where the coset leaders are the standard basis vectors?

Exercise 6.26. Find an upper bound on the number of operations required to scan the standard decoding table (with syndromes) associated to an *p-ary* $[n, k]$ code to find any $\mathbf{d} \in V(n, 2)$. Compare this result to the number of operations needed to find $\mathbf{d}$ before adding the syndromes.

6.6 Perfect linear codes

Recall from Definition 6.5 that a code $C \subset V(n,p)$ with $d(C) = d > 2$ such that the balls $B_e(\mathbf{c})$ cover $V(n,p)$ as $\mathbf{c}$ ranges through C is called perfect. The first thing we will do in this section is to reformulate this definition for linear codes. It turns out that each coset of a perfect linear code has a unique element of weight less than $e = (d-1)/2$. As we mentioned in the previous section, since the coset leaders of the rows of a standard decoding table represent the common error for its entire coset (i.e. row), one would like the standard decoding table to have the property that the weight of each coset leader is strictly less than the weight of all other elements of its row. This property can't always be arranged, since there can be several minimal weight vectors in a coset, as we saw in Example 6.10.

Proposition 6.14. *A linear code $C \subset V(n,p)$ with $d > 2$ is perfect if and only if every coset $\mathbf{x} + C$ of C contains a unique element of $B_e(\mathbf{0})$.*

Proof. Suppose C is perfect. Then by definition, every coset $\mathbf{x}+C$ contains an element of $B_e(\mathbf{c})$. We need to show that a coset $\mathbf{x}+C$ can't contain two elements of $B_e(\mathbf{0})$. That is, we have to show that if $\mathbf{x}, \mathbf{y} \in B_e(\mathbf{0})$ and $\mathbf{x} \neq \mathbf{y}$, then $\mathbf{x} + C \neq \mathbf{y} + C$. But if $\mathbf{x}, \mathbf{y} \in B_e(\mathbf{0})$, the triangle inequality gives

$$d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{0}) + d(\mathbf{y}, \mathbf{0}) \leq 2e < d.$$

It follows that $\mathbf{x} \neq \mathbf{y}$, then $\mathbf{x} + C \neq \mathbf{y} + C$. Indeed, if $\mathbf{x} + C = \mathbf{y} + C$, then $\mathbf{x} - \mathbf{y} \in C$, so

$$d(\mathbf{x}, \mathbf{y}) = \omega(\mathbf{x} - \mathbf{y}) \geq d.$$

Therefore distinct elements of $B_e(\mathbf{0})$ give distinct cosets, so each coset contains a unique element of $B_e(\mathbf{0})$. To prove the converse, let $\mathbf{x} \in V(n,p)$ be arbitrary, and consider its coset $\mathbf{x} + C$. By assumption, $\mathbf{x} + C$ meets $B_e(\mathbf{0})$ in a unique element, say $\mathbf{x} + \mathbf{c}$. It follows that $\mathbf{x} \in B_e(-\mathbf{c})$, so since $\mathbf{x}$ is arbitrary, (6.2) holds. Thus C is perfect, and the proof is finished. □

Example 6.12. The binary code $C_3 = \{000, 111\}$ is perfect. Note $d = 3$ so $e = 1$. Thus the perfection is clear since any element of $V(3,2)$ is one unit away from a codeword (namley 000 if it has two 0's and 111 if it has two 1's. The generating matrix of C_3 is $M = (I_1|1\,1)$. Thus the parity check matrix is

$$H = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

The syndromes are given by the product

$$\begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix},$$

which gives the standard decoding table with syndromes as

000	111	00
100	011	11
010	101	10
011	110	01

We will give a less trivial example in the next subsection. Let's also make a couple of useful observations. For example, perfect codes $C \subset V(n,p)$ with $d = 3$ or 4 and hence $e = 1$ have the property that every vector is of distance one from a unique codeword, which is the minimal possible distance. In particular, if C is also linear, there exists an obvious choice for the standard decoding table , the one for which the coset leaders lie in $B_e(\mathbf{0})$. But the elements of $B_e(\mathbf{0})$ of weight 1 are the $p-1$ multiples (using $\mathbb{F}_p \setminus \{0\}$) of the standard basis vectors of $V(n,p)$. This fact has the consequence that for any standard decoding table for C with syndromes, the coset of any $\mathbf{v} \in V(n,p)$ can be immediately located by comparing its syndrome with the syndromes of the standard basis vectors and their multiples.

6.6.1 Testing for perfection

It turns out that there is a simple way of testing when a binary linear code is perfect.

Proposition 6.15. *Suppose $C \subset V(n,2)$ is a linear code with $\dim C = k$. Then C is perfect if and only if $|B_e(\mathbf{0})| = 2^{(n-k)}$. Put another way, C is perfect if and only if*

$$\sum_{i=0}^{e} \binom{n}{i} = 2^{(n-k)}.$$

In particular, if $e = 1$, then C is perfect if and only if

$$(1+n)2^k = 2^n.$$

Proof. The first statement follows from the fact that $2^{(n-k)}$ is the number of cosets of C. But we already saw that the number of elements in $B_e(\mathbf{0})$ is

given by

$$|B_e(x)| = \sum_{i=0}^{e} \binom{n}{i},$$

so the second statement follows. Applying this formula to the case $e = 1$ gives the second assertion. □

Notice that $|B_e(\mathbf{0})|$ has nothing to do with C. The problem of finding a perfect code is to determine n and k so that $|B_e(\mathbf{0})| = 2^{(n-k)}$ and there exists a k-dimensional subspace C of $V(n, 2)$ with $d(C) = 2e + 1$. If a perfect linear n-bit code C with $e = 1$ exists, then $n = 2^k - 1$ for some m and $n - m = \dim C$. Some possible solutions for these conditions are $n = 3$, $k = 2$ and $n = 7$, $k = 3$. We saw an example of the first case. The next example shows that a perfect code in the latter case (with $n = 7$ and $k = 3$) can be realized.

Example 6.13. Consider the 7-bit code C_7 defined as the row space of the matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

By enumerating the 16 elements of C_7, one sees that $d(C_7) = 3$, so $e = 1$. Since $(7 + 1)2^4 = 2^7$, C_7 is perfect.

6.6.2 The hat problem

The hat problem is an example of an instance where the existence of a particular mathematical structure, in this case, perfect codes with $e = 1$ has a surprising application. Beginning with a simple case, let us describe the hat game. Suppose there are three players each wearing either a white hat or a black hat. Each player can see the hats of the other two players, but cannot see what color her own hat is. Furthermore the players are in sound proof booths and cannot communicate with each other. Each booth has three buttons marked B,W and P (for pass or no guess). At the sound of the buzzer, each player presses one of her three buttons. If nobody guesses incorrectly, and at least one player guesses correctly, then they share a $1,000,000 prize. The problem is this: assuming the players are allowed to formulate their strategy beforehand, how should they proceed to maximize their chances of winning?

Clearly, a pretty good strategy would be to have two players abstain and to have the third make a random guess. Their probability of winning with this strategy is a not bad 1/2. But this strategy doesn't make any use of fact that each player can see the hats of her two teamates. Suppose the following strategy is adopted: if a player sees that the other two players have the same colored hat, she guesses the opposite color. A player who sees different colored hats passes. With this strategy, the only losing hat configurations are BBB or WWW, so they win six out of eight times. Hence the probability of winning is at least a fantastic 3/4.

What does this have to do with perfect codes such that $e = 1$? If we represent Black by 0 and White by 1, then the various hat arrays are represented by the $2^3 = 8$ 3-bit strings. Let C be the 3-bit code $\{000, 111\}$. Thus C is a perfect code with $e = 1$. The above strategy amounts to the following. The three contestants agree ahead of time to assume that the hat configuration isn't in C. The probabilty of this happening is 3/4 since 6 out of 8 configurations aren't in C. Suppose that this assumption is correct. Then two players will see a 0 and a 1. They should automatically pass since there is no way of telling what their hat colors are. The third will see either two 0's or two 1's. If she sees two 0's, then (by assumption) she knows her hat is white and she hits the button marked W (for white). If she sees two 1's, then (by assumption) she knows her hat is black, and she hits the button marked B. This strategy fails only when the configuration lies in C.

Next, let's suppose there are 7 players imaginatively labelled $1, \ldots, 7$. If we still represent Black by 0 and White by 1, then the various hat arrays are represented by the 2^7 7-bit strings. Let's see if the strategy for three hats still works with seven hats. First, all seven players need to memorize the 16 codewords of C_7. The players assume before the game starts to assume that the hat array isn't in C_7. Since $|C_7| = 2^4$, the probability that the hat array is in C_7 is $2^4/2^7 = 1/8$. Suppose (as for three hats) that their assumption is correct. Then the hat array $x_1 \ldots x_7$ differs in one place from a codeword $c_1 \ldots c_7$. Suppose this occurs at x_1. Then $x_2 = c_2, \ldots, x_7 = c_7$. So player #1 sees $c_2 \ldots c_7$ and recognizes that her hat color must be $c_1 + 1$ and guesses accordingly. Player #2 sees $x_1 c_3 \ldots c_7$. But since $d(C_7) = 3$, she knows that whatever x_2 is, $x_1 x_2 c_3 \ldots c_7 \notin C$. Therefore, she has to pass, as do the other five contestants. the chances that they win the million bucks are pretty good (7/8).

Can you devise a strategy for how to proceed if there are 4,5 or 6 players? What about 8 or 9? More information about this problem and other related (and more serious) problems can be found in the article *The hat problem and Hamming codes* by M. Bernstein in Focus Magazine, November 2001.

Exercises

Exercise 6.27. Construct the parity check matrix and syndromes for C_7.

Exercise 6.28. Consider the code $C = \{00000, 11111\} \subset V(5, 2)$.

(i) Determine e.

(ii) Show that C is perfect.

(iii) Does C present any possibilities for a five player hat game?

Exercise 6.29. Show that any binary $[23, 12]$-code with $d = 7$ is perfect.

Exercise 6.30. Show that any binary $[2^k - 1, 2^k - 1 - k]$-code with $d = 3$ is perfect. Notice C_7 is of this type.

Chapter 7

Linear Transformations

In this Chapter, we will define the notion of a linear transformation between two vector spaces V and W which are defined over the same field and prove the most basic properties about them, such as the fact that in the finite dimensional case is that the theory of linear transformations is equivalent to matrix theory. We will also study the geometric properties of linear transformations.

7.1 Definitions and examples

Let V and W be two vector spaces defined over the same field $\mathbb{F}$. To define the notion of a linear transformation $T : V \to W$, we first of all, need to define what a transformation is. A *transformation* $F : V \to W$ is a rule which assigns to every element $\mathbf{v}$ of V (the domain of F) a unique element $\mathbf{w} = F(\mathbf{v})$ in W. We will call W the *target* of F. We often call F a mapping or a vector valued function. If $V = \mathbb{F}^n$ and $W = \mathbb{F}^m$, then a transformation $F : \mathbb{F}^n \to \mathbb{F}^m$ is completely determined by component functions $f_1, \ldots, f_m$ which satisfy

$$F(x_1, x_2, \ldots, x_n) = (f_1(x_1, x_2, \ldots, x_n), f_2(x_1, x_2, \ldots, x_n), \ldots$$

$$\ldots, f_m(x_1, x_2, \ldots, x_n)).$$

The same is true if V and W are finite dimensional vector spaces, since we can choose finite bases of V and W and imitate the above construction where the bases are the standard ones.

7.1.1 The Definition of a Linear Transformation

From the algebraic viewpoint, the most interesting transformations are those which preserve linear combinations. These are called *linear transformations*. We will also, on occasion, call linear transformations linear maps.

Definition 7.1. Suppose V and W are vector spaces over a field $\mathbb{F}$. Then a transformation $T : V \to W$ is said to be *linear* if

(1) for all $\mathbf{x}, \mathbf{y} \in V$, $T(\mathbf{x} + \mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y})$, and

(2) for all $r \in \mathbb{F}$ and all $\mathbf{x} \in V$, $T(r\mathbf{x}) = rT(\mathbf{x})$.

It's obvious that a linear transformation T preserves linear combinations: i.e. for all $r, s \in \mathbb{F}$ and all $\mathbf{x}, \mathbf{y} \in V$

$$T(r\mathbf{x} + s\mathbf{y}) = rT(\mathbf{x}) + sT(\mathbf{y}).$$

Another obvious property is that for any linear transformation $T : V \to W$, $T(\mathbf{0}) = \mathbf{0}$. This follows, for example, from the fact that

$$T(\mathbf{x}) = T(\mathbf{x} + \mathbf{0}) = T(\mathbf{x}) + T(\mathbf{0})$$

for any $\mathbf{x} \in V$. This can only happen if $T(\mathbf{0}) = \mathbf{0}$.

7.1.2 Some Examples

Example 7.1. Let V be any vector space. Then the *identity transformation* is the transformation $Id : V \to V$ defined by $Id(\mathbf{x}) = \mathbf{x}$. The identity transformation is obviously linear.

Example 7.2. If $\mathbf{a} \in \mathbb{R}^n$, the dot product with $\mathbf{a}$ defines a linear transformation $T_{\mathbf{a}} : \mathbb{R}^n \to \mathbb{R}$ by $T_{\mathbf{a}}(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x}$. It turns out that any linear transformation $T : \mathbb{R}^n \to \mathbb{R}$ has the form $T_{\mathbf{a}}$ for some $\mathbf{a}$.

Example 7.3. A linear transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ of the form

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \lambda x \\ \mu y \end{pmatrix},$$

where λ and μ are scalars, will be called a *diagonal transformation*. Since $T(\mathbf{e}_1) = \lambda \mathbf{e}_1$ and $T(\mathbf{e}_2) = \mu \mathbf{e}_2$, whenever both λ and μ are nonzero, T maps a rectangle with sides parallel to $\mathbf{e}_1$ and $\mathbf{e}_2$ onto another such rectangle whose sides have been dilated by λ and μ and whose area has been changed by

the factor $|\lambda\mu|$. Such diagonal transformations also map circles to ellipses. For example, let C denote the unit circle $x^2 + y^2 = 1$, and put $w = \lambda x$ and $z = \mu y$. Then if $\lambda \neq \mu$, the image of T is the ellipse

$$(\frac{w}{\lambda})^2 + (\frac{z}{\mu})^2 = 1.$$

More generally, we will call a linear transformation $T : V \to V$ *diagonalizable* if there exist a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of V such that $T(\mathbf{v}_i) = \lambda_i \mathbf{v}_i$ for each index i, where $\lambda_i \in \mathbb{F}$. Diagonalizable linear transformations will also be called *semi-simple*. It turns out that one of the main problems in the theory of linear transformations is how to determine when a linear transformation is diagonalizable. This question will be taken up when we study eigentheory.

FIGURE
(DIAGONAL TRANSFORMATION)

Example 7.4. The cross product gives a pretty example of a linear transformation on $\mathbb{R}^3$. Let $\mathbf{a} \in \mathbb{R}^3$ and define $C_\mathbf{a} : \mathbb{R}^3 \to \mathbb{R}^3$ by

$$C_\mathbf{a}(\mathbf{v}) = \mathbf{a} \times \mathbf{v}.$$

Notice that $C_\mathbf{a}(\mathbf{a}) = \mathbf{0}$, and that $C_\mathbf{a}(\mathbf{x})$ is orthogonal to $\mathbf{a}$ for any $\mathbf{x}$. The transformation $C_\mathbf{a}$ is used in mechanics to express angular momentum.

Example 7.5. Suppose $V = C(a, b)$, the space of continuous real valued functions on the closed interval $[a, b]$. Then the definite integral over $[a, b]$ defines a linear transformation

$$\int_a^b : V \to \mathbb{R}$$

by the rule $f \mapsto \int_a^b f(t)\mathrm{d}t$. The assertion that $\int_a^b$ is a linear transformation is just the fact that for all $r, s \in \mathbb{R}$ and $f, g \in V$,

$$\int_a^b (rf + sg)(t)\mathrm{d}t = r\int_a^b f(t)\mathrm{d}t + s\int_a^b g(t)\mathrm{d}t.$$

This example is the analogue for $C(a, b)$ of the linear transformation $T_\mathbf{a}$ on $\mathbb{R}^n$ defined in Example 7.3, where $\mathbf{a}$ is the constant function 1, since, by definition,

$$\int_a^b f(t)\mathrm{d}t = (f, 1).$$

Example 7.6. Let V be a vector space over $\mathbb{F}$, and let W be a subspace of V. Let $\pi : V \to V/W$ be the map defined by

$$\pi(\mathbf{v}) = \mathbf{v} + W.$$

We call π the *quotient map*. Then π is a linear map. We leave the details as an exercise.

7.1.3 The Algebra of Linear Transformations

Linear transformations may be added using pointwise addition, and they can be multiplied by scalars in a similar way. That is, if $F, G : V \to W$ are two linear transformations, we form their sum $F + G$ by setting

$$(F + G)(\mathbf{v}) = F(\mathbf{v}) + G(\mathbf{v}).$$

If $a \in \mathbb{F}$, we put

$$(aF)(\mathbf{v}) = aF(\mathbf{v}).$$

Thus, we can take linear combinations of linear transformations, where the domain and target are two $\mathbb{F}$ vector spaces V and W respectively.

Proposition 7.1. *Let V and W be vector spaces over $\mathbb{F}$. Then any linear combination of linear transformations with domain V and target W is also linear. In fact, the set $L(V, W)$ of all linear transformations $T : V \to W$ is a vector space over $\mathbb{F}$.*

Proposition 7.2. *Suppose $\dim V = n$ and $\dim W = m$. Then $L(V, W)$ has finite dimension mn.*

Exercises

Exercise 7.1. Show that every linear function $T : \mathbb{R} \to \mathbb{R}$ has the form $T(x) = ax$ for some $a \in \mathbb{R}$.

Exercise 7.2. Determine whether the following are linear or not:

(i) $f(x_1, x_2) = x^2 - x_2$.

(ii) $g(x_1, x_2) = x_1 - x_2$.

(iii) $f(x) = e^x$.

Exercise 7.3. Prove the following:

Proposition 7.3. *Suppose $T : \mathbb{F}^n \to \mathbb{F}^m$ is an arbitrary transformation and write*

$$T(\mathbf{v}) = (f_1(\mathbf{v}), f_2(\mathbf{v}), \ldots, f_m(\mathbf{v})).$$

Then T is linear if and only if each component f_i is a linear function. In particular, T is linear if and only if there exist $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ in $\mathbb{F}^n$ such that for all $\mathbf{v} \in \mathbb{F}^n$,

$$T(\mathbf{v}) = (\mathbf{a}_1 \cdot \mathbf{v}, \mathbf{a}_2 \cdot \mathbf{v}, \ldots, \mathbf{a}_m \cdot \mathbf{v}).$$

Exercise 7.4. Prove Proposition 7.2.

Exercise 7.5. Let V be a vector space over $\mathbb{F}$, and let W be a subspace of V. Let $\pi : V \to V/W$ be the quotient map defined by $\pi(\mathbf{v}) = \mathbf{v} + W$. Show that π is linear.

Exercise 7.6. Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be a linear map with matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. The purpose of this exercise is to determine when T is linear over $\mathbb{C}$. That is, since, by definition, $\mathbb{C} = \mathbb{R}^2$ (with complex multiplication), we may ask when $T(\alpha\beta) = \alpha T(\beta)$ for all $\alpha, \beta \in \mathbb{C}$. Show that a necessary and sufficient condition is that $a = d$ and $b = -c$.

7.2 Matrix Transformations and Multiplication

7.2.1 Matrix Linear Transformations

Every $m \times n$ matrix A over $\mathbb{F}$ defines linear transformation $T_A : \mathbb{F}^n \to \mathbb{F}^m$ via matrix multiplication. We define T_A by the rule $T_A(\mathbf{x}) = A\mathbf{x}$. If we express A in terms of its columns as $A = (\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_n)$, then

$$T_A(\mathbf{x}) = A\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{a}_i.$$

Hence the value of T_A at $\mathbf{x}$ is the linear combination of the columns of A which is the ith component x_i of $\mathbf{x}$ as the coefficient of the ith column $\mathbf{a}_i$ of A. The distributive and scalar multiplication laws for matrix multiplication imply that T_A is indeed a linear transformation.

In fact, we will now show that every linear transformations from $\mathbb{F}^n$ to $\mathbb{F}^m$ is a matrix linear transformation.

Proposition 7.4. *Every linear transformation $T : \mathbb{F}^n \to \mathbb{F}^m$ is of the form T_A for a unique $m \times n$ matrix A. The ith column of A is $T(\mathbf{e}_i)$, where $\mathbf{e}_i$ is the ith standard basis vector, i.e. the ith column of I_n.*

Proof. The point is that any $\mathbf{x} \in \mathbb{F}^n$ has the unique expansion

$$\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}_i,$$

so,

$$T(\mathbf{x}) = T\big(\sum_{i=1}^{n} x_i \mathbf{e}_i\big) = \sum_{i=1}^{n} x_i T(\mathbf{e}_i) = A\mathbf{x},$$

where A is the $m \times n$ matrix $\big(T(\mathbf{e}_1) \dots T(\mathbf{e}_n)\big)$. If A and B are $m \times n$ and $A \neq B$, then $A\mathbf{e}_i \neq B\mathbf{e}_i$ for some i, so $T_A(\mathbf{e}_i) \neq T_B(\mathbf{e}_i)$. Hence different matrices define different linear transformations, so the proof is done. qed

Example 7.7. For example, the matrix of the identity transformation $Id : \mathbb{F}^n \to \mathbb{F}^n$ is the identity matrix I_n.

A linear transformation $T : \mathbb{F}^n \to \mathbb{F}$ is called a *linear function*. If $a \in \mathbb{F}$, then the function $T_a(x) := ax$ is a linear function $T : \mathbb{F} \to \mathbb{F}$; in fact, every such linear function has this form. If $\mathbf{a} \in \mathbb{F}^n$, set $T_{\mathbf{a}}(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x} = \mathbf{a}^T\mathbf{x}$. Then we have

Proposition 7.5. *Every linear function $T : \mathbb{F}^n \to \mathbb{F}$ has $T_{\mathbf{a}}$ for some $\mathbf{a} \in \mathbb{F}^n$. That is, there exist $a_1, a_2, \ldots, a_n \in \mathbb{F}$ such that*

$$T\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \mathbf{a} \cdot \mathbf{x} = \sum_{i=1}^{n} a_i x_i.$$

Proof. Just set $a_i = T(\mathbf{e}_i)$. □

Example 7.8. Let $\mathbf{a} = (1, 2, 0, 1)^T$. Then the linear function $T_{\mathbf{a}} : \mathbb{F}^4 \to \mathbb{F}$ has the explicit form

$$T_{\mathbf{a}}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = x_1 + 2x_2 + x_4.$$

7.2.2 Composition and Multiplication

So far, matrix multiplication has been a convenient tool, but we have never given it a natural interpretation. For just such an interpretation, we need to consider the operation of composing transformations. Suppose $S : \mathbb{F}^p \to \mathbb{F}^n$ and $T : \mathbb{F}^n \to \mathbb{F}^m$. Since the target of S is the domain of T, one can compose S and T to get a transformation $T \circ S : \mathbb{F}^p \to \mathbb{F}^m$ which is defined by

$$T \circ S(\mathbf{x}) = T\big(S(\mathbf{x})\big).$$

The following Proposition describes the composition.

Proposition 7.6. *Suppose $S : \mathbb{F}^p \to \mathbb{F}^n$ and $T : \mathbb{F}^n \to \mathbb{F}^m$ are linear transformations with matrices $A = M_S$ and $B = M_B$ respectively. Then the composition $T \circ S : \mathbb{F}^p \to \mathbb{F}^m$ is also linear, and the matrix of $T \circ S$ is BA. In other words,*

$$T \circ S = T_B \circ T_A = T_{BA}.$$

Furthermore, letting M_T denote the matrix of T,

$$M_{T \circ S} = M_T M_S.$$

Proof. To prove $T \circ S$ is linear, note that

$$\begin{aligned} T \circ S(r\mathbf{x} + s\mathbf{y}) &= T\big(S(r\mathbf{x} + s\mathbf{y})\big) \\ &= T\big(rS(\mathbf{x}) + sS(\mathbf{y})\big) \\ &= rT\big(S(\mathbf{x})\big) + sT\big(S(\mathbf{y})\big). \end{aligned}$$

In other words, $T \circ S(r\mathbf{x} + s\mathbf{y}) = rT \circ S(\mathbf{x}) + sT \circ S(\mathbf{y})$, so $T \circ S$ is linear as claimed. To find the matrix of $T \circ S$, we observe that

$$T \circ S(\mathbf{x}) = T(A\mathbf{x}) = B(A\mathbf{x}) = (BA)\mathbf{x}.$$

This implies that the matrix of $T \circ S$ is the product BA as asserted. The rest of the proof now follows easily. □

Note that the key fact in this proof is that matrix multiplication is associative. In fact, the main observation is that $T \circ S(\mathbf{x}) = T(A\mathbf{x}) = B(A\mathbf{x})$. Given this, it is immediate that $T \circ S$ is linear, so the first step in the proof was actually unnecessary.

7.2.3 An Example: Rotations of $\mathbb{R}^2$

A nice way of illustrating the previous discussion is by considering rotations of the plane. Let $\mathcal{R}_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ stand for the counter-clockwise rotation of $\mathbb{R}^2$ through θ. Computing the images of $\mathcal{R}(\mathbf{e}_1)$ and $\mathcal{R}(\mathbf{e}_2)$, we have

$$\mathcal{R}_\theta(\mathbf{e}_1) = \cos\theta\mathbf{e}_1 + \sin\theta\mathbf{e}_2,$$

and

$$\mathcal{R}_\theta(\mathbf{e}_2) = -\sin\theta\mathbf{e}_1 + \cos\theta\mathbf{e}_2.$$

FIGURE

I claim that rotations are linear. This can be seen as follows. Suppose $\mathbf{x}$ and $\mathbf{y}$ are any two non collinear vectors in $\mathbb{R}^2$, and let P be the parallelogram they span. Then $\mathcal{R}_\theta$ rotates the whole parallelogram P about $\mathbf{0}$ to a new parallelogram $\mathcal{R}_\theta(P)$. The edges of $\mathcal{R}_\theta(P)$ at $\mathbf{0}$ are $\mathcal{R}_\theta(\mathbf{x})$ and $\mathcal{R}_\theta(\mathbf{y})$. Hence, the diagonal $\mathbf{x} + \mathbf{y}$ of P is rotated to the diagonal of $\mathcal{R}_\theta(P)$. Thus

$$\mathcal{R}_\theta(\mathbf{x} + \mathbf{y}) = \mathcal{R}_\theta(\mathbf{x}) + \mathcal{R}_\theta(\mathbf{y}).$$

Similarly, for any scalar r,

$$\mathcal{R}_\theta(r\mathbf{x}) = r\mathcal{R}_\theta(\mathbf{x}).$$

Therefore $\mathcal{R}_\theta$ is linear, as claimed. Putting our rotation into the form of a matrix transformation gives

$$\begin{aligned} \mathcal{R}_\theta \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \end{pmatrix} \\ &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \end{aligned}$$

Thus the matrix of $\mathcal{R}_\theta$ is Let's now illustrate a consequence of Proposition 7.6. If one first applies the rotation $\mathcal{R}_\psi$ and follows that by the rotation $\mathcal{R}_\theta$, the outcome is the rotation $\mathcal{R}_{\theta+\psi}$ through $\theta + \psi$ (why?). In other words,

$$\mathcal{R}_{\theta+\psi} = \mathcal{R}_\theta \circ \mathcal{R}_\psi.$$

Therefore, by Proposition 7.6, we see that

$$\begin{pmatrix} \cos(\theta+\psi) & -\sin(\theta+\psi) \\ \sin(\theta+\psi) & \cos(\theta+\psi) \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{pmatrix}.$$

Expanding the product gives the angle sum formulas for $\cos(\theta + \psi)$ and $\sin(\theta + \psi)$. Namely,

$$\cos(\theta+\psi) = \cos\theta\cos\psi - \sin\theta\sin\psi,$$

and

$$\sin(\theta+\psi) = \sin\theta\cos\psi + \cos\theta\sin\psi.$$

Thus the angle sum formulas for cosine and sine can be seen via matrix algebra and the fact that rotations are linear.

Exercises

Exercise 7.7. Find the matrix of the following transformations:

(i) $F(x_1, x_2, x_3) = (2x_1 - 3x_3, x_1 + x_2 - x_3, x_1, x_2 - x_3)^T$.

(ii) $G(x_1, x_2, x_3, x_4) = (x_1 - x_2 + x_3 + x_4, x_2 + 2x_3 - 3x_4)^T$.

(iii) The matrix of $G \circ F$.

Exercise 7.8. Find the matrix of

(i) The rotation $R_{-\pi/4}$ of $\mathbb{R}^2$ through $-\pi/4$.

(ii) The reflection H of $\mathbb{R}^2$ through the line $x = y$.

(iii) The matrices of $H \circ R_{-\pi/4}$ and $R_{-\pi/4} \circ H$, where H is the reflection of part (ii).

(iv) The rotation of $\mathbb{R}^3$ through $\pi/3$ about the z-axis.

Exercise 7.9. Let $V = \mathbb{C}$, and consider the transformation $R : V \to V$ defined by $R(z) = e^{i\theta}z$. Interpret R as a transformation from $\mathbb{R}^2$ to $\mathbb{R}^2$. Compare your answer with the result of Exercise 7.6.

Exercise 7.10. Suppose $T : \mathbb{F}^n \to \mathbb{F}^n$ is linear. When does the inverse transformation T^{-1} exist?

7.3 Some Geometry of Linear Transformations on $\mathbb{R}^n$

As illustrated by the last section, linear transformations $T : \mathbb{R}^n \to \mathbb{R}^n$ have a very rich geometry. In this section we will discuss some these geometric aspects.

7.3.1 Transformations on the Plane

We know that a linear transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ is determined by $T(\mathbf{e}_1)$ and $T(\mathbf{e}_2)$, and so if $T(\mathbf{e}_1)$ and $T(\mathbf{e}_2)$ are non-collinear, then T sends each one of the coordinate axes $\mathbb{R}\mathbf{e}_i$ to the line $\mathbb{R}T(\mathbf{e}_i)$. Furthermore, T transforms the square $\mathbf{S}$ spanned by $\mathbf{e}_1$ and $\mathbf{e}_2$ onto the parallelogram $\mathbf{P}$ with edges $T(\mathbf{e}_1)$ and $T(\mathbf{e}_2)$. Indeed,

$$\mathbf{P} = \{rT(\mathbf{e}_1) + sT(\mathbf{e}_2) \mid 0 \leq r, s \leq 1\},$$

and since $T(r\mathbf{e}_1 + s\mathbf{e}_2) = rT(\mathbf{e}_1) + sT(\mathbf{e}_2)$, $T(\mathbf{S}) = \mathbf{P}$. More generally, T sends the parallelogram with sides $\mathbf{x}$ and $\mathbf{y}$ to the parallelogram with sides $T(\mathbf{x})$ and $T(\mathbf{y})$. Note that we implicitly already used this fact in the last section.

We next consider a slightly different phenomenon.

Example 7.9 (Projections). Let $\mathbf{a} \in \mathbb{R}^2$ be non-zero. Recall that the transformation

$$P_{\mathbf{a}}(\mathbf{x}) = \frac{\mathbf{a} \cdot \mathbf{x}}{\mathbf{a} \cdot \mathbf{a}}\mathbf{a}$$

is called the projection on the line $\mathbb{R}\mathbf{a}$ spanned by $\mathbf{a}$. In an exercise in the first chapter, you actually showed $P_{\mathbf{a}}$ is linear. If you skipped this, it is proved as follows.

$$P_{\mathbf{a}}(\mathbf{x} + \mathbf{y}) = \frac{\mathbf{a} \cdot (\mathbf{x} + \mathbf{y})}{\mathbf{a} \cdot \mathbf{a}}\mathbf{a} = \big(\frac{\mathbf{a} \cdot \mathbf{x} + \mathbf{a} \cdot \mathbf{y}}{\mathbf{a} \cdot \mathbf{a}}\big)\mathbf{a} = P_{\mathbf{a}}(\mathbf{x}) + P_{\mathbf{a}}(\mathbf{y}).$$

In addition, for any scalar r,

$$P_{\mathbf{a}}(r\mathbf{x}) = \frac{\mathbf{a} \cdot (r\mathbf{x})}{\mathbf{a} \cdot \mathbf{a}}\mathbf{a} = r\big(\frac{\mathbf{a} \cdot \mathbf{x}}{\mathbf{a} \cdot \mathbf{a}}\big)\mathbf{a} = rP_{\mathbf{a}}(\mathbf{x}).$$

This verifies the linearity of any projection. Using the formula, we get the explicit expression

$$P_{\mathbf{a}}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} (\frac{a_1x_1 + a_2x_2}{a_1^2 + a_2^2})a_1 \\ (\frac{a_1x_1 + a_2x_2}{a_1^2 + a_2^2})a_2 \end{pmatrix},$$

where $\mathbf{a} = (a_1, a_2)^T$ and $\mathbf{x} = (x_1, x_2)^T$.

Hence

$$P_{\mathbf{a}} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{a_1^2 + a_2^2} \begin{pmatrix} a_1^2 & a_1 a_2 \\ a_1 a_2 & a_2^2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Thus the matrix of $P_{\mathbf{a}}$ is

$$\frac{1}{a_1^2 + a_2^2} \begin{pmatrix} a_1^2 & a_1 a_2 \\ a_1 a_2 & a_2^2 \end{pmatrix}.$$

Of course, projections don't send parallelograms to parallelograms, since any two values $P_{\mathbf{b}}(\mathbf{x})$ and $P_{\mathbf{b}}(\mathbf{y})$ are collinear. Nevertheless, projections have another interesting geometric property. Namely, each vector on the line spanned by $\mathbf{b}$ is preserved by $P_{\mathbf{b}}$, and every vector orthogonal to $\mathbf{b}$ is mapped by $P_{\mathbf{b}}$ to $\mathbf{0}$.

7.3.2 Orthogonal Transformations

Orthogonal transformations are the linear transformations associated with orthogonal matrices (see §3.5.3). They are closely related with Euclidean geometry. Orthogonal transformations are characterized by the property that they preserve angles and lengths. Rotations are specific examples. Reflections are another class of examples. Your reflection is the image you see when you look in a mirror. The reflection is through the plane of the mirror.

Let us analyze reflections carefully, starting with the case of the plane $\mathbb{R}^2$. Consider a line ℓ in $\mathbb{R}^2$ through the origin. The reflection of $\mathbb{R}^2$ through ℓ acts as follows: every point on ℓ is left fixed, and the points on the line $\ell^\perp$ through the origin orthogonal to ℓ are sent to their negatives.

FIGURE FOR REFLECTIONS

Perhaps somewhat surprisingly, reflections are linear. We will show this by deriving a formula. Let $\mathbf{b}$ be any non-zero vector on $\ell^\perp$, and let $H_{\mathbf{b}}$ denote the reflection through ℓ. Choose an arbitrary $\mathbf{v} \in \mathbb{R}^2$, and consider its orthogonal decomposition (see Chapter 1)

$$\mathbf{v} = P_{\mathbf{b}}(\mathbf{v}) + \mathbf{c}$$

with $\mathbf{c}$ on ℓ. By the parallelogram law,

$$H_{\mathbf{b}}(\mathbf{v}) = \mathbf{c} - P_{\mathbf{b}}(\mathbf{v}).$$

Replacing $\mathbf{c}$ by $\mathbf{v} - P_{\mathbf{b}}(\mathbf{v})$ gives the formula

$$\begin{aligned} H_{\mathbf{b}}(\mathbf{v}) &= \mathbf{v} - 2P_{\mathbf{b}}(\mathbf{v}) \\ &= \mathbf{v} - 2\big(\frac{\mathbf{v}\cdot\mathbf{b}}{\mathbf{b}\cdot\mathbf{b}}\big)\mathbf{b}. \end{aligned}$$

Expressing this in terms of the unit vector $\widehat{\mathbf{b}}$ determined by $\widehat{\mathbf{b}}$ gives us the simpler expression

$$H_{\mathbf{b}}(\mathbf{v}) = \mathbf{v} - 2(\mathbf{v}\cdot\widehat{\mathbf{b}})\widehat{\mathbf{b}}. \tag{7.1}$$

Certainly $H_{\mathbf{b}}$ has the properties we sought: $H_{\mathbf{b}}(\mathbf{v}) = \mathbf{v}$ if $\mathbf{b}\cdot\mathbf{v} = 0$, and $H_{\mathbf{b}}(\mathbf{b}) = -\mathbf{b}$. Moreover, $H_{\mathbf{b}}$ can be expressed as $I_2 - 2P_{\mathbf{b}}$, so it is linear since any linear combination of linear transformations is linear.

The above expression of a reflection goes through not just for $\mathbb{R}^2$, but for $\mathbb{R}^n$ for any $n \geq 3$ as well. Let $\mathbf{b}$ be any nonzero vector in $\mathbb{R}^n$, and let W be the hyperplane in $\mathbb{R}^n$ consisting of all vectors orthogonal to $\mathbf{b}$. Then the transformation $H : \mathbb{R}^n \to \mathbb{R}^n$ defined by (7.1) is the reflection of $\mathbb{R}^n$ through W.

Example 7.10. Let $\mathbf{b} = (1,1)^T$, so $\widehat{\mathbf{b}} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T$. Then $H_{\mathbf{b}}$ is the reflection through the line $x = -y$. We have

$$\begin{aligned} H_{\mathbf{b}}\begin{pmatrix} a \\ b \end{pmatrix} &= \begin{pmatrix} a \\ b \end{pmatrix} - 2\Big(\begin{pmatrix} a \\ b \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}\Big) \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \\ &= \begin{pmatrix} a - (a+b) \\ b - (a+b) \end{pmatrix} \\ &= \begin{pmatrix} -b \\ -a \end{pmatrix}. \end{aligned}$$

There are several worthwhile consequences of formula (7.1). All reflections are linear, and reflecting $\mathbf{v}$ twice returns $\mathbf{v}$ to itself, i.e. $H_{\mathbf{b}} \circ H_{\mathbf{b}} = I_2$. Furthermore, reflections preserve inner products. That is, for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^2$,

$$H_{\mathbf{b}}(\mathbf{v}) \cdot H_{\mathbf{b}}(\mathbf{w}) = \mathbf{v}\cdot\mathbf{w}.$$

We will leave these properties as an exercise.

A consequence of the last property is that since lengths, distances and angles between vectors are expressed in terms of the dot product, reflections preserve all these quantites. In other words, a vector and its reflection have the same length, and the angle (measured with respect to the origin) between a vector and the reflecting line is the same as the angle between the reflection and the reflecting line. This motivates the following

Definition 7.2. A linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ is said to be *orthogonal* if it preserves the dot product. That is, T is orthogonal if and only if

$$T(\mathbf{v}) \cdot T(\mathbf{w}) = \mathbf{v} \cdot \mathbf{w}$$

for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$.

Proposition 7.7. *If a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ is orthogonal, then for any $\mathbf{v} \in \mathbb{R}^n$, $|T(\mathbf{v})| = |\mathbf{v}|$. In particular, if $\mathbf{v} \neq \mathbf{0}$, then $T(\mathbf{v}) \neq \mathbf{0}$. Moreover, the angle between any two nonzero vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ is the same as the angle between the vectors $T(\mathbf{v})$ and $T(\mathbf{w})$, which are both nonzero by the last assertion.*

We also have

Proposition 7.8. *A linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ is orthogonal if and only if its matrix M_T is orthogonal.*

We leave the proofs of the previous two propositions as exercises. By Proposition7.8, every rotation of $\mathbb{R}^2$ is orthogonal, since a rotation matrix is clearly orthogonal. Recall that $O(2, \mathbb{R})$ is the matrix group consisting of all 2×2 orthogonal matrices. We can now prove the following pretty fact.

Proposition 7.9. *Every orthogonal transformation of $\mathbb{R}^2$ is a reflection or a rotation. In fact, the reflections are those orthogonal transformations T for which M_T is symmetric but $M_T \neq I_2$. The rotations $\mathcal{R}_\theta$ are those such that $M_{\mathcal{R}} = I_2$ or $M_{\mathcal{R}}$ is not symmetric.*

Proof. It is not hard to check that any 2×2 orthogonal matrix has the form

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

or

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

The former are rotations (including I_2) and the latter are symmetric, but do not include I_2. The transformations H_θ are in fact reflections. We leave it as an exercise to check that H_θ is the reflection through the line spanned by $(\cos(\theta/2), \sin(\theta/2))^T$. In Chapter 8, we will give a simple geometric proof using eigentheory that H_θ is a reflection. □

The structure of orthogonal transformations in higher dimensions is more complicated. For example, the rotations and reflections of $\mathbb{R}^3$ do not give all the possible orthogonal linear transformations of $\mathbb{R}^3$.

7.3.3 Gradients and differentials

Since arbitrary transformations can be very complicated, we should view linear transformations as one of the simplest are types of transformations. In fact, we can make a much more precise statement about this. One of the most useful principals about smooth transformations is that no matter how complicated such transformations are, they admit linear approximations, which means that one certain information may be obtained by constructing a taking partial derivatives. Suppose we consider a transformation $F : \mathbb{R}^n \to \mathbb{R}^m$ such that each component function f_i of F has continuous first partial derivatives throughout $\mathbb{R}^n$, that is F is smooth. Then it turns out that in a sense which can be made precise, the differentials of the components f_i are the best linear approximations to the f_i. Recall that if $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function, the differential $df(\mathbf{x})$ of f at $\mathbf{x}$ is the linear function $df(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ whose value at $\mathbf{v} \in \mathbb{R}^n$ is

$$df(\mathbf{x})\mathbf{v} = \nabla f(\mathbf{x}) \cdot \mathbf{v}.$$

Here,

$$\nabla f(\mathbf{x}) = (\frac{\partial f}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{x})) \in \mathbb{R}^n$$

is the called the gradient of f at $\mathbf{x}$. In other words,

$$df(\mathbf{x})\mathbf{v} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{x})\mathbf{v}_i,$$

so the differential is the linear transformation induced by the gradient and the dot product. Note that in the above formula, $\mathbf{x}$ is not a variable. It represents the point at which the differential of f is being computed.

The differential of the transformation F at $\mathbf{x}$ is the linear function $DF(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^n$ defined by $DF(\mathbf{x}) = (df_1(\mathbf{x}), df_2(\mathbf{x}), \ldots, df_m(\mathbf{x}))$. The components of DF at $\mathbf{x}$ are the differentials of the components of F at $\mathbf{x}$.

We will have to leave further discussion of the differential for a course in vector analysis.

Exercises

Exercise 7.11. Verify from the formula that the projection $P_{\mathbf{b}}$ fixes every vector on the line spanned by $\mathbf{b}$ and sends every vector orthogonal to $\mathbf{b}$ to $\mathbf{0}$.

Exercise 7.12. Let $H_{\mathbf{b}} : \mathbb{R}^2 \to \mathbb{R}^2$ be the reflection of $\mathbb{R}^2$ through the line orthogonal to $\mathbf{b}$. Recall that $H_{\mathbf{b}}(\mathbf{v}) = \mathbf{v} - 2\big(\frac{\mathbf{v}\cdot\mathbf{b}}{\mathbf{b}\cdot\mathbf{b}}\big)\mathbf{b}$.

(i) Use this formula to show that every reflection is linear.

(ii) Show also that $H_{\mathbf{b}}(H_{\mathbf{b}}(\mathbf{x})) = \mathbf{x}$.

(iii) Find formulas for $H_{\mathbf{b}}((1,0))$ and $H_{\mathbf{b}}((0,1))$.

Exercise 7.13. Consider the transformation $C_{\mathbf{a}} : \mathbb{R}^3 \to \mathbb{R}^3$ defined by

$$C_{\mathbf{a}}(\mathbf{v}) = \mathbf{a} \times \mathbf{v}.$$

(i) Show that $C_{\mathbf{a}}$ is linear.

(ii) Descibe the set of vectors $\mathbf{x}$ such that $C_{\mathbf{a}}(\mathbf{x}) = \mathbf{0}$.

Exercise 7.14. Let $\mathbf{u}$ and $\mathbf{v}$ be two orthogonal unit length vectors in $\mathbb{R}^2$. Show that the following formulas hold for all $\mathbf{x} \in \mathbb{R}^2$:

(a) $P_{\mathbf{u}}(\mathbf{x}) + P_{\mathbf{v}}(\mathbf{x}) = \mathbf{x}$, and

(b) $P_{\mathbf{u}}(P_{\mathbf{v}}(\mathbf{x})) = P_{\mathbf{v}}(P_{\mathbf{u}}(\mathbf{x})) = \mathbf{0}$.

Conclude from (a) that $\mathbf{x} = (\mathbf{x}\cdot\mathbf{u})\mathbf{u} + (\mathbf{x}\cdot\mathbf{v})\mathbf{v}$.

Exercise 7.15. Suppose $T : \mathbb{R}^2 \to \mathbb{R}^2$ is a linear transformation which sends any two non collinear vectors to non collinear vectors. Suppose $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^2$ are non collinear. Show that T sends any parallelogram with sides parallel to $\mathbf{x}$ and $\mathbf{y}$ to another parallelogram with sides parallel to $T(\mathbf{x})$ and $T(\mathbf{y})$.

Exercise 7.16. Show that all reflections are orthogonal linear transformations. In other words, show that for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$,

$$H_{\mathbf{b}}(\mathbf{x}) \cdot H_{\mathbf{b}}(\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}.$$

Exercise 7.17. Show that rotations $\mathcal{R}_\theta$ of $\mathbb{R}^2$ also give orthogonal linear transformations.

Exercise 7.18. Show that every orthogonal linear transformation not only preserves dot products, but also lengths of vectors and angles and distances between two distinct vectors. Do reflections and rotations preserve lengths and angles and distances?

Exercise 7.19. Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is a transformation with the property that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$F(\mathbf{x}) \cdot F(\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}.$$

(a) Show that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $||F(\mathbf{x}+\mathbf{y}) - F(\mathbf{x}) - F(\mathbf{y})||^2 = 0$.

(b) Show similarly that for all $\mathbf{x} \in \mathbb{R}^n$ and $r \in \mathbb{R}$, $||F(r\mathbf{x}) - rF(\mathbf{x})||^2 = 0$ Conclude that F is in fact linear. Hence F is an orthogonal linear transformation.

Exercise 7.20. Find the reflection of $\mathbb{R}^3$ through the plane P if:

(a) P is the plane $x + y + z = 0$; and

(b) P is the plane $ax + by + cz = 0$.

Exercise 7.21. Which of the following statements are true? Explain.

(i) The composition of two rotations is a rotation.

(ii) The composition of two reflections is a reflection.

(iii) The composition of a reflection and a rotation is a rotation.

Exercise 7.22. Find a formula for the composition of two rotations. That is, compute $\mathcal{R}_\theta \circ \mathcal{R}_\mu$ in terms of sines and cosines. Give an interpretation of the result.

Exercise 7.23. * Let $f(x_1, x_2) = x_1^2 + 2x_2^2$.

(a) Find both the gradient and differential of f at $(1, 2)$.

(b) If $\mathbf{u} \in \mathbb{R}^2$ is a unit vector, then $\mathrm{d}f(1,2)\mathbf{u}$ is called the directional derivative of f at $(1, 2)$ in the direction $\mathbf{u}$. Find the direction $\mathbf{u} \in \mathbb{R}^2$ which maximizes the value of $\mathrm{d}f(1,2)\mathbf{u}$.

(c) What has your answer in part (b) got to do with the length of the gradient of f at $(1, 2)$?

Exercise 7.24. Let $V = \mathbb{C}$ and consider the transformation $H : V \to V$ defined by $H(z) = \overline{z}$. Interpret H as a transformation from $\mathbb{R}^2$ to $\mathbb{R}^2$.

Exercise 7.25. *. Find the differential at any (x_1, x_2) of the polar coordinate map of Example 3.7.

7.4 Matrices With Respect to an Arbitrary Basis

Let V and W be finite dimensional vector spaces over $\mathbb{F}$, and suppose $T : V \to W$ is linear. The purpose of this section is to define the matrix of T with respect to arbitrary bases of the domain V and the target W.

7.4.1 Coordinates With Respect to a Basis

We will first define the coordinates of a vector with respect to an arbitrary basis. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be a basis of V, and let $\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be a basis of V. Then every $\mathbf{w} \in V$ has a unique expression

$$\mathbf{w} = r_1\mathbf{v}_1 + r_2\mathbf{v}_2 + \cdots + r_n\mathbf{v}_n,$$

so we will make the following definition.

