MA251: Algebra I – Advanced Linear Algebra

Dmitriy Rumynin and Adam Thomas
Based on notes of Derek Holt and David Loeffler

Term 1, 2020-21

Contents

0 Blurb

Week 1

As its title suggests, this module is a continuation of last year's MA106 Linear Algebra module; we'll be studying vector spaces, linear maps, and their properties in a bit more detail. Later in the module we'll think a bit about matrices whose entries lie not in a field but in the integers $\mathbb{Z}$, and we'll see what our methods have to tell us in that case.

1 Review of some MA106 material

In this section, we'll recall some ideas from the first year MA106 Linear Algebra module. This will just be a brief reminder; for detailed statements and proofs, go back to your MA106 notes! (You didn't throw them away, did you? I hope not.)

1.1 Fields

Recall that a *field* is a number system where we know how to do all of the basic arithmetic operations: we can add, subtract, multiply and divide (as long as we're not trying to divide by zero).

Examples.

- A non-example is $\mathbb{Z}$, the integers. Here we can add, subtract, and multiply, but we can't always divide without jumping out of $\mathbb{Z}$ into some bigger world.
- The real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$ are fields, and these are perhaps the most familiar ones.
- The rational numbers $\mathbb{Q}$ are also a field.
- A more subtle example: if p is a prime number, the integers mod p are a field, written as $\mathbb{Z}/p\mathbb{Z}$ or $\mathbb{F}_p$.

There are lots of fields out there, and the reason we take the axiomatic approach is that we know that everything we prove will be applicable to any field we like, as long as we've only used the field axioms in our proofs (rather than any specific properties of the fields we happen to most like). We don't have to know all the fields in existence and check that our proofs are valid for each one separately.

1.2 Vector spaces

Let K be a field[1]. A *vector space* over K is a set V together with two extra pieces of structure. Firstly, it has to have a notion of *addition*: we need to know what $v + w$ means if v and w are in V. Secondly, it has to have a notion of *scalar multiplication*: we need to know what λv means if v is in V and λ is in K. These have to satisfy some axioms, for which I'm going to refer you again to your MA106 notes.

A *basis* of a vector space is a subset $B \subset V$ such that every $v \in V$ can be written as a finite linear combination of elements of B,

$$v = \lambda_1 b_1 + \cdots + \lambda_n b_n,$$

for some $n \in \mathbb{N}$ and some $\lambda_1, \ldots, \lambda_n \in K$; and for each $v \in V$, we can do this in one and only one way. Another way of saying this is that B is a linearly independent set which spans V,

[1] It's conventional to use K as the letter to denote a field; the K stands for the German word "Körper".

which is the definition you had in MA106. We say V is *finite-dimensional* if there is a finite basis of V. You saw last year that if V has one basis which is finite, then every basis of V is finite, and they all have the same size; and we define the *dimension* of V to be this magic number which is the size of any basis of V.

Examples. Let $K = \mathbb{R}$.

- The space of polynomials in x with coefficients in $\mathbb{R}$ is certainly a vector space over $\mathbb{R}$; but it's not finite-dimensional (rather obviously).
- For any $d \in \mathbb{N}$, the set $\mathbb{R}^d$ of column vectors with d real entries is a vector space over $\mathbb{R}$ (which, not surprisingly, has dimension d).
- The set
$$\left\{ \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \in \mathbb{R}^3 : x_1 + x_2 + x_3 = 0 \right\}$$
is a vector space over $\mathbb{R}$.

The third example above is an interesting one because there's no "natural choice" of basis. It certainly has bases, e.g. the set
$$\left\{ \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \right\},$$
but there's no reason why that's better than any other one. This is one of the reasons why we need to worry about the choice of basis – if you want to tell someone else all the wonderful things you've found out about this vector space, you might get into a total muddle if you insisted on using one particular basis and they preferred another different one.

1.3 Linear maps

If V and W are vector spaces (over the same field K), then a *linear map* from V to W is a map $T : V \to W$ which "respects the vector space structure". That is, we know two things that we can do with vectors in a vector space – add them, and multiply them by scalars; and a linear map is a map where adding or scalar-multiplying on the V side, then applying the map T, is the same as applying the map T, then adding or multiplying on the W side. Formally, for T to be a linear map means that we must have
$$T(v_1 + v_2) = T(v_1) + T(v_2) \quad \forall\, v_1, v_2 \in V$$
and
$$T(\lambda v_1) = \lambda T(v_1) \quad \forall\, \lambda \in K, v_1 \in V.$$

1.4 The matrix of a linear map with respect to a choice of bases

Let V and W be vector spaces over a field K. Let $T : V \to W$ be a linear map, where $\dim(V) = n$, $\dim(W) = m$. Choose a basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of V and a basis $\mathbf{f}_1, \ldots, \mathbf{f}_m$ of W.

Now, for $1 \leq j \leq n$, $T(\mathbf{e}_j) \in W$, 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 $1 \leq i \leq m$, $1 \leq j \leq n$) are uniquely determined.

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

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

over K. Then A is called the matrix of the linear map T with respect to the chosen bases of V and W. Note that the columns of A are the images $T(\mathbf{e}_1), \ldots, T(\mathbf{e}_n)$ of the basis vectors of V represented as column vectors with respect to the basis $\mathbf{f}_1, \ldots, \mathbf{f}_m$ of W.

It was shown in MA106 that T is uniquely determined by A, and so there is a one-one correspondence between linear maps $T : V \to W$ and $m \times n$ matrices over K, which depends on the choice of bases of V and W.

For $\mathbf{v} \in V$, we can write $\mathbf{v}$ uniquely as a linear combination of the basis vectors $\mathbf{e}_i$; that is, $\mathbf{v} = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n$, where the x_i are uniquely determined by $\mathbf{v}$ and the basis $\mathbf{e}_i$. We shall call x_i the *coordinates* of $\mathbf{v}$ with respect to the basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$. We associate the column vector

$$\underline{\mathbf{v}} = \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{pmatrix} \in K^{n,1},$$

to $\mathbf{v}$, where $K^{n,1}$ denotes the space of $n \times 1$-column vectors with entries in K.

It was proved in MA106 that if A is the matrix of the linear map T, then for $\mathbf{v} \in V$, we have $T(\mathbf{v}) = \mathbf{w}$ if and only if $A\underline{\mathbf{v}} = \underline{\mathbf{w}}$, where $\underline{\mathbf{w}} \in K^{m,1}$ is the column vector associated with $\mathbf{w} \in W$.

1.5 Change of basis

Let V be a vector space of dimension n over a field K, and let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ and $\mathbf{e}'_1, \ldots, \mathbf{e}'_n$ be two bases of V. Then there is an invertible $n \times n$ matrix $P = (p_{ij})$ such that

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

Note that the columns of P are the new basis vectors $\mathbf{e}'_i$ written as column vectors in the old basis vectors $\mathbf{e}_i$. (Recall also that P is the matrix of the identity map $V \to V$ using basis $\mathbf{e}'_1, \ldots, \mathbf{e}'_n$ in the domain and basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ in the codomain.)

Usually the original basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ will be the standard basis of K^n.

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

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

The following result was proved in MA106.

Proposition 1.5.1. *With the above notation, let $\mathbf{v} \in V$, and let $\underline{\mathbf{v}}$ and $\underline{\mathbf{v}}'$ denote the column vectors associated with $\mathbf{v}$ when we use the bases $\mathbf{e}_1, \ldots, \mathbf{e}_n$ and $\mathbf{e}'_1, \ldots, \mathbf{e}'_n$, respectively. Then $P\underline{\mathbf{v}}' = \underline{\mathbf{v}}$.*

So, in the example above, if we take $\mathbf{v} = \begin{pmatrix}1\\-2\\4\end{pmatrix}$, then we have $\mathbf{v} = \mathbf{e}_1 - 2\mathbf{e}_2 + 4\mathbf{e}_3$ (obviously); so the coordinates of $\mathbf{v}$ in the basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ are $\underline{\mathbf{v}} = \begin{pmatrix}1\\-2\\4\end{pmatrix}$.

On the other hand, we also have $\mathbf{v} = 2\mathbf{e}'_1 - 2\mathbf{e}'_2 - 3\mathbf{e}'_3$, so the coordinates of $\mathbf{v}$ in the basis $\{\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3\}$ are

$$\underline{\mathbf{v}}' = \begin{pmatrix}2\\-2\\-3\end{pmatrix},$$

and you can check that

$$P\underline{\mathbf{v}}' = \begin{pmatrix}0 & 1 & -1\\ 1 & 2 & 0\\ 2 & 0 & 0\end{pmatrix}\begin{pmatrix}2\\-2\\-3\end{pmatrix} = \begin{pmatrix}1\\-2\\4\end{pmatrix} = \underline{\mathbf{v}},$$

just as Proposition 1.5.1 says.

This equation $P\underline{\mathbf{v}}' = \underline{\mathbf{v}}$ describes the change of coordinates associated with our basis change. If we want to compute the new coordinates from the old ones, we need to use the inverse matrix: $\underline{\mathbf{v}}' = P^{-1}\underline{\mathbf{v}}$. Thus, to enable calculations in the new basis we need both matrices P and P^{-1}. We'll be using this relationship over and over again, so make sure you're happy with it!

Which matrix, P or P^{-1} should be called the *basis change matrix* or *transition matrix* from the original basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ to the new basis $\mathbf{e}'_1, \ldots, \mathbf{e}'_n$?

Well, the books are split on this. As a historic quirk, the *basis change matrix* in Algebra-1 was always P and the *basis change matrix* in Linear Algebra was P^{-1} since around 2011. We continue with this noble tradition of calling P the *basis change matrix* because, otherwise, we risk introducing typos throughout the text.

Now let $T : V \to W$, $\mathbf{e}_i$, $\mathbf{f}_i$ and A be as in Subsection 1.4 above, and choose new bases $\mathbf{e}'_1, \ldots, \mathbf{e}'_n$ of V and $\mathbf{f}'_1, \ldots, \mathbf{f}'_m$ of W. Then

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

where $B = (\beta_{ij})$ is the $m \times n$ matrix of T with respect to the bases $\{\mathbf{e}'_i\}$ and $\{\mathbf{f}'_i\}$ of V and W. Let the $n \times n$ matrix $P = (p_{ij})$ be the basis change matrix for the original basis $\{\mathbf{e}_i\}$ and new basis $\{\mathbf{e}'_i\}$, and let the $m \times m$ matrix $Q = (q_{ij})$ be the basis change matrix for original basis $\{\mathbf{f}_i\}$ and new basis $\{\mathbf{f}'_i\}$. The following theorem was proved in MA106:

Theorem 1.5.2. *With the above notation, we have $AP = QB$, or equivalently $B = Q^{-1}AP$.*

In most of the applications in this module we will have $V = W$ $(= K^n)$, $\{\mathbf{e}_i\} = \{\mathbf{f}_i\}$, and $\{\mathbf{e}'_i\} = \{\mathbf{f}'_i\}$. So $P = Q$, and hence $B = P^{-1}AP$.

2 The Jordan Canonical Form

2.1 Introduction

Throughout this section V will be a vector space of dimension n over a field K, $T : V \to V$ will be a linear map, and A will be the matrix of T with respect to a fixed basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of V. Our aim is to find a new basis $\mathbf{e}'_1, \ldots, \mathbf{e}'_n$ for V, such that the matrix of T with respect to the new basis is as simple as possible. Equivalently (by Theorem 1.5.2), we want to find an invertible matrix P (the associated basis change matrix) such that $P^{-1}AP$ is a simple as possible.

(Recall that if B is a matrix which can be written in the form $B = P^{-1}AP$, we say B is *similar* to A. So a third way of saying the above is that we want to find a matrix that's similar to A, but which is as nice as possible.)

Our preferred form of a matrix is a diagonal matrix. So we'd really rather like it if every matrix was similar to a diagonal matrix. But this won't work: we saw in MA106 that the matrix $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, for example, is not similar to a diagonal matrix. (We say this matrix is not *diagonalizable*.)

The point of this section of the module is to show that although we can't always get to a diagonal matrix, we can get pretty close (at least if K is $\mathbb{C}$). Under this assumption, it can be proved that A is always similar to a matrix B of a certain type (called the *Jordan canonical form* or sometimes *Jordan normal form* of the matrix), which is not far off being diagonal: its only non-zero entries are on the diagonal or just above it.

2.2 Eigenvalues and eigenvectors

We start by summarising some of what we know from MA106 which is going to be relevant to us here.

If we can find some $0 \neq \mathbf{v} \in V$ and $\lambda \in K$ such that $T\mathbf{v} = \lambda\mathbf{v}$, or equivalently $A\underline{\mathbf{v}} = \lambda\underline{\mathbf{v}}$, then λ is an *eigenvalue*, and $\mathbf{v}$ a corresponding *eigenvector* of T (or of A).

