MA106 Linear Algebra lecture notes

Lecturers: Agelos Georgakopoulos
Derek Holt

10th May 2016

Contents

1 Introduction

In the first term of your studies you were introduced to the concepts of groups, rings and fields. Linear algebra comes to add a further concept, that of a *vector space*. If we think of the concept of groups as a general setting in which we can add and subtract, rings as a general setting in which we can add, subtract, and multiply, and fields as a general setting in which we can add, subtract, multiply, and divide, then a vector space can be thought of as a general setting in which we can add, subtract, and *scale*.

These two operations are modelled according to our intuition about vectors in the plane: vectors are added using the 'parallelogram rule', or by adding their coordinates separately, and they can be stretched or contracted by multiplication with a real number.

We defined groups in such a way that we can use our intuition about addition of numbers to prove results about groups, but we made the definition general enough that these proofs apply to many situations. In other words, we wanted a large variety of objects to be accepted as groups, to maximise the benefit from proving theorems about all groups.

Likewise, now that we are about to define our new concept of vector space, we want to do it in such a way that, on one hand, our intuition about vectors in the plane can help us prove theorems, and on the other hand, as many objects as possible are accepted as vector spaces.

And indeed, vector spaces are abundant across mathematics, and their study —Linear algebra— has found a large variety of applications, both theoretical and practical. Some of them will be mentioned in the course, and many others can be found in the many textbooks about the topic.

2 Vector spaces

2.1 Definition of a vector space

We now give the definition of a vector space, making the idea sketched in the Introduction more precise.

Definition. A *vector space* over a field K, is a set V endowed with two operations, $+ : V \times V \to V$ (vector addition) and $\cdot : K \times V \to V$ (scalar multiplication), satisfying the following requirements for all $\alpha, \beta \in K$ and all $\mathbf{u}, \mathbf{v} \in V$.

(i) Vector addition $+$ satisfies axioms A1, A2, A3 and A4.

(ii) $\alpha \cdot (\mathbf{u} + \mathbf{v}) = \alpha \cdot \mathbf{u} + \alpha \cdot \mathbf{v}$;

(iii) $(\alpha + \beta) \cdot \mathbf{v} = \alpha \cdot \mathbf{v} + \beta \cdot \mathbf{v}$;

(iv) $(\alpha\beta) \cdot \mathbf{v} = \alpha \cdot (\beta \cdot \mathbf{v})$;

(v) $1 \cdot \mathbf{v} = \mathbf{v}$.

Elements of the field K will be called *scalars*, while elements of V will be called *vectors*. We will use boldface letters like $\mathbf{v}$ to denote vectors. The zero vector in V will be written as $\mathbf{0}_V$, or usually just $\mathbf{0}$. This is different from the zero scalar $0 = 0_K \in K$.

The notation $+ : V \times V \to V$ we used above means that our addition takes two elements of V as arguments and returns another element of V. Some authors formulate this by saying that V is closed under addition. Likewise, scalar multiplication takes an element of K and an element of V as arguments, and always returns an element of V.

Nearly all results in this course are not less interesting if we assume that K is the field $\mathbb{R}$ of real numbers. So you may find it helpful to first assume $K = \mathbb{R}$ when solving problems, and later let K be an arbitrary field once you understand the real case.

2.2 Addition and multiplication axioms and fields

Let K be a set, and suppose we have an operation $+ : K \times K \to K$. If we want to call this operation 'addition', then it is natural to ask that it satisfies the following requirements, called *axioms for addition*.

Axioms for addition.

A1. $(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$ for all $\alpha, \beta, \gamma \in K$.

A2. There is an element $0 \in K$ such that $\alpha + 0 = 0 + \alpha = \alpha$ for all $\alpha \in K$.

A3. For each $\alpha \in K$ there exists an element $-\alpha \in K$ such that $\alpha + (-\alpha) = (-\alpha) + \alpha = 0$.

A4. $\alpha + \beta = \beta + \alpha$ for all $\alpha, \beta \in K$.

For example, in $\mathbb{N}$, A1 and A4 hold but A2 and A3 do not hold. A1–A4 all hold in $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ and $\mathbb{C}$.

Note that A1–A3 say that $(K, +)$ is a group, while A4 says that this group is abelian.

Similarly, if we have an operation $\cdot : K \times K \to K$, then we are justified to call it a multiplication operation if it satisfies the following:

Axioms for multiplication.

M1. $(\alpha \cdot \beta) \cdot \gamma = \alpha \cdot (\beta \cdot \gamma)$ for all $\alpha, \beta, \gamma \in K$.

M2. There is an element $1 \in K$ such that $\alpha \cdot 1 = 1 \cdot \alpha = \alpha$ for all $\alpha \in K$.

M3. For each $\alpha \in K$ with $\alpha \neq 0$, there exists an element $\alpha^{-1} \in K$ such that $\alpha \cdot \alpha^{-1} = \alpha^{-1} \cdot \alpha = 1$.

M4. $\alpha \cdot \beta = \beta \cdot \alpha$ for all $\alpha, \beta \in K$.

In $\mathbb{N}$ and $\mathbb{Z}$, M1,M2 and M4 hold but M3 does not. M1–M4 all hold in $\mathbb{Q}, \mathbb{R}$ and $\mathbb{C}$.

Using these axioms, we can formulate the definition of a field as follows

Definition. Let K be a set on which two operations $+\colon K \times K \to K$ and $\cdot\colon K \times K \to K$ (called addition and multiplication) are defined. Then K is called a *field* if it satisfies each of the axioms A1–A4 and M1–M4, as well as $1 \neq 0$, and the following axiom:

D. $(\alpha + \beta) \cdot \gamma = \alpha \cdot \gamma + \beta \cdot \gamma$ for all $\alpha, \beta, \gamma \in K$.

Roughly speaking, K is a field if addition, subtraction, multiplication and division (except by zero) are all possible in K. We will usually use the letter K for a general field.

Example. $\mathbb{N}$ and $\mathbb{Z}$ with the usuall operations are not fields, but $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are.

There are many other fields, including some finite fields. For example, for each prime number p, there is a field $\mathbb{F}_p = \{0, 1, 2, \ldots, p-1\}$ with p elements, where addition and multiplication are carried out modulo p. Thus, in $\mathbb{F}_7$, we have $5+4 = 2$, $5 \times 4 = 6$ and $5^{-1} = 3$ because $5 \times 3 = 1$. The smallest such field $\mathbb{F}_2$ has just two elements 0 and 1, where $1 + 1 = 0$. This field is extremely important in Computer Science since an element of $\mathbb{F}_2$ represents one bit of information.

2.3 Examples of vector spaces

1. $K^n = \{(\alpha_1, \alpha_2, \ldots, \alpha_n) \mid \alpha_i \in K\}$. This is the space of row vectors. Addition and scalar multiplication are defined by the obvious rules:

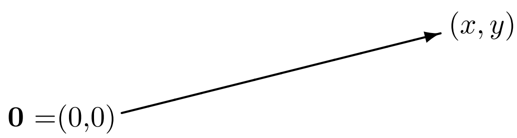
$$(\alpha_1, \alpha_2, \ldots, \alpha_n) + (\beta_1, \beta_2, \ldots, \beta_n) = (\alpha_1 + \beta_1, \alpha_2 + \beta_2, \ldots, \alpha_n + \beta_n);$$
$$\lambda(\alpha_1, \alpha_2, \ldots, \alpha_n) = (\lambda\alpha_1, \lambda\alpha_2, \ldots, \lambda\alpha_n).$$

   The most familiar examples are

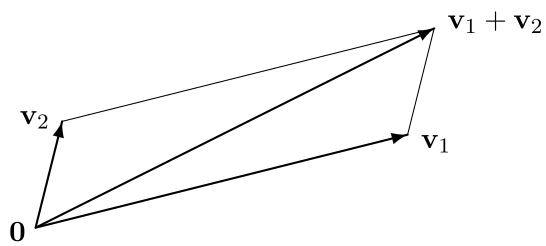
$$\mathbb{R}^2 = \{(x, y) \mid x, y \in \mathbb{R}\} \quad \text{and} \quad \mathbb{R}^3 = \{(x, y, z) \mid x, y, z \in \mathbb{R}\},$$

   which we can think of geometrically as the points in ordinary 2- and 3-dimensional space, equipped with a coordinate system.

   Vectors in $\mathbb{R}^2$ and $\mathbb{R}^3$ can also be thought of as directed lines joining the origin to the points with coordinates (x, y) or (x, y, z).



   Addition of vectors is then given by the parallelogram law.



   Note that K^1 is essentially the same as K itself.

2. Let $K[x]$ be the set of polynomials in an indeterminate x with coefficients in the field K. That is,

$$K[x] = \{\alpha_0 + \alpha_1 x + \cdots + \alpha_n x^n \mid n \geq 0, \alpha_i \in K\}.$$

   Then $K[x]$ is a vector space over K.

3. Let $K[x]_{\leq n}$ be the set of polynomials over K of degree at most n, for some $n \geq 0$. Then $K[x]_{\leq n}$ is also a vector space over K; in fact it is a *subspace* of $K[x]$.

   Note that the polynomials of degree exactly n do not form a vector space. (Why not?)

4. Let $K = \mathbb{R}$ and let V be the set of n-times differentiable functions $f\colon \mathbb{R} \to \mathbb{R}$ which are solutions of the differential equation

$$\lambda_0 \frac{d^n f}{dx^n} + \lambda_1 \frac{d^{n-1} f}{dx^{n-1}} + \cdots + \lambda_{n-1} \frac{df}{dx} + \lambda_n f = 0.$$

   for fixed $\lambda_0, \lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Then V is a vector space over $\mathbb{R}$, for if $f(x)$ and $g(x)$ are both solutions of this equation, then so are $f(x) + g(x)$ and $\alpha f(x)$ for all $\alpha \in \mathbb{R}$.

5. The previous example is a space of functions. There are many such examples that are important in Analysis. For example, the set $C^k((0,1),\mathbb{R})$, consisting of all functions $f\colon (0,1) \to \mathbb{R}$ such that the kth derivative $f^{(k)}$ exists and is continuous, is a vector space over $\mathbb{R}$ with the usual pointwise definitions of addition and scalar multiplication of functions.

6. Any n bits of information can be thought of as a vector in $\mathbb{F}_2^n$.

Facing such a variety of vector spaces, a mathematician wants to derive useful methods of handling all these vector spaces. If work out techniques for dealing with a single example, say $\mathbb{R}^3$, how can we be certain that our methods will also work for $\mathbb{R}^8$ or even $\mathbb{C}^8$? That is why we use the *axiomatic approach* to developing mathematics. We must use only arguments based on the vector space axioms. We have to avoid making any other assumptions. This ensures that everything we prove is valid for all vector spaces, not just the familiar ones like $\mathbb{R}^3$.

We shall be assuming the following additional simple properties of vectors and scalars from now on. They can all be deduced from the axioms (and it is a useful exercise to do so).

(i) $\alpha\mathbf{0} = \mathbf{0}$ for all $\alpha \in K$

(ii) $0\mathbf{v} = \mathbf{0}$ for all $\mathbf{v} \in V$

(iii) $-(\alpha\mathbf{v}) = (-\alpha)\mathbf{v} = \alpha(-\mathbf{v})$, for all $\alpha \in K$ and $\mathbf{v} \in V$.

(iv) if $\alpha\mathbf{v} = \mathbf{0}$ then $\alpha = 0$ or $\mathbf{v} = \mathbf{0}$.

3 Linear independence, spanning and bases of vector spaces

3.1 Linear dependence and independence

Definition. Let V be a vector space over the field K. The vectors $\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_n \in V$ are said to be *linearly dependent* if there exist scalars $\alpha_1, \alpha_2, \dots, \alpha_n \in K$, not all zero, such that

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n = \mathbf{0}.$$

If the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_n$ are not linearly dependent, they are said to be *linearly independent*. In other words, they are linearly independent if the only scalars $\alpha_1, \alpha_2, \dots, \alpha_n \in K$ that satisfy the above equation are $\alpha_1 = 0$, $\alpha_2 = 0$, $\dots$, $\alpha_n = 0$.

Definition. Vectors of the form $\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n$ for $\alpha_1, \alpha_2, \dots, \alpha_n \in K$ are called *linear combinations* of $\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_n$.

Example. Let $V = \mathbb{R}^2$, $\mathbf{v}_1 = (1,3)$, $\mathbf{v}_2 = (2,5)$.

Then $\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 = (\alpha_1 + 2\alpha_2, 3\alpha_1 + 5\alpha_2)$, which is equal to $\mathbf{0} = (0,0)$ if and only if $\alpha_1 + 2\alpha_2 = 0$ and $3\alpha_1 + 5\alpha_2 = 0$. Thus we have a pair of simultaneous equations in α_1, α_2 and the only solution is $\alpha_1 = \alpha_2 = 0$, so $\mathbf{v}_1$, $\mathbf{v}_2$ are linearly independent.

Example. Let $V = \mathbb{R}^2$, $\mathbf{v}_1 = (1,3)$, $\mathbf{v}_2 = (2,6)$.

This time the equations are $\alpha_1 + 2\alpha_2 = 0$ and $3\alpha_1 + 6\alpha_2 = 0$, and there are non-zero solutions, such as $\alpha_1 = -2$, $\alpha_2 = 1$, and so $\mathbf{v}_1$, $\mathbf{v}_2$ are linearly dependent.

Lemma 3.1. $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n \in V$ *are linearly dependent if and only if either* $\mathbf{v}_1 = \mathbf{0}$ *or, for some* r, $\mathbf{v}_r$ *is a linear combination of* $\mathbf{v}_1, \dots, \mathbf{v}_{r-1}$.

Proof. If $\mathbf{v}_1 = \mathbf{0}$ then by putting $\alpha_1 = 1$ and $\alpha_i = 0$ for $i > 1$ we get $\alpha_1\mathbf{v}_1 + \cdots + \alpha_n\mathbf{v}_n = \mathbf{0}$, so $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in V$ are linearly dependent.

If $\mathbf{v}_r$ is a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_{r-1}$, then $\mathbf{v}_r = \alpha_1\mathbf{v}_1 + \cdots + \alpha_{r-1}\mathbf{v}_{r-1}$ for some $\alpha_1, \ldots, \alpha_{r-1} \in K$ and so we get $\alpha_1\mathbf{v}_1 + \cdots + \alpha_{r-1}\mathbf{v}_{r-1} - 1\,\mathbf{v}_r = \mathbf{0}$ and again $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in V$ are linearly dependent.

Conversely, suppose that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in V$ are linearly dependent, and α_i are scalars, not all zero, satisfying $\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n = \mathbf{0}$. Let r be maximal with $\alpha_r \neq 0$; then $\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_r\mathbf{v}_r = \mathbf{0}$. If $r = 1$ then $\alpha_1\mathbf{v}_1 = \mathbf{0}$ which, by (iv) above, is only possible if $\mathbf{v}_1 = \mathbf{0}$. Otherwise, we get

$$\mathbf{v}_r = -\frac{\alpha_1}{\alpha_r}\mathbf{v}_1 - \cdots - \frac{\alpha_{r-1}}{\alpha_r}\mathbf{v}_{r-1}.$$

In other words, $\mathbf{v}_r$ is a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_{r-1}$. □

3.2 Spanning vectors

Definition. We say that the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$ *span* V, if every vector $\mathbf{v} \in V$ is a linear combination $\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n$ of $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

3.3 Bases of vector spaces

Definition. The vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ in V form a *basis* of V, if they are linearly independent and span V.

Proposition 3.2. *The vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ form a basis of V if and only if for every $\mathbf{v} \in V$ there is a unique sequence of scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that $\mathbf{v} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n$.*

Proof. Suppose that $\mathbf{v}_1, \ldots, \mathbf{v}_n$ form a basis of V. Then they span V, so certainly every $\mathbf{v} \in V$ can be written as $\mathbf{v} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n$. Suppose now there is a further sequence of scalars $\beta_i \in K$ such that $\mathbf{v} = \beta_1\mathbf{v}_1 + \beta_2\mathbf{v}_2 + \cdots + \beta_n\mathbf{v}_n$. Then we have

$$\begin{aligned}\mathbf{0} = \mathbf{v} - \mathbf{v} &= (\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n) - (\beta_1\mathbf{v}_1 + \beta_2\mathbf{v}_2 + \cdots + \beta_n\mathbf{v}_n)\\ &= (\alpha_1 - \beta_1)\mathbf{v}_1 + (\alpha_2 - \beta_2)\mathbf{v}_2 + \cdots + (\alpha_n - \beta_n)\mathbf{v}_n\end{aligned}$$

and so

$$(\alpha_1 - \beta_1) = (\alpha_2 - \beta_2) = \cdots = (\alpha_n - \beta_n) = 0$$

by linear independence of $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Hence $\alpha_i = \beta_i$ for all i, which means that the sequence α_i is indeed unique.

Conversely, suppose that every $\mathbf{v} \in V$ can be written uniquely as $\mathbf{v} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n$. Then $\mathbf{v}_1, \ldots, \mathbf{v}_n$ certainly span V. If $\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n = \mathbf{0}$, then

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n = 0\mathbf{v}_1 + 0\mathbf{v}_2 + \cdots + 0\mathbf{v}_n.$$

Applying the uniqueness assumption to the vector $\mathbf{0}$ yields that $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$, and so $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent. Hence they form a basis of V. □

Definition. The scalars $\alpha_1, \ldots, \alpha_n$ in the statement of the proposition are called the *coordinates* of $\mathbf{v}$ with respect to the basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

With respect to a different basis, $\mathbf{v}$ will have different coordinates. Thus, a basis for a vector space can be thought of as a choice of a system of coordinates.

Examples Here are some examples of bases of vector spaces.

1. $(1, 0)$ and $(0, 1)$ form a basis of K^2. This follows from Proposition 3.2, because each element $(\alpha_1, \alpha_2) \in K^2$ can be written as $\alpha_1(1, 0) + \alpha_2(0, 1)$, and this expression is clearly unique.

2. More generally, $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$ form a basis of K^3, $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, $(0, 0, 0, 1)$ form a basis of K^4 and so on. This is called the *standard basis* of K^n for $n \in \mathbb{N}$.

   (To be precise, the standard basis of K^n is $\mathbf{v}_1, \ldots, \mathbf{v}_n$, where $\mathbf{v}_i$ is the vector with a 1 in the ith position and a 0 in all other positions.)

3. There are many other bases of K^n. For example $(1, 0), (1, 1)$ form a basis of K^2, because $(\alpha_1, \alpha_2) = (\alpha_1 - \alpha_2)(1, 0) + \alpha_2(1, 1)$, and this expression is unique. In fact, any two non-zero vectors such that one is not a scalar multiple of the other form a basis for K^2.

4. The way we defined a basis, it has to consist of a finite number of vectors. Not every vector space has a finite basis. For example, let $K[x]$ be the space of polynomials in x with coefficients in K. Let $p_1(x), p_2(x), \ldots, p_n(x)$ be any finite collection of polynomials in $K[x]$. Then, if d is the maximum degree of $p_1(x), p_2(x), \ldots, p_n(x)$, any linear combination of $p_1(x), p_2(x), \ldots, p_n(x)$ has degree at most d, and so $p_1(x), p_2(x), \ldots, p_n(x)$ cannot span $K[x]$. On the other hand, it is possible (with a little care) to define what it means for an infinite set of vectors to be a basis of a vector space; in fact the infinite sequence of vectors $1, x, x^2, x^3, \ldots, x^n, \ldots$ is a basis of $K[x]$.

A vector space is called *finite-dimensional* if it has a finite basis. Nearly all of this course will be about finite-dimensional spaces, but it is important to remember that these are not the only examples. The spaces of functions mentioned in Example 5. of Section 2 typically have no countable basis.

Theorem 3.3 (The basis theorem)**.** *Suppose that $\mathbf{v}_1, \ldots, \mathbf{v}_m$ and $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are both bases of the vector space V. Then $m = n$. In other words, all finite bases of V contain the same number of vectors.*

For the proof of this theorem we will need some intermediate facts and the concept of *sifting* which we introduce after the next lemma.

Definition. The number n of vectors in a basis of the finite-dimensional vector space V is called the *dimension* of V, and we write $\dim(V) = n$.

Thus, as we might expect, K^n has dimension n. $K[x]$ is infinite-dimensional, but the space $K[x]_{\leq n}$ of polynomials of degree at most n has basis $1, x, x^2, \ldots, x^n$, so its dimension is $n + 1$ (not n).

Note that the dimension of V depends on the field K. Thus the complex numbers $\mathbb{C}$ can be considered as

- a vector space of dimension 1 over $\mathbb{C}$, with one possible basis being the single element 1;
- a vector space of dimension 2 over $\mathbb{R}$, with one possible basis given by the two elements $1, i$;
- a vector space of infinite dimension over $\mathbb{Q}$.

The first step towards proving the basis theorem is to be able to remove unnecessary vectors from a spanning set of vectors.

Lemma 3.4. *Suppose that the vectors* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n, \mathbf{w}$ *span* V *and that* $\mathbf{w}$ *is a linear combination of* $\mathbf{v}_1, \ldots, \mathbf{v}_n$. *Then* $\mathbf{v}_1, \ldots, \mathbf{v}_n$ *span* V.

Proof. Since $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n, \mathbf{w}$ span V, any vector $\mathbf{v} \in V$ can be written as

$$\mathbf{v} = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n + \beta \mathbf{w},$$

But $\mathbf{w}$ is a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_n$, so $\mathbf{w} = \gamma_1 \mathbf{v}_1 + \cdots + \gamma_n \mathbf{v}_n$ for some scalars γ_i; replacing in the above equation we obtain

$$\begin{aligned} \mathbf{v} &= \alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n + \beta(\gamma_1 \mathbf{v}_1 + \cdots + \gamma_n \mathbf{v}_n) \\ &= (\alpha_1 + \beta\gamma_1)\mathbf{v}_1 + \cdots + (\alpha_n + \beta\gamma_n)\mathbf{v}_n. \end{aligned}$$

Thus $\mathbf{v}$ is a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_n$, which therefore span V. □

There is an important process, called *sifting*, which can be applied to any sequence of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ in a vector space V, as follows. We consider each vector $\mathbf{v}_i$ in turn. If it is zero, or a linear combination of the preceding vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}$, then we remove it from the list.

Example. Let us sift the following sequence of vectors in $\mathbb{R}^3$.

$$\begin{array}{llll} \mathbf{v}_1 = (0,0,0) & \mathbf{v}_2 = (1,1,1) & \mathbf{v}_3 = (2,2,2) & \mathbf{v}_4 = (1,0,0) \\ \mathbf{v}_5 = (3,2,2) & \mathbf{v}_6 = (0,0,0) & \mathbf{v}_7 = (1,1,0) & \mathbf{v}_8 = (0,0,1) \end{array}$$

$\mathbf{v}_1 = \mathbf{0}$, so we remove it. $\mathbf{v}_2$ is non-zero so it stays. $\mathbf{v}_3 = 2\mathbf{v}_2$ so it is removed. $\mathbf{v}_4$ is clearly not a linear combination of $\mathbf{v}_2$, so it stays.

We have to decide next whether $\mathbf{v}_5$ is a linear combination of $\mathbf{v}_2, \mathbf{v}_4$. If so, then $(3,2,2) = \alpha_1(1,1,1) + \alpha_2(1,0,0)$, which (fairly obviously) has the solution $\alpha_1 = 2$, $\alpha_2 = 1$, so remove $\mathbf{v}_5$. Then $\mathbf{v}_6 = \mathbf{0}$ so that is removed too.

Next we try $\mathbf{v}_7 = (1,1,0) = \alpha_1(1,1,1) + \alpha_2(1,0,0)$, and looking at the three components, this reduces to the three equations

$$1 = \alpha_1 + \alpha_2; \qquad 1 = \alpha_1; \qquad 0 = \alpha_1.$$

The second and third of these equations contradict each other, and so there is no solution. Hence $\mathbf{v}_7$ is not a linear combination of $\mathbf{v}_2, \mathbf{v}_4$, and it stays.

Finally, we need to try

$$\mathbf{v}_8 = (0,0,1) = \alpha_1(1,1,1) + \alpha_2(1,0,0) + \alpha_3(1,1,0)$$

leading to the three equations

$$0 = \alpha_1 + \alpha_2 + \alpha_3 \qquad 0 = \alpha_1 + \alpha_3; \qquad 1 = \alpha_1$$

and solving these in the normal way, we find a solution $\alpha_1 = 1, \alpha_2 = 0, \alpha_3 = -1$. Thus we delete $\mathbf{v}_8$ and we are left with just $\mathbf{v}_2, \mathbf{v}_4, \mathbf{v}_7$.

Of course, the vectors that are removed during the sifting process depends very much on the order of the list of vectors. For example, if $\mathbf{v}_8$ had come at the beginning of the list rather than at the end, then we would have kept it.

Definition. The vectors remaining after applying the process of sifting to a sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ is called the *sifted subsequence* of $\mathbf{v}_1, \ldots, \mathbf{v}_r$.

The idea of sifting allows us to prove the following theorem, stating that every finite sequence of vectors which spans a vector space V actually contains a basis for V.

Theorem 3.5. *Suppose that the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ span the vector space V. Then the sifted subsequence of $\mathbf{v}_1, \ldots, \mathbf{v}_r$ forms a basis of V.*

Proof. We sift the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$. The vectors that we remove are linear combinations of the preceding vectors, and so applying Lemma 3.4 after each step, we see that the remaining vectors still span V. After sifting, no vector is zero or a linear combination of the preceding vectors (otherwise it would have been removed), so by Lemma 3.1, the remaining vectors are linearly independent. Hence they form a basis of V. □

The last theorem tells us that any vector space with a finite spanning set is finite-dimensional, and indeed the spanning set contains a basis. We now prove the dual result: any linearly independent set is contained in a basis.