Definition 7.3. We will call $r_1, r_2, \ldots, r_n$ the *coordinates* of $\mathbf{w}$ with respect to $\mathcal{B}$, and we will write $\mathbf{w} =< r_1, r_2, \ldots, r_n >$. If there is a possibility of confusion, will write the coordinates as $< r_1, r_2, \ldots, r_n >_{\mathcal{B}}$.

Notice that the notion of coordinates assumes that the basis is ordered. Finding the coordinates of a vector with respect to a given basis of is a familiar problem.

Example 7.11. Suppose $\mathbb{F} = \mathbb{R}$, and consider two bases of $\mathbb{R}^2$, say

$$\mathcal{B} = \{(1,2)^T, (0,1)^T\} \quad \text{and} \quad \mathcal{B}' = \{(1,1)^T, (1,-1)^T\}.$$

Expanding $\mathbf{e}_1 = (1,0)^T$ in terms of these two bases gives two different sets of coordinates for $\mathbf{e}_1$. By inspection,

$$\begin{pmatrix}1\\0\end{pmatrix} = 1\begin{pmatrix}1\\2\end{pmatrix} - 2\begin{pmatrix}0\\1\end{pmatrix},$$

and

$$\begin{pmatrix}1\\0\end{pmatrix} = \frac{1}{2}\begin{pmatrix}1\\1\end{pmatrix} + \frac{1}{2}\begin{pmatrix}1\\-1\end{pmatrix}.$$

Thus the coordinates of $\mathbf{e}_1$ with respect to $\mathcal{B}$ are $< 1, -2 >$, and with respect to $\mathcal{B}'$ they are $< \frac{1}{2}, \frac{1}{2} >$.

Now consider how two different sets of coordinates for the same vector are related. In fact, we can set up a system to decide this. For example, using the bases of $\mathbb{R}^2$ in the above example, we expand the second basis $\mathcal{B}'$ in terms of the first $\mathcal{B}$. That is, write

$$\begin{pmatrix}1\\1\end{pmatrix} = a\begin{pmatrix}1\\2\end{pmatrix} + b\begin{pmatrix}0\\1\end{pmatrix},$$

and

$$\begin{pmatrix}1\\-1\end{pmatrix} = c\begin{pmatrix}1\\2\end{pmatrix} + d\begin{pmatrix}0\\1\end{pmatrix}.$$

These equations are expressed in matrix form as:

$$\begin{pmatrix}1 & 1\\1 & -1\end{pmatrix} = \begin{pmatrix}1 & 0\\2 & 1\end{pmatrix}\begin{pmatrix}a & c\\b & d\end{pmatrix}.$$

Now suppose $\mathbf{p}$ has coordinates $< r, s >$ in terms of the first basis and coordinates $< x, y >'$ in terms of the second. Then

$$\mathbf{p} = \begin{pmatrix}1 & 0\\2 & 1\end{pmatrix}\begin{pmatrix}r\\s\end{pmatrix} = \begin{pmatrix}1 & 1\\1 & -1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}.$$

Hence

$$\begin{pmatrix}r\\s\end{pmatrix} = \begin{pmatrix}1 & 0\\2 & 1\end{pmatrix}^{-1}\begin{pmatrix}1 & 1\\1 & -1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}.$$

Therefore,

$$\begin{pmatrix}r\\s\end{pmatrix} = \begin{pmatrix}1 & 1\\-1 & -3\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}.$$

We can imitate this in the general case. Let

$$\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\},$$

and

$$\mathcal{B}' = \{\mathbf{v}'_1, \mathbf{v}'_2, \ldots, \mathbf{v}'_n\}$$

be two bases of V. Define the *change of basis matrix* $\mathcal{M}^{\mathcal{B}}_{\mathcal{B}'} \in \mathbb{F}^{n\times n}$ to be the matrix (a_{ij}) with entries determined by

$$\mathbf{v}'_j = \sum_{i=1}^{n} a_{ij}\mathbf{v}_i.$$

To see how this works, consider the case $n = 2$. We have

$$\mathbf{v}'_1 = a_{11}\mathbf{v}_1 + a_{21}\mathbf{v}_2$$

$$\mathbf{v}'_2 = a_{12}\mathbf{v}_1 + a_{22}\mathbf{v}_2.$$

It's convenient to write this in matrix form

$$(\mathbf{v}_1' \ \mathbf{v}_2') = (\mathbf{v}_1 \ \mathbf{v}_2)\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = (\mathbf{v}_1 \ \mathbf{v}_2)\mathcal{M}_{\mathcal{B}'}^{\mathcal{B}},$$

where

$$\mathcal{M}_{\mathcal{B}'}^{\mathcal{B}} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

Notice that $(\mathbf{v}_1 \ \mathbf{v}_2)$ is a generalized matrix in the sense that it is a 1×2 matrix with vector entries. A nice property of this notation is that if $(\mathbf{v}_1 \ \mathbf{v}_2)A = (\mathbf{v}_1 \ \mathbf{v}_2)B$, then $A = B$. This is due to the fact that expressions in terms of bases are unique and holds for any $n > 2$ also.

In general, we can express this suggestively as

$$\mathcal{B}' = \mathcal{B}\mathcal{M}_{\mathcal{B}'}^{\mathcal{B}}. \tag{7.2}$$

Also note that

$$\mathcal{M}_{\mathcal{B}}^{\mathcal{B}} = I_n.$$

In the above example,

$$\mathcal{M}_{\mathcal{B}'}^{\mathcal{B}} = \begin{pmatrix} 1 & 1 \\ -1 & -3 \end{pmatrix}.$$

Proposition 7.10. *Let $\mathcal{B}$ and $\mathcal{B}'$ be bases of V. Then*

$$(\mathcal{M}_{\mathcal{B}'}^{\mathcal{B}})^{-1} = \mathcal{M}_{\mathcal{B}}^{\mathcal{B}'}.$$

Proof. We have

$$(\mathbf{v}_1 \ \mathbf{v}_2) = (\mathbf{v}_1' \ \mathbf{v}_2')\mathcal{M}_{\mathcal{B}}^{\mathcal{B}'} = (\mathbf{v}_1 \ \mathbf{v}_2)\mathcal{M}_{\mathcal{B}'}^{\mathcal{B}}\mathcal{M}_{\mathcal{B}}^{\mathcal{B}'}.$$

Thus, since $\mathcal{B}$ is a basis,

$$\mathcal{M}_{\mathcal{B}'}^{\mathcal{B}}\mathcal{M}_{\mathcal{B}}^{\mathcal{B}'} = I_2.$$

□

Now what happens if a third basis $\mathcal{B}'' = \{\mathbf{v}_1'', \mathbf{v}_2''\}$ is thrown in? If we iterate the expression in (7.3), we get

$$(\mathbf{v}_1'' \ \mathbf{v}_2'') = (\mathbf{v}_1' \ \mathbf{v}_2')\mathcal{M}_{\mathcal{B}''}^{\mathcal{B}'} = (\mathbf{v}_1 \ \mathbf{v}_2)\mathcal{M}_{\mathcal{B}'}^{\mathcal{B}}\mathcal{M}_{\mathcal{B}''}^{\mathcal{B}'}.$$

Thus

$$\mathcal{M}_{\mathcal{B}''}^{\mathcal{B}} = \mathcal{M}_{\mathcal{B}'}^{\mathcal{B}}\mathcal{M}_{\mathcal{B}''}^{\mathcal{B}'}.$$

This generalizes immediately to the n-dimensional case, so we have

Proposition 7.11. *Let $\mathcal{B}, \mathcal{B}'$ and $\mathcal{B}''$ be bases of V. Then*

$$\mathcal{M}_{\mathcal{B}''}^{\mathcal{B}} = \mathcal{M}_{\mathcal{B}'}^{\mathcal{B}}\mathcal{M}_{\mathcal{B}''}^{\mathcal{B}'}.$$

7.4.2 Change of Basis for Linear Transformations

As above, let V and W be finite dimensional vector spaces over $\mathbb{F}$, and suppose $T : V \to W$ is linear. The purpose of this section is to define the matrix of T with respect to arbitrary bases of V and W. Fix a basis

$$\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$$

of V and a basis

$$\mathcal{B}' = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$$

of W. Suppose

$$T(\mathbf{v}_j) = \sum_{i=1}^{m} c_{ij}\mathbf{w}_i.$$

Definition 7.4. The *matrix of T with respect to the bases $\mathcal{B}$ and $\mathcal{B}'$* is defined to be the $m \times n$ matrix $\mathcal{M}^{\mathcal{B}}_{\mathcal{B}'}(T) = (c_{ij})$.

This notation is set up so that if $V = \mathbb{F}^n$ and $W = \mathbb{F}^m$ and $T = T_A$ for an $m \times n$ matrix A, we have $\mathcal{M}^{\mathcal{B}}_{\mathcal{B}'}(T) = A$ when $\mathcal{B}$ and $\mathcal{B}'$ are the standard bases since $T_A(\mathbf{e}_j)$ is the jth column of A. We remark that

$$\mathcal{M}^{\mathcal{B}}_{\mathcal{B}'}(Id) = \mathcal{M}^{\mathcal{B}}_{\mathcal{B}'}$$

where $Id : V \to V$ is the identity.

Now suppose $V = W$. In this case, we want to express the matrix of T in a single basis and then find its expression in another basis. So let $\mathcal{B}$ and $\mathcal{B}'$ be bases of V. As above, for simplicity, we assume $n = 2$ and $\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2\}$ and $\mathcal{B}' = \{\mathbf{v}'_v, \mathbf{v}'_2\}$. Hence $(\mathbf{v}'_1\ \mathbf{v}'_2) = (\mathbf{v}_1\ \mathbf{v}_2)\mathcal{M}^{\mathcal{B}}_{\mathcal{B}'}$. Applying T, we obtain

$$\begin{aligned}(T(\mathbf{v}'_1)\ T(\mathbf{v}'_2)) &= (T(\mathbf{v}_1)\ T(\mathbf{v}_2))\mathcal{M}^{\mathcal{B}}_{\mathcal{B}'} \\ &= (\mathbf{v}_1\ \mathbf{v}_2)\mathcal{M}^{\mathcal{B}}_{\mathcal{B}}(T)\mathcal{M}^{\mathcal{B}}_{\mathcal{B}'} \\ &= (\mathbf{v}'_1\ \mathbf{v}'_2)\mathcal{M}^{\mathcal{B}'}_{\mathcal{B}}\mathcal{M}^{\mathcal{B}}_{\mathcal{B}}(T)\mathcal{M}^{\mathcal{B}}_{\mathcal{B}'}.\end{aligned}$$

Hence,

$$\mathcal{M}^{\mathcal{B}'}_{\mathcal{B}'}(T) = \mathcal{M}^{\mathcal{B}'}_{\mathcal{B}}\mathcal{M}^{\mathcal{B}}_{\mathcal{B}}(T)\mathcal{M}^{\mathcal{B}}_{\mathcal{B}'}.$$

Putting $P = \mathcal{M}^{\mathcal{B}'}_{\mathcal{B}}$, we therefore see that

$$\mathcal{M}^{\mathcal{B}'}_{\mathcal{B}'}(T) = P\mathcal{M}^{\mathcal{B}}_{\mathcal{B}}(T)P^{-1}.$$

We have therefore shown

Proposition 7.12. *Let $T : V \to V$ be linear and let $\mathcal{B}$ and $\mathcal{B}'$ be bases of V. Then*

$$\mathcal{M}_{\mathcal{B}'}^{\mathcal{B}'}(T) = \mathcal{M}_{\mathcal{B}}^{\mathcal{B}'}\mathcal{M}_{\mathcal{B}}^{\mathcal{B}}(T)\mathcal{M}_{\mathcal{B}'}^{\mathcal{B}}. \tag{7.3}$$

Thus, if $P = \mathcal{M}_{\mathcal{B}}^{\mathcal{B}'}$, we have

$$\mathcal{M}_{\mathcal{B}'}^{\mathcal{B}'}(T) = P\mathcal{M}_{\mathcal{B}}^{\mathcal{B}}(T)P^{-1}. \tag{7.4}$$

Example 7.12. Consider the linear transformation T of $\mathbb{R}^2$ whose matrix with respect to the standard basis is

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

Let's find the matrix B of T with respect to the basis $(1,1)^T$ and $(1,-1)^T$. Calling this basis $\mathcal{B}'$ and the standard basis $\mathcal{B}$, formula (7.3) says

$$B = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -4 & 3 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{-1}.$$

Computing the product gives

$$B = \begin{pmatrix} 0 & -3 \\ 1 & -1 \end{pmatrix}.$$

Definition 7.5. Let A and B be $n \times n$ matrices over $\mathbb{F}$. Then we say A *is similar to* B if and only if there exists an invertible $P \in \mathbb{F}^{n\times n}$ such that $B = PAP^{-1}$.

It is not hard to see that similarity is an equivalence relation on $\mathbb{F}^{n\times n}$ (Exercise: check this). An equivalence class for this equivalence relation is called a *conjugacy class*. Hence,

Proposition 7.13. *The matrices which represent a given linear transformation T form a conjugacy class in $\mathbb{F}^{n\times n}$.*

Example 7.13. Let $\mathbb{F} = \mathbb{R}$ and suppose $\mathbf{v}_1$ and $\mathbf{v}_2$ denote $(1,2)^T$ and $(0,1)^T$ respectively. Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be the linear transformation such that $T(\mathbf{v}_1) = \mathbf{v}_1$ and $T(\mathbf{v}_2) = 3\mathbf{v}_2$. By Proposition 7.14, T exists and is unique. Now the matrix of T with respect to the basis $\mathbf{v}_1, \mathbf{v}_2$ is

$$\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}.$$

Thus T has a diagonal matrix in the $\mathbf{v}_1, \mathbf{v}_2$ basis.

Exercises

Exercise 7.26. Find the coordinates of $\mathbf{e}_1,\ \mathbf{e}_2,\ \mathbf{e}_3$ of $\mathbb{R}^3$ in terms of the basis $(1,1,1)^T,\ (1,0,1)^T,\ (0,1,1)^T$. Then find the matrix of the linear transformation $T(x_1,x_2,x_3) = (4x_1 + x_2 - x_3, x_1 + 3x_3, x_2 + 2x_3)^T$ with respect to this basis.

Exercise 7.27. Consider the basis $(1,1,1)^T,\ (1,0,1)^T$, and $(0,1,1)^T$ of $\mathbb{R}^3$. Find the matrix of the linear transformation $T : \mathbb{R}^3 \to \mathbb{R}^3$ defined by $T(\mathbf{x}) = (1,1,1)^T \times \mathbf{x}$ with respect to this basis.

Exercise 7.28. Let $H : \mathbb{R}^2 \to \mathbb{R}^2$ be the reflection through the line $x = y$. Find a basis of $\mathbb{R}^2$ such that the matrix of H is diagonal.

Exercise 7.29. Show that any projection $P_{\mathbf{a}} : \mathbb{R}^2 \to \mathbb{R}^2$ is diagonalizable. That is, there exists a basis for which the matrix of $P_{\mathbf{a}}$ is diagonal.

Exercise 7.30. Let R_θ be any rotation of $\mathbb{R}^2$. Does there exist a basis of $\mathbb{R}^2$ for which the matrix of R_θ is diagonal. That is, is there an invertible 2×2 matrix P such that $R_\theta = PDP^{-1}$.

Exercise 7.31. A rotation $\mathcal{R}_\theta$ defines a linear map from $\mathbb{R}^2$ to itself. Show that $\mathcal{R}_\theta$ also defines a $\mathbb{C}$-linear map $\mathcal{R}_\theta : \mathbb{C} \to \mathbb{C}$. Describe this map in terms of the complex exponential.

Exercise 7.32. Show that similarity is an equivalence relation an $\mathbb{F}^{n\times n}$.

7.5 Further Results on Linear Transformations

The purpose of this chapter is to develop some more of the tools necessary to get a better understanding of linear transformations.

7.5.1 An Existence Theorem

To begin, we will prove an extremely fundamental, but very simple, existence theorem about linear transformations. In essence, this result tells us that given a basis of a finite dimensional vector space V over $\mathbb{F}$ and any other vector space W over $\mathbb{F}$, there exists a unique linear transformation $T : V \to W$ taking whatever values we wish on the given basis. We will then derive a few interesting consequences of this fact.

Proposition 7.14. *Let V and W be any finite dimensional vector space over $\mathbb{F}$. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be any basis of V, and let $\mathbf{w}_1, \ldots, \mathbf{w}_n$ be arbitrary vectors in W. Then there exists a unique linear transformation $T : V \to W$ such that $T(\mathbf{v}_i) = \mathbf{w}_i$ for each i. In other words a linear transformation is uniquely determined by giving its values on a basis.*

Proof. The proof is surprisingly simple. Since every $\mathbf{v} \in V$ has a unique expression

$$\mathbf{v} = \sum_{i=1}^{n} r_i \mathbf{v}_i,$$

where $r_1, \ldots r_n \in \mathbb{F}$, we can define

$$T(\mathbf{v}) = \sum_{i=1}^{n} r_i T(\mathbf{v}_i).$$

This certainly defines a transformation, and we can easily show that T is linear. Indeed, if $\mathbf{v} = \sum \alpha_i \mathbf{v}_i$ and $\mathbf{w} = \sum \beta_i \mathbf{v}_i$, then $\mathbf{v} + \mathbf{w} = \sum(\alpha_i + \beta_i)\mathbf{v}_i$, so

$$T(\mathbf{v} + \mathbf{w}) = \sum(\alpha_i + \beta_i)T(\mathbf{v}_i) = T(\mathbf{v}) + T(\mathbf{w}).$$

Similarly, $T(r\mathbf{v}) = rT(\mathbf{v})$. Moreover, T is unique, since a linear transformationis determined on a basis. □

If $V = \mathbb{F}^n$ and $W = \mathbb{F}^m$, there is an even simpler proof by appealing to matrix theory. Let $B = (\mathbf{v}_1\ \mathbf{v}_2\ \ldots \mathbf{v}_n)$ and $C = (\mathbf{w}_1\ \mathbf{w}_2\ \ldots \mathbf{w}_n)$. Then the matrix A of T satisfies $AB = C$. But B is invertible since $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a basis of $\mathbb{F}^n$, so $A = CB^{-1}$.

7.5.2 The Kernel and Image of a Linear Transformation

Let $T : V \to W$ be a linear transformation.

Definition 7.6. The *kernel* of T, is defined to be the set $\ker(T)$ consisting of all $\mathbf{v} \in V$ such that $T(\mathbf{v}) = \mathbf{0}$. The *image* of T is the set $\mathrm{Im}(T)$ consisting of all $\mathbf{w} \in W$ such that $T(\mathbf{v}) = \mathbf{w}$ for some $\mathbf{v} \in V$.

If $V = \mathbb{F}^n$, $W = \mathbb{F}^m$ and $T = T_A$, then of course, $\ker(T) = \mathcal{N}(A)$ and $\mathrm{Im}(T) = \mathrm{col}(A)$. Hence the problem of finding $\ker(T)$ is the same as finding the solution space of an $m \times n$ homogeneous linear system.

Proposition 7.15. *The kernel and image of a linear transformation $T : V \to W$ are subspaces of V and W are respectively. T is one to one if and only if* $\ker(T) = \{\mathbf{0}\}$.

Proof. The first assertion is obvious. Suppose that T is one to one. Then, since $T(\mathbf{0}) = \mathbf{0}$, $\ker(T) = \{\mathbf{0}\}$. Conversely, suppose $\ker(T) = \{\mathbf{0}\}$. If $\mathbf{x}, \mathbf{y} \in V$ are such that $T(\mathbf{x}) = T(\mathbf{y})$, then

$$T(\mathbf{x}) - T(\mathbf{y}) = T(\mathbf{x} - \mathbf{y}).$$

Therefore,

$$T(\mathbf{x} - \mathbf{y}) = \mathbf{0}.$$

Thus $\mathbf{x} - \mathbf{y} \in \ker(T)$, so $\mathbf{x} - \mathbf{y} = \{\mathbf{0}\}$. Hence $\mathbf{x} = \mathbf{y}$, and we conclude T is one to one. □

Example 7.14. Let W be any subspace of V. Let's use Proposition 7.14 to show that there exists a linear transformation $T : V \to V$ whose kernel is W. Choose a basis $\mathbf{v}_1, \ldots, \mathbf{v}_k$ of W and extend this basis to a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of V. Define a linear transformation $T : V \to V$ by putting $T(\mathbf{v}_i) = \mathbf{0}$ if $1 \leq i \leq k$ and putting $T(\mathbf{v}_j) = \mathbf{v}_j$ if $k + 1 \leq j \leq n$. Then $\ker(T) = W$. For if $\mathbf{v} = \sum_{i=1}^{n} a_i \mathbf{v}_i \in \ker(T)$, we have

$$T(\mathbf{v}) = T(\sum_{i=1}^{n} a_i \mathbf{v}_i) = \sum_{i=1}^{n} a_i T(\mathbf{v}_i) = \sum_{j=k+1}^{n} a_j \mathbf{v}_j = \mathbf{0},$$

so $a_{k+1} = \cdots = a_n = 0$. Hence $\mathbf{v} \in W$, so $\ker(T) \subset W$. Since we designed T so that $W \subset \ker(T)$, we are through.

The main result on the kernel and image is the following.

Theorem 7.16. *Suppose $T : V \to W$ is a linear transformation where $\dim V = n$. Then*

$$\dim \ker(T) + \dim \mathrm{Im}(T) = n. \tag{7.5}$$

In fact, there exists a basis $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ of V so that

(i) $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_k$ is a basis of $\ker(T)$ and

(ii) $T(\mathbf{v}_{k+1}), T(\mathbf{v}_{k+2}), \ldots, T(\mathbf{v}_n)$ is a basis of $\mathrm{im}(T)$.

Proof. Choose any basis $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_k$ of $\ker(T)$, and extend it to a basis $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ of V. I claim that $T(\mathbf{v}_{k+1}), T(\mathbf{v}_{k+2}), \ldots, T(\mathbf{v}_n)$ span $\mathrm{Im}(T)$. Indeed, if $\mathbf{w} \in \mathrm{Im}(T)$, then $\mathbf{w} = T(\mathbf{v})$ for some $\mathbf{v} \in V$. But then $v = \sum a_i \mathbf{v}_i$, so

$$T(\mathbf{v}) = \sum a_i T(\mathbf{v}_i) = \sum_{i=k+1}^{n} a_i T(\mathbf{v}_i),$$

by the choice of the basis. To see that $T(\mathbf{v}_{k+1}), T(\mathbf{v}_{k+2}), \ldots, T(\mathbf{v}_n)$ are independent, let

$$\sum_{i=k+1}^{n} a_i T(\mathbf{v}_i) = \mathbf{0}.$$

Then $T(\sum_{i=k+1}^{n} a_i \mathbf{v}_i) = \mathbf{0}$, so $\sum_{i=k+1}^{n} a_i \mathbf{v}_i \in \ker(T)$. But if $\sum_{i=k+1}^{n} a_i \mathbf{v}_i \neq \mathbf{0}$, the $\mathbf{v}_i$ $(1 \leq i \leq n)$ cannot be a basis, since every vector in $\ker(T)$ is a linear combination of the $\mathbf{v}_i$ with $1 \leq i \leq k$. This shows that $\sum_{i=k+1}^{n} a_i \mathbf{v}_i = \mathbf{0}$, so each $a_i = 0$. □

This Theorem is the final version of the basic principle that in a linear system, the number of free variables plus the number of corner variables is the total number of variables stated in (3.4).

7.5.3 Vector Space Isomorphisms

One of the nicest applications of Proposition7.14 is the result that if V and W are two vector spaces over $\mathbb{F}$ having the same dimension, then there exists a one to one linear transformation $T : V \to W$ such that $T(V) = W$. Hence, in a sense, we can't distinguish finite dimensional vector spaces over a field if they have the same dimension. To construct this T, choose a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of V and a basis $\mathbf{w}_1, \ldots, \mathbf{w}_n$ of W. All we have to do is let $T : V \to W$ be the unique linear transformation such that $T(\mathbf{v}_i) = \mathbf{w}_i$ if $1 \leq i \leq k$. We leave it as an exercise to show that T satisfies all our requirements. Namely, T is one to one and onto, i.e. $T(V) = W$.

Definition 7.7. Let V and W be two vector spaces over $\mathbb{F}$. A linear transformation $S : V \to W$ which is both one to one and onto (i.e. $\mathrm{im}(T) = W$) is called an *isomorphism* between V and W.

The argument above shows that every pair of subspaces of $\mathbb{F}^n$ and $\mathbb{F}^m$ of the same dimension are isomorphic. (Thus a plane is a plane is a plane.) The converse of this assertion is also true.

Proposition 7.17. *Any two finite dimensional vector spaces over the same field which are isomorphic have the same dimension.*

We leave the proof as an exercise.

Example 7.15. Let's compute the dimension and a basis of the space $L(\mathbb{F}^3, \mathbb{F}^3)$. Consider the transformation $\Phi : L(\mathbb{F}^3, \mathbb{F}^3) \to \mathbb{F}^{3\times 3}$ defined by $\Phi(T) = M_T$, the matrix of T. We have already shown that Φ is linear. In fact, Proposition 7.4 tells us that Φ is one to one and $\mathrm{Im}(\Phi) = \mathbb{F}^{3\times 3}$. Hence Φ is an isomorphism. Thus $\dim L(\mathbb{F}^3, \mathbb{F}^3) = 9$. To get a basis of $L(\mathbb{F}^3, \mathbb{F}^3)$, all we have to do is find a basis of $\mathbb{F}^{3\times 3}$. But a basis is of $\mathbb{F}^{3\times 3}$ given by the matrices E_{ij} such that E_{ij} has a one in the (i, j) position and zeros elsewhere. Every $A \in \mathbb{F}^{3\times 3}$ has the form $A = (a_{ij}) = \sum_{i,j} a_{ij} E_{ij}$, so the E_{ij} span. If $\sum_{i,j} a_{ij} E_{ij}$ is the zero matrix, then obviously each $a_{ij} = 0$, so the E_{ij} are also independent. Thus we have a basis of $\mathbb{F}^{3\times 3}$, and hence we also have one for $L(\mathbb{F}^3, \mathbb{F}^3)$.

This example can easily be extended to the space $L(\mathbb{F}^n, \mathbb{F}^m)$. In particular, we have just given a proof of Proposition 7.2.

Exercises

Exercise 7.33. Suppose $T : V \to V$ is a linear transformation, where V is finite dimensional over $\mathbb{F}$. Find the relationship between between $\mathcal{N}(\mathcal{M}_{\mathcal{B}}^{\mathcal{B}})$ and $\mathcal{N}(\mathcal{M}_{\mathcal{B}'}^{\mathcal{B}'})$, where $\mathcal{B}$ and $\mathcal{B}'$ are any two bases of V.

Exercise 7.34. Find a description of both the column space and null space of the matrix

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

Exercise 7.35. Using only the basic definition of a linear transformation, show that the image of a linear transformation T is a subspace of the target of T. Also, show that if V is a finite dimensional vector space , then $\dim T(V) \leq \dim V$.

Exercise 7.36. Let A and B be $n \times n$ matrices.

(a) Explain why the null space of A is contained in the null space of BA.

(b) Explain why the column space of A contains the column space of AB.

(c) If $AB = O$, show that the column space of B is contained in $\mathcal{N}(A)$.

Exercise 7.37. Consider the subspace W of $\mathbb{R}^4$ spanned by $(1, 1, -1, 2)^T$ and $(1, 1, 0, 1)^T$. Find a system of homogeneous linear equations whose solution space is W.

Exercise 7.38. What are the null space and image of

(i) a projection $P_{\mathbf{b}} : \mathbb{R}^2 \to \mathbb{R}^2$,

(ii) the cross product map $T(\mathbf{x}) = \mathbf{x} \times \mathbf{v}$.

Exercise 7.39. What are the null space and image of a reflection $H_{\mathbf{b}} : \mathbb{R}^2 \to \mathbb{R}^2$. Ditto for a rotation $\mathcal{R}_\theta : \mathbb{R}^2 \to \mathbb{R}^2$.

Exercise 7.40. Ditto for the projection $P : \mathbb{R}^3 \to \mathbb{R}^3$ defined by

$$P(x, y, z)^T = (x, y)^T.$$

Exercise 7.41. Let A be a real 3×3 matrix such that the first row of A is a linear combination of A's second and third rows.

(a) Show that $\mathcal{N}(A)$ is either a line through the origin or a plane containing the origin.

(b) Show that if the second and third rows of A span a plane P, then $\mathcal{N}(A)$ is the line through the origin orthogonal to P.

Exercise 7.42. Let $T : \mathbb{F}^n \to \mathbb{F}^n$ be a linear transformation such that $\mathcal{N}(T) = \mathbf{0}$ and $\text{Im}(T) = \mathbb{F}^n$. Prove the following statements.

(a) There exists a transformation $S : \mathbb{F}^n \to \mathbb{F}^n$ with the property that $S(\mathbf{y}) = \mathbf{x}$ if and only if $T(\mathbf{x}) = \mathbf{y}$. Note: S is called the *inverse* of T.

(b) Show that in addition, S is also a linear transformation.

(c) If A is the matrix of T and B is the matrix of S, then $BA = AB = I_n$.

Exercise 7.43. Let $F : \mathbb{R}^2 \to \mathbb{R}^2$ be the linear transformation given by $F\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ x - y \end{pmatrix}$.

(a) Show that F has an inverse and find it.

(b) Verify that if A is the matrix of F, then $AB = BA = I_2$ if B is a matrix of the inverse of F.

Exercise 7.44. Let $S : \mathbb{R}^n \to \mathbb{R}^m$ and $T : \mathbb{R}^m \to \mathbb{R}^p$ be two linear transformations both of which are one to one. Show that the composition $T \circ S$ is also one to one. Conclude that if A is $m \times n$ has $\mathcal{N}(A) = \{\mathbf{0}\}$ and B is $n \times p$ has $\mathcal{N}(B) = \{\mathbf{0}\}$, then $\mathcal{N}(BA) = \{\mathbf{0}\}$ too.

Exercise 7.45. If A is any $n \times n$ matrix over a field $\mathbb{F}$, then A is said to be *invertible* with *inverse* B if B is an $n \times n$ matrix over $\mathbb{F}$ such that $AB = BA = I_n$. In other words, A is invertible if and only if its associated linear transformation is. Show that if $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then A is invertible provided $ad - bc \neq 0$ and that in this case, the matrix $(ad-bc)^{-1}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ is an inverse of A.

Exercise 7.46. Prove Proposition 7.17.

Exercises

Exercise 7.47. Find a bases for the row space and the column space of each of the matrices in Exercise 3.21.

Exercise 7.48. In this problem, the field is $\mathbb{F}_2$. Consider the matrix

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix}.$$

(a) Find a basis of $\text{row}(A)$.
(b) How many elements are in $\text{row}(A)$?
(c) Is (01111) in $\text{row}(A)$?

Exercise 7.49. Suppose A is any real $m \times n$ matrix. Show that when we view both $\text{row}(A)$ and $\mathcal{N}(A)$ as subspaces of $\mathbb{R}^n$,

$$\text{row}(A) \cap \mathcal{N}(A) = \{\mathbf{0}\}.$$

Is this true for matrices over other fields, eg $\mathbb{F}_2$ or $\mathbb{C}$?

Exercise 7.50. Show that if A is any symmetric real $n \times n$ matrix, then $\text{col}(A) \cap \mathcal{N}(A) = \{\mathbf{0}\}$.

Exercise 7.51. Suppose A is a square matrix over an arbitrary field such that $A^2 = O$. Show that $\text{col}(A) \subset \mathcal{N}(A)$. Is the converse true?

Exercise 7.52. Suppose A is a square matrix over an arbitrary field. Show that if $A^k = O$ for some positive integer k, then $\dim \mathcal{N}(A) > 0$.

Exercise 7.53. Suppose A is a symmetric real matrix so that $A^2 = O$. Show that $A = O$. In fact, show that $\text{col}(A) \cap \mathcal{N}(A) = \{\mathbf{0}\}$.

Exercise 7.54. Find a non zero 2×2 symmetric matrix A over $\mathbb{C}$ such that $A^2 = O$. Show that no such a matrix exists if we replace $\mathbb{C}$ by $\mathbb{R}$.

Exercise 7.55. For two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$, the dot product $\mathbf{x} \cdot \mathbf{y}$ can be expressed as $\mathbf{x}^T\mathbf{y}$. Use this to prove that for any real matrix A, A^TA and A have the same nullspace. Conclude that A^TA and A have the same rank. (Hint: consider $\mathbf{x}^TA^TA\mathbf{x}$.)

Exercise 7.56. In the proof of Proposition 5.11, we showed that $\text{row}(A) = \text{row}(EA)$ for any elementary matrix E. Why does this follow once we know $\text{row}(EA) \subset \text{row}(A)$?

7.6 Summary

A linear transformation between two vector spaces V and W over the same field (the domain and target respectively) is a transformation $T : V \to W$ which has the property that $T(a\mathbf{x} + b\mathbf{y}) = aT(\mathbf{x}) + bT(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y}$ in V and a, b in $\mathbb{F}$. In other words, the property defining a linear transformation is that it preserves all linear combinations. Linear transformations are a way of using the linear properties of V to study W. The set of all linear transformations with domain V and target W is another vector space over $\mathbb{F}$ denoted by $L(V, W)$. For example, if V and W are real inner product spaces, we can consider linear transformations which preserve the inner product. Such linear transformations are called orthogonal. The two fundamental spaces associated with a linear transformation are its kernel $\ker(T)$ and its image $\mathrm{Im}(T)$. If the domain V is finite dimensional, then the fundamental relationship from Chapter 3 which said that in a linear system, the number of variables equals the number of free variables plus the number of corner variables takes its final form in the identity which says $\dim V = \dim \ker(T) + \dim \mathrm{Im}(T)$. If V and W are both finite dimensional, $\dim \ker(T) = 0$ and $\dim \mathrm{Im}(T) = \dim W$, then T is one to one and onto. In this case, it is called an isomorphism. We also showed that given an arbitrary basis of V, there exists a linear transformation $T : V \to W$ taking arbitrarily preassigned values on the basis. This is a useful existence theorem, and it also demonstrates how different linear transformations are from arbitrary transformations.

If $V = \mathbb{F}^n$ and $W = \mathbb{F}^m$, then a linear transformation $T : V \to W$ is nothing but an $m \times n$ matrix over $\mathbb{F}$, i.e. an element of M_T of $\mathbb{F}^{m \times n}$. Conversely, every element of $\mathbb{F}^{m \times n}$ defines such a linear transformation. Thus $L(\mathbb{F}^n, \mathbb{F}^m) = \mathbb{F}^{m \times n}$. If V and W are finite dimensional, then whenever we are given bases $\mathcal{B}$ of V and $\mathcal{B}'$ of W, we can associate a unique matrix $\mathcal{M}_{\mathcal{B}'}^{\mathcal{B}}(T)$ to T. There are certain rules for manipulating these matrices whicch we won't repeat here. They amount to the rule $M_{T \circ S} = M_T M_S$ when $S : \mathbb{F}^p \to \mathbb{F}^n$ and $T : \mathbb{F}^n \to \mathbb{F}^m$ are both linear. If we express a linear transformation $T : V \to V$ in terms of two bases $\mathcal{B}$ and $\mathcal{B}'$ of V, then $\mathcal{M}_{\mathcal{B}'}^{\mathcal{B}'}(T) = P \mathcal{M}_{\mathcal{B}}^{\mathcal{B}}(T) P^{-1}$ where $P = \mathcal{M}_{\mathcal{B}}^{\mathcal{B}'}$ is the change of basis matrix $\mathcal{M}_{\mathcal{B}}^{\mathcal{B}'}(I_n)$.

As we have mentioned before, one of the main general questions about linear transformations is this: when is a linear transformation $T : V \to V$ semi-simple? That is, when does there exist a basis $\mathcal{B}$ of V for which $\mathcal{M}_{\mathcal{B}}^{\mathcal{B}}(T)$ is diagonal. Put another way, when is $\mathcal{M}_{\mathcal{B}}^{\mathcal{B}}(T)$ similar to a diagonal matrix? We will solve it in Chapter **??** when $\mathbb{F}$ is algebraically closed.

Chapter 8

An Introduction to the Theory of Determinants

8.1 Introduction

The determinant is a function defined on the set of $n \times n$ matrices $\mathbb{F}^{n\times n}$ over a field $\mathbb{F}$, whose definition goes back to the 18th century. It is one of the richest and most interesting functions in matrix theory, if not all of mathematics. The list of its applications is long and distinguished, including for example the eigentheory of square matrices, the change of variables formula for integration in several variables, and the notion of area in n dimensions. For an interesting sample of 19th century mathematics, take a glance at the remarkable work Theory of Determinants, by J. Muir.

8.2 The Definition of the Determinant

8.2.1 Some comments

If $A \in \mathbb{F}^{n\times n}$, the determinant of A, which is denoted either by $\det(A)$ or by $\det(A)$, is an element of $\mathbb{F}$ with the remarkable property that if $B \in \mathbb{F}^{n\times n}$ also, then $\det(AB) = \det(A)\det(B)$. Moreover, $\det(A) \neq 0$ if and only if A^{-1} exists. Thus, for example, the nullspace of A has positive dimension if and only if $\det(A) = 0$. This is the standard numerical criterion for determining if A is singular. The general definition of $\det(A)$ (8.2) is a sum with $n!$ terms, At first glance, any attempt at finding a general method for computing $\det(A)$ will appear hopeless. Fortunately, we will see that $\det(A)$ can be computed by bringing A into upper triangular form by row

operations.

Besides having many mathematically beautiful properties, the determinant also has basic many applications. For example, it is used to define the characteristic polynomial of a square matrix A. The roots of this polynomial are the eigenvalues of A. Nowadays, there are now many powerful readily available tools for computing determinants and for approximating eigenvalues, but, nevertheless, the characteristic polynomial is still an important basic concept.

8.2.2 The 2×2 case

The first case, of course, is the 1×1 case. Here, we can simply put $\det(a) = a$. Hence we can begin with the 2×2 case. Suppose $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Define the determinant of A to be

$$\det(A) = ad - bc.$$

Sometimes $\det(A)$ is also denoted by $\det(A)$. It is not hard to see that $ad - bc = 0$ if and only if the rows of A are proportional. Thus, A has rank 2 if and only if $ad - bc \neq 0$. Since A has maximal rank, it is invertible, and (as we have shown elsewhere),

$$A^{-1} = \frac{1}{(ad - bc)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Proposition 8.1. *The determinant function* $\det : \mathbb{F}^{2\times 2} \to \mathbb{F}$ *has the following properties:*

(1) $\det(A) \neq 0$ *if and only if* A *is invertible;*

(2) $\det(I_2) = 1$*; and*

(3) $\det(AB) = \det(A)\det(B)$ *for any* $A, B \in \mathbb{F}^{2\times 2}$.

Proof. We indicated the proof of (1) above. Statement (2) is obvious, and (3) can be checked by a direct calculation. □

The remarkable property is (3), which, as mentioned above, holds for all $A, B \in \mathbb{F}^{n\times n}$ for all n.

8.2.3 Some Combinatorial Preliminaries

Unlike many situations we have seen, the definition of $\det(A)$ for 2×2 matrices gives only the slightest hint of how to extend the definition to

the $n \times n$ matrices so that Proposition 8.1 still holds. To give the general definition, we first need to study some elementary combinatorics.

First of all, let X denote any set.

Definition 8.1. A *permutation* or *bijection* of X is a mapping $\sigma : X \to X$ which is both one to one and onto.

Let us put $X_n = \{1, 2, \ldots, n\}$. First of all, we state the

Lemma 8.2. *A mapping $\sigma : X_n \to X_n$ which is either one to one or onto is a permutation. Moreover, the set $S(n)$ of all permutations of X_n has exactly $n!$ elements. Finally, if $\pi, \sigma \in S(n)$, then the composition $\sigma\pi$ and the inverse σ^{-1} are also in $S(n)$.*

Proof. The proof is an application of elementary combinatorics. We will leave as an exercise. □

In order to define the determinant, we need to define what is known as the signature of a permutation. The definition of the signature goes as follows.

Definition 8.2. Let $\sigma \in S(n)$. Then we define the *signature* of σ to be the rational number

$$\operatorname{sgn}(\sigma) = \prod_{i<j} \frac{\sigma(i) - \sigma(j)}{i - j}.$$

Sometimes the signature of σ is also called the *sign* of σ. First of all, notice that $\operatorname{sgn}(\sigma)$ is a nonzero rational number. In fact, we have

Proposition 8.3. *For any $\sigma \in S(n)$, $\operatorname{sgn}(\sigma) = \pm 1$.*

Proof. Since σ is a bijection of X_n, and

$$\frac{\sigma(i) - \sigma(j)}{i - j} = \frac{\sigma(j) - \sigma(j)}{j - i},$$

we see that

$$(\operatorname{sgn}(\sigma))^2 = \prod_{i \neq j} \frac{\sigma(i) - \sigma(j)}{i - j}.$$

Moreover, since

$$\sigma(\sigma^{-1}(i)) - \sigma(\sigma^{-1}(j)) = i - j,$$

each possible value of $i-j$ occurs the same number of times in the numerator as in the denominator. Hence $\operatorname{sgn}(\sigma)^2 = 1$. Thus $\operatorname{sgn}(\sigma) = \pm 1$, so the Propositionis proven. □.

If $i < j$, then

$$\frac{\sigma(i) - \sigma(j)}{i - j} > 0$$

exactly when $\sigma(i) < \sigma(j)$. Therefore

$$\operatorname{sgn}(\sigma) = (-1)^{m(\sigma)},$$

where

$$m(\sigma) = |\{(i, j) \mid i < j,\, \sigma(i) > \sigma(j)\}|.$$

The simplest permutations are what are called transpositions. A *transposition* is a permutation σ such that there are two distinct $a, b \in X_n$ with $\sigma(a) = b,\ \sigma(b) = a$ and $\sigma(i) = i$ if $i \neq a, b$. We will denote the transposition which interchanges $a \neq b$ in X_n by σ_{ab}.

Example 8.1. For example, σ_{12} interchanges 1 and 2 and leaves every integer between 3 and n alone. I claim $m(\sigma_{12}) = 1$. For, the only pair (i, j) such that $i < j$ for which $\sigma_{12}(i) > \sigma_{12}(j)$ is the pair $(1, 2)$. Hence $\operatorname{sgn}(\sigma_{12}) = -1$.

We will need the explicit value for the signature of an arbitrary transposition. This is one of the results in the main theorem on the signature, which we now state and prove.

Theorem 8.4. *The signature mapping* $\operatorname{sgn} : S(n) \to \{\pm 1\}$ *satisfies the following properties:*

(1) *for all* $\sigma, \tau \in S(n)$, *then* $\operatorname{sgn}(\tau\sigma) = \operatorname{sgn}(\tau)\operatorname{sgn}(\sigma)$,

(2) *if* σ *is a transposition, then* $\operatorname{sgn}(\sigma) = -1$, *and*

(3) *if* σ *is the identity, then* $\operatorname{sgn}(\sigma) = 1$.

Proof. First consider $\operatorname{sgn}(\tau\sigma)$. We have

$$\begin{aligned}
\operatorname{sgn}(\tau\sigma) &= \prod_{i<j} \frac{\tau\sigma(i) - \tau\sigma(j)}{i - j} \\
&= \prod_{i<j} \frac{\tau(\sigma(i)) - \tau(\sigma(j))}{\sigma(i) - \sigma(j)} \prod_{i<j} \frac{\sigma(i) - \sigma(j)}{i - j} \\
&= \prod_{r<s} \frac{\tau(r) - \tau(s)}{r - s} \prod_{i<j} \frac{\sigma(i) - \sigma(j)}{i - j}.
\end{aligned}$$

The third equality follows since

$$\frac{\tau(\sigma(i)) - \tau(\sigma(j))}{\sigma(i) - \sigma(j)} = \frac{\tau(\sigma(j)) - \tau(\sigma(i))}{\sigma(j) - \sigma(i)},$$

and since σ is a permutation of $1, 2, \ldots, n$. Hence $\text{sgn}(\tau\sigma) = \text{sgn}(\tau)\text{sgn}(\sigma)$, so (1) is proven.

We now prove (2), using the result of Example 8.1. Consider an arbitrary transposition σ_{ab}, where $1 \leq a < b \leq n$. I claim

$$\sigma_{ab} = \sigma_{1b}\sigma_{2a}\sigma_{12}\sigma_{2a}\sigma_{1b}. \tag{8.1}$$

We leave this as an exercise. By (1),

$$\text{sgn}(\sigma_{ab}) = \text{sgn}(\sigma_{1b})\text{sgn}(\sigma_{2a})\text{sgn}(\sigma_{12})\text{sgn}(\sigma_{2a})\text{sgn}(\sigma_{1b}).$$

We know from Example 8.1 that $\text{sgn}(\sigma_{12}) = -1$. Clearly,

$$\text{sgn}(\sigma_{1b})\text{sgn}(\sigma_{2a})\text{sgn}(\sigma_{2a})\text{sgn}(\sigma_{1b}) > 0,$$

hence $\text{sgn}(\sigma_{ab}) = -1$. The last claim is obvious, so the proof is finished. □

8.2.4 Permutations and Permutation Matrices

Permutations and permutation matrices are closely related. Since any permutation matrix P is a square matrix of zeros and ones so that each row and each column contains exactly one non-zero entry, P is uniquely determined by a permutation σ. Namely, if the ith column of P contains a 1 in the jth row, we put $\sigma(i) = j$. If P is $n \times n$, this defines a unique element $\sigma \in S(n)$. Conversely, given $\sigma \in S(n)$, let us define P_σ by putting

$$P_\sigma = (\mathbf{e}_{\sigma(1)}\ \mathbf{e}_{\sigma(2)}\ \cdots\ \mathbf{e}_{\sigma(n)}),$$

where $\mathbf{e}_{\sigma(i)}$ is the vector with 0 in each component except the $\sigma(i)$th component, which is 1.

Proposition 8.5. *The mapping $\sigma \to P_\sigma$ defines a one to one corresponence between $S(n)$ and the set $P(n)$ of $n \times n$ permutation matrices.*

Proof. This is obvious consequence of the definitions. □

Example 8.2. In order to have some notation for a permutation, we will represent $\sigma \in S(n)$ by the symbol

$$[\sigma(1), \sigma(2), \ldots, \sigma(n)].$$

For example, the identity permutation of $S(3)$ is $[1, 2, 3]$. It corresponds to I_3. The permutation $[2, 3, 1] \in S(3)$ corresponds to

$$P_{[2,3,1]} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

and so forth. Note that the non-zero element of the ith column of P_σ is $p_{\sigma(i)i}$. Let's see what what happens when we form the product $P_{[2,3,1]}A$. We have

$$P_{[2,3,1]}A = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \mathbf{a}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{a}_3 \\ \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix}.$$

To give another example, we have

$$P_{[3,1,2]}A = \begin{pmatrix} \mathbf{a}_2 \\ \mathbf{a}_3 \\ \mathbf{a}_1 \end{pmatrix}.$$

More generally, let $P_{[i,j,k]}$ be the 3×3 permutation matrix in which the first row of I_3 is in the ith row, the second in the jth row and the third in the kth row. Then $P_{[i,j,k]}A$ is obtained from A by permutating the rows of A by the permutation $[i, j, k]$.

8.2.5 The General Definition of $\det(A)$

Now that we have the signature of a permutation, we can give the definition of $\det(A)$ for any n.

Definition 8.3. Let $\mathbb{F}$ be any field, and let $A \in \mathbb{F}^{n\times n}$. Then the determinant of A is defined to be

$$\det(A) := \sum_{\pi \in S(n)} \operatorname{sgn}(\pi)\ a_{\pi(1)1} a_{\pi(2)2} \cdots a_{\pi(n)n} \tag{8.2}$$

Example 8.3. Obviously, $\det(A) = a$ if A is the 1×1 matrix (a). Let's compute $\det(A)$ in the 2×2 cases. There are only two elements $\sigma \in S(2)$, namely the identity and σ_{12}. By definition,

$$\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = +a_{11}a_{22} + \operatorname{sgn}(\sigma_{12}) a_{21}a_{12} = a_{11}a_{22} - a_{21}a_{12}.$$

Thus the expression (8.2) agrees with the original definition.