From MA106, you have a theorem that tells you when a matrix is diagonalizable:

Proposition 2.2.1. *Let $T : V \to V$ be a linear map. Then the matrix of T is diagonal with respect to some basis of V if and only if V has a basis consisting of eigenvectors of T.*

This is a nice theorem, but it doesn't tell you how you might find such a basis! But there's one case where it's easy, as another theorem from MA106 tells us:

Proposition 2.2.2. *Let $\lambda_1, \ldots, \lambda_r$ be distinct eigenvalues of $T : V \to V$, and let $\mathbf{v}_1, \ldots, \mathbf{v}_r$ be corresponding eigenvectors. (So $T(\mathbf{v}_i) = \lambda_i\mathbf{v}_i$ for $1 \leq i \leq r$.) Then $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are linearly independent.*

Corollary 2.2.3. *If the linear map $T : V \to V$ (or equivalently the $n \times n$ matrix A) has n distinct eigenvalues, where $n = \dim(V)$, then T (or A) is diagonalizable.*

2.3 The minimal polynomial

Week 2

We start this section with what might look like a bit of a digression; but it'll be really relevant later! We'll spend a little while thinking about *polynomials* in a single variable x with coefficients in our field K; e.g. if K is $\mathbb{C}$ (or $\mathbb{Q}$) we could take $p = p(x) = 2x^2 - \frac{3}{2}x + 11$. The set of all such polynomials is denoted by $K[x]$.

If $A \in K^{n,n}$ is a square $n \times n$ matrix over K, and $p \in K[x]$ is a polynomial, then we know what $p(A)$ means: we just calculate the powers of A in the usual way, and then plug them into the formula defining p, interpreting the constant term as a multiple of I_n.

For instance, if $K = \mathbb{Q}$, p is the polynomial I just wrote down, and $A = \begin{pmatrix} 2 & 3 \\ 0 & 1 \end{pmatrix}$, then $A^2 = \begin{pmatrix} 4 & 9 \\ 0 & 1 \end{pmatrix}$, and

$$\begin{aligned} p(A) &= 2\begin{pmatrix} 4 & 9 \\ 0 & 1 \end{pmatrix} - \tfrac{3}{2}\begin{pmatrix} 2 & 3 \\ 0 & 1 \end{pmatrix} + 11\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 16 & 27/2 \\ 0 & 23/2 \end{pmatrix}. \end{aligned}$$

Warning. Notice that this is *not* the same as the matrix $\begin{pmatrix} p(2) & p(3) \\ p(0) & p(1) \end{pmatrix}$.

Theorem 2.3.1. *Let $A \in K^{n,n}$. Then there is some non-zero polynomial $p \in K[x]$ of degree at most n^2 such that $p(A)$ is the $n \times n$ zero matrix $\mathbf{0}_n$.*

Proof. This is easy, but requires a slightly odd mental contortion. The key thing to observe is that $K^{n,n}$, the space of $n \times n$ matrices over K, is itself a vector space over K. Its dimension is n^2.

Let's consider the set $\{I_n, A, A^2, \ldots, A^{n^2}\} \subset K^{n,n}$. Since this is a set of $n^2 + 1$ vectors in an n^2-dimensional vector space, there had better be a linear relation between them. That is, we can find constants $\lambda_0, \lambda_1, \ldots, \lambda_{n^2}$, not all zero, such that

$$\lambda_0 I_n + \cdots + \lambda_{n^2} A^{n^2} = \mathbf{0}_n.$$

Now we define the polynomial $p = \lambda_0 + \lambda_1 x + \cdots + \lambda_{n^2} x^{n^2}$. This isn't zero, and its degree is at most n^2. (It might be less, since λ_{n^2} might be 0.) Then that's it! We've just shown that $p(A) = \mathbf{0}_n$, so we're done! We've plucked a polynomial that A satisfies out of thin air. □

Now, the problem with this argument is that there's no reason why the polynomial it produces should be unique. In general, there might be many linear relations you could write down between those $n^2 + 1$ vectors, so there could be lots of polynomials satisfied by A. Is there a way of finding a particularly nice polynomial that A satisfies? To answer that question, we'll have to think a little bit about arithmetic in $K[x]$.

Note that we can do "division" with polynomials, a bit like with integers. We can divide one polynomial p (with $p \neq 0$) into another polynomial q and get a remainder with degree less than p. For example, if $q = x^5 - 3$, $p = x^2 + x + 1$, then we find $q = sp + r$ with $s = x^3 - x^2 + 1$ and $r = -x - 4$.

If the remainder is 0, so $p = qr$ for some r, we say "p divides q" and write this relation as $p \mid q$.

Finally, a polynomial with coefficients in a field K is called *monic* if the coefficient of the highest power of x is 1. So, for example, $x^3 - 2x^2 + x + 11$ is monic, but $2x^2 - x - 1$ is not.

Theorem 2.3.2. *Let A be an $n \times n$ matrix over K representing the linear map $T : V \to V$. Then*

(i) There is a unique monic non-zero polynomial $p(x)$ with minimal degree and coefficients in K such that $p(A) = 0$.

(ii) If $q(x)$ is any polynomial with $q(A) = 0$, then $p \mid q$.

Proof. (i) If we have any polynomial $p(x)$ with $p(A) = 0$, then we can make p monic by multiplying it by a constant. By Theorem 2.3.1, there exists such a $p(x)$, so there exists one of minimal degree. If we had two distinct monic polynomials $p_1(x)$, $p_2(x)$ of the same minimal degree with $p_1(A) = p_2(A) = 0$, then $p = p_1 - p_2$ would be a non-zero polynomial of smaller degree with $p(A) = 0$, contradicting the minimality of the degree, so p is unique.

(ii) Let $p(x)$ be the minimal monic polynomial in (i) and suppose that $q(A) = 0$. As we saw above, we can write $q = sp + r$ where r has smaller degree than p. If r is non-zero, then $r(A) = q(A) - s(A)p(A) = 0$ contradicting the minimality of p, so $r = 0$ and $p \mid q$. □

Definition 2.3.3. The unique monic polynomial $\mu_A(x)$ of minimal degree with $\mu_A(A) = 0$ is called the *minimal polynomial* of A.

We know that for $p \in K[x]$, $p(T) = \mathbf{0}_V$ if and only if $p(A) = \mathbf{0}_n$; so μ_A is also the unique monic polynomial of minimal degree such that $\mu_A(T) = 0$ (the minimal polynomial of T.) In particular, since similar matrices A and B represent the same linear map T, and their minimal polynomial is the same as that of T, we have

Proposition 2.3.4. *Similar matrices have the same minimal polynomial.*

By Theorem 2.3.1 and Theorem 2.3.2 (ii), we have

Corollary 2.3.5. *The minimal polynomial of an $n \times n$ matrix A has degree at most n^2.*

(In the next section, we'll see that we can do much better than this.)

Example. If D is a diagonal matrix, say

$$D = \begin{pmatrix} d_{11} & & \\ & \ddots & \\ & & d_{nn} \end{pmatrix},$$

then for any polynomial p we see that $p(D)$ is the diagonal matrix with entries

$$\begin{pmatrix} p(d_{11}) & & \\ & \ddots & \\ & & p(d_{nn}) \end{pmatrix}.$$

Hence $p(D) = 0$ if and only if $p(d_{ii}) = 0$ for all i. So for instance if

$$D = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

the minimal polynomial of D is the smallest-degree polynomial which has 2 and 3 as roots, which is clearly $\mu_D(x) = (x-2)(x-3) = x^2 - 5x + 6$.

We can generalize this example as follows

Proposition 2.3.6. *Let D be any diagonal matrix and let $\{\delta_1, \ldots, \delta_r\}$ be the* set *of diagonal entries of D (i.e. without any repetitions, so the values $\delta_1, \ldots, \delta_r$ are all different). Then we have*

$$\mu_D(x) = (x - \delta_1)(x - \delta_2) \ldots (x - \delta_r).$$

Proof. As in the example, we have $p(D) = 0$ if and only if $p(\delta_i) = 0$ for all $i \in \{1, \ldots, r\}$. The smallest-degree monic polynomial vanishing at these points is clearly the polynomial above. □

Corollary 2.3.7. *If A is any diagonalizable matrix, then $\mu_A(x)$ is a product of distinct linear factors.*

Proof. Clear from Proposition 2.3.6 and Proposition 2.3.4. □

Remark. *We'll see later in the course that this is a necessary and sufficient condition: A is diagonalizable* ***if and only if*** *$\mu_A(x)$ is a product of distinct linear factors. But we don't have enough tools to prove this theorem yet – be patient!*

2.4 The Cayley–Hamilton theorem

In this section we'll prove one of the main big theorems of the course, which shows that the minimal polynomial of an $n \times n$ matrix has degree $\leq n$, and tells us a lot more besides:

Theorem 2.4.1 (Cayley–Hamiton)**.** *Let $c_A(x)$ be the characteristic polynomial of the $n \times n$ matrix A over an arbitrary field K. Then $c_A(A) = \mathbf{0}$.*

Proof. (Non-examinable) Let's agree to drop the various subscripts and bold zeroes – it'll be obvious from context when we mean a zero matrix, zero vector, zero linear map, etc.

Recall from MA106 that, if B is any $n \times n$ matrix, the "adjugate matrix" of B is another matrix $\mathrm{adj}(B)$ which was constructed along the way to constructing the inverse of B. The entries of $\mathrm{adj}(B)$ are the "cofactors" of B: the (i,j) entry of B is $(-1)^{i+j}c_{ji}$ (note the transposition of indices here!), where $c_{ji} = \det(B_{ji})$, B_{ij} being the $(n-1) \times (n-1)$ matrix obtained by deleting the j-th row and the i-th column of B. The key property of $\mathrm{adj}(B)$ is that it satisfies

$$B\,\mathrm{adj}(B) = \mathrm{adj}(B)B = (\det B)I_n.$$

(Notice that if B is invertible, this just says that $\mathrm{adj}(B) = (\det B)B^{-1}$, but the adjugate matrix still makes sense even if B is not invertible.)

Let's apply this to the matrix $B = A - xI_n$. By definition, $\det(B)$ is the characteristic polynomial $c_A(x)$, so

$$(A - xI_n)\,\mathrm{adj}(A - xI_n) = c_A(x)I_n. \tag{1}$$

The key to this proof is that the entries of the adjugate $\mathrm{adj}(B)$ are made up from determinants of $(n-1) \times (n-1)$ submatrices of B. Since the entries of B are linear or constant polynomials in x, the entries of $\mathrm{adj}(B)$ are polynomials of degree $\leq (n-1)$. We can make a new matrix by taking the jth coefficient of each entry of $\mathrm{adj}(B)$, and thus write

$$\mathrm{adj}(A - xI_n) = B_0 + B_1x + \cdots + B_{n-1}x^{n-1},$$

where each B_i is an $n \times n$ matrix with entries in K. We'll also give names to the coefficients of $c_A(x)$: let's say $c_A(x) = a_0 + a_1x + \cdots + a_nx^n$. So we can write equation (1) as

$$(A - xI_n)(B_0 + B_1x + \cdots + B_{n-1}x^{n-1}) = (a_0 + a_1x + \cdots + a_nx^n)I_n.$$

Since this is a polynomial identity, we can equate coefficients of the powers of x on the left and right hand sides. In the list of equations below, the equations on the left are the result of equating coefficients of x^i for $0 \leq i \leq n$, and those on right are obtained by multiplying A^i by the corresponding left hand equation.

$$\begin{aligned}
AB_0 &= a_0I_n && \Rightarrow && AB_0 &= a_0I_n,\\
AB_1 - B_0 &= a_1I_n, && \Rightarrow && A^2B_1 - AB_0 &= a_1A,\\
AB_2 - B_1 &= a_2I_n, && \Rightarrow && A^3B_2 - A^2B_1 &= a_2A^2,\\
&\vdots && \Rightarrow && &\vdots\\
AB_{n-1} - B_{n-2} &= a_{n-1}I_n && \Rightarrow && A^nB_{n-1} - A^{n-1}B_{n-2} &= a_{n-1}A^{n-1},\\
-B_{n-1} &= a_nI_n && \Rightarrow && -A^nB_{n-1} &= a_nA^n.
\end{aligned}$$

Now summing all of the equations in the right hand column gives

$$0 = a_0I_n + a_1A + \ldots + a_{n-1}A^{n-1} + a_nA^n$$

which says exactly that $c_A(A) = 0$. □

Corollary 2.4.2. *For any $A \in K^{n,n}$, we have $\mu_A \mid c_A$, and in particular* $\deg(\mu_A) \leq n$.