Theorem 3.6. *Let V be a vector space over K which has a finite spanning set, and suppose that the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are linearly independent in V. Then we can extend the sequence to a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of V, where $n \geq r$.*

Proof. Let $\mathbf{w}_1, \ldots, \mathbf{w}_q$ be a spanning set for V. We sift the combined sequence

$$\mathbf{v}_1, \ldots, \mathbf{v}_r, \mathbf{w}_1, \ldots, \mathbf{w}_q.$$

Since $\mathbf{w}_1, \ldots, \mathbf{w}_q$ span V, the whole sequence spans V. Sifting results in a basis for V by Theorem 3.5. Since $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are linearly independent, none of them can be a linear combination of the preceding vectors by Lemma 3.1, and hence none of the $\mathbf{v}_i$ are deleted in the sifting process. Thus the resulting basis contains $\mathbf{v}_1, \ldots, \mathbf{v}_r$. □

Example. The vectors $\mathbf{v}_1 = (1,2,0,2), \mathbf{v}_2 = (0,1,0,2)$ are linearly independent in $\mathbb{R}^4$. Let us extend them to a basis of $\mathbb{R}^4$. The easiest thing is to append the standard basis of $\mathbb{R}^4$, giving the combined list of vectors

$$\mathbf{v}_1 = (1,2,0,2), \qquad \mathbf{v}_2 = (0,1,0,2), \qquad \mathbf{w}_1 = (1,0,0,0),$$
$$\mathbf{w}_2 = (0,1,0,0), \qquad \mathbf{w}_3 = (0,0,1,0), \qquad \mathbf{w}_4 = (0,0,0,1),$$

which we shall sift. We find that $(1,0,0,0) = \alpha_1(1,2,0,2) + \alpha_2(0,1,0,2)$ has no solution, so $\mathbf{w}_1$ stays. However, $\mathbf{w}_2 = \mathbf{v}_1 - \mathbf{v}_2 - \mathbf{w}_1$ so $\mathbf{w}_2$ is deleted. It is clear that $\mathbf{w}_3$ is not a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{w}_1$, because all of those have a 0 in their third component. Hence $\mathbf{w}_3$ remains. Now we have four linearly independent vectors, so must have a basis at this stage, and we can stop the sifting early by Theorem 3.3 (which we haven't finished proving yet). The resulting basis is

$$\mathbf{v}_1 = (1,2,0,2), \qquad \mathbf{v}_2 = (0,1,0,2), \qquad \mathbf{w}_1 = (1,0,0,0), \qquad \mathbf{w}_3 = (0,0,1,0).$$

We are now ready to prove Theorem 3.3. Since bases of V are both linearly independent and span V, the following proposition implies that any two bases contain the same number of vectors.

Proposition 3.7 (The exchange lemma)**.** *Suppose that vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ span V and that vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m \in V$ are linearly independent. Then $m \leq n$.*

Proof. The idea is to place the $\mathbf{w}_i$ one by one in front of the sequence $\mathbf{v}_1, \ldots, \mathbf{v}_n$, sifting each time.

Since $\mathbf{v}_1, \ldots, \mathbf{v}_n$ span V, $\mathbf{w}_1, \mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly dependent, so when we sift, at least one $\mathbf{v}_j$ is deleted. We then place $\mathbf{w}_2$ in front of the resulting sequence and sift again. Then we put $\mathbf{w}_3$ in from of the result, and sift again, and carry on doing this for each $\mathbf{w}_i$ in turn. Since $\mathbf{w}_1, \ldots, \mathbf{w}_m$ are linearly independent none of them are ever deleted. We claim that, on the other hand, at least one $\mathbf{v}_j$ is deleted in each step.

To see this, note that the sifted subsequence we obtain after each step spans V, which can be proved by inductively applying Theorem 3.5. Thus at the beginning of each step we place a vector in front of a sequence which spans V, and so the extended sequence is linearly dependent; hence at least one $\mathbf{v}_j$ gets eliminated each time.

Now in total, we append m vectors $\mathbf{w}_i$, and each time at least one $\mathbf{v}_j$ is eliminated, so we must have $m \leq n$. □

Corollary 3.8. *Let V be a vector space of dimension n over K. Then any n vectors which span V form a basis of V, and no $n-1$ vectors can span V.*

Proof. After sifting a spanning sequence, the remaining vectors form a basis by Theorem 3.5. So by Theorem 3.3, there must be precisely $n = \dim(V)$ vectors remaining. The result is now clear. □

Corollary 3.9. *Let V be a vector space of dimension n over K. Then any n linearly independent vectors form a basis of V and no $n+1$ vectors can be linearly independent.*

Proof. By Theorem 3.6 any linearly independent set is contained in a basis but by Theorem 3.3, there must be precisely $n = \dim(V)$ vectors in that basis. The result is now clear. □

3.4 Existence of a basis

Although we have studied bases quite carefully in the previous section, we have not addressed the following fundamental question. Let V be a vector space. Does it contain a basis?

Theorem 3.5 gives a partial answer that is good for many practical purposes. Let us formulate it as a corollary.

Corollary 3.10. *If a non-trivial vector space V is spanned by a finite number of vectors, then it has a basis.*

In fact, if we define the idea of an infinite basis carefully, then it can be proved that *any* vector space has a basis. That result will not be proved in this course. Its proof, which necessarily deals with infinite sets, requires a subtle result in axiomatic set theory called Zorn's lemma.

4 Subspaces

Let V be a vector space over the field K. Certain subsets of V have the nice property of being *closed* under addition and scalar multiplication; that is, adding or taking scalar multiples of vectors in the subset gives vectors which are again in the subset. We call such a subset a *subspace*:

Definition. A *subspace* of V is a non-empty subset $W \subseteq V$ such that

(i) W is closed under addition: $\mathbf{u}, \mathbf{v} \in W \Rightarrow \mathbf{u} + \mathbf{v} \in W$;

(ii) W is closed under scalar multiplication: $\mathbf{v} \in W,\ \alpha \in K \Rightarrow \alpha\mathbf{v} \in W$.

These two conditions can be replaced with a single condition

$$\mathbf{u}, \mathbf{v} \in W, \alpha, \beta \in K \Rightarrow \alpha\mathbf{u} + \beta\mathbf{v} \in W.$$

A subspace W is itself a vector space over K under the operations of vector addition and scalar multiplication in V. Notice that all vector space axioms of W hold automatically. (They are inherited from V.)

Example. The subset of $\mathbb{R}^2$ given by

$$W = \{(\alpha, \beta) \in \mathbb{R}^2 \mid \beta = 2\alpha\},$$

that is, the subset consisting of all row vectors whose second entry is twice their first entry, is a subspace of $\mathbb{R}^2$. You can check that adding two vectors of this form always gives another vector of this form; and multiplying a vector of this form by a scalar always gives another vector of this form.

For any vector space V, V is always a subspace of itself. Subspaces other than V are sometimes called *proper* subspaces. We also always have a subspace $\{\mathbf{0}\}$ consisting of the zero vector alone. This is called the *trivial* subspace, and its dimension is 0, because it has no linearly independent sets of vectors at all.

Intersecting two subspaces gives a third subspace:

Proposition 4.1. *If W_1 and W_2 are subspaces of V then so is $W_1 \cap W_2$.*

Proof. Let $\mathbf{u}, \mathbf{v} \in W_1 \cap W_2$ and $\alpha \in K$. Then $\mathbf{u} + \mathbf{v} \in W_1$ (because W_1 is a subspace) and $\mathbf{u} + \mathbf{v} \in W_2$ (because W_2 is a subspace). Hence $\mathbf{u} + \mathbf{v} \in W_1 \cap W_2$. Similarly, we get $\alpha\mathbf{v} \in W_1 \cap W_2$, so $W_1 \cap W_2$ is a subspace of V. □

It is **not** necessarily true that $W_1 \cup W_2$ is a subspace, as the following example shows.

Example. Let $V = \mathbb{R}^2$, let $W_1 = \{(\alpha, 0) \mid \alpha \in \mathbb{R}\}$ and $W_2 = \{(0, \alpha) \mid \alpha \in \mathbb{R}\}$. Then W_1, W_2 are subspaces of V, but $W_1 \cup W_2$ is not a subspace, because $(1, 0), (0, 1) \in W_1 \cup W_2$, but $(1, 0) + (0, 1) = (1, 1) \notin W_1 \cup W_2$.

Note that any subspace of V that contains W_1 and W_2 has to contain all vectors of the form $\mathbf{u} + \mathbf{v}$ for $\mathbf{u} \in W_1,\ \mathbf{v} \in W_2$. This motivates the following definition.

Definition. Let W_1, W_2 be subspaces of the vector space V. Then $W_1 + W_2$ is defined to be the set of vectors $\mathbf{v} \in V$ such that $\mathbf{v} = \mathbf{w}_1 + \mathbf{w}_2$ for some $\mathbf{w}_1 \in W_1,\ \mathbf{w}_2 \in W_2$. Or, if you prefer, $W_1 + W_2 = \{\mathbf{w}_1 + \mathbf{w}_2 \mid \mathbf{w}_1 \in W_1, \mathbf{w}_2 \in W_2\}$.

Do not confuse $W_1 + W_2$ with $W_1 \cup W_2$.

Proposition 4.2. *If W_1, W_2 are subspaces of V then so is $W_1 + W_2$. In fact, it is the smallest subspace that contains both W_1 and W_2.*

Proof. Let $\mathbf{u}, \mathbf{v} \in W_1 + W_2$. Then $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$ for some $\mathbf{u}_1 \in W_1,\ \mathbf{u}_2 \in W_2$ and $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$ for some $\mathbf{v}_1 \in W_1,\ \mathbf{v}_2 \in W_2$. Then $\mathbf{u} + \mathbf{v} = (\mathbf{u}_1 + \mathbf{v}_1) + (\mathbf{u}_2 + \mathbf{v}_2) \in W_1 + W_2$. Similarly, if $\alpha \in K$ then $\alpha\mathbf{v} = \alpha\mathbf{v}_1 + \alpha\mathbf{v}_2 \in W_1 + W_2$. Thus $W_1 + W_2$ is a subspace of V.

Any subspace of V that contains both W_1 and W_2 must contain $W_1 + W_2$ by closedness, so the latter is the smallest such subspace. □

Definition. Two subspaces W_1, W_2 of V are called *complementary* if $W_1 \cap W_2 = \{\mathbf{0}\}$ and $W_1 + W_2 = V$.

Proposition 4.3. *Let W_1, W_2 be subspaces of V. Then W_1, W_2 are complementary subspaces if and only if each vector $\mathbf{v} \in V$ can be written in a unique way as $\mathbf{v} = \mathbf{w}_1 + \mathbf{w}_2$ with $\mathbf{w}_1 \in W_1$ and $\mathbf{w}_2 \in W_2$.*

Proof. Suppose first that W_1, W_2 are complementary subspaces and let $\mathbf{v} \in V$. Then $W_1 + W_2 = V$, so we can find $\mathbf{w}_1 \in W_1$ and $\mathbf{w}_2 \in W_2$ with $\mathbf{v} = \mathbf{w}_1 + \mathbf{w}_2$. If we also had $\mathbf{v} = \mathbf{w}_1' + \mathbf{w}_2'$ with $\mathbf{w}_1' \in W_1,\ \mathbf{w}_2' \in W_2$, then we would have $\mathbf{w}_1 - \mathbf{w}_1' = \mathbf{w}_2' - \mathbf{w}_2$. The left-hand side lies in W_1 and the right-hand side lies in W_2, and so both sides (being equal) must lie in $W_1 \cap W_2 = \{\mathbf{0}\}$. Hence both sides are zero, which means $\mathbf{w}_1 = \mathbf{w}_1'$ and $\mathbf{w}_2 = \mathbf{w}_2'$, so the expression is unique.

Conversely, suppose that every $\mathbf{v} \in V$ can be written uniquely as $\mathbf{v} = \mathbf{w}_1 + \mathbf{w}_2$ with $\mathbf{w}_1 \in W_1$ and $\mathbf{w}_2 \in W_2$. Then certainly $W_1 + W_2 = V$. If $\mathbf{v}$ was a non-zero vector in $W_1 \cap W_2$, then in fact $\mathbf{v}$ would have two distinct expressions as $\mathbf{w}_1 + \mathbf{w}_2$ with $\mathbf{w}_1 \in W_1$ and $\mathbf{w}_2 \in W_2$, one with $\mathbf{w}_1 = \mathbf{v},\ \mathbf{w}_2 = \mathbf{0}$ and the other with $\mathbf{w}_1 = \mathbf{0},\ \mathbf{w}_2 = \mathbf{v}$. Hence $W_1 \cap W_2 = \{\mathbf{0}\}$, and W_1 and W_2 are complementary. □

Examples We give some examples of complementary subspaces.

1. As in the previous example, let $V = \mathbb{R}^2$, $W_1 = \{(\alpha, 0) \mid \alpha \in \mathbb{R}\}$ and $W_2 = \{(0, \alpha) \mid \alpha \in \mathbb{R}\}$. Then W_1 and W_2 are complementary subspaces.

2. Let $V = \mathbb{R}^3$, $W_1 = \{(\alpha, 0, 0) \mid \alpha \in \mathbb{R}\}$ and $W_2 = \{(0, \alpha, \beta) \mid \alpha, \beta \in \mathbb{R}\}$. Then W_1 and W_2 are complementary subspaces.

3. Let $V = \mathbb{R}^2$, $W_1 = \{(\alpha, \alpha) \mid \alpha \in \mathbb{R}\}$ and $W_2 = \{(-\alpha, \alpha) \mid \alpha \in \mathbb{R}\}$. Then W_1 and W_2 are complementary subspaces.

Another way to form subspaces is to take linear combinations of some given vectors:

Proposition 4.4. *Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be vectors in the vector space V. Then the set of all linear combinations $\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n$ of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ forms a subspace of V.*

The proof of this is completely routine and will be omitted. The subspace in this proposition is known as the subspace *spanned* by $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

The following nice fact shows that dimension is monotone with respect to subspaces:

Proposition 4.5. *Let V be a finite-dimensional vector space and W a subspace of V. Then* $\dim(W) \leq \dim(V)$.

Proof. If $\dim(W) > \dim(V)$ then, easily, W has a linearly independent subset F with more than $\dim(V)$ elements. But F is also a linearly independent subset of V, contradicting Corollary 3.9. □

We finish the section with our deepest theorem about subspaces. In a sense, it corresponds to the following basic fact about sets: if A, B are finite sets, then $|A \cup B| = |A| + |B| - |A \cap B|$.

Theorem 4.6. *Let V be a finite-dimensional vector space, and let W_1, W_2 be subspaces of V. Then*

$$\dim(W_1 + W_2) = \dim(W_1) + \dim(W_2) - \dim(W_1 \cap W_2).$$

Proof. First note that any subspace W of V is finite-dimensional by Proposition 4.5.

Let $\dim(W_1 \cap W_2) = r$ and let $\mathbf{e}_1, \ldots, \mathbf{e}_r$ be a basis of $W_1 \cap W_2$. Then $\mathbf{e}_1, \ldots, \mathbf{e}_r$ is a linearly independent set of vectors, so by Theorem 3.6 it can be extended to a basis $\mathbf{e}_1, \ldots, \mathbf{e}_r, \mathbf{f}_1, \ldots, \mathbf{f}_s$ of W_1 where $\dim(W_1) = r + s$, and it can also be extended to a basis $\mathbf{e}_1, \ldots, \mathbf{e}_r, \mathbf{g}_1, \ldots, \mathbf{g}_t$ of W_2, where $\dim(W_2) = r + t$.

To prove the theorem, we need to show that $\dim(W_1 + W_2) = r + s + t$, and to do this, we shall show that

$$\mathbf{e}_1, \ldots, \mathbf{e}_r, \mathbf{f}_1, \ldots, \mathbf{f}_s, \mathbf{g}_1, \ldots, \mathbf{g}_t$$

is a basis of $W_1 + W_2$. Certainly they all lie in $W_1 + W_2$.

First we show that they span $W_1 + W_2$. Any $\mathbf{v} \in W_1 + W_2$ is equal to $\mathbf{w}_1 + \mathbf{w}_2$ for some $\mathbf{w}_1 \in W_1,\ \mathbf{w}_2 \in W_2$. So we can write

$$\mathbf{w}_1 = \alpha_1\mathbf{e}_1 + \cdots + \alpha_r\mathbf{e}_r + \beta_1\mathbf{f}_1 + \cdots + \beta_s\mathbf{f}_s$$

for some scalars $\alpha_i, \beta_j \in K$, and

$$\mathbf{w}_2 = \gamma_1\mathbf{e}_1 + \cdots + \gamma_r\mathbf{e}_r + \delta_1\mathbf{g}_1 + \cdots + \delta_t\mathbf{g}_t$$

for some scalars $\gamma_i, \delta_j \in K$. Then

$$\mathbf{v} = (\alpha_1 + \gamma_1)\mathbf{e}_1 + \cdots + (\alpha_r + \gamma_r)\mathbf{e}_r + \beta_1\mathbf{f}_1 + \cdots + \beta_s\mathbf{f}_s + \delta_1\mathbf{g}_1 + \cdots + \delta_t\mathbf{g}_t$$

and so $\mathbf{e}_1, \ldots, \mathbf{e}_r, \mathbf{f}_1, \ldots, \mathbf{f}_s, \mathbf{g}_1, \ldots, \mathbf{g}_t$ span $W_1 + W_2$.

Finally we have to show that $\mathbf{e}_1, \ldots, \mathbf{e}_r, \mathbf{f}_1, \ldots, \mathbf{f}_s, \mathbf{g}_1, \ldots, \mathbf{g}_t$ are linearly independent. Suppose that

$$\alpha_1\mathbf{e}_1 + \cdots + \alpha_r\mathbf{e}_r + \beta_1\mathbf{f}_1 + \cdots + \beta_s\mathbf{f}_s + \delta_1\mathbf{g}_1 + \cdots + \delta_t\mathbf{g}_t = \mathbf{0} \qquad (\diamond)$$

for some scalars $\alpha_i, \beta_j, \delta_k \in K$. Then

$$\alpha_1\mathbf{e}_1 + \cdots + \alpha_r\mathbf{e}_r + \beta_1\mathbf{f}_1 + \cdots + \beta_s\mathbf{f}_s = -\delta_1\mathbf{g}_1 - \cdots - \delta_t\mathbf{g}_t \qquad (*)$$

The left-hand side of this equation lies in W_1 and the right-hand side of this equation lies in W_2. Since the two sides are equal, both must in fact lie in $W_1 \cap W_2$. Since $\mathbf{e}_1, \ldots, \mathbf{e}_r$ is a basis of $W_1 \cap W_2$, we can write the right-hand vector as

$$-\delta_1\mathbf{g}_1 - \cdots - \delta_t\mathbf{g}_t = \gamma_1\mathbf{e}_1 + \cdots + \gamma_r\mathbf{e}_r$$

for some $\gamma_i \in K$, and so

$$\gamma_1\mathbf{e}_1 + \cdots + \gamma_r\mathbf{e}_r + \delta_1\mathbf{g}_1 + \cdots + \delta_t\mathbf{g}_t = \mathbf{0}.$$

But, $\mathbf{e}_1, \ldots, \mathbf{e}_r, \mathbf{g}_1, \ldots, \mathbf{g}_t$ form a basis of W_2, so they are linearly independent, and hence $\gamma_i = 0$ for $1 \le i \le r$ and $\delta_i = 0$ for $1 \le i \le t$. But now, from the equation $(*)$ above, we get

$$\alpha_1\mathbf{e}_1 + \cdots + \alpha_r\mathbf{e}_r + \beta_1\mathbf{f}_1 + \cdots + \beta_s\mathbf{f}_s = \mathbf{0}.$$

Now $\mathbf{e}_1, \ldots, \mathbf{e}_r, \mathbf{f}_1, \ldots, \mathbf{f}_s$ form a basis of W_1, so they are linearly independent, and hence $\alpha_i = 0$ for $1 \le i \le r$ and $\beta_i = 0$ for $1 \le i \le s$. Thus all coefficients in equation $(\diamond)$ are zero, proving that the vectors $\mathbf{e}_1, \ldots, \mathbf{e}_r, \mathbf{f}_1, \ldots, \mathbf{f}_s, \mathbf{g}_1, \ldots, \mathbf{g}_t$ are linearly independent, which completes the proof that they form a basis of $W_1 + W_2$.

Hence

$$\dim(W_1+W_2) = r+s+t = (r+s)+(r+t)-r = \dim(W_1)+\dim(W_2)-\dim(W_1\cap W_2).$$

$\square$

5 Linear transformations

When you study sets, the notion of function is extremely important. There is little to say about a single isolated set, while functions allow you to link different sets. Similarly, in Linear Algebra, a single isolated vector space is not the end of the story. Things become more interesting when we connect different vector spaces by functions, especially when these functions respect the vector space operations in some way.

5.1 Definition and examples

Often in mathematics, it is as important to study special classes of functions as it is to study special classes of objects. Usually these are functions which preserve certain properties or structures. For example, continuous functions preserve which points are close to which other points. In linear algebra, the functions which preserve the vector space structure are called linear transformations; they are defined as follows

Definition. Let U, V be two vector spaces over the same field K. A *linear transformation*, or *linear map*, from U to V is a function $T\colon U \to V$ such that

(i) $T(\mathbf{u}_1 + \mathbf{u}_2) = T(\mathbf{u}_1) + T(\mathbf{u}_2)$ for all $\mathbf{u}_1, \mathbf{u}_2 \in U$;

(ii) $T(\alpha\mathbf{u}) = \alpha T(\mathbf{u})$ for all $\alpha \in K$ and $\mathbf{u} \in U$.

Notice that these two conditions are equivalent to the following single condition

$$T(\alpha\mathbf{u}_1 + \beta\mathbf{u}_2) = \alpha T(\mathbf{u}_1) + \beta T(\mathbf{u}_2) \text{ for all } \mathbf{u}_1, \mathbf{u}_2 \in U, \alpha, \beta \in K.$$

First let us state a couple of easy consequences of the definition:

Lemma 5.1. *Let $T\colon U \to V$ be a linear map. Then*

(i) $T(\mathbf{0}_U) = \mathbf{0}_V$*;*

(ii) $T(-\mathbf{u}) = -T(\mathbf{u})$ *for all* $\mathbf{u} \in U$*.*

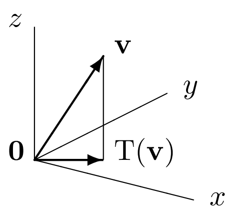
Proof. For (i), the definition of linear map gives

$$T(\mathbf{0}_U) = T(\mathbf{0}_U + \mathbf{0}_U) = T(\mathbf{0}_U) + T(\mathbf{0}_U),$$

and therefore $T(\mathbf{0}_U) = \mathbf{0}_V$. For (ii), just put $\alpha = -1$ in the definition of linear map. □

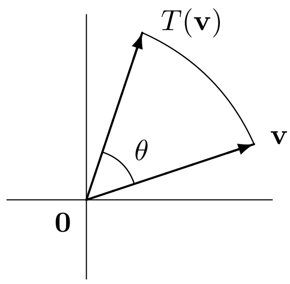
Examples Many familiar geometrical transformations, such as projections, rotations, reflections and magnifications are linear maps, and the first three examples below are of this kind. Note, however, that a nontrivial translation is not a linear map, because it does not satisfy $T(\mathbf{0}_U) = \mathbf{0}_V$.

1. Let $U = \mathbb{R}^3$, $V = \mathbb{R}^2$ and define $T\colon U \to V$ by $T((\alpha, \beta, \gamma)) = (\alpha, \beta)$. Then T is a linear map. This type of map is known as a *projection*, because of the geometrical interpretation.



   Note: In the future we shall just write $T(\alpha, \beta, \gamma)$ instead of $T((\alpha, \beta, \gamma))$.

2. Let $U = V = \mathbb{R}^2$. We interpret $\mathbf{v}$ in $\mathbb{R}^2$ as a directed line vector from $\mathbf{0}$ to $\mathbf{v}$ (see the examples in Section 2), and let $T(\mathbf{v})$ be the vector obtained by rotating $\mathbf{v}$ through an angle θ anti-clockwise about the origin.

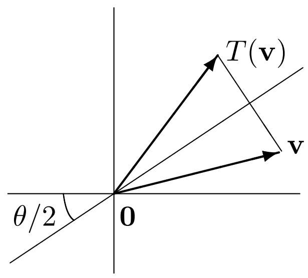


It is easy to see geometrically that $T(\mathbf{u}_1 + \mathbf{u}_2) = T(\mathbf{u}_1) + T(\mathbf{u}_2)$ and $T(\alpha\mathbf{u}) = \alpha T(\mathbf{u})$ (because everything is simply rotated about the origin), and so T is a linear map. By considering the unit vectors, we have $T(1,0) = (\cos\theta, \sin\theta)$ and $T(0,1) = (-\sin\theta, \cos\theta)$, and hence

$$T(\alpha,\beta) = \alpha T(1,0) + \beta T(0,1) = (\alpha\cos\theta - \beta\sin\theta, \alpha\sin\theta + \beta\cos\theta).$$

(*Exercise*: Show this directly.)

3. Let $U = V = \mathbb{R}^2$ again. Now let $T(\mathbf{v})$ be the vector resulting from reflecting $\mathbf{v}$ through a line through the origin that makes an angle $\theta/2$ with the x-axis.