Example 8.4. For the 3×3 case, we begin by listing the elements $\sigma \in S(3)$ and their signatures. Denoting each σ by the triple $[\sigma(1), \sigma(2), \sigma(3)]$, we get the table

$$\begin{array}{cccccccc} \pi & [1,2,3] & [1,3,2] & [2,3,1] & [2,1,3] & [3,1,2] & [3,2,1] \\ \operatorname{sgn}(\pi) & 1 & -1 & +1 & -1 & +1 & -1 \end{array}.$$

Hence, if $A = (a_{ij})$, then

$$\begin{aligned} \det(A) = a_{11}a_{22}a_{33} - a_{11}a_{32}a_{23} + a_{21}a_{32}a_{13} \\ -a_{21}a_{12}a_{33} + a_{31}a_{12}a_{23} - a_{31}a_{32}a_{13},, \end{aligned}$$

which is the standard formula for a 3×3 determinant.

8.2.6 The Determinant of a Permutation Matrix

The first result about determinants is

Proposition 8.6. *If $P \in P(n)$ has the form P_σ, then* $\det(P) = \operatorname{sgn}(\sigma)$.

Proof. We know that the nonzero entries of P_σ are the entries of the form $p_{\sigma(i)i}$, all of which are 1 (see Example 8.2. Applying the definition of det, namely

$$\det(P_\sigma) = \sum_{\pi \in S(n)} \operatorname{sgn}(\pi)\ p_{\pi(1)1} p_{\pi(2)2} \cdots p_{\pi(n)n}, \tag{8.3}$$

we see that the only non-zero term is

$$\operatorname{sgn}(\sigma) p_{\sigma(1)1} p_{\sigma(2)2} \cdots p_{\sigma(n)n} = \operatorname{sgn}(\sigma).$$

Therfore, $\det(P_\sigma) = \operatorname{sgn}(\sigma)$, as claimed. □

Notice that the row swap matrices of row operations are the permutation matices which come from transpositions. Hence, we get

Corollary 8.7. *If S is a row swap matrix, then* $\det(S) = -1$. *Furthermore, if a permutation matrix P is written as a product of row swaps matrices, say $P = S_1 S_2 \cdots S_m$, then* $\det(P) = (-1)^m$.

Proof. This follows from Proposition8.6. □

We can interpret the m of this Corollary as the number of row swaps needed to bring P back to the identity. Another consequence is that for any row swap matrix S and permutation matrix P, $\det(SP) = -\det(P)$ since SP involves one more row swap than P Yet another consequence is that the number of row swaps needed to express P is always even or always odd.

Exercises

Exercise 8.1. Write down two 4×4 matrices A, B each with at most two zero entries such that $\det(A)|B| \neq 0$.

Exercise 8.2. If A is $n \times n$ and r is a scalar, find $|rA|$.

Exercise 8.3. In the 2×2 case, find:

(a) the determinant of a reflection matrix;

(b) the determinant of a rotation matrix;

(c) classify the elements of $O(2, \mathbb{R})$ according to the determinant.

Exercise 8.4. Prove Equation (8.1).

8.3 Determinants and Row Operations

So far, we've given very little insight as to how determinants are actually computed, but it's clear we aren't going to get very far by trying to compute them from the definition. We now show how to compute $\det(A)$ via elementary row operations.

It will simplify matters (especially for studying the laplace expansion) to have a slight reformulation of the definition of $\det(A)$, which we will give first. For any $A \in \mathbb{F}^{n\times n}$, put

$$\delta(A) = \prod_{i=1}^{n} a_{ii}.$$

Thus $\delta(A)$ is the product of all the diagonal entries of A.

For example, if A is upper or lower triangular, then $\det(A) = \delta(A)$.

Proposition 8.8. *If $A \in \mathbb{F}^{n\times n}$, then*

$$\begin{aligned}\det(A): &= \sum_{\sigma\in S(n)} \operatorname{sgn}(\sigma)\ a_{\sigma(1)1}a_{\sigma(2)2}\cdots a_{\sigma(n)n}\\ &= \sum_{\sigma\in S(n)} \det(P_\sigma)\delta(P_\sigma A),\end{aligned}$$

Before giving the proof, let's calculate a 3×3 example. Consider $\delta(PA)$, where P is the matrix $P_{[2,3,1]}$. Thus

$$P. = \begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix}, \quad \text{and} \quad A = \begin{pmatrix}\mathbf{a}_1\\\mathbf{a}_2\\\mathbf{a}_3\end{pmatrix}.$$

Recall from Example 8.2 that

$$P_{[2,3,1]}A = \begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix}\begin{pmatrix}\mathbf{a}_1\\\mathbf{a}_2\\\mathbf{a}_3\end{pmatrix} = \begin{pmatrix}\mathbf{a}_3\\\mathbf{a}_1\\\mathbf{a}_2\end{pmatrix}.$$

Hence

$$\delta(P_{[2,3,1]}A) = a_{31}a_{12}a_{23}.$$

Thus, since $\sigma(1) = 2$, $\sigma(2) = 3$ and $\sigma(3) = 1$, we see that

$$\delta(P_\sigma A) = a_{\sigma^{-1}(1)1}a_{\sigma^{-1}(2)2}a_{\sigma^{-1}(3)3}.$$

Let's now give the proof of Proposition8.8.

Proof. From the above calculation,

$$\delta(P_\sigma A) = a_{\sigma^{-1}(1)1} a_{\sigma^{-1}(2)2} \cdots a_{\sigma^{-1}(n)n}.$$

But as $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\sigma^{-1})$ (why?), we have

$$\operatorname{sgn}(\sigma)\delta(P_\sigma A) = \operatorname{sgn}(\sigma^{-1}) a_{\sigma^{-1}(1)1} a_{\sigma^{-1}(2)2} \cdots a_{\sigma^{-1}(n)n}.$$

Now, as σ varies over all of $S(n)$, so does σ^{-1}, hence we see that

$$\begin{aligned}
\det(A) &= \sum_{\sigma \in S(n)} \operatorname{sgn}(\sigma)\ a_{\sigma(1)1} a_{\sigma(2)2} \cdots a_{\sigma(n)n} \\
&= \sum_{\sigma \in S(n)} \operatorname{sgn}(\sigma^{-1})\ a_{\sigma^{-1}(1)1} a_{\sigma^{-1}(2)2} \cdots a_{\sigma^{-1}(n)n} \\
&= \sum_{\sigma \in S(n)} \operatorname{sgn}(\sigma)\delta(P_\sigma A) \\
&= \sum_{\sigma \in S(n)} \det(P_\sigma)\delta(P_\sigma A).
\end{aligned}$$

□

8.3.1 The Main Result

The strategy for computing determinants is explained by first considering the triangular case.

Proposition 8.9. *Suppose A is $n \times n$ and upper triangular, that is, every element in A below the diagonal is zero. Then*

$$\det(A) = \delta(A) = a_{11} a_{22} \cdots a_{nn}.$$

The same formula also holds for a lower triangular matrix.

The point is that the only nonzero term in $\det(A)$ is $a_{11} a_{22} \cdots a_{nn}$. For if P is a permutation matrix different from the identity, then there has to be an index i so that the ith row of PA is different from the ith row of A. But that means PA has to have a 0 on the diagonal, so $\delta(PA) = 0$.

Hence the key to computing higher order determinants is to use row operations to bring A into triangular form. Thus we need to investigate how $\det(A)$ changes after a row operation is performed on A. The following properties of the determinant function give the rather pleasant answer.

Theorem 8.10. *Let $\mathbb{F}$ be an arbitrary field, and suppose $n \geq 2$. Then the following properties hold for any $A, B \in \mathbb{F}^{n\times n}$:*

(Det I) *if B is obtained from A by a row swap, then*

$$\det(B) = -\det(A), \tag{8.4}$$

(Det II) *if B is obtained from A by multiplying a row of A by a (possibly zero) scalar r, then*

$$\det(B) = r\det(A), \tag{8.5}$$

(Det III) *if B is obtained from A by replacing the ith row by itself plus a multiple of the jth row, $i \neq j$, then*

$$\det(B) = \det(A), \tag{8.6}$$

and

(Det IV) $\det(I_n) = 1$.

Proof. Suppose, $B = SA$, where S swaps two rows, but leaves all the other rows alone. By the previous Proposition,

$$\det(B) = \det(SA) = \sum_{P\in P(n)} \det(P)\delta(P(SA)).$$

Since S and P are permutation matrices, so is $Q = PS$. Moreover, if we hold S fixed and vary P over $P(n)$, then $Q = PS$ varies over all of $P(n)$ also. Therefore,

$$\det(SA) = \sum_{P\in P(n)} \det(PS)\delta(PSSA) = \sum_{P\in P(n)} -\det(P)\delta(PS^2A).$$

But $S^2 = I_n$, and $\det(PS) = -\det(P)$, there being one more row swap in PS than in P. Hence, by Proposition8.8, $\det(SA) = -\det(A)$.

The proof of **(Det II)** is obvious. If E multiplies the i-th row of A by the scalar r, then for every permutation matrix P, $\delta(P(EA)) = r\delta(PA)$. Thus $\det(EA) = r\det(A)$. □

(Det III) follows from two facts. First of all, suppose $A, A' \in \mathbb{F}^{n\times n}$ coincide in all but one row, say the kth row. That is, if $A = (a_{ij})$ and $A' = (a'_{ij})$, then $a_{ij} = a'_{ij}$ as long as $i \neq k$. Now define a matrix $B = (b_{ij})$

by $b_{ij} = a_{ij} = a'_{ij}$ if $i \neq k$, and $b_{kj} = a_{kj} + a'_{kj}$. Then it follows from the definition of det that

$$\det(B) = \det(A) + \det(A'). \tag{8.7}$$

To prove (8.7), consider any term

$$\operatorname{sgn}(\pi)\ b_{\pi(1)1} b_{\pi(2)2} \cdots b_{\pi(n)n}$$

of $\det(B)$. Let j be the index such that $\pi(j) = k$. Then, by the definition of B, $b_{\pi(j)j} = a_{\pi(j)j} + a'_{\pi(j)j}$. Hence,

$$\begin{aligned} \operatorname{sgn}(\pi)\ b_{\pi(1)1} b_{\pi(2)2} \cdots b_{\pi(n)n} &= \operatorname{sgn}(\pi)\ b_{\pi(1)1} b_{\pi(2)2} \cdots b_{\pi(j)j} \cdots b_{\pi(n)n} \\ &= \operatorname{sgn}(\pi)\ a_{\pi(1)1} a_{\pi(2)2} \cdots a_{\pi(n)n} + \\ &\quad \operatorname{sgn}(\pi)\ a'_{\pi(1)1} a'_{\pi(2)2} \cdots a'_{\pi(n)n} \end{aligned}$$

This implies (8.7).

Next, suppose $C \in \mathbb{F}^{n\times n}$ has two identical rows. There is an easy way to see $\det(C) = 0$ provided the characteristic of $\mathbb{F}$ is different from 2. For, if the rth and sth rows of C coincide, then $P_\sigma C = C$, where $\sigma = \sigma_{rs}$. Hence, by **(Det I)**, $\det(C) = -\det(C)$, which implies $2\det(C) = 0$. Thus $\det(C) = 0$ unless the characteristic of $\mathbb{F}$ is 2. We can give a proof that $\det(C) = 0$ in any characteristic by applying the Laplace expansion. This is proved in §8.3.4, so we will postpone the proof until then.

Now suppose E is the elementary matrix of type III is obtained from I_n by replacing the ith row of I_n by itself plus a times the jth row, where $i \neq j$. Thus $B = EA$. We know from (8.7) and **(Det II)** that

$$\det(B) = \det(A) + a\det(C),$$

where the ith and jth rows of C are equal. But $\det(C) = 0$, so $\det(B) = \det(A)$, and hence the proof of **(Det III)** is finished.

(Det IV) is obvious, so the Theorem is proved. □

In particular, since $\det(I_n) = 1$, **Det I-Det III** imply that if E is an elementary matrix, then

$$\det(E) = \begin{cases} -1 & \text{if } E \text{ is of type I,} \\ r & \text{if } E \text{ is of type II,} \\ 1 & \text{if } E \text{ is of type III} \end{cases}$$

Therefore we can summarize **Det I-Det III** in the following way.

Corollary 8.11. *If $A, E \in \mathbb{F}^{n\times n}$ and E is an elementary, then*

$$\det(EA) = \det(E)\det(A). \tag{8.8}$$

Proof. If E is a row swap , then $\det(EA) = -\det(A)$ by **Det I**. If E is of type II, say $\det(E) = r$, then by **Det II**, $\det(EA) = r\det(A) = \det(E)\det(A)$. If E is of type III, then $\det(E) = 1$ while $\det(EA) = \det(A)$ by **Det III**. □

8.3.2 Properties and consequences

The way one uses Theorem 8.10 to evaluate $\det(A)$ is clear. First find elementary matrices $E_1, \ \ldots \ , E_k$ such that $U = E_k \cdots E_1 A$ is upper triangular. By Proposition 8.9 and repeated application of Theorem 8.10,

$$\det(U) = \det(E_1)\cdots\det(E_k)\det(A) = u_{11}u_{22}\cdots u_{nn},$$

where the u_{ii} are the diagonal entries of U. Since no $\det(E_i) = 0$,

$$\det(A) = \frac{u_{11}u_{22}\cdots u_{nn}}{\det(E_1)\cdots\det(E_k)}. \tag{8.9}$$

Example 8.5. Let us compute $\det(A)$, where

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

taking the field of coefficients to be $\mathbb{Q}$. We can make the following sequence of row operations, all of type III except for the last, which is a row swap.

$$A \to \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \end{pmatrix} \to$$

$$\begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

Thus $\det(A) = -1$.

Example 8.6. Let us next compute $\det(A)$, where A is the matrix of the previous example, this time taking the field of coefficients to be $\mathbb{Z}_2$. First add the first row to the third and fourth rows successively. Then we get

$$\det(A) = \det \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}.$$

Since the field is $\mathbb{Z}_2$, row swaps also leave $\det(A)$ unchanged. Thus

$$\det(A) = \det \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}.$$

Adding the second row to the third row and the fourth row successively, we get

$$\det(A) = \det \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Finally, switching the last two rows, we get

$$\det(A) = \det \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = 1.$$

One can simplify evaluating $\det(A)$ even more in some special cases. For example, if A has the form

$$\begin{pmatrix} B & C \\ O & D \end{pmatrix}, \tag{8.10}$$

where the submatrices B and C are square, then $\det(A) = \det(B)\det(D)$. The proof is similar to the proof of (8.9).

To see what the determinant sees, notice that (8.9) implies $\det(A) \neq 0$ iff each $u_{ii} \neq 0$. Since this is the case precisely when A has maximal rank, we get

Proposition 8.12. *If A is $n \times n$, then $\det(A) \neq 0$ iff when A has rank n. Moreover, if A has rank n and we write $A = E_1 \cdots E_k$ where the E_i are elementary, then $\det(A) = \det(E_1) \cdots \det(E_k)$.*

Proof. The first statement is verified before the proposition. The second follows immediately from (8.9) and the fact noted above that if E is elementary, then $\det(E^{-1}) = \det(E)^{-1}$. □

We can now prove the product formula (Proposition 8.1, part (2)) to the $n \times n$.

Theorem 8.13. *If A and B are any $n \times n$ matrices over $\mathbb{F}$, then* $\det(AB) = \det(A)\det(B)$.

Proof. If A and B both have rank n, each one of them can be expressed as a product of elementay matrices, say $A = E_1 \cdots E_k$ and $B = E_{k+1} \cdots E_m$. Then $AB = (E_1 \cdots E_k)(E_{k+1} \cdots E_m)$, so

$$\det(AB) = \det(E_1) \cdots \det(E_k) \det(E_{k+1}) \cdots \det(E_m) = \det(A)\det(B).$$

by Proposition 8.12. To finish the proof, we have to show that if either $\det(A)$ or $\det(B)$ is zero, then $\det(AB) = 0$. However, if $\det(AB) \neq 0$, then AB is invertible, and we know this implies both A and B are invertible. Hence if either $\det(A)$ or $\det(B)$ is zero, then $\det(AB) = 0$, and the proof is complete. □

Proposition 8.14. *If A is invertible, then* $\det(A^{-1}) = \det(A)^{-1}$.

Proof. This follows from the previous result since $AA^{-1} = I_n$ implies

$$\det(AA^{-1}) = \det(A)\det(A^{-1}) = 1.$$

□

Another remarkable property of the determinant is that both A and A^T have the same determinant.

Proposition 8.15. *If A is any square matrix, then* $\det(A) = \det(A^T)$.

Proof. We know that A and A^T have the same rank, so the result is true if $\det(A) = 0$. Hence suppose A has maximal rank. Express A as a product of elementary matrices, say $A = E_1 E_2 \cdots E_k$. This gives

$$A^T = (E_1 E_2 \cdots E_k)^T = E_k^T E_{k-1}^T \cdots E_1^T,$$

by the rule for transposing a product. Thus it suffices to show that for any elementary matrix E, we have

$$\det(E^T) = \det(E).$$

This is clear if E is of type II or III, since elementary matrices of type II are symmetric and the transpose of a type III elementary matrix is also of type III, so in both cases the determinant is 1. If E is of type I, then $E^{-1} = E^T$, so $\det(E) = \det(E^{-1}) = \det(E^T)$, the common value being -1. □

REMARK: Finally, one can ask the question of how unique the determinant is. As we have remarked above, it is a staightforward consequence of the definition that $\det(A)$ is an $\mathbb{F}$-linear function of the rows of A. That is, if we hold all rows of A fixed except the ith, the the determinant is then a linear function of the ith row, and this is true for any i. Then we have

Theorem 8.16. *The determinant is the only function $F : \mathbb{F}^{n\times n} \to \mathbb{F}$ such that:*

(1) *F is $\mathbb{F}$-linear in each row,*

(2) *$F(B) = -F(A)$ if B is obtained from A by a row swap,*

(3) *$F(A) = 0$ if two rows of A are equal, and*

(4) *$F(I_n) = 1$.*

Then $F(A) = \det(A)$for all $A \in \mathbb{F}^{n\times n}$. In fact, if the characteristic of $\mathbb{F}$ is different from 2, then condition (3) can be dropped.

Proof. In fact, these conditions tell us that for any elementary matrix E, $F(EA)$ is computed from $F(A)$ in exactly the same way $\det(EA)$ is computed from $\det(A)$. □

8.3.3 The determinant of a linear transformation

Here is an application of the product theorem. Let V be a finite dimensional vector space over $\mathbb{F}$, and suppose $T : V \to V$ is linear. We now make the following definition. To do so, let

Definition 8.4. The determinant $\det(T)$ of T is defined to be $\det(A)$, where $A \in \mathbb{F}^{n\times n}$ is any matrix representation of T with respect to a basis.

To see that $\det(T)$ is well defined we have to check the if B is some other matrix representation, then $\det(A) = \det(B)$. But we know from Proposition **??** that if A and B are matrices of T with respect to a basis, then A and B are similar, i.e. there exists an invertible $P \in \mathbb{F}^{n\times n}$ such that $B = P^{-1}AP$. Thus

$$\det(B) = \det(P^{-1}AP) = \det(P^{-1})\det(A)\det(P) = \det(A)$$

so $\det(T)$ is indeed well defined.

8.3.4 The Laplace Expansion

The determinant is frequently defined in terms of the Laplace expansion. The Laplace expansion is useful for evaluating $\det(A)$ when A has entries which are functions. In fact, this situation will arise as soon as we introduce the characteristic polynomial of A.

Suppose A is $n \times n$, and let A_{ij} denote the $(n-1) \times (n-1)$ submatrix obtained from A by deleting its ith row and jth column.

Theorem 8.17. *For any $A \in \mathbb{F}^{n\times n}$, we have*

$$\det(A) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij}). \tag{8.11}$$

This is the Laplace expansion along the jth column. The corresponding Laplace expansion of along the ith row is

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij}). \tag{8.12}$$

Proof. Since $\det(A) = \det(A^T)$, it suffices to prove (8.11) . For simplicity, we will assume $j = 1$, the other cases being similar. Now,

$$\begin{aligned}
\det(A) &= \sum_{\sigma\in S(n)} \operatorname{sgn}(\sigma)\ a_{\sigma(1)1} a_{\sigma(2)2} \cdots a_{\sigma(n)n} \\
&= a_{11} \sum_{\sigma(1)=1} \operatorname{sgn}(\sigma) a_{\sigma(2)2} \cdots a_{\sigma(n)n} + \\
&\quad a_{21} \sum_{\sigma(1)=2} \operatorname{sgn}(\sigma) a_{\sigma(2)2} \cdots a_{\sigma(n)n} + \\
&\quad + \cdots + a_{n1} \sum_{\sigma(1)=n} \operatorname{sgn}(\sigma) a_{\sigma(2)2} \cdots a_{\sigma(n)n}
\end{aligned}$$

If $\sigma \in S(n)$, let P'_σ denote the element of $\mathbb{F}^{n\times n}$ obtained from P_σ by deleting the first column and the $\sigma(1)$st row. Since $p_{\sigma(i)i} = 1$, it follows that $P'_\sigma \in P(n-1)$ (why?). Note that $\det(P_\sigma) = (-1)^{(\sigma(1)-1)} \det(P'_\sigma)$, since if bringing P'_σ to I_{n-1} by row swaps uses t steps, one needs $t + \sigma(1) - 1$ row swaps to bring P_σ to the identity. Next, recall that

$$\det(A) = \sum_{\sigma\in S(n)} \det(P_\sigma) \delta(P_\sigma A).$$

Since $\det(P_\sigma) = (-1)^{(\sigma(1)-1)} \det(P'_\sigma)$, we see that for each r with $1 \leq r \leq n$,

$$\sum_{\substack{\sigma \in S(n) \\ \sigma(1)=r}} \det(P_\sigma)\delta(P_\sigma A) = (-1)^{(r-1)} a_{r1} \sum_{\substack{\sigma \in S(n) \\ \sigma(1)=r}} \det(P'_\sigma)\delta(P'_\sigma A_{r1}).$$

But the right hand side is certainly $(-1)^{(r-1)} a_{r1} \det(A_{r1})$, since every element of $P(n-1)$ is P'_σ for exactly one $\sigma \in S(n)$ with $\sigma(1) = r$. Therefore,

$$\det(A) = \sum_{i=1}^{n} (-1)^{i-1} a_{i1} \det(A_{i1}),$$

which is the desired formula. □

Example 8.7. If A is 3×3, expanding $\det(A)$ along the first column gives

$$\det(A) = a_{11}(a_{22}a_{33} - a_{32}a_{23}) - a_{21}(a_{12}a_{23} - a_{13}a_{32}) + a_{31}(a_{12}a_{23} - a_{13}a_{22}).$$

This is the well known formula for the triple product $\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)$ of the rows of A.

Example 8.8. Here is an example where the Laplace expansion is useful. Suppose we want to find all values of x such that the matrix

$$C_x = \begin{pmatrix} 1-x & 2 & 0 \\ 2 & 1-x & -1 \\ 0 & -1 & 2-x \end{pmatrix}$$

has rank less that 3, i.e. is singular. Hence we will try to solve the equation $\det(C_x) = 0$ for x. Clearly, row operations aren't going to be of much help in finding $\det(C_x)$, so we will use Laplace, as in the previous example. Expanding along the first column gives

$$\begin{aligned} \det(C_x) &= (1-x)\big((1-x)(2-x) - (-1)(-1)\big) - 2\big(2(2-x) - 0(-1)\big) \\ &= -x^3 + 4x - 7 \end{aligned}$$

Hence C_x is singular at the three roots of $x^3 - 4x + 7 = 0$.

A moment's consideration is all that is needed to see that the Laplace expansion isn't even in the same ballpark as row ops when it comes to giving an efficient a procedure for computing $\det(A)$. All that the Laplace expansion is doing is giving a systematic way of organizing all the terms. Using Laplace to evaluate even a 20×20 determinant is impractical, except

possibly for a super computer (note 20! = 2432902008176640000). Yet in applications of linear algebra to biotechnology, one might need to evaluate a 2000×2000 determinant. In fact, calculations involving genomes routinely require evaluating much larger determinants.

REMARK: Recall that we still need to show that $\det(C) = 0$ if two rows of C coincide (or are proportional) in every characteristic. We now fill in this gap.

Proposition 8.18. *Suppose $\mathbb{F}$ is a field of arbitrary characteristic and $n > 1$. Then if $C \in \mathbb{F}^{n\times n}$ has two equal rows, $\det(C) = 0$.*

Proof. This is an ideal situation in which to use mathematical induction. We know that the result is true if $n = 2$. Now make the inductive hypothesis that the result is true for all $C \in \mathbb{F}^{m\times m}$, where $m \leq n-1$. Next, let $C \in \mathbb{F}^{n\times n}$ and suppose C has two equal rows, say the ith and jth, We may in fact suppose, without any loss of generality, that $i, j \neq 1$ (why?). Applying the Laplace expansion along the first row, we see that $\det(C)$ is a sum of $(n-1) \times (n-1)$ determinants, each of which has two equal rows. By the inductive hypothesis, each of these $(n-1) \times (n-1)$ determinants is 0. Therefore $\det(C) = 0$. □

8.3.5 Cramer's rule

Another reason the Laplace expansion is important, at least from a theoretical point of view, is that it gives a closed formula for the inverse of a matrix. For example, if A is 2×2,

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}.$$

Inspecting this formula may suggest the correct formula for A^{-1} in the general case.

Proposition 8.19. *Suppose $\det(A) \neq 0$ and let $b_{ij} = (-1)^{i+j} \det(A_{ji})$, where, as above, A_{ij} denotes the $(n-1) \times (n-1)$ submatrix of A obtained by deleting A's ith row and jth column. Then*

$$A^{-1} = \frac{1}{\det(A)}(b_{ij}).$$

For example, the inverse of an invertible 3×3 matrix A is given as follows:

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} A_{11} & -A_{21} & A_{31} \\ -A_{12} & A_{22} & -A_{23} \\ A_{13} & -A_{23} & A_{33} \end{pmatrix}.$$

Exercises

Exercise 8.5. Find all possible values the determinant of an arbitrary $n \times n$ orthogonal matrix Q.

Exercise 8.6. Two square matrices A and B are said to be similar if there exists a matrix M so that $B = MAM^{-1}$. Show that similar matrices have the same determinants.

Exercise 8.7. Suppose P is an $n \times n$ matrix so that $PP = P$. What is $\det(P)$? What if $P^4 = P^{-1}$?

Exercise 8.8. Find all values $x \in \mathbb{R}$ for which

$$A(x) = \begin{pmatrix} 1 & x & 2 \\ x & 1 & x \\ 2 & 3 & 1 \end{pmatrix}$$

is singular, that is, not invertible.

Exercise 8.9. Do the same as Problem 5 for the matrix

$$B(x) = \begin{pmatrix} 1 & x & 1 & x \\ 1 & 0 & x & 1 \\ 0 & x & 1 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}.$$

(Suggestion: use the Laplace expansion to evaluate.)

Exercise 8.10. Suppose that Q is orthogonal. Find the possible values of $\det(Q)$.

Exercise 8.11. Which of the following statements are true. Give your reasoning.

(a) The determinant of a real symmetric matrix is always non negative.

(b) If A is any 2×3 real matrix, then $\det(AA^T) \geq 0$.

(c) If A is a square real matrix, then $\det(AA^T) \geq 0$.

Exercise 8.12. An $n \times n$ matrix A is called *skew symmetric* if $A^T = -A$. Show that if A is a skew symmetric $n \times n$ matrix and n is odd, then A cannot be invertible.

Exercise 8.13. A complex $n \times n$ matrix U is called *unitary* if $U^{-1} = \overline{U}^T$, where $\overline{U}$ is the matrix obtained by conjugating each entry of U. What are the possible values of the determinant of $\det(U)$ of a unitary matrix U.

Exercise 8.14. Compute

$$\begin{pmatrix} 1 & 2 & -1 & 0 \\ 2 & 1 & 1 & 1 \\ 0 & -1 & 2 & 0 \\ 1 & 1 & -1 & 1 \end{pmatrix}$$

in two cases: first when the field is $\mathbb{Q}$ and secondly when the field is $\mathbb{Z}_5$.

8.4 Geometric Applications of the Determinant

Now let us mention some of the geometric applications of determinants.

8.4.1 Cross and vector products

Recall that the cross product is expressed as a determinant with vector entries; namely,

$$\mathbf{x} \times \mathbf{y} = \det \begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{pmatrix}.$$

A frequently asked question is why the cross product exists only in three dimensions. In fact there is an n dimensional generalization of the cross product called the *vector product*. The vector product assigns to any $(n-1)$ vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-1} \in \mathbb{R}^n$ a vector

$$[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-1}] \in \mathbb{R}^n$$

orthogonal to each of the $\mathbf{x}_i$. It is defined by the follows determinantal expression:

$$[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-1}] = \det \begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \ldots & \mathbf{e}_n \\ x_{11} & x_{12} & \ldots & x_{1n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ x_{n-1,1} & x_{n-1,2} & \ldots & x_{n-1,n} \end{pmatrix}.$$

The fact that

$$\mathbf{x}_i \cdot [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-1}] = 0$$

for each i, $1 \leq i \leq n-1$, is simply due to the fact that a determinant with two equal rows is 0.

8.4.2 Determinants and volumes

There is an important geometric interpretation of determinants for matrices over $\mathbb{R}$. Consider a basis $\mathbf{w}_1, \ \ldots \ , \ \mathbf{w}_n$ of $\mathbb{R}^n$. Then $\mathbf{w}_1, \ \ldots \ , \ \mathbf{w}_n$ span an n-dimensional solid parallelogram $< \mathbf{w}_1, \ \ldots \ , \ \mathbf{w}_n >$. By definition

$$< \mathbf{w}_1, \ \ldots \ , \ \mathbf{w}_n >= \{\sum_{i=1}^{n} t_i \mathbf{w}_i \mid 0 \leq t_i \leq 1\}.$$

It can be shown that the volume of $< \mathbf{w}_1, \ \ldots \ , \ \mathbf{w}_n >$ is given by the formula

$$\mathbf{Vol}(< \mathbf{w}_1, \ \ldots \ , \ \mathbf{w}_n >) = |\det(\mathbf{w}_1 \ \mathbf{w}_2 \ \ldots \ \mathbf{w}_n)|.$$

Note that here the vertical bars denote the absolute value. To connect this with matrices, consider the linear transformation of $A : \mathbb{R}^n \to \mathbb{R}^n$ such that $A(\mathbf{e}_i) = \mathbf{w}_i$. In other words, the ith column of the matrix of A is $\mathbf{w}_i$. Thus $|\det(A)|$ is the volume of the image under A of the unit cube spanned by the standard basis vectors $\mathbf{e}, \ldots, \mathbf{e}_n$, i.e.

$$|\det(A)| = \mathbf{Vol}(< \mathbf{w}_1, \ \ldots \ , \ \mathbf{w}_n >).$$

Let us make a couple of comments about determinants and the geometry of linear transformations. The upshot of the above remarks is that the linear transformation associated to a real matrix having determinant of absolute value 1 preserves the volume of a cube. A linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ whose determinant is 1 is called *unimodular*. We say that the matrix of T is unimodular. The set of all unimodular real $n \times n$ matrices is denoted by $SL(n, \mathbb{R})$. It is called the special linear group.

Proposition 8.20. *Products and inverses of unimodular real matrices are also unimodular.*

Unimodular matrices have another property that is a little more subtle. For concreteness, let's assume for the moment that $n = 3$. If the matrix A of a linear transformation $T : \mathbb{R}^3 \to \mathbb{R}^3$ has positive determinant, then the linear transformation preserves right handed systems. Thus if $\det(T) > 0$, then T is orientation preserving.

If A has determinant -1, then A preserves volumes but reverses orientation. The possibility of having volume preserving orientation reversing transformations is what makes it necessary to put the absolute value in the change of variables formula below.

8.4.3 Change of variables formula

Now consider a smooth transformation $F : \mathbb{R}^n \to \mathbb{R}^n$ of the form

$$F(x_1, \ldots, x_n) = (f_1(x_1, \ldots, x_n), \ldots, f_n(x_1, \ldots, x_n)).$$

The *Jacobian* JF of F is defined to be the determinant of the differential of the transformation F. That is,

$$JF = \frac{\partial(f_1, \ldots, f_n)}{\partial(x_1, \ldots, x_n)} = \det \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}.$$

It is an easy exercise in partial differentiation that if F is linear and A is its associated matrix, then the Jacobian of F satisfies $JF = \det(A)$.

Now the change of variables formula says that if U is a closed n dimensional solid, then

$$\int_{F(U)} g(y)dV = \int_U g(F(x))|JF(x)|dV$$

for all functions g that are continuous on $F(U)$. The intuition for this is based on the observation that $|JF(x)|dV$ is a good approximation of the volume of the (curvilinear) solid parallelogram $F(U)$, where U denotes a small cube in $\mathbb{R}^n$ and $x \in U$ is arbitrary.

A specific application of this formula is that the volume of $F(U)$ is the integral of the absolute value of the Jacobian of F over U. If F is a linear transformation, then

$$\mathbf{Vol}(F(U)) = \det(A)\mathbf{Vol}(U).$$

Using this identity, it is not hard to give a geometric proof that if A, B are two $n \times n$ matrices over $\mathbb{R}$, then

$$|\det(AB)| = |\det(A)||\det(B)|,$$

where we have used the det notation to avoid confusion with absolute values.

8.4.4 Lewis Carroll's identity

The notion of the determinant of a matrix was introduced in the 19th century. At that time, many mathematicians were amateurs. For them, mathematics was a hobby, not a vocation. The determinant held great fascination in those days. It is a complicated concept, yet it is also a concrete object that can be calculated with and manipulated.

One 19th century mathematician who studied determinants was Charles Dodgson, a professor of mathematics at Oxford who is better known by his pseudonym, Lewis Carroll, and for authoring Alice in Wonderland. Dodgson discoverered and published of an amusing identity, now known as Lewis Carroll's Identity, which is reminiscent of the 2×2 case. Suppose A is an $n \times n$ matrix. Let A_C be the $(n-2) \times (n-2)$ submatrix in the middle of A obtained by deleting the first and last rows and the first and last columns. Also, let A_{NW} the $(n-1) \times (n-1)$ submatrix in the upper left hand corner of A, and define A_{NE}, A_{SW} and A_{SE} to be the $(n-1) \times (n-1)$ submatrices in the other three corners of A. If $n = 2$, put $\det(A_C) = 1$. Then Lewis Carroll's Identity says:

$$\det(A)\det(A_C) = \det(A_{NW})\det(A_{SE}) - \det(A_{NE})\det(A_{SW}) \qquad (8.13)$$

(see C.L. Dodgson, Proc. Royal Soc. London **17**, 555-560 (1860)). Interestingly, Lewis Carroll's Identity has recently reappeared in the modern setting of semi-simple Lie algebras.

Exercises

Exercise 8.15. Verify Lewis Carroll's Identity for the matrix

$$\begin{pmatrix} 1 & 2 & -1 & 0 \\ 2 & 1 & 1 & 1 \\ 0 & -1 & 2 & 0 \\ 1 & 1 & -1 & 1 \end{pmatrix}.$$

Exercise 8.16. Under what condition does Lewis Carroll's Identity make it possible to evaluate $\det(A)$?

Exercise 8.17. Suppose A is an $n \times n$ matrix over $\mathbb{R}$.

(a) First show that the Jacobian of the linear transformation defined by A is $\det(A)$.

(b) Use this to give a verification of the identity that says

$$|\det(AB)| = |\det(A)||\det(B)|.$$

Exercise 8.18. List all the 3×3 permutation matrices that lie in $SL(3, \mathbb{R})$.

Exercise 8.19. Prove Proposition 8.20.

Exercise 8.20. Prove that $SL(3, \mathbb{R})$ is a subgroup of $GL(3, \mathbb{R})$, where

$$GL(n, \mathbb{R}) = \{A \in \mathbb{R}^{n \times n} \mid \det(A) \neq 0\}.$$

Exercise 8.21. *. If G and H are groups, then a mapping $\varphi : G \to H$ is called a *group homomorphism* if for any $a, b \in G$, $\varphi(ab) = \varphi(a)\varphi(b)$. Explain how the determinant can be viewed as a group homomorphism if we choose the group G to be $GL(n, \mathbb{F})$, where $\mathbb{F}$ is any field.

8.5 Summary

The theory of determinants was one of the main topics of interest to the mathematicians working in the 19th century. The determinant of a square matrix over an arbitrary field $\mathbb{F}$ is an $\mathbb{F}$-valued function det that with the property that $\det(AB) = \det(A)\det(B)$, $\det(I_n) = 1$ and $\det(A) \neq 0$ if and only if A is invertible. In this book, it is defined combinatorially, and its properties are rigorously proved. In many texts, the determinant is defined inductively, by the Laplace expansion. The problem with this approach is that it is very messy to show that the definition is the same for all possible Laplace expansions. The determinant of a linear transformation $T : V \to V$ is defined as the determinant any matrix representing T.

In addition to its importance in algebra, which will be amply demonstrated in the next chapter, the determinant also has many important geometric applications. This stems from theh fact that $|\det(A)|$ is, by definition, the n-dimensional volume of the solid parallelogram spanned by $A\mathbf{e}_1, A\mathbf{e}_2, \ldots, A\mathbf{e}_n$. Thus the determinant appears in the change of variables theorem for multiple integrals, though in that context, it is called a Jacobian.

Chapter 9

Eigentheory

Let V be a finite dimensional vector space over a field $\mathbb{F}$, and let $T : V \to V$ be a linear transformation. In particular, if $V = \mathbb{F}^n$, then T is an element of $\mathbb{F}^{n\times n}$, i.e. an $n \times n$ matrix over $\mathbb{F}$. One of the most basic questions about T is for which non-zero vectors $\mathbf{v} \in V$ is it true that there exists a scalar $\lambda \in \mathbb{F}$ for which $T(\mathbf{v}) = \lambda\mathbf{v}$. A pair $(\lambda, \mathbf{v})$ for which this happens is called an *eigenpair* for T.

The purpose of this chapter is to study this eigenvalue problem. In particular, we will develop the tools for finding the eigenpairs for T and give examples of how they are used. The fundamental question is when does V have a basis consisting of eigenvectors. If such an *eigenbasis* $\mathcal{B}$ exists, then T is what we have called a semi-simple transformation. It turns out that if T is semi-simple, then the matrix $\mathcal{M}_{\mathcal{B}}^{\mathcal{B}}(T)$ of T is diagonal. For example, if $V = \mathbb{F}^n$, then T itself is a matrix, and an eigenbasis exists if and only if we can write $T = MDM^{-1}$ for an invertible $M \in \mathbb{F}^{n\times n}$. Finding M and D is the diagonalization problem, and determining when it has a solution is the main goals of this chapter (and subsequent chapters).

9.1 An Overview

The purpose of this section is to give a quick introduction to the eigentheory of matrices and linear transformations. We will give a much more complete treatment in the subsequent sections. The key concepts in eigentheory are the eigenvalues and eigenvectors of a linear transformation $T : V \to V$, where V is a finite dimensional vector space over $\mathbb{F}$. In this introduction, we will concentrate just on matrices, keeping in mind that a square matrix A over $\mathbb{F}$ is a linear transformation from $\mathbb{F}^n$ to itself. The eigenvalues and

eigenvectors enable us to understand how this linear transformation acts.

9.1.1 An Example: Dynamical Systems

One of the many applications of eigentheory is the decoupling of a dynamical system, which entails solving the above diagonalization problem.

Let us begin with a Fibonacci sequence. Recall that this is any sequence (a_k) in which a_0 and a_1 are arbitrary non-negative integers and $a_k = a_{k-1} + a_{k-2}$ if $k \geq 2$. We already noticed that the Fibonacci sequence leads to a matrix equation

$$\begin{pmatrix} a_{k+1} \\ a_k \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_k \\ a_{k-1} \end{pmatrix},$$

hence putting

$$F = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix},$$

we see by iterating that

$$\begin{pmatrix} a_{k+1} \\ a_k \end{pmatrix} = F \begin{pmatrix} a_k \\ a_{k-1} \end{pmatrix} = F^2 \begin{pmatrix} a_{k-1} \\ a_{k-2} \end{pmatrix} = \cdots = F^k \begin{pmatrix} a_1 \\ a_0 \end{pmatrix}.$$

Putting

$$\mathbf{v}_0 = \begin{pmatrix} a_1 \\ a_0 \end{pmatrix} \quad \text{and} \quad \mathbf{v}_k = \begin{pmatrix} a_{k+1} \\ a_k \end{pmatrix},$$

we can therefore express the Fibonacci sequence in the form $\mathbf{v}_k = F^k \mathbf{v}_0$.

The Fibonacci sequence is therefore an example of a *dynamical system*. Suppose in general that $A \in \mathbb{R}^{n \times n}$, and fix an arbitrary vector $\mathbf{v}_0 \in \mathbb{R}^n$. Then the dynamical system associated to A having initial value $\mathbf{v}_0$ is the sequence $(\mathbf{v}_k)$ with

$$\mathbf{v}_k = A\mathbf{v}_{k-1}, \quad k = 1, 2, \ldots$$

Thus

$$\mathbf{v}_1 = A\mathbf{v}_0, \ \mathbf{v}_2 = A\mathbf{v}_1 = A^2\mathbf{v}_0, \ldots, \ \mathbf{v}_k = A^k\mathbf{v}_0.$$

This sequence is easy to analyze if A is diagonal, say $A = \operatorname{diag}(d_1, d_2, \ldots, d_n)$. Indeed, A's Nth power A^N is just the diagonal matrix

$$A^N = \operatorname{diag}(d_1^N, d_2^N, \ldots, d_n^N).$$

Thus, if $\mathbf{v}_0 = (v_1, v_2, \ldots, v_n)$, then

$$\mathbf{v}_N = ((d_1)^N v_1, (d_2)^N v_2, \ldots, (d_n)^N v_n).$$

Fortunately, this isn't the only situation where we can compute A^N. Suppose $\mathbf{v} \in \mathbb{R}^n$ satisfies the condition $A\mathbf{v} = \lambda\mathbf{v}$ for some scalar $\lambda \in \mathbb{R}$. We will call such a pair $(\lambda, \mathbf{v})$ an *eigenpair* for A as long as $\mathbf{v} \neq \mathbf{0}$. Given an eigenpair, we see that

$$A^N\mathbf{v} = A^{N-1}(A\mathbf{v}) = A^{N-1}(\lambda\mathbf{v}) = \lambda A^{N-1}\mathbf{v}.$$

By iteration, we therefore see that $A^N\mathbf{v} = \lambda^N\mathbf{v}$, for all $N > 0$, so we know the dynamical system determined by A with initial value $\mathbf{v}$. Now suppose there exists a basis $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_n$ of $\mathbb{F}^n$ such that $A\mathbf{v}_i = \lambda_i\mathbf{v}_i$ for $1 \leq i \leq n$. Since we can expand an arbitrary $\mathbf{v} \in \mathbb{F}^n$ as $\mathbf{v} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n$, it follows that

$$A\mathbf{v} = \sum a_i A\mathbf{v}_i = \sum a_i\lambda\mathbf{v}_i.$$

Given an eigenbasis $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_n$ for A, we therefore get

$$A^N\mathbf{v} = \sum a_i A^N\mathbf{v}_i = \sum a_i\lambda^N\mathbf{v}_i. \tag{9.1}$$

Hence an eigenbasis for A is also one for A^N. The key to understanding a dynamical system with arbitrary initial value $\mathbf{v}_0$ is therefore to find an eigenbasis for A, expand $\mathbf{v}_0$ in terms of this basis and apply (9.1).

We'll finish this discussion below after we've said more about the eigenvalue problem.

We can interpret this as follows. let $T_A : \mathbb{F}^n \to \mathbb{F}^n$ be the linear transformation

9.1.2 The Eigenvalue Problem

Let A be a square matrix over the field $\mathbb{F}$, i.e. $A \in \mathbb{F}^{n\times n}$. We now want to consider the eigenvalue problem for A. That is, we want to find all $\lambda \in \mathbb{F}$ for which there exists a $\mathbf{v} \in \mathbb{F}^n$ such that $\mathbf{v} \neq \mathbf{0}$ and

$$A\mathbf{v} = \lambda\mathbf{v}.$$

Notice that the way this problem is imposed presents a difficulty: the variables being λ and the components of $\mathbf{v}$, the right hand side is nonlinear since λ multiplies the components of $\mathbf{v}$. Since we only know how to treat linear equations, we may have a problem. However, a slight tweak of the problem gives us the much better form

$$(A - \lambda I_n)\mathbf{v} = \mathbf{0}. \tag{9.2}$$

This is a homogeneous linear system. Moreover, it tells us that the eigenvalue problem actually consists breaks up into two parts:

(1) find those $\lambda \in \mathbb{F}$ such that the matrix $A - \lambda I_n$ has a nontrivial null space, and

(2) given λ satisfying (1), find $\mathcal{N}(A - \lambda I_n)$.

For the first part, we consider the equation

$$\det(A - \lambda I_n) = 0, \tag{9.3}$$

which is called the *characteristic equation* of A. Indeed, $\lambda \in \mathbb{F}$ belongs to an eigenpair $(\lambda, \mathbf{v})$ if and only if $A - \lambda I_n$ has rank less than n if and only if $\det(A - \lambda I_n) = 0$. As we will see below, this is the nonlinear part of the eigenvalue problem. Once we have a $\lambda \in \mathbb{F}$ satisfying (9.3), the second problem is a straightforward linear problem, as we only need to find $\mathcal{N}(A - \lambda I_n)$.

Example 9.1. Let's consider the real matrix

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

The eigenvalues of A are the real numbers λ such that

$$A - \lambda I_2 = \begin{pmatrix} 1-\lambda & 2 \\ 2 & 1-\lambda \end{pmatrix}$$

has rank 0 or 1. This happens exactly when $\det(A - \lambda I_2) = 0$. Now

$$\det(A - \lambda I_2) = (1-\lambda)^2 - 2 \cdot 2 = \lambda^2 - 2\lambda - 3 = 0.$$

Since $\lambda^2 - 2\lambda - 3 = (\lambda - 3)(\lambda + 1)$, the eigenvalues of A are 3 and -1. We can now proceed to finding corresponding eigenvectors. For this we need to find the null spaces $\mathcal{N}(A - 3I_2)$ and $\mathcal{N}(A + I_2)$. Clearly,

$$\mathcal{N}(A - 3I_2) = \mathcal{N}(\begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix}) = \mathbb{R} \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

and

$$\mathcal{N}(A + I_2) = \mathcal{N}(\begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}) = \mathbb{R} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

We now want to point out a consequence of this calculation. We can combine everything into a matrix equation

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 3 & -1 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix}.$$

This is an expression in the form $AP = PD$, where the columns of P are two independent eigenvectors. Thus P is invertible, and we get the factorization $A = PDP^{-1}$. At this point, we say A has been diagonalized.

Let's apply this to our dynamical system problem of taking powers of A. For example,

$$A^3 = (PDP^{-1})(PDP^{-1})(PDP^{-1}) = PDI_2DI_2DP^{-1} = PD^3P^{-1}.$$

Generalizing this to any positive integer N, we get the formula

$$A^N = PD^NP^{-1}.$$

9.1.3 Dynamical Systems Revisted

Recall that analyzing the dynamical system $A\mathbf{v}_k = \mathbf{v}_{k+1}$ required that we find the powers A^k of k. Whenever A can be diagonalized, that is expressed as PDP^{-1}, then A^k can easily be computed. For example,

$$A^3 = (PDP^{-1})(PDP^{-1})(PDP^{-1}) = PDI_nDI_nDP^{-1} = PD^3P^{-1}.$$

Generalizing this to any positive integer N, we get the formula

$$A^N = PD^NP^{-1}.$$

We can now finish our discussion of the Fibonacci sequence. Thus we have to solve the eigenvalue problem for F. Clearly, the characteristic equation of F is $\lambda^2 - \lambda + 1 = 0$. Using the quadratic formula, we obtain the roots

$$\tau = \frac{1+\sqrt{5}}{2}, \quad \mu = \frac{1-\sqrt{5}}{2},$$

and since both roots are real, F has two real eigenvalues. I leave it to you to check that $\mathcal{N}(F - \tau I_2) = \mathbb{R}(\tau, 1)^T$, and $\mathcal{N}(F - \mu I_2) = \mathbb{R}(\mu, 1)^T$. Therefore, as in the previous example,

$$F = \begin{pmatrix} \tau & \mu \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \tau & 0 \\ 0 & \mu \end{pmatrix} \begin{pmatrix} \tau & \mu \\ 1 & 1 \end{pmatrix}^{-1}.$$

Hence

$$\begin{pmatrix} a_{m+1} \\ a_m \end{pmatrix} = F^m \begin{pmatrix} a_1 \\ a_0 \end{pmatrix} = \begin{pmatrix} \tau & \mu \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \tau^m & 0 \\ 0 & \mu^m \end{pmatrix} \begin{pmatrix} \tau & \mu \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} a_1 \\ a_0 \end{pmatrix}.$$

To take a special case, let $a_0 = 0$ and $a_1 = 1$. Then this leaves us with the identity

$$a_m = \frac{\tau^m - \mu^m}{\tau - \mu} = \frac{1}{\sqrt{5} \cdot 2^m}((1+\sqrt{5})^m - (1-\sqrt{5})^m). \tag{9.4}$$

Taking the ratio a_{m+1}/a_m and letting m tend to ∞, we obtain

$$\lim_{m\to\infty} \frac{a_{m+1}}{a_m} = \lim_{m\to\infty} \frac{\tau^{m+1} - \mu^{m+1}}{\tau^m - \mu^m} = \tau,$$

since $\lim_{m\to\infty}(\mu/\tau)^m = 0$. Therefore, for large m, the ratio a_{m+1}/a_m is approximately τ. Since $-1 < \mu < 0$ (in fact $\mu \approx -0.618034$) and

$$a_m + \mu^m/\sqrt{5} = \tau^m/\sqrt{5},$$

it follows that

$$a_{2m} = \left[\frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{2m}\right] \quad \text{and} \quad a_{2m+1} = \left[\frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{2m+1}\right] + 1,$$

where $[k]$ denotes the integral part of k.