Example. Let D be the diagonal matrix $\begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}$ from the previous example. We saw above that $\mu_A(x) = (x-2)(x-3)$. However, it's easy to see that

$$c_A(x) = \begin{vmatrix} 3-x & 0 & 0 \\ 0 & 3-x & 0 \\ 0 & 0 & 2-x \end{vmatrix} = -(x-2)(x-3)^2.$$

How NOT to prove the Cayley–Hamilton theorem It is very tempting to try and prove the Cayley–Hamilton theorem as follows: we know that

$$c_A(x) = \det(A - xI_n),$$

so shouldn't we have

$$c_A(A) = \det(A - AI_n) = \det(A - A) = \det(0) = 0?$$

This is **wrong**: the argument is not valid. We know that $c_A(x) = \det(A - xI_n)$ holds for all $x \in K$, but it doesn't hold in general when we replace x by a matrix. Indeed it doesn't even make sense: if B is some other matrix, then $\det(A - BI_n) = \det(A - B)$ is an element of K, but $c_A(B)$ is a matrix, so it's meaningless to expect that

$$c_A(B) = \det(A - BI_n)$$

holds for all B. This bogus argument is a good way to remember what the Cayley–Hamilton theorem says, but it's not a proof.

2.5 Calculating the minimal polynomial

The Cayley–Hamilton theorem gives us a rather powerful tool for calculating the minimal polynomial of a matrix. To make it even easier, here's a little lemma:

Lemma 2.5.1. *Let λ be any eigenvalue of A. Then $\mu_A(\lambda) = 0$.*

Proof. Let $\underline{\mathbf{v}} \in K^{n,1}$ be an eigenvector corresponding to λ. Then $A^n\underline{\mathbf{v}} = \lambda^n\underline{\mathbf{v}}$, and hence for any polynomial $p \in K[x]$, we have

$$p(A)\underline{\mathbf{v}} = p(\lambda)\underline{\mathbf{v}}.$$

We know that $\mu_A(A)\underline{\mathbf{v}} = 0$, since $\mu_A(A)$ is the zero matrix. Hence $\mu_A(\lambda)\underline{\mathbf{v}} = 0$, and since $\underline{\mathbf{v}} \neq 0$ and $\mu_A(\lambda)$ is an element of K (not a matrix!), this can only happen if $\mu_A(\lambda) = 0$. □

This lemma, together with Cayley–Hamilton, give us very, very few possibilities for μ_A. Let's look at an example.

Example. Take $K = \mathbb{C}$ and let

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

This is rather large, but it has a fair few zeros, so you can calculate its characteristic polynomial fairly quickly by hand and find out that

$$c_A(x) = x^4 - 11x^3 + 45x^2 - 81x + 54.$$

Some trial and error shows that 2 is a root of this, and we find that

$$c_A(x) = (x-2)(x^3 - 9x^2 + 27x - 27) = (x-2)(x-3)^3.$$

So $\mu_A(x)$ divides $(x-2)(x-3)^3$. On the other hand, the eigenvalues of A are the roots of $c_A(x)$, namely $\{2, 3\}$; and we know that μ_A must have each of these as roots. So the only possibilities for $\mu_A(x)$ are:

$$\mu_A(x) \in \left\{ \begin{matrix} (x-2)(x-3), \\ (x-2)(x-3)^2, \\ (x-2)(x-3)^3. \end{matrix} \right\}.$$

Some slightly tedious calculation shows that $(A-2)(A-3)$ isn't zero, and nor is $(A-2)(A-3)^2$, and so it must be the case that $(x-2)(x-3)^3$ is the minimal polynomial of A.

Remark. *In practice, one rarely has to go so far as this – you have to be pretty unlucky to get a minimal polynomial with a repeated factor, and unluckier still to get a factor that comes in with multiplicity 3.*

2.6 Jordan chains and Jordan blocks

We'll now consider some special vectors attached to our matrix, which satisfy a condition a little like eigenvectors (but weaker). These will be the stepping-stones towards the Jordan canonical form.

Definition 2.6.1. A non-zero vector $\mathbf{v} \in K^{n,1}$ such that $(A - \lambda I_n)^i \mathbf{v} = 0$, for some $i > 0$, is called a *generalized eigenvector* of A with respect to the eigenvalue λ.

Note that, for fixed $i > 0$, $\{ \mathbf{v} \in V \mid (A - \lambda I_n)^i \mathbf{v} = 0 \}$ is the nullspace of $(A - \lambda I_n)^i$, and is called the *generalized eigenspace of index i* of A with respect to λ.

The generalized eigenspace of index 1 is just called the *eigenspace* of A w.r.t. λ; it consists of the eigenvectors of A w.r.t. λ, together with the zero vector. We sometimes also consider the *full generalized eigenspace* of A w.r.t. λ, which is the set of all generalized eigenvectors together with the zero vector; this is the union of the generalized eigenspaces of index i over all $i \in \mathbb{N}$.

We can arrange generalized eigenvectors into "chains":

Definition 2.6.2. A *Jordan chain of length k* is a sequence of non-zero vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k \in K^{n,1}$ that satisfies

$$A\mathbf{v}_1 = \lambda\mathbf{v}_1, \;\; A\mathbf{v}_i = \lambda\mathbf{v}_i + \mathbf{v}_{i-1}, \;\; 2 \leq i \leq k,$$

for some eigenvalue λ of A.

Equivalently, $(A - \lambda I_n)\mathbf{v}_1 = \mathbf{0}$ and $(A - \lambda I_n)\mathbf{v}_i = \mathbf{v}_{i-1}$ for $2 \leq i \leq k$, so $(A - \lambda I_n)^i \mathbf{v}_i = \mathbf{0}$ for $1 \leq i \leq k$. Thus all of the vectors in a Jordan chain are generalized eigenvectors, and $\mathbf{v}_i$ lies in the generalized eigenspace of index i.

For example, take $K = \mathbb{C}$ and consider the matrix

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

We see that, for $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ the standard basis of $\mathbb{C}^{3,1}$, we have $A\mathbf{b}_1 = 3\mathbf{b}_1$, $A\mathbf{b}_2 = 3\mathbf{b}_2 + \mathbf{b}_1$, $A\mathbf{b}_3 = 3\mathbf{b}_3 + \mathbf{b}_2$, so $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ is a Jordan chain of length 3 for the eigenvalue 3 of A. The generalized eigenspaces of index 1, 2, and 3 are respectively $\langle \mathbf{b}_1 \rangle$, $\langle \mathbf{b}_1, \mathbf{b}_2 \rangle$, and $\langle \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3 \rangle$.

Note that this isn't the only possible Jordan chain. Obviously, $\{17\mathbf{b}_1, 17\mathbf{b}_2, 17\mathbf{b}_3\}$ would be a Jordan chain; but there are more devious possibilities – you can check that $\{\mathbf{b}_1, \mathbf{b}_1 + \mathbf{b}_2, \mathbf{b}_2 + \mathbf{b}_3\}$ is a Jordan chain, so there can be several Jordan chains with the same first vector. On the other hand, two Jordan chains with the same *last* vector are the same and in particular have the same length.

What are the generalized eigenspaces here? The only eigenvalue is 3. For this eigenvalue, the generalized eigenspace of index 1 is $\langle \mathbf{b}_1 \rangle$ (the linear span of $\mathbf{b}_1$); the generalized eigenspace of index 2 is $\langle \mathbf{b}_1, \mathbf{b}_2 \rangle$; and the generalized eigenspace of index 3 is the whole space $\langle \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3 \rangle$. So the dimensions are $(1, 2, 3)$.

In general, you can think of the dimensions of the generalized eigenspaces of a matrix as a sort of "fingerprint". (Since we have these numbers for each λ, perhaps one should think of the eigenvalues as the "fingers"). We'll see later that these dimensions are very important in the theory of Jordan canonical form. For now, we'll stick to proving one proposition about them:

Proposition 2.6.3. *The dimensions of corresponding generalized eigenspaces of similar matrices are the same.*

Proof. Notice that the dimension of a generalized eigenspace of A is the nullity of $(T - \lambda I_V)^i$, which depends only on the linear map T associated with A. Therefore it's independent of the choice of basis. Since similar matrices represent the same linear map in different bases, the proposition follows. □

In the example we had earlier, the standard basis of $K^{n,1}$ was a Jordan chain, and this means that the matrix A had a rather special form. We'll give a name to matrices of this type:

Definition 2.6.4. We define the *Jordan block* of degree k with eigenvalue λ to be the $k \times k$ matrix $J_{\lambda,k}$ whose entries are given by

$$\gamma_{ij} = \begin{cases} \lambda & \text{if } j = i \\ 1 & \text{if } j = i+1 \\ 0 & \text{otherwise.} \end{cases}$$

So, for example,

$$J_{1,2} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad J_{\lambda,3} = \begin{pmatrix} \frac{3-i}{2} & 1 & 0 \\ 0 & \frac{3-i}{2} & 1 \\ 0 & 0 & \frac{3-i}{2} \end{pmatrix}, \quad \text{and} \quad J_{0,4} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

are Jordan blocks, where $\lambda = \frac{3-i}{2}$ in the second example.

It should be clear that the matrix of T with respect to the basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of K^n is a Jordan block of degree n if and only if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a Jordan chain for A.

Note that the minimal polynomial of $J_{\lambda,k}$ is equal to $(x-\lambda)^k$, and the characteristic polynomial is $(\lambda - x)^k$.

Warning. Some authors put the 1's below rather than above the main diagonal in a Jordan block. This corresponds to writing the Jordan chain in reverse order.

2.7 Jordan bases and the Jordan canonical form

Definition 2.7.1. A *Jordan basis* for A is a basis of K^n consisting of one or more Jordan chains strung together.

Such a basis will look like

$$w_{11}, \ldots, w_{1k_1}, w_{21}, \ldots, w_{2k_2}, \ldots, w_{s1}, \ldots, w_{sk_s},$$

where, for $1 \leq i \leq s$, $w_{i1}, \ldots, w_{ik_i}$ is a Jordan chain (for some eigenvalue λ_i).

We denote the $m \times n$ matrix in which all entries are 0 by $\mathbf{0}_{m,n}$. If A is an $m \times m$ matrix and B an $n \times n$ matrix, then we denote the $(m+n) \times (m+n)$ matrix with block form

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

by $A \oplus B$, the *direct sum* of A and B. For example

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

It's clear that the matrix of T with respect to a Jordan basis is the direct sum $J_{\lambda_1,k_1} \oplus J_{\lambda_2,k_2} \oplus \cdots \oplus J_{\lambda_s,k_s}$ of the corresponding Jordan blocks.

We can now state the main theorem of this section, which says that if K is the complex numbers $\mathbb{C}$, then Jordan bases exist.

Theorem 2.7.2. *Let A be an $n \times n$ matrix over $\mathbb{C}$. Then there exists a Jordan basis for A, and hence A is similar to a matrix J which is a direct sum of Jordan blocks. The Jordan blocks occurring in J are uniquely determined by A.*

The matrix J in the theorem is said to be the *Jordan canonical form* (JCF) or sometimes *Jordan normal form* of A. It is uniquely determined by A up to the order of the blocks.

Remark. *The only reason we need $K = \mathbb{C}$ in this theorem is to ensure that A has at least one eigenvalue. If $K = \mathbb{R}$ (or $\mathbb{Q}$), we'd run into trouble with $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$; this matrix has* no *eigenvalues, since $c_A(x) = x^2 + 1$ has no roots in K. So it certainly has no Jordan chains.*

We will prove the theorem later, in Section 2.11. First we derive some consequences and study methods for calculating the JCF of a matrix. Note that, by Theorem 1.5.2, if P is the matrix having the Jordan basis as columns, then $P^{-1}AP = J$.

Theorem 2.7.3 (Consequences of the JCF)**.** *Let $A \in \mathbb{C}^{n,n}$, and $\{\lambda_1, \ldots, \lambda_r\}$ be the set of eigenvalues of A.*

(i) The characteristic polynomial of A is

$$(-1)^n \prod_{i=1}^{r} (x - \lambda_i)^{a_i},$$

where a_i is the sum of the degrees of the Jordan blocks of A of eigenvalue λ_i.

(ii) The minimal polynomial of A is

$$\prod_{i=1}^{r} (x - \lambda_i)^{b_i},$$

where b_i is the largest among the degrees of the Jordan blocks of A of eigenvalue λ_i.

(iii) A is diagonalizable if and only if $\mu_A(x)$ has no repeated factors.

Proof. This follows from what we know about the minimal and characteristic polynomials of Jordan blocks, together with the fact that if $C = A \oplus B$, then $c_C(x)$ is the product of $c_A(x)$ and $c_B(x)$, and $\mu_C(x)$ is the LCM of $\mu_A(x)$ and $\mu_B(x)$. For the last part, notice that if A is diagonalizable, the JCF of A is just the diagonal form of A; since the JCF is unique, it follows that A is diagonalizable if and only if every Jordan block for A has size 1, so all of the numbers b_i are 1. □

Week 3

2.8 The JCF when n=2 and 3

When $n = 2$ and $n = 3$, the JCF can be deduced just from the minimal and characteristic polynomials. Let us consider these cases.