This is again a linear map. We find that $T(1,0) = (\cos\theta, \sin\theta)$ and $T(0,1) = (\sin\theta, -\cos\theta)$, and so

$$T(\alpha,\beta) = \alpha T(1,0) + \beta T(0,1) = (\alpha\cos\theta + \beta\sin\theta, \alpha\sin\theta - \beta\cos\theta).$$

4. Let $U = V = \mathbb{R}[x]$, the set of polynomials over $\mathbb{R}$, and let T be differentiation; i.e. $T(p(x)) = p'(x)$ for $p \in \mathbb{R}[x]$. This is easily seen to be a linear map.

5. Let $U = K[x]$, the set of polynomials over K. Every $\alpha \in K$ gives rise to two linear maps, shift $S_\alpha\colon U \to U$, $S_\alpha(f(x)) = f(x-\alpha)$ and evaluation $E_\alpha\colon U \to K$, $E_\alpha(f(x)) = f(\alpha)$.

The next two examples seem dull but are important!

6. For any vector space V, we define the *identity map* $I_V\colon V \to V$ by $I_V(\mathbf{v}) = \mathbf{v}$ for all $\mathbf{v} \in V$. This is a linear map.

7. For any vector spaces U, V over the field K, we define the *zero map* $\mathbf{0}_{U,V}\colon U \to V$ by $\mathbf{0}_{U,V}(\mathbf{u}) = \mathbf{0}_V$ for all $\mathbf{u} \in U$. This is also a linear map.

One of the most useful properties of linear maps is that, if we know how a linear map $U \to V$ acts on a basis of U, then we know how it acts on the whole of U.

Proposition 5.2 (Linear maps are uniquely determined by their action on a basis)**.** *Let U, V be vector spaces over K, let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be a basis of U and let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be any sequence of n vectors in V. Then there is a unique linear map $T\colon U \to V$ with $T(\mathbf{u}_i) = \mathbf{v}_i$ for $1 \le i \le n$.*

Proof. Let $\mathbf{u} \in U$. Then, since $\mathbf{u}_1, \ldots, \mathbf{u}_n$ is a basis of U, by Proposition 3.2, there exist uniquely determined $\alpha_1, \ldots, \alpha_n \in K$ with $\mathbf{u} = \alpha_1\mathbf{u}_1 + \cdots + \alpha_n\mathbf{u}_n$. Hence, if T exists at all, then we must have

$$T(\mathbf{u}) = T(\alpha_1\mathbf{u}_1 + \cdots + \alpha_n\mathbf{u}_n) = \alpha_1\mathbf{v}_1 + \cdots + \alpha_n\mathbf{v}_n,$$

and so T is uniquely determined.

On the other hand, it is routine to check that the map $T\colon U \to V$ defined by the above equation is indeed a linear map, so T does exist. $\square$

5.2 Kernels and images

To any linear map $U \to V$, we can associate a subspace of U and a subspace of V containing important information about the map.

Definition. Let $T: U \to V$ be a linear map. The *image* of T, written as $\mathrm{im}(T)$, is the set of vectors $\mathbf{v} \in V$ such that $\mathbf{v} = T(\mathbf{u})$ for some $\mathbf{u} \in U$.

Definition. The *kernel* of T, written as $\ker(T)$, is the set of vectors $\mathbf{u} \in U$ such that $T(\mathbf{u}) = \mathbf{0}_V$.

If you prefer:

$$\mathrm{im}(T) = \{T(\mathbf{u}) \mid \mathbf{u} \in U\}; \qquad \ker(T) = \{\mathbf{u} \in U \mid T(\mathbf{u}) = \mathbf{0}_V\}.$$

Examples Let us consider the examples 1–7 above.

- In example 1, $\ker(T) = \{(0, 0, \gamma) \mid \gamma \in \mathbb{R}\}$, and $\mathrm{im}(T) = \mathbb{R}^2$.
- In example 2 and 3, $\ker(T) = \{\mathbf{0}\}$ and $\mathrm{im}(T) = \mathbb{R}^2$.
- In example 4, $\ker(T)$ is the set of all constant polynomials (i.e. those of degree 0), and $\mathrm{im}(T) = \mathbb{R}[x]$.
- In example 5, $\ker(S_\alpha) = \{\mathbf{0}\}$, and $\mathrm{im}(S_\alpha) = K[x]$, while $\ker(E_\alpha)$ is the set of all polynomials divisible by $x - \alpha$, and $\mathrm{im}(E_\alpha) = K$.
- In example 6, $\ker(I_V) = \{\mathbf{0}\}$ and $\mathrm{im}(T) = V$.
- In example 7, $\ker(\mathbf{0}_{U,V}) = U$ and $\mathrm{im}(\mathbf{0}_{U,V}) = \{\mathbf{0}\}$.

Proposition 5.3. *Let $T: U \to V$ be a linear map. Then*

(i) $\mathrm{im}(T)$ *is a subspace of* V*;*

(ii) $\ker(T)$ *is a subspace of* U*.*

Proof. For (i), we must show that $\mathrm{im}(T)$ is closed under addition and scalar multiplication. Let $\mathbf{v}_1, \mathbf{v}_2 \in \mathrm{im}(T)$. Then $\mathbf{v}_1 = T(\mathbf{u}_1)$, $\mathbf{v}_2 = T(\mathbf{u}_2)$ for some $\mathbf{u}_1, \mathbf{u}_2 \in U$. By the definition of a linear map, we have

$$\mathbf{v}_1 + \mathbf{v}_2 = T(\mathbf{u}_1) + T(\mathbf{u}_2) = T(\mathbf{u}_1 + \mathbf{u}_2) \in \mathrm{im}(T)$$

and

$$\alpha\mathbf{v}_1 = \alpha T(\mathbf{u}_1) = T(\alpha\mathbf{u}_1) \in \mathrm{im}(T),$$

so $\mathrm{im}(T)$ is a subspace of V.

Let us now prove (ii). Similarly, we must show that $\ker(T)$ is closed under addition and scalar multiplication. Let $\mathbf{u}_1, \mathbf{u}_2 \in \ker(T)$. Then, by the linearity of T, we have

$$T(\mathbf{u}_1 + \mathbf{u}_2) = T(\mathbf{u}_1) + (\mathbf{u}_2) = \mathbf{0}_V + \mathbf{0}_V = \mathbf{0}_V$$

and

$$T(\alpha\mathbf{u}_1) = \alpha T(\mathbf{u}_1) = \alpha\mathbf{0}_V = \mathbf{0}_V,$$

so $\mathbf{u}_1 + \mathbf{u}_2$, $\alpha\mathbf{u}_1 \in \ker(T)$ and hence $\ker(T)$ is a subspace of U. $\square$

Proposition 5.4. *Let $T: U \to V$ be a linear map. Then T is injective if and only if* $\ker(T) = \{\mathbf{0}_U\}$.

Proof. Proof of $\Rightarrow$. Suppose that T is injective and let $\mathbf{u} \in \ker(T)$. Then $T(\mathbf{0}_U) = \mathbf{0}_V = T(\mathbf{u})$, so $T(\mathbf{0}_U) = T(\mathbf{u})$. But T is injective so $\mathbf{0}_U = \mathbf{u}$. This proves $\Rightarrow$.

Proof of $\Leftarrow$. Let $\ker(T) = 0$ and assume that $\mathbf{u}_1, \mathbf{u}_2 \in U$ are such that $T(\mathbf{u}_1) = T(\mathbf{u}_2)$. Then $\mathbf{0}_V = T(\mathbf{u}_1) - T(\mathbf{u}_2) = T(\mathbf{u}_1 - \mathbf{u}_2)$. So $\mathbf{u}_1 - \mathbf{u}_2 \in \ker(T) = 0$ and therefore $\mathbf{u}_1 - \mathbf{u}_2 = \mathbf{0}_U$, that is, $\mathbf{u}_1 = \mathbf{u}_2$. This proves $\Leftarrow$. $\square$

5.3 Rank and nullity

The dimensions of the kernel and image of a linear map contain important information about it, and are related to each other.

Definition. let $T : U \to V$ be a linear map.

(i) $\dim(\mathrm{im}(T))$ is called the *rank* of T;

(ii) $\dim(\ker(T))$ is called the *nullity* of T.

Theorem 5.5 (The rank-nullity theorem)**.** *Let U, V be vector spaces over K with U finite-dimensional, and let $T : U \to V$ be a linear map. Then*

$$\mathrm{rank}(T) + \mathrm{nullity}(T) = \dim(U).$$

Proof. Since U is finite-dimensional and $\ker(T)$ is a subspace of U, $\ker(T)$ is finite-dimensional. Let $\mathrm{nullity}(T) = s$ and let $\mathbf{e}_1, \ldots, \mathbf{e}_s$ be a basis of $\ker(T)$. By Theorem 3.6, we can extend $\mathbf{e}_1, \ldots, \mathbf{e}_s$ to a basis $\mathbf{e}_1, \ldots, \mathbf{e}_s, \mathbf{f}_1, \ldots, \mathbf{f}_r$ of U. Then $\dim(U) = s + r$, so to prove the theorem we have to prove that $\dim(\mathrm{im}(T)) = r$.

Clearly $T(\mathbf{e}_1), \ldots, T(\mathbf{e}_s), T(\mathbf{f}_1), \ldots, T(\mathbf{f}_r)$ span $\mathrm{im}(T)$, and since

$$T(\mathbf{e}_1) = \cdots = T(\mathbf{e}_s) = \mathbf{0}_V$$

this implies that $T(\mathbf{f}_1), \ldots, T(\mathbf{f}_r)$ span $\mathrm{im}(T)$. We shall show that $T(\mathbf{f}_1), \ldots, T(\mathbf{f}_r)$ are linearly independent.

Suppose that, for some scalars α_i, we have

$$\alpha_1 T(\mathbf{f}_1) + \cdots + \alpha_r T(\mathbf{f}_r) = \mathbf{0}_V.$$

Then $T(\alpha_1 \mathbf{f}_1 + \cdots + \alpha_r \mathbf{f}_r) = \mathbf{0}_V$, so $\alpha_1 \mathbf{f}_1 + \cdots + \alpha_r \mathbf{f}_r \in \ker(T)$. But $\mathbf{e}_1, \ldots, \mathbf{e}_s$ is a basis of $\ker(T)$, so there exist scalars β_i with

$$\alpha_1 \mathbf{f}_1 + \cdots + \alpha_r \mathbf{f}_r = \beta_1 \mathbf{e}_1 + \cdots + \beta_s \mathbf{e}_s \Longrightarrow \alpha_1 \mathbf{f}_1 + \cdots + \alpha_r \mathbf{f}_r - \beta_1 \mathbf{e}_1 - \cdots - \beta_s \mathbf{e}_s = \mathbf{0}_U.$$

But we know that $\mathbf{e}_1, \ldots, \mathbf{e}_s, \mathbf{f}_1, \ldots, \mathbf{f}_r$ form a basis of U, so they are linearly independent, and hence

$$\alpha_1 = \cdots = \alpha_r = \beta_1 = \cdots = \beta_s = 0,$$

and we have proved that $T(\mathbf{f}_1), \ldots, T(\mathbf{f}_r)$ are linearly independent.

Since $T(\mathbf{f}_1), \ldots, T(\mathbf{f}_r)$ both span $\mathrm{im}(T)$ and are linearly independent, they form a basis of $\mathrm{im}(U)$, and hence $\dim(\mathrm{im}(T)) = r$, which completes the proof. □

Examples Once again, we consider examples 1–7 above. Since we only want to deal with finite-dimensional spaces, we restrict to an $(n+1)$-dimensional space $K[x]_{\leq n}$ in examples 4 and 5, that is, we consider $T : \mathbb{R}[x]_{\leq n} \to \mathbb{R}[x]_{\leq n}$, $S_\alpha : K[x]_{\leq n} \to K[x]_{\leq n}$, and $E_\alpha : K[x]_{\leq n} \to K$ correspondingly. Let $n = \dim(U) = \dim(V)$ in 6 and 7.

Example	rank(T)	nullity(T)	dim(U)
1	2	1	3
2	2	0	2
3	2	0	2
4	n	1	$n+1$
5 S_α	$n+1$	0	$n+1$
5 E_α	1	n	$n+1$
6	n	0	n
7	0	n	n

Corollary 5.6. *Let* $T\colon U \to V$ *be a linear map, and suppose that* $\dim(U) = \dim(V) = n$*. Then the following properties of* T *are equivalent:*

(i) T *is surjective;*

(ii) $\mathrm{rank}(T) = n$*;*

(iii) $\mathrm{nullity}(T) = 0$*;*

(iv) T *is injective;*

(v) T *is bijective;*

Proof. That T is surjective means precisely that $\mathrm{im}(T) = V$, so (i) $\Rightarrow$ (ii). But if $\mathrm{rank}(T) = n$, then $\dim(\mathrm{im}(T)) = \dim(V)$ so (by Corollary 3.9) a basis of $\mathrm{im}(T)$ is a basis of V, and hence $\mathrm{im}(T) = V$. Thus (ii) $\Leftrightarrow$ (i).

That (ii) $\Leftrightarrow$ (iii) follows directly from Theorem 5.5.

The equivalence (iii) $\Leftrightarrow$ (iv) is Proposition 5.4.

Finally, (v) is equivalent to (i) and (iv), which we have shown are equivalent to each other. $\square$

Definition. If the conditions in the above corollary are met, then T is called a *non-singular* linear map. Otherwise, T is called *singular*. Notice that the terms singular and non-singular are only used for linear maps $T\colon U \to V$ for which U and V have *the same dimension*.

5.4 Operations on linear maps

We can define the operations of *addition*, *scalar multiplication* and *composition* on linear maps.

Let $T_1\colon U \to V$ and $T_2\colon U \to V$ be two linear maps, and let $\alpha \in K$ be a scalar.

Definition (Addition of linear maps)**.** We define a map

$$T_1 + T_2\colon U \to V$$

by the rule $(T_1 + T_2)(\mathbf{u}) = T_1(\mathbf{u}) + T_2(\mathbf{u})$ for $\mathbf{u} \in U$.

Definition (Scalar multiplication of linear maps)**.** We define a map

$$\alpha T_1\colon U \to V$$

by the rule $(\alpha T_1)(\mathbf{u}) = \alpha T_1(\mathbf{u})$ for $\mathbf{u} \in U$.

Now let $T_1\colon U \to V$ and $T_2\colon V \to W$ be two linear maps.

Definition (Composition of linear maps)**.** We define a map

$$T_2T_1\colon U \to W$$

by $(T_2T_1)(\mathbf{u}) = T_2(T_1(\mathbf{u}))$ for $\mathbf{u} \in U$.

In particular, we define $T^2 = TT$ and $T^{i+1} = T^iT$ for $i > 2$ whenever T is a map from U to itself.

It is routine to check that $T_1 + T_2$, αT_1 and T_2T_1 are themselves all linear maps (but you should do it!).

For fixed vector spaces U and V over K, we denote by $\mathrm{Hom}_K(U, V)$ the set of all linear maps from U to V. Note that the operations of addition and scalar multiplication on elements of $\mathrm{Hom}_K(U, V)$, i.e. linear maps, that we just defined make $\mathrm{Hom}_K(U, V)$ into a vector space over K.

Given a vector space U over a field K, the vector space $U^* = \mathrm{Hom}_K(U, K)$ plays a special role. It is often called the *dual space* or *the space of covectors* of U. One can think of coordinates as elements of U^*. Indeed, let $\mathbf{e}_i$ be a basis of U. Every $\mathbf{x} \in U$ can be uniquely written as

$$\mathbf{x} = \alpha_1 \mathbf{e}_1 + \dots \alpha_n \mathbf{e}_n, \ \alpha_i \in K.$$

The elements α_i depend on $\mathbf{x}$ as well as on a choice of the basis, so for each i one can write the coordinate function

$$\mathbf{e}^i \colon U \to K, \ \mathbf{e}^i(\mathbf{x}) = \alpha_i.$$

It is routine to check that $\mathbf{e}^i$ is a linear map, and indeed the functions $\mathbf{e}^i$ form a basis of the dual space U^*.

6 Matrices

The material in this section will be familiar to many of you already, at least when K is the field of real numbers.

Definition. Let K be a field and $m, n \in \mathbb{N}$. An *$m \times n$ matrix* A over K is an $m \times n$ rectangular array of numbers (i.e. scalars) in K. The entry in row i and column j is often written α_{ij}. (We use the corresponding Greek letter.) We write $A = (\alpha_{ij})$ to make things clear.

For example, we could take

$$K = \mathbb{R}, \qquad m = 3, \ n = 4, \qquad A = (\alpha_{ij}) = \begin{pmatrix} 2 & -1 & -\pi & 0 \\ 3 & -3/2 & 0 & 6 \\ -1.23 & 0 & 10^{10} & 0 \end{pmatrix},$$

and then $\alpha_{13} = -\pi$, $\alpha_{33} = 10^{10}$, $\alpha_{34} = 0$, and so on.

Having defined what matrices are, we want to be able add them, multiply them by scalars, and multiply them by each other. You probably already know how to do this, but we will define these operations anyway.

Definition (Addition of matrices)**.** Let $A = (\alpha_{ij})$ and $B = (\beta_{ij})$ be two $m \times n$ matrices over K. We define $A + B$ to be the $m \times n$ matrix $C = (\gamma_{ij})$, where $\gamma_{ij} = \alpha_{ij} + \beta_{ij}$ for all i, j.

Example.

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

Definition (Scalar multiplication of matrices)**.** Let $A = (\alpha_{ij})$ be an $m \times n$ matrix over K and let $\beta \in K$ be a scalar. We define the scalar multiple βA to be the $m \times n$ matrix $C = (\gamma_{ij})$, where $\gamma_{ij} = \beta \alpha_{ij}$ for all i, j.

Definition (Multiplication of matrices)**.** Let $A = (\alpha_{ij})$ be an $l \times m$ matrix over K and let $B = (\beta_{ij})$ be an $m \times n$ matrix over K. The product AB is an $l \times n$ matrix $C = (\gamma_{ij})$ where, for $1 \le i \le l$ and $1 \le j \le n$,

$$\gamma_{ij} = \sum_{k=1}^{m} \alpha_{ik} \beta_{kj} = \alpha_{i1}\beta_{1j} + \alpha_{i2}\beta_{2j} + \dots + \alpha_{im}\beta_{mj}.$$

It is essential that the number m of columns of A is equal to the number of rows of B; otherwise AB makes no sense.

If you are familiar with scalar products of vectors, note also that γ_{ij} is the scalar product of the ith row of A with the jth column of B.

Example. Let

$$A = \begin{pmatrix} 2 & 3 & 4 \\ 1 & 6 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 6 \\ 3 & 2 \\ 1 & 9 \end{pmatrix}.$$

Then

$$AB = \begin{pmatrix} 2\times 2+3\times 3+4\times 1 & 2\times 6+3\times 2+4\times 9 \\ 1\times 2+6\times 3+2\times 1 & 1\times 6+6\times 2+2\times 9 \end{pmatrix} = \begin{pmatrix} 17 & 54 \\ 22 & 36 \end{pmatrix},$$

$$BA = \begin{pmatrix} 10 & 42 & 20 \\ 8 & 21 & 16 \\ 11 & 57 & 22 \end{pmatrix}.$$

Let $C = \begin{pmatrix} 2 & 3 & 1 \\ 6 & 2 & 9 \end{pmatrix}$. Then AC and CA are not defined.

Let $D = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$. Then AD is not defined, but $DA = \begin{pmatrix} 4 & 15 & 8 \\ 1 & 6 & 2 \end{pmatrix}$.

Proposition 6.1. *Matrices satisfy the following laws whenever the sums and products involved are defined:*

(i) $A + B = B + A$;

(ii) $(A + B)C = AC + BC$;

(iii) $C(A + B) = CA + CB$;

(iv) $(\lambda A)B = \lambda(AB) = A(\lambda B)$;

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

Proof. These are all routine checks that the entries of the left-hand sides are equal to the corresponding entries on the right-hand side. Let us do (v) as an example.

Let A, B and C be $l \times m$, $m \times n$ and $n \times p$ matrices, respectively. Then $AB = D = (\delta_{ij})$ is an $l \times n$ matrix with $\delta_{ij} = \sum_{s=1}^{m} \alpha_{is}\beta_{sj}$, and $BC = E = (\varepsilon_{ij})$ is an $m \times p$ matrix with $\varepsilon_{ij} = \sum_{t=1}^{n} \beta_{it}\gamma_{tj}$. Then $(AB)C = DC$ and $A(BC) = AE$ are both $l \times p$ matrices, and we have to show that their coefficients are equal. The (i, j)-coefficient of DC is

$$\sum_{t=1}^{n} \delta_{it}\gamma_{tj} = \sum_{t=1}^{n}(\sum_{s=1}^{m} \alpha_{is}\beta_{st})\gamma_{tj} = \sum_{s=1}^{m} \alpha_{is}(\sum_{t=1}^{n} \beta_{st}\gamma_{tj}) = \sum_{s=1}^{m} \alpha_{is}\varepsilon_{sj}$$

which is the (i, j)-coefficient of AE. Hence $(AB)C = A(BC)$. □

There are some useful matrices to which we give names.

Definition. The $m \times n$ *zero matrix* $\mathbf{0}_{mn}$ over any field K has all of its entries equal to 0.

Definition. The $n \times n$ *identity matrix* $I_n = (\alpha_{ij})$ over any field K has $\alpha_{ii} = 1$ for $1 \le i \le n$, but $\alpha_{ij} = 0$ when $i \ne j$.

Example.

$$I_1 = (1), \quad I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Note that $I_nA = A$ for any $n \times m$ matrix A and $AI_n = A$ for any $m \times n$ matrix A.

The set of all $m \times n$ matrices over K will be denoted by $K^{m,n}$. Note that $K^{m,n}$ is itself a vector space over K using the operations of addition and scalar multiplication defined above, and it has dimension mn. (This should be obvious – is it?)

A $1 \times n$ matrix is called a *row vector*. We will regard $K^{1,n}$ as being the same as K^n.

A $n \times 1$ matrix is called a *column vector*. We will denote the the space $K^{n,1}$ of all column vectors by $K^{n,1}$. In matrix calculations, we will use $K^{n,1}$ more often than K^n.

7 Linear transformations and matrices

We shall see in this section that, for fixed choice of bases, there is a very natural one-one correspondence between linear maps and matrices, such that the operations on linear maps and matrices defined in Chapters 5 and 6 also correspond to each other. This is perhaps the most important idea in linear algebra, because it enables us to deduce properties of matrices from those of linear maps, and vice-versa. It also explains why we multiply matrices in the way we do.

7.1 Setting up the correspondence

Let $T\colon U \to V$ be a linear map, where $\dim(U) = n,\ \dim(V) = m$. Suppose that we are given a basis $E = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ of U and a basis $F = \{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$ of V.

Now, for $1 \le j \le n$, the vector $T(\mathbf{e}_j)$ lies in V, so $T(\mathbf{e}_j)$ can be written uniquely as a linear combination of $\mathbf{f}_1, \ldots, \mathbf{f}_m$. Let

$$\begin{aligned}
T(\mathbf{e}_1) &= \alpha_{11}\mathbf{f}_1 + \alpha_{21}\mathbf{f}_2 + \cdots + \alpha_{m1}\mathbf{f}_m \\
T(\mathbf{e}_2) &= \alpha_{12}\mathbf{f}_1 + \alpha_{22}\mathbf{f}_2 + \cdots + \alpha_{m2}\mathbf{f}_m \\
&\cdots \\
T(\mathbf{e}_n) &= \alpha_{1n}\mathbf{f}_1 + \alpha_{2n}\mathbf{f}_2 + \cdots + \alpha_{mn}\mathbf{f}_m
\end{aligned}$$

where the coefficients $\alpha_{ij} \in K$ (for $1 \le i \le m,\ 1 \le j \le n$) are uniquely determined. Putting it more compactly, we define scalars α_{ij} by

$$T(\mathbf{e}_j) = \sum_{i=1}^{m} \alpha_{ij}\mathbf{f}_i \text{ for } 1 \le j \le n.$$

The coefficients α_{ij} form an $m \times n$ matrix

$$A = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn} \end{pmatrix}$$

over K.

Definition. This matrix A is called the matrix of the linear map T with respect to the chosen bases E of U and F of V. We will denote A by $[F, T, E]$, or $[FTE]$.

In general, different choice of bases gives different matrices. We shall address this issue later in the course, in Section 11.

Notice the role of individual columns in A. The jth column of A consists of the coordinates of $T(\mathbf{e}_j)$ with respect to the basis $\mathbf{f}_1, \ldots, \mathbf{f}_m$ of V.

Theorem 7.1. *Let U, V be vector spaces over K of dimensions n, m, respectively. Then, for a given choice of bases of U and V, there is a one-one correspondence between the set $\mathrm{Hom}_K(U, V)$ of linear maps $U \to V$ and the set $K^{m,n}$ of $m \times n$ matrices over K.*

Proof. As we saw above, any linear map $T: U \to V$ determines an $m \times n$ matrix A over K.

Conversely, let $A = (\alpha_{ij})$ be an $m \times n$ matrix over K. Then, by the existence statement in Proposition 5.2, there is a linear map $T: U \to V$ with $T(\mathbf{e}_j) = \sum_{i=1}^{m} \alpha_{ij}\mathbf{f}_i$ for $1 \le j \le n$. This shows that the above correspondence is onto. Moreover, by the uniqueness statement in Proposition 5.2, this correspondence is injective. Thus it is one-one. $\square$

Examples Once again, we consider our examples from Section 5.