The eigenvalue τ is the so called golden number $\frac{1+\sqrt{5}}{2}$ which was known to the early Greeks and, in fact, used in their architecture. It was also encountered in the discussion of the icosahedron and Buckminsterfullerene. There is an interesting observation in botany, namely that certain observed ratios in plant growth are approximated by quotients of Fibonacci numbers. For example, on some types of pear trees, every eight consecutive leaves make three turns around the stem. (Think of a spiral staircase making three complete turns around its axis that has eight steps and you have the leaves on the stem of a pear tree.) There are more examples and references of this is the book Geometry, by H.M.S. Coxeter. The implication is that the golden number may have some properties that influence biological patterns, just as it has properties that affect molecular structures such as buckminsterfullerene.

In the above two examples, we solved the diagonalization problem. That is, given A we constructed a matrix P such that $A = PDP^{-1}$.

Definition 9.1. Two matrices A and B in $\mathbb{F}^{n\times n}$ are said to be *similar* if there exists an invertible $M \in \mathbb{F}^{n\times n}$ so that $B = MAM^{-1}$.

Thus the diagonalization problem for A is to find a diagonal matrix similar to A.

Now suppose that V is an arbitrary vector space over $\mathbb{R}$ and $T : V \to V$ is a linear transformation. In the infinite dimensional setting, it is customary to call a linear transformation a *linear operator*. Then the eigenvalue problem for T is still the same: to find scalars $\lambda \in \mathbb{R}$ so that there exists a non zero $\mathbf{v} \in V$ such that $T(\mathbf{v}) = \lambda\mathbf{v}$. However, since V is not assumed to be finite dimensional, the above method involving the characteristic equation won't work. One way to proceed is to try to find finite dimensional subspaces W of V so that $A(W) \subset W$. But in the infinite dimensional setting, there is no such simple technique for finding eigenvalues of T. This has lead to the development of many different techniques. In the next example, we consider a case which arises in differential equations.

Example 9.2. Let $V = C^\infty(\mathbb{R})$ be the space of real valued functions on $\mathbb{R}$ which have derivatives of all orders. Since the derivative of such a function also has derivatives of all orders, differentiation defines a linear operator $D : C^\infty(\mathbb{R}) \to C^\infty(\mathbb{R})$. That is, $D(f) = f'$. It is clear that the exponential function $f(x) = e^{rx}$ is an eigenvector of D with corresponding eigenvalue r. Thus (r, e^{rx}) form an eigenpair for D. In this context, eigenvectors are usually called *eigenfunctions*. Considering D^2 instead of D, we easily see that for any integers m and n, $\cos mx$ and $\sin nx$ are also eigenfunctions with corresponding eigenvalues $-m^2$ and $-n^2$ respectively.

We will return to this example after the Principal Axis Theorem.

9.2 The Characteristic Polynomial

Having now seen some of the basic definitions and some applications and examples, we will next consider eigentheory in greater detail.

9.2.1 Basic Definitions and Properties

Let $\mathbb{F}$ be a field, and suppose V is a finite dimensional vector space over $\mathbb{F}$.

Definition 9.2. Suppose $T : V \to V$ is a linear map. Then a pair $(\lambda, \mathbf{v})$, where $\lambda \in \mathbb{F}$ and $\mathbf{v} \in V$, is called an *eigenpair* for T if $\mathbf{v} \neq \mathbf{0}$ and

$$T(\mathbf{v}) = \lambda\mathbf{v}. \tag{9.5}$$

If $(\lambda, \mathbf{v})$ is an eigenpair for T, we call λ an $\mathbb{F}$-*eigenvalue*, or, simply, an *eigenvalue* of T and $\mathbf{v}$ an *eigenvector* of T corresponding to λ.

As we discussed above, the fundamental question is whether T has an eigenbasis.

It turns out that it is most convenient to first treat the eigenvalue problem for matrices. Thus we will consider matrices first and return to linear transformations after that. Here are a couple of observations that follow directly form the definitions.

First, let us make some simple observations. Not every square matrix has an eigenvalue.

Example 9.3. The characteristic equation of $J = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ is $\lambda^2 + 1 = 0$. Since the roots of $\lambda^2 + 1 = 0$ are $\pm i$, there are no *real* eigenvalues. This isn't surprising since J is the matrix of the rotation $R_{\pi/2} : \mathbb{R}^2 \to \mathbb{R}^2$ of $\mathbb{R}^2$ through $\pi/2$, and thus there are no eigenvectors in $\mathbb{R}^2$. On the other hand, if we think of J as a complex matrix (as we may since $\mathbb{R} \subset \mathbb{C}$), the eigenvalues of J are $\pm i$. Solving the for corresponding eigenvectors gives eigenpairs $(i, (-1, i)^T)$ and $(-i, (1, i)^T)$. Thus

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix} \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}.$$

Hence $JM = MD$ or $J = MDM^{-1}$.

This Example brings up a somewhat subtle point. When considering the eigenvalue problem for an $A \in \mathbb{R}^{n \times n}$, we know, by the Fundamental Theorem of Algebra (Theorem 4.11), that the characteristic equation of A has n complex roots. But only the real roots determine eigenpairs for A when we are taking $\mathbb{F} = \mathbb{R}$. However, as $\mathbb{R}$ is a subfield of $\mathbb{C}$, so we can also consider A as a matrix in $\mathbb{C}^{n \times n}$, and thus every root μ of the characteristic equation of A gives us an eigenpair $(\mu, \mathbf{v})$ with $\mathbf{v} \in \mathbb{C}^n$.

Let us now give some of the basic general properties of eigenpairs.

Proposition 9.1. *Suppose A is a square matrix over $\mathbb{F}$ and $(\lambda, \mathbf{v})$ is an eigenpair for A. Then for any scalar $r \in \mathbb{F}$, $(r\lambda, \mathbf{v})$ is an eigenpair for rA. Moreover, for any positive integer k, $(\lambda^k, \mathbf{v})$ is an eigenpair for A^k. Finally, A has an eigenpair of the form $(0, \mathbf{v})$ if and only if $\mathcal{N}(A)$ is nontrivial.*

Proof. The proof is left as an exercise. □

Recall that we defined the *characteristic equation* of $A \in \mathbb{F}^{n \times n}$ in the previous section to be $\det(A - \lambda I_n) = 0$. We will frequently denote the determinant $\det(A - \lambda I_n)$ by $|A - \lambda I_n|$.

Proposition 9.2. *If $A \in \mathbb{F}^{n\times n}$, then $\det(A-\lambda I_n)$ is a polynomial in λ over $\mathbb{F}$ of degree n. Its leading term is $(-1)^n\lambda^n$ and its constant term is $\det(A)$. The eigenvalues of A are the roots of $\det(A-\lambda I_n)=0$ in $\mathbb{F}$.*

Proof. This is obvious from the definition of the determinant. □

Definition 9.3. Given $A \in \mathbb{F}^{n\times n}$, we call $p_A(\lambda) = \det(A-\lambda I_n)$ the *characteristic polynomial* of A.

The next Proposition gives an important property of the characteristic polynomial .

Proposition 9.3. *Two similar matrices have the same characteristic polynomial.*

Proof. Suppose A and B are similar, say $B = MAM^{-1}$. Then

$$\begin{aligned}\det(B-\lambda I_n) &= \det(MAM^{-1}-\lambda I_n)\\ &= \det(M(A-\lambda I_n)M^{-1})\\ &= \det(M)\det(A-\lambda I_n)\det(M^{-1})\end{aligned}$$

Since $\det(M^{-1}) = \det(M)^{-1}$, the proof is done. □

We can now extend these definitions to linear transformations.

Definition 9.4. If V is a finite dimensional vector space over $\mathbb{F}$ and $T : V \to V$ is linear, then we define the *characteristic equation* of T to be the characteristic equation $p_A(\lambda) = 0$ of any matrix $A \in \mathbb{F}^{n\times n}$ which represents T. Similarly, we call $p_A(\lambda)$ the *characteristic polynomial* of T.

Since any two matrices $\mathcal{M}_{\mathcal{B}}^{\mathcal{B}}(T)$ representing T are similar, Proposition 9.2 tells us that the the characteristic equation and characteristic polynomial of T are well defined. Now we need to show that an eigenpair for $\mathcal{M}_{\mathcal{B}}^{\mathcal{B}}(T)$ gives an eigenpair for T, and conversely.

Proposition 9.4. *Let V be a finite dimensional vector space over $\mathbb{F}$ and suppose $T : V \to V$ is linear. Then any root $\mu \in \mathbb{F}$ of the characteristic equation of T is an eigenvalue of T, and conversely.*

Proof. Let A be the matrix representing T for a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of V. Suppose $\mu \in \mathbb{F}$ is a root of the characteristic equation $\det(A-\mu I_n) = 0$, and let

$(\mu, \mathbf{x})$ be an eigenpair for A. Thus $A\mathbf{x} = \mu\mathbf{x}$. Let $\mathbf{x} = (x_1, \ldots, x_n)^T$ and put $\mathbf{v} = \sum_i x_i \mathbf{v}_i$. Then

$$\begin{aligned} T(\mathbf{v}) &= \sum_i x_i T(\mathbf{v}_i) \\ &= \sum_{i,j} x_i a_{ji} \mathbf{v}_j \\ &= \sum_j \mu x_j \mathbf{v}_j \\ &= \mu\mathbf{v} \end{aligned}$$

Since some $x_j \neq 0$, it follows that $\mathbf{v}$ is non zero, so $(\mu, \mathbf{v})$ is an eigenpair for T. For the converse, just reverse the argument. □

The previous two Propositionshave the consequence that from now on, we can concentrate on the eigentheory of $n \times n$ matrices and ignore the more viewpoint of linear maps. Put in other terms, the eigenvalue problem for linear transformations on a finite dimensional vector space reduces completely to the case of matrices.

Let's now consider some examples.

Example 9.4. If $A = \left(\begin{smallmatrix} 1 & 2 \\ 2 & -1 \end{smallmatrix}\right)$, then $A - \lambda I_2 = \left(\begin{smallmatrix} 1-\lambda & 2 \\ -2 & -1-\lambda \end{smallmatrix}\right)$, so the characteristic polynomial of A is $|A - \lambda I_2| = (1-\lambda)(-1-\lambda) - (2)(2) = \lambda^2 - 5$. The eigenvalues of A are $\pm\sqrt{5}$. Both eigenvalues are real.

Example 9.5. As we saw above, the matrix $J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ has no real eigenvalues. However, it does have two distinct eigenvalues in $\mathbb{C}$.

Example 9.6. Let $K = \left(\begin{smallmatrix} 0 & -i \\ i & 0 \end{smallmatrix}\right)$. The characteristic polynomial of K is $\lambda^2 - 1$, so the eigenvalues of K are ± 1. Thus K is a complex matrix with real eigenvalues. Notice that $K = iJ$, so Proposition 9.1 in fact tells us a priori that its eigenvalues are i times those of J.

9.2.2 Formulas for the Characteristic Polynomial

The characteristic polynomial $p_A(\lambda)$ of a 2×2 matrix $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ over an arbitrary field $\mathbb{F}$ has a nice form, which we will generalize. First, define the *trace* of A to be the sum of the diagonal elements of A, i.e. $\mathrm{Tr}(A) = a + d$. Then

$$p_A(\lambda) = (a-\lambda)(d-\lambda) - bc = \lambda^2 - (a+d)\lambda + (ad - bc).$$

Hence,

$$p_A(\lambda) = \lambda^2 + \mathrm{T}r(A)\lambda + \det(A). \tag{9.6}$$

The quadratic formula gives the eigenvalues of A in the form

$$\lambda = \frac{1}{2}\big(-(a+d) \pm \sqrt{(a+d)^2 - 4(ad-bc)}\big)$$

which can be rewritten as

$$\lambda = \frac{1}{2}\big(-\mathrm{Tr}(A) \pm \sqrt{\mathrm{Tr}(A)^2 - 4\det(A)}\big). \tag{9.7}$$

Hence if A is real, it has real eigenvalues if and only if the discriminant $\Delta(A) :=\mathrm{Tr}(A)^2 - 4\det(A)$ is non negative: $\Delta(A) \geq 0$. If $\Delta(A) = 0$, the roots are real but repeated. If $\Delta(A) < 0$, the roots are complex and unequal. In this case, the roots are conjugate complex numbers. By factoring the characteristic polynomial as

$$(\lambda - \lambda_1)(\lambda - \lambda_2) = \lambda^2 - (\lambda_1 + \lambda_2)\lambda + \lambda_1\lambda_2$$

and comparing coefficients, we immediately see that:

(i) the trace of A is the sum of the eigenvalues of A:

$$\mathrm{T}r(A) = \lambda_1 + \lambda_2,$$

(ii) the determinant of A is the product of the eigenvalues of A:

$$\det(A) = \lambda_1\lambda_2.$$

For $n > 2$, the characteristic polynomial is more difficult to compute. Using row operations to compute a characteristic polynomial isn't very practical (see Example 8.8), so when computing by hand, it is almost necessary to also use the Laplace expansion. There is an important warning that needs to be issued here. **Whenever you are computing the characteristic polynomial of a matrix A, never, repeat, never row reduce A or partly row reduce A before computing the characteristic polynomial**. There is absolutely no reason the characteristic polynomials of A and EA should have any common roots. If they did all invertible matrices would have the same eigenvalues.

On the other hand, there is a beautiful formula for the characteristic polynomial involving the so called *principal minors* of A. Since $p_A(\lambda)$ is a

polynomial in λ of degree n with leading coefficient $(-1)^n\lambda^n$ and constant term $\det(A)$, we can write

$$p_A(\lambda) = (-1)^n\lambda^n + (-1)^{n-1}\sigma_1(A)\lambda^{n-1} + (-1)^{n-2}\sigma_2(A)\lambda^{n-2} +$$

$$+ \cdots + (-1)^1\sigma_{n-1}(A)\lambda + \det(A), \tag{9.8}$$

where the $\sigma_i(A)$, $1 \leq i \leq n-1$, are the remaining coefficients.

Theorem 9.5. *The coefficients $\sigma_i(A)$ for $1 \leq i \leq n$ are given by*

$$\sigma_i(A) := \sum \big(\textit{all principal } i \times i \textit{ minors of } A\big), \tag{9.9}$$

where the principal $i \times i$ minors of A are defined to be the determinants of the $i \times i$ submatrices of A obtained by deleting $n-i$ rows of A and then the same $n-i$ columns.

We will omit the proof.

By definition, the principal 1×1 minors are just the diagonal entries of A, since deleting $(n-1)$ of the rows and the same $(n-1)$ columns just leaves the diagonal entry a_{ii} in the unique row that wasn't deleted. Hence

$$\sigma_1(A) = a_{11} + a_{22} + \cdots + a_{nn}$$

so

$$\sigma_1(A) = \mathrm{Tr}(A).$$

Of course, formula (9.8) says that the constant term is the determinant of A, i.e. $\sigma_n(A) = \det(A)$. But this is clear. In general, the number of $j \times j$ minors of A is the binomial coefficient $\binom{n}{j} = \frac{n!}{j!(n-j)!}$. Thus, the characteristic polynomial of a 4×4 matrix will involve four 1×1 principal minors, six 2×2 principal minors, four 3×3 principal minors and a single 4×4 principal minor. But using Theorem 9.5 is still much simpler than expanding d$et(A - \lambda I_n)$ via Laplace.

Example 9.7. For example, let

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

Then

$$\det(A - \lambda I_3) = -\lambda^3 + (-1)^2(3 - 1 + 0)\lambda^2 +$$

$$(-1)^1(\det\begin{pmatrix}3 & -2\\ 3 & -1\end{pmatrix} + \det\begin{pmatrix}-1 & -3\\ -2 & 0\end{pmatrix} + \det\begin{pmatrix}3 & -2\\ 1 & 0\end{pmatrix})\lambda + \det(A).$$

Thus the characteristic polynomial of A is

$$p_A(\lambda) = -\lambda^3 + 2\lambda^2 + \lambda - 2.$$

A natural question is how to find the roots of the characteristic polynomial. Except in the 2×2, 3×3 and 4×4 cases, where there are general formulas, there aren't any general methods for finding the roots of a polynomial. Solving the eigenvalue problem for a given square matrix is a problem which is usually approached by other methods, such as tNewton's method or the QR algorithm, which we will discuss later.

In particular examples, the **rational root test** can be helpful, since most examples deal with matrices with integer entries. Hence the characteristic polynomial has integer coefficients. The rational root test treats such polynomials. It says that if

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

is a polynomial with integer coefficients $a_0, \ldots a_n$, then the only possible rational roots have the form p/q, where p and q are integers without any common factors, p divides a_0 and q divides a_n. In particular, if the leading coefficient $a_n = 1$, then $q = \pm 1$, so the only possible rational roots are the inegers which divide the constant term a_0. Therefore we obtain the following

Proposition 9.6. *If A is a matrix with integer entries, then the only possible rational eigenvalues are the integers which divide* $\det(A)$.

Since the characteristic polynomial of the matrix A in the previous example is $-\lambda^3 + 2\lambda^2 + \lambda - 2$ the only possible rational eigenvalues are the divisors of 2, that is ± 1 and ± 2. Checking these possibilities, we find that ± 1 and 2 are roots, so these are the eigenvalues of A.

Note also that the coefficients of $p_A(\lambda)$ are certain explicit functions of its roots. For if $p_A(\lambda)$ has roots $\lambda_1, \ldots, \lambda_n$, then

$$\begin{aligned} p_A(\lambda) &= (\lambda_1 - \lambda)(\lambda_1 - \lambda)\ldots(\lambda_1 - \lambda) \\ &= (-1)^n(\lambda)^n + (-1)^{n-1}(\lambda_1 + \lambda_2 + \ldots \lambda_n)\lambda^{n-1} + \cdots + \lambda_1\lambda_2\cdots\lambda_n \end{aligned}$$

Thus we obtain a generalization of what we showed in the $2\times$ case.

Proposition 9.7. *The trace of a matrix A is the sum of the roots of its characteristic polynomial, and similarly, the determinant is the product of the roots of its characteristic polynomial.*

There are other functions $\sigma_i(\lambda_1, \ldots, \lambda_n)$ expressing the coefficients $\sigma_i(A)$ as functions of $\lambda_1, \ldots, \lambda_n$. These are called the elementary symmetric functions. You might be amused to find some expressions for them on your own. For example,

$$\sigma_2(\lambda_1, \ldots, \lambda_n) = \sigma_2(A) = \sum_{i<j} \lambda_i \lambda_j.$$

Exercises

In the following exercises, A and B are assumed to be square matrices of the same size over

either $\mathbb{R}$ or $\mathbb{C}$. Recall that an eigenbasis for a real $n \times n$ matrix A is a basis of $\mathbb{R}^n$ consisting of eigenvectors.

Exercise 9.1. Find the characteristic polynomial, eigenvalues and if possible, an eigenbasis for:

(i) the X-files matrix

$$X = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix},$$

(ii) the checkerboard matrix

$$C = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

(iii) the 4×4 X-files matrix

$$\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix},$$

(iv) the 4×4 checkerboard matrix

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Exercise 9.2. Find the characteristic polynomial and eigenvalues of

$$\begin{pmatrix} -3 & 0 & -4 & -4 \\ 0 & 2 & 1 & 1 \\ 4 & 0 & 5 & 4 \\ -4 & 0 & -4 & -3 \end{pmatrix}$$

in two ways, one using the Laplace expansion and the other using principal minors.

Exercise 9.3. The following matrix A was on a blackboard in the movie Good Will Hunting:

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 0 & 2 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}.$$

Find the characteristic polynomial of A and try to decide how many real eigenvalues there are.

Exercise 9.4. Find the characteristic polynomial of a 4×4 matrix A if you know that three eigenvalues of A are ± 1 and 2 and you also know that $\det(A) = 6$.

Exercise 9.5. Using only the definitions, prove Proposition 9.1.

Exercise 9.6. Suppose $A \in \mathbb{F}^{n \times n}$ has the property that $A = A^{-1}$. Show that if λ is an eigenvalue of A, then so is λ^{-1}.

Exercise 9.7. Show that two similar matrices have the same trace and determinant.

Exercise 9.8. True or False: If two matrices have the same characteristic polynomial , they are similar.

Exercise 9.9. If A is a square matrix, determine whether or not A and A^T have the same characteristic polynomial, hence the same eigenvalues.

Exercise 9.10. Show that 0 is an eigenvalue of A if and only if A is singular, that is, A^{-1} does not exist.

Exercise 9.11. True or False: If λ is an eigenvalue of A and μ is an eigenvalue of B, then $\lambda + \mu$ is an eigenvalue of $A + B$.

Exercise 9.12. An $n \times n$ matrix such that $A^k = O$ for some positive integer k is called *nilpotent*.

(a) Show all eigenvalues of a nilpotent matrix A are 0.

(b) Hence conclude that the characteristic polynomial of A is $(-1)^n \lambda^n$. In particular, the trace of a nilpotent matrix is 0.

(c) Find a 3×3 matrix A so that $A^2 \neq O$, but $A^3 = O$. (Hint: look for an upper triangular example.)

Exercise 9.13. Find the characteristic polynomial of the X-Files matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Exercise 9.14. Show that the complex eigenvalues of a real $n \times n$ matrix occur in conjugate pairs λ and $\overline{\lambda}$. (Note: the proof of this we gave for $n = 2$ does not extend. First observe that if $p(x)$ is a polynomial with real coefficients, then $\overline{p(x)} = p(\overline{x})$.)

Exercise 9.15. Conclude from the previous exercise that a real $n \times n$ matrix, where n is odd, has at least one real eigenvalue. In particular, every 3×3 real matrix has a real eigenvalue.

Exercise 9.16. Find eigenpairs for the two eigenvalues of the rotation R_θ of $\mathbb{R}^2$. (Note, the eigenvalues are complex.)

Exercise 9.17. Show that in general, the only possible real eigenvalues of an $n \times n$ real orthogonal matrix are ± 1.

Exercise 9.18. Suppose A is $n \times n$ and invertible. Show that for any $n \times n$ matrix B, AB and BA have the same characteristic polynomial.

Exercise 9.19. * Find the characteristic polynomial of

$$\begin{pmatrix} a & b & c \\ b & c & a \\ c & a & b \end{pmatrix},$$

where a, b, c are all real. (Note that the second matrix in Problem 2 is of this type. What does the fact that the trace is an eigenvalue say?)

Exercise 9.20. Find the elementary symmetric functions $\sigma_i(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ for $i = 1, 2, 3, 4$ by expanding $(x - \lambda_1)(x - \lambda_2)(x - \lambda_3)(x - \lambda_4)$. Deduce an expression for all $\sigma_i(A)$ for an arbitrary 4×4 matrix A.

Exercise 9.21. Show directly that

$$\frac{\tau^m - \mu^m}{\tau - \mu} = \frac{1}{\sqrt{5} \cdot 2^m}((1 + \sqrt{5})^m - (1 - \sqrt{5})^m)$$

is an integer, thus explaining the strange expression in Section 9.1.

9.3 Eigenvectors and Diagonalizability

We now begin our study of the question of when an eigenbasisexists.

9.3.1 Eigenspaces

Let $A \in \mathbb{F}^{n\times n}$ and suppose $\lambda \in \mathbb{F}$ is an eigenvalue.

Definition 9.5. The null space $\mathcal{N}(A - \lambda I_n)$ is called the *eigenspace* of A corresponding to λ. Similarly, if $T : V \to V$ is linear and λ is an eigenvalue of T, then the *eigenspace* of T corresponding to λ is the subspace

$$E_\lambda = \{\mathbf{v} \in V \mid T(\mathbf{v}) = \lambda\mathbf{v}\}.$$

Example 9.8. Consider a simple example, say

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

Then $p_A(\lambda) = \lambda^2 - 2\lambda - 3 = (\lambda - 3)(\lambda + 1)$, so the eigenvalues of A are $\lambda = 3, -1$. Now

$$A - (-1)I_2 = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix} \quad \text{and} \quad A - 3I_2 = \begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix}.$$

Thus $E_{-1} = \mathcal{N}(A + I_2) = \mathbb{R}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $E_3 = \mathcal{N}(A - 3I_2) = \mathbb{R}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Notice that we have found an eigenbasis of $\mathbb{R}^2$.

Definition 9.6. Let $A \in \mathbb{F}^{n\times n}$. Then we say A is *diagonalizable over* $\mathbb{F}$, or simply *diagonalizable* if there is no danger of confusion, if and only if there exist eigenpairs $(\lambda_1, \mathbf{w}_1), \ldots, (\lambda_n, \mathbf{w}_n)$ so that $\mathbf{w}_1, \ \ldots \ , \ \mathbf{w}_n$ are a basis of $\mathbb{F}^n$.

Note that in this definition, some eigenvalues may appear more than once. The next proposition describes the set of diagonalizable matrices.

Proposition 9.8. *Assume A is an $n \times n$ matrix over $\mathbb{F}$ which is diagonalizable, and let $\mathbf{w}_1, \ \ldots \ , \ \mathbf{w}_n \in \mathbb{F}^n$ be an eigenbasis with eigenvalues $\lambda_1, \ \ldots \ , \ \lambda_n$ in $\mathbb{F}$. Then $A = PDP^{-1}$, where $P = (\mathbf{w}_1 \ \ldots \ \mathbf{w}_n)$ and $D = \operatorname{diag}(\lambda_1, \ \ldots \ , \ \lambda_n)$. Conversely, if $A = PDP^{-1}$, where $P \in \mathbb{F}^{n\times n}$ is nonsingular and $D = \operatorname{diag}(\lambda_1, \ \ldots \ , \ \lambda_n)$, then the columns of P are an eigenbasis of $\mathbb{F}^n$ for A and the diagonal entries of D are the corresponding eigenvalues. That is, if the ith column of P is $\mathbf{w}_i$, then $(\lambda_i, \mathbf{w}_i)$ is an eigenpair for A.*

Proof. Suppose an eigenbasis$\mathbf{w}_1 \ \dots \ \mathbf{w}_n$ is given and $P = (\mathbf{w}_1 \ \dots \ \mathbf{w}_n)$. Then

$$AP = (\lambda_1\mathbf{w}_1 \ \dots \ \lambda_n\mathbf{w}_n = (\mathbf{w}_1 \ \dots \ \mathbf{w}_n)\text{diag}(\lambda_1, \ \dots \ , \ \lambda_n).$$

Since the columns of P are an eigenbasis, P is invertible, hence A is diagonalizable. The converse is proved in the same way. □

Example 9.9. Here is another calculation. Let

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

Then $p_A(\lambda) = -\lambda^3+2\lambda^2+\lambda-2$. The eigenvalues of A are $\pm 1, \ 2$. Thus $\mathcal{N}(A-I_3) = \mathbb{R}(1,0,1)^T$, $\mathcal{N}(A+I_3) = \mathbb{R}(1,1,1)^T$, and $\mathcal{N}(A-2I_3) = \mathbb{R}(0,1,-1)^T$, where $\mathbb{R}\mathbf{v}$ is the line spanned by $\mathbf{v}$. Hence an eigenbasis exists and A is diagonalizable.

You may have noticed that in all the above examples of diagonalizable matrices, the matrices have distinct eigenvalues. No eigenvalues are repeated. Matrices with this property are always diagonalizable.

Proposition 9.9. *An $n \times n$ matrix A over $\mathbb{F}$ with n distinct eigenvalues in $\mathbb{F}$ is diagonalizable. More generally, if V be a finite dimensional vector space over $\mathbb{F}$ and $T : V \to V$ is a linear transformation with distinct eigenvalues, then V admits an eigenbasis for T.*

The proof entails showing that if $\mathbf{w}_1, \ \dots \ , \ \mathbf{w}_n$ are eigenvectors for A (or T) corresponding to the n distinct eigenvalues, then $\mathbf{w}_1, \ \dots \ , \ \mathbf{w}_n$ are linearly independent. This assures us that $P = (\mathbf{w}_1 \ \dots \mathbf{w}_n)$ is invertible. The idea for $n = 2$ is that since the eigenvalues are distinct, corresponding eigenvectors cannot be proportional. This idea can be extended to the $n \times n$ case, but we won't give the proof since we will prove a more general fact in the next section.

Consider another example.

Example 9.10. The counting matrix

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

has characteristic polynomial $p_C(x) = -\lambda^3 + 15\lambda^2 - 21\lambda$, hence eigenvalues 0 and $\frac{1}{2}(15 \pm \sqrt{151})$. The eigenvalues of C are real and distinct, hence C is diagonalizable over $\mathbb{R}$.

A simple way of determining whether A has distinct eigenvaluesis to apply the repeated root test. A polynomial $p(x)$ has a double root if and only if $p(x)$ and $p'(x)$ have a common root. Thus we get

Proposition 9.10. *A square matrix A has non repeated eigenvalues exactly when $p_A(\lambda)$ and $p'_A(\lambda)$ have no common roots.*

Example 9.11. By the previous example, the characteristic polynomial of the counting matrix C is $p_C(\lambda) = -\lambda^3 + 15\lambda^2 - 21\lambda$. Since $p'_C(\lambda) = -3(\lambda^2 - 10\lambda + 21)$ has roots $\lambda = 3, 7$, $p_C(\lambda)$ has simple roots since neither 3 nor 7 is one of its roots. Thus, C has distinct eigenvalues.

Exercises

Exercise 9.22. Diagonalize the following matrices if possible:

$$A = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} -3 & 0 & -4 & -4 \\ 0 & 2 & 1 & 1 \\ 4 & 0 & 5 & 4 \\ -4 & 0 & -4 & -3 \end{pmatrix}.$$

Exercise 9.23. Decide whether the Good Will Hunting matrix (cf Exercise 9.3) can be diagonalized.

Exercise 9.24. Suppose A and B are similar and $\mathbf{v}$ is an eigenvector of A. Find an eigenvector of B.

Exercise 9.25. Let A be a real 3×3 matrix so that A and $-A$ are similar. Show that

(a) $\det(A) = \mathrm{Tr}(A) = 0$,

(b) 0 is an eigenvalue of A, and

(c) if some eigenvalue of A is non-zero, then A is diagonalizable over $\mathbb{C}$.

Exercise 9.26. Find an example of two real matrices which have the same characteristic polynomial but which are not similar.

Exercise 9.27. A 4×4 matrix has eigenvalues ± 1, trace 3 and determinant 0. Can A be diagonalized?

Exercise 9.28. Let A be a 3×3 matrix whose characteristic polynomial has the form $-x^3 + 7x^2 - bx + 8$. Suppose that the eigenvalues of A are integers.

(i) Find the eigenvalues of A.

(ii) Find the value of b.

Exercise 9.29. What is the characteristic polynomial of A^3 in terms of that of A?

Exercise 9.30. Diagonalize $J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. More generally do the same for R_θ for all $\theta \neq 0$.

Exercise 9.31. Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, where a, b, c, d are all positive real numbers. Show that A is diagonalizable.

Exercise 9.32. Suppose A is the matrix in Exercise 9.31. Show that A has an eigenvector in the first quadrant and another in the third quadrant.

Exercise 9.33. We say that two $n \times n$ matrices A and B are simultaneously diagonalizable if they are diagonalized by the same matrix M. Show that two simultaneously diagonalizable matrices A and B commute. That is, show that $AB = BA$.

Exercise 9.34. This is the converse to Exercise 9.33. Suppose that two $n \times n$ matrices A and B commute. Show that if both A and B are diagonalizable, then they are simultaneously diagonalizable. That is, they share a common eigenbasis.

9.4 Is Every Matrix Diagonalizable?

9.4.1 A Sufficient Condition

In the last section, we stated a result which says that every matrix with distinct eigenvalues is diagonalizable. In particular, if A is a real $n \times n$ matrix with n distinct real eigenvalues, then there is an eigenbasis of $\mathbb{R}^n$ for A. If A has some complex entries or some complex eigenvalues, then we are guaranteed an eigenbasis of $\mathbb{C}^n$. Thus, to answer to the question in the title, we have to see what happens when A has repeated eigenvalues. Before considering that question, we will give a characterization of the matrices are diagonalizable in terms of their eigenspaces. First, let us prove the generalization of Proposition 9.9 mentioned above.

Proposition 9.11. *Suppose $A \in \mathbb{F}^{n\times n}$, and $\lambda_1, \ldots, \lambda_k \in \mathbb{F}$ are distinct eigenvalues of A. Choose an eigenpair $(\lambda_i, \mathbf{w}_i)$ for each λ_i. Then the vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$ are linearly independent. Moreover, if we choose a set of linearly independent eigenvectors in E_{λ_i} for each λ_i, $1 \leq i \leq k$, then the union of these k linearly independent sets of eigenvectors is linearly independent.*

Proof. Let W be the subspace of $\mathbb{F}^n$ spanned by $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$. We already know from Chapter 5 that some subset of $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$ gives a basis of W. Hence, suppose, after renumbering that $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m$ is this basis, where $m \leq k$. If $m < k$, then we may write

$$\mathbf{w}_{m+1} = \sum_{i=1}^{m} a_i \mathbf{w}_i. \tag{9.10}$$

Applying A, we obtain $A(\mathbf{w}_{m+1}) = \sum_{i=1}^{m} a_i A(\mathbf{w}_i)$, so

$$\lambda_{m+1}\mathbf{w}_{m+1} = \sum_{i=1}^{m} a_i \lambda_i \mathbf{w}_i. \tag{9.11}$$

Multiplying (9.10) by λ_{m+1} and subtracting (9.11), we obtain

$$\sum_{i=1}^{m} (\lambda_{m+1} - \lambda_i) a_i \mathbf{w}_i = \mathbf{0}.$$

But $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m$ are independent, so $(\lambda_{m+1} - \lambda_i)a_i = 0$ provided $1 \leq i \leq m$. Since each $(\lambda_{m+1} - \lambda_i) \neq 0$, all $a_i = 0$, which contradicts the fact that $\mathbf{w}_{m+1} \neq \mathbf{0}$. Thus $m < k$ leads to a contradiction, so $m = k$, and we have verified the first claim of the Proposition.

To verify the second assertion, suppose we take a set of linearly independent vectors from each E_{λ_i}, and consider a linear combination of all these vectors which gives $\mathbf{0}$. Let $\mathbf{v}_i$ be the part of this sum which lies in E_{λ_i}. Hence we have that

$$\sum_{i=1}^{k} \mathbf{v}_i = \mathbf{0}.$$

It follows from the what we just proved that each $\mathbf{v}_i = \mathbf{0}$. But this implies that all the coefficients in the part of the sum involving vectors from E_{λ_i} are zero. Since i is arbitrary, all coefficients in the original sum are zero, proving the independence and finishing the proof. □

Using this, we obtain a new characterization of the set of all diagonalizable matrices.

Proposition 9.12. *Let A be an $n \times n$ matrix over $\mathbb{F}$, and suppose that $\lambda_1, \ \ldots \ \lambda_k$ are the distinct eigenvalues of A in $\mathbb{F}$. Then A is diagonalizable if and only if*

$$\sum_{i=1}^{k} \dim E_{\lambda_i} = n. \tag{9.12}$$

In that case,

$$\mathbb{F}^n = \sum_{i=1}^{k} E_{\lambda_i},$$

where the above sum is direct. Moreover, if the equality (9.12) *holds, the union of the bases of the E_{λ_i} is an eigenbasis of $\mathbb{F}^n$.*

Similarly, for a linear transformation $T : V \to V$, where V is a finite dimensional vector space , we obtain that a linear map $T : V \to V$ is semi-simple if and only if $\dim V = \sum_{i=1}^{k} \dim E_{\lambda_i}$, where the E_{λ_i} are the eigenspaces of T.

Given Proposition 9.12, it is easy to deduce the diagonalizability of a matrix with distinct eigenvalues, which was asserted in the last section.

Corollary 9.13. *An $n \times n$ matrix A over $\mathbb{F}$ with n distinct eigenvalues in $\mathbb{F}$ is diagonalizable.*

9.4.2 Do Non-diagonalizable Matrices Exist?

We now have a criterion which can answer the question of whether non-diagonalizable matrices can exist. Of course, we have seen that there exist real matrices which aren't diagonalizable over $\mathbb{R}$ since some of their eigenvalues are complex. But these matrices might all be diagonalizable over $\mathbb{C}$.

However, it turns out that there are matrices for which (9.12) isn't satisfied. Such matrices can't be diagonalizable. In fact, examples are quite easy to find.

Example 9.12. Consider the real 2×2 matrix

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Clearly $p_A(\lambda) = \lambda^2$, so 0 is the only eigenvalue of A. It is a repeated eigenvalue. Clearly, $E_0 = \mathcal{N}(A) = \mathbb{R}\mathbf{e}_1$, i.e. E_0 has dimension one. Therefore, A cannot have two linearly independent eigenvectors hence cannot be diagonalized over $\mathbb{R}$. For the same reason, it can't be diagonalized over $\mathbb{C}$ either.

Another way of seeing A isn't diagonalizable is to suppose it is. Then $A = MDM^{-1}$ for some invertible M. Since A's eigenvalues are both 0 and two similar matrices have he same eigenvalues, $D = \operatorname{diag}(0, 0)$. This leads to the equation $A = MDM^{-1} = O$, where O is the zero matrix. But $A \neq O$.

The above example is a nice illustration of a general fact, which we will prove in due course. Namely, the multiplicity of an eigenvalue as a root of the characteristic polynomial is at least the dimension of the corresponding eigenspace.

On the other hand, having repeated eigenvalues does not preclude diagonalizability, which is the point of Proposition 9.12. Here is an example.

Example 9.13. Here is an example with repeated eigenvalues that is rather fun to analyze. Let B denote the 4×4 all ones matrix

$$B = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Now 0 is an eigenvalue of B. In fact, B has rank 1, so the system $B\mathbf{x} = \mathbf{0} = 0\mathbf{x}$ has three linearly independent solutions. Thus the eigenspace E_0 of 0 has dimension 3. All eigenvectors for 0 satisfy the equation $x_1 + x_2 + x_3 + x_4 = 0$, which has fundamental solutions $\mathbf{f}_1 = (-1, 1, 0, 0)^T$, $\mathbf{f}_2 = (-1, 0, 1, 0)^T$, $\mathbf{f}_3 = (-1, 0, 0, 1)^T$. Another eigenvalue can be found by inspection, if we notice a special property of B. Every row of B adds up to 4. Thus, $\mathbf{f}_4 = (1, 1, 1, 1)^T$ is another eigenvector for $\lambda = 4$. By Proposition 9.11, we now have four linearly independent eigenvectors, hence an eigenbasis. Therefore B is diagonalizable; in fact, B is similar to $D = \operatorname{diag}(0, 0, 0, 4)$.

In the above example, the fourth eigenvalue of B was found by noticing a special property of B. A more methodical way to find λ_4 would have been to use the fact that the trace of a square matrix is the sum of its eigenvalues. Hence if all but one eigenvalue is known, the final eigenvalue can be found immediately. In our case, three eigenvalues are 0, hence the fourth must be the trace, which is 4.

Example 9.14. Recall the Good Will Hunting matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 0 & 2 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}.$$

The characteristic polynomial of A is

$$p(\lambda) = \lambda^4 - 7\lambda^2 - 2\lambda + 4.$$

This polynomial has -1 as a root and factors as

$$(\lambda + 1)(\lambda^3 - \lambda^2 - 6\lambda + 4).$$

We want to see if it's possible to show that $p(\lambda)$ has four real distinct roots. It's clear that -1 is not a root of $q(\lambda) = \lambda^3 - \lambda^2 - 6\lambda + 4 = 0$ Now $q'(\lambda) = 3\lambda^2 - 2\lambda - 6$ which has roots

$$r = \frac{2 \pm \sqrt{76}}{6}.$$

Now

$$q(\frac{2 + \sqrt{76}}{6}) < 0,$$

while

$$q(\frac{2 - \sqrt{76}}{6}) > 0.$$

Since the points where $q' = 0$ are not zeros of q, the graph of q crosses the real axis at three distinct points. Therefore, A has 4 distinct real eigenvalues, so it is diagonalizable.

9.4.3 The Cayley-Hamilton Theorem

We conclude this section by stating a famous result called the Cayley-Hamilton Theorem, which gives an important relationship between a matrix and its characteristic polynomial.

Theorem 9.14. *Let A be an $n \times n$ matrix over an arbitrary field $\mathbb{F}$. Then $p_A(A) = O$. That is, A satisfies its own characteristic polynomial.*

Note that by $p_A(A) = O$ we mean

$$(-1)^n A^n + (-1)^{n-1}\mathrm{Tr}(A)A^{n-1} + \cdots + \det(A)I_n = O.$$

Note that here we have put $A^0 = I_n$. For example, the characteristic polynomial of the matrix $J = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ is $\lambda^2 + 1$. By Cayley-Hamilton, $J^2 + I_2 = O$, which is easy to check directly.

There is a deceptively attractive but definitely incorrect proof of Cayley-Hamilton that goes as follows. Consider the characteristic equation $\det(A - \lambda I_n) = 0$ of A. If we set $\lambda = A$, then we get the equation $\det(A - A) = 0$, which is obviously valid. Hence A satisfies its own characteristic equation. Be sure to find the flaw in this proof?

We can outline a correct proof if $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, but we won't be able to give all the details. We will however give a complete proof in Chapter **??**. The first thing to notice is that if A is diagonal, say $A = \mathrm{diag}(d_1, \ldots, d_n)$, then $p_A(A) = \mathrm{diag}(p(d_1), p(d_2), \ldots, p(d_n))$. But the diagonal entries of a diagonal matrix are the eigenvalues, so the conclusion $p_A(A) = O$ is clear. Now if A is diagonalizable, say $A = MDM^{-1}$, then

$$p_A(A) = p_A(MDM^{-1}) = Mp_A(D)M^{-1} = MOM^{-1} = O.$$

Thus we are done if A is diagonalizable. To finish the proof, one can use limits. If A is any matrix over $\mathbb{R}$ or $\mathbb{C}$, one can show there is a sequence of diagonalizable matrices A_k such that

$$\lim_{k\to\infty} A_k = A.$$

Letting p_k denote the characteristic polynomial of A_k, we then have

$$\lim_{k\to\infty} p_k(A_k) = p(A) = O$$

since each $p_k(A_k) = O$.

Exercises

Exercise 9.35. Suppose Q is an orthogonal matrix that is diagonalizable (over $\mathbb{R}$), say $Q = PDP^{-1}$. Show that the diagonal matrix D has only ± 1 on its diagonal.

Exercise 9.36. Determine which of the following matrices is diagonalizable over the reals:

$$A = \begin{pmatrix} 1 & 0 & -1 \\ -1 & 1 & 1 \\ 2 & -1 & -2 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} 2 & -1 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -2 \end{pmatrix}.$$

Exercise 9.37. Does

$$C = \begin{pmatrix} 1 & 0 & -1 \\ -1 & 1 & 1 \\ 2 & -1 & -2 \end{pmatrix}$$

have distinct eigenvalues? Is it diagonalizable?

Exercise 9.38. Show from first principals that if λ and μ are distinct eigenvalues of A, then $E_\lambda \cap E_\mu = \{\mathbf{0}\}$.

Exercise 9.39. Find an example of a non-diagonalizable 3×3 matrix A with real entries which is not upper or lower triangular such that every eigenvalue of A is 0.

Exercise 9.40. Recall that a square matrix A is called *nilpotent* if $A^k = O$ for some integer $k > 0$.

(i) Show that if A is nilpotent, then all eigenvalues of A are 0.

(ii) Prove that a nonzero nilpotent matrix cannot be diagonalizable.

(iii) Show, conversely, that if all the eigenvalues of A are 0, then A is nilpotent. (Hint: Consider the characteristic polynomial.)

Exercise 9.41. Show that if an $n \times n$ matrix A is nilpotent, then in fact $A^n = O$.

Exercise 9.42. Let A be a 3×3 matrix with eigenvalues 0,0,1. Show that $A^3 = A^2$.

Exercise 9.43. Let A be a 2×2 matrix so that $A^2 + 3A + 2I_2 = O$. Show that $-1, -2$ are eigenvalues of A.

Exercise 9.44. Suppose A is a 2×2 matrix so that $A^2 + A - 3I_2 = O$. Show that A is diagonalizable.

Exercise 9.45. Let U be an upper triangular matrix over $\mathbb{F}$ with distinct entries on its diagonal. Show that U is diagonalizable.

Exercise 9.46. Suppose that a 3×3 matrix A with real entries satisfies the equation $A^3 + A^2 - A + 2I_3 = O$.

(i) Find the eigen-values of A.

(ii) Is A diagable? Explain.

(iii) How do you know A isn't symmetric.

Exercise 9.47. Next, let U be an arbitrary upper triangular matrix over $\mathbb{F}$ possibly having repeated diagonal entries. Show by example that U may not be diagonalizable, and give a condition to guarantee it will be diagonalizable without changing any diagonal entries.

Exercise 9.48. Give the proofs of Corollary 9.13 and Proposition 9.12.

Exercise 9.49. Prove the Cayley-Hamilton Theorem for diagonal matrices A by proving that if $p(x)$ is any polynomial, then

$$p(\mathrm{d}iag(\lambda_1, \lambda_2, \ldots, \lambda_n)) = \mathrm{d}iag(p(\lambda_1), p(\lambda_2), \ldots, p(\lambda_n)).$$

Deduce the theorem for all diagonalizable matrices from this.

Exercise 9.50. Prove the Cayley-Hamilton Theorem for upper triangular matrices, and then deduce the general case from Schur's Theorem (which you will have to look up).

Exercise 9.51. The following proof of the Cayley-Hamilton Theorem appears in at least one algebra book. Since setting $\lambda = A$ in $p(\lambda) = \det(A - \lambda I_n)$ gives $\det(A - AI_n) = 0$, it follows that $p(A) = O$. Comment on this "proof".

Exercise 9.52. Use the Cayley-Hamilton Theorem to deduce that a 2×2 matrix A is nilpotent if and only if $\mathrm{Tr}(A) = \det(A) = 0$. Generalize this result to 3×3 matrices.

Exercise 9.53. * Fill in the details of the proof of the Cayley-Hamilton Theorem suggested above using sequences. That is, show that any matrix is the limit of a sequence of diagonalizable matrices.

9.5 Matrix powers and the exponential of a matrix

We now return to the topic of the powers of a square matrix A, expanding somewhat on the remarks in §9.1. We will also extend the ordinary the *exponential* function e^x to be a function on $\mathbb{R}^{n\times n}$. We could also deal with complex matrices, but some care must be taken in defining the derivative of the exponential in the complex case. Thus we will omit it.

9.5.1 Powers of Matrices

Suppose $A \in \mathbb{R}^{n\times n}$ can be diagonalized, say $A = MDM^{-1}$. Then we saw that for any $k > 0$, $A^k = MD^kM^{-1}$. For example:

(1) If A is a diagonalizable real matrix with non negative eigenvalues, then A has a square root; in fact kth roots of A for all positive integers k are given by $A^{\frac{1}{k}} = MD^{\frac{1}{k}}M^{-1}$;

(2) If A is diagonalizable and none of the eigenvalues of A are 0, then the negative powers of A are found from the formula $A^{-k} = MD^{-k}M^{-1}$. Here, A^{-k} means $(A^{-1})^k$.

(3) If all the eigenvalues λ of A satisfy $0 \leq \lambda \leq 1$, then $\lim_{m\to\infty} A^m$ exists, and if no $\lambda = 1$, this limit is O.