When $n = 2$, we have either two distinct eigenvalues λ_1, λ_2, or a single repeated eigenvalue λ_1. If the eigenvalues are distinct, then by Corollary 2.2.3 A is diagonalizable and the JCF is the diagonal matrix $J_{\lambda_1,1} \oplus J_{\lambda_2,1}$.

Example 1. $A = \begin{pmatrix} 1 & 4 \\ 1 & 1 \end{pmatrix}$. We calculate $c_A(x) = x^2 - 2x - 3 = (x-3)(x+1)$, so there are two distinct eigenvalues, 3 and -1. Associated eigenvectors are $\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} -2 \\ 1 \end{pmatrix}$, so we put $P = \begin{pmatrix} 2 & -2 \\ 1 & 1 \end{pmatrix}$ and then $P^{-1}AP = \begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix}$.

If the eigenvalues are equal, then there are two possible JCFs, $J_{\lambda_1,1} \oplus J_{\lambda_1,1}$, which is a scalar matrix, and $J_{\lambda_1,2}$. The minimal polynomial is respectively $(x - \lambda_1)$ and $(x - \lambda_1)^2$ in these two cases. In fact, these cases can be distinguished without any calculation whatsoever, because in the first case A is a scalar multiple of the identity, and in particular A is already in JCF.

In the second case, a Jordan basis consists of a single Jordan chain of length 2. To find such a chain, let $\mathbf{v}_2$ be any vector for which $(A - \lambda_1 I_2)\mathbf{v}_2 \neq \mathbf{0}$ and let $\mathbf{v}_1 = (A - \lambda_1 I_2)\mathbf{v}_2$. (Note that, in practice, it is often easier to find the vectors in a Jordan chain in reverse order.)

Example 2. $A = \begin{pmatrix} 1 & 4 \\ -1 & -3 \end{pmatrix}$. We have $c_A(x) = x^2 + 2x + 1 = (x+1)^2$, so there is a single eigenvalue -1 with multiplicity 2. Since the first column of $A + I_2$ is non-zero, we can choose $\mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\mathbf{v}_1 = (A + I_2)\mathbf{v}_2 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$, so $P = \begin{pmatrix} 2 & 1 \\ -1 & 0 \end{pmatrix}$ and $P^{-1}AP = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}$.

Now let $n = 3$. If there are three distinct eigenvalues, then A is diagonalizable.

Suppose that there are two distinct eigenvalues, so one has multiplicity 2, and the other has multiplicity 1. Let the eigenvalues be $\lambda_1, \lambda_1, \lambda_2$, with $\lambda_1 \neq \lambda_2$. Then there are two possible JCFs for A, $J_{\lambda_1,1} \oplus J_{\lambda_1,1} \oplus J_{\lambda_2,1}$ and $J_{\lambda_1,2} \oplus J_{\lambda_2,1}$, and the minimal polynomial is $(x - \lambda_1)(x - \lambda_2)$ in the first case and $(x - \lambda_1)^2(x - \lambda_2)$ in the second.

In the first case, a Jordan basis is a union of three Jordan chains of length 1, each of which consists of an eigenvector of A.

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

$$c_A(x) = (2-x)[(5-x)(-2-x) + 12] = (2-x)(x^2 - 3x + 2) = (2-x)^2(1-x).$$

We know from the theory above that the minimal polynomial must be $(x-2)(x-1)$ or $(x-2)^2(x-1)$. We can decide which simply by calculating $(A-2I_3)(A-I_3)$ to test whether or not it is 0. We have

$$A-2I_3=\begin{pmatrix}0&0&0\\1&3&2\\-2&-6&-4\end{pmatrix},\quad A-I_3=\begin{pmatrix}1&0&0\\1&4&2\\-2&-6&-3\end{pmatrix},$$

and the product of these two matrices is 0, so $\mu_A=(x-2)(x-1)$.

The eigenvectors $\mathbf{v}$ for $\lambda_1=2$ satisfy $(A-2I_3)\mathbf{v}=\mathbf{0}$, and we must find two linearly independent solutions; for example we can take $\mathbf{v}_1=\begin{pmatrix}0\\2\\-3\end{pmatrix}$, $\mathbf{v}_2=\begin{pmatrix}1\\-1\\1\end{pmatrix}$. An eigenvector for the eigenvalue 1 is $\mathbf{v}_3=\begin{pmatrix}0\\1\\-2\end{pmatrix}$, so we can choose

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

and then $P^{-1}AP$ is diagonal with entries $2,2,1$.

In the second case, there are two Jordan chains, one for λ_1 of length 2, and one for λ_2 of length 1. For the first chain, we need to find a vector $\mathbf{v}_2$ with $(A-\lambda_1 I_3)^2\mathbf{v}_2=\mathbf{0}$ but $(A-\lambda_1 I_3)\mathbf{v}_2\neq\mathbf{0}$, and then the chain is $\mathbf{v}_1=(A-\lambda_1 I_3)\mathbf{v}_2,\mathbf{v}_2$. For the second chain, we simply need an eigenvector for λ_2.

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

$$c_A(x)=(3-x)[(3-x)(-1-x)+4]-2+(3-x)=-x^3+5x^2-8x+4=(2-x)^2(1-x),$$

as in Example 3. We have

$$A-2I_3=\begin{pmatrix}1&2&1\\0&1&1\\-1&-4&-3\end{pmatrix},\quad (A-2I_3)^2=\begin{pmatrix}0&0&0\\-1&-3&-2\\2&6&4\end{pmatrix},\quad (A-I_3)=\begin{pmatrix}2&2&1\\0&2&1\\-1&-4&-2\end{pmatrix}.$$

and we can check that $(A-2I_3)(A-I_3)$ is non-zero, so we must have $\mu_A=(x-2)^2(x-1)$.

For the Jordan chain of length 2, we need a vector with $(A-2I_3)^2\mathbf{v}_2=\mathbf{0}$ but $(A-2I_3)\mathbf{v}_2\neq\mathbf{0}$, and we can choose $\mathbf{v}_2=\begin{pmatrix}2\\0\\-1\end{pmatrix}$. Then $\mathbf{v}_1=(A-2I_3)\mathbf{v}_2=\begin{pmatrix}1\\-1\\1\end{pmatrix}$. An eigenvector for the

eigenvalue 1 is $\mathbf{v}_3 = \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix}$, so we can choose

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

and then

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

Finally, suppose that there is a single eigenvalue, λ_1, so $c_A = (\lambda_1 - x)^3$. There are three possible JCFs for A, $J_{\lambda_1,1} \oplus J_{\lambda_1,1} \oplus J_{\lambda_1,1}$, $J_{\lambda_1,2} \oplus J_{\lambda_1,1}$, and $J_{\lambda_1,3}$, and the minimal polynomials in the three cases are $(x - \lambda_1)$, $(x - \lambda_1)^2$, and $(x - \lambda_1)^3$, respectively.

In the first case, J is a scalar matrix, and $A = PJP^{-1} = J$, so this is recognisable immediately.

In the second case, there are two Jordan chains, one of length 2 and one of length 1. For the first, we choose $\mathbf{v}_2$ with $(A - \lambda_1 I_3)\mathbf{v}_2 \neq 0$, and let $\mathbf{v}_1 = (A - \lambda_1 I_3)\mathbf{v}_2$. (This case is easier than the case illustrated in Example 4, because we have $(A - \lambda_1 I_3)^2\mathbf{v} = \mathbf{0}$ for all $\mathbf{v} \in \mathbb{C}^{3,1}$.) For the second Jordan chain, we choose $\mathbf{v}_3$ to be an eigenvector for λ_1 such that $\mathbf{v}_2$ and $\mathbf{v}_3$ are linearly independent.

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

$$c_A(x) = -x[(3+x)x + 2] - 2(x+1) - 2 + (3+x) = -x^3 - 3x^2 - 3x - 1 = -(1+x)^3.$$

We have

$$A + I_3 = \begin{pmatrix} 1 & 2 & 1 \\ -1 & -2 & -1 \\ 1 & 2 & 1 \end{pmatrix},$$

and we can check that $(A + I_3)^2 = 0$. The first column of $A + I_3$ is non-zero, so $(A + I_3)\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \neq \mathbf{0}$, and we can choose $\mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ and $\mathbf{v}_1 = (A + I_3)\mathbf{v}_2 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$. For $\mathbf{v}_3$ we need to choose a vector which is not a multiple of $\mathbf{v}_1$ such that $(A + I_3)\mathbf{v}_3 = \mathbf{0}$, and we can choose $\mathbf{v}_3 = \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix}$.

So we have

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

and then

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

In the third case, there is a single Jordan chain, and we choose $\mathbf{v}_3$ such that $(A - \lambda_1 I_3)^2 \mathbf{v}_3 \neq 0$, $\mathbf{v}_2 = (A - \lambda_1 I_3)\mathbf{v}_3$, $\mathbf{v}_1 = (A - \lambda_1 I_3)^2 \mathbf{v}_3$.

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

$$c_A(x) = -x[(2+x)(1+x)] - (2+x) + 1 = -(1+x)^3.$$

We have

$$A + I_3 = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \\ 1 & 0 & -1 \end{pmatrix}, \quad (A + I_3)^2 = \begin{pmatrix} 0 & 1 & 1 \\ 0 & -1 & -1 \\ 0 & 1 & 1 \end{pmatrix},$$

so $(A + I_3)^2 \neq 0$ and $\mu_A = (x+1)^3$. For $\mathbf{v}_3$, we need a vector that is not in the nullspace of $(A + I_3)^2$. Since the second column, which is the image of $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ is non-zero, we can choose $\mathbf{v}_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, and then $\mathbf{v}_2 = (A + I_3)\mathbf{v}_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ and $\mathbf{v}_1 = (A + I_3)\mathbf{v}_2 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$. So we have

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

and then

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

2.9 The general case

In the examples above, we could tell what the sizes of the Jordan blocks were for each eigenvalue from the dimensions of the eigenspaces, since the dimension of the eigenspace for each eigenvalue λ is the number of blocks for that eigenvalue. This doesn't work for $n = 4$: for instance, the matrices

$$A_1 = J_{\lambda,2} \oplus J_{\lambda,2}$$

and

$$A_2 = J_{\lambda,3} \oplus J_{\lambda,1}$$

both have only one eigenvalue (λ) with the eigenspace being of dimension 2.

(Knowing the minimal polynomial helps, but it's a bit of a pain to calculate – generally the easiest way to find the minimal polynomial is to calculate the JCF first! Worse still, it still doesn't uniquely determine the JCF in large dimensions, since

$$A_3 = J_{\lambda,3} \oplus J_{\lambda,3} \oplus J_{\lambda,1}$$

and

$$A_4 = J_{\lambda,3} \oplus J_{\lambda,2} \oplus J_{\lambda,2}$$

have the same minimal polynomial, the same characteristic polynomial, and the same number of blocks.)

In general, we can compute the JCF from the dimensions of the generalized eigenspaces (what I called the "fingerprint" of the matrix). Notice that the matrices A_1 and A_2 have different fingerprints: the generalized eigenspace for λ of index 2 has dimension 4 for A_1 (it's the whole space) but dimension only 3 for A_2.

Theorem 2.9.1. *Let λ be an eigenvalue of a matrix $A \in \mathbb{C}^{n,n}$, and let J be the JCF of A. Then*

(i) The number of Jordan blocks of J with eigenvalue λ is equal to $\text{nullity}(A - \lambda I_n)$.

(ii) More generally, for $i > 0$, the number of Jordan blocks of J with eigenvalue λ and degree at least i is equal to $\text{nullity}((A - \lambda I_n)^i) - \text{nullity}((A - \lambda I_n)^{i-1})$.

Note that this proves the uniqueness part of Theorem 2.7.2: the theorem says that the block sizes of the Jordan form of A are determined by the fingerprint of A, so any two Jordan canonical forms for A must have the same blocks (possibly ordered differently).

Proof. By Proposition 2.6.3, the corresponding generalized eigenspaces of A and J have the same dimensions, so we may assume WLOG that $A = J$. So A is a direct sum of several Jordan blocks $J_{\lambda_1,k_1} \oplus \cdots \oplus J_{\lambda_s,k_s}$.

However, it's easy to see that the dimension of the generalized λ-eigenspace of index i of a direct sum $A \oplus B$ is the sum of the dimensions of the generalized λ eigenspaces of index i of A and of B. Hence it suffices to prove the theorem for a single Jordan block $J_{\lambda,k}$.