1. $T: \mathbb{R}^3 \to \mathbb{R}^2$, $T(\alpha, \beta, \gamma) = (\alpha, \beta)$. Usually, we choose the standard bases of K^m and K^n, which in this case are $\mathbf{e}_1 = (1, 0, 0)$, $\mathbf{e}_2 = (0, 1, 0)$, $\mathbf{e}_3 = (0, 0, 1)$ and $\mathbf{f}_1 = (1, 0)$, $\mathbf{f}_2 = (0, 1)$. We have $T(\mathbf{e}_1) = \mathbf{f}_1$, $T(\mathbf{e}_2) = \mathbf{f}_2$, $T(\mathbf{e}_3) = \mathbf{0}$, and the matrix is
$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$
But suppose we chose different bases, say $\mathbf{e}_1 = (1, 1, 1)$, $\mathbf{e}_2 = (0, 1, 1)$, $\mathbf{e}_3 = (1, 0, 1)$, and $\mathbf{f}_1 = (0, 1)$, $\mathbf{f}_2 = (1, 0)$. Then we have $T(\mathbf{e}_1) = (1, 1) = \mathbf{f}_1 + \mathbf{f}_2$, $T(\mathbf{e}_2) = (0, 1) = \mathbf{f}_1$, $T(\mathbf{e}_3) = (1, 0) = \mathbf{f}_2$, and the matrix is
$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

2. $T: \mathbb{R}^2 \to \mathbb{R}^2$, T is a rotation through θ anti-clockwise about the origin. We saw that $T(1, 0) = (\cos\theta, \sin\theta)$ and $T(0, 1) = (-\sin\theta, \cos\theta)$, so the matrix using the standard bases is
$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

3. $T: \mathbb{R}^2 \to \mathbb{R}^2$, T is a reflection through the line through the origin making an angle $\theta/2$ with the x-axis. We saw that $T(1, 0) = (\cos\theta, \sin\theta)$ and $T(0, 1) = (\sin\theta, -\cos\theta)$, so the matrix using the standard bases is
$$\begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}.$$

4. This time we take the differentiation map T from $\mathbb{R}[x]_{\le n}$ to $\mathbb{R}[x]_{\le n-1}$. Then, with respect to the bases $1, x, x^2, \ldots, x^n$ and $1, x, x^2, \ldots, x^{n-1}$ of $\mathbb{R}[x]_{\le n}$ and $\mathbb{R}[x]_{\le n-1}$, respectively, the matrix of T is
$$\begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & n-1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & n \end{pmatrix}.$$

5. Let $S_\alpha\colon K[x]_{\le n} \to K[x]_{\le n}$ be the shift. With respect to the basis $1, x, x^2, \dots, x^n$ of $K[x]_{\le n}$, we calculate $S_\alpha(x^n) = (x-\alpha)^n$. The binomial formula gives the matrix of S_α,
$$\begin{pmatrix} 1 & -\alpha & \alpha^2 & \cdots & (-1)^n\alpha^n \\ 0 & 1 & -2\alpha & \cdots & (-1)^{n-1}n\alpha^{n-1} \\ 0 & 0 & 1 & \cdots & (-1)^{n-2}\frac{n(n-1)}{2}\alpha^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -n\alpha \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$
In the same basis of $K[x]_{\le n}$ and the basis 1 of K, $E_\alpha(x^n) = \alpha^n$. The matrix of E_α is
$$\begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n-1} & \alpha^n \end{pmatrix}.$$

6. $T\colon V \to V$ is the identity map. Notice that $U = V$ in this example. Provided that we choose the same basis for U and V, then the matrix of T is the $n \times n$ identity matrix I_n. We shall be considering the situation where we use different bases for the domain and range of the identity map in Section 11.

7. $T\colon U \to V$ is the zero map. The matrix of T is the $m \times n$ zero matrix $\mathbf{0}_{mn}$, regardless of what bases we choose. (The coordinates of the zero vector are all zero in any basis.)

We now connect how a linear transformation acts on elements of a vector space to how its matrix acts on their coordinates.

For the given basis $\mathbf{e}_1, \dots, \mathbf{e}_n$ of U and a vector $\mathbf{u} = \lambda_1\mathbf{e}_1 + \cdots + \lambda_n\mathbf{e}_n \in U$, let $\underline{\mathbf{u}}$ denote the column vector
$$\underline{\mathbf{u}} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix} \in K^{n,1},$$
whose entries are the coordinates of $\mathbf{u}$ with respect to that basis. Similarly, for the given basis $\mathbf{f}_1, \dots, \mathbf{f}_m$ of V and a vector $\mathbf{v} = \mu_1\mathbf{f}_1 + \cdots + \mu_m\mathbf{f}_m \in V$, let $\underline{\mathbf{v}}$ denote the column vector
$$\underline{\mathbf{v}} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \in K^{m,1}$$
whose entries are the coordinates of $\mathbf{v}$ with respect to that basis.

Proposition 7.2. *Let $T\colon U \to V$ be a linear map, let E be a basis of U, and F a basis of V. Let $\underline{\mathbf{u}}$ and $\underline{\mathbf{v}}$ be the column vectors of coordinates of two vectors $\mathbf{u} \in U$ and $\mathbf{v} \in V$ with respect to the basis E and F respectively. Then $T(\mathbf{u}) = \mathbf{v}$ if and only if $[F,T,E]\underline{\mathbf{u}} = \underline{\mathbf{v}}$. In other words, $\underline{T(\mathbf{u})} = [F,T,E]\underline{\mathbf{u}}$.*

Proof. Writing the entries of the matrix $[F,T,E]$ as α_{ij}, we have
$$T(\mathbf{u}) = T(\sum_{j=1}^n \lambda_j\mathbf{e}_j) = \sum_{j=1}^n \lambda_j T(\mathbf{e}_j) = \sum_{j=1}^n \lambda_j(\sum_{i=1}^m \alpha_{ij}\mathbf{f}_i) = \sum_{i=1}^m(\sum_{j=1}^n \alpha_{ij}\lambda_j)\mathbf{f}_i = \sum_{i=1}^m \mu_i\mathbf{f}_i,$$
where $\mu_i = \sum_{j=1}^n \alpha_{ij}\lambda_j$ is the entry in the ith row of the column vector $[F,T,E]\underline{\mathbf{u}}$. This proves the result. $\square$

What is this proposition really telling us? One way of looking at it is this. Choosing a basis for U gives every vector in U a unique set of coordinates. Choosing a basis for V gives every vector in V a unique set of coordinates. Now applying the linear transformation T to $\mathbf{u} \in U$ is "the same" as multiplying its column vector of coordinates by the matrix $[F, T, E]$ representing T, as long as we interpret the resulting column vector as coordinates in V with respect to our chosen basis.

Of course, choosing different bases will change the matrix $[F, T, E]$, and will change the coordinates of both $\mathbf{u}$ and $\mathbf{v}$. But it will change all of these quantities in exactly the right way that the theorem still holds.

7.2 The correspondence between operations on linear maps and matrices

Let U, V and W be vector spaces over the same field K, let $\dim(U) = n$, $\dim(V) = m$, $\dim(W) = l$, and choose fixed bases $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of U and $\mathbf{f}_1, \ldots, \mathbf{f}_m$ of V, and $\mathbf{g}_1, \ldots, \mathbf{g}_l$ of W. All matrices of linear maps between these spaces will be written with respect to these bases.

We have defined addition and scalar multiplication of linear maps, and we have defined addition and scalar multiplication of matrices. We have also defined a way to associate a matrix to a linear map. It turns out that all these operations behave together in the way we might hope.

Proposition 7.3. *1. Let $T_1, T_2 \colon U \to V$ be linear maps with corresponding matrices A, B respectively. Then the matrix of $T_1 + T_2$ is $A + B$. In other words, we have*

$$[F, T_1 + T_2, E] = [F, T_1, E] + [F, T_2, E].$$

2. Let $T \colon U \to V$ be a linear map with corresponding matrix A and let $\lambda \in K$ be a scalar. Then the matrix of λT is λA. In other words, we have

$$[F, \lambda T, E] = \lambda [F, T, E].$$

Proof. These are both straightforward to check, using the definitions, as long as you keep your wits about you. Checking them is a useful exercise, and you should do it. □

Note that the above two properties imply that the natural correspondence between linear maps and matrices is actually itself a linear map from $\mathrm{Hom}_K(U, V)$ to $K^{m,n}$.

Composition of linear maps corresponds to matrix multiplication. This time the correspondence is less obvious, and we state it as a theorem.

Theorem 7.4. *Let $T_1 \colon V \to W$ and $T_2 \colon U \to V$ be linear maps. Fix bases E, F and G of the spaces U, V and W respectively, and let $A = (\alpha_{ij}) = [G, T_1, F]$ and $B = (\beta_{ij}) = [F, T_2, E]$ be the corresponding matrices. Then the matrix of the composite map $T_1 T_2 \colon U \to W$ is AB. In other words, we have*

$$[G, T_1 T_2, E] = [G, T_1, F][F, T_2, E].$$

Proof. Let $n := dim(U), m := dim(V), l := dim(W)$. Let AB be the $l \times n$ matrix (γ_{ij}). Then by the definition of matrix multiplication, we have $\gamma_{ik} = \sum_{j=1}^{m} \alpha_{ij} \beta_{jk}$ for $1 \le i \le l, 1 \le k \le n$.

Let us calculate the matrix of $T_1 T_2$. We have

$$T_1 T_2(\mathbf{e}_k) = T_1(T_2(\mathbf{e}_k)) = T_1(\sum_{j=1}^{m} \beta_{jk} \mathbf{f}_j) = \sum_{j=1}^{m} \beta_{jk} T_1(\mathbf{f}_j) = \sum_{j=1}^{m} \beta_{jk} \sum_{i=1}^{l} \alpha_{ij} \mathbf{g}_i$$

$$= \sum_{i=1}^{l} (\sum_{j=1}^{m} \alpha_{ij} \beta_{jk}) \mathbf{g}_i = \sum_{i=1}^{l} \gamma_{ik} \mathbf{g}_i,$$

so the matrix of T_1T_2 is $(\gamma_{ik}) = AB$ as claimed. $\square$

Exercise: Give another proof of the last theorem using Proposition 7.2.

Examples Let us look at some examples of matrices corresponding to the composition of two linear maps.

1. Let $R_\theta\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a rotation through an angle θ anti-clockwise about the origin. We have seen that the matrix of R_θ (using the standard basis) is $\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}$. Now clearly R_θ followed by R_ϕ is equal to $R_{\theta+\phi}$. We can check the corresponding result for matrices:

$$\begin{aligned}\begin{pmatrix}\cos\phi & -\sin\phi\\ \sin\phi & \cos\phi\end{pmatrix}&\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}\\ &= \begin{pmatrix}\cos\phi\cos\theta - \sin\phi\sin\theta & -\cos\phi\sin\theta - \sin\phi\cos\theta\\ \sin\phi\cos\theta + \cos\phi\sin\theta & -\sin\phi\sin\theta + \cos\phi\cos\theta\end{pmatrix}\\ &= \begin{pmatrix}\cos(\phi+\theta) & -\sin(\phi+\theta)\\ \sin(\phi+\theta) & \cos(\phi+\theta)\end{pmatrix}.\end{aligned}$$

   Note that in this case $T_1T_2 = T_2T_1$. This actually gives an alternative way of deriving the addition formulae for sin and cos.

2. Let R_θ be as in Example 1, and let $M_\theta\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a reflection through a line through the origin at an angle $\theta/2$ to the x-axis. We have seen that the matrix of M_θ is $\begin{pmatrix}\cos\theta & \sin\theta\\ \sin\theta & -\cos\theta\end{pmatrix}$. What is the effect of doing first R_θ and then M_ϕ? In this case, it might be easier (for some people) to work it out using the matrix multiplication! We have

$$\begin{aligned}\begin{pmatrix}\cos\phi & \sin\phi\\ \sin\phi & -\cos\phi\end{pmatrix}&\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}\\ &= \begin{pmatrix}\cos\phi\cos\theta + \sin\phi\sin\theta & -\cos\phi\sin\theta + \sin\phi\cos\theta\\ \sin\phi\cos\theta - \cos\phi\sin\theta & -\sin\phi\sin\theta - \cos\phi\cos\theta\end{pmatrix}\\ &= \begin{pmatrix}\cos(\phi-\theta) & \sin(\phi-\theta)\\ \sin(\phi-\theta) & -\cos(\phi-\theta)\end{pmatrix},\end{aligned}$$

   which is the matrix of $M_{\phi-\theta}$.

   We get a different result if we do first M_ϕ and then R_θ. What do we get then?

7.3 Linear equations and the inverse image problem

The study and solution of systems of simultaneous linear equations is the main motivation behind the development of the theory of linear algebra and of matrix operations. Let us consider a system of m equations in n unknowns $x_1, x_2 \dots x_n$, where $m, n \geq 1$.

$$\left\{\begin{array}{ccccccccc}\alpha_{11}x_1 & + & \alpha_{12}x_2 & + & \cdots & + & \alpha_{1n}x_n & = & \beta_1\\ \alpha_{21}x_1 & + & \alpha_{22}x_2 & + & \cdots & + & \alpha_{2n}x_n & = & \beta_2\\ & & & & & & & \vdots & \\ \alpha_{m1}x_1 & + & \alpha_{m2}x_2 & + & \cdots & + & \alpha_{mn}x_n & = & \beta_m\end{array}\right. \tag{1}$$

All coefficients α_{ij} and β_i belong to K. Solving this system means finding all collections $x_1, x_2 \dots x_n \in K$ such that the equations (1) hold.

Let $A = (\alpha_{ij}) \in K^{m,n}$ be the $m \times n$ matrix of coefficients. The crucial step is to introduce the column vectors

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in K^{n,1} \qquad \text{and} \qquad \mathbf{b} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} \in K^{m,1}.$$

This allows us to rewrite system (1) as a single equation

$$A\mathbf{x} = \mathbf{b} \tag{2}$$

where the coefficient A is a matrix, the right hand side $\mathbf{b}$ is a vector in $K^{m,1}$ and the unknown $\mathbf{x}$ is a vector $K^{n,1}$.

Using the notation of linear maps, we have just reduced solving a system of linear equations to the *inverse image* problem. That is, given a linear map $T \colon U \to V$, and a fixed vector $\mathbf{v} \in V$, find all $\mathbf{u} \in U$ such that $T(\mathbf{u}) = \mathbf{v}$.

In fact, these two problems are equivalent! In the opposite direction, let us first forget all about A, $\mathbf{x}$ and $\mathbf{b}$, and suppose that we are given an inverse image problem to solve. Choose bases E of U and F of V and denote by A the corresponding matrix $[F, T, E]$. Also denote the row vector of coordinates of $\mathbf{u}$ (with respect to E) by $\mathbf{x}$ and the row vector of coordinates of $\mathbf{v}$ (with respect to F) by $\mathbf{b}$. Proposition 7.2 then says that $T(\mathbf{u}) = \mathbf{v}$ if and only if $A\mathbf{x} = \mathbf{b}$. This reduces the inverse image problem to solving a system of linear equations.

Let us make several easy observations about the inverse image problem.

The case when $\mathbf{v} = \mathbf{0}$ or, equivalently when $\beta_i = 0$ for $1 \le i \le m$, is called the *homogeneous* case. Here the set of solutions is $\{\mathbf{u} \in U \mid T(\mathbf{u}) = \mathbf{0}\}$, which is precisely the kernel $\ker(T)$ of T. The corresponding set of column vectors $\mathbf{x} \in K^{n,1}$ with $A\mathbf{x} = \mathbf{0}$ is called the *nullspace* of the matrix A. These column vectors are the coordinates of the vectors in the kernel of T, with respect to our chosen basis for U. So the nullity of A is the dimension of its nullspace.

In general, it is easy to see (and you should work out the details) that if $\mathbf{x}$ is one solution to a system of equations, then the complete set of solutions is equal to

$$\mathbf{x} + \mathrm{nullspace}(A) = \{\mathbf{x} + \mathbf{y} \mid \mathbf{y} \in \mathrm{nullspace}(A)\}.$$

It is possible that there are no solutions at all; this occurs when $\mathbf{v} \notin \mathrm{im}(T)$. If there are solutions, then there is a unique solution precisely when $\ker(T) = \{\mathbf{0}\}$, or equivalently when $\mathrm{nullspace}(A) = \{\mathbf{0}\}$. If the field K is infinite and there are solutions but $\ker(T) \neq \{\mathbf{0}\}$, then there are infinitely many solutions.

Now we would like to develop methods for solving the inverse image problem.

8 Elementary operations and the rank of a matrix

8.1 Gauss transformations

There are two standard high school methods for solving linear systems: the *substitution method* (where you express variables in terms of the other variables and substitute the result in the remaining equations) and the *elimination method* (sometimes called the *Gauss method*). The latter is usually faster, so we will concentrate on it. Let us recall how it is done.

Examples Here are some examples of solving systems of linear equations by the elimination method.

1.

$$
\begin{aligned}
2x + y &= 1 && (1)\\
4x + 2y &= 1 && (2)
\end{aligned}
$$

Replacing (2) by (2) $-$ 2 $\times$ (1) gives $0 = -1$. This means that there are *no solutions*.

2.

$$
\begin{aligned}
2x + y &= 1 && (1)\\
4x + y &= 1 && (2)
\end{aligned}
$$

Replacing (2) by (2) $-$ (1) gives $2x = 0$, and so $x = 0$. Replacing (1) by (1) $-$ 2 $\times$ (new 2) gives $y = 1$. Thus, $(0, 1)$ is *a unique solution*.

3.

$$
\begin{aligned}
2x + y &= 1 && (1)\\
4x + 2y &= 2 && (2)
\end{aligned}
$$

This time (2) $-$ 2 $\times$ (1) gives $0 = 0$, so (2) is redundant.

After reduction, there is no equation with leading term y, which means that y can take on any value, say $y = \alpha$. The first equation determines x in terms of y, giving $x = (1 - \alpha)/2$. So the *general solution* is $(x, y) = ((1 - \alpha)/2, \alpha)$, meaning that for each $\alpha \in \mathbb{R}$ we find one solution (x, y). There are *infinitely many solutions*.

Notice also that one solution is $(x, y) = (1/2, 0)$, and the general solution can be written as $(x, y) = (1/2, 0) + \alpha(-1/2, 1)$, where $\alpha(-1/2, 1)$ is the solution of the corresponding *homogeneous system* $2x + y = 0;\ 4x + 2y = 0$.

4.

$$
\begin{array}{rcrcrcll}
x & + & y & + & z & = & 1 & \qquad (1)\\
x & + & & & z & = & 2 & \qquad (2)\\
x & - & y & + & z & = & 3 & \qquad (3)\\
3x & + & y & + & 3z & = & 5 & \qquad (4)
\end{array}
$$

Now replacing (2) by (2) $-$ (1) and then multiplying by -1 gives $y = -1$. Replacing (3) by (3) $-$ (1) gives $-2y = 2$, and replacing (4) by (4) $-$ 3 $\times$ (1) also gives $-2y = 2$. So (3) and (4) both then reduce to $0 = 0$, and they are redundant.

z does not occur as a leading term, so it can take any value, say α, and then (2) gives $y = -1$ and (1) gives $x = 1 - y - z = 2 - \alpha$, so the general solution is

$$(x, y, z) = (2 - \alpha, -1, \alpha) = (2, -1, 0) + \alpha(-1, 0, 1).$$

8.2 Elementary row operations

Many types of calculations with matrices can be carried out in a computationally efficient manner by the use of certain types of operations on rows and columns. We shall see a little later that these are really the same as the operations used in solving sets of simultaneous linear equations.

Let A be an $m \times n$ matrix over K with rows $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m \in K^{1,n}$. The three types of elementary row operations on A are defined as follows.

(R1) For some $i \neq j$, add a multiple of $\mathbf{r}_j$ to $\mathbf{r}_i$.

$$\text{Example:} \quad \begin{pmatrix} 3 & 1 & 9 \\ 4 & 6 & 7 \\ 2 & 5 & 8 \end{pmatrix} \xrightarrow{\mathbf{r}_3 \to \mathbf{r}_3 - 3\mathbf{r}_1} \begin{pmatrix} 3 & 1 & 9 \\ 4 & 6 & 7 \\ -7 & 2 & -19 \end{pmatrix}$$

(R2) Interchange two rows

(R3) Multiply a row by a *non-zero* scalar.

$$\text{Example:} \quad \begin{pmatrix} 2 & 0 & 5 \\ 1 & -2 & 3 \\ 5 & 1 & 2 \end{pmatrix} \xrightarrow{\mathbf{r}_2 \to 4\mathbf{r}_2} \begin{pmatrix} 2 & 0 & 5 \\ 4 & -8 & 12 \\ 5 & 1 & 2 \end{pmatrix}$$

8.3 The augmented matrix

We would like to make the process of solving a system of linear equations more mechanical by forgetting about the variable names w, x, y, z, etc. and doing the whole operation as a matrix calculation. For this, we use the *augmented matrix* of the system of equations, which is constructed by "glueing" an extra column on the right-hand side of the matrix representing the linear transformation, as follows. For the system $A\underline{\mathbf{x}} = \underline{\beta}$ of m equations in n unknowns, where A is the $m \times n$ matrix (α_{ij}) is defined to be the $m \times (n+1)$ matrix

$$A = \left(\begin{array}{cccc|c} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} & \beta_1 \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} & \beta_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn} & \beta_m \end{array}\right).$$

The vertical line in the matrix is put there just to remind us that the rightmost column arised from the constants on the right hand side of the equations.

Let us look at the following system of linear equations over $\mathbb{R}$: suppose that we want to find all $w, x, y, z \in \mathbb{R}$ satisfying the equations.

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

Elementary row operations on A are precisely Gauss transformations of the corresponding linear system. Thus, the solution can be carried out mechanically as follows:

Matrix	**Operation(s)**
$\left(\begin{array}{rrrr\|r} 2 & -1 & 4 & -1 & 1 \\ 1 & 2 & 1 & 1 & 2 \\ 1 & -3 & 3 & -2 & -1 \\ -3 & -1 & -5 & 0 & -3 \end{array}\right)$	$\mathbf{r}_1 \to \mathbf{r}_1/2$
$\left(\begin{array}{rrrr\|r} 1 & -1/2 & 2 & -1/2 & 1/2 \\ 1 & 2 & 1 & 1 & 2 \\ 1 & -3 & 3 & -2 & -1 \\ -3 & -1 & -5 & 0 & -3 \end{array}\right)$	$\mathbf{r}_2 \to \mathbf{r}_2 - \mathbf{r}_1$, $\mathbf{r}_3 \to \mathbf{r}_3 - \mathbf{r}_1$, $\mathbf{r}_4 \to \mathbf{r}_4 + 3\mathbf{r}_1$

Matrix	Operation(s)
$\left(\begin{array}{cccc\|c} 1 & -1/2 & 2 & -1/2 & 1/2 \\ 0 & 5/2 & -1 & 3/2 & 3/2 \\ 0 & -5/2 & 1 & -3/2 & -3/2 \\ 0 & -5/2 & 1 & -3/2 & -3/2 \end{array}\right)$	$\mathbf{r}_3 \to \mathbf{r}_3 + \mathbf{r}_2$, $\mathbf{r}_4 \to \mathbf{r}_4 + \mathbf{r}_2$
$\left(\begin{array}{cccc\|c} 1 & -1/2 & 2 & -1/2 & 1/2 \\ 0 & 5/2 & -1 & 3/2 & 3/2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$	$\mathbf{r}_2 \to 2\mathbf{r}_2/5$
$\left(\begin{array}{cccc\|c} 1 & -1/2 & 2 & -1/2 & 1/2 \\ 0 & 1 & -2/5 & 3/5 & 3/5 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$	$\mathbf{r}_1 \to \mathbf{r}_1 + \mathbf{r}_2/2$
$\left(\begin{array}{cccc\|c} 1 & 0 & 9/5 & -1/5 & 4/5 \\ 0 & 1 & -2/5 & 3/5 & 3/5 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$	

The original system has been transformed to the following equivalent system, that is, both systems have the same solutions.

$$\left\{\begin{array}{rcrcrcl} w & + & 9y/5 & - & z/5 & = & 4/5 \\ x & - & 2y/5 & + & 3z/5 & = & 3/5 \end{array}\right.$$

In a solution to the latter system, variables y and z can take arbitrary values in $\mathbb{R}$; say $y = \alpha$, $z = \beta$. Then the equations tell us that $w = -9\alpha/5 + \beta/5 + 4/5$ and $x = 2\alpha/5 - 3\beta/5 + 3/5$ (be careful to get the signs right!), and so the complete set of solutions is

$$\begin{aligned} (w, x, y, z) &= (-9\alpha/5 + \beta/5 + 4/5, 2\alpha/5 - 3\beta/5 + 3/5, \alpha, \beta) \\ &= (4/5, 3/5, 0, 0) + \alpha(-9/5, 2/5, 1, 0) + \beta(1/5, -3/5, 0, 1). \end{aligned}$$

8.4 Row reducing a matrix

Let $A = (\alpha_{ij})$ be an $m \times n$ matrix over the field K. For the ith row, let $c(i)$ denote the position of the first (leftmost) non-zero entry in that row. In other words, $\alpha_{i,c(i)} \neq 0$ while $\alpha_{ij} = 0$ for all $j < c(i)$. It will make things a little easier to write if we use the convention that $c(i) = \infty$ if the ith row is entirely zero.

We will describe a procedure, analogous to solving systems of linear equations by elimination, which starts with a matrix, performs certain row operations, and finishes with a new matrix in a special form. After applying this procedure, the resulting matrix $A = (\alpha_{ij})$ will have the following properties.

(i) All zero rows are below all non-zero rows.

(ii) Let $\mathbf{r}_1, \ldots, \mathbf{r}_s$ be the non-zero rows. Then each $\mathbf{r}_i$ with $1 \leq i \leq s$ has 1 as its first non-zero entry. In other words, $\alpha_{i,c(i)} = 1$ for all $i \leq s$.

(iii) The first non-zero entry of each row is strictly to the right of the first non-zero entry of the row above: that is, $c(1) < c(2) < \cdots < c(s)$.