We can in general obtain kth roots of a real matrix A as long as A is diagonalizable. One can't expect these matrices to be real however. For example, if A is diagonalizable but has a negative eigenvalue, then A cannot have any real square roots. However, these comments aren't valid for non-diagonalizable matrices.

9.5.2 The Exponential

Let $A \in \mathbb{R}^{n\times n}$. The exponential $\exp(A)$ of A is defined to be the matrix obtained by plugging A into the usual exponential series

$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \dots.$$

Thus the exponential $\exp(A)$ of A is given by the infinite series

$$\exp(A) = I_n + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \cdots = I_n + \sum_{m=1}^{\infty} \frac{1}{m!}A^m. \tag{9.13}$$

It can be shown that for any A, every component of the exponential series converges, a fact we will simply assume. The matrix exponential behaves just like the ordinary exponential e^x in a number of ways, but the identity $e^{(x+y)} = e^x e^y$ no longer always holds. The reason for this is that although real number multiplication is commutative, matrix multiplication definitely isn't. In fact, we have

Proposition 9.15. *If $AB = BA$, then $\exp(A+B) = \exp(A)\exp(B)$.*

If A is diagonalizable, then the matter of finding $\exp(A)$ is easily settled by the following:

Proposition 9.16. *Suppose A is a diagonalizable $n \times n$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$, say $A = M\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)M^{-1}$. Then*

$$\exp(A) = M\exp(D)M^{-1} = M\mathrm{diag}(e^{\lambda_1}, \ldots, e^{\lambda_n})M^{-1}.$$

In other words, if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is an eigenbasis of $\mathbb{R}^n$ for A, then it is also an eigenbasis for $\exp(A)$, and the eigenvector $\mathbf{v}_i$ has eigenvalue e^{λ_i}. Thus, for each $1 \leq i \leq n$,

$$\exp(A)\mathbf{v}_i = e^{\lambda_i}\mathbf{v}_i.$$

In particular, if $\mathbf{w} = \sum_{i=1}^n a_i\mathbf{v}_i$, then

$$\exp(A)\mathbf{w} = \sum_{i=1}^n a_i e^{\lambda_i}\mathbf{v}_i.$$

9.5.3 Uncoupling systems

One of the main applications of the exponential is to solve first order linear systems of differential equations. A typical application is the exponential growth problem solved in calculus. Assume $a(t)$ denotes the amount of a substance at time t that obeys the law $a'(t) = ka(t)$, where k is a constant. Then $a(t) = a_0 e^{kt}$ for all t, where a_0 is the initial amount of a.

The general form of this problem is the *first order linear system*

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathbf{x}(t) = A\mathbf{x}(t),$$

where $A \in M_n(\mathbb{R})$ and $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^T$.

The geometric interpretation of this is that $\mathbf{x}(t)$ traces out a curve in $\mathbb{R}^n$, whose velocity vector at every time t is $A\mathbf{x}(t)$. It turns out that to solve this system, we consider the derivative with respect to t of $\exp(tA)$. First notice that by Proposition 9.15,

$\exp((s+t)A) = \exp(sA)\exp(tA)$. Thus

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}\exp(tA) &= \lim_{s\to 0}\frac{1}{s}(\exp((t+s)A) - \exp(tA)) \\ &= \exp(tA)\lim_{s\to 0}\frac{1}{s}(\exp(sA) - I_n).\end{aligned}$$

It follows from the definition of $\exp(tA)$ that

$$\lim_{s\to 0}\frac{1}{s}(\exp(sA) - I_n) = A,$$

so we have the (not unexpected) formula

$$\frac{\mathrm{d}}{\mathrm{dt}}\exp(tA) = \exp(tA)A = A\exp(tA).$$

This implies that if we set $\mathbf{x}(t) = \exp(tA)\mathbf{v}$, then

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathbf{x}(t) = \frac{\mathrm{d}}{\mathrm{dt}}\exp(tA)\mathbf{v} = A\exp(tA)\mathbf{v} = A\mathbf{x}(t).$$

Hence $\mathbf{x}(t)$ is a solution curve or trajectory of our given first order system. Since $\mathbf{x}(0) = \mathbf{v}$ is the initial value of $\mathbf{x}(t)$, it follows that the initial value of the trajectory $\mathbf{x}(t)$ can be arbutrarily prescribed, so a solution curve $\mathbf{x}(t)$ can be found passing through any given initial point $\mathbf{x}(0)$.

Example 9.15. Consider the system

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

The matrix $A = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ can be written $A = M\mathrm{diag}(3,1)M^{-1}$, so

$$\exp(tA) = M\exp\begin{pmatrix} 3t & 0 \\ 0 & t \end{pmatrix}M^{-1} = M\begin{pmatrix} e^{3t} & 0 \\ 0 & e^{t} \end{pmatrix}M^{-1}.$$

Therefore, using the value of M already calculated, our solution to the system is

$$\mathbf{x}(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} e^{3t} & 0 \\ 0 & e^{t} \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} x(0) \\ y(0) \end{pmatrix}.$$

The final expression for $\mathbf{x}(t)$ is therefore

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \frac{1}{2}\begin{pmatrix} e^{3t}+e^{t} & e^{3t}-e^{t} \\ e^{3t}+e^{t} & e^{3t}-e^{t} \end{pmatrix}\begin{pmatrix} x(0) \\ y(0) \end{pmatrix}.$$

If the matrix A is nilpotent, then the system $\mathbf{x}'(t) = A\mathbf{x}(t)$ is still solved by exponentiating tA. The only difference is that A is no longer diagonalizable unless $A = O$. However, since A is nilpotent, $A^k = O$ for some $k > 0$, and so the infinite series is actually a finite sum. More generally, if A is an arbitrary real $n \times n$ matrix, it turns out that A is similar to a matrix of the form $D + N$, where D is diagonal, N is upper triangular and $DN = ND$. But then the exponential of $D+N$ is easily computed from Proposition 9.15. Namely,

$$\exp(D + N) = \exp(D)\exp(N).$$

This factorization $A = P(D+N)P^{-1}$ is known as the *Jordan Decomposition* of A. We will establish the existence of the Jordan decomposition in Chapter **??**.

Example 9.16. Consider the system

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

Notice that the matrix of the system is in the $D + N$ form above. Now

$$\exp(t \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}) = \exp(t \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix})\exp(t \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}).$$

Thus

$$\exp(t \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}) = \begin{pmatrix} e^t & 0 \\ 0 & e^t \end{pmatrix} \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} e^t & te^t \\ 0 & e^t \end{pmatrix}.$$

Finally,

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} e^t & te^t \\ 0 & e^t \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix},$$

where the point $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ gives the initial position of the solution. Therefore

$$\begin{array}{rcl} x(t) & = & x_0 e^t + y_0 t e^t \\ y(t) & = & y_0 e^t \end{array}$$

Exercises

Exercise 9.54. Find all possible square roots of the following matrices if any exist:

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}.$$

Exercise 9.55. Do the same as in Problem 9.54 for the 4×4 all 1's matrix.

Exercise 9.56. Calculate the exponentials of the matrices of Problem 9.54. What are the eigenvalues of their exponentials?

Exercise 9.57. Show that if A is diagonalizable, then $\det(\exp(A)) = e^{\mathrm{Tr}(A)}$. Conclude that $\det(\exp(A)) > 0$ for any diagonalizable A.

Exercise 9.58. Verify that $\exp(A + B) = \exp(A)\exp(B)$ if A and B are diagonal matrices. Use this formula to find the inverse of $\exp(A)$ for any square matrix A over $\mathbb{R}$.

Exercise 9.59. Recall that a square matrix A is called *nilpotent* if $A^k = O$ for some integer $k > 0$. Find a formula for the exponential of a nilpotent 3×3 matrix A such that $A^2 = O$.

Exercise 9.60. Solve the first order system $\mathbf{x}'(t) = A\mathbf{x}(t)$ with $\mathbf{x}(0) = (a, b)^T$ for the following matrices A:

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}.$$

For the last matrix, you may need to compute the exponential directly.

Exercise 9.61. Compute the nth power of all the matrices of Exercise 2 and also the 3×3 all 1's matrix.

Chapter 10

The Orthogonal Geometry of $\mathbb{R}^n$

As the title indicates, the goal of this chapter is to study some geometric problems arising in $\mathbb{R}^n$. The first of these problems is to find the minimal distance from a point of $\mathbb{R}^n$ to a subspace. This is called the least squares problem. We will also do some preparation for the Principal Axis Theorem, which is proved in the next chapter. In particular, we will show that every subspace of $\mathbb{R}^n$ has an orthonormal basis. The last section is devoted to studying the rotations of $\mathbb{R}^3$ and to giving examples of rotation groups.

10.1 Orthogonal Projection on a Subspace

Let us start with an elementary problem. Suppose W is a subspace of $\mathbb{R}^n$ and $\mathbf{x} \in \mathbb{R}^n$. What is the formula for the distance from from $\mathbf{x}$ to W? Note that by the distance from $\mathbf{x}$ to W, we mean the minimum distance $d(\mathbf{x}, \mathbf{w}) = |\mathbf{x} - \mathbf{w}|$ as $\mathbf{w}$ varies over W.

This is a problem we solved geometrically for a plane through the origin in $\mathbb{R}^3$. Recall that the distance d from $(x_0, y_0, z_0) \in \mathbb{R}^3$ to the plane $ax + by + cz = 0$ is given by

$$d = \frac{|ax_0 + by_0 + cz_0|}{(a^2 + b^2 + c^2)^{1/2}}. \tag{10.1}$$

What this formula represents of course is the length of the projection of (x_0, y_0, z_0) onto the line through the origin normal to the plane.

Before taking this up, we will recall the notion of the orthogonal complement of a subspace.

10.1.1 The orthogonal complement of a subspace

Consider a subspace W of $\mathbb{R}^n$. Recall from Example 5.24 that the *orthogonal complement* of W is the subspace $W^\perp$ of $\mathbb{R}^n$ defined as the set of all vectors $\mathbf{v} \in \mathbb{R}^n$ which are orthogonal to every vector in W. That is,

$$W^\perp = \{\mathbf{v} \in \mathbb{R}^n \mid \mathbf{v} \cdot \mathbf{w} = 0 \ \forall \ \mathbf{w} \in W\} \tag{10.2}$$

Note that $W^\perp$ is defined whether or not W is a subspace, and $W^\perp$ is always a subspace of $\mathbb{R}^n$. This is easy to visalize in terms of matrices. Suppose W is the column space of an $n \times k$ real matrix A. Then clearly $W^\perp = \mathcal{N}(A^T)$. In other words, ignoring the distinction between row and column vectors, it is clear that the row space and null space of a matrix are the orthogonal complements of one another. Applying our old principle that the number of variables in a homogeneous linear system is the number of corner variables plus the number of free variables, and using the fact that A and A^T have the same rank, we thus see that

$$\dim(W) + \dim(W^\perp) = n.$$

This leads to a basic result.

Proposition 10.1. *Let W be a subspace of $\mathbb{R}^n$ and $W^\perp$ its orthogonal complement. Then*

(i) $W \cap W^\perp = \{\mathbf{0}\}$.

(ii) *Every $\mathbf{x} \in \mathbb{R}^n$ can be orthogonally decomposed in exactly one way as $\mathbf{x} = \mathbf{w} + \mathbf{y}$, where $\mathbf{w} \in W$ and $\mathbf{y} \in W^\perp$. In particular, $W \oplus W^\perp = \mathbb{R}^n$.*

(iii) $(W^\perp)^\perp = W$.

Proof. Part (i) follows immediately from the fact that if $\mathbf{v} \in W \cap W^\perp$, then $\mathbf{v} \cdot \mathbf{v} = 0$, so $\mathbf{v} = \mathbf{0}$. The proof of (ii) is harder. Let a basis for W be $\mathbf{w}_1, \ldots, \mathbf{w}_r$ and a basis for $W^\perp$ be $\mathbf{v}_1, \ldots, \mathbf{v}_s$. We showed above that $r + s = n$, so all we have to show is that $\mathbf{w}_1, \ldots, \mathbf{w}_r, \mathbf{v}_1, \ldots, \mathbf{v}_s$ are independent since n independent vectors in $\mathbb{R}^n$ form a basis. But we know from (i) that if we have a sum $\mathbf{w} + \mathbf{v} = \mathbf{0}$, where $\mathbf{w} \in W$, and $\mathbf{v} \in W^\perp$, then $\mathbf{w} = \mathbf{v} = \mathbf{0}$. Thus if a linear combination

$$\sum a_i \mathbf{w}_i + \sum b_j \mathbf{v}_j = \mathbf{0},$$

then the a_i are all 0 and similarly, the b_j are all 0. Therefore we have the independence so $W \oplus W^\perp = \mathbb{R}^n$.

We leave (iii) as an exercise.

□

Definition 10.1. Let $\mathbf{x} = \mathbf{w} + \mathbf{y}$ as in Proposition10.1 (ii). Then we'll call $\mathbf{w}$ the component of $\mathbf{x}$ in W.

10.1.2 A Fundamental Subspace Problem

We can now solve the following fundamental
Problem: Let W be a subspace of $\mathbb{R}^n$, and let $\mathbf{x} \in \mathbb{R}^n$ be arbitrary. Find the minimum distance $d(\mathbf{x}, \mathbf{y})$, where $\mathbf{y}$ is an arbitrary vector in W. The minimum is called the *distance from* $\mathbf{x}$ *to* W.

Observe that minimizing the distance $d(\mathbf{x}, \mathbf{y})$ is equivalent to minimizing the sum of squares $|\mathbf{x} - \mathbf{y}|^2$. This is a convenient simplification so that we can use the Pythagorean property of the inner product. By part (ii) of Proposition 10.1, we may break $\mathbf{x}$ down uniquely as $\mathbf{x} = \mathbf{w} + \mathbf{v}$ with $\mathbf{w} \in W$ and $\mathbf{v} \in W^\perp$. Then for any $\mathbf{y} \in W$, $(\mathbf{x} - \mathbf{w}) \cdot (\mathbf{w} - \mathbf{y}) = 0$, so

$$\begin{aligned} |\mathbf{x} - \mathbf{y}|^2 &= |(\mathbf{x} - \mathbf{w}) + (\mathbf{w} - \mathbf{y})|^2 \\ &= |\mathbf{x} - \mathbf{w}|^2 + |\mathbf{w} - \mathbf{y}|^2 \\ &\geq |\mathbf{x} - \mathbf{w}|^2. \end{aligned}$$

Thus the minimum distance is realized by the component $\mathbf{w}$ of $\mathbf{x}$ in W.

Proposition 10.2. *The distance from $\mathbf{x} \in \mathbb{R}^n$ to the subspace W is $|\mathbf{x} - \mathbf{w}|$, where $\mathbf{w}$ is the component of $\mathbf{x}$ in W. Put another way, the distance from $\mathbf{x}$ to W is the length of the component of $\mathbf{x}$ in $W^\perp$.*

10.1.3 The Projection on a Subspace

In view of our analysis of the above least squares problem, it is clear that the next step is to find an expression for the component $\mathbf{w}$ of $\mathbf{x}$ in W. Let us begin with the following definition.

Definition 10.2. The *orthogonal projection* of $\mathbb{R}^n$ onto a subspace W is the transformation $P_W : \mathbb{R}^n \to W$ defined by $P_W(\mathbf{x}) = \mathbf{w}$, where $\mathbf{w}$ is the component of $\mathbf{x}$ in W.

Thus $P_W(\mathbf{x})$ is the unique solution of the least squares problem for the subspace W. Usually we will simply call P_W the *projection* of $\mathbb{R}^n$ onto W.

We now derive a method for finding P_W. First, choose a basis of W, say $\mathbf{w}_1, \ldots, \mathbf{w}_m$, and put $A = (\mathbf{w}_1 \quad \cdots \quad \mathbf{w}_m)$, so that W is the column space $\mathrm{col}(A)$ of A. Note that $A \in \mathbb{R}^{n \times m}$. Then $\mathbf{w}$ satisfies two conditions called the *normal equations*:

$$\text{for some } \mathbf{y} \in \mathbb{R}^m, \quad \mathbf{w} = A\mathbf{y} \quad \text{and} \quad A^T(\mathbf{x} - \mathbf{w}) = 0. \tag{10.3}$$

Now the first normal equation says $\mathbf{w}$ is a linear combination of $\mathbf{w}_1, \ldots, \mathbf{w}_m$, while the second says that $\mathbf{x} - \mathbf{w} \in W^\perp$, since the rows of A^T span W. Proposition10.1 (ii) implies the normal equations can be solved. The only question is whether the solution can be expressed elegantly (and usefully).

Multiplying the normal equations by A^T and combining leads to the single equation

$$A^T\mathbf{x} = A^T\mathbf{w} = A^TA\mathbf{y}.$$

We now call on the fact, left as an exercise (see Exercise 10.13), that if A is a real matrix with independent columns, then A^TA is invertible. Given this, we can uniquely solve for $\mathbf{y}$. Indeed,

$$\mathbf{y} = (A^TA)^{-1}A^T\mathbf{x}.$$

Multiplying by A, we get $\mathbf{w}$ in the sought after elegant form:

$$\mathbf{w} = A\mathbf{y} = A(A^TA)^{-1}A^T\mathbf{x}. \tag{10.4}$$

Therefore,

$$P_W(\mathbf{x}) = A(A^TA)^{-1}A^T\mathbf{x}. \tag{10.5}$$

Let's next consider some examples.

Example 10.1. Let W be the line in $\mathbb{R}^n$ spanned by $\mathbf{w}$. Here the projection P_W is simply

$$P_W = \mathbf{w}\big(\mathbf{w}^T\mathbf{w}\big)^{-1}\mathbf{w}^T.$$

This is a formula we already saw in Chapter 1.

Example 10.2. Let

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

Then A has rank 2 and we find by direct computation that

$$P_W = A(A^TA)^{-1}A^T = \frac{1}{17}\begin{pmatrix} 14 & 1 & 5 & 4 \\ 1 & 11 & 4 & 7 \\ 5 & 4 & 3 & 1 \\ 4 & 7 & 1 & 6 \end{pmatrix}.$$

Example 10.3. Suppose $W = \mathbb{R}^n$. Then clearly, $P_W = I_n$. In this case, A has rank n, so A and A^T are both invertible. Thus $P_W = A(A^TA)^{-1}A^T = A(A^{-1}(A^T)^{-1})A^T = I_n$ as desired.

The reader has undoubtedly noticed that the wild card is the choice of A. Some choices of A are more natural than others. We will see in the next section that an optimal choice of A is achieved by making the columns of A orthonormal. The important thing, however, is that we now have an explicit method for finding the component $P_W(\mathbf{x})$ of any $\mathbf{x} \in \mathbb{R}^n$ in W.

The following Propositionsummarizes the basic properties of projections.

Proposition 10.3. *The projection $P_W : \mathbb{R}^n \to W$ is a linear transformation, and*

$$\mathbf{x} = P_W(\mathbf{x}) + (\mathbf{x} - P_W(\mathbf{x}))$$

is the orthogonal sum decomposition of $\mathbf{x}$ into the sum of a component in W and a component in $W^\perp$. In addition, P_W has the following properties:

(*i*) *if $\mathbf{w} \in W$, then $P_W(\mathbf{w}) = \mathbf{w}$;*

(*ii*) *$P_W P_W = P_W$; and finally,*

(*iii*) *the matrix $A(A^TA)^{-1}A^T$ of P_W is symmetric.*

Proof. The fact that P_W is linear follows from the fact that it is defined by the matrix $A(A^TA)^{-1}A^T$. We already showed that $\mathbf{x} = P_W(\mathbf{x}) + (\mathbf{x} - P_W(\mathbf{x}))$ is the unique orthogonal sum decomposition of $\mathbf{x}$, so it remains to show (i)-(iii). If $\mathbf{w} \in W$, then $\mathbf{w} = \mathbf{w} + \mathbf{0}$ is an orthogonal sum decomposition of $\mathbf{w}$ with one component in W and the other in $W^\perp$, so it follows from the uniqueness of such a decomposition that $P_W(\mathbf{w}) = \mathbf{w}$. (One can also show $A(A^TA)^{-1}A^T\mathbf{w} = \mathbf{w}$, but this is a little harder.) Part (ii) follows immediately from (i) by setting $\mathbf{w} = P_W(\mathbf{x})$. We leave part (iii) as an exercise. □

In the next section, we will express P_W in terms of an orthonormal basis. This expression is theoretically much more important, because it is tied up with Fourier series.

Exercises

Exercise 10.1. Find:

(i) the component of $\mathbf{x} = (1, 1, 2)^T$ on the line $\mathbb{R}(2, -1, 0)^T$, and

(ii) the minimum distance from $\mathbf{x} = (1, 1, 2)^T$ to this line.

Exercise 10.2. Find:

(i) the component of $\mathbf{x} = (1, 1, 2, 1)^T$ in the subspace of $\mathbb{R}^4$ spanned by $(1, 1, -1, 1)-^T$ and $(2, -1, 0, 1)^T$, and

(ii) the minimum distance from $\mathbf{x} = (1, 1, 2, 1)^T$ to this subspace.

Exercise 10.3. What are the eigenvalues of a projection P_W?

Exercise 10.4. True or False: The matrix of a projection can always be diagonalized, i.e. there always exists an eigenbasis of $\mathbb{R}^n$ for every P_W.

Exercise 10.5. Suppose $n = 3$ and W is a line or a plane. Is it True or False that there exists an orthonormal eigenbasis for P_W? See §10.2 for the definition of an orthonormal basis.

Exercise 10.6. Finish the proof of Proposition10.3 by showing that every projection matrix is symmetric.

Exercise 10.7. Diagonalize (if possible) the matrix P_W in Example 10.6.

Exercise 10.8. Recall from (10.13) that the reflection $H_{\mathbf{u}}$ through the hyperplane orthogonal to a unit vector $\mathbf{u} \in \mathbb{R}^n$ is given by the formula $H_{\mathbf{u}} = I_n - 2P_W$, where $W = \mathbb{R}\mathbf{u}$ (so $H = W^{\perp}$). Find the matrix of $H_{\mathbf{u}}$ in the following cases:

(a) $\mathbf{u}$ is a unit normal to the hyperplane $x_1 + \sqrt{2}x_2 + x_3 = 0$ in $\mathbb{R}^3$.

(b) $\mathbf{u}$ is a unit normal to the hyperplane $x_1 + x_2 + x_3 + x_4 = 0$ in $\mathbb{R}^4$.

Exercise 10.9. Show that the matrix $H_{\mathbf{u}}$ defined in (10.13) is a symmetric orthogonal matrix such that $H_{\mathbf{u}}\mathbf{u} = -\mathbf{u}$ and $H_{\mathbf{u}}\mathbf{x} = \mathbf{x}$ if $\mathbf{x} \cdot \mathbf{u} = 0$.

Exercise 10.10. Let Q be the matrix of the reflection $H_{\mathbf{b}}$.

(a) What are the eigenvalues of Q?

(b) Use the result of (a) to show that $\det(Q) = -1$.

(c) Show that Q can be diagonalized by explicitly finding an eigenbasis of $\mathbb{R}^n$ for Q.

Exercise 10.11. Prove the Pythagorean relation used to prove Proposition 10.1. That is, show that if $\mathbf{p} \cdot \mathbf{q} = 0$, then

$$|\mathbf{p} + \mathbf{q}|^2 = |\mathbf{p} - \mathbf{q}|^2 = |\mathbf{p}|^2 + |\mathbf{q}|^2.$$

Conversely, if this identity holds for $\mathbf{p}$ and $\mathbf{q}$, show that $\mathbf{p}$ and $\mathbf{q}$ are orthogonal.

Exercise 10.12. Let A be the matrix

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

(i) Find the projection P_W of $\mathbb{R}^4$ onto the column space W of A.

(ii) Find the projection of $(2, 1, 1, 1)^T$ onto W.

(iii) Find the projection of $(2, 1, 1, 1)^T$ onto $W^\perp$.

Exercise 10.13. Suppose $A \in \mathbb{R}^{n \times m}$ has rank m. Show that $A^T A$ has rank m. (Hint: consider $\mathbf{x}^T A^T A \mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^m$.)

Exercise 10.14. Show that the result of Exercise 10.13 does not always hold if $\mathbb{R}$ is replaced with $\mathbb{Z}_2$ (or another $\mathbb{Z}_p$) by giving an explicit example of a 3×2 matrix A over $\mathbb{Z}_2$ of rank 2 so that $A^T A$ has rank 0 or 1.

Exercise 10.15. Suppose H is a hyperplane in $\mathbb{R}^n$ with normal line L. Interpret each of $P_H + P_L$, $P_H P_N$ and $P_N P_H$ by giving a formula for each.

10.2 Orthonormal Sets

In this section, we will study properties of orthonormal bases and give some basic applications, such as the defining pseudo-inverse of a matrix, which is based on the subspace problem we solved in the previous section. We will also use orthonormal bases to obtain another expression for the projection P_W.

10.2.1 Orthonormal Bases

We begin with the definition.

Definition 10.3. A set O of unit vectors in $\mathbb{R}^n$ is called *orthonormal* if any two distinct elements of O are orthogonal. An *orthonormal basis* of $\mathbb{R}^n$ is a basis which is orthonormal. More generally, an orthonormal basis of a subspace W of $\mathbb{R}^n$ is a basis of W which is orthonormal.

Proposition 10.4. *The vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ give an orthonormal basis of $\mathbb{R}^n$ if and only if the matrix $U = (\mathbf{u}_1\ \mathbf{u}_2 \ldots \mathbf{u}_n)$ is orthogonal.*

Proof. This is clear since $U^TU = I_n$ if and only if $(\mathbf{u}_i^T\mathbf{u}_j) = (\mathbf{u}_i \cdot \mathbf{u}_j) = I_n$. □

Proposition 10.5. *Any orthonormal set in $\mathbb{R}^n$ is linearly independent. In particular, an orthonormal subset of $\mathbb{R}^n$ cannot contain more than n elements. In addition, if W is a subspace of $\mathbb{R}^n$ such that $\dim W = m$, then any orthonormal set in W having m elements is an orthonormal basis of W.*

Proof. We leave this is an exercise. □

We now establish a basic fact: every subspace of $\mathbb{R}^n$ admits an orthonormal basis.

Proposition 10.6. *Every non trivial subspace of $\mathbb{R}^n$ admits an orthonormal basis.*

Proof. We prove this by induction on $\dim W$. If $\dim W = 1$, the result is clear since either vector in W of length one is an orthonormal basis. Thus suppose $\dim W = m > 1$ and the result is true for any subspace of W of dimension $m - 1$. Let $\mathbf{u}$ be any unit vector in W and let $H = (\mathbb{R}\mathbf{u})^{\perp}$. Then $\dim H = m - 1$, so we know H admits an orthonormal basis. But this orthonormal basis, together with W give m orthonormal elements of W, hence, by the previous Proposition, are an orthonormal basis of W. □

Example 10.4 (Some orthonormal bases). Here are some examples.

(a) The standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ is an orthonormal basis of $\mathbb{R}^n$.

(b) $\mathbf{u}_1 = \frac{1}{\sqrt{3}}(1,1,1)^T$, $\mathbf{u}_2 = \frac{1}{\sqrt{6}}(1,-2,1)^T$, $\mathbf{u}_3 = \frac{1}{\sqrt{2}}(1,0,-1)^T$ are an orthonormal basis of $\mathbb{R}^3$. The first two basis vectors are an orthonormal basis of the plane $x - z = 0$.

(c) The columns of an orthogonal matrix Q are an orthonormal basis of $\mathbb{R}^n$. Using the fact that the matrix

$$Q = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \end{pmatrix},$$

is orthogonal, can you produce two distinct orthonormal bases of $\mathbb{R}^4$.

10.2.2 Fourier Coefficients and the Projection Formula

Suppose $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ is an orthonormal basis of $\mathbb{R}^n$ and $\mathbf{w} \in \mathbb{R}^n$. How do we find the scalars $a_1, \ldots, a_n$ such that $\mathbf{w} = a_1\mathbf{u}_1 + \cdots + a_n\mathbf{u}_n$? Of course we know that the a_i can be found by solving a system of n equations in n variables. However, as our next Propositionshows, there is a much neater solution.

Proposition 10.7 (Projection Formula). *Suppose $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ is any orthonormal basis of $\mathbb{R}^n$, and let $\mathbf{w} \in \mathbb{R}^n$ be arbitrary. Then $\mathbf{w}$ has the unique expansion*

$$\mathbf{w} = \sum_{1=i}^{n} (\mathbf{w} \cdot \mathbf{u}_i)\mathbf{u}_i = \sum_{1=i}^{n} (\mathbf{w}^T\mathbf{u}_i)\mathbf{u}_i. \tag{10.6}$$

Before giving the proof, let's make some commments. First of all, the coefficients in (10.6) have a particularly simple form. In fact, we make the following definition.

Definition 10.4. The scalars $\mathbf{w} \cdot \mathbf{u}_i = \mathbf{w}^T\mathbf{u}_i$ are called the *Fourier coefficients* of $\mathbf{w}$ with respect to the orthonormal basis $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$.

The reason this result is called the projection formula for $\mathbb{R}^n$ is that it says every vector in $\mathbb{R}^n$ can be expressed as the sum of its projections on any given orthonormal basis.

The projection formula can also be stated in a matrix form, namely

$$I_n = \sum_{1=i}^{n} \mathbf{u}_i\mathbf{u}_i^T, \tag{10.7}$$

which says the sum of the projections on an orthonormal basis is the identity.

Example 10.5. For example,

$$(1,0,0,0) = \frac{1}{2}(1,1,1,1) - \frac{1}{2}(-1,1,-1,1) + \frac{1}{2}(1,-1,-1,1) + \frac{1}{2}(1,1,-1,-1).$$

Let us now prove the formula.

Proof. To begin, write

$$\mathbf{w} = \sum_{1=i}^{n} x_i \mathbf{u}_i.$$

To find the x_i, consider the system

$$Q\mathbf{x} = \mathbf{w},$$

where $Q = (\mathbf{u}_1\ \mathbf{u}_2 \dots \mathbf{u}_n)$. Since Q is orthogonal, the unique solution is

$$\mathbf{x} = Q^T\mathbf{w}.$$

But this that says that for each i,

$$x_i = \mathbf{u}_i^T\mathbf{w} = \mathbf{u}_i \cdot \mathbf{w} = \mathbf{w} \cdot \mathbf{u}_i,$$

which is the desired formula. □

More generally, suppose W is a subspace of $\mathbb{R}^n$ with an orthonormal basis $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m$. By a similar argument, each $\mathbf{w} \in W$ has the unique expansion

$$\mathbf{w} = \sum_{1=i}^{m} (\mathbf{w} \cdot \mathbf{u}_i)\mathbf{u}_i. \tag{10.8}$$

To see this, first write

$$\mathbf{w} = \sum_{1=i}^{m} x_i \mathbf{u}_i,$$

and repeat the argument above.

I claim that the formula for the projection P_W is

$$P_W(\mathbf{x}) = \sum_{i=1}^{m} (\mathbf{x} \cdot \mathbf{u}_i)\mathbf{u}_i. \tag{10.9}$$

There are two ways to show this. The first is to use the fact that $P_W(\mathbf{x})$ is characterized by the property that $\mathbf{x} - P_W(\mathbf{x}) \in W^\perp$. But $\mathbf{y} = \mathbf{x} - \sum_{1=i}^{m} (\mathbf{x} \cdot \mathbf{u}_i)\mathbf{u}_i \in W^\perp$. Indeed, $\mathbf{y} \cdot \mathbf{u}_i = 0$ for each i. Hence $P_W(\mathbf{x}) = \mathbf{y}$.

The second proof of (10.9) is to calculate the matrix of P_W directly, using our orthonormal basis. That is, suppose $Q = (\mathbf{u}_1 \cdots \mathbf{u}_m)$. Since $\text{col}(Q) = W$, $P_W = Q(Q^TQ)^{-1}Q^T$. But since $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are orthonormal, $Q^TQ = I_m$ (check this), so $P_W = QI_mQ^T = QQ^T$.

Hence we have proven

Proposition 10.8. *Let $\mathbf{u}_1, \ldots, \mathbf{u}_m$ be an orthonormal basis for a subspace W of $\mathbb{R}^n$. Then the projection $P_W : \mathbb{R}^n \to W$ is given by*

$$P_W(\mathbf{x}) = \sum_{i=1}^{m} (\mathbf{x} \cdot \mathbf{u}_i)\mathbf{u}_i. \tag{10.10}$$

If $Q \in \mathbb{R}^{n\times m}$ is the matrix $Q = (\mathbf{u}_1 \ \cdots \ \mathbf{u}_m)$ with orthonormal columns, then

$$P_W = QQ^T. \tag{10.11}$$

Put another way,

$$P_W = \sum_{i=1}^{m} \mathbf{u}_i\mathbf{u}_i^T. \tag{10.12}$$

Equation (10.11) is the optimal expression for P_W we mentioned above.

Example 10.6. Let $W = \text{span}\{(1,1,1,1)^T, (1,-1,-1,1)^T\}$. To find the matrix of P_W the old fashioned way, we would compute $P_W(\mathbf{e}_i)$ for $i = 1,2,3,4$. Observe that $\mathbf{u}_1 = 1/2(1,1,1,1)^T$ and $\mathbf{u}_2 = 1/2(1,-1,-1,1)^T$ are an orthonormal basis of W. Now, by a straightforward computation,

$$P_W(\mathbf{e}_1) = \frac{1}{2}(1,0,0,1)^T, \quad P_W(\mathbf{e}_2) = \frac{1}{2}(0,1,1,0)^T.$$

By inspection, $P_W(\mathbf{e}_3) = P_W(\mathbf{e}_2)$ and $P_W(\mathbf{e}_4) = P_W(\mathbf{e}_1)$. Hence the matrix A of P_W is

$$A = \frac{1}{2}\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

The simpler method would have been to calculate QQ^T, where $Q = (\mathbf{u}_1 \ \mathbf{u}_2)$. Now

$$QQ^T = 1/4 \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix},$$

which of course gives the same result as the old fashioned method, but more efficiently.

The projection onto a hyperplane W in $\mathbb{R}^n$ with unit normal $\mathbf{u} \neq \mathbf{0}$ has the form

$$P_W(\mathbf{x}) = (I_n - \mathbf{u}\mathbf{u}^T)\mathbf{x}.$$

This shows that to find P_W, one doesn't need an orthonormal basis for W as long as an orthonormal basis for $W^\perp$ is given. This makes computing reflections a simple matter. Using the same reasoning as in Chapter 1, the reflection through a hyperplane W is the linear transformation

$$H_{\mathbf{u}} = I_n - 2P_{\mathbf{u}}, \tag{10.13}$$

where $\mathbf{u}$ is a unit vector in the line $W^\perp$. Note $H_{\mathbf{u}}(\mathbf{u}) = -\mathbf{u}$ and $H_{\mathbf{u}}(\mathbf{w}) = \mathbf{w}$ for all $\mathbf{w} \in W$. That is, $H_{\mathbf{u}}$ has the properties expected of a reflection. We leave it as an exercise to show that $H_{\mathbf{u}}$ is a symmetric orthogonal matrix.

The notion of Fourier coefficients and orthogonal projections are important in infinite dimensional situations also. For example, recall that $C[a, b]$ is an inner product space with the inner product

$$(f, g) = \int_a^b f(t)g(t)\mathrm{d}t.$$

A set S of functions in $C[a, b]$ is orthonormal if for any $f, g \in S$,

$$(f, g) = \begin{cases} 0 & \text{if } f \neq g \\ 1 & \text{if } f = g \end{cases}$$

The formula for projecting $C[a, b]$ onto the subspace W spanned by a finite set of functions is exactly as given above, provided an orthonormal basis of W has been found. The next Section is devoted to describing a natural method for producing such an orthonormal basis. Before that, however, we give some further considerations on the method of least squares.

10.2.3 The Pseudo-Inverse and Least Squares

Suppose $A \in \mathbb{R}^{n\times m}$ has independent columns, and W denotes the column space of A. Then the matrix $A^+ = (A^TA)^{-1}A^T$ is called the *pseudo-inverse* of A. If $m = n$ so that A is square, then A and A^T are both invertible, and $A^+ = A^{-1}$. In general, A^+ is always a left inverse of A. That is, $A^+A = I_m$.

To see a little better what is going on, look at A as a linear transformation $A : \mathbb{R}^m \to \mathbb{R}^n$. Since the columns of A are independent, $\mathcal{N}(A) = \mathbf{0}$, hence A is one to one. Thus, $A\mathbf{x} = A\mathbf{y}$ implies $\mathbf{x} = \mathbf{y}$. By the result of Exercise 10.25, every one to one linear map $T : \mathbb{F}^m \to \mathbb{F}^n$ has at least one left inverse

$B : \mathbb{F}^n \to \mathbb{F}^m$, which is also linear. (Here, $\mathbb{F}$ is an arbitrary field.) However, when $m < n$, the inverse left inverse B, turns out not to be unique.

The pseudo-inverse A^+ is a choice of left inverse with certain useful properties. In particular, not only is $A^+A = I_m$, but $AA^+ = P_W$. Thus A^+ solves an important least squares problem for W: given $\mathbf{b} \in \mathbb{R}^n$, find $\mathbf{x} \in \mathbb{R}^m$ so that $A\mathbf{x}$ is the element of W nearest $\mathbf{b}$. The solution is $\mathbf{x} = A^+\mathbf{b}$, and

$$A\mathbf{x} = AA^+\mathbf{b} = P_W(\mathbf{b}).$$

This amounts to a method for finding an optimal solution of an inconsistent system. The system $A\mathbf{x} = \mathbf{b}$ if inconsistent, can be replaced by the nearest consistent system

$$A\mathbf{x} = AA^+\mathbf{b} = P_W(\mathbf{b}).$$

since $P_W(\mathbf{b})$ is the vector in the column space W of A nearest $\mathbf{b}$, and luckily we already know that the solution to this system is $\mathbf{x} = A^+\mathbf{b}$. One can easily envision that this idea has error correcting possibilities.

Notice that if the original system $A\mathbf{x} = \mathbf{b}$ had been consistent, that is $\mathbf{b} \in W$, then first projecting wouldn't have changed anything since $P_W(\mathbf{b}) = \mathbf{b}$.

Let's consider a typical application. Suppose one has m points (a_i, b_i) in $\mathbb{R}^2$, which represent the outcome of an experiment. Typically, one wants to find the line $y = cx + d$ fitting these points as well as possible. The points (a_i, b_i) are all lined up (so to speak) if and only if $b_i = ca_i + d$ for all $i = 1, \ldots, m$. When this happens, we get the matrix equation

$$A\mathbf{x} = \begin{pmatrix} a_1 & 1 \\ a_2 & 1 \\ \cdots & \cdots \\ a_m & 1 \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \cdot\,\cdot \\ b_m \end{pmatrix},$$

with the unknowns c and d in the components of $\mathbf{x}$. Obviously, the property of being lined up has no real chance of happening, so we need to apply least squares. Hence we replace $b_1, \ldots, b_m$ by $c_1, \ldots, c_m$ so that all (a_i, c_i) lie on a line and the sum

$$\sum_{i=1}^{m} (b_i - c_i)^2$$

is minimized. Using the pseudo-inverse A^+, we get

$$\mathbf{x} = \begin{pmatrix} c \\ d \end{pmatrix} = A^+\mathbf{b} = (A^TA)^{-1}A^T\mathbf{b}.$$

Thus,

$$\begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} \sum a_i^2 & \sum a_i \\ \sum a_i & m \end{pmatrix}^{-1} \begin{pmatrix} \sum a_i b_i \\ \sum b_i \end{pmatrix}.$$

Note that the 2×2 matrix in this solution is invertible just as long as we don't have all $a_i = 0$ or all $a_i = 1$.

The problem of fitting a set of points (a_i, b_i, c_i) to a plane is similar. The method can also be adapted to the problem of fitting a set of points in $\mathbb{R}^2$ to a nonlinear curve, such as an ellipse. This is apparently the origin of the least squares method. Its inventor, the renowned mathematician Gauss, astonished the astronomical world in 1801 by his prediction, based on approximately 9° of observed orbit, of the position where astronomers would find an obscure astroid named Ceres a full 11 months after his initial calculations had been made.

Least squares can apply to function spaces such as $C[a, b]$ as well.

Example 10.7. Suppose we want to minimize the integral

$$\int_{-1}^{1} (\cos x - (a + bx + cx^2))^2 \mathrm{d}x.$$

The solution proceeds exactly as in the Euclidean situation. The problem is to minimize the square of the distance from the function $\cos x$ on $[-1, 1]$ to the subspace of $C[-1, 1]$ spanned by 1, x and x^2. We first apply Gram-Schmidt to $1, x, x^2$ on $[-1, 1]$ to obtain orthonormal polynomials f_0, f_1, f_2 on $[-1, 1]$, and then compute the Fourier coefficients of $\cos x$ with respect to the f_i. Clearly $f_0(x) = \frac{1}{\sqrt{2}}$. Since x is odd and the interval is symmetric about the origin, $(x, f_0) = 0$. Hence

$$f_1(x) = \frac{x}{\sqrt{(x, x)}} = \sqrt{\frac{2}{3}} x.$$

To get f_2, we calculate $x^2 - (x^2, f_0) f_0 - (x^2, f_1) f_1$ which turns out to be $x^2 - \frac{1}{3}$. Computing $(x^2 - \frac{1}{3}, x^2 - \frac{1}{3})$, we get $\frac{8}{45}$, so

$$f_2(x) = \frac{\sqrt{45}}{\sqrt{8}} (x^2 - \frac{1}{3}).$$

The Fourier coefficients $(\cos x, f_i) = \int_{-1}^{1} \cos x f_i(x) \mathrm{d}x$ turn out to be $\frac{2 \cos 1}{\sqrt{2}}, 0$ and $\frac{\sqrt{45}}{\sqrt{8}} (4 \cos 1 - \frac{8}{3} \sin 1)$. Thus the best least squares approximation is

$$\cos 1 + \frac{\sqrt{45}}{\sqrt{8}} (4 \cos 1 - \frac{8}{3} \sin 1)(x^2 - \frac{1}{3}).$$

The calculation was greatly simplified by the fact that $[-1, 1]$ is symmetric about 0, since x and x^2 are already orthogonal on $[-1, 1]$, as are any two polynomials such that one is even and the other odd.

Exercises

Exercise 10.16. Expand $(1,0,0)^T$ using the orthonormal basis consisting of the columns of the matrix Q of Example 10.4(b). Do the same for $(1,0,0,0)$ using the rows of U.

Exercise 10.17. Find an orthonormal basis for the plane $x - 2y + 3z = 0$ in $\mathbb{R}^3$. Now extend this orthonormal set in $\mathbb{R}^3$ to an orthonormal basis of $\mathbb{R}^3$.

Exercise 10.18. Prove Proposition10.5. That is, show that any orthonormal set in $\mathbb{R}^n$ is linearly independent and cannot contain more than n elements. (For the first part, use the projection formula.)

Exercise 10.19. Let $Q = (\mathbf{u}_1\mathbf{u}_2\cdots\mathbf{u}_n) \in O(n,\mathbb{R})$. If Q is not symmetric, show how to produce a new orthonormal basis of $\mathbb{R}^n$ from the columns of Q. What new orthonormal basis of $\mathbb{R}^4$ does one obtain from the orthonormal basis in Example 10.4, part (c)?

Exercise 10.20. Assume $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ is an orthonormal basis of $\mathbb{R}^n$. Show that $I_n = \sum_{1=i}^{n} \mathbf{u}_i\mathbf{u}_i^T$. This verifies the identity in (10.7).

Exercise 10.21. Let $\mathbf{u}$ be a unit vetor in $\mathbb{R}^n$. Show that the reflection $H_{\mathbf{u}}$ through the hyperplane H orthogonal to $\mathbf{u}$ admits an orthonormal eigenbasis.

Exercise 10.22. Find the line that best fits the points $(-1,1)$, $(0,.5)$, $(1,2)$, and $(1.5,2.5)$.

Exercise 10.23. Suppose coordinates have been put on the universe so that the sun's position is $(0,0,0)$. Four observations of a planet orbiting the sun tell us that the planet passed through the points $(5,.1,0)$, $(4.2,2,1.4)$, $(0,4,3)$, and $(-3.5,2.8,2)$. Find the plane (through the origin) that best fits the planet's orbit.

Exercise 10.24. Find the pseudo-inverse of the matrix

$$\begin{pmatrix} 1 & 0 \\ 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

Exercise 10.25. Let $\mathbb{F}$ be any field. Show that every one to one linear map $T : \mathbb{F}^m \to \mathbb{F}^n$ has at least one left inverse $B : \mathbb{F}^n \to \mathbb{F}^m$.

Exercise 10.26. Show that if A has independent columns, then any left inverse of A has the form $A^+ + C$, where $CA = O$. (Note: $CA = O$ is equivalent to $\text{col}(A) \subset \mathcal{N}(C)$. If $CA = O$, what is $(A^+ + C)A$? And conversely?)

Exercise 10.27. Suppose A has independent columns and let $A = QR$ be the QR factorization of A.

(i) Find the pseudo-inverse A^+ of A in terms of Q and R; and

(ii) Find a left inverse of A in terms of Q and R.

Exercise 10.28. Consider the matrix

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

(a) Find the pseudo-inverse A^+ of A, and

(b) Compute the QR factorization of A and use the result to find another left inverse of A.

Exercise 10.29. Let W be a subspace of $\mathbb{R}^n$ with basis $\mathbf{w}_1, \ldots, \mathbf{w}_k$ and put $A = (\mathbf{w}_1 \ \cdots \ \mathbf{w}_k)$. Show that A^TA is always invertible. (HINT: It is sufficient to show that $A^TA\mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$ (why?). Now consider $\mathbf{x}^TA^TA\mathbf{x}$.)

10.3 Gram-Schmidt and the QR Factorization

10.3.1 The Gram-Schmidt Method

We are next going to give a constructive method for finding orthonormal bases. Given a subspace W with a basis $\mathbf{w}_1, \ \dots \ , \mathbf{w}_m$, we will in fact construct an orthonormal basis $\mathbf{u}_1, \ \dots \ , \mathbf{u}_m$ of W such that for each index for $k = 1, \dots m$,

$$\text{span}\{\mathbf{u}_1, \ \dots \ , \mathbf{u}_k\} = \text{span}\{\mathbf{w}_1, \ \dots \ , \mathbf{w}_k\}.$$

This is referred to as the *Gram-Schmidt method* (just Gram-Schmidt for short). In fact, Gram-Schmidt is simply a repeated application of Proposition 10.3 (ii).

The reason we treat Gram-Schmidt in such detail is because it's applicable in more general settings than $\mathbb{R}^n$. That is, Gram-Schmidt gives a constructive method for constructing an orthonormal basis of any finite dimensional subspace W of an arbitrary inner product space, such as $C[a, b]$.

Consider a subspace W of $\mathbb{R}^n$ having a basis $\mathbf{w}_1, \ \dots \ , \mathbf{w}_m$. Recall that no proper subset of this basis can span W. For each index j, let $W_j = \text{span}\{\mathbf{w}_1, \dots, \mathbf{w}_j\}$. Clearly $W_j \subset W_{j+1}$, $\dim W_j = j$ and $W_m = W$. Now proceed as follows: as $\mathbf{w}_1 \neq \mathbf{0}$, we may put

$$\mathbf{u}_1 = |\mathbf{w}_1|^{-1}\mathbf{w}_1.$$

Then $\mathbf{u}_1$ is an orthonormal basis of W_1. Next, we need a non zero vector $\mathbf{v}_2$ in W_2 orthogonal to $\mathbf{u}_1$. In fact, we can just let $\mathbf{v}_2$ be the component of $\mathbf{w}_2$ orthogonal to $\mathbf{u}_1$. Since $\mathbf{w}_2 = P_{W_1}(\mathbf{w}_2) + (\mathbf{w}_2 - P_{W_1})$ is the decomposition of $\mathbf{w}_2$ into orthogonal components (by Proposition 10.3), we put

$$\mathbf{v}_2 := \mathbf{w}_2 - P_{W_1}(\mathbf{w}_2) = \mathbf{w}_2 - (\mathbf{w}_2 \cdot \mathbf{u}_1)\mathbf{u}_1.$$

Clearly $\mathbf{v}_2 \neq \mathbf{0}$, so put

$$\mathbf{u}_2 := |\mathbf{v}_2|^{-1}\mathbf{v}_2.$$

Then $\mathbf{u}_1$ and $\mathbf{u}_2$ are orthogonal, so they are an orthonormal basis of W_2.