But we know that $(J_{\lambda,k} - \lambda I_k)^i$ has a single diagonal line of zeroes i places above the diagonal, for $i < k$, and is 0 for $i \geq k$. Hence the dimension of its kernel is i for $0 \leq i \leq k$ and k for $i \geq k$. This clearly implies the theorem when A is a single Jordan block, and hence for any A. □

2.10 Examples

Example 7. $A = \begin{pmatrix} -1 & -3 & -1 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 3 & 1 & -1 \end{pmatrix}$. Then $c_A(x) = (-1-x)^2(2-x)^2$, so there are two eigenvalues $-1, 2$, both with multiplicity 2. There are four possibilities for the JCF (one or two

blocks for each of the two eigenvalues). We could determine the JCF by computing the minimal polynomial μ_A but it is probably easier to compute the nullities of the eigenspaces and use Theorem 2.9.1. We have

$$A + I_4 = \begin{pmatrix} 0 & -3 & -1 & 0 \\ 0 & 3 & 1 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 3 & 1 & 0 \end{pmatrix}, \quad (A - 2I_4) = \begin{pmatrix} -3 & -3 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 3 & 1 & -3 \end{pmatrix},$$

$$(A - 2I_4)^2 = \begin{pmatrix} 9 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -9 & 0 & 9 \end{pmatrix}.$$

The rank of $A + I_4$ is clearly 2, so its nullity is also 2, and hence there are two Jordan blocks with eigenvalue -1. The three non-zero rows of $(A - 2I_4)$ are linearly independent, so its rank is 3, hence its nullity 1, so there is just one Jordan block with eigenvalue 2, and the JCF of A is $J_{-1,1} \oplus J_{-1,1} \oplus J_{2,2}$.

For the two Jordan chains of length 1 for eigenvalue -1, we just need two linearly independent eigenvectors, and the obvious choice is $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$, $\mathbf{v}_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$. For the Jordan chain $\mathbf{v}_3, \mathbf{v}_4$ for eigenvalue 2, we need to choose $\mathbf{v}_4$ in the nullspace of $(A - 2I_4)^2$ but not in the nullspace of $A - 2I_4$. (This is why we calculated $(A - 2I_4)^2$.) An obvious choice here is $\mathbf{v}_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$, and then $\mathbf{v}_3 = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 1 \end{pmatrix}$, and to transform A to JCF, we put

$$P = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}, \quad P^{-1} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad P^{-1}AP = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

Example 8. $A = \begin{pmatrix} -2 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & -2 & 0 \\ 1 & 0 & -2 & -2 \end{pmatrix}$. Then $c_A(x) = (-2 - x)^4$, so there is a single eigenvalue -2 with multiplicity 4. We find $(A + 2I_4) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & -2 & 0 \end{pmatrix}$, and $(A + 2I_4)^2 = 0$, so

$\mu_A = (x+2)^2$, and the JCF of A could be $J_{-2,2} \oplus J_{-2,2}$ or $J_{-2,2} \oplus J_{-2,1} \oplus J_{-2,1}$.

To decide which case holds, we calculate the nullity of $A + 2I_4$ which, by Theorem 2.9.1, is equal to the number of Jordan blocks with eigenvalue -2. Since $A + 2I_4$ has just two non-zero rows, which are distinct, its rank is clearly 2, so its nullity is $4 - 2 = 2$, and hence the JCF of A is $J_{-2,2} \oplus J_{-2,2}$.

A Jordan basis consists of a union of two Jordan chains, which we will call $\mathbf{v}_1, \mathbf{v}_2$, and $\mathbf{v}_3, \mathbf{v}_4$, where $\mathbf{v}_1$ and $\mathbf{v}_3$ are eigenvectors and $\mathbf{v}_2$ and $\mathbf{v}_4$ are generalized eigenvectors of index 2. To find such chains, it is probably easiest to find $\mathbf{v}_2$ and $\mathbf{v}_4$ first and then to calculate $\mathbf{v}_1 = (A + 2I_4)\mathbf{v}_2$ and $\mathbf{v}_3 = (A + 2I_4)\mathbf{v}_4$.

Although it is not hard to find $\mathbf{v}_2$ and $\mathbf{v}_4$ in practice, we have to be careful, because they need to be chosen so that no linear combination of them lies in the nullspace of $(A + 2I_4)$. In fact, since this nullspace is spanned by the second and fourth standard basis vectors, the obvious choice is $\mathbf{v}_2 = \begin{pmatrix}1\\0\\0\\0\end{pmatrix}$, $\mathbf{v}_4 = \begin{pmatrix}0\\0\\1\\0\end{pmatrix}$, and then $\mathbf{v}_1 = (A + 2I_4)\mathbf{v}_2 = \begin{pmatrix}0\\0\\0\\1\end{pmatrix}$, $\mathbf{v}_3 = (A + 2I_4)\mathbf{v}_4 = \begin{pmatrix}0\\1\\0\\-2\end{pmatrix}$, so to transform A to JCF, we put

$$P = \begin{pmatrix}0&1&0&0\\0&0&1&0\\0&0&0&1\\1&0&-2&0\end{pmatrix}, \quad P^{-1} = \begin{pmatrix}0&2&0&1\\1&0&0&0\\0&1&0&0\\0&0&1&0\end{pmatrix}, \quad P^{-1}AP = \begin{pmatrix}-2&1&0&0\\0&-2&0&0\\0&0&-2&1\\0&0&0&-2\end{pmatrix}.$$

2.11 Proof of Theorem 2.7.2

This proof is too hard.

Well, the right work is tedious. We agree **not to examine it**, unless it is an open book exam.

We proceed by induction on $n = \dim(V)$. The case $n = 1$ is clear.

Let λ be an eigenvalue of T and let $U = \mathrm{im}(T - \lambda I_V)$ and $m = \dim(U)$. Then $m = \mathrm{rank}(T - \lambda I_V) = n - \mathrm{nullity}(T - \lambda I_V) < n$, because the eigenvectors for λ lie in the nullspace of $T - \lambda I_V$. For $\mathbf{u} \in U$, we have $\mathbf{u} = (T - \lambda I_V)(\mathbf{v})$ for some $\mathbf{v} \in V$, and hence $T(\mathbf{u}) = T(T - \lambda I_V)(\mathbf{v}) = (T - \lambda I_V)T(\mathbf{v}) \in U$. So T restricts to $T_U : U \to U$, and we can apply our inductive hypothesis to T_U to deduce that U has a basis $\mathbf{e}_1, \ldots, \mathbf{e}_m$, which is a disjoint union of Jordan chains for T_U.

We now show how to extend the Jordan basis of U to one of V. We do this in two stages. For the first stage, suppose that l of the Jordan chains of T_U are for the eigenvalue λ (possibly $l = 0$). For each such chain $\mathbf{v}_1, \ldots, \mathbf{v}_k$ with $T(\mathbf{v}_1) = \lambda\mathbf{v}_1$, $T(\mathbf{v}_i) = \lambda\mathbf{v}_i + \mathbf{v}_{i-1}$, $2 \leq i \leq k$, since $\mathbf{v}_k \in U = \mathrm{im}(T - \lambda I_V)$, we can find $\mathbf{v}_{k+1} \in V$ with $T(\mathbf{v}_{k+1}) = \lambda\mathbf{v}_{k+1} + \mathbf{v}_k$, thereby extending the chain by an extra vector. So far we have adjoined l new vectors to the basis, by extending the length l Jordan chains by 1. Let us call these new vectors $\mathbf{w}_1, \ldots, \mathbf{w}_l$.

For the second stage, observe that the first vector in each of the l chains lies in the eigenspace of T_U for λ. We know that the dimension of the eigenspace of T for λ is the nullspace of $(T - \lambda I_V)$, which has dimension $n - m$. So we can adjoin $(n - m) - l$ further eigenvectors of T to the l that we have already to complete a basis of the nullspace of $(T - \lambda I_V)$. Let us call these $(n - m) - l$ new vectors $\mathbf{w}_{l+1}, \ldots, \mathbf{w}_{n-m}$. They are adjoined to our basis of V in the second stage. They each form a Jordan chain of length 1, so we now have a collection of n vectors which form a disjoint union of Jordan chains.

To complete the proof, we need to show that these n vectors form a basis of V, for which is it is enough to show that they are linearly independent.

Partly because of notational difficulties, we provide only a sketch proof of this, and leave the details to the student. Suppose that $\alpha_1 \mathbf{w}_1 + \cdots + \alpha_{n-m} \mathbf{w}_{n-m} + \mathbf{x} = \mathbf{0}$, where $\mathbf{x}$ is a linear combination of the basis vectors of U. Applying $T - \lambda I_n$ gives

$$\alpha_1 (T - \lambda I_n)(\mathbf{w}_1) + \cdots + \alpha_l (T - \lambda I_n)(\mathbf{w}_l) + (T - \lambda I_n)(\mathbf{x}).$$

Each of $\alpha_i (T - \lambda I_n)(\mathbf{w}_i)$ for $1 \leq i \leq l$ is the last member of one of the l Jordan chains for T_U. When we apply $(T - \lambda I_n)$ to one of the basis vectors of U, we get a linear combination of the basis vectors of U other than $\alpha_i (T - \lambda I_n)(\mathbf{w}_i)$ for $1 \leq i \leq l$. Hence, by the linear independence of the basis of U, we deduce that $\alpha_i = 0$ for $1 \leq i \leq l$ This implies that $(T - \lambda I_n)(\mathbf{x}) = \mathbf{0}$, so $\mathbf{x}$ is in the eigenspace of T_U for the eigenvalue λ. But, by construction, $\mathbf{w}_{l+1}, \ldots, \mathbf{w}_{n-m}$ extend a basis of this eigenspace of T_U to that the eigenspace of V, so we also get $\alpha_i = 0$ for $l + 1 \leq i \leq n - m$, which completes the proof.

3 Functions of matrices

In this section, we're going to use what we know about Jordan normal forms of matrices to do analysis with matrices.

3.1 Powers of matrices

The theory we developed can be used to compute powers of matrices efficiently. Suppose we need to compute A^{2020} where $A = \begin{pmatrix} -2 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & -2 & 0 \\ 1 & 0 & -2 & -2 \end{pmatrix}$ is the matrix from Example 8 in 2.10.

There are two practical ways of computing M^n by hand for a general matrix and a very large n. The first one involves Jordan forms. If $J = P^{-1}MP$ is the JCF of M then it is sufficient to compute J^n because of the telescoping product:

$$M^n = (PJP^{-1})^n = PJP^{-1}PJP^{-1}P\dots JP^{-1} = PJ^nP^{-1}.$$

$$\text{If } J = J_{k_1,\lambda_1} \oplus \dots \oplus J_{k_t,\lambda_t} \text{ then } J^n = J^n_{k_1,\lambda_1} \oplus \dots \oplus J^n_{k_t,\lambda_t}\,. \tag{2}$$

Finally, the power of an individual Jordan block can be computed as

$$J^n_{k,\lambda} = \begin{pmatrix} \lambda^n & n\lambda^{n-1} & \dots & \binom{n}{k-2}\lambda^{n-k+2} & \binom{n}{k-1}\lambda^{n-k+1} \\ 0 & \lambda^n & \dots & \binom{n}{k-3}\lambda^{n-k+3} & \binom{n}{k-2}\lambda^{n-k+2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda^n & n\lambda^{n-1} \\ 0 & 0 & \dots & 0 & \lambda^n \end{pmatrix} \tag{3}$$

where $\binom{n}{t} = \frac{n!}{(n-t)!t!}$ is the choose-function (or binomial coefficient), interpreted as $\binom{n}{t} = 0$ whenever $t > n$.

Let us apply it to the matrix A above:

$$A^n = PJ^nP^{-1} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & -2 & 0 \end{pmatrix} \begin{pmatrix} -2 & 1 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & -2 \end{pmatrix}^n \begin{pmatrix} 0 & 2 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} =$$

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & -2 & 0 \end{pmatrix} \begin{pmatrix} (-2)^n & n(-2)^{n-1} & 0 & 0 \\ 0 & (-2)^n & 0 & 0 \\ 0 & 0 & (-2)^n & n(-2)^{n-1} \\ 0 & 0 & 0 & (-2)^n \end{pmatrix} \begin{pmatrix} 0 & 2 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} =$$

$$\begin{pmatrix} (-2)^n & 0 & 0 & 0 \\ 0 & (-2)^n & n(-2)^{n-1} & 0 \\ 0 & 0 & (-2)^n & 0 \\ n(-2)^{n-1} & 0 & n(-2)^n & (-2)^n \end{pmatrix}.$$

The second method of computing M^n uses Lagrange's interpolation polynomial. It is less labour intensive and more suitable for pen-and-paper calculations. Suppose $\psi(M) = 0$ for a polynomial $\psi(z)$, in practice, $\psi(z)$ is either the minimal, or the characteristic polynomial. Dividing with a remainder $z^n = q(z)\psi(z) + h(z)$, we conclude that

$$M^n = q(M)\psi(M) + h(M) = h(M).$$

Division with a remainder may appear problematic[2] for large n but there is a shortcut. If we know the roots of $\psi(z)$, say $\alpha_1, \ldots, \alpha_k$ with their multiplicities $m_1, \ldots, m_k$, then $h(z)$ can be found by solving the system of simultaneous equations in coefficients of $h(z)$:

$$f^{(t)}(\alpha_j) = h^{(t)}(\alpha_j),\ 1 \le j \le k,\ 0 \le t < m_j$$

where $f(z) = z^n$ and $f^{(t)} = f^{(t-1)\prime}$ is the t-th derivative. In other words, $h(z)$ is Lagrange's interpolation polynomial for the function z^n at the roots of $\psi(z)$.