(iv) If row i is non-zero, then all entries below the first non-zero entry of row i are zero: $\alpha_{k,c(i)} = 0$ for all $k > i$.

Definition. A matrix satisfying properties (i)–(iv) above is said to be in *upper echelon form*.

Example. The matrix we came to at the end of the previous example was in upper echelon form.

There is a stronger version of the last property:

(v) If row i is non-zero, then all entries both above and below the first non-zero entry of row i are zero: $\alpha_{k,c(i)} = 0$ for all $k \neq i$.

Definition. A matrix satisfying properties (i)–(v) is said to be in *row reduced form*.

An upper echelon form of a matrix will be used later to calculate the rank of a matrix. The row reduced form (the use of the definite article is intended: this form is, indeed, unique, though we shall not prove this) is used to solve systems of linear equations. In this light, the following theorem says that every system of linear equations can be solved by the Gauss (Elimination) method.

Theorem 8.1. *Every matrix can be brought to row reduced form by elementary row transformations.*

Proof. We describe an algorithm to achieve this. For a formal proof, we have to show:

(i) after termination the resulting matrix has a row reduced form;

(ii) the algorithm terminates after finitely many steps.

Both of these statements are clear from the nature of the algorithm. Make sure that you understand why they are clear!

At any stage in the procedure we will be looking at the entry α_{ij} in a particular position (i, j) of the matrix. We will call (i, j) the *pivot* position, and α_{ij} the *pivot* entry. We start with $(i, j) = (1, 1)$ and proceed as follows.

1. If α_{ij} and all entries below it in its column are zero (i.e. if $\alpha_{kj} = 0$ for all $k \geq i$), then move the pivot one place to the right, to $(i, j + 1)$ and repeat Step 1, or terminate if $j = n$.

2. If $\alpha_{ij} = 0$ but $\alpha_{kj} \neq 0$ for some $k > i$ then apply row operation `(R2)` to interchange $\mathbf{r}_i$ and $\mathbf{r}_k$.

3. At this stage $\alpha_{ij} \neq 0$. If $\alpha_{ij} \neq 1$, then apply row operation `(R3)` to multiply $\mathbf{r}_i$ by α_{ij}^{-1}.

4. At this stage $\alpha_{ij} = 1$. If, for any $k \neq i$, $\alpha_{kj} \neq 0$, then apply row operation `(R1)`, and subtract α_{kj} times $\mathbf{r}_i$ from $\mathbf{r}_k$.

5. At this stage, $\alpha_{kj} = 0$ for all $k \neq i$. If $i = m$ or $j = n$ then terminate. Otherwise, move the pivot diagonally down to the right to $(i+1, j+1)$, and go back to Step 1.

$\square$

If one needs only an upper echelon form, this can done faster by replacing steps 4 and 5 with weaker and faster steps as follows.

4a. At this stage $\alpha_{ij} = 1$. If, for any $k > i$, $\alpha_{kj} \neq 0$, then apply `(R1)`, and subtract α_{kj} times $\mathbf{r}_i$ from $\mathbf{r}_k$.

5a. At this stage, $\alpha_{kj} = 0$ for all $k > i$. If $i = m$ or $j = n$ then terminate. Otherwise, move the pivot diagonally down to the right to $(i+1, j+1)$, and go back to Step 1.

In the example below, we find an upper echelon form of a matrix by applying the faster algorithm. The number in the 'Step' column refers to the number of the step applied in the description of the procedure above.

Example Let $A = \begin{pmatrix} 0 & 0 & 1 & 2 & 1 \\ 2 & 4 & 2 & -4 & 2 \\ 3 & 6 & 3 & -6 & 3 \\ 1 & 2 & 3 & 3 & 3 \end{pmatrix}$.

Matrix	**Pivot**	**Step**	**Operation**
$\begin{pmatrix} 0 & 0 & 1 & 2 & 1 \\ 2 & 4 & 2 & -4 & 2 \\ 3 & 6 & 3 & -6 & 3 \\ 1 & 2 & 3 & 3 & 3 \end{pmatrix}$	$(1, 1)$	**2**	$\mathbf{r}_1 \leftrightarrow \mathbf{r}_2$
$\begin{pmatrix} 2 & 4 & 2 & -4 & 2 \\ 0 & 0 & 1 & 2 & 1 \\ 3 & 6 & 3 & -6 & 3 \\ 1 & 2 & 3 & 3 & 3 \end{pmatrix}$	$(1, 1)$	**3**	$\mathbf{r}_1 \to \mathbf{r}_1/2$
$\begin{pmatrix} 1 & 2 & 1 & -2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 3 & 6 & 3 & -6 & 3 \\ 1 & 2 & 3 & 3 & 3 \end{pmatrix}$	$(1, 1)$	**4**	$\mathbf{r}_3 \to \mathbf{r}_3 - 3\mathbf{r}_1$ $\mathbf{r}_4 \to \mathbf{r}_4 - \mathbf{r}_1$
$\begin{pmatrix} 1 & 2 & 1 & -2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 5 & 2 \end{pmatrix}$	$(1, 1) \to (2, 2) \to (2, 3)$	**5, 1**	
$\begin{pmatrix} 1 & 2 & 1 & -2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 5 & 2 \end{pmatrix}$	$(2, 3)$	**4**	$\mathbf{r}_4 \to \mathbf{r}_4 - 2\mathbf{r}_2$
$\begin{pmatrix} 1 & 2 & 1 & -2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$	$(2, 3) \to (3, 4)$	**5, 2**	$\mathbf{r}_3 \leftrightarrow \mathbf{r}_4$
$\begin{pmatrix} 1 & 2 & 1 & -2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$	$(3, 4) \to (4, 5) \to$ **stop**	**5, 1**	

This matrix is now in upper echelon form.

8.5 Elementary column operations

In analogy to elementary row operations, one can introduce elementary column operations. Let A be an $m \times n$ matrix over K with columns $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n$ as above. The three types of elementary column operations on A are defined as follows.

`(C1)` For some $i \neq j$, add a multiple of $\mathbf{c}_j$ to $\mathbf{c}_i$.

$$\text{Example:} \quad \begin{pmatrix} 3 & 1 & 9 \\ 4 & 6 & 7 \\ 2 & 5 & 8 \end{pmatrix} \xrightarrow{\mathbf{c}_3 \to \mathbf{c}_3 - 3\mathbf{c}_1} \begin{pmatrix} 3 & 1 & 0 \\ 4 & 6 & -5 \\ 2 & 5 & 2 \end{pmatrix}$$

`(C2)` Interchange two columns.

`(C3)` Multiply a column by a *non-zero* scalar.

$$\text{Example:} \quad \begin{pmatrix} 2 & 0 & 5 \\ 1 & -2 & 3 \\ 5 & 1 & 2 \end{pmatrix} \xrightarrow{\mathbf{c}_2 \to 4\mathbf{c}_2} \begin{pmatrix} 2 & 0 & 5 \\ 1 & -8 & 3 \\ 5 & 4 & 2 \end{pmatrix}$$

Elementary column operations change a linear system and cannot be applied to solve a system of linear equations. However, they are useful for reducing a matrix to a very nice form.

Theorem 8.2. *By applying elementary row and column operations, a matrix can be brought into the block form*

$$\left(\begin{array}{c|c} I_s & \mathbf{0}_{s,n-s} \\ \hline \mathbf{0}_{m-s,s} & \mathbf{0}_{m-s,n-s} \end{array} \right),$$

where, as in Section 6, I_s denotes the $s \times s$ identity matrix, and $\mathbf{0}_{kl}$ the $k \times l$ zero matrix.

Proof. First, use elementary row operations to reduce A to row reduced form.

Now all $\alpha_{i,c(i)} = 1$. We can use these leading entries in each row to make all the other entries zero: for each $\alpha_{ij} \neq 0$ with $j \neq c(i)$, replace $\mathbf{c}_j$ with $\mathbf{c}_j - \alpha_{ij}\mathbf{c}_{c(i)}$.

Finally the only nonzero entries of our matrix are $\alpha_{i,c(i)} = 1$. Now for each number i starting from $i = 1$, exchange $\mathbf{c}_i$ and $\mathbf{c}_{c(i)}$, putting all the zero columns at the right-hand side. $\square$

Definition. The matrix in Theorem 8.2 is said to be in *row and column reduced form*, which is also called *Smith normal form.*

Let us look at an example of the second stage of procedure, that is, after reducing the matrix to the row reduced form.

Matrix	**Operation**
$\begin{pmatrix} 1 & 2 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$	$\mathbf{c}_2 \to \mathbf{c}_2 - 2\mathbf{c}_1$ $\mathbf{c}_5 \to \mathbf{c}_5 - \mathbf{c}_1$
$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$	$\mathbf{c}_2 \leftrightarrow \mathbf{c}_3$ $\mathbf{c}_5 \to \mathbf{c}_5 - 3\mathbf{c}_4$

Matrix	**Operation**
$\begin{pmatrix}1&0&0&0&0\\0&1&0&0&2\\0&0&0&1&0\\0&0&0&0&0\end{pmatrix}$	$\mathbf{c}_3 \leftrightarrow \mathbf{c}_4$ $\mathbf{c}_5 \to \mathbf{c}_5 - 2\mathbf{c}_2$
$\begin{pmatrix}1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&0&0\end{pmatrix}$	

Now we would like to discuss the number s that appears in Theorem 8.2, that is, the number of non-zero entries in the Smith normal form. Does the initial matrix uniquely determine this number? Although we have an algorithm for reducing a matrix to Smith normal form, there will be other sequences of row and column operations which also put the matrix into Smith normal form. Could we maybe end up with a different number of non-zero entries depending on the row and column operations used?

8.6 The rank of a matrix

Let A be an $m \times n$ matrix over K. We shall denote the m rows of A, which are row vectors in K^n, by $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m$, and similarly, we denote the n columns of A, which are column vectors in $K^{m,1}$, by $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n$.

Definition.

1. The *row space* of A is the subspace of K^n spanned by the rows $\mathbf{r}_1, \ldots, \mathbf{r}_m$ of A. The *row rank* of A is the dimension of the row space of A. Equivalently, by Corollary 3.9, the row rank of A is equal to the size of the largest linearly independent subset of $\mathbf{r}_1, \ldots, \mathbf{r}_m$.
2. The *column space* of A is the subspace of $K^{m,1}$ spanned by the columns $\mathbf{c}_1, \ldots, \mathbf{c}_n$ of A. The *column rank* of A is the dimension of the column space of A. Equivalently, the column rank of A is equal to the size of the largest linearly independent subset of $\mathbf{c}_1, \ldots, \mathbf{c}_n$.

There is no obvious reason why there should be any particular relationship between the row and column ranks, but in fact it will turn out that they are always equal.

Example Consider the matrix

$$A = \begin{array}{c} \begin{pmatrix} 1 & 2 & 0 & 1 & 1 \\ 2 & 4 & 1 & 3 & 0 \\ 4 & 8 & 0 & 4 & 4 \end{pmatrix} \begin{array}{l} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \mathbf{r}_3 \end{array} \\ \begin{array}{ccccc} \mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_3 & \mathbf{c}_4 & \mathbf{c}_5 \end{array} \end{array}$$

We can calculate the row and column ranks by applying the sifting process (described in Section 3) to the row and column vectors, respectively.

Doing rows first, $\mathbf{r}_1$ and $\mathbf{r}_2$ are linearly independent, but $\mathbf{r}_3 = 4\mathbf{r}_1$, so the row rank is 2.

Now doing columns, $\mathbf{c}_2 = 2\mathbf{c}_1$, $\mathbf{c}_4 = \mathbf{c}_1 + \mathbf{c}_3$ and $\mathbf{c}_5 = \mathbf{c}_1 - 2\mathbf{c}_3$, so the column rank is also 2.

We now show that the column rank is the same as the rank of the associated linear map T.

Theorem 8.3. *Suppose that the matrix A corresponds to a linear map $T : U \to V$. Then* rank(T) *is equal to the column rank of A.*

Proof. Let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ be a basis of U and $\mathbf{f}_1, \ldots, \mathbf{f}_n$ a basis of V. By the definition of the correspondence between linear maps and matrices, which we saw in Section 7.1, the columns $\mathbf{c}_1, \ldots, \mathbf{c}_n$ of A are precisely the column vectors of coordinates of the vectors $T(\mathbf{e}_1), \ldots, T(\mathbf{e}_n)$, with respect to the basis $\mathbf{f}_1, \ldots, \mathbf{f}_n$ of V.

We claim that the vectors $T(\mathbf{e}_1), \ldots, T(\mathbf{e}_n)$ span $\mathrm{im}(T)$. Indeed, for every $\mathbf{v} \in \mathrm{im}(T)$, there is $\mathbf{u} \in U$ such that $\mathbf{v} = T(\mathbf{u})$. Writing $\mathbf{u}$ in terms of the basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$, and using the definition of a linear map, we have

$$\mathbf{v} = T(\mathbf{u}) = T(\alpha_1\mathbf{e}_1 + \ldots + \alpha_n\mathbf{e}_n) = \alpha_1 T(\mathbf{e}_1) + \ldots + \alpha_n T(\mathbf{e}_n),$$

which means that $\mathbf{v}$ is a linear combination of the $T(\mathbf{e}_i)$, proving our claim.

By Theorem 3.5, the sifted subsequence of the vectors $T(\mathbf{e}_i)$ forms a basis of $\mathrm{im}(T)$. By definition, rank(T) is the size of that subsequence. Similarly, the sifted subsequence of the column vectors $\mathbf{c}_i$ forms a basis of the column space of A. By definition, the column rank of A is the size of that subsequence. Since the $\mathbf{c}_i$ are precisely the column vectors corresponding to the $T(\mathbf{e}_i)$, the sifted subsequences coincide, proving the statement. □

Theorem 8.4. *Applying elementary row operations* (R1), (R2) *or* (R3) *to a matrix does not change the row or column rank. The same is true for elementary column operations* (C1), (C2) *and* (C3).

Proof. We will prove first that the elementary row operations do not change either the row rank or column rank.

The row rank of a matrix A is the dimension of the row space of A, which is the space of linear combinations $\lambda_1\mathbf{r}_1 + \cdots + \lambda_m\mathbf{r}_m$ of the rows of A. It is easy to see that (R1), (R2) and (R3) do not change this space, so they do not change the row-rank. (But notice that the scalar in (R3) must be non-zero for this to be true!)

The column rank of $A = (\alpha_{ij})$ is the size of the largest linearly independent subset of $\mathbf{c}_1, \ldots, \mathbf{c}_n$. Let $\{\mathbf{c}_1, \ldots, \mathbf{c}_s\}$ be some subset of the set $\{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ of columns of A. (We have written this as though the subset consisted of the first s columns, but this is just to keep the notation simple; it could be any subset of the columns.)

Then $\mathbf{c}_1, \ldots, \mathbf{c}_s$ are linearly dependent if and only if there exist scalars $x_1, \ldots, x_s \in K$, not all zero, such that $x_1\mathbf{c}_1 + x_2\mathbf{c}_2 + \cdots + x_s\mathbf{c}_s = \mathbf{0}$. If we write out the m components of this vector equation, we get a system of m linear equations in the scalars x_i (which is why we have suddenly decided to call the scalars x_i rather than λ_i).

$$\begin{aligned}
\alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1s}x_s &= 0\\
\alpha_{21}x_1 + \alpha_{22}x_2 + \cdots + \alpha_{2s}x_s &= 0\\
&\vdots\\
\alpha_{m1}x_1 + \alpha_{m2}x_2 + \cdots + \alpha_{ms}x_s &= 0
\end{aligned}$$

Now if we perform (R1), (R2) or (R3) on A, then we perform the corresponding operation on this system of equations. That is, we add a multiple of one equation to another, we interchange two equations, or we multiply one equation by a non-zero scalar. None of these operations change the set of solutions of the equations. Hence if they have some solution with the x_i not all zero before the operation, then they have the same solution after the operation, and the other way round: if they have some non-zero solution after the operation, then they had the same solution before the operation. In other words, the elementary row operations do not change the linear dependence or independence of the set of columns $\{\mathbf{c}_1, \dots, \mathbf{c}_s\}$. Thus they do not change the size of the largest linearly independent subset of $\mathbf{c}_1, \dots, \mathbf{c}_n$, so they do not change the column rank of A.

The proof for the column operations `(C1)`, `(C2)` and `(C3)` is the same with rows and columns interchanged. $\square$

Corollary 8.5. *Let s be the number of non-zero rows in the Smith normal form of a matrix A (see Theorem 8.2). Then both the row rank of A and the column rank of A are equal to s.*

Proof. Since elementary operations preserve ranks, it suffices to find both ranks of a matrix in Smith normal form. But it is easy to see that the row space is precisely the space spanned by the first s standard vectors and hence has dimension s. Similarly the column space has dimension s. $\square$

In particular, Corollary 8.5 establishes that the row rank is always equal to the column rank. This allows us to forget this distinction. From now we shall just talk about *the rank of a matrix.*

Corollary 8.6. *The rank of a matrix A is equal to the number of non-zero rows after reducing A to upper echelon form or row reduced form.*

Proof. Firstly, note that the number of non-zero rows after reducing A to upper echelon form is the same as the number of non-zero rows after reducing to row reduced form.

Having row reduced A, we can reach the Smith normal form by applying some of the column operations `(C1)`, `(C2)` and `(C3)` (see the proof of Theorem 8.2). We claim that these operations cannot change the number of non-zero rows. Indeed, it is easy to check that if a row r_i has at least one non-zero entry, then it will still have at least one one non-zero entry after applying any of the operations `(C1)`, `(C2)` and `(C3)`, and if r_i has only zero entries then it will still have only zero entries after the operation.

This means that the number of non-zero rows in the row reduced form of A equals the number of non-zero rows in the Smith normal form of A. By Corollary 8.5 and the remark after it, this number is the rank of A.

$\square$

Corollary 8.6 gives the most efficient way of computing the rank of a matrix. For instance, let us look at $A = \begin{pmatrix} 1 & 2 & 0 & 1 & 1 \\ 2 & 4 & 1 & 3 & 0 \\ 4 & 8 & 1 & 5 & 2 \end{pmatrix}$.

Matrix	**Operation**
$\begin{pmatrix} 1 & 2 & 0 & 1 & 1 \\ 2 & 4 & 1 & 3 & 0 \\ 4 & 8 & 1 & 5 & 2 \end{pmatrix}$	$\mathbf{r}_2 \to \mathbf{r}_2 - 2\mathbf{r}_1$ $\mathbf{r}_3 \to \mathbf{r}_3 - 4\mathbf{r}_1$
$\begin{pmatrix} 1 & 2 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & -2 \\ 0 & 0 & 1 & 1 & -2 \end{pmatrix}$	$\mathbf{r}_3 \to \mathbf{r}_3 - \mathbf{r}_2$
$\begin{pmatrix} 1 & 2 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$	

Since the resulting matrix in upper echelon form has 2 nonzero rows, $\mathrm{rank}(A) = 2$.

8.7 The rank criterion

The following theorem is proved in Assignment Sheet 6.

Theorem 8.7. *Let A be the augmented $n \times (m+1)$ matrix of a linear system. Let B be the $n \times m$ matrix obtained from A by removing the last column. The system of linear equations has a solution if and only if* $\mathrm{rank}(A) = \mathrm{rank}(B)$.

9 The inverse of a linear transformation and of a matrix

9.1 Definitions

Let $T : U \to V$ be a linear map. If there is a linear map $T^{-1} : V \to U$ with $TT^{-1} = I_V$ and $T^{-1}T = I_U$ then T is said to be *invertible*, and T^{-1} is called the *inverse* of T.

Similarly, if A is an $n \times m$ matrix, and A^{-1} an $m \times n$ matrix such that $AA^{-1} = I_n$ and $A^{-1}A = I_m$, we call A *invertible*, and call A^{-1} the *inverse* of A.

Lemma 9.1. *Let A be a matrix corresponding to the linear map T. Then T is invertible if and only if A is invertible. The inverses T^{-1} and A^{-1} are unique.*

Proof. Recall that, under the bijection between matrices and linear maps, multiplication of matrices corresponds to composition of linear maps (Theorem 7.4). It now follows immediately from the definitions above that invertible matrices correspond to invertible linear maps. This establishes the first statement.

Since the inverse map of a bijection is unique, T^{-1} is unique. Under the bijection between matrices and linear maps, A^{-1} must be the matrix of T^{-1}. Thus, A^{-1} is unique as well. □

Theorem 9.2. *A linear map T is invertible if and only if T is non-singular. In particular, if T is invertible then $m = n$, so only square matrices can be invertible.*

See Corollary 5.6 for the definition of non-singular linear maps. We may also say that the matrix A is non-singular if T is; but by this theorem, this is equivalent to A being invertible.

Proof. If any function T has a left and right inverse, then it must be a bijection. Hence $\ker(T) = \{\mathbf{0}\}$ and $\mathrm{im}(T) = V$, so $\mathrm{nullity}(T) = 0$ and $\mathrm{rank}(T) = \dim(V) = m$. But by Theorem 5.5, we have

$$n = \dim(U) = \mathrm{rank}(T) + \mathrm{nullity}(T) = m + 0 = m$$

and we see from the definition that T is non-singular.

Conversely, if $n = m$ and T is non-singular, then by Corollary 5.6 T is a bijection, and so it has an inverse $T^{-1} : V \to U$ as a function. However, we still have to show that T^{-1} is a *linear* map. Let $\mathbf{v}_1, \mathbf{v}_2 \in V$. Then there exist $\mathbf{u}_1, \mathbf{u}_2 \in U$ with $T(\mathbf{u}_1) = \mathbf{v}_1$, $T(\mathbf{u}_2) = \mathbf{v}_2$. So $T(\mathbf{u}_1 + \mathbf{u}_2) = \mathbf{v}_1 + \mathbf{v}_2$ and hence $T^{-1}(\mathbf{v}_1 + \mathbf{v}_2) = \mathbf{u}_1 + \mathbf{u}_2$. If $\alpha \in K$, then

$$T^{-1}(\alpha\mathbf{v}_1) = T^{-1}(T(\alpha\mathbf{u}_1)) = \alpha\mathbf{u}_1 = \alpha T^{-1}(\mathbf{v}_1),$$

so T^{-1} is linear, which completes the proof. □

Example. Let $A = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 0 & 1 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & -2 \\ 0 & 1 \\ -2 & 5 \end{pmatrix}$. Then $AB = I_2$, but $BA \neq I_3$, so a non-square matrix can have a right inverse which is not a left inverse. However, it can be deduced from Corollary 5.6 that if A is a square $n \times n$ matrix and $AB = I_n$ then A is non-singular, and then by multiplying $AB = I_n$ on the left by A^{-1}, we see that $B = A^{-1}$ and so $BA = I_n$.

This technique of multiplying on the left or right by A^{-1} is often used for transforming matrix equations. If A is invertible, then $AX = B \iff X = A^{-1}B$ and $XA = B \iff X = BA^{-1}$.

Lemma 9.3. *If A and B are invertible $n \times n$ matrices, then AB is invertible, and $(AB)^{-1} = B^{-1}A^{-1}$.*

Proof. This is clear, because $ABB^{-1}A^{-1} = B^{-1}A^{-1}AB = I_n$. □

9.2 Matrix inversion by row reduction

Two methods for finding the inverse of a matrix will be studied in this course. The first, using row reduction, which we shall look at now, is an efficient practical method similar to that used by computer packages. The second, using determinants, is of more theoretical interest, and will be done later in Section 10.

First, we claim that if an $n \times n$ matrix A is invertible, then it has rank n. Indeed, this follows by combining Lemma 9.1, Theorem 9.2, Corollary 5.6 (ii) and Theorem 8.3. Consider the row reduced form $B = (\beta_{ij})$ of A. As we saw in Section 8.6, we have $\beta_{ic(i)} = 1$ for $1 \leq i \leq n$ (since $\mathrm{rank}(A) = \mathrm{rank}(B) = n$), where $c(1) < c(2) < \cdots < c(n)$, and this is only possible without any zero rows if $c(i) = i$ for $1 \leq i \leq n$. Then, since all other entries in column $c(i)$ are zero, we have $B = I_n$. We have therefore proved:

Proposition 9.4. *The row reduced form of an invertible $n \times n$ matrix A is I_n.*

To compute A^{-1}, we reduce A to its row reduced form I_n, using elementary row operations, while simultaneously applying the same row operations, but starting with the identity matrix I_n. It turns out that these operations transform I_n to A^{-1}.

In practice, we might not know whether or not A is invertible before we start, but we will find out while carrying out this procedure because, if A is not invertible, then its rank will be less than n, and it will not row reduce to I_n.

First we will do an example to demonstrate the method, and then we will explain why it works. In the table below, the row operations applied are given in the middle column. The results of applying them to the matrix

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 4 & 1 & 3 \\ 2 & 1 & 6 \end{pmatrix}$$

are given in the left column, and the results of applying them to I_3 in the right column. So A^{-1} should be the final matrix in the right column.