To continue, consider W_3. Again, by Proposition 10.3, the vector

$$\mathbf{v}_3 = \mathbf{w}_3 - P_{W_2}(\mathbf{w}_3) = \mathbf{w}_3 - (\mathbf{w}_3 \cdot \mathbf{u}_1)\mathbf{u}_1 - (\mathbf{w}_3 \cdot \mathbf{u}_2)\mathbf{u}_2$$

is orthogonal to W_2. Moreover, $\mathbf{v}_3 \in W_3$ and $\mathbf{v}_3 \neq \mathbf{0}$ (why?). As above, putting

$$\mathbf{u}_3 = |\mathbf{v}_3|^{-1}\mathbf{v}_3.$$

gives an orthonormal basis $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ of W_3.

In general, if $1 < j \leq m$, suppose an orthonormal basis $\mathbf{u}_1, \ \dots \ , \mathbf{u}_{j-1}$ of W_{j-1} is already given. Then $\mathbf{v}_j := \mathbf{w}_j - P_{W_{j-1}}(\mathbf{w}_j)$ is orthogonal to W_{j-1} and $\mathbf{v}_j \neq \mathbf{0}$ since $\mathbf{w}_j \notin W_{j-1}$. Hence,

$$\mathbf{v}_j = \mathbf{w}_j - (\mathbf{w}_j \cdot \mathbf{u}_1)\mathbf{u}_1 - (\mathbf{w}_j \cdot \mathbf{u}_2)\mathbf{u}_2 - \cdots - (\mathbf{w}_j \cdot \mathbf{u}_{j-1})\mathbf{u}_{j-1}$$

and $\mathbf{u}_1, \dots \mathbf{u}_{j-1}$ give an orthonormal set in W_j. Thus putting

$$\mathbf{u}_j = |\mathbf{v}_j|^{-1}\mathbf{v}_j$$

gives an orthonormal basis $\mathbf{u}_1, \ \dots \ , \mathbf{u}_{j-1}, \mathbf{u}_j$of W_j since $\dim W_j = j$. Continuing in this manner, will yield an orthonormal basis $\mathbf{u}_1, \ \dots \ , \mathbf{u}_m$ of W with the desired property that the span of $\mathbf{u}_1, \ \dots \ , \mathbf{u}_j$ is W_j for each j. Hence we have shown

Proposition 10.9. *Suppose* $\mathbf{w}_1, \ \dots \ , \mathbf{w}_m \in \mathbb{R}^n$ *are linearly independent. Then Gram-Schmidt produces an orthonormal subset* $\{\mathbf{u}_1, \ \dots \ , \mathbf{u}_m\} \subset W$ *such that*

$$\text{span}\{\mathbf{u}_1, \ \dots \ , \mathbf{u}_k\} = \text{span}\{\mathbf{w}_1, \ \dots \ , \mathbf{w}_k\}$$

for each k *with* $1 \leq k \leq m$.

10.3.2 The QR Decomposition

The Gram-Schmidt method can be stated in an alternate form which is quite important in applied linear algebra, for example as the basis of the the QR algorithm. This is known as the QR *decomposition*.

Suppose a matrix $A = (\mathbf{w}_1 \ \mathbf{w}_2 \ \mathbf{w}_3)$ with independent columns is given. Applying Gram-Schmidt to $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ gives an orthonormal basis $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ of the column space $\text{col}(A)$. Moreover, by construction, the following matrix identity holds:

$$(\mathbf{w}_1 \ \mathbf{w}_2 \ \mathbf{w}_3) = (\mathbf{u}_1 \ \mathbf{u}_2 \ \mathbf{u}_3) \begin{pmatrix} \mathbf{w}_1 \cdot \mathbf{u}_1 & \mathbf{w}_2 \cdot \mathbf{u}_1 & \mathbf{w}_3 \cdot \mathbf{u}_1 \\ 0 & \mathbf{w}_2 \cdot \mathbf{u}_2 & \mathbf{w}_3 \cdot \mathbf{u}_2 \\ 0 & 0 & \mathbf{w}_3 \cdot \mathbf{u}_3 \end{pmatrix}.$$

In general, if $A = (\mathbf{w}_1 \ \cdots \ \mathbf{w}_m)$ is an $n \times m$ matrix over $\mathbb{R}$ with linearly independent columns, let $Q = (\mathbf{u}_1 \ \cdots \ \mathbf{u}_m)$ be the associated $n \times m$ matrix produced by the Gram-Schmidt method and R is the $m \times m$ upper triangular matrix of Fourier coefficients. Then

$$A = QR. \tag{10.14}$$

Summarizing (and adding a bit more), we have

Proposition 10.10. *Every $A \in \mathbb{R}^{n\times m}$ of rank m can be factored $A = QR$, where $Q \in \mathbb{R}^{n\times m}$ has orthonormal columns and $R \in \mathbb{R}^{m\times m}$ is invertible and upper triangular.*

Proof. The only assertion we have left to show is that R is invertible. Thus we have to show that $\mathbf{w}_i \cdot \mathbf{u}_i \neq 0$ for each i. But if $\mathbf{w}_i \cdot \mathbf{u}_i = 0$, then $\mathbf{u}_i \in W_{i-i}$ (here we set $W_0 = \{\mathbf{0}\}$ if $i = 1$). But the Gram-Schmidt method guarantees this can't happen, so we are through. □

If A is square, then both Q and R are square. In particular, Q is an orthogonal matrix. The factorization $A = QR$ is the first step in the QR algorithm, which is an important method for approximating the eigenvalues of A. For more details, see Chapter 12.

Exercises

Exercise 10.30. Let $W \subset \mathbb{R}^4$ be the span of $(1,0,1,1)$, $(-1,1,0,0)$, and $(1,0,1,-1)$.

(i) Find an orthonormal basis of W.

(ii) Expand $(0,0,0,1)$ and $(1,0,0,0)$ in terms of this basis.

Exercise 10.31. Let

$$A := \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & -1 \end{pmatrix}.$$

Find the QR factorization of A.

Exercise 10.32. Find a 4×4 orthogonal matrix Q whose first three columns are the columns of A in the previous problem.

Exercise 10.33. What would happen if the Gram-Schmidt method were applied to a set of vectors that were not lineary independent? In other words, why can't we produce an orthonormal basis from nothing?

Exercise 10.34. In the QR decomposition, we claimed that the diagonal entries of R are non zero, hence R is invertible. Explain why they are indeed non zero.

Exercise 10.35. Suppose $A = QDQ^{-1}$ with Q orthogonal and D diagonal. Show that A is always symmetric and that A is orthogonal if and only if all diagonal entries of D are either

± 1. Show that A is the matrix of a reflection $H_{\mathbf{u}}$ precisely when $D = \operatorname{diag}(-1,1,\ldots,1)$, that is exactly one diagonal entry of D is -1 and all others are $+1$.

Exercise 10.36. How would you define the reflection H through a subspace W of $\mathbb{R}^n$? What properties should the matrix of H have? For example, what should the eigenvalues of H be?

Exercise 10.37. Check directly that if $R = I_n - P_W$, then $R^2 = R$. Verify also that the eigenvalues of R are 0 and 1 and that $E_0 = W$ and $E_1 = W^{\perp}$.

Exercise 10.38. Show that for any subspace W of $\mathbb{R}^n$, P_W can be expressed as $P_W = QDQ^T$, where D is diagonal and Q is orthogonal. Find the diagonal entries of D, and describe Q.

Exercise 10.39. Find an orthonormal basis of the plane W in $\mathbb{R}^4$ spanned by $(0,1,0,1)$ and $(1,-1,0,0)$. Do the same for $W^\perp$. Now find an orthonormal basis of $\mathbb{R}^4$ containing the orthonormal bases of W and $W^\perp$.

Exercise 10.40. Let A have independent columns. Verify the formula $P = QQ^T$ using $A = QR$.

Exercise 10.41. Suppose A has independent columns and let $A = QR$ be the QR factorization of A.

(i) Find the pseudo-inverse A^+ of A in terms of Q and R; and

(ii) Find a left inverse of A in terms of Q and R.

Exercise 10.42. The Gram-Schmidt method applies to the inner product on $C[a,b]$ as well.

(a) Apply Gram-Schmidt to the functions $1,\ x,\ x^2$ on the interval $[-1,1]$ to produce an orthonormal basis of the set of polynomials on $[-1,1]$ of degree at most two. The resulting functions P_0, P_1, P_2 are the first three normalized *orthogonal polynomials* of Legendre type.

(b) Show that your nth polynomial P_n satisfies the differential equation

$$(1-x^2)y'' - 2xy' + n(n+1)y = 0.$$

(c) The nth degree Legendre polynomial satisfies this second order differential equation equation for all $n \geq 0$. This and the orthogonality condition can be used to generate all the Legendre polynomials. Find P_3 and P_4 without GS.

Exercise 10.43. Using the result of the previous exercise, find the projection of $x^4 + x$ on the subspace of $C[-1,1]$ spanned by $1,\ x,\ x^2$.

10.4 The group of rotations of $\mathbb{R}^3$

In crystallography, the study of the molecular structure of crystals, one of the basic problems is to determine the set of rotational symmetries of a particular crystal. More generally, one may also consider the problem of determining the set of all rotational symmetries $\mathrm{Rot}(B)$ of an arbitrary solid B in $\mathbb{R}^3$. The set of these symmetries is known as the *rotation group* of B. One of the first problems which one thinks of is to find the rotation groups of the Platonic solids in $\mathbb{R}^3$. A *Platonic solid* in $\mathbb{R}^3$ is a solid whose boundary is made up of plane polygons in such a way that all the polygons that make up the boundary are congruent. It has been known since the time of the Greeks that there are exactly five Platonic solids: a cube, a regular quadrilateral, a regular tetrahedron, a regular dodecahedron and a regular icosahedron. We will take these up again in Chapter 16.

10.4.1 Rotations of $\mathbb{R}^3$

The first question we need to consider is what a rotation of $\mathbb{R}^3$ is. We will use a characterization, due to Euler, that says that a rotation ρ of $\mathbb{R}^3$ is characterized as a transformation R which fixes every point on some axis through the origin, and rotates every plane orthogonal to this axis through the same fixed angle θ.

Using this as the basic definition, we will now show that all rotations are linear and describe them in terms of matrix theory. It is clear that a rotation R of $\mathbb{R}^3$ about $\mathbf{0}$ should preserve lengths and angles. Recalling that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$,

$$\mathbf{x} \cdot \mathbf{y} = |\mathbf{x}||\mathbf{y}| \cos \alpha,$$

we see that any transformation of $\mathbb{R}^3$ preserving both lengths and angles also preserves the all dot products. Thus if $\rho \in \mathrm{Rot}(\mathbb{R}^3)$,

$$\rho(\mathbf{x}) \cdot \rho(\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}. \tag{10.15}$$

Therefore, every rotation is given by an orthogonal matrix, and we see that $\mathrm{Rot}(\mathbb{R}^3) \subset O(3, \mathbb{R})$, the set of 3×3 orthogonal matrices.

We now need the following fact.

Proposition 10.11. *A transformation $\rho : \mathbb{R}^n \to \mathbb{R}^n$ satisfying*

$$\rho(\mathbf{x}) \cdot \rho(\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is linear. Hence, ρ is given by an orthogonal linear transformationand its matrix is orthogonal.

Proof. If we can show ρ is linear, we will be done since from Proposition7.8, the matrix of ρ is orthogonal. We have to show two things: for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$|\rho(\mathbf{x}+\mathbf{y}) - \rho(\mathbf{x}) - \rho(\mathbf{y})|^2 = 0$$

and, in addition, for all $r \in \mathbb{R}$,

$$|\rho(r\mathbf{x}) - r\rho(\mathbf{x})|^2 = 0.$$

For the first,

$$|\rho(\mathbf{x}+\mathbf{y})-\rho(\mathbf{x})-\rho(\mathbf{y})|^2 = (\rho(\mathbf{x}+\mathbf{y})-\rho(\mathbf{x})-\rho(\mathbf{y}))\cdot(\rho(\mathbf{x}+\mathbf{y})-\rho(\mathbf{x})-\rho(\mathbf{y})),$$

so expanding one gets

$$|\rho(\mathbf{x}+\mathbf{y}) - \rho(\mathbf{x}) - \rho(\mathbf{y})|^2 = \rho(\mathbf{x}+\mathbf{y})\cdot\rho(\mathbf{x}+\mathbf{y})$$

$$-2\rho(\mathbf{x}+\mathbf{y})\cdot\rho(\mathbf{x}) - 2\rho(\mathbf{x}+\mathbf{y})\cdot\rho(\mathbf{y}) - \rho(\mathbf{x})\cdot\rho(\mathbf{x}) - \rho(\mathbf{y})\cdot\rho(\mathbf{y}).$$

But, by assumption, the right hand side is

$$(\mathbf{x}+\mathbf{y})\cdot(\mathbf{x}+\mathbf{y}) - 2(\mathbf{x}+\mathbf{y})\cdot\mathbf{x} - 2(\mathbf{x}+\mathbf{y})\cdot\mathbf{y} - \mathbf{x}\cdot\mathbf{x} - \mathbf{y}\cdot\mathbf{y},$$

which, as it is easy to show, clearly 0. The proof that $|\rho(r\mathbf{x}) - r\rho(\mathbf{x})|^2 = 0$ is similar and is left to the reader. □

In particular, every rotation ρ of $\mathbb{R}^3$ is an orthogonal linear transformation. However, not every orthogonal 3×3 matrix gives rise to a rotation. For example, a reflection of $\mathbb{R}^3$ through a plane through the origin clearly isn't a rotation, because if a rotation fixes two orthogonal vectors in $\mathbb{R}^3$, it fixes all of $\mathbb{R}^3$. On the other hand, a reflection does fix two orthogonal vectors without fixing $\mathbb{R}^3$. In fact, a reflection has eigenvalues1,1,-1, so the determinant of a reflection is -1.

I claim that every rotation ρ of $\mathbb{R}^3$ has a positive determinant. Indeed, ρ fixes a line L through the origin pointwise, so ρ has eigenvalue 1. Moreover, the plane orthogonal to L is rotated through an angle θ, so there exists an orthonormal basis of $\mathbb{R}^3$ for which the matrix of ρ has the form

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}.$$

Hence $\det(\rho) = 1$.

We now bring in $SO(3)$. Recall that $SL(3,\mathbb{R})$ denotes the set of all 3×3 real matrices of determinant 1. Put $SO(3) = SL(3,\mathbb{R}) \cap O(3,\mathbb{R})$. It follows that $\mathrm{Rot}(\mathbb{R}^3) \subset SO(3)$. In fact, we will now show

Theorem 10.12. $\mathrm{Rot}(\mathbb{R}^3) = SO(3)$.

Proof. We only need to show $SO(3) \subset \mathrm{Rot}(\mathbb{R}^3)$, i.e. every element of $SO(3)$ is a rotation. Note that by our definition, the identity transformation I_3 is a rotation. Namely I_3 is the rotation which fixes any line L through $\mathbf{0}$ and rotates every plane parallel to $L^\perp$ through zero degrees.

I claim that if $\sigma \in SO(3)$, then 1 is an eigenvalue of σ, and moreover, if $\sigma \neq I_3$, the eigenspace E_1 of 1 has dimension 1. That is, E_1 is a line.

We know that every 3×3 real matrix has a real eigenvalue, and we also know that the real eigenvalues of an orthogonal matrix are either 1 or -1. Hence, if $\sigma \in SO(3)$, the eigenvalues of σ are one of the following possibilities:

(i) 1 of multiplicity three,

(ii) $1, -1$, where -1 has multiplicity two, and

(iii) 1, λ, $\overline{\lambda}$, where $\lambda \neq \overline{\lambda}$ (since the complex roots of the characteristic polynomial of a real matrix occur in conjugate pairs).

Hence, 1 is always an eigenvalue of σ, so $\dim E_1 \geq 1$. I claim that if $\sigma \in SO(3)$ and $\sigma \neq I_3$, then $\dim E_1 = 1$. Indeed, $\dim E_1 = 3$, is impossible since $\sigma \neq I_3$. If $\dim E_2 = 2$, then σ fixes the plane E_2 pointwise. Since σ preserves angles, it also has to send the line $L = E_2^\perp$ to itself. Thus L is an eigenspace. But the only real eigenvalue different from 1 is -1, so if $\sigma \neq I_3$, there is a basis of $\mathbb{R}^3$ so that the matrix of σ is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

But then $\det(\sigma) = -1$, so $\dim E_1 = 2$ cannot happen. This gives the claim that $\dim E_1 = 1$ if $\sigma \neq I_3$.

Therefore σ fixes every point on a unique line L through the origin and maps the plane $L^\perp$ orthogonal to L into itself. We now need to show σ rotates $L^\perp$. Let $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ be an orthonormal basis in $\mathbb{R}^3$ such that $\mathbf{u}_1, \mathbf{u}_2 \in L^\perp$ and $\sigma(\mathbf{u}_3) = \mathbf{u}_3$. Let $Q = (\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3)$. Since $\sigma\mathbf{u}_1$ and $\sigma\mathbf{u}_2$ are orthogonal unit vectors on $L^\perp$, we can choose an angle θ such that

$$\sigma\mathbf{u}_1 = \cos\theta\mathbf{u}_1 + \sin\theta\mathbf{u}_2$$

and

$$\sigma\mathbf{u}_2 = \pm(\sin\theta\mathbf{u}_1 - \cos\theta\mathbf{u}_2).$$

In matrix terms, this says

$$\sigma Q = Q \begin{pmatrix} \cos\theta & \pm\sin\theta & 0 \\ \sin\theta & \mp\cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Since $\det(\sigma) = 1$ and $\det(Q) \neq 0$, it follows that

$$\det(\sigma) = \det \begin{pmatrix} \cos\theta & \pm\sin\theta & 0 \\ \sin\theta & \mp\cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} = 1.$$

The only possibility is that

$$\sigma = Q \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} Q^{-1}. \tag{10.16}$$

This tells us that σ rotates the plane $L^{\perp}$ through θ, hence $\sigma \in \mathrm{Rot}(\mathbb{R}^3)$. This completes the proof that $SO(3,\mathbb{R}) = \mathrm{Rot}(\mathbb{R}^3)$. □

We get a surprising conclusion.

Corollary 10.13. *$Rot(\mathbb{R}^3)$ is a matrix group. In particular, the composition of two rotations of $\mathbb{R}^3$ is another rotation.*

Proof. This is clear since $SO(3)$ is a matrix group: the product of two elements of $SO(3)$ is another element of $SO(3)$ and $SO(3)$ is closed under taking inverses. Indeed, $SO(3) = SL(3,\mathbb{R}) \cap O(3,\mathbb{R})$, and, by the product theorem for determinants, the product of two elements of $SL(3,\mathbb{R})$ is another element of $SL(3,\mathbb{R})$. Moreover, we also know that the product of two elements of $O(3,\mathbb{R})$ is also in $O(3,\mathbb{R})$. □

The fact that the composition of two rotations is a rotation is anything but obvious from the definition. For example, how does one describe the unique line fixed pointwise by the composition of two rotations?

Notices that the matrix Q defined above may be chosen so as to be a rotation. Therefore, the above argument gives another result.

Proposition 10.14. *The matrix of a rotation $\sigma \in SO(3)$ is similar via another rotation Q to a matrix of the form*

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

10.4.2 Rotation Groups of Solids

We begin with a definition.

Definition 10.5. Let S be a solid in $\mathbb{R}^3$. The *rotation group* of S is defined to be the set of all $\sigma \in SO(3)$ such that $\sigma(S) = S$. We denote the rotation group of S by $\mathrm{Rot}(S)$.

Proposition 10.15. *Let S be a solid in $\mathbb{R}^3$. If σ and τ are rotations of S, then so are $\sigma\tau$ and σ^{-1}. Hence the rotation group of S is a matrix group.*

Proof. Clearly $\sigma^{-1} \in SO(3)$. By Corollary 10.13, $\sigma\tau \in SO(3)$ as well. It's also clear that $\sigma\tau(S) = S$. Since $I_3 \in Rot(S)$ as well, the proof is finished. □

Example 10.8. Let S denote the cube with vertices at the points (A, B, C), where $A, B, C = \pm 1$. Let us find $Rot(S)$. Every rotation of $\mathbb{R}^3$ which maps S to itself maps each one of its six faces to another face. Moreover, since any face contains a basis of $\mathbb{R}^3$, each $\sigma \in Rot(S)$ is completely determined by how it acts on any face. Let F denote one of the faces. Given any one of the six faces F' of S, there is at least one σ such that $\sigma(F) = F'$. Furthermore, each face has four rotations, so by Proposition10.15, we have found that $rot(S)$ has at least 24 elements.

Now consider the 4 diagonals of S, i.e. the segments which join a vertex (A, B, C) to $(-A, -B, -C)$. Every rotation of S permutes these segments, and two rotations which define the same permutation of the diagonals coincide (why?). Since the number of permutations of 4 objects is $4! = 24$, it follows that $\mathrm{Rot}(S)$ has at most 24 elements. Therefore, there are exactly 24 rotations of S, and, incidentally, these 24 rotations are realized by the 24 permutations of the diagonals.

Example 10.9. Consider the set consisting of the midpoints of the 6 faces of the cube S. The solid polygon S' determined by these 6 points is called the regular octahedron. It is a solid with 8 triangular faces all congruent to each other. The cube and the regular octahedron are two of the 5 Platonic solids, which we will consider in Chapter 16. Since each element of $\mathrm{Rot}(S)$ must also send midpoint to another midpoint, it follows that $\mathrm{Rot}(S) \subset \mathrm{Rot}(S')$. But the other containment clearly also holds, so we deduce that $\mathrm{Rot}(S) = \mathrm{Rot}(S')$.

10.4.3 Reflections of $\mathbb{R}^3$

We now know that rotations of $\mathbb{R}^3$ are characterized by the property that their determinants are $+1$, and we know that the determinant of any element of $O(3,\mathbb{R})$ is ± 1. Hence every element of $O(3,\mathbb{R})$ that isn't a rotation has determinant -1. We also know that every orthogonal 2×2 matrix is either a rotation or a reflection: a rotation when the determinant is $+1$ and a reflection when the determinant is -1. A natural is whether this is also true in $O(3,\mathbb{R})$. It turns out that the determinant of a reflection of $\mathbb{R}^3$ is indeed -1. This is due to the fact that a reflection leaves a plane pointwise fixed and maps every vector orthogonal to the plane to its negative. Thus, for a reflection, $\dim E_1 = 2$ and $\dim E_{-1} = 1$, so the determinant is -1.

It turns out, however, that there exist elements $\sigma \in O(3,\mathbb{R})$ $\det(\sigma) = -1$ which are not reflections. For example, such a σ has eigenvalues $-1, \lambda, \overline{\lambda}$. It is left as an easy exercise to describe how σ acts on $\mathbb{R}^3$. As to reflections, we have the following fact.

Proposition 10.16. *An element $Q \in O(3,\mathbb{R})$ is a reflection if and only if Q is symmetric and* $\det(Q) = -1$.

We leave the proof as an exercise. It is useful to recall a reflection can be expressed as $I_3 - 2P_L$, where P_L is the projection on the line L orthogonal to the plane E_1 of the reflection. One final comment is that every reflection of $\mathbb{R}^2$ actually defines a rotation of $\mathbb{R}^3$. For if σ reflects $\mathbb{R}^2$ through a line L, the rotation ρ of $\mathbb{R}^3$ through π with L as the axis of rotation acts the same way as σ on $\mathbb{R}^2$, hence the claim. Note: the eigenvalues of ρ are $1, -1, -1$, that is -1 occurs with multiplicity two.

Remark: The abstract definition of the term group as in "rotation group" is given in Chapter 16. In essence, a group is a set that has a structure like that matrix group. In particular, elements can be multiplied, there is an identity and every element has an inverse.

Exercises

Exercise 10.44. Prove Proposition 10.16.

Exercise 10.45. Let S be a regular quadrilateral in $\mathbb{R}^3$, that is S has 4 faces made up of congruent triangles. How many elements does $\mathrm{Sym}(S)$ have?

Exercise 10.46. Compute $\mathrm{Rot}(S)$ in the following cases:

(a) S is the half ball $\{x^2 + y^2 + z^2 \leq 1, z \geq 0\}$, and

(b) S is the solid rectangle $\{-1 \leq x \leq 1, -2 \leq y \leq 2, -1 \leq z \leq 1\}$.

Chapter 11

The Diagonalization Theorems

Let V be a finite dimensional vector space and $T : V \to V$ be linear. The most basic question one can ask is whether T is semi-simple, that is, whether it admits an eigenbasis or, equivalently, whether T is represented by a diagonal matrix. The purpose of this chapter is to study this question. We will prove several theorems: the Principal Axis (or Spectral) Theorem, Schur's Theorem and the Jordan Decomposition Theorem. We will also draw a number of consequences of these results, such as the Cayley-Hamilton Theorem.

Our first topic is the Principal Axis Theorem, the fundamental result that says every Hermitian matrix admits a Hermitian orthonormal eigenbasis, hence is orthogonally diagonalizable. The reason this result is so important is that so many basic eigenvalue problems in mathematics or in the physical sciences involve real symmetric or or their complex analogues Hermitian matrices. In the real case, this result says that every symmetric matrix admits an orthonormal eigenbasis. As we will also point out, the general finite dimensional version of the Principal Axis Theorem is the result that every self adjoint linear transformation $T : V \to V$ on a finite dimensional inner product space V admits an orthogonal eigenbasis. In particular, T is semi-simple.

After we obtain the Principal Axis Theorem, we will extend it to normal matrices. We will then study linear transformations and matrices over an arbitrary algebraically closed field $\mathbb{F}$. It turns out there that every $T : V \to V$, equivalently every $A \in M_n(\mathbb{F})$, has a unique expression $T = S + N$ with S semi-simple, N nilpotent and $SN = NS$ with a corresponding version

for A. This fact is the Jordan Decomposition Theorem. It opens the door to a number of other results about linear transformations and matrices, for example the Cayley-Hamilton Theorem and the fact that T is semi-simple if and only if its minimal polynomial has simple roots.

11.1 The Principal Axis Theorem in the Real Case

In this section, we will show that all symmetric matrices $A \in \mathbb{R}^{n\times n}$ are orthogonally diagonalizable. More precisely, we will prove

Theorem 11.1. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric. Then all eigenvalues of A are real, and there exists an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of A. Consequently, there exists an orthogonal matrix Q such that*

$$A = QDQ^{-1} = QDQ^T,$$

where $D \in \mathbb{R}^{n\times n}$ is diagonal. Conversely, if $A = QDQ^{-1}$, where Q is orthogonal, then A is symmetric.

Notice that the last assertion is rather obvious since any matrix of the form CDC^T is symmetric, and $Q^{-1} = Q^T$ for all $Q \in O(n, \mathbb{R})$.

11.1.1 The Basic Properties of Symmetric Matrices

One of the problems in understanding symmetric matrices (or self adjoint operators in general) is to understand the geometric significance of the condition $a_{ij} = a_{ji}$. It turns out that there are several key geometric properties, the first of which is that every eigenvalue of a symmetric matrix is real. The second key fact is that two eigenvectors corresponding to different eigenvalues are orthogonal. These two facts are all we need for the first proof of the Principal Axis Theorem. We will also give a second proof which gives a more complete understanding of the geometric principles behind the result.

We will first formulate the condition $a_{ij} = a_{ji}$ in a more useful form.

Proposition 11.2. *$A \in \mathbb{R}^{n\times n}$ is symmetric if and only if*

$$\mathbf{v}^T A\mathbf{w} = \mathbf{w}^T A\mathbf{v}$$

for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$.

Proof. To see this, notice that since $\mathbf{v}^T A\mathbf{w}$ is a scalar, it equals its own transpose. Thus

$$\mathbf{v}^T A\mathbf{w} = (\mathbf{v}^T A\mathbf{w})^T = \mathbf{w}^T A^T\mathbf{v}.$$

So if $A = A^T$, then

$$\mathbf{v}^T A\mathbf{w} = \mathbf{w}^T A\mathbf{v}.$$

For the converse, use the fact that

$$a_{ij} = \mathbf{e}_i^T A\mathbf{e}_j,$$

so if $\mathbf{e}_i^T A\mathbf{e}_j = \mathbf{e}_j^T A\mathbf{e}_i$, then $a_{ij} = a_{ji}$. □

In other words, $A \in \mathbb{R}^{n\times n}$ is symmetric if and only if $T_A(\mathbf{v})\cdot\mathbf{w} = \mathbf{v}\cdot T_A(\mathbf{w})$ for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, where $T_A : \mathbb{R}^n \to \mathbb{R}^n$ is the linear transformation associated to A.

Definition 11.1. A linear map $T : \mathbb{R}^n \to \mathbb{R}^n$ is called *self adjoint* if and only if $T(\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot T(\mathbf{w})$ for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$.

In other words, symmetric matrices are the same as self adjoint linear transformations. We now establish the two basic properties mentioned above. First, we show

Proposition 11.3. *Eigenvectors of a real symmetric $A \in \mathbb{R}^{n\times n}$ corresponding to different eigenvalues are orthogonal.*

Proof. Let $\mathbf{u}$ and $\mathbf{v}$ be eigenvectors corresponding to distinct eigenvalues $\lambda \neq \mu$. Then

$$\mathbf{u}^T A\mathbf{v} = \mathbf{u}^T \mu\mathbf{v} = \mu\mathbf{u}^T\mathbf{v}$$

while

$$\mathbf{v}^T A\mathbf{u} = \mathbf{v}^T \lambda\mathbf{u} = \lambda\mathbf{v}^T\mathbf{u}.$$

Since A is symmetric, $\mathbf{u}^T A\mathbf{v} = \mathbf{v}^T A\mathbf{u}$, so $\lambda\mathbf{u}^T\mathbf{v} = \mu\mathbf{v}^T\mathbf{u}$. But $\mathbf{u}^T\mathbf{v} = \mathbf{v}^T\mathbf{u}$, so $(\lambda - \mu)\mathbf{u}^T\mathbf{v} = 0$. Since $\lambda \neq \mu$, we infer $\mathbf{u}^T\mathbf{v} = 0$, which finishes the proof. □

We next show that the second property.

Proposition 11.4. *All eigenvalues of a real symmetric matrix are real.*

Proof. Not unexpectedly, this proof requires complex numbers. In fact, we will establish a general identity also needed in the in the Hermitian case.

Lemma 11.5. *If $A \in \mathbb{R}^{n\times n}$ is symmetric, then*

$$\overline{\mathbf{v}}^T A\mathbf{v} \in \mathbb{R}$$

for all $\mathbf{v} \in \mathbb{C}^n$.

Proof. Recall that $\alpha \in \mathbb{C}$ is real if and only if $\overline{\alpha} = \alpha$. Keeping in mind that $A \in \mathbb{R}^{n\times n}$, $\overline{\alpha\beta} = \overline{\alpha}\,\overline{\beta}$ for all $\alpha, \beta \in \mathbb{C}$ and $\overline{\mathbf{v}}^T A\mathbf{v} \in \mathbb{C}$, we see that

$$\begin{aligned}\overline{\overline{\mathbf{v}}^T A\mathbf{v}} &= \mathbf{v}^T\overline{A}\overline{\mathbf{v}} \\ &= \left(\mathbf{v}^T A\overline{\mathbf{v}}\right)^T \\ &= \overline{\mathbf{v}}^T A^T\mathbf{v}.\end{aligned}$$

Since $A = A^T$, it follows that $\overline{\overline{\mathbf{v}}^T A\mathbf{v}} = \overline{\mathbf{v}}^T A\mathbf{v}$, so the proof is finished. □

Now let $A \in \mathbb{R}^{n\times n}$ be symmetric. Since the characteristic polynomial of A is a real polynomial of degree n, the Fundamental Theorem of Algebra implies it has n roots in $\mathbb{C}$. Suppose that $\lambda \in \mathbb{C}$ is a root. Then there exists a $\mathbf{v} \neq \mathbf{0}$ in $\mathbb{C}^n$ so that $A\mathbf{v} = \lambda\mathbf{v}$. Hence

$$\overline{\mathbf{v}}^T A\mathbf{v} = \overline{\mathbf{v}}^T\lambda\mathbf{v} = \lambda\overline{\mathbf{v}}^T\mathbf{v}.$$

We may obviously assume $\lambda \neq 0$, so the right hand side is nonzero. Indeed, if $\mathbf{v} = (v_1, v_2, \ldots v_n)^T \neq \mathbf{0}$, then

$$\overline{\mathbf{v}}^T\mathbf{v} = \sum_{i=1}^{n} \overline{v_i}v_i = \sum_{i=1}^{n} |v_i|^2 > 0.$$

Since $\overline{\mathbf{v}}^T A\mathbf{v} \in \mathbb{R}$, λ is a quotient of two reals, so $\lambda \in \mathbb{R}$. This completes the proof of the Proposition. □

11.1.2 Some Examples

Example 11.1. Let H denote a 2×2 reflection matrix. Then H has eigenvalues ± 1. Either unit vector $\mathbf{u}$ on the reflecting line together with either unit vector $\mathbf{v}$ orthogonal to the reflecting line form an orthonormal eigenbasis of $\mathbb{R}^2$ for H. Thus $Q = (\mathbf{u}\ \mathbf{v})$ is orthogonal and

$$H = Q\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}Q^{-1} = Q\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}Q^T.$$

Note that there are only four possible choices for Q. All 2×2 reflection matrices are similar to $\mathrm{diag}[1, -1]$. The only thing that can vary is Q.

Here is another example.

Example 11.2. Let B be the all ones 4×4 matrix. We already saw that the eigenvalues of B are 0 and 4. The eigenspace $E_0 = \mathcal{N}(B) = (\mathbb{R}(1, 1, 1, 1)^T)^\perp$, so to check the Principal Axis Theorem, we need to find an orthonormal

basis of $(\mathbb{R}(1,1,1,1)^T)^\perp$. We will do this by inspection rather than Gram-Schmidt, since it is easy to find vectors orthogonal to $(1,1,1,1)^T$. In fact, $\mathbf{v}_1 = (1,-1,0,0)^T$, $\mathbf{v}_2 = (0,0,1,-1)^T$, and $\mathbf{v}_3 = (1,1,-1,-1)^T$ give an orthonormal basis after we normalize. Then $\mathbf{v}_4 = (1,1,1,1)^T \in E_4$ is a fourth eigenvector, which, by our construction is orthogonal to E_0. Of course the real reason $\mathbf{v}_4$ is orthogonal to E_0 is that 0 and 4 are distinct eigenvalues of the symmetric matrix B. Thus we get the following expression for B as QDQ^T:

$$\frac{1}{4}\begin{pmatrix} 2\sqrt{2} & 0 & 0 & 1 \\ -2\sqrt{2} & 0 & 1 & 1 \\ 1 & 2\sqrt{2} & -1 & 1 \\ 0 & -2\sqrt{2} & -1 & 1 \end{pmatrix}\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}\begin{pmatrix} 2\sqrt{2} & -2\sqrt{2} & 1 & 0 \\ 0 & 0 & 2\sqrt{2} & -2\sqrt{2} \\ 0 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

This is somewhat amusing since 3 of D's diagonal entries are 0.

11.1.3 The First Proof

Since we know that all eigenvalues of a symmetric matrix A are real, and eigenvectors corresponding to different eigenvalues are orthogonal, there is nothing to prove when all the eigenvalues are distinct. The difficulty is that if A has repeated eigenvalues, say $\lambda_1, \ldots \lambda_m$, then one has to show

$$\sum_{i=1}^{m} \dim E_{\lambda_i} = n.$$

Our first proof doesn't actually require us to overcome this difficulty. It rests on the group theoretic property that the product of two orthogonal matrices is orthogonal.

To keep the notation simple and since we will also give a second proof, let us just do the 3×3 case, which, in fact, involves all the essential ideas. Thus let A be real 3×3 symmetric, and consider an eigenpair $(\lambda_1, \mathbf{u}_1)$ where $\mathbf{u}_1 \in \mathbb{R}^3$ is a unit vector. By the Gram-Schmidt process, we can include $\mathbf{u}_1$ in an orthonormal basis $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$ of $\mathbb{R}^3$. Let $Q_1 = (\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3)$. Then Q_1 is orthogonal and

$$AQ_1 = (A\mathbf{u}_1\ A\mathbf{u}_2\ A\mathbf{u}_3) = (\lambda_1\mathbf{u}_1\ A\mathbf{u}_2\ A\mathbf{u}_3).$$

Now

$$Q_1^T AQ_1 = \begin{pmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \mathbf{u}_3^T \end{pmatrix}(\lambda_1\mathbf{u}_1\ A\mathbf{u}_2\ A\mathbf{u}_3) = \begin{pmatrix} \lambda_1\mathbf{u}_1^T\mathbf{u}_1 & * & * \\ \lambda_1\mathbf{u}_2^T\mathbf{u}_1 & * & * \\ \lambda_1\mathbf{u}_3^T\mathbf{u}_1 & * & * \end{pmatrix}.$$

But since $Q_1^T A Q_1$ is symmetric (since A is), and since $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$ are orthonormal, we see that

$$Q_1^T A Q_1 = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & * & * \\ 0 & * & * \end{pmatrix}.$$

It is clear that the 2×2 matrix in the lower right hand corner of A is symmetric. Calling this matrix B, we can find, by repeating the construction just given, a 2×2 orthogonal matrix Q' so that

$$Q'^T B Q' = \begin{pmatrix} \lambda_2 & 0 \\ 0 & \lambda_3 \end{pmatrix}.$$

Putting $Q' = \begin{pmatrix} r & s \\ t & u \end{pmatrix}$, it follows that

$$Q_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r & s \\ 0 & t & u \end{pmatrix}$$

is orthogonal, and in addition

$$Q_2^T Q_1^T A Q_1 Q_2 = Q_2^T \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & * & * \\ 0 & * & * \end{pmatrix} Q_2 = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}.$$

But $Q_1 Q_2$ is orthogonal, and so $(Q_1 Q_2)^{-1} = Q_2^{-1} Q_1^{-1} = Q_2^T Q_1^T$. Therefore, putting $Q = Q_1 Q_2$ and $D = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$, we get $A = QDQ^{-1} = QAQ^T$, so A has been orthogonally diagonalized and the first proof is done. □

Note, using mathematical induction, this proof extends immediately to prove the general case. The drawback of the above technique is that it requires a repeated application of the Gram-Schmidt process. What is nice is that the argument is completely transparent.

11.1.4 Proof Number Two

Our second proof, more geometric proof, seems to me to give more insight. It relies on the two basic properties of symmetric matrices discussed above and also one more.

Proposition 11.6. *If $A \in \mathbb{R}^{n \times n}$ is symmetric and W is a nonzero subspace of $\mathbb{R}^n$ with the property that $T_A(W) = AW \subset W$, then W contains an eigenvector of A.*

Proof. Pick an orthonormal basis $\mathcal{Q} = \{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ of W. As $T_A(W) \subset W$, there exist scalars r_{ij} $(1 \leq i, j \leq m)$, such that

$$T_A(\mathbf{u}_j) = A\mathbf{u}_j = \sum_{i=1}^{m} r_{ij}\mathbf{u}_i.$$

This defines an $m \times m$ matrix R, which I claim is symmetric. Indeed, since T_A is self adjoint,

$$r_{ij} = \mathbf{u}_i \cdot A\mathbf{u}_j = A\mathbf{u}_i \cdot \mathbf{u}_j = r_{ji}.$$

Let $(\lambda, (x_1, \ldots, x_m)^T)$ be an eigenpair for R. Putting $\mathbf{w} := \sum_{j=1}^{m} x_j\mathbf{u}_j$, I claim that $(\lambda, \mathbf{w})$ forms an eigenpair for A. In fact,

$$\begin{aligned} T_A(\mathbf{w}) &= \sum_{j=1}^{m} x_j T_A(\mathbf{u}_j) \\ &= \sum_{j=1}^{m} x_j (\sum_{i=1}^{m} r_{ij}\mathbf{u}_i) \\ &= \sum_{i=1}^{m} (\sum_{j=1}^{m} r_{ij}x_j)\mathbf{u}_i \\ &= \sum_{i=1}^{m} \lambda x_i \mathbf{u}_i \\ &= \lambda\mathbf{w}. \end{aligned}$$

This finishes the proof. □

We can now complete the second proof. As above, let $\lambda_1, \ldots \lambda_k$ be the distinct eigenvalues of A. Then the corresponding eigenspaces E_{λ_i} are orthogonal to each other, so if we combine orthonormal bases of each E_{λ_i}, we obtain an orthonormal set in $\mathbb{R}^n$. Putting $\dim E_{\lambda_i} = k_i$, it follows the span of this orthonormal set is a subspace E of $\mathbb{R}^n$ of dimension $e := \sum_{i=1}^{k} k_i$. Combining all the eigenvectors gives an orthonormal basis of E. To show $e = n$, let $W = E^{\perp}$. If $E \neq \mathbb{R}^n$, then by Proposition 10.1, $\dim W = n - e > 0$. Clearly, $T_A(E_{\lambda_i}) \subset E_{\lambda_i}$ (check). Hence $T_A(E) \subset E$. I claim that we also have that $T_A(W) \subset W$. To see this, we have to show that if $\mathbf{v} \in E$ and $\mathbf{w} \in W$, then $\mathbf{v} \cdot A\mathbf{w} = \mathbf{v}^T A\mathbf{w} = 0$. But since $T_A(E) \subset E$ and A is symmetric,

$$\mathbf{v}^T A\mathbf{w} = \mathbf{w}^T A\mathbf{v} = 0.$$

Hence the claim. But W has positive dimension, so it contains an eigenvector $\mathbf{w}$ of A. But all eigenvectors of A are, by definition, elements of E, so E

and W both contain a non zero vector. This contradicts Proposition10.1, so $W = \{\mathbf{0}\}$. Therefore, $E = \mathbb{R}^n$ and the second proof is done. □

Note also that in the course of the proof of the Principal Axis Theorem, we discovered a fourth interesting geometric property of symmetric matrices:

Proposition 11.7. *If $A \in \mathbb{R}^{n\times n}$ is symmetric and $A(W) \subset W$ for a subspace W of $\mathbb{R}^n$, then $A(W^\perp) \subset W^\perp$ too.*

11.1.5 A Projection Formula for Symmetric Matrices

Sometimes it's useful to express the Principal Axis Theorem as a projection formula for symmetric matrices. Let A be symmetric, let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be an orthonormal eigenbasis of $\mathbb{R}^n$ for A, and suppose $(\lambda_i, \mathbf{u}_i)$ is an eigenpair. Suppose $\mathbf{x} \in \mathbb{R}^n$. By the projection formula of Chapter 10,

$$\mathbf{x} = (\mathbf{u}_1^T\mathbf{x})\mathbf{u}_1 + \cdots + (\mathbf{u}_n^T\mathbf{x})\mathbf{u}_n,$$

hence

$$A\mathbf{x} = \lambda_1(\mathbf{u}_1^T\mathbf{x})\mathbf{u}_1 + \cdots + \lambda_n(\mathbf{u}_n^T\mathbf{x})\mathbf{u}_n.$$

This amounts to writing

$$A = \lambda_1\mathbf{u}_1\mathbf{u}_1^T + \cdots + \lambda_n\mathbf{u}_n\mathbf{u}_n^T. \tag{11.1}$$

Recall that $\mathbf{u}_i\mathbf{u}_i^T$ is the matrix of the projection of $\mathbb{R}^n$ onto the line $\mathbb{R}\mathbf{u}_i$, so (11.1) expresses A as a sum of orthogonal projections.

Exercises

Exercise 11.1. Orthogonally diagonalize the following matrices:

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 3 \\ 1 & 3 & 1 \\ 3 & 1 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

I claim that you can diagonalize the first and second matrices, and a good deal (if not all) of the third, without pencil and paper.

Exercise 11.2. Prove Proposition11.7.

Exercise 11.3. Answer either T or F. If T, give a brief reason. If F, give a counter example.

(a) The sum and product of two symmetric matrices is symmetric.

(b) For any real matrix A, the eigenvalues of A^TA are all real.

(c) For A as in (b), the eigenvalues of A^TA are all non negative.

(d) If two symmetric matrices A and B have the same eigenvalues, counting multiplicities, then A and B are orthogonally similar (that is, $A = QBQ^T$ where Q is orthogonal).

Exercise 11.4. Recall that two matrices A and B which have a common eigenbasis commute. Conclude that if A and B have a common eigenbasis and are symmetric, then AB is symmetric.

Exercise 11.5. Describe the orthogonal diagonalization of a reflection matrix.

Exercise 11.6. Let W be a hyperplane in $\mathbb{R}^n$, and let H be the reflection through W.

(a) Express H in terms of P_W and $P_{W^\perp}$.

(b) Show that $P_W P_{W^\perp} = P_{W^\perp} P_W$.

(c) Simultaneously orthogonally diagonalize P_W and $P_{W^\perp}$.

Exercise 11.7. * Diagonalize

$$\begin{pmatrix} a & b & c \\ b & c & a \\ c & a & b \end{pmatrix},$$

where a, b, c are all real. (Note that the second matrix in Problem 1 is of this type. What does the fact that the trace is an eigenvalue say?)

Exercise 11.8. * Diagonalize

$$A = \begin{pmatrix} aa & ab & ac & ad \\ ba & bb & bc & bd \\ ca & cb & cc & cd \\ da & db & dc & dd \end{pmatrix},$$

where a, b, c, d are arbitrary real numbers. (Note: thimk!)

Exercise 11.9. Prove that a real symmetric matrix A whose only eigenvalues are ± 1 is orthogonal.

Exercise 11.10. Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric. Show the following.

(i) $\mathcal{N}(A)^{\perp} = \mathrm{Im}(A)$.

(ii) $\mathrm{Im}(A)^{\perp} = \mathcal{N}(A)$.

(iii) $\mathrm{col}(A) \cap \mathcal{N}(A) = \{\mathbf{0}\}$.

(iv) Conclude form (iii) that if $A^k = O$ for some $k > 0$, then $A = O$.

Exercise 11.11. Give a proof of the Principal Axis Theorem from first principles in the 2×2 case.

Exercise 11.12. Show that two symmetric matrices A and B that have the same characteristic polynomial are orthogonally similar. That is, $A = QBQ^{-1}$ for some orthogonal matrix Q.

Exercise 11.13. Let $A \in \mathbb{R}^{n\times n}$ be symmetric, and let λ_m and λ_M be its minimum and maximum eigenvalues respectively.

(a) Show that for every $\mathbf{x} \in \mathbb{R}^n$, we have

$$\lambda_m \mathbf{x}^T\mathbf{x} \le \mathbf{x}^T A\mathbf{x} \le \lambda_M \mathbf{x}^T\mathbf{x}.$$

(b) Use this inequality to find the maximum and minimum values of $|A\mathbf{x}|$ on the ball $|\mathbf{x}| \le 1$.

11.2 Self Adjoint Maps

The purpose of this section is to formulate the Principal Axis Theorem for an arbitrary finite dimensional inner product space V. In order to do this, we have to make some preliminary comments about this class of spaces.

11.2.1 Inner Product Spaces and Isometries

Recall that a real vector space$V$ with an inner product is called an inner product space. Simple examples include of course $\mathbb{R}^n$ with the dot product and $C[a,b]$ with the inner product $(f,g) = \int_a^b f(x)g(x)\mathrm{d}x$. In fact, we have

Proposition 11.8. *Every finite dimensional vector space over $\mathbb{R}$ admits an inner product.*

Proof. Select a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of V. Then, if $\mathbf{x} = \sum x_i\mathbf{v}_i$ and $\mathbf{y} = \sum y_j\mathbf{v}_j$, put

$$(\mathbf{x}, \mathbf{y}) = \sum x_i y_i.$$

Then it is easy to see that $(\ ,\ )$ is an inner product on V. □

Let V be a finite dimensional inner product space, and let $(\ ,\ )$ denote its inner product. In the above proof, the inner product is defined so that $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is an orthonormal basis. On the other hand, the techniques of the previous chapter extend immediately to an arbitrary finite dimensional inner product space, so we have the analogue of Proposition10.6: every finite dimensional inner product space has an orthonormal basis.

Definition 11.2. Let U and V be finite dimensional inner product spaces, and suppose $\Phi : U \to V$ is a transformation such that

$$(\Phi(\mathbf{x}), \Phi(\mathbf{y})) = (\mathbf{x}, \mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in U$. Then Φ is called an *isometry*.

The notion of an isometry was briefly touched on in Chapter 3.3. In particular, we showed that every element of $O(n, \mathbb{R})$ defines an isometry: see (3.8).