We know the minimal polynomial $\mu_A(z) = (z+2)^2$ for the matrix A above. Suppose the Lagrange interpolation of z^n at the roots of $(z+2)^2$ is $h(z) = \alpha z + \beta$. The condition on the coefficients is given by

$$\begin{cases} (-2)^n & = & h(-2) & = & -2\alpha + \beta \\ n(-2)^{n-1} & = & h'(-2) & = & \alpha \end{cases}$$

Solving them gives $\alpha = n(-2)^{n-1}$ and $\beta = (1-n)(-2)^n$. It follows that

$$A^n = n(-2)^{n-1}A + (1-n)(-2)^n I = \begin{pmatrix} (-2)^n & 0 & 0 & 0 \\ 0 & (-2)^n & n(-2)^{n-1} & 0 \\ 0 & 0 & (-2)^n & 0 \\ n(-2)^{n-1} & 0 & n(-2)^n & (-2)^n \end{pmatrix}.$$

3.2 Applications to difference equations

Week 4

Let us consider *an initial value problem* for an *autonomous* system with discrete time:

$$\mathbf{x}(n+1) = A\mathbf{x}(n),\ n \in \mathbb{N},\ \mathbf{x}(0) = w.$$

Here $\mathbf{x}(n) \in K^m$ is a sequence of vectors in a vector space over a field K. One thinks of $\mathbf{x}(n)$ as a state of the system at time n. The initial state is $\mathbf{x}(0) = w$. The $n \times n$-matrix A with coefficients in K describes the evolution of the system. The adjective *autonomous* means that the evolution equation does not change with the time[3].

It takes longer to formulate this problem then to solve it. The solution is straightforward:

$$\mathbf{x}(n) = A\mathbf{x}(n-1) = A^2\mathbf{x}(n-2) = \ldots = A^n\mathbf{x}(0) = A^n w. \tag{4}$$

[2] Try to divide z^{2020} by $z^2 + z + 1$ without reading any further.

[3] A nonautonomous system would be described by $\mathbf{x}(n+1) = A(n)\mathbf{x}(n)$ here.

As a working example, let us consider Fibonacci and Lucas numbers:

$$F_0 = 0,\ F_1 = 1 \text{ and } F_n = F_{n-1} + F_{n-2}\ (n \geq 2),$$

$$L_0 = 2,\ L_1 = 1 \text{ and } L_n = L_{n-1} + L_{n-2}\ (n \geq 2).$$

The recursion relations for them turn into

$$\begin{pmatrix} F_n \\ F_{n+1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} F_{n-1} \\ F_n \end{pmatrix} \text{ and } \begin{pmatrix} L_n \\ L_{n+1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} L_{n-1} \\ L_n \end{pmatrix}$$

so that (4) immediately yields a general solution

$$\begin{pmatrix} F_n \\ F_{n+1} \end{pmatrix} = A^n \begin{pmatrix} 0 \\ 1 \end{pmatrix},\ \begin{pmatrix} L_n \\ L_{n+1} \end{pmatrix} = A^n \begin{pmatrix} 2 \\ 1 \end{pmatrix} \text{ where } A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}. \tag{5}$$

We compute the characteristic polynomial of A to be $c_A(z) = z^2 - z - 1$. Its discriminant is 5. The roots of $c_A(z)$ are the golden ratio $\lambda = (1+\sqrt{5})/2$ and $1-\lambda = (1-\sqrt{5})/2$. It is useful to observe that

$$2\lambda - 1 = \sqrt{5} \quad \text{and} \quad \lambda(1-\lambda) = -1.$$

Let us introduce the number $\mu_n = \lambda^n - (1-\lambda)^n$. Suppose the Lagrange interpolation of z^n at the roots of $z^2 - z - 1$ is $h(z) = \alpha z + \beta$. The condition on the coefficients is given by

$$\begin{cases} \lambda^n & = & h(\lambda) & = & \alpha\lambda + \beta \\ (1-\lambda)^n & = & h(1-\lambda) & = & \alpha(1-\lambda) + \beta \end{cases}$$

Solving them gives

$$\alpha = \mu_n/\sqrt{5} \text{ and } \beta = \mu_{n-1}/\sqrt{5}\,.$$

It follows that

$$A^n = \alpha A + \beta = \mu_n/\sqrt{5}A + \mu_{n-1}/\sqrt{5}I_2 = \begin{pmatrix} \mu_{n-1}/\sqrt{5} & \mu_n/\sqrt{5} \\ \mu_n/\sqrt{5} & (\mu_n + \mu_{n-1})/\sqrt{5} \end{pmatrix}.$$

Equation (5) immediately implies that

$$F_n = \mu_n/\sqrt{5} \text{ and } A^n = \begin{pmatrix} F_{n-1} & F_n \\ F_n & F_{n+1} \end{pmatrix}.$$

Similarly for the Lucas numbers, we get

$$L_n = 2F_{n-1} + F_n = F_{n-1} + F_{n+1} = (\mu_{n-1} + \mu_{n+1})/\sqrt{5}.$$

If we try and do this for more complicated difference equations, we could meet matrices which aren't diagonalizable. Here's an example (taken from the book by Kaye and Wilson, §14.11), done using Jordan canonical form.

Example. Let x_n, y_n, z_n be sequences of complex numbers satisfying

$$\begin{cases} x_{n+1} &= 3x_n + z_n, \\ y_{n+1} &= -x_n + y_n - z_n, \\ z_{n+1} &= y_n + 2z_n. \end{cases}$$

with $x_0 = y_0 = z_0 = 1$.

We can write this as

$$\mathbf{v}_{n+1} = \begin{pmatrix} 3 & 0 & 1 \\ -1 & 1 & -1 \\ 0 & 1 & 2 \end{pmatrix} \mathbf{v}_n.$$

So we have

$$\mathbf{v}_n = A^n \mathbf{v}_0 = A^n \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

where A is the 3×3 matrix above.

We find that the JCF of A is $J = P^{-1}DP$ where

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

The formula for the entries of J^k for J a Jordan block tells us that

$$\begin{aligned} J^n &= \begin{pmatrix} 2^n & n2^{n-1} & \binom{n}{2}2^{n-2} \\ 0 & 2^n & n2^{n-1} \\ 0 & 0 & 2^n \end{pmatrix} \\ &= 2^n \begin{pmatrix} 1 & \frac{1}{2}n & \frac{1}{4}\binom{n}{2} \\ 0 & 1 & \frac{1}{2}n \\ 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

We therefore have

$$\begin{aligned} A^n &= PJ^nP^{-1} \\ &= 2^n \begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{2}n & \frac{1}{4}\binom{n}{2} \\ 0 & 1 & \frac{1}{2}n \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & -1 \\ 0 & -1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \\ &= 2^n \begin{pmatrix} 1 & 1+\frac{1}{2}n & 1+\frac{1}{2}n+\frac{1}{4}\binom{n}{2} \\ 0 & -1 & -\frac{1}{2}n \\ -1 & -\frac{1}{2}n & -\frac{1}{4}\binom{n}{2} \end{pmatrix} \begin{pmatrix} 0 & 0 & -1 \\ 0 & -1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \\ &= 2^n \begin{pmatrix} 1+\frac{1}{2}n+\frac{1}{4}\binom{n}{2} & \frac{1}{4}\binom{n}{2} & \frac{1}{2}n+\frac{1}{4}\binom{n}{2} \\ -\frac{1}{2}n & 1-\frac{1}{2}n & -\frac{1}{2}n \\ -\frac{1}{4}\binom{n}{2} & \frac{1}{2}n-\frac{1}{4}\binom{n}{2} & 1-\frac{1}{4}\binom{n}{2} \end{pmatrix} \end{aligned}$$

Finally, we obtain

$$A^n \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 2^n \begin{pmatrix} 1 + n + \frac{3}{4}\binom{n}{2} \\ 1 - \frac{3}{2}n \\ 1 + \frac{1}{2}n - \frac{3}{4}\binom{n}{2} \end{pmatrix}$$

or equivalently, using the fact that $\binom{n}{2} = \frac{n(n-1)}{2}$,

$$\begin{cases} x_n &= 2^n(\frac{3}{4}n^2 + \frac{5}{8}n + 1), \\ y_n &= 2^n(1 - \frac{3}{2}n), \\ z_n &= 2^n(-\frac{3}{4}n^2 + \frac{7}{8}n + 1). \end{cases}$$

3.3 Motivation: systems of differential equations

Suppose I want to solve a system of first-order simultaneous differential equations, say

$$\begin{aligned} \frac{\mathrm{d}a}{\mathrm{d}t} &= 3a - 4b + 8c, \\ \frac{\mathrm{d}b}{\mathrm{d}t} &= a - c, \\ \frac{\mathrm{d}c}{\mathrm{d}t} &= a + b + c. \end{aligned}$$

You'll have seen things like this in your Differential Equations course last year. Now, I want to write this in a bit of a different form. If we write $\mathbf{v}(t) = \begin{pmatrix} a(t) \\ b(t) \\ c(t) \end{pmatrix}$ – a vector-valued function of time – we can write the above system as

$$\frac{d\mathbf{v}}{dt} = A\mathbf{v}$$

where A is the matrix

$$\begin{pmatrix} 3 & -4 & 8 \\ 1 & 0 & -1 \\ 1 & 1 & 1 \end{pmatrix}.$$

"Aha!" we say. "I know what the solution is: it's $\mathbf{v}(t) = e^{tA}\mathbf{v}(0)$!" But then we pause, and say "Hang on, what does e^{tA} actually mean?" In this section, we'll use what we now know about special forms of matrices to work out how to define e^{tA}, and other functions of a matrix, in a sensible way that will make this actually work; and having got our definition, we'll work out how to calculate with it. It turns out that Jordan canonical form will play a key role in making the latter possible.

3.4 Definition of a function of a matrix

Suppose we have a "nice" one variable complex-valued function $f(z)$. What is $f(M)$? We know the answer for $f(z) = z^n$ already (see (2) and(3)). Let us rewrite it, using the function $f(z)$:

$$f(M) = Pf(J)P^{-1},\ f(J_{k_1,\lambda_1} \oplus \cdots \oplus J_{k_t,\lambda_t}) = f(J_{k_1,\lambda_1}) \oplus \cdots \oplus f(J_{k_t,\lambda_t}), \tag{6}$$

$$f(J_{k,\lambda}) = \begin{pmatrix} f(\lambda) & f'(\lambda) & \dots & f^{[k-2]}(\lambda) & f^{[k-1]}(\lambda) \\ 0 & f(\lambda) & \dots & f^{[k-3]}(\lambda) & f^{[k-2]}(\lambda) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & f(\lambda) & f'(\lambda) \\ 0 & 0 & \dots & 0 & f(\lambda) \end{pmatrix},$$

where we use "the divided powers derivatives"

$$f^{[k]}(x) := \frac{1}{k!}f^{(k)}(x),\ \text{e.g.,}\ f^{[1]} = f',\ f^{[2]} = \frac{1}{2}f'',\ f^{[3]} = \frac{1}{6}f^{(''')},\ \text{etc.}$$

to make the last formula less cluttered.

There is no "natural" $f(M)$. *We need to choose what we want it to be.* Thus, we **define** $f(M)$ by the equations in (6).

Let the roots of the minimal polynomial $\mu_M(z)$ be $\alpha_1, \dots, \alpha_k$ with multiplicities[4] $m_1, \dots, m_k$.

Since the multiplicity m_i is the maximal size of a Jordan block with of M with eigenvalue α_i, we can make the following conclusion:

Lemma 3.4.1. *(i) If the values[5] of f and its derivatives*

$$f(\alpha_1), f'(\alpha_1), \dots, f^{m_1-1}(\alpha_1), \dots, f(\alpha_k), f'(\alpha_k), \dots, f^{m_k-1}(\alpha_k)$$

are defined, then $f(M)$ is defined.

(ii) If $h(z)$ is another function with the same values on the spectrum of M, then $f(M) = h(M)$.

The upshot of Lemma 3.4.1 is that Lagrange's interpolation polynomial of $f(z)$ at the spectrum of M computes $f(M)$, in general. As an example, let us compute $f(A)$ for $f(z) = \sqrt{-z}$ for the matrix A from Example 8, section 2.10. Suppose the Lagrange interpolation of $\sqrt{-z}$ at the roots of $\mu_A(z) = (z+2)^2$ (or the spectrum of A) is $h(z) = \alpha z + \beta$. Since $f'(z) = \frac{-1}{2\sqrt{-z}}$, the condition on the coefficients is given by

$$\begin{cases} \sqrt{2} &= h(-2) &= -2\alpha + \beta \\ \frac{-1}{2\sqrt{2}} &= h'(-2) &= \alpha \end{cases}$$

[4] This is called the spectrum of M.