Matrix 1	**Operation(s)**	**Matrix 2**
$\begin{pmatrix} 3 & 2 & 1 \\ 4 & 1 & 3 \\ 2 & 1 & 6 \end{pmatrix}$		$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
$\downarrow$	$\mathbf{r}_1 \to \mathbf{r}_1/3$	$\downarrow$

Matrix 1	**Operation(s)**	**Matrix 2**
$\begin{pmatrix}1 & 2/3 & 1/3\\ 4 & 1 & 3\\ 2 & 1 & 6\end{pmatrix}$		$\begin{pmatrix}1/3 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}$
$\vdots$ $\downarrow$	$\mathbf{r}_2 \to \mathbf{r}_2 - 4\mathbf{r}_1$, $\mathbf{r}_3 \to \mathbf{r}_3 - 2\mathbf{r}_1$	$\vdots$ $\downarrow$
$\begin{pmatrix}1 & 2/3 & 1/3\\ 0 & -5/3 & 5/3\\ 0 & -1/3 & 16/3\end{pmatrix}$		$\begin{pmatrix}1/3 & 0 & 0\\ -4/3 & 1 & 0\\ -2/3 & 0 & 1\end{pmatrix}$
$\downarrow$	$\mathbf{r}_2 \to -3\mathbf{r}_2/5$	$\downarrow$
$\begin{pmatrix}1 & 2/3 & 1/3\\ 0 & 1 & -1\\ 0 & -1/3 & 16/3\end{pmatrix}$		$\begin{pmatrix}1/3 & 0 & 0\\ 4/5 & -3/5 & 0\\ -2/3 & 0 & 1\end{pmatrix}$
$\vdots$ $\downarrow$	$\mathbf{r}_1 \to \mathbf{r}_1 - 2\mathbf{r}_2/3$, $\mathbf{r}_3 \to \mathbf{r}_3 + \mathbf{r}_2/3$	$\vdots$ $\downarrow$
$\begin{pmatrix}1 & 0 & 1\\ 0 & 1 & -1\\ 0 & 0 & 5\end{pmatrix}$		$\begin{pmatrix}-1/5 & 2/5 & 0\\ 4/5 & -3/5 & 0\\ -2/5 & -1/5 & 1\end{pmatrix}$
$\downarrow$	$\mathbf{r}_3 \to \mathbf{r}_3/5$	$\downarrow$
$\begin{pmatrix}1 & 0 & 1\\ 0 & 1 & -1\\ 0 & 0 & 1\end{pmatrix}$		$\begin{pmatrix}-1/5 & 2/5 & 0\\ 4/5 & -3/5 & 0\\ -2/25 & -1/25 & 1/5\end{pmatrix}$
$\vdots$ $\downarrow$	$\mathbf{r}_1 \to \mathbf{r}_1 - \mathbf{r}_3$, $\mathbf{r}_2 \to \mathbf{r}_2 + \mathbf{r}_3$	$\vdots$ $\downarrow$
$\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}$		$\begin{pmatrix}-3/25 & 11/25 & -1/5\\ 18/25 & -16/25 & 1/5\\ -2/25 & -1/25 & 1/5\end{pmatrix}$

So

$$A^{-1} = \begin{pmatrix}-3/25 & 11/25 & -1/5\\ 18/25 & -16/25 & 1/5\\ -2/25 & -1/25 & 1/5\end{pmatrix}.$$

It is always a good idea to check the result afterwards. This is easier if we remove the common denominator 25, and we can then easily check that

$$\begin{pmatrix}3 & 2 & 1\\ 4 & 1 & 3\\ 2 & 1 & 6\end{pmatrix}\begin{pmatrix}-3 & 11 & -5\\ 18 & -16 & 5\\ -2 & -1 & 5\end{pmatrix} = \begin{pmatrix}-3 & 11 & -5\\ 18 & -16 & 5\\ -2 & -1 & 5\end{pmatrix}\begin{pmatrix}3 & 2 & 1\\ 4 & 1 & 3\\ 2 & 1 & 6\end{pmatrix} = \begin{pmatrix}25 & 0 & 0\\ 0 & 25 & 0\\ 0 & 0 & 25\end{pmatrix}$$

which confirms the result!

9.3 Elementary matrices

We shall now explain why the above method of calculating the inverse of a matrix works. Each elementary row operation on a matrix can be achieved by multiplying the matrix on the left by a corresponding matrix known as an *elementary matrix*. There are three types of these, all being slightly different from the identity matrix.

1. $E(n)^1_{\lambda,i,j}$ (where $i \neq j$) is the an $n \times n$ matrix equal to the identity, but with an additional non-zero entry λ in the (i,j) position.

2. $E(n)^2_{i,j}$ is the $n \times n$ identity matrix with its ith and jth rows interchanged.

3. $E(n)^3_{\lambda,i}$ (where $\lambda \neq 0$) is the $n \times n$ identity matrix with its (i,i) entry replaced by λ.

Example. Some elementary matrices:

$$E(3)^1_{\frac{1}{3},1,3} = \begin{pmatrix} 1 & 0 & \frac{1}{3} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \ E(4)^2_{2,4} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \ E(3)^3_{-4,3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -4 \end{pmatrix}.$$

Let A be any $m \times n$ matrix. Then $E(m)^1_{\lambda,i,j}A$ is the result we get by adding λ times the jth row of A to the ith row of A. Similarly $E(m)^2_{i,j}A$ is equal to A with its ith and jth rows interchanged, and $E(m)^3_{\lambda,i}$ is equal to A with its ith row multiplied by λ. You need to work out a few examples to convince yourself that this is true. For example

$$E(4)^1_{-2,4,2} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -2 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

So, in the matrix inversion procedure, the effect of applying elementary row operations to reduce A to the identity matrix I_n is equivalent to multiplying A on the left by a sequence of elementary matrices. In other words, we have $E_rE_{r-1}\dots E_1A = I_n$, for certain elementary $n \times n$ matrices $E_1, \dots, E_r$. Since we are assuming that A is invertible, we can multiply both sides with A^{-1} to obtain $E_rE_{r-1}\dots E_1 = A^{-1}$. But when we apply the same elementary row operations to I_n, then we end up with $E_rE_{r-1}\dots E_1I_n = A^{-1}$. This explains why the method works.

Notice also that the inverse of an elementary row matrix is another one of the same type. In fact it is easily checked that the inverses of $E(n)^1_{\lambda,i,j}$, $E(n)^2_{i,j}$ and $E(n)^3_{\lambda,i}$ are respectively $E(n)^1_{-\lambda,i,j}$, $E(n)^2_{i,j}$ and $E(n)^3_{\lambda^{-1},i}$. Hence, if $E_rE_{r-1}\dots E_1A = I_n$ as in the preceding paragraph, then by using Lemma 9.3 we find that

$$A = (E_rE_{r-1}\dots E_1)^{-1} = E_1^{-1}E_2^{-1}\dots E_r^{-1},$$

which is itself a product of elementary matrices. We have proved:

Theorem 9.5. *An invertible matrix is a product of elementary matrices.*

9.4 Application to linear equations

The most familiar examples of simultaneous equations are those where we have the same number n of equations as unknowns. However, even in that case, there is no guarantee that there is a unique solution; there can still be zero, one or many solutions (for instance, see examples in section 8.1). The case of a unique solution occurs exactly when the matrix A is non-singular.

Theorem 9.6. *Let A be an $n \times n$ matrix. Then*

(i) the homogeneous system of equations $A\underline{\mathbf{x}} = \mathbf{0}$ has a non-zero solution if and only if A is singular;

(ii) the equation system $A\underline{\mathbf{x}} = \underline{\beta}$ has a unique solution if and only if A is non-singular.

Proof. We first prove (i). The solution set of the equations is exactly nullspace(A). Let $T : K^{n,1} \to K^{n,1}$ be the linear map given by $\underline{\mathbf{v}} \mapsto A\underline{\mathbf{v}}$. It is easy to check that A is the matrix corresponding to T with respect to the standard basis of $K^{n,1}$.

By Corollary 5.6,

$$\text{nullspace}(A) = \ker(T) = \{\mathbf{0}\} \iff \text{nullity}(T) = 0 \iff T \text{ is non-singular},$$

and so there are non-zero solutions if and only if T and hence A is singular.

Now (ii). If A is singular then its nullity is greater than 0 and so its nullspace is not equal to $\{\mathbf{0}\}$, and contains more than one vector. Either there are no solutions, or the solution set is $\underline{\mathbf{x}} + \text{nullspace}(A)$ for some specific solution $\underline{\mathbf{x}}$, in which case there is more than one solution. Hence there cannot be a unique solution when A is singular.

Conversely, if A is non-singular, then it is invertible by Theorem 9.2, and one solution is $\underline{\mathbf{x}} = A^{-1}\underline{\beta}$. Since the complete solution set is then $\underline{\mathbf{x}} + \text{nullspace}(A)$, and nullspace$(A) = \{\mathbf{0}\}$ in this case, the solution is unique. □

In general, it is more efficient to solve the equations $A\underline{\mathbf{x}} = \underline{\beta}$ by elementary row operations rather than by first computing A^{-1} and then $A^{-1}\underline{\beta}$. However, if A^{-1} is already known for some reason, then this is a useful method.

Example. Consider the system of linear equations

$$\begin{aligned} 3x + 2y + z &= 0 && (1)\\ 4x + y + 3z &= 2 && (2)\\ 2x + y + 6z &= 6. && (3) \end{aligned}$$

Here $A = \begin{pmatrix} 3 & 2 & 1\\ 4 & 1 & 3\\ 2 & 1 & 6 \end{pmatrix}$, and we computed $A^{-1} = \begin{pmatrix} -3/25 & 11/25 & -1/5\\ 18/25 & -16/25 & 1/5\\ -2/25 & -1/25 & 1/5 \end{pmatrix}$ in Section 9. Computing $A^{-1}\underline{\beta}$ with $\underline{\beta} = \begin{pmatrix} 0\\ 2\\ 6 \end{pmatrix}$ yields the solution $x = -8/25$, $y = -2/25$, $z = 28/25$. If we had not already known A^{-1}, then it would have been quicker to solve the linear equations directly rather than computing A^{-1} first.

10 The determinant of a matrix

10.1 Definition of the determinant

Let A be an $n \times n$ matrix over the field K. The *determinant* of A, which is written as $\det(A)$ or sometimes as $|A|$, is a certain scalar that is defined from A in a rather complicated way. The definition for small values of n might already be familiar to you.

$n = 1$	$A = (\alpha)$	$\det(A) = \alpha$
$n = 2$	$A = \begin{pmatrix} \alpha_{11} & \alpha_{12}\\ \alpha_{21} & \alpha_{22} \end{pmatrix}$	$\det(A) = \alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}$

And, for $n = 3$, we have

$$A = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13}\\ \alpha_{21} & \alpha_{22} & \alpha_{23}\\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix}$$

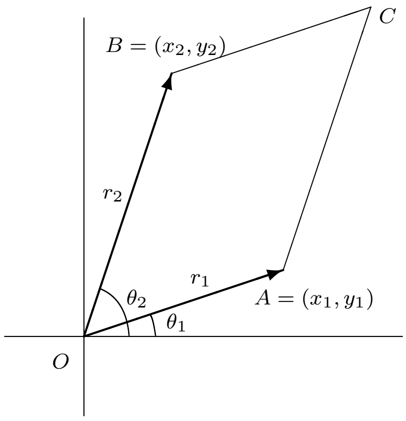
and

$$\det(A) = \alpha_{11}\begin{vmatrix} \alpha_{22} & \alpha_{23} \\ \alpha_{32} & \alpha_{33} \end{vmatrix} - \alpha_{12}\begin{vmatrix} \alpha_{21} & \alpha_{23} \\ \alpha_{31} & \alpha_{33} \end{vmatrix} + \alpha_{13}\begin{vmatrix} \alpha_{21} & \alpha_{22} \\ \alpha_{31} & \alpha_{32} \end{vmatrix}$$
$$= \alpha_{11}\alpha_{22}\alpha_{33} - \alpha_{11}\alpha_{23}\alpha_{32} - \alpha_{12}\alpha_{21}\alpha_{33} + \alpha_{12}\alpha_{23}\alpha_{31} + \alpha_{13}\alpha_{21}\alpha_{32} - \alpha_{13}\alpha_{22}\alpha_{31}.$$

Where do these formulae come from, and why are they useful?

The geometrical motivation for the determinant is that it represents area or volume. For $n = 2$, consider the position vectors of two points (x_1, y_1), (x_2, y_2) in the plane. Then, in the diagram below, the area of the parallelogram $OABC$ enclosed by these two vectors is

$$r_1r_2\sin(\theta_2 - \theta_1) = r_1r_2(\sin\theta_2\cos\theta_1 - \sin\theta_1\cos\theta_2) = x_1y_2 - x_2y_1 = \begin{vmatrix} x_1 & x_2 \\ y_1 & y_2 \end{vmatrix}$$



Similarly, when $n = 3$ the volume of the parallelepiped enclosed by the three position vectors in space is equal to (plus or minus) the determinant of the 3×3 matrix defined by the coordinates of the three points.

Now we turn to the general definition for $n\times n$ matrices. Suppose that we take the product of n entries from the matrix, where we take exactly one entry from each row and one from each column. Such a product is called an *elementary product*. There are $n!$ such products altogether (we shall see why shortly) and the determinant is the sum of $n!$ terms, each of which is plus or minus one of these elementary products. We say that it is a sum of $n!$ *signed elementary products*. You should check that this holds in the 2 and 3-dimensional cases written out above.

Before we can be more precise about this, and determine which signs we choose for which elementary products, we need to make a short digression to study permutations of finite sets. A permutation of a set, which we shall take here to be the set $X_n = \{1, 2, 3, \dots, n\}$, is simply a bijection from X_n to itself. The set of all such permutations of X_n is called the *symmetric group* S_n. There are $n!$ permutations altogether, so $|S_n| = n!$.

Now an elementary product contains one entry from each row of A, so let the entry in the product from the ith row be $\alpha_{i\phi(i)}$, where ϕ is some as-yet unknown function from X_n to X_n. Since the product also contains exactly one entry from each column, each integer $j \in X_n$ must occur exactly once as $\phi(i)$. But this is just saying that $\phi\colon X_n \to X_n$ is a bijection; that is $\phi \in S_n$. Conversely, any $\phi \in S_n$ defines an elementary product in this way.

So an elementary product has the general form $\alpha_{1\phi(1)}\alpha_{2\phi(2)}\dots\alpha_{n\phi(n)}$ for some $\phi \in S_n$, and there are $n!$ elementary products altogether. We want to define

$$\det(A) = \sum_{\phi\in S_n} \pm\alpha_{1\phi(1)}\alpha_{2\phi(2)}\dots\alpha_{n\phi(n)},$$

but we still have to decide which of the elementary products has a plus sign and which has a minus sign. In fact this depends on the *sign* of the permutation ϕ, which we must now define.

A *transposition* is a permutation of X_n that interchanges two numbers i and j in X_n and leaves all other numbers fixed. It is written as (i, j). There is a theorem, which is quite easy, but we will not prove it here because it is a theorem in Group Theory, that says that every permutation can be written as a composite of transpositions. For example, if $n = 5$, then the permutation ϕ defined by

$$\phi(1) = 4,\ \phi(2) = 5,\ \phi(3) = 3,\ \phi(4) = 2,\ \phi(5) = 1$$

is equal to the composite $(1, 4) \circ (2, 4) \circ (2, 5)$. (Remember that permutations are functions $X_n \to X_n$, so this means first apply the function $(2, 5)$ (which interchanges 2 and 5) then apply $(2, 4)$ and finally apply $(1, 4)$.)

Definition. Now a permutation ϕ is said to be *even*, and to have sign $+1$, if ϕ is a composite of an even number of transpositions; and ϕ is said to be *odd*, and to have sign -1, if ϕ is a composite of an odd number of transpositions.

For example, the permutation ϕ defined on X_n above is a composite of 3 transpositions, so ϕ is odd and $\text{sign}(\phi) = -1$. The identity permutation, which leaves all points fixed, is even (because it is a composite of 0 transpositions).

Now at last we can give the general definition of the determinant.

Definition. The *determinant* of a $n \times n$ matrix $A = (\alpha_{ij})$ is the scalar quantity

$$\det(A) = \sum_{\phi \in S_n} \text{sign}(\phi)\alpha_{1\phi(1)}\alpha_{2\phi(2)} \dots \alpha_{n\phi(n)}.$$

(**Note:** You might be worrying about whether the same permutation could be both even and odd. Well, there is a moderately difficult theorem in Group Theory, which we shall not prove here, that says that this cannot happen; in other words, the concepts of even and odd permutation are *well-defined*.)

10.2 The effect of matrix operations on the determinant

Theorem 10.1. *Elementary row operations affect the determinant of a matrix as follows.*

(i) $\det(I_n) = 1$.

(ii) Let B result from A by applying `(R2)` *(interchanging two rows). Then* $\det(B) = -\det(A)$.

(iii) If A has two equal rows then $\det(A) = 0$.

(iv) Let B result from A by applying `(R1)` *(adding a multiple of one row to another). Then* $\det(B) = \det(A)$.

(v) Let B result from A by applying `(R3)` *(multiplying a row by a scalar λ). Then* $\det(B) = \lambda \det(A)$.

Proof. (i) If $A = I_n$ then $\alpha_{ij} = 0$ when $i \neq j$. So the only non-zero elementary product in the sum occurs when ϕ is the identity permutation. Hence $\det(A) = \alpha_{11}\alpha_{22} \dots \alpha_{nn} = 1$.

(ii) To keep the notation simple, we shall suppose that we interchange the first two rows, but the same argument works for interchanging any pair of rows. Then if $B = (\beta_{ij})$, we have $\beta_{1j} = \alpha_{2j}$ and $\beta_{2j} = \alpha_{1j}$ for all j. Hence

$$\det(B) = \sum_{\phi \in S_n} \operatorname{sign}(\phi)\beta_{1\phi(1)}\beta_{2\phi(2)} \dots \beta_{n\phi(n)}$$
$$= \sum_{\phi \in S_n} \operatorname{sign}(\phi)\alpha_{1\phi(2)}\alpha_{2\phi(1)}\alpha_{3\phi(3)} \dots \alpha_{n\phi(n)}.$$

For $\phi \in S_n$, let $\psi = \phi \circ (1, 2)$, so $\phi(1) = \psi(2)$ and $\phi(2) = \psi(1)$, and $\operatorname{sign}(\psi) = -\operatorname{sign}(\phi)$. Now, as ϕ runs through all permutations in S_n, so does ψ (but in a different order), so summing over all $\phi \in S_n$ is the same as summing over all $\psi \in S_n$. Hence

$$\det(B) = \sum_{\phi \in S_n} -\operatorname{sign}(\psi)\alpha_{1\psi(1)}\alpha_{2\psi(2)} \dots \alpha_{n\psi(n)}$$
$$= \sum_{\psi \in S_n} -\operatorname{sign}(\psi)\alpha_{1\psi(1)}\alpha_{2\psi(2)} \dots \alpha_{n\psi(n)} = -\det(A).$$

(iii) Again to keep notation simple, assume that the equal rows are the first two. Using the same notation as in (ii), namely $\psi = \phi \circ (1, 2)$, the two elementary products:

$$\alpha_{1\psi(1)}\alpha_{2\psi(2)} \dots \alpha_{n\psi(n)} \quad \text{and} \quad \alpha_{1\phi(1)}\alpha_{2\phi(2)} \dots \alpha_{n\phi(n)}$$

are equal. This is because $\alpha_{1\psi(1)} = \alpha_{2\psi(1)}$ (first two rows equal) and $\alpha_{2\psi(1)} = \alpha_{2\phi(2)}$ (because $\phi(2) = \psi(1)$); hence $\alpha_{1\psi(1)} = \alpha_{2\phi(2)}$. Similarly $\alpha_{2\psi(2)} = \alpha_{1\phi(1)}$ and the two products differ by interchanging their first two terms. But $\operatorname{sign}(\psi) = -\operatorname{sign}(\phi)$ so the two corrresponding signed products cancel each other out. Thus each signed product in $\det(A)$ cancels with another and the sum is zero.

(iv) Again, to simplify notation, suppose that we replace the second row $\mathbf{r}_2$ by $\mathbf{r}_2 + \lambda\mathbf{r}_1$ for some $\lambda \in K$. Then

$$\begin{aligned}\det(B) &= \sum_{\phi \in S_n} \operatorname{sign}(\phi)\alpha_{1\phi(1)}(\alpha_{2\phi(2)} + \lambda\alpha_{1\phi(2)})\alpha_{3\phi(3)} \dots \alpha_{n\phi(n)} \\ &= \sum_{\phi \in S_n} \operatorname{sign}(\phi)\alpha_{1\phi(1)}\alpha_{2\phi(2)} \dots \alpha_{n\phi(n)} \\ &\quad + \lambda \sum_{\phi \in S_n} \operatorname{sign}(\phi)\alpha_{1\phi(1)}\alpha_{1\phi(2)} \dots \alpha_{n\phi(n)}.\end{aligned}$$

Now the first term in this sum is $\det(A)$, and the second is $\lambda \det(C)$, where C is a matrix in which the first two rows are equal. Hence $\det(C) = 0$ by (iii), and $\det(B) = \det(A)$.

(v) Easy. Note that this holds even when the scalar $\lambda = 0$.

$\square$

Definition. A matrix is called *upper triangular* if all of its entries below the main diagonal are zero; that is, (α_{ij}) is upper triangular if $\alpha_{ij} = 0$ for all $i > j$.

The matrix is called *diagonal* if all entries not on the main diagonal are zero; that is, $\alpha_{ij} = 0$ for $i \neq j$.

Example. $\begin{pmatrix} 3 & 0 & -1/2 \\ 0 & -1 & -11 \\ 0 & 0 & -2/5 \end{pmatrix}$ is upper triangular, and $\begin{pmatrix} 0 & 0 & 0 \\ 0 & 17 & 0 \\ 0 & 0 & -3 \end{pmatrix}$ is diagonal.

Corollary 10.2. *If $A = (\alpha_{ij})$ is upper triangular, then $\det(A) = \alpha_{11}\alpha_{22}\dots\alpha_{nn}$ is the product of the entries on the main diagonal of A.*

Proof. This is not hard to prove directly from the definition of the determinant. Alternatively, we can apply row operations `(R1)` to reduce the matrix to row reduced form. If any of the diagonal elements of A were 0, then the row reduced form has a zero row and hence has zero determinant. Otherwise, the row reduced form is a diagonal matrix with the identity, and the the result follows from parts (i) and (v) of the theorem. □

The above theorem and corollary provide the most efficient way of computing $\det(A)$, at least for $n \geq 3$. (For $n = 2$, it is easiest to do it straight from the definition.) Use row operations `(R1)` and `(R2)` to reduce A to upper triangular form, keeping track of changes of sign in the determinant resulting from applications of `(R2)`, and then use Corollary 10.2.

Example.

$$\begin{vmatrix} 0 & 1 & 1 & 2 \\ 1 & 2 & 1 & 1 \\ 2 & 1 & 3 & 1 \\ 1 & 2 & 4 & 2 \end{vmatrix} \overset{\mathbf{r}_2 \leftrightarrow \mathbf{r}_1}{=} - \begin{vmatrix} 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 2 & 1 & 3 & 1 \\ 1 & 2 & 4 & 2 \end{vmatrix} \overset{\mathbf{r}_3 \to \mathbf{r}_3 - 2\mathbf{r}_1}{=} - \begin{vmatrix} 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & -3 & 1 & -1 \\ 1 & 2 & 4 & 2 \end{vmatrix}$$

$$\overset{\mathbf{r}_4 \to \mathbf{r}_4 - \mathbf{r}_1}{=} - \begin{vmatrix} 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & -3 & 1 & -1 \\ 0 & 0 & 3 & 1 \end{vmatrix} \overset{\mathbf{r}_3 \to \mathbf{r}_3 + 3\mathbf{r}_2}{=} - \begin{vmatrix} 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 4 & 5 \\ 0 & 0 & 3 & 1 \end{vmatrix} \overset{\mathbf{r}_4 \to \mathbf{r}_4 - 3\mathbf{r}_3/4}{=} - \begin{vmatrix} 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 4 & 5 \\ 0 & 0 & 0 & -\frac{11}{4} \end{vmatrix}$$

$$= 11$$

We could have been a little more clever, and stopped the row reduction one step before the end, noticing that the determinant was equal to $\left|\begin{smallmatrix} 4 & 5 \\ 3 & 1 \end{smallmatrix}\right| = 11$.

Definition. Let $A = (\alpha_{ij})$ be an $m \times n$ matrix. We define the *transpose* A^{T} of A to be the $n \times m$ matrix (β_{ij}), where $\beta_{ij} = \alpha_{ji}$ for $1 \leq i \leq n$, $1 \leq j \leq m$.

For example, $\begin{pmatrix} 1 & 3 & 5 \\ -2 & 0 & 6 \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} 1 & -2 \\ 3 & 0 \\ 5 & 6 \end{pmatrix}$.