Proposition 11.9. *Let U and V be finite dimensional inner product spaces of the same positive dimension. Then every isometry $\Phi : U \to V$ is a linear isomorphism. Moreover, $\Phi : U \to V$ is an isometry if and only if Φ carries any orthonormal basis of U to an orthonormal basis of V. In particular, any inner product space of dimension n is isometric to $\mathbb{R}^n$.*

Proof. To see that $\Phi(\mathbf{x}+\mathbf{y}) = \Phi(\mathbf{x}) + \Phi(\mathbf{y})$, it suffices to show

$$(\Phi(\mathbf{x}+\mathbf{y}) - \Phi(\mathbf{x}) - \Phi(\mathbf{y}), \Phi(\mathbf{x}+\mathbf{y}) - \Phi(\mathbf{x}) - \Phi(\mathbf{y}) = 0,$$

The proof of this is exactly the same as in the proof of Proposition10.11. One sees that $\Phi(r\mathbf{x}) = r\Phi(\mathbf{x})$ in a similar way. Hence, every isometry is linear. An isometry Φ is also one to one since if $\mathbf{x} \neq \mathbf{0}$, then $\Phi(\mathbf{x}) \neq \mathbf{0}$ (why?). Hence Φ is an isomorphism since it is one to one and $\dim U = \dim V$. We leave the rest of the proof as an exercise. □

A couple of more comments about isometries are in order. First, the matrix of an isometry $\Phi : U \to U$ with respect to an orthonormal basis is orthogonal. (See the comment above about $O(n, \mathbb{R})$.) Also, since isometries preserve inner products, so they also preserve lengths and angles and angles between vectors.

11.2.2 Self Adjoint Operators

In the previous section, we defined the notion of a self adjoint linear map $T : \mathbb{R}^n \to \mathbb{R}^n$. The notion of a self adjoint operator on an arbitrary inner product space is exactly the same. We will treat this in a slightly more general way, however. First we make the following definition.

Definition 11.3. Let V be a real inner product space and suppose $T : V \to V$ is linear. Define the *adjoint* of T to be the map $T^* : V \to V$ determined by the condition that

$$(T^*(\mathbf{x}), \mathbf{y}) = (\mathbf{x}, T(\mathbf{y}))$$

for all $\mathbf{x}, \mathbf{y} \in V$. Then we say that T is *self adjoint* if and only if $T = T^*$.

Proposition 11.10. *Let V be a real inner product space and suppose $T : V \to V$ is linear. Then the adjoint $T^* : V \to V$ is also a well defined linear transformation. If V is finite dimensional, then T is self adjoint if and only if for every orthonormal basis $\mathcal{Q}$ of V, the matrix $\mathcal{M}_{\mathcal{Q}}^{\mathcal{Q}}(T)$ is symmetric. More generally, the matrix $\mathcal{M}_{\mathcal{Q}}^{\mathcal{Q}}(T^*)$ is $\mathcal{M}_{\mathcal{Q}}^{\mathcal{Q}}(T)^T$.*

Proof. The proof is left as an exercise. □

Hence a symmetric matrix is a self adjoint linear transformation from $\mathbb{R}^n$ to itself and conversely. Therefore the eigenvalue problem for self adjoint maps on a finite dimensional inner product space reduces to the eigenvalue problem for symmetric matrices on $\mathbb{R}^n$.

Here is a familiar example.

Example 11.3. Let W be a subspace of $\mathbb{R}^n$. Then the projection P_W is self adjoint. In fact, we know that its matrix with respect to the standard basis has the form $C(CC^T)^{-1}C^T$, which is clearly symmetric. Another way to see the self adjointness is to choose an orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of $\mathbb{R}^n$ so that $\mathbf{u}_1, \ldots, \mathbf{u}_m$ span W. Then, by the projection formula, $P_W(\mathbf{x}) = \sum_{i=1}^{k} (\mathbf{x} \cdot \mathbf{u}_i)\mathbf{u}_i$. It follows easily that $P_W(\mathbf{u}_i) \cdot \mathbf{u}_j = \mathbf{u}_i \cdot P_W(\mathbf{u}_j)$ for all indices i and j. Hence P_W is self adjoint.

To summarize the Principal Axis Theorem for self adjoint operators, we state

Theorem 11.11. *Let V be a finite dimensional inner product space, and let $T : V \to V$ be self adjoint. Then there exists an orthonormal eigenbasis $\mathcal{Q}$ of V consisting of eigenvectors of T. Thus T is semi-simple, and the matrix $\mathcal{M}_{\mathcal{Q}}^{\mathcal{Q}}(T)$ is diagonal.*

Proof. Let $\mathcal{B} = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ denote an orthonormal basis of V. The map $\Phi : V \to \mathbb{R}^n$ defined by $\Phi(\mathbf{u}_i) = \mathbf{e}_i$ is an isometry (see Proposition10.11). Now $S = \Phi T \Phi^{-1}$ is a self adjoint map of $\mathbb{R}^n$ (check), hence S has an orthonormal eigenbasis $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^n$. Since Φ is an isometry, $\mathbf{v}_1 = \Phi^{-1}(\mathbf{x}_1), \ldots, \mathbf{v}_n = \Phi^{-1}(\mathbf{x}_n)$ form an orthonormal basis of V. Moreover, the $\mathbf{v}_i$ are eigenvectors of T. For, if $S(\mathbf{x}_i) = \lambda_i \mathbf{x}_i$, then

$$T(\mathbf{v}_i) = \Phi^{-1} S \Phi(\mathbf{v}_i) = \Phi^{-1} S(\mathbf{x}_i) = \Phi^{-1}(\lambda_i \mathbf{x}_i) = \lambda_i \Phi^{-1}(\mathbf{x}_i) = \lambda_i \mathbf{v}_i.$$

Thus T admits an orthonormal eigenbasis, as claimed. □

11.2.3 An Infinite Dimensional Self Adjoint Operator

We now give an example of a self adjoint operator (or linear transformation) in the infinite dimensional setting. As mentioned in the introduction, self adjoint operators are frequently encountered in mathematical, as well as physical, problems.

We will consider a certain subspace of function space $C[a, b]$ of all continuous functions $f : [a, b] \to \mathbb{R}$ with the usual inner product

$$(f, g) = \int_a^b f(x)g(x)\mathrm{d}x.$$

The condition for a linear transformation $T : C[a, b] \to C[a, b]$ to be self adjoint is that satisfies the condition $(Tf, g) = (f, Tg)$ for all f, g, that is

$$\int_a^b T(f)(x)g(x)\mathrm{d}x = \int_a^b f(x)T(g)(x)\mathrm{d}x.$$

Now let $[a,b] = [0,2\pi]$, and let $\mathcal{P}$ (for periodic) denote the subspace of $C[0,2\pi]$ consisting of all functions f which have derivatives of all orders on $[0,2\pi]$ and satisfy the further condition that

$$f^{(i)}(0) = f^{(i)}(2\pi) \qquad \text{if} \qquad i = 0,1,2,\ldots,$$

where $f^{(i)}$ denotes the ith derivative of f. Among the functions in $\mathcal{P}$ are the trigonometric functions $\cos\lambda x$ and $\sin\lambda x$ for all $\lambda \in \mathbb{R}$. We will show below that these functions are linearly independent if $\lambda > 0$, so $\mathcal{P}$ is an infinite dimensional space.

We next give an example of a self adjoint operator on $\mathcal{P}$. Thus symmetric matrices can have infinite dimensional analogues. By the definition of $\mathcal{P}$, it is clear that if $f \in \mathcal{P}$, then $f^{(i)} \in \mathcal{P}$ for all $i \geq 1$. Hence the derivative operator $D(f) = f'$ defines a linear transformation $D : \mathcal{P} \to \mathcal{P}$. I claim the second derivative $D^2(f) = f''$ is self adjoint. To prove this, we have to show $(D^2(f), g) = (f, D^2(g))$ for all $f, g \in \mathcal{P}$. This follows from integration by parts. For we have

$$\begin{aligned}(D^2(f), g) &= \int_0^{2\pi} f''(t)g(t)\mathrm{d}t \\ &= f'(2\pi)g(2\pi) - f'(0)g(0) - \int_0^{2\pi} f'(t)g'(t)\mathrm{d}t.\end{aligned}$$

But by the definition of $\mathcal{P}$, $f'(2\pi)g(2\pi) - f'(0)g(0) = 0$, so

$$(D^2(f), g) = -\int_0^{2\pi} f'(t)g'(t)\mathrm{d}t.$$

Since this expression for $(D^2(f), g)$ is symmetric in f and g, it follows that

$$(D^2(f), g) = (f, D^2(g)),$$

so D^2 is self adjoint, as claimed.

We can now ask for the eigenvalues and corresponding eigenfunctions of D^2. There is no general method for finding the eigenvalues of a linear operator on an infinite dimensional space, but one can easily see that the trig functions $\cos\lambda x$ and $\sin\lambda x$ are eigenfunctions for $-\lambda^2$ if $\lambda \neq 0$. Now there is a general theorem in differential equations that asserts that if $\mu > 0$, then any solution of the equation

$$D^2(f) + \mu f = 0$$

has the form $f = a\cos\sqrt{\mu}x + b\sin\sqrt{\mu}x$ for some $a, b \in \mathbb{R}$. Moreover, $\lambda = 0$ is an eigenvalue for eigenfunction $1 \in \mathcal{P}$. Note that although $D^2(x) = 0$, x is not an eigenfunction since $x \notin \mathcal{P}$.

To summarize, D^2 is a self adjoint operator on $\mathcal{P}$ such that every non positive real number is an ev. The corresponding eigenspaces are $E_0 = \mathbb{R}$ and $E_{-\lambda} = \mathbb{R}\cos\sqrt{\lambda}x + \mathbb{R}\sin\sqrt{\lambda}x$ if $\lambda > 0$. We can also draw some other consequences. For any positive $\lambda_1, \ldots \lambda_k$ and any $f_i \in E_{-\lambda_i}$, $f_1, \ldots, f_k$ are linearly independent. Therefore, the dimension of $\mathcal{P}$ cannot be finite, i.e. $\mathcal{P}$ is infinite dimensional.

Recall that distinct eigenvalues of a symmetric matrix have orthogonal eigenspaces. Thus distinct eigenvalues of a self adjoint linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ have orthogonal eigenspaces. The proof of this goes over unchanged to $\mathcal{P}$, so if $\lambda, \mu > 0$ and $\lambda \neq \mu$, then

$$\int_0^{2\pi} f_\lambda(t) f_\mu(t) \mathrm{d}t = 0,$$

where f_λ and f_μ are any eigenfunctions for $-\lambda$ and $-\mu$ respectively. In particular,

$$\int_0^{2\pi} \sin\sqrt{\lambda}t \sin\sqrt{\mu}t \mathrm{d}t = 0,$$

with corresponding identities for the other pairs of eigenfunctions f_λ and f_μ. In addition, $\cos\sqrt{\lambda}x$ and $\sin\sqrt{\lambda}x$ are also orthogonal.

The next step is to normalize the eigenfunctions to obtain an orthonormal set. Clearly if $\lambda \neq 0$, $|f_\lambda|^2 = (f_\lambda, f_\lambda) = \pi$, while $|f_0|^2 = 2\pi$. Hence the functions

$$\frac{1}{\sqrt{2\pi}}, \quad \frac{1}{\sqrt{\pi}}\cos\sqrt{\lambda}x, \quad \frac{1}{\pi}\sin\sqrt{\lambda}x,$$

where $\lambda > 0$ are a family of ON functions in $\mathcal{P}$. It turns out that one usually considers only the eigenfunctions where λ is a positive integer. The *Fourier series* of a function $f \in C[0, 2\pi]$ such that $f(0) = f(2\pi)$ is the infinite series development

$$f(x) \approx \frac{1}{\pi}\sum_{m=1}^{\infty} a_m \cos mx + \frac{1}{\pi}\sum_{m=1}^{\infty} b_m \sin mx,$$

where a_m and b_m are the Fourier coefficients encountered in §33. In particular,

$$a_m = \frac{1}{\sqrt{\pi}}\int_0^{2\pi} f(t)\cos mt \mathrm{d}t$$

and

$$b_m = \frac{1}{\sqrt{\pi}} \int_0^{2\pi} f(t) \sin mt \mathrm{d}t.$$

For a precise interpretation of the meaning $\approx$, we refer to a text on Fourier series. The upshot of this example is that Fourier series are an important tool in partial differential equations, mathematical physics and many other areas.

Exercises

Exercise 11.14. Show that if V is a finite dimensional inner product space, then $T \in L(V)$ is self adjoint if and only if for every orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$ OF V, $(T(\mathbf{u}_i), \mathbf{u}_j) = (\mathbf{u}_i, T(\mathbf{u}_j))$ for all indices i and j.

Exercise 11.15. Let U and V be inner product spaces of the same dimension. Show that a linear transformation $\Phi : U \to V$ is an isometry if and only if Φ carries every orthonormal basis of U onto an orthonormal basis of V.

Exercise 11.16. Give an example of a linear transformation$\Phi : \mathbb{R}^2 \to \mathbb{R}^2$ that isn't an isometry.

Exercise 11.17. Show that the matrix of an isometry $\Phi : U \to U$ with respect to and orthonormal basis is orthogonal. Conversely show that given an orthonormal basis, any orthogonal matrix defines an isometry form U to itself.

Exercise 11.18. Give the proof of Proposition11.10 .

Exercise 11.19. Describe all isometries $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$.

Exercise 11.20. Prove Proposition11.10.

11.3 The Principal Axis Theorem Hermitian Matrices

The purpose of this section is to extend the Principal Axis Theorem to Hermitian matrices, which is one of the fundamental mathematical facts needed in quantum theory.

11.3.1 Hermitian Inner Products and Hermitian Matrices

Let $A = (\alpha_{ij}) \in \mathbb{C}^{n\times n}$.

Definition 11.4. The *Hermitian transpose* of $A = (\alpha_{ij})$ is defined to be the matrix

$$A^H := \overline{A}^T,$$

where $\overline{A}$ is the matrix $(\overline{\alpha_{ij}})$ obtained by conjugating the entries of A.

Definition 11.5. A square matrix A over $\mathbb{C}$ is called *Hermitian* if $A^H = A$. In other words, A is Hermitian if and only if $\alpha_{ij} = \overline{\alpha_{ji}}$.

Example 11.4. For example,

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

is Hermitian.

Clearly, real Hermitian matrices are symmetric, and Hermitian matrices have real diagonal entries.

To extend the Principal Axis Theorem, we need to modify our notion of an inner product to take in vector spaces over $\mathbb{C}$. Indeed, the usual real inner product extended to $\mathbb{C}^n$ has the problem that non-zero vectors may have negative or zero length, e.g. $(1, i)^T$. Instead, we introduce the *Hermitian inner product* on $\mathbb{C}^n$. For $\mathbf{w}, \mathbf{z} \in \mathbb{C}^n$, put

$$\mathbf{w} \cdot \mathbf{z} := \overline{w_1} z_1 + \overline{w_2} z_2 + \cdots + \overline{w_n} z_n.$$

In other words,

$$\mathbf{w} \cdot \mathbf{z} := \overline{\mathbf{w}}^T \mathbf{z} = \mathbf{w}^H \mathbf{z}.$$

Obviously the Hermitian inner product coincides with the usual inner product if $\mathbf{w}, \mathbf{z} \in \mathbb{R}^n \subset \mathbb{C}^n$. It is easy to see that the distributivity property still

holds: that is, $\mathbf{w}\cdot(\mathbf{x}+\mathbf{y}) = \mathbf{w}\cdot\mathbf{x}+\mathbf{w}\cdot\mathbf{y}$, and $(\mathbf{w}+\mathbf{x})\cdot\mathbf{y} = \mathbf{w}\cdot\mathbf{y}+\mathbf{x}\cdot\mathbf{y}$. However, the scalar multiplication properties are slightly different. For example,

$$(\alpha\mathbf{w})\cdot\mathbf{z} = \overline{\alpha}(\mathbf{w}\cdot\mathbf{z}),$$

but

$$\mathbf{w}\cdot(\alpha\mathbf{z}) = \alpha(\mathbf{w}\cdot\mathbf{z}).$$

Another difference is that

$$\mathbf{z}\cdot\mathbf{w} = \overline{\mathbf{w}\cdot\mathbf{z}}.$$

The length of $\mathbf{z}\in\mathbb{C}^n$ can be written in several ways:

$$|\mathbf{z}| := (\mathbf{z}\cdot\mathbf{z})^{1/2} = (\mathbf{z}^H\mathbf{z})^{1/2} = \Big(\sum_{i=1}^{n}|z_i|^2\Big)^{1/2}.$$

Note that $|\mathbf{z}|>0$ unless $\mathbf{z}=0$ and $|\alpha\mathbf{z}| = |\alpha||\mathbf{z}|$. As usual, $\mathbf{z}\in\mathbb{C}^n$ is called a *unit vector* if $|\mathbf{z}|=1$. Note also that

$$(1,i)^T\cdot(1,i)^T = (1,-i)(1,i)^T = 1-i^2 = 2$$

so $(1,i)^T$ is no longer orthogonal to itself and has positive length $\sqrt{2}$.

11.3.2 Hermitian orthonormal Bases

Two vectors $\mathbf{w}$ and $\mathbf{z}$ are said to be *Hermitian orthogonal* if and only if $\mathbf{w}\cdot\mathbf{z} = \mathbf{w}^H\mathbf{z} = 0$. For example,

$$\begin{pmatrix}1\\ i\end{pmatrix}^H\begin{pmatrix}1\\ -i\end{pmatrix} = 0.$$

A set of mutually Hermitian orthogonal unit vectors is called *Hermitian orthonormal*. For example, $\frac{1}{\sqrt{2}}(1,i)^T$ and $\frac{1}{\sqrt{2}}(1,-i)^T$ are a Hermitian orthonormal basis of $\mathbb{C}^2$.

There is a projection formula in the Hermitian case, but it is slightly different from the real case. Suppose $\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_n$ form a Hermitian orthonormal basis of $\mathbb{C}^n$. Then any $\mathbf{z}\in\mathbb{C}^n$ can be written $\mathbf{z} = \sum_{i=1}^n\alpha_i\mathbf{u}_i$. Then

$$\mathbf{u}_j\cdot\mathbf{z} = \mathbf{u}_j\cdot\sum_{i=1}^{n}\alpha_i\mathbf{u}_i = \sum_{i=1}^{n}\alpha_i\mathbf{u}_j^H\mathbf{u}_i = \alpha_j.$$

Thus we have

Proposition 11.12. *Let* $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ *be a Hermitian orthonormal basis of* $\mathbb{C}^n$ *and suppose* $\mathbf{z} \in \mathbb{C}^n$*. Then*

$$\mathbf{z} = \sum_{i=1}^{n} (\mathbf{u}_i \cdot \mathbf{z})\mathbf{u}_i = \sum_{i=1}^{n} (\mathbf{u}_i^H \mathbf{z})\mathbf{u}_i.$$

Projections and the Gram-Schmidt method work for complex subspaces of $\mathbb{C}^n$ with the obvious modifications.

Proposition 11.13. *Every complex subspace* W *of* $\mathbb{C}^n$ *admits a Hermitian orthonormal basis, which can be included in a Hermitian orthonormal basis of* $\mathbb{C}^n$*. If* $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ *is a Hermitian orthonormal basis of* W*, then the projection of* $\mathbf{z} \in \mathbb{C}^n$ *onto* W *is given by*

$$P_W(\mathbf{z}) = \sum_{i=1}^{k} (\mathbf{u}_i \cdot \mathbf{z})\mathbf{u}_i = \sum_{i=1}^{n} (\mathbf{u}_i^H \mathbf{z})\mathbf{u}_i.$$

That is, $\mathbf{z}$ *has the unique orthogonal decomposition*

$$\mathbf{z} = P_W(\mathbf{z}) + (\mathbf{z} - P_W(\mathbf{z}))$$

into a component in W *and a component orthogonal to* W*.*

The proof is just like the proofs of the corresponding statements in the real case.

Definition 11.6. We will say that $U \in \mathbb{C}^{n \times n}$ is *unitary* if $U^H U = I_n$. The set of all $n \times n$ unitary matrices will be denoted by $U(n, \mathbb{C})$.

Proposition 11.14. $U \in \mathbb{C}^{n \times n}$ *is unitary if and only if* $U^{-1} = U^H$ *if and only if the columns of* U *are a Hermitian orthonormal basis of* $\mathbb{C}^n$*. Moreover,* $U(n, \mathbb{C})$ *is a matrix group containing* $O(n, \mathbb{R})$*.*

Some of the properties of unitary matrices are given in the next

Proposition 11.15. *Let* U *be unitary. Then:*

(i) every eigenvalue of U *has absolute value 1,*

(ii) $|\det(U)| = 1$*, and*

(iii) if U *is diagonal, then each diagonal entry has absolute value 1.*

Proof. This is left as an exercise. □

11.3.3 Properties of Hermitian matrices

Hermitian matrices are exactly those matrices which satisfy complex version of the fundamental property symmetric matrices (cf. Proposition11.2). Namely,

Proposition 11.16. *Let $K \in \mathbb{C}^{n\times n}$, Then $K = K^H$ if and only if $\mathbf{u}^H K\mathbf{v} = \mathbf{v}^H K\mathbf{u}$ if and only if $\mathbf{u}\cdot K\mathbf{v} = K\mathbf{u}\cdot\mathbf{v}$ for all $\mathbf{u},\mathbf{v}\in\mathbb{C}^n$. Moreover, if K is Hermitian, then $\mathbf{u}^H K\mathbf{u}$ is real for all $\mathbf{u}\in\mathbb{C}^n$.*

We also get Hermitian versions of the fundamental properties of symmetric matrices.

Proposition 11.17. *Let $K \in \mathbb{C}^{n\times n}$ be Hermitian. Then the eigen-values of K are real and eigen-vectors corresponding to distinct eigen-values are Hermitian orthogonal. Moreover, if K leaves a complex subspace W of $\mathbb{C}^n$ stable, that is $T_K(W)\subset W$, then K has an eigen-vector in W.*

The proof is identical to that of the real case, exept that here it is more natural, since we applied the Hermitian identity $A^H = A$ and (11.16) for real symmetric matrices without mentioning it.

Example 11.5. Consider $J = \left(\begin{smallmatrix} 0 & i \\ -i & 0\end{smallmatrix}\right)$. The characteristic polynomial of J is $\lambda^2 - 1$ so the eigenvalues are ± 1. Clearly $E_1 = \mathbb{C}(-i,1)^T$ and $E_{-1} = \mathbb{C}(i,1)^T$. Normalizing gives the Hermitian orthonormal eigenbasis $\mathbf{u}_1 = \frac{1}{\sqrt{2}}(-i,1)^T$ and $\mathbf{u}_2 = \frac{1}{\sqrt{2}}(i,1)^T$. Hence $JU = U\mathrm{diag}(1,-1)$, where $U = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} -i & i \\ 1 & 1\end{smallmatrix}\right)$. Therefore $J = U\mathrm{diag}(1,-1)U^H$.

11.3.4 Principal Axis Theorem for Hermitian Matrices

Theorem 11.18. *Every Hermitian matrix is similar to a real diagonal matrix via a unitary matrix. That is, if K is Hermitian, there exist a unitary U and a real diagonal D such that $K = UDU^{-1} = UDU^H$. Equivalently, every Hermitian matrix has a Hermitian ON eigen-basis.*

Since the proof is exactly the same as in the real symmetric case, so we don't need to repeat it. It would be a good exercise to reconstruct the proof for the Hermitian case in order to check that it does indeed work.

Note that in the complex case, the so called *principal axes* are actually one dimensional complex subspaces of $\mathbb{C}^n$. Hence the principal axes are real two planes (an $\mathbb{R}^2$) instead of lines as in the real case.

11.3.5 Self Adjointness in the Complex Case

A *Hermitian inner product* (,) on a general $\mathbb{C}$-vector space V is defined in the same way as a Hermitian inner product on $\mathbb{C}^n$. The key difference from the real case is that

$$(\alpha\mathbf{x}, \mathbf{y}) = \overline{\alpha}(\mathbf{x}, \mathbf{y})$$

and

$$(\mathbf{x}, \alpha\mathbf{y}) = \alpha(\mathbf{x}, \mathbf{y}).$$

Note that a Hermitian inner product is still an inner product in the real sense. Every finite dimensional $\mathbb{C}$-vector space admits a Hermitian inner product. If (,) is a Hermitian inner product on V, then a $\mathbb{C}$-linear transformation $T : V \to V$ is said to be *Hermitian self adjoint* if and only if

$$(T(\mathbf{x}), \mathbf{y}) = (\mathbf{x}, T(\mathbf{y}))$$

for all $\mathbf{x}, \mathbf{y} \in V$.

The Principal Axis Theorem for Hermitian self adjoint operators is the same as in Theorem 11.11, except for the obvious modifications.

Exercises

Exercise 11.21. Find the eigen-values of $K = \begin{pmatrix} 2 & 3+4i \\ 3-4i & -2 \end{pmatrix}$ and diagonalize K.

Exercise 11.22. Unitarily diagonalize $R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$.

Exercise 11.23. Show that the trace and determinant of a Hermitian matrix are real. In fact, show that the characteristic polynomial of a Hermitian matrix has real coefficients.

Exercise 11.24. Prove that the Hermitian matrices are exactly the complex matrices with real eigen-values that can be diagonalized using a unitary matrix.

Exercise 11.25. Show that $U(n, \mathbb{C})$ is a matrix group. Can you find a general description for $U(2, \mathbb{C})$?

Exercise 11.26. Show that two unit vectors in $\mathbb{C}^n$ coincide if and only if their dot product is 1.

Exercise 11.27. Give a description of the set of all 1×1 unitary matrices. That is, describe $U(1, \mathbb{C})$.

Exercise 11.28. Consider a 2×2 unitary matrix U such that one of U's columns is in $\mathbb{R}^2$. Is U orthogonal?

Exercise 11.29. Prove assertions (i)-(iii) in Proposition11.14.

Exercise 11.30. Supppose W is a complex subspace of $\mathbb{C}^n$. Show that the projection P_W is Hermitian.

Exercise 11.31. How does one adjust the formula $P_W = A(AA^T)^{-1}A^T$ to get the formula for the projection of a complex subspace W of $\mathbb{C}^n$?

Exercise 11.32. Give a direct proof of the Principal Axis Theorem in the 2×2 Hermitian case.

11.4 Normal Matrices and Schur's Theorem

The result that any Hermitian matrix K can be expressed in the form $K = UDU^H$, where D is real diagonal and U unitary, suggests that we can ask which other matrices $A \in \mathbb{C}^{n\times n}$ can be unitarily diagonalized. To answer leads us to a beautiful class of matrices.

11.4.1 Normal matrices

Theorem 11.19. *An $n \times n$ matrix A over $\mathbb{C}$ is unitarily diagonalizable if and only if*

$$AA^H = A^H A. \tag{11.2}$$

Definition 11.7. A matrix $A \in \mathbb{C}^{n\times n}$ for which (11.2) holds is said to be *normal*.

The only if part of the above theorem is straightforward, so we'll omit the proof. The if statement will follow from Schur's Theorem, proved below.

Clearly Hermitian matrices are normal. We also obtain more classes of normal matrices by putting various conditions on D. One of the most interesting is given in the following

Example 11.6. Suppose the diagonal of D is pure imaginary. Then $N = UDU^H$ satisfies $N^H = UD^HU^H = -UDU^H = -N$. A matrix S such that $S^H = -S$ is called *skew Hermitian*. Skew Hermitian matrices are clearly normal, and writing $N = UDU^H$, the condition $N^H = -N$ obviously implies $D^H = -D$, i.e. the diagonal of D to be pure imaginary. Therefore, a matrix N is skew Hermitian if and only if iN is Hermitian.

Example 11.7. A real skew Hermitian matrix is called *skew symmetric*. In other words, a real matrix S is skew symmetric if $S^T = -S$. For example, let

$$S = \begin{pmatrix} 0 & 1 & 2 \\ -1 & 0 & 2 \\ -2 & -2 & 0 \end{pmatrix}.$$

The determinant of a skew symmetric matrix of odd order is 0 (see Exercise 11.33 below). The trace is obviously also 0, since all diagonal entries of a skew symmetric matrix are 0. Since S is 3×3, t its characteristic polynomial is determined by the sum $\sigma_2(S)$ of the principal 2×2 minors of S. Here, $\sigma_2(S) = 9$, so the characteristic polynomial of S up to sign is $\lambda^3 - 9\lambda$. Thus the eigenvalues of S are $0, \pm 3i$.

Since the characteristic polynomial of a skew symmetric matrix S is real, the nonzero eigenvalues of S are pure imaginary and they occur in conjugate pairs. Hence the only possible real eigenvalue is 0. Recall that a polynomial $p(x)$ is called even if $p(-x) = p(x)$ and odd if $p(-x) = -p(x)$. Only even powers of x occur in an even polynomial, and only odd powers occur in an odd one.

Proposition 11.20. *Let A be $n \times n$ and skew symmetric. Then the characteristic polynomial is of A is even or odd according to whether n is even or odd.*

Proof. Since the characteristic polynomial is real, if n is even, the eigenvalues occur in pairs $\mu \neq \overline{\mu}$. Thus the characteristic polynomial $p_A(\lambda)$ factors into products of the form $\lambda^2 - |\mu|^2$, $p_A(\lambda)$ involves only even powers. If n is odd, then the characteristic polynomial has a real root μ, which has to be 0 since 0 is the only pure imaginary real number. Hence $p_A(\lambda) = \lambda q_A(\lambda)$, where q_A is even, which proves the result. □

Example 11.8. Let $A = UDU^H$, where every diagonal entry of D is a unit complex number. Then D is unitary, hence so is A. Conversely, every unitary matrix is normal and the eigenvalues of a unitary matrix have modulus one (see Exercise 11.35), so every unitary matrix has this form. For example, the skew symmetric matrix

$$U = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

is orthogonal. U has eigenvalues $\pm i$, and we can easily compute that $E_i = \mathbb{C}(1, -i)^T$ and $E_{-i} = \mathbb{C}(1, i)^T$. Thus

$$U = U_1 D U_1^H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix} \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix}.$$

As a complex linear transformation of $\mathbb{C}^2$, the way U acts can be interpreted geometrically as follows. U rotates vectors on the principal axis $\mathbb{C}(1, i)^T$ spanned by $(1, i)^T$ through $\frac{\pi}{2}$ and rotates vectors on the orthogonal principal axis spanned by $(1, -i)^T$ by $-\frac{\pi}{2}$. Note, $U = R_{\pi/2}$ considered as a transformation on $\mathbb{C}^2$.

The abstract formulation of the notion of a normal matrix of course uses the notion of the adjoint of a linear transformation.

Definition 11.8. Let V be a Hermitian inner product space with inner product $(\ ,\ )$, and let $T : V \to V$ be $\mathbb{C}$-linear. Then T is said to be *normal* if and only if $TT^* = T^*T$, where T^* is the adjoint of T.

We leave it to the reader to formulate the appropriate statement of Theorem 11.19 for a normal operator T.

11.4.2 Schur's Theorem

The Theorem on normal matrices, Theorem 11.19, is a consequence of a very useful general result known as Schur's Theorem.

Theorem 11.21. *Let A be any $n \times n$ complex matrix. Then there exists an $n \times n$ unitary matrix U and an upper triangular T so that $A = UTU^{-1}$.*

Schur's Theorem can also be formulated abstractly as follows:

Theorem 11.22. *If V is a finite dimensional $\mathbb{C}$-vector space and $T : V \to V$ is linear over $\mathbb{C}$, then there exists a Hermitian orthonormal basis $\mathcal{U}$ of V for which the matrix $\mathcal{M}_{\mathcal{U}}^{\mathcal{U}}(T)$ of T is upper triangular.*

We will leave the proof Theorem 11.21 as an exercise. The idea is to apply the same method used in the first proof of the Principal Axis Theorem. The only essential facts are that A has an eigenpair $(\lambda_1, \mathbf{u}_1)$, where $\mathbf{u}_1$ can be included in a Hermitian orthonormal basis of $\mathbb{C}^n$, and the product of two unitary matrices is unitary. The reader is encouraged to write out a complete proof using induction on n.

11.4.3 Proof of Theorem 11.19

We will now finish this section by proving Theorem 11.19. Let A be normal. By Schur's Theorem, we may write $A = UTU^H$, where U is unitary and T is upper triangular. We claim that T is in fact diagonal. To see this, note that since $A^HA = AA^H$, it follows that $TT^H = T^HT$ (why?). Hence we need to show that an upper triangular normal matrix is diagonal. The key is to compare the diagonal entries of TT^H and TT^H. Let t_{ii} be the ith diagonal entry of T, and let $\mathbf{a}_i$ denote its ith row. Now the diagonal entries of TT^H are $|\mathbf{a}_1|^2, |\mathbf{a}_2|^2, \ldots |\mathbf{a}_n|^2$. On the other hand, the diagonal entries of T^HT are $|t_{11}|^2, |t_{22}|^2, \ldots, |t_{nn}|^2$. It follows that $|\mathbf{a}_i|^2 = |t_{ii}|^2$ for each i, and consequently T has to be diagonal. Therefore A is unitarily diagonalizable, and the proof is complete. □

Exercises

Exercise 11.33. Unitarily diagonalize the skew symmetric matrix of Example 11.7.

Exercise 11.34. Let S be a skew Hermitian $n \times n$ matrix. Show the following:

(a) Every diagonal entry of S is pure imaginary.

(b) All eigenvalues of S are pure imaginary.

(c) If n is odd, then $|S|$ is pure imaginary, and if n is even, then $|S|$ is real.

(d) If S is skew symmetric, then $|S| = 0$ if n is odd, and $|S| \geq 0$ if n is even.

Exercise 11.35. Let U be any unitary matrix. Show that

(a) $|U|$ has modulus 1.

(b) Every eigenvalue of U also has modulus 1.

(c) Show that U is normal.

Exercise 11.36. Are all complex matrices normal? (Sorry)

Exercise 11.37. Formulate the appropriate statement of Theorem 11.19 for a normal operator T.

Exercise 11.38. The Principle of Mathematical Induction says that if a S(n) is statement about every positive integer n, then S(n) is true for all positive integers n provided:

(a) S(1) is true, and

(b) the truth of S($n-1$) implies the truth of S(n).

Give another proof of Schur's Theorem using induction. That is, if the theorem is true for B when B is $(n-1) \times (n-1)$, show that it immediately follows for A. (Don't forget the 1×1 case.)

11.5 The Jordan Decomposition

The purpose of this section is to prove a fundamental result about an arbitrary linear transformation $T : V \to V$, where V is a finite dimensional vector space over an algebraically closed field $\mathbb{F}$. What we will show is that T can be uniquely decomposed into a sum $S+N$ of two linear transformations S and N, where S is a semi-simple and N is nilpotent and, moreover, T commutes with both S and N. This decomposition, which is called the *Jordan decomposition* of T, is a very useful tool for understanding the structure of T and has many applications. One of the main applications is, in fact, the Cayley-Hamilton Theorem, which says that T satisfies its own characteristic polynomial .

11.5.1 The Main Result

Let T be a linear transformation from V to itself, that is, an element of $L(V)$. Recall that we call T *semi-simple* if it admits an eigen-basis or, equivalently, if the matrix $\mathcal{M}_{\mathcal{B}}^{\mathcal{B}}(T)$ is diagonalizable. Recall that T is called *nilpotent* if $T^k = 0$ for some integer $k > 0$. We will prove the following fundamental

Theorem 11.23. *Let $\mathbb{F}$ be an arbitrary algebraically closed field, and consider a linear transformation $T : V \to V$, where V is a finite dimensional vector space over $\mathbb{F}$ of dimension n, say. Let $\lambda_1, \ldots, \lambda_m$ be the distinct eigenvalues of T, and suppose μ_i denotes the multiplicity of λ_i. Thus*

$$p_T(x) = (-1)^n(x - \lambda_1)^{\mu_1} \cdots (x - \lambda_m)^{\lambda_m}.$$

Then there exist subspaces $C_1, \ldots, C_m$ of V with the following properties:

(1) $\dim C_i = \mu_i$ *and V is the direct sum*

$$V = C_1 \oplus C_2 \oplus \cdots \oplus C_m.$$

In particular, $\dim V = \sum_{i=1}^{m} \mu_i$.

(2) We may define a semi-simple element $S \in L(V)$ by the condition that $S(\mathbf{v}) = \lambda_i \mathbf{v}$ if $\mathbf{v} \in C_i$. Furthermore, the element $N = T - S$ of $L(V)$ is nilpotent, that is $N^k = 0$ for some $k > 0$.

(3) S and N commute, and both commute with T: $SN = NS$, $NT = TN$ and $ST = TS$.

(4) Finally, the decomposition $T = S + N$ of T into the sum of a semi-simple transformation and a nilpotent transformation which commute is unique.

Definition 11.9. The decomposition $T = S + N$ is called the *Jordan decomposition* of T.

For example, if T is semi-simple, then the uniqueness (4) says that its nilpotent part $N = O$, while if T is nilpotent, its semi-simple part is O. We also know, for example, that if T has distinct eigenvalues, then T is semi-simple. The Jordan decomposition is necessary when T has repeated eigenvalues. Of course, if $\mathbb{F} = \mathbb{C}$ and $T \in M_n(\mathbb{C})$ is a Hermitian matrix, then we know that T is semi-simple.

Before proving the result, let's consider an example.

Example 11.9. Let $\mathbb{F} = \mathbb{C}$ take $V = \mathbb{F}^3$. Let T be the matrix linear transformation

$$T = \begin{pmatrix} 5 & 12 & 6 \\ -2 & -5 & -3 \\ 1 & 4 & 4 \end{pmatrix}.$$

The characteristic polynomial of T is $-(x-1)^2(x-2)$, so the eigenvalues are 1 and 2, which is repeated. Now the matrices $T - I_3$ and $T - 2I_3$ row reduce to

$$\begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 0 & -3 \\ 0 & 1 & 3/2 \\ 0 & 0 & 0 \end{pmatrix}$$

respectively. Hence T is not semi-simple. eigenvectors for 2 and 1 are

$$\begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 3 \\ -3/2 \\ 1 \end{pmatrix}.$$

It isn't clear how to proceed, so we will return to the example after we give the proof.

Proof of Theorem 11.23. First of all, notice that if $R : V \to V$ is linear, then $\ker(R^r) \subset \ker(R^s)$ for all positive integers $r < s$. Since V is finite dimensional, , it follows that for some $r > 0$, $\ker(R^r) = \ker(R^{r+1})$, and thus $\ker(R^r) = \ker(R^s)$ if $r \leq s$. In this case, we say that $\ker(R^r)$ is *stable. Now define*

$$C_i = \ker(R_i^{r_i}),$$

where $R_i = T - \lambda_i I_n$ *and* r_i *is large enough so that* $\ker(R_i^{r_i})$ *is stable.*

Lemma 11.24. *Let* $U_i = R_i^{r_i}$*, and choose* $k > 0$ *such that* $(U_1 \cdots U_m)^k = (U_1 \cdots U_m)^{k+1}$*. Then the element* $U = (U_1 \cdots U_m)^k$ *of* $L(V)$ *is zero.*

Proof. Let $W = \ker(U)$. Obviously, $T(W) \subset W$, so T induces a linear transformation $T' : V/W \to V/W$. Suppose $W \neq V$, so that $\dim V/W > 0$. Since $\mathbb{F}$ is algebraically closed, it follows that T' has an eigenvalue in $\mathbb{F}$, hence an eigenvector in V/W. By definition, it follows that there is a $\mu \in \mathbb{F}$ and $\mathbf{v} \in V$ such that

$$T'(\mathbf{v} + W) = \mu(\mathbf{v} + W) = (\mu\mathbf{v}) + W = T(\mathbf{v}) + W.$$

It follows from this that $\mathbf{x} = T(\mathbf{v}) - \mu\mathbf{v} \in W$. If μ is not an eigenvalue of T, then $(T - \mu I_n)^{-1}$ exists and

$$(T - \mu I_n)^{-1}(\mathbf{x}) \in W,$$

since $(T - \mu I_n)(W) \subset W$ (why?). But this is impossible, since it implies $(T - \mu I_n)^{-1}(\mathbf{x}) = \mathbf{v} \in W$, which we know is impossible (for otherwise $\mathbf{v} + W = W$). It follows that μ is an eigenvalue of T. But then the fact that $T(\mathbf{v}) - \mu\mathbf{v} \in W$ gives us that $U^2(\mathbf{v}) = \mathbf{0}$, which again says that $\mathbf{v} \in W$. This is another contradiction, so $V = W$. □

Lemma 11.25. *If $i \neq j$, then R_i is one to one on* $\ker(U_j)$.

Proof. Let $\mathbf{v} \in \ker(R_i) \cap \ker(U_j)$. By the kernel criterion for one to oneness, it suffices to show $\mathbf{v} = \mathbf{0}$. Note first that $\ker(R_i) \cap \ker(R_j) = \mathbf{0}$. But since $R_i R_j = R_j R_i$, $R_j(\ker(R_i) \subset R_i$. Thus, R_j is a one to one linear transformation of $\ker(R_i)$. Hence so is R_j^s for any $s > 0$. Therefore $\ker(R_i) \cap \ker(U_j) = \mathbf{0}$. □

It follows immediately that U_i is one to one on $\ker(U_j)$. We now come to the final Lemma.

Lemma 11.26. *Suppose $Q_1, \ldots, Q_m \in L(V)$ are such that $Q_1 \cdots Q_m = \mathbf{0}$ in $L(V)$, and, for all $i \neq j$, $\ker(Q_i) \cap \ker(Q_j) = \mathbf{0}$. Then*

$$V = \sum_{i=1}^{m} \ker(Q_i).$$

Proof. We will prove the Lemma by induction. The key step is the following **Claim:** Suppose P and Q are elements of $L(V)$ such that $PQ = O$ and $\ker(P) \cap \ker(Q) = \mathbf{0}$. Then $V = \ker(P) + \ker(Q)$.

To prove the claim, we will show that $\dim(\ker(P) + \ker(Q)) \geq \dim V$. By the Hausdorff Intersection Formula,

$$\dim \ker(P) + \dim \ker(Q) = \dim(\ker(P) + \ker(Q)) - \dim(\ker(P) \cap \ker(Q)).$$

Since $\dim(\ker(P) \cap \ker(Q)) = \mathbf{0}$, all we have to show is that $\dim\ker(P) + \dim\ker(Q) \geq \dim V$. Now

$$\dim V = \dim\ker(Q) + \dim\mathrm{Im}(Q).$$

But as $PQ = O$, $\mathrm{Im}(Q) \subset \ker(P)$, so we indeed have the claim.

To finish the proof, we induct on m. If $m = 1$, there is nothing to prove, so suppose the Lemma holds for $m - 1$. Then

$$\ker(Q_2 \cdots Q_m) = \sum_{i=2}^{m} \ker(Q_i).$$

Thus the result follows from the claim applied with $P = Q_1$. □

Thus we get the crucial fact that $V = \sum_{i=1}^{m} C_i$. In order to prove (1), we have to show that the sum is direct. By definition, we therefore have to show that if

$$\sum_{i=1}^{m} \mathbf{v}_i = \mathbf{0},$$

where each $\mathbf{v}_i \in C_i$, then in fact, every $\mathbf{v}_i = \mathbf{0}$. If there is such a sum where some $\mathbf{v}_i \neq \mathbf{0}$, let $\sum_{i=1}^{m} \mathbf{v}_i = \mathbf{0}$ be such a sum where the number of non-zero components is minimal, and let $\mathbf{v}_r$ be any non-zero component in this sum. Since there have to be at least two non-zero components, and since U_r is one to one on C_j if $j \neq r$ but $U_r(\mathbf{v}_r) = \mathbf{0}$, $U_r(\sum_{i=1}^{m} \mathbf{v}_i)$ has one less non-zero component. This contradicts the minimality of $\sum_{i=1}^{m} \mathbf{v}_i$, hence the sum $V = \sum_{i=1}^{m} C_i$ is direct. Now let $\nu_i = \dim C_i$. Then $\sum_{i=1}^{m} \mu_i = \dim V$.

To finish the proof of (1), we have to show that ν_i is the multiplicity of λ_i as an eigenvalue, i.e. $\nu_i = \mu_i$. By choosing a basis of each C_i, we get a basis $\mathcal{B}$ of V for which the matrix $A = \mathcal{M}_{\mathcal{B}}^{\mathcal{B}}(T)$ has block form

$$A = \begin{pmatrix} A_1 & O & \cdots & O \\ O & A_2 & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & \cdots & O & A_m \end{pmatrix}, \tag{11.3}$$

where A_i is $\nu_i \times \nu_i$, and each O is a zero matrix. It follows from this that

$$p_A(x) = p_{A_1}(x) \cdots p_{A_m}(x). \tag{11.4}$$

But the only eigenvalue of A_i is λ_i since $R_j = T - \lambda_j I_n$ is one to one on C_i if $i \neq j$. Thus $p_{A_i}(x) = (x - \lambda_i)^{\nu_i}$, so we conclude that the multiplicity of λ_i is ν_i, which proves (1). Note that since $p_T(x) = p_A(x)$, we have shown

$$p_T(x) = (x - \lambda_1)^{\mu_1} \cdots (x - \lambda_m)^{\mu_m}. \tag{11.5}$$

We next prove (2). The transformation S is well defined, and, by definition, $S(C_i) \subset C_i$ for all i. Clearly, $T(C_i) \subset C_i$ also, thus $N(C_i) \subset C_i$. To show N is nilpotent, we only need to show that N is nilpotent on each C_i. But for $\mathbf{v} \in C_i$, we have

$$N^{r_i}(\mathbf{v}) = (T - S)^{r_i}(\mathbf{v}) = (T - \lambda_i I_n)^{r_i}(\mathbf{v}) = \mathbf{0},$$

by the definition of C_i. Hence N is nilpotent.

To prove (3), it suffices to show that if $\mathbf{v} \in C_i$, then $NS(\mathbf{v}) = SN(\mathbf{v})$. But this is obvious, since $S(\mathbf{v}) = \lambda_i \mathbf{v}$.

To finish the proof, suppose $T = S' + N'$ is another decomposition of T, where S' is semi-simple, N' is nilpotent, and $S'N' = N'S'$. Now as S' is semi-simple, we can write

$$V = \sum_{i=1}^{k} E_{\mu_i}(S'), \tag{11.6}$$

where $\mu_1, \ldots, \mu_k$ are the distinct eigenvalues of S' and $E_{\mu_i}(S')$ denotes the eigenspace of S' for μ_i. Since N' and S' commute, $N'(E_{\mu_i}(S')) \subset E_{\mu_i}(S')$. Therefore, we can assert that for any $\mathbf{v} \in E_{\mu_i}(S')$,

$$N'^{r}(\mathbf{v}) = (T - S')^{r}(\mathbf{v}) = (T - \mu_i I_n)^{r}(\mathbf{v}) = \mathbf{0}$$

if r is sufficiently large. But this says μ_i is an eigenvalueλ_j of T, and furthermore, $E_{\mu_i}(S') \subset C_j$. Thus, $S = S'$ on $E_{\mu_i}(S')$, and therefore (11.6) implies $S' = S$. Hence, $N = N'$ too, and the proof is complete. □

Definition 11.10. The subspaces $C_1, \ldots, C_m$ associated to $T : V \to V$ are called the *invariant subspaces* of T.

Corollary 11.27. *Any $n \times n$ matrix A over $\mathbb{F}$ can be expressed in one and only one way as a sum $A = S + N$ of two commuting matrices S and N in $M_n(\mathbb{F})$, where S is diagonalizable and N is nilpotent.*

Proof. This follows immediately from the Theorem. □

Let's compute an example.