[5] And that is called the value of f on the spectrum of M.

Solving them gives $\alpha = \frac{-1}{2\sqrt{2}}$ and $\beta = \frac{1}{\sqrt{2}}$. It follows that

$$f(A) = \frac{-1}{2\sqrt{2}}A + \frac{1}{\sqrt{2}}I = \begin{pmatrix} \sqrt{2} & 0 & 0 & 0 \\ 0 & \sqrt{2} & \frac{-1}{2\sqrt{2}} & 0 \\ 0 & 0 & \sqrt{2} & 0 \\ \frac{-1}{2\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} & \sqrt{2} \end{pmatrix}.$$

 It is worth noticing that the square $f(A)^2$ is, indeed, equal to $-A$.

3.5 Power series

Let us consider a power series $\sum_n a_n x^n$ where $a_n \in \mathbb{C}$ with a positive radius of convergence ε. It defines a function on the ball

$$f : \{z \in \mathbb{C} \mid |z| < \varepsilon\} \to \mathbb{C} \text{ by } f(z) = \sum_{n=0}^{\infty} a_n z^n$$

and the power series is the Taylor's series of $f(x)$ at zero. In particular, $a_n = f^{(n)}(0)/n!$.

 Can we just define $f(M)$ by

$$f^\sharp(M) = \sum_{n=0}^{\infty} a_n M^n \ ? \tag{7}$$

Well, not quite. We do not know

- whether the formula defines anything at all,
- and, even if it defines something, whether it defines $f(M)$.

This is why we write $f^\sharp(M)$ for now. The right hand side of this formula is a matrix series. We can write this as a matrix, whose individual entries are series. But all of these series need to converge for the right hand side to define a matrix.

Chasing up the individual entry series in (7) is a hard task. We should use the *matrix norms*[6] instead. Suppose $M \in \mathbb{C}^{n,n}$. Let

$$|\mathbf{x}| = \sqrt{\mathbf{x}^*\mathbf{x}} \quad \text{or} \quad |(x_i)| = \sqrt{\sum |x_i|^2}$$

be the usual euclidean norm (distance to zero) on $\mathbb{C}^n$. Define the *2-norm*

$$\|\cdot\|_2 : \mathbb{C}^n \to \mathbb{R}_{\geq 0}\,, \ \|A\|_2 := \sup\{\frac{|A\mathbf{x}|}{|\mathbf{x}|} \mid \mathbf{x} \in \mathbb{C}^n \setminus \{0\}\}.$$

Let us summarise the key properties of this norm, not to be proved in this module:

[6]The full scope of this belongs to the modules you can do next term: *Norms, Metrics and Topologies* and *Metric Spaces*.

- (*norm*) $\|A\|_2 = 0$ iff $A = 0$, $\|\alpha A\|_2 = |\alpha|\|A\|_2$, $\|A + B\|_2 \leq \|A\|_2 + \|B\|_2$,
- (*submultiplicativity*) $\|AB\|_2 \leq \|A\|_2 \cdot \|B\|_2$,
- (*completeness*) every Cauchy sequence converges.

Now we can pinpoint what is going on:

Theorem 3.5.1. *If $\|M\|_2 < \varepsilon$, then the following statements hold.*

(i) If λ is an eigenvalue of M, then $\lambda < \varepsilon$.

(ii) Formula (6) defines $f(M)$.

(iii) Power series (7) converges to $f(M)$.

Proof. Let $\mathbf{y}$ be an eigenvector, corresponding to λ. Then $\frac{|A\mathbf{y}|}{|\mathbf{y}|} = |\lambda| \in \{\frac{|A\mathbf{x}|}{|\mathbf{x}|} \mid \mathbf{x} \in \mathbb{C}^n \setminus \{0\}\}$. Thus, $|\lambda| \leq \|M\| < \varepsilon$.

 The next parts are not "Algebra".

Whatever! The algebra[7] is in the eye of beholder. ◡̈

It follows from Complex Analysis[8] that the term-wise derivative of the power series

$$(\sum_{n=0}^{\infty} a_n x^n)^{(m)} = \sum_{n=0}^{\infty} a_n (x^n)^{(m)} = \sum_{n=m}^{\infty} a_n \frac{n!}{(n-m)!} x^{n-m} \tag{8}$$

converges to $f^{(m)}(z)$ on the ball $\{z \in \mathbb{C} \mid |z| < \varepsilon\}$. In particular, all higher derivatives exist on the ball.

By part (i), all eigenvalues of M are in the ball. Thus, f is defined on the spectrum of M and $f(M)$ is defined.

Let $\gamma > 0$ be a real number such that $\|M\|_2 < \gamma < \epsilon$. We have just proved that γ is bigger than the absolute values of all eigenvalues of M. Let $\delta = (\gamma + \varepsilon)/2$. Consider the series

$$\sum_{k=0}^{m} |a_k \gamma^k| = \sum_{k=0}^{m} |a_k| \gamma^k = \sum_{k=0}^{m} \left(\frac{|a_k|}{a_k} \left(\frac{\gamma}{\delta} \right)^k \right) \cdot (a_k \delta^k) \tag{9}$$

where we set $\frac{0}{0}$ equal to 1, to avoid worrying about zeroes. The sequence $\frac{|a_k|}{a_k} \left(\frac{\gamma}{\delta}\right)^k$ converges to zero, while the partial sums $\sum_{k=1}^{m} a_k \delta^k$ are bounded. By Dirichlet's test, the series (9) converges so that the series $\sum_{k=1}^{\infty} a_k \gamma^k$ converges absolutely. Consider the sequence of partial matrix sums

$$M_t := \sum_{k=0}^{t} a_k M^k \,.$$

[7] Having said that, this is a module in Algebra. Some of you are not doing Analysis-3 or Mathematical Analysis-3. Hence, you will not need to know how to justify rigorously the analytic points but you *will* need to be able to work with the functions of matrices and compute examples.

[8] This fact is covered in Analysis-3

This sequence is Cauchy because

$$\|M_s - M_t\|_2 = \|\sum_{k=t+1}^{s} a_k M^k\|_2 \leq \sum_{k=t+1}^{s} |a_k| \|M^k\|_2 \leq \sum_{k=t+1}^{s} |a_k| \gamma^k$$

for all $s > t$. Thus, the series (7) converges by completeness of the norm.

Thus, the power series (7) converges and defines a new function $f^\sharp(M)$. We need to show that the function $f^\sharp(M)$ satisfies all the properties, outlined in (6). This implies that $f^\sharp(M) = f(M)$.

For the first property in (6): since $M^k = PJ^kP^{-1}$ for any $k \in \mathbb{N}$, $h(M) = Ph(J)P^{-1}$ for any polynomial $h(z)$. So if we consider the sum of the first m terms of the series, we have

$$\sum_{k=0}^{m} a_k M^k = \sum_{k=0}^{m} a_k P J^k P^{-1} = P \left(\sum_{k=0}^{m} a_k J^k \right) P^{-1}.$$

Taking the limit as $m \to \infty$, we deduce the first property

$$f(M) = \sum_{k=0}^{\infty} a_k M^k = P \left(\sum_{k=0}^{\infty} a_k J^k \right) P^{-1} = P f(J) P^{-1} .$$

The second property is obvious.

The proof of the third property is similar to the first one. The formula for $h(J_{k,\lambda})$ can be checked directly for powers $h(z) = z^m$ and polynomials. Let $f_m = \sum_{k=0}^{m} a_k z^k$. We need to establish convergence

$$\begin{pmatrix} f_m(\lambda) & \dots & f_m^{(k-2)}(\lambda) & f_m^{(k-1)}(\lambda) \\ 0 & \dots & f_m^{(k-3)}(\lambda) & f_m^{(k-2)}(\lambda) \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & f_m(\lambda) & f_m'(\lambda) \\ 0 & \dots & 0 & f_m(\lambda). \end{pmatrix} \longrightarrow \begin{pmatrix} f(\lambda) & f'(\lambda) & \dots & f^{(k-2)}(\lambda) & f^{(k-1)}(\lambda) \\ 0 & f(\lambda) & \dots & f^{(k-3)}(\lambda) & f^{(k-2)}(\lambda) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & f(\lambda) & f'(\lambda) \\ 0 & 0 & \dots & 0 & f(\lambda). \end{pmatrix}$$

as $m \to \infty$. This is exactly the property stated above: the term-wise derivative of the power series converges to $f^{(m)}(z)$ (see (8)). □

3.6 The exponential of a matrix

The Taylor's series at zero of the exponential function is $\sum_{k=0}^{\infty} \frac{z^k}{k!}$. Hence, given a matrix A, the matrix exponent is

$$e^A = I_n + A + \frac{A^2}{2} + \frac{A^3}{6} + \dots = \sum_{k=0}^{\infty} \frac{A^k}{k!}.$$

It is *not* true that $e^{A+B} = e^A e^B$ for general matrices A and B; we will see an example of this on the problem sheet. The precise connection between e^{A+B}, e^A and e^B is a subject of the Baker-Campbell-Hausdorff formula, beyond the remit of this module.

Lemma 3.6.1.

(i) We have $e^{A+B} = e^A e^B$ **if** $A, B \in \mathbb{C}^{n,n}$ *satisfy* $AB = BA$.

(ii) If we consider the matrix-valued function $t \mapsto e^{tA}$, *for* $t \in \mathbb{R}$, *then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{tA}\right) = Ae^{tA}.$$

Sketch proof. For part (i), it's easily checked by induction on k that if $AB = BA$, then the binomial expansion of $(A+B)^k$ works: we have

$$(A+B)^k = \sum_{j=0}^{k} \binom{k}{j} A^{k-j} B^j.$$

Substituting this into the power series definition of e^{A+B}, and grouping terms appropriately gives the result.

For part (ii), we need to show that

$$\lim_{h\to 0}\left(\frac{e^{(t+h)A} - e^{tA}}{h}\right) = Ae^{tA}.$$

Using part (i), we know that $e^{(t+h)A} = e^{tA}e^{hA}$, so we just need to show that

$$\lim_{h\to 0}\left(\frac{e^{hA} - I_n}{h}\right) = A.$$

Substituting in the series expansion of e^{hA}, we get

$$\frac{e^{Ah} - I_n}{h} = A + \frac{A^2}{2}h + \frac{A^3}{3!}h^2 + \ldots$$

and, clearly, the limit of the right-hand side is A. □

Let us do several examples.

Example 9. Consider $A = \begin{pmatrix} 1 & 4 \\ 1 & 1 \end{pmatrix}$. This was Example 1 from Section 2.8 above, and we saw that $P^{-1}AP = D$ where

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

Hence

$$\begin{aligned} e^{tA} &= Pe^{Dt}P^{-1} \\ &= \begin{pmatrix} 2 & -2 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} e^{3t} & 0 \\ 0 & e^{-t} \end{pmatrix}\begin{pmatrix} 2 & -2 \\ 1 & 1 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \frac{1}{2}e^{3t} + \frac{1}{2}e^{-t} & e^{3t} - e^{-t} \\ \frac{1}{4}e^{3t} - \frac{1}{4}e^{-t} & \frac{1}{2}e^{3t} + \frac{1}{2}e^{-t} \end{pmatrix}. \end{aligned}$$

Example 10. Let

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

Using the methods of the last chapter we can check that its JCF is $J = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and the basis change matrix P such that $J = P^{-1}AP$ is given by $P = \begin{pmatrix} 3 & 0 & 2 \\ 3 & 1 & 1 \\ -1 & -1 & 0 \end{pmatrix}$.

Applying the argument above, we see that $e^{tA} = Pe^{Jt}P^{-1}$ where

$$e^{Jt} = \begin{pmatrix} e^{2t} & te^{2t} & 0 \\ 0 & e^{2t} & 0 \\ 0 & 0 & e^{t} \end{pmatrix}.$$

We can now calculate e^{tA} explicitly by doing the matrix multiplication to get the entries of $Pe^{Jt}P^{-1}$, as we did in the 2×2 example above.


It looks messy. Do we really want to write it down here?

Well, let us not do it. In a pen-and-paper calculation, except a few cases (for example, diagonal matrices) it is simpler to use Lagrange's interpolation.