Theorem 10.3. *Let $A = (\alpha_{ij})$ be an $n \times n$ matrix. Then $\det(A^{\mathrm{T}}) = \det(A)$.*

Proof. Let $A^{\mathrm{T}} = (\beta_{ij})$ where $\beta_{ij} = \alpha_{ji}$. Then

$$\begin{aligned} \det(A^{\mathrm{T}}) &= \sum_{\phi \in S_n} \mathrm{sign}(\phi)\beta_{1\phi(1)}\beta_{2\phi(2)}\dots\beta_{n\phi(n)} \\ &= \sum_{\phi \in S_n} \mathrm{sign}(\phi)\alpha_{\phi(1)1}\alpha_{\phi(2)2}\dots\alpha_{\phi(n)n}. \end{aligned}$$

Now, by rearranging the terms in the elementary product, we have

$$\alpha_{\phi(1)1}\alpha_{\phi(2)2}\dots\alpha_{\phi(n)n} = \alpha_{1\phi^{-1}(1)}\alpha_{2\phi^{-1}(2)}\dots\alpha_{n\phi^{-1}(n)},$$

where ϕ^{-1} is the *inverse* permutation to ϕ. Notice also that if ϕ is a composite $\tau_1 \circ \tau_2 \circ \dots \circ \tau_r$ of transpositions τ_i, then $\phi^{-1} = \tau_r \circ \dots \circ \tau_2 \circ \tau_1$ (because each $\tau_i \circ \tau_i$ is the identity permutation). Hence $\mathrm{sign}(\phi) = \mathrm{sign}(\phi^{-1})$. Also, summing over all $\phi \in S_n$ is the same as summing over all $\phi^{-1} \in S_n$, so we have

$$\begin{aligned} \det(A^{\mathrm{T}}) &= \sum_{\phi \in S_n} \mathrm{sign}(\phi)\alpha_{1\phi^{-1}(1)}\alpha_{2\phi^{-1}(2)}\dots\alpha_{n\phi^{-1}(n)} \\ &= \sum_{\phi^{-1} \in S_n} \mathrm{sign}(\phi^{-1})\alpha_{1\phi^{-1}(1)}\alpha_{2\phi^{-1}(2)}\dots\alpha_{n\phi^{-1}(n)} = \det(A). \end{aligned}$$

□

If you find proofs like the above, where we manipulate sums of products, hard to follow, then it might be helpful to write it out in full in a small case, such as $n = 3$. Then

$$\begin{aligned}\det(A^{\mathrm{T}}) &= \beta_{11}\beta_{22}\beta_{33} - \beta_{11}\beta_{23}\beta_{32} - \beta_{12}\beta_{21}\beta_{33}\\ &\qquad + \beta_{12}\beta_{23}\beta_{31} + \beta_{13}\beta_{21}\beta_{32} - \beta_{13}\beta_{22}\beta_{31}\\ &= \alpha_{11}\alpha_{22}\alpha_{33} - \alpha_{11}\alpha_{32}\alpha_{23} - \alpha_{21}\alpha_{12}\alpha_{33}\\ &\qquad + \alpha_{21}\alpha_{32}\alpha_{13} + \alpha_{31}\alpha_{12}\alpha_{23} - \alpha_{31}\alpha_{22}\alpha_{13}\\ &= \alpha_{11}\alpha_{22}\alpha_{33} - \alpha_{11}\alpha_{23}\alpha_{32} - \alpha_{12}\alpha_{21}\alpha_{33}\\ &\qquad + \alpha_{12}\alpha_{23}\alpha_{31} + \alpha_{13}\alpha_{21}\alpha_{32} - \alpha_{13}\alpha_{22}\alpha_{31}\\ &= \det(A).\end{aligned}$$

Corollary 10.4. *All of Theorem 10.1 remains true if we replace rows by columns.*

Proof. This follows from Theorems 10.1 and 10.3, because we can apply column operations to A by transposing it, applying the corresponding row operations, and then re-transposing it. □

We are now ready to prove one of the most important properties of the determinant.

Theorem 10.5. *For an $n \times n$ matrix A, $\det(A) = 0$ if and only if A is singular.*

Proof. A can be reduced to row reduced echelon form by using row operations `(R1)`, `(R2)` and `(R3)`. By Theorem 8.4, none of these operations affect the rank of A, and so they do not affect whether or not A is singular (remember 'singular' means $\mathrm{rank}(A) < n$; see definition after Corollary 5.6). By Theorem 10.1, they do not affect whether or not $\det(A) = 0$. So we can assume that A is in row reduced form.

Then $\mathrm{rank}(A)$ is the number of non-zero rows of A, so if A is singular then it has some zero rows. But then $\det(A) = 0$. On the other hand, if A is nonsingular then, as we saw in Section 9.2, the fact that A is in row reduced form implies that $A = I_n$, so $\det(A) = 1 \neq 0$. □

10.3 The determinant of a product

Example. Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$ and $B = \begin{pmatrix} -1 & -1 \\ 2 & 0 \end{pmatrix}$. Then $\det(A) = -4$ and $\det(B) = 2$. We have $A+B = \begin{pmatrix} 0 & 1 \\ 5 & 2 \end{pmatrix}$ and $\det(A+B) = -5 \neq \det(A)+\det(B)$. In fact, in general there is no simple relationship between $\det(A+B)$ and $\det(A)$, $\det(B)$.

However, $AB = \begin{pmatrix} 3 & -1 \\ 1 & -3 \end{pmatrix}$, and $\det(AB) = -8 = \det(A)\det(B)$.

In this subsection, we shall prove that this simple relationship holds in general.

Recall from Section 9.3 the definition of an *elementary* matrix E, and the property that if we multiply a matrix B on the left by E, then the effect is to apply the corresponding elementary row operation to B. This enables us to prove:

Lemma 10.6. *If E is an $n \times n$ elementary matrix, and B is any $n \times n$ matrix, then $\det(EB) = \det(E)\det(B)$.*

Proof. E is one of the three types $E(n)^1_{\lambda,ij}$, $E(n)^2_{ij}$ or $E(n)^3_{\lambda,i}$, and multiplying B on the left by E has the effect of applying `(R1)`, `(R2)` or `(R3)` to B, respectively. Hence, by Theorem 10.1, $\det(EB) = \det(B), -\det(B)$, or $\lambda\det(B)$, respectively. But by considering the special case $B = I_n$, we see that $\det(E) = 1, -1$ or λ, respectively, and so $\det(EB) = \det(E)\det(B)$ in all three cases. □

Theorem 10.7. *For any two $n \times n$ matrices A and B, we have*

$$\det(AB) = \det(A)\det(B).$$

Proof. We first dispose of the case when $\det(A) = 0$. Then we have $\operatorname{rank}(A) < n$ by Theorem 10.5. Let $T_1, T_2 \colon V \to V$ be linear maps corresponding to A and B, where $\dim(V) = n$. Then AB corresponds to T_1T_2 (by Theorem 7.4). By Corollary 5.6, $\operatorname{rank}(A) = \operatorname{rank}(T_1) < n$ implies that T_1 is not surjective. But then T_1T_2 cannot be surjective, so $\operatorname{rank}(T_1T_2) = \operatorname{rank}(AB) < n$. Hence $\det(AB) = 0$ so $\det(AB) = \det(A)\det(B)$.

On the other hand, if $\det(A) \neq 0$, then A is nonsingular, and hence invertible, so by Theorem 9.5 A is a product $E_1E_2 \dots E_r$ of elementary matrices E_i. Hence $\det(AB) = \det(E_1E_2 \dots E_rB)$. Now the result follows from the above lemma, because

$$\begin{aligned}
\det(AB) &= \det(E_1)\det(E_2 \cdots E_rB) \\
&= \det(E_1)\det(E_2)\det(E_3 \cdots E_rB) \\
&= \det(E_1)\det(E_2)\cdots\det(E_r)\det(B) \\
&= \det(E_1E_2 \cdots E_r)\det(B) \\
&= \det(A)\det(B).
\end{aligned}$$

□

10.4 Minors and cofactors

Definition. Let $A = (\alpha_{ij})$ be an $n \times n$ matrix. Let A_{ij} be the $(n-1) \times (n-1)$ matrix obtained from A by deleting the ith row and the jth column of A. Now let $M_{ij} = \det(A_{ij})$. Then M_{ij} is called the (i,j)th *minor* of A.

Example. If $A = \begin{pmatrix} 2 & 1 & 0 \\ 3 & -1 & 2 \\ 5 & -2 & 0 \end{pmatrix}$, then $A_{12} = \begin{pmatrix} 3 & 2 \\ 5 & 0 \end{pmatrix}$ and $A_{31} = \begin{pmatrix} 1 & 0 \\ -1 & 2 \end{pmatrix}$, and so $M_{12} = -10$ and $M_{31} = 2$.

Definition. We define c_{ij} to be equal to M_{ij} if $i+j$ is even, and to $-M_{ij}$ if $i+j$ is odd. Or, more concisely,

$$c_{ij} = (-1)^{i+j}M_{ij} = (-1)^{i+j}\det(A_{ij}).$$

Then c_{ij} is called the (i,j)th *cofactor* of A.

Example. In the example above,

$$\begin{aligned}
c_{11} &= \begin{vmatrix} -1 & 2 \\ -2 & 0 \end{vmatrix} = 4, & c_{12} &= -\begin{vmatrix} 3 & 2 \\ 5 & 0 \end{vmatrix} = 10, & c_{13} &= \begin{vmatrix} 3 & -1 \\ 5 & -2 \end{vmatrix} = -1, \\
c_{21} &= -\begin{vmatrix} 1 & 0 \\ -2 & 0 \end{vmatrix} = 0, & c_{22} &= \begin{vmatrix} 2 & 0 \\ 5 & 0 \end{vmatrix} = 0, & c_{23} &= -\begin{vmatrix} 2 & 1 \\ 5 & -2 \end{vmatrix} = 9, \\
c_{31} &= \begin{vmatrix} 1 & 0 \\ -1 & 2 \end{vmatrix} = 2, & c_{32} &= -\begin{vmatrix} 2 & 0 \\ 3 & 2 \end{vmatrix} = -4, & c_{33} &= \begin{vmatrix} 2 & 1 \\ 3 & -1 \end{vmatrix} = -5.
\end{aligned}$$

The cofactors give us a useful way of expressing the determinant of a matrix in terms of determinants of smaller matrices.

Theorem 10.8. *Let A be an $n \times n$ matrix.*

(i) (Expansion of a determinant by the ith row.) For any i with $1 \leq i \leq n$, we have

$$\det(A) = \alpha_{i1}c_{i1} + \alpha_{i2}c_{i2} + \cdots + \alpha_{in}c_{in} = \sum_{j=1}^{n} \alpha_{ij}c_{ij}.$$

(ii) (Expansion of a determinant by the jth column.) For any j with $1 \leq j \leq n$, we have

$$\det(A) = \alpha_{1j}c_{1j} + \alpha_{2j}c_{2j} + \cdots \alpha_{nj}c_{nj} = \sum_{i=1}^{n} \alpha_{ij}c_{ij}.$$

For example, expanding the determinant of the matrix A above by the first row, the third row, and the second column give respectively:

$$\begin{array}{rcccl} \det(A) & = & 2 \times 4 + 1 \times 10 + 0 \times -1 & = & 18 \\ \det(A) & = & 5 \times 2 + -2 \times -4 + 0 \times -5 & = & 18 \\ \det(A) & = & 1 \times 10 + -1 \times 0 + -2 \times -4 & = & 18. \end{array}$$

Proof of Theorem 10.8. By definition, we have

$$\det(A) = \sum_{\phi \in S_n} \operatorname{sign}(\phi)\alpha_{1\phi(1)}\alpha_{2\phi(2)} \ldots \alpha_{n\phi(n)} \qquad (*)$$

Step 1. We first find the sum of all of those signed elementary products in the sum $(*)$ that contain α_{nn}. These arise from those permutations ϕ with $\phi(n) = n$; so the required sum is

$$\sum_{\substack{\phi \in S_n \\ \phi(n)=n}} \operatorname{sign}(\phi)\alpha_{1\phi(1)}\alpha_{2\phi(2)} \ldots \alpha_{n\phi(n)}$$

$$= \alpha_{nn} \sum_{\phi \in S_{n-1}} \operatorname{sign}(\phi)\alpha_{1\phi(1)}\alpha_{2\phi(2)} \ldots \alpha_{n-1\phi(n-1)}$$

$$= \alpha_{nn}M_{nn} = \alpha_{nn}c_{nn}.$$

Step 2. Next we fix any i and j with $1 \leq i, j \leq n$, and find the sum of all of those signed elementary products in the sum $(*)$ that contain α_{ij}. We move row $\mathbf{r}_i$ of A to $\mathbf{r}_n$ by interchanging $\mathbf{r}_i$ with $\mathbf{r}_{i+1}, \mathbf{r}_{i+2}, \ldots, \mathbf{r}_n$ in turn. This involves $n - i$ applications of `(R2)`, and leaves the rows of A other than $\mathbf{r}_i$ in their original order. We then move column $\mathbf{c}_j$ to $\mathbf{c}_n$ in the same way, by applying `(C2)` $n - j$ times. Let the resulting matrix be $B = (\beta_{ij})$ and denote its minors by N_{ij}. Then $\beta_{nn} = \alpha_{ij}$, and $N_{nn} = M_{ij}$. Furthermore,

$$\det(B) = (-1)^{2n-i-j} \det(A) = (-1)^{i+j} \det(A),$$

because $(2n - i - j) - (i + j)$ is even.

Now, by the result of Step 1, the sum of terms in $\det(B)$ involving β_{nn} is

$$\beta_{nn}N_{nn} = \alpha_{ij}M_{ij} = (-1)^{i+j}\alpha_{ij}c_{ij},$$

and hence, since $\det(B) = (-1)^{i+j} \det(A)$, the sum of terms involving α_{ij} in $\det(A)$ is $\alpha_{ij}c_{ij}$.

Step 3. The result follows from Step 2, because every signed elementary product in the sum $(*)$ involves exactly one array element α_{ij} from each row and from each column. Hence, for any given row or column, we get the full sum $(*)$ by adding up the total of those products involving each individual element in that row or column. $\square$

Example. Expanding by a row and column can sometimes be a quick method of evaluating the determinant of matrices containing a lot of zeros. For example, let

$$A = \begin{pmatrix} 9 & 0 & 2 & 6 \\ 1 & 2 & 9 & -3 \\ 0 & 0 & -2 & 0 \\ -1 & 0 & -5 & 2 \end{pmatrix}.$$

Then, expanding by the third row, we get $\det(A) = -2 \begin{vmatrix} 9 & 0 & 6 \\ 1 & 2 & -3 \\ -1 & 0 & 2 \end{vmatrix}$, and then expanding by the second column, $\det(A) = -2 \times 2 \begin{vmatrix} 9 & 6 \\ -1 & 2 \end{vmatrix} = -96$.

10.5 The inverse of a matrix using determinants

Definition. Let A be an $n \times n$ matrix. We define the *adjugate* matrix $\mathrm{adj}(A)$ of A to be the $n \times n$ matrix of which the (i, j)th element is the cofactor c_{ji}. In other words, it is the transpose of the matrix of cofactors.

The adjugate is also sometimes called the *adjoint*. However, the word "adjoint" is also used with other meanings, so to avoid confusion we will use the word "adjugate".

Example. In the example above,

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 3 & -1 & 2 \\ 5 & -2 & 0 \end{pmatrix}, \qquad \mathrm{adj}(A) = \begin{pmatrix} 4 & 0 & 2 \\ 10 & 0 & -4 \\ -1 & 9 & -5 \end{pmatrix}.$$

The adjugate is almost an inverse to A, as the following theorem shows.

Theorem 10.9. $A\,\mathrm{adj}(A) = \det(A)I_n = \mathrm{adj}(A)\,A$

Proof. Let $B = A\,\mathrm{adj}(A) = (\beta_{ij})$. Then $\beta_{ii} = \sum_{k=1}^{n} \alpha_{ik}c_{ik} = \det(A)$ by Theorem 10.8 (expansion by the ith row of A). For $i \neq j$, $\beta_{ij} = \sum_{k=1}^{n} \alpha_{ik}c_{jk}$, which is the determinant of a matrix C obtained from A by substituting the ith row of A for the jth row. But then C has two equal rows, so $\beta_{ij} = \det(C) = 0$ by Theorem 10.1(iii). (This is sometimes called an expansion by *alien cofactors*.) Hence $A\,\mathrm{adj}(A) = \det(A)I_n$. A similar argument using columns instead of rows gives $\mathrm{adj}(A)\,A = \det(A)I_n$. □

Example. In the example above, check that $A\,\mathrm{adj}(A) = \mathrm{adj}(A)\,A = 18I_3$.

Corollary 10.10. *If* $\det(A) \neq 0$*, then* $A^{-1} = \frac{1}{\det(A)}\,\mathrm{adj}(A)$.

(Theorems 9.2 and 10.5 imply that A is invertible if and only if $\det(A) \neq 0$.)

Example. In the example above,

$$\begin{pmatrix} 2 & 1 & 0 \\ 3 & -1 & 2 \\ 5 & -2 & 0 \end{pmatrix}^{-1} = \frac{1}{18}\begin{pmatrix} 4 & 0 & 2 \\ 10 & 0 & -4 \\ -1 & 9 & -5 \end{pmatrix},$$

and in the example in Section 9,

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 4 & 1 & 3 \\ 2 & 1 & 6 \end{pmatrix}, \quad \mathrm{adj}(A) = \begin{pmatrix} 3 & -11 & 5 \\ -18 & 16 & -5 \\ 2 & 1 & -5 \end{pmatrix}, \quad \det(A) = -25,$$

and so $A^{-1} = \frac{-1}{25}\,\mathrm{adj}(A)$.

For 2×2 and (possibly) 3×3 matrices, the cofactor method of computing the inverse is often the quickest. For larger matrices, the row reduction method described in Section 9 is quicker.

10.6 Cramer's rule for solving simultaneous equations

Given a system $A\underline{\mathbf{x}} = \underline{\beta}$ of n equations in n unknowns, where $A = (\alpha_{ij})$ is non-singular, the solution is $\underline{\mathbf{x}} = A^{-1}\underline{\beta}$. So the ith component x_i of this column vector is the ith row of $A^{-1}\underline{\beta}$. Now, by Corollary 10.10, $A^{-1} = \frac{1}{\det(A)}\,\mathrm{adj}(A)$, and its (i, j)th entry is $c_{ji}/\det(A)$. Hence

$$x_i = \frac{1}{\det(A)} \sum_{j=1}^{n} c_{ji}\beta_j.$$

Now let A_i be the matrix obtained from A by substituting $\underline{\beta}$ for the ith column of A. Then the sum $\sum_{j=1}^{n} c_{ji}\beta_j$ is precisely the expansion of $\det(A_i)$ by its ith column (see Theorem 10.8). Hence we have $x_i = \det(A_i)/\det(A)$. This is Cramer's rule.

This is more of a curiosity than a practical method of solving simultaneous equations, although it can be quite quick in the 2×2 case. Even in the 3×3 case it is rather slow.

Example. Let us solve the following system of linear equations:

$$\begin{array}{rcrcrcr} 2x & & & + & z & = & 1 \\ & & y & - & 2z & = & 0 \\ x & + & y & + & z & = & -1 \end{array}$$

Cramer's rule gives

$$\det(A) = \begin{vmatrix} 2 & 0 & 1 \\ 0 & 1 & -2 \\ 1 & 1 & 1 \end{vmatrix} = 5, \qquad \det(A_1) = \begin{vmatrix} 1 & 0 & 1 \\ 0 & 1 & -2 \\ -1 & 1 & 1 \end{vmatrix} = 4$$

$$\det(A_2) = \begin{vmatrix} 2 & 1 & 1 \\ 0 & 0 & -2 \\ 1 & -1 & 1 \end{vmatrix} = -6, \qquad \det(A_3) = \begin{vmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & -1 \end{vmatrix} = -3$$

so the solution is $x = 4/5$, $y = -6/5$, $z = -3/5$.

11 Change of basis and equivalent matrices

We have been thinking of matrices as representing linear maps between vector spaces. But don't forget that, when we defined the matrix corresponding to a linear map between vector spaces U and V, the matrix depended on a particular choice of bases for both U and V. In this section, we investigate the relationship between the matrices corresponding to the same linear map $T : U \to V$, but using different bases for the vector spaces U and V. We first discuss the relationship between two different bases of the same space. Assume throughout the section that all vector spaces are over the same field K.

Let U be a vector space of dimension n, and let $E = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ and $E' = \{\mathbf{e}'_1, \ldots, \mathbf{e}'_n\}$ be two bases of U. The matrix $P = [E', I_U, E]$ of the identity map $I_U : U \to U$ using the basis E in the domain and E' in the range is called the *change of basis matrix* from the basis E to the basis E'.

Let us look carefully what this definition says. Taking $P = (\sigma_{ij})$, we obtain from Section 7.1

$$I_U(\mathbf{e}_j) = \mathbf{e}_j = \sum_{i=1}^{n} \sigma_{ij}\mathbf{e}'_i \text{ for } 1 \le j \le n. \qquad (*)$$

In other words, the columns of P are the coordinates of the "old" basis vectors $\mathbf{e}_i$ with respect to the "new" basis $\mathbf{e}'_i$.

The name "change of basis matrix" is justified by the following proposition.

Proposition 11.1. *With the above notation, let $\mathbf{v} \in U$, and let $\underline{\mathbf{v}}$ and $\underline{\mathbf{v}'}$ denote the column vectors of coordinates of $\mathbf{v}$ with respect to the bases E and E', respectively. Then $P\underline{\mathbf{v}} = \underline{\mathbf{v}'}$.*

Proof. This follows immediately from Proposition 7.2 applied to the identity map I_U. ☐

This gives a useful way to think about the change of basis matrix: it is the matrix which turns a vector's coordinates with respect to the "old" basis into the same vector's coordinates with respect to the "new" basis.

Proposition 11.2. *The change of basis matrix is invertible. More precisely, if P is the change of basis matrix from the basis E to the basis E' and Q is the change of basis matrix from the basis E' to the basis E, then $P = Q^{-1}$.*

Proof. Consider the composition of linear maps $I_U : U \xrightarrow{I_U} U \xrightarrow{I_U} U$ using the basis E' for the first and the third copy of U and the basis E for the middle copy of U. The composition has matrix I_n because the same basis is used for both domain and range. But the first I_U has matrix Q (change of basis from E' to E) and the second I_U similarly has matrix P. Therefore by Theorem 7.4, $I_n = PQ$.

Similarly, $I_n = QP$. Consequently, $P = Q^{-1}$. □

Example. Let $U = \mathbb{R}^3$, $\mathbf{e}'_1 = (1,0,0)$, $\mathbf{e}'_2 = (0,1,0)$, $\mathbf{e}'_3 = (0,0,1)$ (the standard basis) and $\mathbf{e}_1 = (0,2,1)$, $\mathbf{e}_2 = (1,1,0)$, $\mathbf{e}_3 = (1,0,0)$. Then

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

The columns of P are the coordinates of the "old" basis vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ with respect to the "new" basis $\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3$.

The converse of Proposition 11.2 is also true:

Proposition 11.3. *Every invertible matrix is a change of basis matrix. More precisely, if U is a vector space of dimension n over K, and if P is an invertible matrix in $K^{n,n}$, then there are bases E and E' of U such that P is the change of basis matrix from E to E'.*

Proof. To prove that $P = (\sigma_{ij})$ is a change of basis matrix, let $F' = \{f'_1, \dots f'_n\}$ be an arbitrary basis of U. Define $f_i := \sigma_{1i} f'_1 + \sigma_{2i} f'_2 + \ldots + \sigma_{ni} f'_n$ for every $i \in \{1, \dots, n\}$. Note that if the set of vectors $F = \{f_1, \dots f_n\}$ is a basis of U, then P is by definition the change of basis matrix from F to F' as $(*)$ is satisfied (this is why we defined the f_i the way we did). It only remains to check that F is indeed a basis. To see this, note that as P is invertible, its rank is n by Theorem 9.2 and Corollary 5.6. This means that its columns are linearly independent (recall the definition of column rank). But this is equivalent to saying that the vectors F are linearly independent and so they form a basis of U by Corollary 3.9. □

Now we will turn to the effect of change of basis on linear maps. let $T : U \to V$ be a linear map, where $\dim(U) = n$, $\dim(V) = m$. Choose a basis $E = \{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ of U and a basis $F = \{\mathbf{f}_1, \dots, \mathbf{f}_m\}$ of V. Then, from Section 7.1, we have

$$T(\mathbf{e}_j) = \sum_{i=1}^{m} \alpha_{ij} \mathbf{f}_i \text{ for } 1 \le j \le n$$

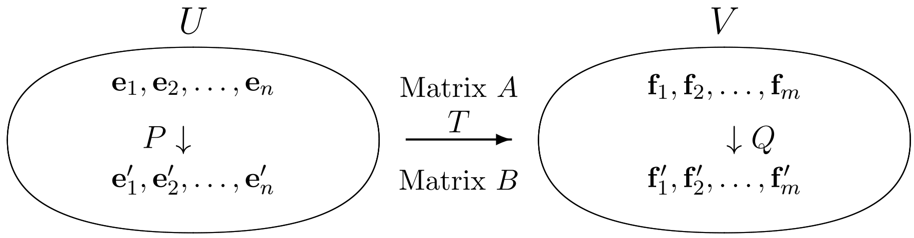
where $A = (\alpha_{ij}) = [F, T, E]$ is the $m \times n$ matrix of T with respect to the bases E and F of U and V.

Now choose new bases $E' = \{\mathbf{e}'_1, \dots, \mathbf{e}'_n\}$ of U and $F' = \{\mathbf{f}'_1, \dots, \mathbf{f}'_m\}$ of V. There is now a new matrix representing the linear transformation T:

$$T(\mathbf{e}'_j) = \sum_{i=1}^{m} \beta_{ij} \mathbf{f}'_i \text{ for } 1 \le j \le n,$$

where $B = (\beta_{ij}) = [F', T, E']$ is the $m \times n$ matrix of T with respect to the bases E' and F'. Our objective is to find the relationship between A and B in terms of the change of basis matrices.

Let the $n \times n$ matrix $P = (\sigma_{ij})$ be the change of basis matrix from E to E', and let the $m \times m$ matrix $Q = (\tau_{ij})$ be the change of basis matrix from F to F'.



Theorem 11.4. *With the above notation, we have $BP = QA$, or equivalently $B = QAP^{-1}$.*

Proof. By Theorem 7.4, BP represents the composite of the linear maps I_U using bases E and E' and T using bases E' and F'. So BP represents T using bases E and F'. Similarly, QA represents the composite of T using bases E and F and I_V using bases F and F', so QA also represents T using bases E and F'. Hence $BP = QA$. □

Another way to think of this is the following. The matrix B should be the matrix which, given the coordinates of a vector $\mathbf{u} \in U$ with respect to the basis E', produces the coordinates of $T(\mathbf{u}) \in V$ with respect to the basis F'. On the other hand, suppose we already know the matrix A, which performs the corresponding task with the "old" bases E and F. Now, given the coordinates of some vector $\mathbf{u}$ with respect to the "new" basis, we need to:

(i) Find the coordinates of $\mathbf{u}$ with respect to the "old" basis of U: this is done by multiplying by the change of basis matrix from E' to E, which is P^{-1};

(ii) find the coordinates of $T(\mathbf{u})$ with respect to the "old" basis of V: this is what multiplying by A does;

(iii) translate the result into coordinates with respect to the "new" basis for V; this is done by multiplying by the change of basis matrix Q.