Example 11.10. Let $T : \mathbb{C}^3 \to \mathbb{C}^3$ be the matrix linear transformation of Example 11.9, and recall that $p_T(x) = -(x-1)^2(x-2)$. We have to find the invariant subspaces. Since 2 is a simple root, its invariant subspace is simply the line $E_2(T) = \mathbb{C}(2, -1, 1)^T$. Now

$$(T - I_3)^2 = \begin{pmatrix} -2 & 0 & 6 \\ 1 & 0 & -3 \\ -1 & 0 & 3 \end{pmatrix},$$

which clearly has rank one. Its kernel, which is spanned by

$$\begin{pmatrix}0\\1\\0\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}3\\0\\1\end{pmatrix},$$

is therefore T's other invariant subspace. Hence the semi-simple linear transformation S is determined

$$S(\begin{pmatrix}0\\1\\0\end{pmatrix}) = \begin{pmatrix}0\\1\\0\end{pmatrix}, \quad S(\begin{pmatrix}3\\0\\1\end{pmatrix}) = \begin{pmatrix}3\\0\\1\end{pmatrix}, \quad S(\begin{pmatrix}2\\-1\\1\end{pmatrix}) = 2\begin{pmatrix}2\\-1\\1\end{pmatrix}.$$

The matrix of S (as usual found by $SP = PD$) is therefore

$$M_S = \begin{pmatrix}-1 & 0 & 6\\1 & 1 & -3\\-1 & 0 & 4\end{pmatrix},$$

and we get N by subtraction:

$$N = \begin{pmatrix}6 & 12 & 0\\-3 & -6 & 0\\2 & 4 & 0\end{pmatrix}.$$

Thus the decomposition of T as the sum of commuting diagonalizable and a nilpotent matrices is

$$\begin{pmatrix}5 & 12 & 6\\-2 & -5 & -3\\1 & 4 & 4\end{pmatrix} = \begin{pmatrix}-1 & 0 & 6\\1 & 1 & -3\\-1 & 0 & 4\end{pmatrix} + \begin{pmatrix}6 & 12 & 0\\-3 & -6 & 0\\2 & 4 & 0\end{pmatrix}.$$

Notice that if P is the matrix which diagonalizes S, i.e.

$$P = \begin{pmatrix}0 & 3 & 2\\1 & 0 & -1\\0 & 1 & 1\end{pmatrix},$$

then

$$P^{-1}TP = \begin{pmatrix}-5 & -9 & 0\\4 & 7 & 0\\0 & 0 & 2\end{pmatrix}.$$

This gives us the matrix $\mathcal{M}_{\mathcal{B}}^{\mathcal{B}}(T)$ of T consisting of blocks down the diagonal. We will also see that by choosing P more carefully, we can even guarantee that $P^{-1}TP$ is upper triangular.

11.5.2 The Cayley-Hamilton Theorem

The Jordan decomposition has at least one exciting consequence.

Theorem 11.28. *Let $\mathbb{F}$ be algebraically closed, and let V be finite dimensional over $\mathbb{F}$. Then every $T \in L(V)$ satisfies its own characteristic polynomial .*

Proof. This follows almost immediately from the Jordan decomposition. We have to show that for every $\mathbf{v} \in V$,

$$(T - \lambda_1 I_n)^{\mu_1} \cdots (T - \lambda_m I_n)^{\mu_m}(\mathbf{v}) = \mathbf{0}. \tag{11.7}$$

Since $V = \sum C_i$, it is enough to show (11.7) if $\mathbf{v} \in C_i$ for some i. What we need is

Lemma 11.29. *Let W be a finite dimensional vector space and assume $T \in L(W)$ is nilpotent. Then $T^{\dim W} = O$.*

Proof. We will leave this as an exercise. □

To finish the proof, suppose $\mathbf{v} \in C_i$. Then as $\dim C_i = \mu_i$, the Lemma says that $(T - \lambda_i I_n)^{\mu_i}(\mathbf{v}) = \mathbf{0}$. But this implies (11.7) for $\mathbf{v}$ since the operators $(T - \lambda_i I_n)^{\mu_i}$ commute. Hence the proof is complete. □

Corollary 11.30. *A linear transformation $T : V \to V$ is nilpotent if and only if every eigenvalue of T is 0.*

Proof. We leave this as an exercise. □

One of the big questions is how do you tell whether a linear transformation $T : V \to V$ is semi-simple. In fact, there's a very simple characterization of this property. As usual, let $\lambda_1, \ldots, \lambda_m$ be the distinct eigenvalues of T.

Theorem 11.31. *A linear transformation $T : V \to V$ is semi-simple if and only if*

$$(T - \lambda_1 I_n) \cdots (T - \lambda_m I_n) = O. \tag{11.8}$$

Proof. If T is semi-simple, then in the Jordan decomposition $T = S + N$, we have $N = O$. Therefore, by the definition of N given in the proof of Theorem 11.23, $T - \lambda_i I_n = O$ on C_i. Hence $(T - \lambda_1 I_n) \cdots (T - \lambda_m I_n) = O$ on V. Conversely, if $(T - \lambda_1 I_n) \cdots (T - \lambda_m I_n) = O$, I claim $N(\mathbf{v}) = \mathbf{0}$ for all $\mathbf{v} \in V$. As always, it suffices show this for $\mathbf{v} \in C_i$ for some i. To simplify the notation, suppose $i = 1$. Then

$$\begin{aligned}(T - \lambda_1 I_n) \cdots (T - \lambda_m I_n)(\mathbf{v}) &= (T - \lambda_2 I_n) \cdots (T - \lambda_m I_n)(T - \lambda_1 I_n)(\mathbf{v}) \\ &= (T - \lambda_2 I_n) \cdots (T - \lambda_m I_n) N(\mathbf{v}) \\ &= \mathbf{0}.\end{aligned}$$

But $(T - \lambda_2 I_n) \cdots (T - \lambda_m I_n)$ is one to one on C_1, so $N(\mathbf{v}) = \mathbf{0}$. Thus, $N = O$ and the proof is done. □

Example 11.11. Let's reconsider $T : \mathbb{C}^3 \to \mathbb{C}^3$ from Example 11.9. By direct computation,

$$(T - I_3)(T - 2I_2) = \begin{pmatrix} -6 & -12 & 0 \\ 3 & 6 & 0 \\ -2 & -4 & 0 \end{pmatrix}.$$

This tells us that T is not semi-simple, which of course, we already knew.

Notice that the Cayley-Hamilton Theorem tells us that there is always a polynomial $p(x) \in \mathbb{F}[x]$ for which $p(T) = O$.

Definition 11.11. Let $T : V \to V$ be a non-zero linear transformation. Then the non-zero polynomial $p(x) \in \mathbb{F}[x]$ of least degree and leading coefficient one such that $p(T) = O$ is called the *minimal polynomial* of T.

Of course, it isn't clear that a unique minimal polynomial exists. However, let p_1 and p_2 each be a minimal polynomial. By a general property of $\mathbb{F}[x]$, we can find polynomials $q(x)$ and $r(x)$ in $\mathbb{F}[x]$ such that

$$p_2(x) = q(x)p_1(x) + r(x),$$

where either $r = 0$ or the degree of r is less than the degree of p_2. But as $p_1(T) = p_2(T) = O$, it follows that $r(T) = O$ also. Since either $r = 0$ or the degree of r is smaller than the degree of p_2, we conclude that it must be the case that $r = 0$. But then $q(x)$ is a constant since p_1 and p_2 have to have the same degree. Thus $q = 1$ since p_1 and p_2 each have leading coefficient one. Hence $p_1 = p_2$.

Proposition 11.32. *Suppose $T : V \to V$ is a non-zero linear transformation, and assume its distinct eigenvalues are $\lambda_1, \ldots, \lambda_m$. The minimal polynomial $p(x)$ of T is unique, it divides the characteristic polynomial $p_T(x)$, and finally, $(x - \lambda_1) \cdots (x - \lambda_m)$ divides $p(x)$.*

Proof. The uniqueness was already shown. By the Cayley-Hamilton Theorem, $p_T(T) = O$. Hence writing $p_T(x) = q(x)p(x) + r(x)$ as above and repeating the argument, we get $r = 0$. The fact that $(x - \lambda_1) \cdots (x - \lambda_m)$ divides $p(x)$ is clear from the proof of Theorem 11.23. Indeed, we can factor $p(x)$ into linear factors $p(x) = (x - a_1) \cdots (x - a_k)$ where all $a_i \in \mathbb{F}$. If a λ_j is not among the a_i, we know $p(T)$ cannot be zero on C_j. Hence each $(x - \lambda_j)$ has to be a factor. □

Corollary 11.33. *A nonzero linear transformation $T : V \to V$ is semi-simple if and only if its minimal polynomial is $(x - \lambda_1) \cdots (x - \lambda_m)$.*

Proof. Just apply Theroem 11.31 and PropositionrefMINPOLYPROP. □

Example 11.12. Let's reconsider $T : \mathbb{C}^3 \to \mathbb{C}^3$ from Example 11.9. We have $P_T(x) = -(x-1)^2(x-2)$. Now by direct computation,

$$(T - I_3)(T - 2I_2) = \begin{pmatrix} -6 & -12 & 0 \\ 3 & 6 & 0 \\ -2 & -4 & 0 \end{pmatrix}.$$

This tells us that T is not semi-simple, which of course, we already knew.

11.5.3 The Jordan Canonical Form

The Jordan decomposition $T = S + N$ of a $T \in L(V)$ can be extensively improved. The first step is to find a basis for which T is upper triangular. In fact, it will suffice to show that there exists a basis of each C_i for which N is upper triangular.

For this we may as well suppose $C_i = V$. Let k be the least positive integer for which $N^k = O$. Now

$$\ker(N) \subset \ker(N^2) \subset \cdots \subset \ker(N^k) = V.$$

Since k is the least integer such that $N^k = O$, each of the above inclusions is proper. Notice that for each $r > 0$, $N(\ker(N^r)) \subset \ker(N^{r-1})$. Thus we can construct a basis of V by first selecting a basis of $\ker(N)$, extending this basis to a basis of $\ker(N^2)$, extending the second basis to a basis of $\ker(N^3)$ and so forth until a basis $\mathcal{B}$ of $V = \ker(N^k)$ is obtained. Since $N(\ker(N^r)) \subset \ker(N^{r-1})$, it follows that the matrix $\mathcal{M}_{\mathcal{B}}^{\mathcal{B}}(N)$ is upper triangular. Since S is a multiple of $\lambda_i I_{\mu_i}$ on C_i, we can infer

Proposition 11.34. *Every $n \times n$ matrix over an algebraically closed field $\mathbb{F}$ is similar to a upper trianglar matrix over $\mathbb{F}$.*

This result is reminiscent of Schur's Theorem and could in fact have been proven in a similar way. We next introduce the famous Jordan Canonical Form of a matrix.

Theorem 11.35 (The Jordan Canonical Form). *As usual, let V be a finite dimensional vector space over the algebraically closed field $\mathbb{F}$, and*

suppose $T \in L(V)$. Then there exists a basis $\mathcal{B}$ of V for which

$$\mathcal{M}_{\mathcal{B}}^{\mathcal{B}}(T) = \begin{pmatrix} J_1 & O & \cdots & O \\ O & J_2 & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & \cdots & O & J_s \end{pmatrix}, \tag{11.9}$$

where the matrices J_i (the Jordan blocks) have the form

$$J_i = \lambda I_{n_i} + N_i,$$

where N_i is the upper trianglar $n_i \times n_i$ matrix with 0's on the diagonal, 1's on the super diagonal and 0's above the super diagonal. Furthermore, we may suppose $n_1 \geq n_2 \geq \cdots \geq n_s$. In particular, when $V = \mathbb{F}^n$, we get the result that every $A \in M_n(\mathbb{F})$ is similar to a matrix having the form (11.9).

The proof requires that we play around a bit more in the manner of the discussion before Proposition11.34. We will skip the details. Note that there is no connection between the n_i and the eigenvalues λ_j of T, except that if $J_i = \lambda I_{n_i} + N_i$, then n_i cannot exceed the multiplicity of λ as a root of $p_T(x)$. note also that each eigenvalue λ_i of T appears μ_i times on the diagonal of $\mathcal{M}_{\mathcal{B}}^{\mathcal{B}}(T)$.

11.5.4 A Connection With Number Theory

One of the surprising conclusions that can be drawn from the Jordan Canonical Form has to do with the partitions of n.

Definition 11.12. Let n be a positive integer. Then a *partition* of n is a non-increasing sequence of positive integers $a_1, a_2, \cdots, a_r$ such that $\sum_{i=1}^{r} a_i = n$. The partition function $\pi(n)$ is the function which counts the number of partitions of n.

Thus $\pi(1) = 1$, $\pi(2) = 2$, $\pi(3) = 3$ and $\pi(4) = 5$.

Example 11.13. The partitions of 6 are

$$6 = 1+1+1+1+1+1 = 2+1+1+1 = 2+2+1+1 = 3+2+1 =$$

$$2+2+2 = 3+3 = 4+1+1 = 4+2 = 5+1.$$

Thus there are 10 partition of 6, so $\pi(6) = 10$.

The partition function grows very rapidly. The upshot of the Jordan Canonical Form is that to each partition $(n_1, n_2, \ldots, n_r)$ of n, there is a nilpotent matrix of the form (11.9) (with only zeros on the diagonal, of course), and every $n \times n$ nilpotent matrix is similar to one of these matrices. This seemingly accidental connection has lead to some astoundingly deep results in algebra

We intend to revisit the Jordan Decomposition when we take up ring theory.

Exercises

Exercise 11.39. Use the Cayley-Hamilton Theorem to prove directly that the minimal polynomial of a linear transformation $T : V \to V$ divides the characteristic polynomial of T. (Hint: write $p_T(x) = a(x)p(x) + r(x)$ where either $r = 0$ or the degree of $r(x)$ is smaller than the degree of $p(x)$.)

Exercise 11.40. Prove Corollary 11.30 directly using the fact that

$$\ker(T) \subset \ker(T^2) \subset \ker(T^3) \subset \cdots .$$

Exercise 11.41. Compute the minimal polynomials of the following matrices:

Exercise 11.42. Show that if $A = \text{diag}[d_1, \ldots, d_n]$ is a diagonal matrix, then for any polynomial $f(x)$,

$$f(A) = \text{diag}[f(d_1), \ldots, f(d_n)].$$

Use this to conclude the Cayley-Hamilton Theorem for diagonal matrices.

Exercise 11.43. Show that if $\mathbb{F} = \mathbb{C}$, the Cayley-Hamilton Theorem follows from the fact any element of $L(V)$ is the limit of a sequence of semi-simple elements of $L(V)$. How does one construct such a sequence?

Exercise 11.44. List all 10 6×6 nilpotent matrices in Jordan Canonical Form.

Chapter 12

Applications of Symmetric Matrices

The purpose of this chapter is to present an assortment of results and applications for symmetric matrices. We will begin on with the topic of quadratic forms, study the QR algorithm and finish by giving a brief introduction to graph theory.

12.1 Quadratic Forms

12.1.1 The Definition

The most basic functions of several variables in algebra are linear functions. A linear function is a polynomial in several variables in which every term has degree one, such as $f(x_1, \ldots, x_n) = \sum_{i=1}^{n} a_i x_i$. The coefficients $a_1, \ldots, a_n$ are usually interpreted as elements of some field $\mathbb{F}$. Polynomials in which every term has degree two are called *quadratic forms*. An arbitrary quadratic form over a field $\mathbb{F}$ has the form

$$q(x_1, \ldots, x_n) = \sum_{i,j=1}^{n} h_{ij} x_i x_j \tag{12.1}$$

where each $h_{ij} \in \mathbb{F}$. Notice that by putting

$$q_{ij} = 1/2(h_{ij} + h_{ji}),$$

we can always assume that the coefficients of a quadratic form are symmetric in the sense that $q_{ij} = q_{ji}$ for all indices i, j.

The following Propositionpoints out an extremely simple yet very powerful way of representing a quadratic form.

Proposition 12.1. *For every quadratic form q over $\mathbb{F}$, there exists a unique symmetric matrix $A \in \mathbb{F}^{n\times n}$ such that*

$$q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x},$$

for all $\mathbf{x} \in \mathbb{F}^n$. Conversely, every symmetric matrix $A \in \mathbb{F}^{n\times n}$ defines a unique quadratic form over $\mathbb{F}$ by $q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$.

Proof. This is left as an exercise. □

12.1.2 Critical Point Theory

For the remainder of the the section, we will assume $\mathbb{F}$ is the reals $\mathbb{R}$. Every quadratic form $q(x_1, \ldots, q_n)$ has a critical point at the origin $\mathbf{0}$. That is,

$$\frac{\partial q}{\partial x_i}(0, \ldots, 0) = 0,$$

for all i. One of the basic applications of quadratic forms is to the problem of determining the nature of the citical point. In particular, one would usually like to know if $\mathbf{0}$ is a max or a min or neither. In vector calculus, one usually states the second derivative test, which says which of the possibilities occurs.

Let's consider the two variable case. Let $q(x, y) = ax^2 + 2bxy + cy^2$, where $a, b, c \in \mathbb{R}$. Then

$$q(x, y) = (x\ y) \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

The second derivative test says that if $\delta = \det(\begin{pmatrix} a & b \\ b & c \end{pmatrix}) > 0$, then q has a local minimum at $\mathbf{0}$ if $a > 0$ and a local maximum if $a < 0$. Moreover, if $\delta < 0$, there can't be either a local max or min at $\mathbf{0}$. We can easily see what's going on. The first thing to do is to diagonalize our symmetric matrix $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$.

In fact, suppose $A = QDQ^T$, where Q is orthogonal. Then $q(x, y) > 0$ for all $(x, y) \neq \mathbf{0}$ if and only if both eigenvalues of A are positive. Similarly, $q(x, y) < 0$ for all $(x, y) \neq \mathbf{0}$ if and only if both eigenvalues of A are negative. If the one eigenvalue is positive and the other is negative, then there neither

inequality holds for all $(x, y) \neq \mathbf{0}$. Indeed, this follows easily by putting $(u, v) = (x, y)Q$. For then

$$q(x, y) = (u\ v) \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \lambda u^2 + \mu v^2,$$

and $(u, v) \neq 0$ if and only if $(x, y) \neq 0$.

Example 12.1. Let $f(x, y) = x^2 + xy + y^2$. The associated symmetric matrix is

$$A = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix},$$

and $f(x, y) = (x\ y)A(x\ y)^T$. Both rows sum to $3/2$, so $3/2$ is obviously an eigenvalue. Since the trace of A is 2, the other eigenvalue is $1/2$. Then A can then be expressed as QDQ^{-1}, where $Q = \frac{1}{\sqrt{2}} \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$. Putting $(u\ v) = (x\ y)Q$, gives $f(x, y) = 3/2u^2 + 1/2v^2$. Thus $f(x, y)$ can be expressed as the sum of squares $3/2u^2 + 1/2v^2$, so Since the eigenvaluesare both positive f has a local minimum at $(0, 0)$.

Example 12.2. The analysis above also works on the question of determining the nature of the curve $ax^2 + 2bxy + cy^2 = d$. In the above example,

$$x^2 + xy + y^2 = 3/2u^2 + 1/2v^2,$$

provided x, y, u and v are related by

$$(u\ v) = (x\ y)Q.$$

The last equation gives $(x\ y)^T = Q(u\ v)^T$ since Q is orthogonal. Now if we consider the orthonormal basis of $\mathbb{R}^2$ defined by the columns $\mathbf{q}_1$ and $\mathbf{q}_2$ of Q, this tells us

$$\begin{pmatrix} x \\ y \end{pmatrix} = u\mathbf{q}_1 + v\mathbf{q}_2.$$

In other words, u and v are the coordinates of $(x\ y)^T$ with respect to the orthonormal basis $\mathbf{q}_1$ and $\mathbf{q}_2$. What this means is that if instead of using the x and y axis as he coordinate axes, if we use the coordinate axes along $\mathbf{q}_1$ and $\mathbf{q}_2$, then the original curve looks like an ellipse. Even more, one can certainly choose $\mathbf{q}_1$ and $\mathbf{q}_2$ such that $\det(Q) = 1$. Then Q is a rotation, and $\mathbf{q}_1$ and $\mathbf{q}_2$ are gotten by rotating $\mathbf{e}_1$ and $\mathbf{e}_2$ Notice that in this example, $\det(Q) = -1$, so we can for example instead use the rotation

$$Q^* = R_{\pi/4} = \frac{1}{\sqrt{2}} \left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right).$$

The curve $x^2 + xy + y^2 = 1$ has its major and minor axes along the v and u axes respectively.

Passing to the new expression for $f(x, y)$ in terms of u and v is a change of variables. The matrix Q is the change of variables matrix. In the new uv-coordinates, the curve $3/2u^2 + 1/2v^2 = 1$ is an ellipse.

What the above example shows is the following important fact:

Proposition 12.2. *Every real quadratic form $q(x_1, \dots, x_n)$ can be written as a sum of squares*

$$q(x_1, \dots, x_n) = \sum_{i=1}^{n} \lambda_i r_i^2,$$

where $\lambda_1, \dots, \lambda_n$ are the eigenvaluesof the matrix associated to q. The coordinates $r_1, \dots, r_n$ are obtained from a orthogonal change of variables, $(r_1 \ \cdots \ r_n) = (x_1 \ \cdots \ x_n)Q$, i.e. $\mathbf{r} = Q^T\mathbf{x}$.

12.1.3 Positive Definite Matrices

Let's start with a definition.

Definition 12.1. Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric and let $q(\mathbf{x}) = \mathbf{x}A\mathbf{x}^T$ be its associated quadratic form. Then we say A is *positive definite* if and only if $q(\mathbf{x}) > 0$ whenever $\mathbf{x} \in \mathbf{0}$. Similarly, we say q is *negative definite* if and only if $q(\mathbf{x}) < 0$ whenever $\mathbf{x} \in \mathbf{0}$. Otherwise, we say that q is *indefinite*.

Of course, if A is positive definite, then the quadratic form q has a minimum at the origin and a maximum if A is negative definite. As in the above examples, we have the

Proposition 12.3. *Let A be an $n \times n$ symmetric matrix over $\mathbb{R}$. Then the associated quadratic form $q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ is positive definite if and only if all the eigenvalues of A are positive and negative definite if and only if all the eigenvalues of A are negative.*

Proof. This is an execise. □

Example 12.3. Consider

$$A_1 = \begin{pmatrix} 2 & 3 \\ 3 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} -7 & 4 \\ 4 & -3 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix},$$

and denote the associated quadratic polynomials by f_1, f_2 and f_3. Since $|A_1| < 0$, f_1 has one positive and one negative ev, so f_1 has neither a max nor min at $(0, 0)$. Here we say that f_1 has a saddle point at $(0, 0)$. Both

eigenvalues of A_2 are negative and both eigenvalues of A_3 are positive. So f_2 has a local maximum at $(0,0)$ and f_3 has a local minimum at $(0,0)$.

To decide when a real symmetric matrix is positive definite, we have to test whether its eigenvalues are all positive or not. There are tests for determining the signs of the roots of a polynomial, but they are somewhat difficult to implement. For a satisfactory result, we need to return to the LDU decomposition.

12.1.4 Positive Definite Matrices and Pivots

Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric. We now want to determine the signs of the eigenvalues of A. If we recall the result that the characteristic polynomial of A can be written in terms of the principal minors of A, we can ask if there is a way of determining the signs of the eigenvalues from the behavior of the principal minors. It turns out that there is indeed a way of doing this, which involves the LDU decomposition of A.

Recall that if A has an LDU decomposition, say $A = LDU$, then in fact $U = L^T$, so $A = LDL^T$. As a result, we can write $q(\mathbf{x}) = \mathbf{x}^T LDL^T\mathbf{x}$, which will allow us below to express q as a sum of squares but with non-orthogonal axes. Note that the decomposition $A = LDL^T$, which is found simply by row operations, has nothing to do with the Principal Axis Theorem and eigenvalues.

Before proceeding, here are a couple of examples.

Example 12.4. Let $q(x,y,z) = x^2 + 2xy + 4xz + 2y^2 + 6yz + 2z^2$. The associated symmetric matrix is

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 2 \end{pmatrix}.$$

A routine calculation gives

$$L^*A = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & -3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -3 \end{pmatrix} \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus the LDU decomposition is

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -3 \end{pmatrix} \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

and the pivots are 1,1,-3. Changing coordinates according to

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = U \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

gives the sum of squares expression

$$q(x,y,z) = (u\ v\ w) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -3 \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} = u^2 + v^2 - 3w^2,$$

from which it is obvious that A is positive definite.

If we make a slight change, we run into a problem.

Example 12.5. Let $q'(x,y,z) = x^2 + 2xy + 4xz + y^2 + 6yz + 2z^2$. The associated matrix is

$$B = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 1 & 3 \\ 2 & 3 & 2 \end{pmatrix}.$$

Subtracting the first row from the second and third gives

$$\begin{pmatrix} 1 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix}.$$

Thus, to get a symmetric reduction, we need to consider PBP^T, where

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

The reader is asked to finish this example in Exercise 12.6.

In general, suppose a symmetric matrix A can be expressed as $A = LDL^T$ with pivots $d_1, \ldots, d_n$. Putting $\mathbf{u} = L^T\mathbf{x}$, we have

$$q(\mathbf{x}) = \mathbf{x}^T LDL^T \mathbf{x} = \mathbf{u}^T D\mathbf{u} = \sum_{i=1}^{n} d_i u_i^2.$$

From this we easily get

Proposition 12.4. *Consider a real quadratic form $q(\mathbf{x})$ with associated symmetric matrix A, which admits an LDU decomposition $A = LDL^T$ with pivots $d_1, \ldots, d_n$. Then A is positive definite if and only if all the pivots are positive and negative definite if and only if all the pivots are negative.*

This classifies the positive definite matrices with an LDU decoposition. We will now prove a result that implies every positive definite (or negative definite) matrix has an LDU decomposition. Suppose A is positive definite. We want to consider the sequence of matrices A_m consisting of the first m rows and columns of A. Thus A_m is clearly a $m \times m$ symmetric matrix. In fact, a key observation (left as an exercise) is that A_m is also positive definite on $\mathbb{R}^m$.

Before proving the main result, let us make some comments on the pivots of a matrix $A \in \mathbb{F}^{n\times n}$. Recall that if A can be written $A = LDU$, with L, D, U as usual, then the mth pivot of A is the mth diagonal entry of D. Moreover, it's easy to see that $A_m = L_m D_m U_m$. This just uses the fact the L is lower triangular and U is upper triangular. Hence the first m pivots of A are just those of A_m. Moreover, since $\det(L) = \det(U) = 1$, we see that $\det(A) = \det(D)$, so the determinant of A is the product of the pivots of A. More importantly, we have

Proposition 12.5. *Suppose $\det(A_m) \neq 0$ for each m between 1 and n. Then A has an LDU decomposition, and in fact, the mth pivot d_m of A is given by $d_m = \det(A_m)/\det(A_{m-1})$.*

Proof. We will leave this proof as an exercise. □

The main result about positive definite matrices is

Theorem 12.6. *A real symmetric $A \in \mathbb{R}^{n\times n}$ is positive definite if and only if $det(A_m) > 0$ for all indices m, $1 \leq m \leq n$. Similarly, A is negative definite if and only if $(-1)^m det(A_m) > 0$ for all such indices.*

Proof. Suppose first that A is positive definite. Then A_m is positive definite for each m with $1 \leq m \leq n$. But every eigenvalue of a positive definite matrix is positive, so the determinant of a positive definite matrix is positive. Therefore, $\det(A_m) > 0$ for all m.

Conversely, suppose $\det(A_m) > 0$ for all m with $1 \leq m \leq n$. I claim this implies all pivots of A are positive. Let us prove this by induction. Certainly, the first pivot $d_1 = a_{11} > 0$. Since A_{m-1} is positive definite, our induction assumption is that if $m \leq n$, the pivots $d_1, \ldots, d_{m-1}$ of A_{m-1} are positive. We have to show that the mth pivot $d_m > 0$ also. But the pivots of A_m are $d_1, \ldots, d_{m-1}$ and (by Proposition12.5) $d_m = \det(A_m)/\det(A_{m-1})$.

Since each $\det(A_m) > 0$, this shows $d_m > 0$. Therefore we can conclude by induction that all pivots of A are positive. It follows that A is positive definite.

The proof of the negative definite claims are similar and will be omitted. □

Certainly the test for positivity in the above Theorem is much simpler to apply than directly computing the signs of the eigenvalues of A. Keep in mind however, that the simplest test is the positivity of the pivots.

Let us compute another example.

Example 12.6. Consider the matrix

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

By row operations using lower triangular elementary matrices of the third kind, we get

$$L^*A = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

Hence A has an LDU decomposition, but only three pivots are positive. Therefore A is indefinite. In this example, $|A_1| = |A_2| = |A_3| = 1$ and $|A_4| = -1$. But to discover this, we already had to compute the pivots.

Exercises

Exercise 12.1. Show that if $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ satisfies $a > 0$ and $ac - b^2 > 0$, then both eigenvaluesof A are positive. In other words, justify the second derivative test.

Exercise 12.2. Put the matrices A, B, C in the last example into the form LDL^T.

Exercise 12.3. Decide whether $g(x, y, z) = x^2 + 6xy + 2xz + 3y^2 - xz + z^2$ has a max, min or neither at $(0, 0, 0)$.

Exercise 12.4. Suppose A is a symmetric matrix such that $|A| \neq 0$ and A has both positive and negative diagonal entries. Explain why A must be indefinite.

Exercise 12.5. Show that if A is a positive definite 3×3 symmetric matrix, then the coefficients of its characteristic polynomial alternate in sign. Also show that if A is negative definite, the coefficients are all positive.

Exercise 12.6. Determine the signs of the eigenvaluesof B Example 12.5.

Exercise 12.7. Give an example of a 3×3 symmetric matrix A such that the coefficients of the characteristic polynomial of A are all negative, but A is not negative define. (Your answer could be a diagonal matrix.)

Exercise 12.8. Show that if $A \in \mathbb{R}^{n \times n}$ is positive definite, then every diagonal entry of A is positive. Also show that rA is positive definite if $r > 0$ and negative definite if $r < 0$.

Exercise 12.9. Let $A \in \mathbb{R}^{n \times n}$ be positive definite and suppose $S \in \mathbb{R}^{n \times n}$ is nonsingular.

(1) When is SAS^{-1} positive definite?
(2) Is SAS^T positive definite?

Exercise 12.10. Prove Proposition12.5.

Exercise 12.11. Describe the surface $(x\ y\ z)A(x\ y\ z)^T = 1$ for the following choices of A:

$$\begin{pmatrix} 1 & 2 & -1 \\ 2 & 0 & 3 \\ 3 & -1 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 4 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 5 \end{pmatrix}.$$

(Be careful here.)

Exercise 12.12. Suppose that $F(x, y, z)$ is a real valued function that is smooth near the origin and has a critical point at $(0, 0, 0)$. Formulate the second derivative test for F at $(0, 0, 0)$.

Exercise 12.13. Suppose $A_i = 0$ for some $i < n$. Does this mean A has a zero eigenvalue?

Exercise 12.14. Show that if A is positive definite or negative definite, then A has an LDU decomposition.

Exercise 12.15. When is e^A positive definite? Can e^A ever be negative definite or indefinite?

Exercise 12.16. A symmetric real matrix A is called *positive semi-definite* if its quadratic form$q$ satisfies $q(\mathbf{x})$ $ge0$ for all $\mathbf{x} \in \mathbb{R}^n$. Prove that A is positive semi-definite if and only if every eigenvalue of A is non-negative.

12.2 Symmetric Matrices and Graph Theory

12.2.1 Introductory Remarks

The purpose of this section is to give a very brief introduction to the subject of graph theory and to show how symmetric matrices play a fundamental role. A *graph* is a structure that consists of a finite set of vertices and a finite set of bonds or edges joining pairs of vertices. We will always assume that any pair of vertices are joined by at most one edge, and no edge joins a single vertex to itself.

Graphs arise in all sorts of situations. For example, one version of the travelling salesman problem poses the following question: suppose a travelling salesman has to visit n cities any one of which is conected to any other city by a flight. Assuming the cost of the flight between any two cities is the same, find the least expensive route. Another well known problem, this one with an 18th century origin, was the question of whether there exists a path which allows one to cross all seven bridges over the Prugel River in the city of Königsberg without ever crossing the same bridge twice. This was settled in the negative by L. Euler in 1736. Another problem with graph theoretic connections is the problem of electrical networks which is solved by Kirkhoff's Laws. However, in this problem, one has to consider graphs in a slightly different context. For more information about the above topics, I suggest consulting *Introduction to Graph Theory* by B. Bollobás.

Here are some examples of graphs.

FIGURE

12.2.2 The Adjacency Matrix and Regular Graphs

Every graph has an symmetric matrix matrix with 0,1 entries called the *adjacency matrix* of the graph. Let Γ be a graph with vertices labelled $v_1, v_2, \ldots, v_n$. The adjacency matrixA_Γ is the $n \times n$ matrix A_{ij} with A_{ij} is 1 if there exists an edge joining v_i and v_j and 0 if not. It's clear from the definition that $A_{ij} = A_{ji}$, so A is symmetric as claimed. For example, a moment's reflection tells us that the number of 1's in the first row is the number of edges containing v_1. The same holds for any row, in fact. The number of edges $d(v_i)$ at a vertex v_i is called the *degree* of v_i.

Many graphs have the property that any two vertices have the same degree. Duch graphs are called *regular*. More particularly, a graph is called *k-regular* if any every vertex has degree k. To test k-regularity, it is conve-

nient to single out the vector $\mathbf{1_n} \in \mathbb{R}^{\mathbf{n}}$ all of whose components are 1. The next Propositiondetermines which graphs are k-regular.

Proposition 12.7. *A graph Γ with n vertices is k-regular if and only if $(k, \mathbf{1_n})$ is an eigenpair for A_Γ.*

Proof. A graph Γ is k-regular if and only if every row of A_Γ has exactly k 1's, i.e. each row sum is k. This is the same as saying $(k, \mathbf{1_n})$ is an eigenpair. □

We can improve significantly on this result in the following way. A graph Γ is said to be *connected* if any two of its vertices v, v' can be joined by a *path* in Γ. That is, there exists a sequence of vertices $x_0, x_1, \ldots, x_s$ such that $x_0 = v$, $x_s = v'$, and x_i and x_{i+1} are distinct vertices which lie on a common edge for each i with $0 \leq i \leq r-1$.

Let $\Delta(\Gamma)$ denote the largest degree $d(v_i)$, and let $\delta(\Gamma)$ denote the smallest $d(v_i)$. Then we have

Theorem 12.8. *Let Γ be a connected graph with adjacency matrix A_Γ, and let $\Delta(\Gamma)$ and $\delta(\Gamma)$ denote respectively the maximum and minimum of the degrees $d(v_i)$ over all vertices of Γ. Then we have the following.*

(*i*) *Every eigenvalue λ of A_Γ satisfies $|\lambda| \leq \Delta(\Gamma)$.*

(*ii*) *The largest eigenvalue λ_M satisfies $\delta(\Gamma) \leq \lambda_M \leq \Delta(\Gamma)$.*

(*iii*) *Γ is $k = \Delta(\Gamma)$-regular if and only if $\lambda_M = \Delta(\Gamma)$.*

(*iv*) *Finally, if Γ is regular, then multiplicity of $\lambda_M = \Delta(\Gamma)$ as an eigenvalue is 1.*

Proof. Let λ be an eigenvalue and choose an eigenvector $\mathbf{u}$ for λ with the property the $|u_j| \leq 1$ for each component while $u_s = 1$ for some component u_s. Then

$$|\lambda| = |\lambda u_s| = |\sum_j a_{sj} u_j| \leq \sum_j a_{sj} |u_j| \leq \sum_j a_{sj} \leq \Delta(\Gamma).$$

This proves (i). For (ii), recall the result of Exercise 11.13, namely if λ_m is also the smallest eigenvalue of A_Γ, then for any $\mathbf{u} \in \mathbb{R}^n$, we have

$$\lambda_m \mathbf{u}^T \mathbf{u} \leq \mathbf{u}^T A_\Gamma \mathbf{u} \leq \lambda_M \mathbf{u}^T \mathbf{u}.$$

But certainly

$$\mathbf{1}_n^T A_\Gamma \mathbf{1}_n = \sum_{i,j} a_{ij} \geq n\delta(\Gamma).$$

Hence

$$n\lambda_M = \lambda_M \mathbf{1}_n^T \mathbf{1}_n \geq \mathbf{1}_n^T A_\Gamma \mathbf{1}_n \geq n\delta(\Gamma),$$

so we have (ii).

We now prove (iii). (This is the only claim which uses the hypothesis that Γ is connected.) The claim that if Γ is $\Delta(\Gamma)$-regular, then $\lambda_M = \Delta(\Gamma)$ is obvious. Suppose that $\lambda_M = \Delta(\Gamma)$. Using the eigenvector $\mathbf{u}$ chosen above, we have

$$\Delta(\Gamma) = \Delta(\Gamma)u_s = \sum_j a_{sj}u_j \leq \sum_j a_{sj}|u_j| \leq \sum_j a_{sj} \leq \Delta(\Gamma).$$

Hence for every j such that $a_{sj} \neq 0$, we have $u_j = 1$. Since every vertex can be joined to v_s by a path, it follows that $\mathbf{u} = \mathbf{1}_n$. But this implies Γ is $\Delta(\Gamma)$-regular. It also follows that the multiplicity of $\Delta(\Gamma)$ as an eigenvalue is 1, which proves (iv). □

Another nice application of the adjacency matrix is that it answers the question of how many paths join two vertices v_i and v_j of Γ. Let us say that a path $v_i = x_0, x_1, \ldots, x_r = v_j$ with r edges has length r. Here, we don't require that v_i and v_j are distinct.

Proposition 12.9. *The number of paths of length $r \geq 1$ between two not necessarily distinct vertices v_i and v_j of Γ is $(A_\Gamma)_{ij}^r$.*

Proof. This is just a matter of applying the definition of matrix multiplication. □

Example 12.7. Consider the connected graph with two vertices. Its adjacency matrix is $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Now $A^m = I_2$ if m is even and $A^m = A$ if m is odd. Thus, as is easy to see directly, there is one path of any even length from each vertex to itself and one of any odd length from each vertex to the other.

12.3 The QR Algorithm

12.3.1 Introduction to the QR Algorithm

The QR algorithm is an important technique for approximating eigenvalues which is based on the fact that an invertible matrix A can be factored as $A = QR$, where Q is orthogonal and R is upper triangular with nonzero diagonals, i.e. $R \in GL(n, \mathbb{R})$. Although this was only proven for real matrices, we can also use the Gram-Schmidt method for the Hermitian inner product on $\mathbb{C}^n$ to establish the QR factorization for invertible complex matrices. Thus, any element $A \in GL(n, \mathbb{C})$ can be written in the form $A = QR$, where Q is unitary and $R \in GL(n, \mathbb{C})$ is upper triangular.

The QR algorithm starts from the QR factorization and uses the following nice observation. If A and B are a pair of invertible $n \times n$ matrices, then AB and BA have the same eigenvalues. In fact, AB and BA are similar, since

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

Thus, $A = QR$ and $A_1 = RQ$ have the same eigenvalues. Now A_1 is still invertible, so it also has a QR factorization $A_1 = Q_1R_1$, and $A_2 = R_1Q_1$ is similar to A_1, hence A_2 is also similar to A. Continuing in this manner, we get a sequence of similar matrices $A_1, A_2, A_3, \ldots$, which in many cases tends to an upper triangular matrix as i tends to ∞. Thus, the eigenvalues of A will be the elements on the diagonal of the limit.

12.3.2 Proof of Convergence

To understand why the QR algorithm works, we first need to study convergence in the unitary group $U(n, \mathbb{C})$.

Definition 12.2. A sequence $U_m = (u_{ij}^{(m)})$ of unitary matrices is said to converge if and only if all the component sequences $u_{ij}^{(m)}$ converge. Suppose all the component sequences $u_{ij}^{(m)}$ converge, and let $\lim_{m\to\infty} u_{ij}^{(m)} = x_{ij}$. Then we say that $\lim_{m\to\infty} U_m = X$, where $X = (x_{ij})$.

Proposition 12.10. *Let U_m be a sequence of unitary matrices such that $\lim_{m\to\infty} U_m$ exists, say $\lim_{m\to\infty} U_m = X$. Then $X \in U(n, \mathbb{C})$.*

The following proof uses the fact that limits of sequences of matrices behave exactly like limits of sequences of real or complex numbers. In particular, the product rule holds. (This isn't verify, since the sum and product

rules hold in $\mathbb{C}$.) Since $U_m^H U_m = I_n$ for all m, it follows that

$$I_n = \lim_{m\to\infty} U_m^H U_m = \lim_{m\to\infty} U_m^H \lim_{m\to\infty} U_m = X^H X.$$

Hence, X is unitary.

The second fact we need to know is:

Proposition 12.11. *Every sequence of $n \times n$ unitary matrices has a convergent subsequence.*

This proposition follows from the fact that the components of a unitary matrix are bounded. In fact, since the columns of a unitary matrix U are unit vectors in $\mathbb{C}^n$, every component u_{ij} of U must satisfy $|u_{ij}| \leq 1$. Thus, every component sequence has a convergent subsequence, so every sequence of unitary matrices has a convergent subsequence.

Now let us return to the QR algorithm. Let $A = A_0, A_1, A_2, \ldots$ be the sequence matrices similar to A defined above. Thus,

$$A_0 = Q_0 R_0,$$

$$A_1 = Q_0^{-1} A Q_0 = R_0 Q_0 = Q_1 R_1,$$

$$A_2 = Q_1^{-1} Q_0^{-1} A Q_0 Q_1 = R_1 Q_1 = Q_2 R_2, \ldots$$

The $(m+1)$st term of this sequence is

$$A_{m+1} = U_m^{-1} A U_m = Q_{m+1} R_{m+1}, \tag{12.2}$$

where

$$U_m = Q_0 Q_1 \cdots Q_m.$$

Each $U_m \in U(n, \mathbb{C})$, so a subsequence converges, say $\lim_{m\to\infty} U_m = X$. By Proposition 12.10, $X \in U(n, \mathbb{C})$. Again taking the algebraic properties of limits for granted, we see that

$$\lim_{m\to\infty} A_{m+1} = \lim_{m\to\infty} U_m^{-1} A U_m = X^{-1} A X.$$

Now consider the limit on the right hand side of (12.2). Since $Q_{m+1} = U_m^{-1} U_{m+1}$, it follows that

$$\lim_{m\to\infty} Q_{m+1} = \lim_{m\to\infty} U_m \lim_{m\to\infty} U_{m+1} = X^{-1} X = I_n.$$

Since $\lim_{m\to\infty} A_m$ exists, $\lim_{m\to\infty} Q_m R_m$ must also exist, so $\lim_{m\to\infty} R_m$ exists. Call the limit T. It's clear that T is upper triangular since each

R_m is. Hence, we have found a unitary matrix X and an upper triangular matrix T so that

$$X^{-1}AX = T.$$

But this is exactly the conclusion of Schur's Theorem, written in a slightly different form, so we now have the second proof of this result. To summarize, the argument above gives the following result.

Theorem 12.12. *The sequence of unitary matrices U_m obtained in the QR algorithm for $A \in GL(n,\mathbb{C})$ converges to a unitary matrix X so that $X^{-1}AX$ is upper triangular. Moreover, $T = \lim_{m\to\infty} R_m$, so if the diagonal entries of R_m are denoted $r_{ii}^{(m)}$ and we put $\lambda_i = \lim_{m\to\infty} r_{ii}^{(m)}$, then $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A.*

The above result can be modified so that it also applies when A is singular. For if $r \in \mathbb{C}$ is not an eigenvalue of A, then $A' = A - rI_n$ is invertible, so the result applies to A'. But the eigenvalues of A' are just the eigenvalues of A shifted by r.

12.3.3 The Power Method

There is another method for approximating the eigenvalues of a complex matrix called the power method which can also be explained after we make a few more definitions. First, define a *flag* $\mathbf{F}$ in $\mathbb{C}^n$ to be a sequence $(F_1, F_2, \ldots, F_n)$ of subspaces of $\mathbb{C}^n$ such that $F_1 \subset F_2 \subset \cdots \subset F_n$ of $\mathbb{C}^n$ and, for all i, $\dim F_i = i$. An ordered basis $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ of $\mathbb{C}^n$ is a *flag basis* for $\mathbf{F}$ if $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_i$ is a basis of F_i for each i. Given $A \in GL(n,\mathbb{C})$, the associated flag $\mathbf{F}(A)$ is the flag for which F_i is the span of the first i columns of A. Using column operations, it is not hard to see

Proposition 12.13. *Two matrices A and B in $GL(n,\mathbb{C})$ have the same flag if and only if $A = BT$ for an upper triangular matrix $T \in GL(n,\mathbb{C})$.*

Each $A \in GL(n,\mathbb{C})$ acts on a flag $\mathbf{F} = (F_1, F_2, \ldots, F_n)$ by

$$A\mathbf{F} = (AF_1, AF_2, \ldots, AF_n).$$

Thus elements of $GL(n,\mathbb{C})$ permute flags. A flag $\mathbf{F}$ such that $A\mathbf{F} = \mathbf{F}$ is call a *fixed flag* or *eigenflag* for A.

Example 12.8. Let A be diagonal, say $A = \operatorname{diag}(\alpha_1, \alpha_2, \ldots \alpha_n)$, where $\alpha_i \neq \alpha_j$ if $i \neq j$ and all $\alpha_i \neq 0$. Let P be any permutation matrix. Then the flag $\mathbf{F}(P)$ is fixed by A, and since the eigenvalues of A are distinct, the $\mathbf{F}(P)$, where P runs over Π_n (§22) are the only flags fixed by A. Hence A has exactly $n!$ fixed flags.

Theorem 12.14. *Let $A \in GL(n,\mathbb{C})$. Then the sequence of flags $\mathbf{F}(A^k)$ for $k = 0, 1, \ldots$ has the property that*

$$\lim_{k\to\infty} \mathbf{F}(A^k) = \mathbf{F}(X),$$

where $X \in U(n,\mathbb{C})$ is the unitary matrix of Schur's Theorem such that $A = XTX^{-1}$. In particular,

$$\lim_{k\to\infty} \mathbf{F}(A^k)$$

is a fixed flag for A.

Before giving the proof, we have to discuss the meaning of convergence for flags, we will do so now before giving the proof. Every flag $\mathbf{F}$ is the flag $\mathbf{F}(U)$ associated to some unitary matrix U. We will say that a sequence of flags $\mathbf{F}_m$, $\lim_{m\to\infty} \mathbf{F}_m = \mathbf{F}$ if for each $m \geq 0$, there exists a unitary matrix U_m such that:

(i) $\mathbf{F}(U_m) = \mathbf{F}_m$,

(ii) $\lim_{m\to\infty} U_m$ exists, say it is U, and

(iii) $\mathbf{F}(U) = \mathbf{F}$.

Alternatively, we could say $\lim_{m\to\infty} \mathbf{F}_m = \mathbf{F}$ if for each m, there exists a flag basis $\mathbf{v}_1^{(m)}, \mathbf{v}_2^{(m)}, \ldots, \mathbf{v}_n^{(m)}$ of $\mathbf{F}_m$ such that $\mathbf{v}_1^{(m)}, \mathbf{v}_2^{(m)}, \ldots, \mathbf{v}_n^{(m)}$ converge to a flag basis $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ of $\mathbf{F}$ as m tends to ∞.

Let us now prove the theorem. Let A be given as in the previous theorem, and let Q_i, R_i and U_m all be as in that proof. The key point, which is certainly not obvious, is that the QR factorization of A^{m+1} is

$$A^{m+1} = (Q_0 Q_1 \cdots Q_{m-1} Q_m)(R_m R_{m-1} \cdots R_1 R_0). \tag{12.3}$$

Hence, letting $T_m := R_m R_{m-1} \cdots R_1 R_0$, we have

$$A^{m+1} = U_m T_m.$$

Therefore

$$\mathbf{F}(A^{m+1}) = \mathbf{F}(U_m T_m) = \mathbf{F}(U_m),$$

since T_m is upper triangular. But we may assume that

$$\lim_{m\to\infty} \mathbf{F}(U_m) = \mathbf{F}(X).$$

Therefore $\lim_{m\to\infty} \mathbf{F}(A^m) = \mathbf{F}(X)$ as claimed. Since $AX = XT$, it is immediate that $\mathbf{F}(X)$ is a fixed flag for A. To complete the proof, we

must show (12.3). But this is just a matter of using the identities $R_iQ_i = Q_{i+1}R_{i+1}$ for $i = 0, 1, 2, \ldots$ starting from

$$A^{m+1} = (Q_0R_0)^{m+1}.$$

Throughout the above discussion, there was no assumption that A is diagonalizable. However, if $\mathbf{F}(X)$ is a fixed flag for A, then F_1 is spanned by an eigenvector of A for the eigenvalue with the largest modulus. This is, roughly, the power method.

We can end the section with a comment about the $LPDU$ decomposition. The set of all flags in $\mathbb{C}^n$ is usually called the variety of flags in $\mathbb{C}^n$, or, simply, the flag variety of $\mathbb{C}^n$. We will denote it by $\mathcal{F}(n)$. The $LPDU$ decomposition of a matrix A is actually a partition of the set of invertible $n \times n$ matrices into what are called "cells", each cell consisting of all matrices with the same permutation matrix P. The subsets of the form $\mathbf{F}(LPDU) = \mathbf{F}(LP)$ in $\mathcal{F}(n)$ are known as Schubert cells. They contain important geometrical information about $\mathcal{F}(n)$.