Putting these three steps together, we again see that $B = QAP^{-1}$.

Corollary 11.5. *Two $m \times n$ matrices A and B represent the same linear map from an n-dimensional vector space U to an m-dimensional vector space V (with respect to different bases) if and only if there exist invertible $n \times n$ and $m \times m$ matrices P and Q with $B = QAP^{-1}$.*

Proof. For the forward direction, note that it follows from the Theorem 11.4 that A and B represent the same linear map if there exist change of basis matrices P and Q with $B = QAP^{-1}$, and by Proposition 11.2 the change of basis matrices are precisely invertible matrices of the correct size.

For the backward direction, suppose there exist invertible matrices Q, P such that $B = QAP^{-1}$. By Proposition 11.3, Q is a change of basis matrix between two bases F, F' of V, and P is a change of basis matrix between two bases of E, E' of U. As A corresponds to some linear map T with respect to the bases E and F (by Theorem 7.1), it follows from Theorem 11.4 that B is the matrix of the same linear map T corresponding to the bases E' and F'. □

Definition. Two $m \times n$ matrices A and B are said to be *equivalent* if there exist invertible P and Q with $B = QAP$. (Note that this is the same as saying that there exist invertible P and Q with $B = QAP^{-1}$, since P^{-1} is invertible itself.)

It is easy to check that being equivalent is an equivalence relation on the set $K^{m,n}$ of $m \times n$ matrices over K using Corollary 11.5. We shall show now that equivalence of matrices has other characterisations.

Theorem 11.6. *Let A and B be $m \times n$ matrices over K. Then the following conditions on A and B are equivalent.*

(i) A and B are equivalent.

(ii) A and B represent the same linear map with respect to different bases.

(iii) A and B have the same rank.

(iv) B can be obtained from A by application of elementary row and column operations.

Proof. (i) $\Leftrightarrow$ (ii): This is true by Corollary 11.5.

(ii) $\Rightarrow$ (iii): Since A and be both represent the same linear map T, we have $\operatorname{rank}(A) = \operatorname{rank}(T) = \operatorname{rank}(B)$ by Theorem 8.3.

(iii) $\Rightarrow$ (iv): By Theorem 8.2, if A and B both have rank s, then they can both be brought into the Smith normal form

$$E_s = \left(\begin{array}{c|c} I_s & \mathbf{0}_{s,n-s} \\ \hline \mathbf{0}_{m-s,s} & \mathbf{0}_{m-s,n-s} \end{array}\right)$$

by elementary row and column operations. Since these operations are invertible, we can first transform A to E_s and then transform E_s to B.

(iv) $\Rightarrow$ (i): We saw in Section 9.2 that applying an elementary row operation to A can be achieved by multiplying A on the left by an elementary row matrix, and similarly applying an elementary column operation can be done by multiplying A on the right by an elementary column matrix. Hence (iv) implies that there exist elementary row matrices $R_1, \ldots, R_r$ and elementary column matrices $C_1, \ldots, C_s$ with $B = R_r \cdots R_1 A C_1 \cdots C_s$. Since elementary matrices are invertible, $Q = R_r \cdots R_1$ and $P = C_1 \cdots C_s$ are invertible and $B = QAP$. $\square$

In the above proof, we also showed the following:

Proposition 11.7. *Any $m \times n$ matrix is equivalent to the matrix E_s defined above, where $s = \operatorname{rank}(A)$.*

This fact provides a further reason why the Smith normal form, also called the *canonical form* for $m \times n$ matrices under equivalence, is important: it is an easily recognizable representative of its equivalence class. This is one of the many examples in mathematics where it is useful to have a canonical form, i.e. a nice representative, for a class of objects.

12 Similar matrices, eigenvectors and eigenvalues

12.1 Similar matrices

In Section 11 we studied what happens to the matrix of a linear map $T: U \to V$ when we change the bases of U and V. Now we look at the case when $U = V$, where we only have a single vector space V, and a single change of basis. Surprisingly, this turns out to be more complicated than the situation with two different spaces.

Let V be a vector space of dimension n over the field K, and let $T: V \to V$ be a linear map. Now, given any basis for V, there will be a matrix representing T with respect to that basis.

Let $E = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ and $E' = \{\mathbf{e}'_1, \ldots, \mathbf{e}'_n\}$ be two bases of V, and let $A = (\alpha_{ij})$ and $B = (\beta_{ij})$ be the matrices of T with respect to E and E' respectively. Let $P = (\sigma_{ij})$ be the change of basis matrix from E' to E. *Note that this is the opposite change of basis to the one considered in the last section. Different textbooks adopt* *different conventions on which way round to do this; this is how we'll do it in this course.*

Then Theorem 11.4 applies, and with both Q and P replaced by P^{-1} we find:

Theorem 12.1. *With the notation above,* $B = P^{-1}AP$.

Definition. Two $n \times n$ matrices over K are said to be *similar* if there exists an $n \times n$ invertible matrix P with $B = P^{-1}AP$.

So two matrices are similar if and only if they represent the same linear map $T\colon V \to V$ with respect to different bases of V. It is easily checked that similarity is an equivalence relation on the set of $n \times n$ matrices over K.

We saw in Theorem 11.6 that two matrices of the same size are equivalent if and only if they have the same rank. It is more difficult to decide whether two matrices are similar, because we have much less flexibility - there is only one basis to choose, not two. Similar matrices are certainly equivalent, so they have the same rank, but equivalent matrices need not be similar.

Example. Let $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$.

Then A and B both have rank 2, so they are equivalent. However, since $A = I_2$, for any invertible 2×2 matrix P we have $P^{-1}AP = A$, so A is similar only to itself. Hence A and B are not similar.

To decide whether matrices are similar, it would be helpful to have a canonical form, just like we had the canonical form E_s in Section 11 for equivalence. Then we could test for similarity by reducing A and B to canonical form and checking whether we get the same result. But this turns out to be quite difficult, and depends on the field K. For the case $K = \mathbb{C}$ (the complex numbers), we have the *Jordan Canonical Form*, which Maths students learn about in the Second Year.

12.2 Eigenvectors and eigenvalues

In this course, we shall only consider the question of which matrices are similar to a diagonal matrix.

Definition. A matrix which is similar to a diagonal matrix is said to be *diagonalisable.*

(Recall that $A = (\alpha_{ij})$ is diagonal if $\alpha_{ij} = 0$ for $i \neq j$.) We shall see, for example, that the matrix B in the example above is not diagonalisable.

It turns out that the possible entries on the diagonal of a matrix similar to A can be calculated directly from A. They are called *eigenvalues* of A and depend only on the linear map to which A corresponds, and not on the particular choice of basis.

Definition. Let $T\colon V \to V$ be a linear map, where V is a vector space over K. Suppose that for some non-zero vector $\mathbf{v} \in V$ and some scalar $\lambda \in K$, we have $T(\mathbf{v}) = \lambda\mathbf{v}$. Then $\mathbf{v}$ is called an *eigenvector* of T, and λ is called the *eigenvalue* of T corresponding to $\mathbf{v}$.

Note that the zero vector is **not** an eigenvector. (This would not be a good idea, because $T\mathbf{0} = \lambda\mathbf{0}$ for all λ.) However, the zero scalar 0_K may sometimes be an eigenvalue (corresponding to some non-zero eigenvector).

Example. Let $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ be defined by $T(\alpha_1, \alpha_2) = (2\alpha_1, 0)$. Then $T(1, 0) = 2(1, 0)$, so 2 is an eigenvalue and $(1, 0)$ an eigenvector. Also $T(0, 1) = (0, 0) = 0(0, 1)$, so 0 is an eigenvalue and $(0, 1)$ an eigenvector.

In this example, notice that in fact $(\alpha, 0)$ and $(0, \alpha)$ are eigenvectors for any $\alpha \neq 0$. In general, it is easy to see that if $\mathbf{v}$ is an eigenvector of T, then so is $\alpha\mathbf{v}$ for any non-zero scalar α.

In some books, eigenvectors and eigenvalues are called *characteristic vectors* and *characteristic roots*, respectively.

Let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ be a basis of V, and let $A = (\alpha_{ij})$ be the matrix of T with respect to this basis. As in Section 7.1, to each vector $\mathbf{v} = \lambda_1\mathbf{e}_1 + \cdots + \lambda_n\mathbf{e}_n \in V$, we associate its column vector of coordinates

$$\underline{\mathbf{v}} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix} \in K^{n,1}.$$

Then, by Proposition 7.2, for $\mathbf{u}, \mathbf{v} \in V$, we have $T(\mathbf{u}) = \mathbf{v}$ if and only if $A\underline{\mathbf{u}} = \underline{\mathbf{v}}$, and in particular

$$T(\mathbf{v}) = \lambda\mathbf{v} \iff A\underline{\mathbf{v}} = \lambda\underline{\mathbf{v}}. \tag{4}$$

So it will be useful to define the eigenvalues and eigenvectors of a matrix, as well as of a linear map.

Definition. Let A be an $n \times n$ matrix over K. Suppose that, for some non-zero column vector $\underline{\mathbf{v}} \in K^{n,1}$ and some scalar $\lambda \in K$, we have $A\underline{\mathbf{v}} = \lambda\underline{\mathbf{v}}$. Then $\underline{\mathbf{v}}$ is called an *eigenvector* of A, and λ is called the *eigenvalue* of A corresponding to $\underline{\mathbf{v}}$.

It follows from (4) that if the matrix A corresponds to the linear map T, then λ is an eigenvalue of T if and only if it is an eigenvalue of A. Thus similar matrices have the same eigenvalues, because they represent the same linear map with respect to different bases. We shall give another proof of this fact in Theorem 12.3 below.

Given a matrix, how can we compute its eigenvalues? Certainly trying every vector to see whether it is an eigenvector is not a practical approach.

Theorem 12.2. *Let A be an $n \times n$ matrix. Then λ is an eigenvalue of A if and only if* $\det(A - \lambda I_n) = 0$.

Proof. Suppose that λ is an eigenvalue of A. Then $A\underline{\mathbf{v}} = \lambda\underline{\mathbf{v}}$ for some non-zero $\underline{\mathbf{v}} \in K^{n,1}$. This is equivalent to $A\underline{\mathbf{v}} = \lambda I_n\underline{\mathbf{v}}$, or $(A - \lambda I_n)\underline{\mathbf{v}} = \underline{\mathbf{0}}$. But this says exactly that $\underline{\mathbf{v}}$ is a non-zero solution to the homogeneous system of simultaneous equations defined by the matrix $A - \lambda I_n$, and then by Theorem 9.6(i), $A - \lambda I_n$ is singular, and so $\det(A - \lambda I_n) = 0$ by Theorem 10.5.

Conversely, if $\det(A - \lambda I_n) = 0$ then $A - \lambda I_n$ is singular, and so by Theorem 9.6(i) the system of simultaneous equations defined by $A - \lambda I_n$ has nonzero solutions. Hence there exists a non-zero $\underline{\mathbf{v}} \in K^{n,1}$ with $(A - \lambda I_n)\underline{\mathbf{v}} = \underline{\mathbf{0}}$, which is equivalent to $A\underline{\mathbf{v}} = \lambda I_n\underline{\mathbf{v}}$, and so λ is an eigenvalue of A. □

If we treat λ as an unknown, we get a polynomial equation which we can solve to find all the eigenvalues of A:

Definition. For an $n \times n$ matrix A, the equation $\det(A - xI_n) = 0$ is called the *characteristic equation* of A, and $\det(A - xI_n)$ is called the *characteristic polynomial* of A.

Note that the characteristic polynomial of an $n \times n$ matrix is a polynomial of degree n in x.

The above theorem says that the eigenvalues of A are the roots of the characteristic equation, which means that we have a method of calculating them. Once the eigenvalues are known, it is then straightforward to compute the corresponding eigenvectors.

Example. Let $A = \begin{pmatrix} 1 & 2 \\ 5 & 4 \end{pmatrix}$. Then

$$\det(A - xI_2) = \begin{vmatrix} 1-x & 2 \\ 5 & 4-x \end{vmatrix} = (1-x)(4-x) - 10 = x^2 - 5x - 6 = (x-6)(x+1).$$

Hence the eigenvalues of A are the roots of $(x-6)(x+1) = 0$; that is, 6 and -1.

Let us now find the eigenvectors corresponding to the eigenvalue 6. We seek a non-zero column vector $\left(\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right)$ such that

$$\begin{pmatrix} 1 & 2 \\ 5 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 6 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}; \text{ that is, } \begin{pmatrix} -5 & 2 \\ 5 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Solving this easy system of linear equations, we can take $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$ to be our eigenvector; or indeed any non-zero multiple of $\begin{pmatrix} 2 \\ 5 \end{pmatrix}$.

Similarly, for the eigenvalue -1, we want a non-zero column vector $\left(\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right)$ such that

$$\begin{pmatrix} 1 & 2 \\ 5 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = -1 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}; \text{ that is, } \begin{pmatrix} 2 & 2 \\ 5 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

and we can take $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ to be our eigenvector.

Example. This example shows that the eigenvalues can depend on the field K. Let

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \text{ Then } \det(A - xI_2) = \begin{vmatrix} -x & -1 \\ 1 & -x \end{vmatrix} = x^2 + 1,$$

so the characteristic equation is $x^2 + 1 = 0$. If $K = \mathbb{R}$ (the real numbers) then this equation has no solutions, so there are no eigenvalues or eigenvectors. However, if $K = \mathbb{C}$ (the complex numbers), then there are two eigenvalues i and $-i$, and by a similar calculation to the one in the last example, we find that $\left(\begin{smallmatrix} -1 \\ i \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 1 \\ i \end{smallmatrix}\right)$ are eigenvectors corresponding to i and $-i$ respectively.

Theorem 12.3. *Similar matrices have the same characteristic equation and hence the same eigenvalues.*

Proof. Let A and B be similar matrices. Then there exists an invertible matrix P with $B = P^{-1}AP$. Then

$$\begin{aligned} \det(B - xI_n) &= \det(P^{-1}AP - xI_n) \\ &= \det(P^{-1}(A - xI_n)P) \\ &= \det(P^{-1})\det(A - xI_n)\det(P) \text{ (by Theorem 10.3)} \\ &= \det(P^{-1})\det(P)\det(A - xI_n) \\ &= \det(A - xI_n). \end{aligned}$$

Hence A and B have the same characteristic equation. Since the eigenvalues are the roots of the characteristic equation, they have the same eigenvalues. $\square$

Since the different matrices corresponding to a linear map T are all similar, they all have the same characteristic equation, so we can unambiguously refer to it also as the characteristic equation of T if we want to.

There is one case where the eigenvalues can be written down immediately.

Proposition 12.4. *Suppose that the matrix A is upper triangular. Then the eigenvalues of A are just the diagonal entries α_{ii} of A.*

Proof. We saw in Corollary 10.2 that the determinant of A is the product of the diagonal entries α_{ii}. Hence the characteristic polynomial of such a matrix is $\prod_{i=1}^{n}(\alpha_{ii}-x)$, and so the eigenvalues are the α_{ii}. $\square$

Example. Let $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Then A is upper triangular, so its only eigenvalue is 1. We can now see that A cannot be similar to any diagonal matrix B. Such a B would also have just 1 as an eigenvalue, and then, by Corollary 10.2 again, this would force B to be the identity matrix I_2. But $P^{-1}I_2P = I_2$ for any invertible matrix P, so I_2 is not similar to any matrix other than itself! So A cannot be similar to I_2, and hence A is not diagonalisable.

The next theorem describes the connection between diagonalisable matrices and eigenvectors. If you have understood everything so far then its proof should be almost obvious.

Theorem 12.5. *Let $T\colon V \to V$ be a linear map. Then there is a basis of V with respect to which the matrix of T is diagonal if and only if there is a basis of V consisting of eigenvectors of T.*

Equivalently, let A be an $n \times n$ matrix over K. Then A is similar to a diagonal matrix if and only if the space $K^{n,1}$ has a basis consisting of eigenvectors of A.

Proof. The equivalence of the two statements follows directly from the correspondence between linear maps and matrices, and the corresponding definitions of eigenvectors and eigenvalues.

Suppose that the matrix $A = (\alpha_{ij})$ of T is diagonal with respect to the basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of V. Recall from Section 7.1 that the images of the ith basis vector of V is represented by the ith column of A. But since A is diagonal, this column has the single non-zero entry α_{ii}. Hence $T(\mathbf{e}_i) = \alpha_{ii}\mathbf{e}_i$, and so each basis vector $\mathbf{e}_i$ is an eigenvector of A.

Conversely, suppose that $\mathbf{e}_1, \ldots, \mathbf{e}_n$ is a basis of V consisting entirely of eigenvectors of T. Then, for each i, we have $T(\mathbf{e}_i) = \lambda_i\mathbf{e}_i$ for some $\lambda_i \in K$. But then the matrix of T with respect to this basis is the diagonal matrix $A = (\alpha_{ij})$ with $\alpha_{ii} = \lambda_i$ for each i. $\square$

We now show that A is diagonalisable in the case when there are n distinct eigenvalues.

Theorem 12.6. *Let $\lambda_1, \ldots, \lambda_r$ be* distinct *eigenvalues of $T\colon V \to V$, and let $\mathbf{v}_1, \ldots, \mathbf{v}_r$ be corresponding eigenvectors. (So $T(\mathbf{v}_i) = \lambda_i\mathbf{v}_i$ for $1 \le i \le r$.) Then $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are linearly independent.*

Proof. We prove this by induction on r. It is true for $r = 1$, because eigenvectors are non-zero by definition. For $r > 1$, suppose that for some $\alpha_1, \ldots, \alpha_r \in K$ we have

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_r\mathbf{v}_r = \mathbf{0}.$$

Then, applying T to this equation gives

$$\alpha_1\lambda_1\mathbf{v}_1 + \alpha_2\lambda_2\mathbf{v}_2 + \cdots + \alpha_r\lambda_r\mathbf{v}_r = \mathbf{0}.$$

Now, subtracting λ_1 times the first equation from the second gives

$$\alpha_2(\lambda_2 - \lambda_1)\mathbf{v}_2 + \cdots + \alpha_r(\lambda_r - \lambda_1)\mathbf{v}_r = \mathbf{0}.$$

By inductive hypothesis, $\mathbf{v}_2, \ldots, \mathbf{v}_r$ are linearly independent, so $\alpha_i(\lambda_i - \lambda_1) = 0$ for $2 \leq i \leq r$. But, by assumption, $\lambda_i - \lambda_1 \neq 0$ for $i > 1$, so we must have $\alpha_i = 0$ for $i > 1$. But then $\alpha_1 \mathbf{v}_1 = \mathbf{0}$, so α_1 is also zero. Thus $\alpha_i = 0$ for all i, which proves that $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are linearly independent. □

Corollary 12.7. *If the linear map $T\colon V \to V$ (or equivalently the $n \times n$ matrix A) has n distinct eigenvalues, where $n = \dim(V)$, then T (or A) is diagonalisable.*

Proof. Under the hypothesis, there are n linearly independent eigenvectors, which form a basis of V by Corollary 3.9. The result follows from Theorem 12.5. □

Example.

$$A = \begin{pmatrix} 4 & 5 & 2 \\ -6 & -9 & -4 \\ 6 & 9 & 4 \end{pmatrix}. \quad \text{Then } |A - xI_3| = \begin{vmatrix} 4-x & 5 & 2 \\ -6 & -9-x & -4 \\ 6 & 9 & 4-x \end{vmatrix}.$$

To help evaluate this determinant, apply first the row operation $\mathbf{r}_3 \to \mathbf{r}_3 + \mathbf{r}_2$ and then the column operation $\mathbf{c}_2 \to \mathbf{c}_2 - \mathbf{c}_3$, giving

$$|A - xI_3| = \begin{vmatrix} 4-x & 5 & 2 \\ -6 & -9-x & -4 \\ 0 & -x & -x \end{vmatrix} = \begin{vmatrix} 4-x & 3 & 2 \\ -6 & -5-x & -4 \\ 0 & 0 & -x \end{vmatrix},$$

and then expanding by the third row we get

$$|A - xI_3| = -x\big((4-x)(-5-x) + 18\big) = -x(x^2 + x - 2) = -x(x+2)(x-1)$$

so the eigenvalues are 0, 1 and -2. Since these are distinct, we know from the above corollary that A can be diagonalised. In fact, the eigenvectors will be the new basis with respect to which the matrix is diagonal, so we will calculate these.

In the following calculations, we will denote eigenvectors $\underline{\mathbf{v}}_1$, etc. by $\left(\begin{smallmatrix} x_1 \\ x_2 \\ x_3 \end{smallmatrix}\right)$, where x_1, x_2, x_3 need to be calculated by solving simultaneous equations.

For the eigenvalue $\lambda = 0$, an eigenvector $\underline{\mathbf{v}}_1$ satisfies $A\underline{\mathbf{v}}_1 = \underline{\mathbf{0}}$, which gives the three equations:

$$4x_1 + 5x_2 + 2x_3 = 0; \qquad -6x_1 - 9x_2 - 4x_3 = 0; \qquad 6x_1 + 9x_2 + 4x_3 = 0.$$

The third is clearly redundant, and adding twice the first to the second gives $2x_1 + x_2 = 0$ and then we see that one solution is $\underline{\mathbf{v}}_1 = \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix}$.

For $\lambda = 1$, we want an eigenvector $\mathbf{v}_2$ with $A\underline{\mathbf{v}}_2 = \underline{\mathbf{v}}_2$, which gives the equations

$$4x_1 + 5x_2 + 2x_3 = x_1; \qquad -6x_1 - 9x_2 - 4x_3 = x_2; \qquad 6x_1 + 9x_2 + 4x_3 = x_3;$$

or equivalently

$$3x_1 + 5x_2 + 2x_3 = 0; \qquad -6x_1 - 10x_2 - 4x_3 = 0; \qquad 6x_1 + 9x_2 + 3x_3 = 0.$$

Adding the second and third equations gives $x_2 + x_3 = 0$ and then we see that a solution is $\underline{\mathbf{v}}_2 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$.

Finally, for $\lambda = -2$, $A\underline{\mathbf{v}}_3 = -2\underline{\mathbf{v}}_3$ gives the equations

$$6x_1 + 5x_2 + 2x_3 = 0; \qquad -6x_1 - 7x_2 - 4x_3 = 0; \qquad 6x_1 + 9x_2 + 6x_3 = 0,$$

of which one solution is $\underline{\mathbf{v}}_3 = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}$.

Now, if we change basis to $\underline{\mathbf{v}}_1, \underline{\mathbf{v}}_2, \underline{\mathbf{v}}_3$, we should get the diagonal matrix with the eigenvalues $0, 1, -2$ on the diagonal. We can check this by direct calculation. Remember that P is the change of basis matrix from the new basis to the old one and has columns the new basis vectors expressed in terms of the old. But the old basis is the standard basis, so the columns of P are the new basis vectors. Hence

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

and, according to Theorem 12.1, we should have $P^{-1}AP = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}$.

To check this, we first need to calculate P^{-1}, either by row reduction or by the cofactor method. The answer turns out to be

$$P^{-1} = \begin{pmatrix} 0 & 1 & 1 \\ 2 & 1 & 0 \\ -1 & -2 & -1 \end{pmatrix},$$

and now we can check that the above equation really does hold.

Warning! The converse of Corollary 12.7 is not true. If it turns out that there do not exist n distinct eigenvalues, then you cannot conclude from this that the matrix is not diagonalisable. This is really rather obvious, because the identity matrix has only a single eigenvalue, but it is diagonal already. Even so, this is one of the most common mistakes that students make.

If there are fewer than n distinct eigenvalues, then the matrix may or may not be diagonalisable, and you have to test directly to see whether there are n linearly independent eigenvectors. Let us consider two rather similar looking examples:

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

Both matrices are upper triangular, so we know from Proposition 12.4 that both have eigenvalues 1 and -1, with 1 repeated. Since -1 occurs only once, it can only have a single associated linearly independent eigenvector. (Can you prove that?) Solving the equations as usual, we find that A_1 and A_2 have eigenvectors $\begin{pmatrix} 1 \\ -2 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$, respectively, associated with eigenvalue -1.

The repeated eigenvalue 1 is more interesting, because there could be one or two associated linearly independent eigenvectors. The equation $A_1\underline{\mathbf{x}} = \underline{\mathbf{x}}$ gives the equations

$$x_1 + x_2 + x_3 = x_1; \qquad -x_2 + x_3 = x_2; \qquad x_3 = x_3,$$

so $x_2 + x_3 = -2x_2 + x_3 = 0$, which implies that $x_2 = x_3 = 0$. Hence the only eigenvectors are multiples of $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$. Hence A_1 has only two linearly independent eigenvectors in total, and so it cannot be diagonalised.

On the other hand, $A_2\underline{\mathbf{x}} = \underline{\mathbf{x}}$ gives the equations

$$x_1 + 2x_2 - 2x_3 = x_1; \qquad -x_2 + 2x_3 = x_2; \qquad x_3 = x_3,$$

which reduce to the single equation $x_2 - x_3 = 0$. This time there are two linearly independent solutions, giving eigenvectors $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$. So A_2 has three linearly independent eigenvectors in total, and it can be diagonalised. In fact, using the eigenvectors as columns of the change of basis matrix P as before gives

$$P = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 1 & 0 \end{pmatrix} \text{ and we compute } P^{-1} = \begin{pmatrix} 1 & 1 & -1 \\ 0 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}.$$

We can now check that $P^{-1}A_2P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$, as expected.