Example 11. Let us compute e^A for the matrix A from Example 8, Section 2.10, using the Lagrange interpolation. Suppose that $h(z) = \alpha z + \beta$ is the interpolation of e^z at the roots of $\mu_A(z) = (z+2)^2$. The condition on the coefficients is given by

$$\begin{cases} e^{-2} = h(-2) = -2\alpha + \beta \\ e^{-2} = h'(-2) = \alpha \end{cases}$$

Solving them gives $\alpha = e^{-2}$ and $\beta = 3e^{-2}$. It follows that

$$e^A = e^{-2}A + 3e^{-2}I = \begin{pmatrix} e^{-2} & 0 & 0 & 0 \\ 0 & e^{-2} & e^{-2} & 0 \\ 0 & 0 & e^{-2} & 0 \\ e^{-2} & 0 & -2e^{-2} & e^{-2} \end{pmatrix}.$$

Example 12. Let us consider a harmonic oscillator described by the equation $y''(t) + y(t) = 0$. The general solution $y(t) = \alpha \sin(t) + \beta \cos(t)$ is well known. Let us obtain it using matrix exponents. Setting

$$x(t) = \begin{pmatrix} y(t) \\ y'(t) \end{pmatrix}, \; A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

the harmonic oscillator becomes the initial value problem with a solution $x(t) = e^{tA}x(0)$. The eigenvalues of A are i and $-i$. Interpolating e^{zt} at these values of z gives the following condition on $h(z) = \alpha z + \beta$

$$\begin{cases} e^{it} &= h(i) &= \alpha i + \beta \\ e^{-it} &= h(-i) &= -\alpha i + \beta \end{cases}$$

Solving them gives $\alpha = (e^{it} - e^{-it})/2i = \sin(t)$ and $\beta = (e^{it} + e^{-it})/2 = \cos(t)$. It follows that

$$e^{tA} = \sin(t)A + \cos(t)I_2 = \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix}$$

and $y(t) = \cos(t)y(0) + \sin(t)y'(0)$.

Example 13. Let us consider a system of differential equations

$$\begin{cases} y_1' &= y_1 - 3y_3 \\ y_2' &= y_1 - y_2 - 6y_3 \\ y_3' &= -y_1 + 2y_2 + 5y_3 \end{cases}, \text{ with the initial condition } \begin{cases} y_1(0) &= 1 \\ y_2(0) &= 1 \\ y_3(0) &= 0 \end{cases}$$

Using matrices

$$x(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{pmatrix},\ w = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix},\ A = \begin{pmatrix} 1 & 0 & -3 \\ 1 & -1 & -6 \\ -1 & 2 & 5 \end{pmatrix},$$

it becomes an initial value problem. The characteristic polynomial is $c_A(z) = -z^3 + 5z^2 - 8z + 4 = (1-z)(2-z)^2$. We need to interpolate e^{tz} at 1 and 2 by $h(z) = \alpha z^2 + \beta z + \gamma$. At the multiple root 2 we need to interpolate up to order 2 that involves tracking the derivative $(e^{tz})' = te^{tz}$:

$$\begin{cases} e^{t} &= h(1) &= \alpha + \beta + \gamma \\ e^{2t} &= h(2) &= 4\alpha + 2\beta + \gamma \\ te^{2t} &= h'(2) &= 4\alpha + \beta \end{cases}$$

Solving, $\alpha = (t-1)e^{2t} + e^t$, $\beta = (4-3t)e^{2t} - 4e^t$, $\gamma = (2t-3)e^{2t} + 4e^t$. It follows that

$$e^{tA} = e^{2t}\begin{pmatrix} 3t-3 & -6t+6 & -9t+6 \\ 3t-2 & -6t+4 & -9t+3 \\ -t & 2t & 3t+1 \end{pmatrix} + e^t\begin{pmatrix} 4 & -6 & -6 \\ 2 & -3 & -3 \\ 0 & 0 & 0 \end{pmatrix}$$

and

$$x(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{pmatrix} = e^{tA}\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} (3-3t)e^{2t} - 2e^t \\ (2-3t)e^{2t} - e^t \\ te^{2t} \end{pmatrix}.$$

Example 14. Let x and y be functions of time satisfying

$$\frac{\mathrm{d}x}{\mathrm{d}t} = 3x - 4y,$$
$$\frac{\mathrm{d}y}{\mathrm{d}t} = x - y.$$

We can write this as $\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = A\mathbf{v}$, where $\mathbf{v}(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$ and $A = \begin{pmatrix} 3 & -4 \\ 1 & -1 \end{pmatrix}$. Suppose we choose any vector $\mathbf{v}_0 \in \mathbb{C}^{2,1}$; then it's clear from the lemma that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{tA}\mathbf{v}_0\right) = Ae^{tA}\mathbf{v}_0.$$

Hence if we set $\mathbf{v}(t) = e^{tA}\mathbf{v}_0$, we know that this is a solution to the differential equations, with $\mathbf{v}(0) = \mathbf{v}_0$ being the vector we started with. And we can use what we know about Jordan form to calculate e^{tA}: we have

$$e^{tA} = \begin{pmatrix} (1+2t)e^t & -4te^t \\ te^t & (1-2t)e^t \end{pmatrix}.$$

So the general solution of the equations above is

$$\begin{cases} x(t) &= (1+2t)e^t x(0) - 4te^t y(0), \\ y(t) &= te^t x(0) + (1-2t)e^t y(0). \end{cases}$$

4 Bilinear Maps and Quadratic Forms

We'll now introduce another, rather different kind of object you can define for vector spaces: a *bilinear map*. These are a bit different from linear maps: rather than being machines that take a vector and spit out another vector, they take two vectors as input and spit out a number.

4.1 Bilinear maps: definitions

Let V and W be vector spaces over a field K.

Definition 4.1.1. A *bilinear map* on V and W is a map $\tau : V \times W \to K$ such that

(i) $\tau(\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2, \mathbf{w}) = \alpha_1\tau(\mathbf{v}_1, \mathbf{w}) + \alpha_2\tau(\mathbf{v}_2, \mathbf{w})$; and

(ii) $\tau(\mathbf{v}, \alpha_1\mathbf{w}_1 + \alpha_2\mathbf{w}_2) = \alpha_1\tau(\mathbf{v}, \mathbf{w}_1) + \alpha_2\tau(\mathbf{v}, \mathbf{w}_2)$

for all $\mathbf{v}, \mathbf{v}_1, \mathbf{v}_2 \in V$, $\mathbf{w}, \mathbf{w}_1, \mathbf{w}_2 \in W$, and $\alpha_1, \alpha_2 \in K$.

So $\tau(\mathbf{v}, \mathbf{w})$ is linear in $\mathbf{v}$ for each $\mathbf{w}$, and linear in $\mathbf{w}$ for each $\mathbf{v}$ – linear in two different ways, hence the term "bilinear".

Clearly if we fix bases of V and W, a bilinear map will be determined by what it does to the basis vectors. Choose a basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of V and a basis $\mathbf{f}_1, \ldots, \mathbf{f}_m$ of W.

Let $\tau : V \times W \to K$ be a bilinear map, and let $\alpha_{ij} = \tau(\mathbf{e}_i, \mathbf{f}_j)$, for $1 \leq i \leq n$, $1 \leq j \leq m$. Then the $n \times m$ matrix $A = (\alpha_{ij})$ is defined to be the matrix of τ with respect to the bases $\mathbf{e}_1, \ldots, \mathbf{e}_n$ and $\mathbf{f}_1, \ldots, \mathbf{f}_m$ of V and W.

For $\mathbf{v} \in V$, $\mathbf{w} \in W$, let $\mathbf{v} = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n$ and $\mathbf{w} = y_1\mathbf{f}_1 + \cdots + y_m\mathbf{f}_m$, so the coordinates of $\mathbf{v}$ and $\mathbf{w}$ with respect to our bases are

$$\underline{\mathbf{v}} = \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{pmatrix} \in K^{n,1}, \quad \text{and} \quad \underline{\mathbf{w}} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_m \end{pmatrix} \in K^{m,1}.$$

Then, by using the equations (i) and (ii) above, we get

$$\tau(\mathbf{v}, \mathbf{w}) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i\,\tau(\mathbf{e}_i, \mathbf{f}_j)\,y_j = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i\,\alpha_{ij}\,y_j = \underline{\mathbf{v}}^{\mathrm{T}} A \underline{\mathbf{w}}. \tag{2.1}$$

So once we've fixed bases of V and W, every bilinear map on V and W corresponds to an $n \times m$ matrix, and conversely every matrix determines a bilinear map.

For example, let $V = W = \mathbb{R}^2$ and use the natural basis of V. Suppose that $A = \begin{pmatrix} 1 & -1 \\ 2 & 0 \end{pmatrix}$. Then

$$\tau((x_1, x_2), (y_1, y_2)) = (x_1\ x_2) \begin{pmatrix} 1 & -1 \\ 2 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = x_1y_1 - x_1y_2 + 2x_2y_1.$$

4.2 Bilinear maps: change of basis

Week 5

We retain the notation of the previous section, so τ is a bilinear map on V and W, and A is the matrix of τ with respect to some bases $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of V and $\mathbf{f}_1, \ldots, \mathbf{f}_m$ of W.

As in §1.5 of the course, suppose that we choose new bases $\mathbf{e}'_1, \ldots, \mathbf{e}'_n$ of V and $\mathbf{f}'_1, \ldots, \mathbf{f}'_m$ of W, and let P and Q be the associated basis change matrices. Let B be the matrix of τ with respect to these new bases.

Let $\mathbf{v}$ be any vector in V. Then we know (from Proposition 1.5.1) that if $\underline{\mathbf{v}} \in K^{n,1}$ is the column vector of coordinates of $\mathbf{v}$ with respect to the old basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$, and $\underline{\mathbf{v}}'$ the coordinates of $\mathbf{v}$ in the new basis $\mathbf{e}'_1, \ldots, \mathbf{e}'_n$, then we have $P\underline{\mathbf{v}}' = \underline{\mathbf{v}}$. Similarly, for any $\mathbf{w} \in W$, the coordinates $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}'$ of $\mathbf{w}$ with respect to the old and new bases of W are related by $Q\underline{\mathbf{w}}' = \underline{\mathbf{w}}$.

We know that we have

$$\underline{\mathbf{v}}^{\mathrm{T}} A \underline{\mathbf{w}} = \tau(\mathbf{v}, \mathbf{w}) = (\underline{\mathbf{v}}')^{\mathrm{T}} B \underline{\mathbf{w}}'.$$

Substituting in the formulae from Proposition 1.5.1, we have

$$\begin{aligned} (\underline{\mathbf{v}}')^{\mathrm{T}} B \underline{\mathbf{w}}' &= (P\underline{\mathbf{v}}')^{\mathrm{T}} A (Q\underline{\mathbf{w}}') \\ &= (\underline{\mathbf{v}}')^{\mathrm{T}}\, P^{\mathrm{T}} A Q\, \underline{\mathbf{w}}'. \end{aligned}$$

Since this relation must hold for all $\underline{\mathbf{v}}' \in K^{n,1}$ and $\underline{\mathbf{w}}' \in K^{m,1}$, the two matrices in the middle must be equal (exercise!): that is, we have $B = P^{\mathrm{T}} A Q$. So we've proven:

Theorem 4.2.1. *Let A be the matrix of the bilinear map $\tau : V \times W \to K$ with respect to the bases $\mathbf{e}_1, \ldots, \mathbf{e}_n$ and $\mathbf{f}_1, \ldots, \mathbf{f}_m$ of V and W, and let B be its matrix with respect to the bases $\mathbf{e}'_1, \ldots, \mathbf{e}'_n$ and $\mathbf{f}'_1, \ldots, \mathbf{f}'_m$ of V and W. Let P and Q be the basis change matrices, as defined above. Then $B = P^{\mathrm{T}} A Q$.*

Compare this result with Theorem 1.5.2.

We shall be concerned from now on only with the case where $V = W$. A bilinear map $\tau : V \times V \to K$ is called a *bilinear form* on V. Theorem 4.2.1 then becomes:

Theorem 4.2.2. *Let A be the matrix of the bilinear form τ on V with respect to the basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of V, and let B be its matrix with respect to the basis $\mathbf{e}'_1, \ldots, \mathbf{e}'_n$ of V. Let P the basis change matrix with original basis $\{\mathbf{e}_i\}$ and new basis $\{\mathbf{e}'_i\}$. Then $B = P^{\mathrm{T}} A P$.*

Let's give a name to this relation between matrices:

Definition 4.2.3. Two matrices A and B are called *congruent* if there exists an invertible matrix P with $B = P^{\mathrm{T}}AP$.

So congruent matrices represent the same bilinear form in different bases. Notice that congruence is very different from similarity; if τ is a bilinear form on V and T is a linear operator on V, it might be the case that τ and T have the same matrix A in some specific basis of V, but that doesn't mean that they have the same matrix in any other basis – they inhabit different worlds.

So, in the example at the end of Subsection 4.1, if we choose the new basis $\mathbf{e}'_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, $\mathbf{e}'_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ then $P = \begin{pmatrix} 1 & 1 \\ -1 & 0 \end{pmatrix}$, $P^{\mathrm{T}}AP = \begin{pmatrix} 0 & -1 \\ 2 & 1 \end{pmatrix}$, and

$$\tau((y'_1\mathbf{e}'_1 + y'_2\mathbf{e}'_2, x'_1\mathbf{e}'_1 + x'_2\mathbf{e}'_2)) = -y'_1x'_2 + 2y'_2x'_1 + y'_2x'_2.$$

Since P is an invertible matrix, P^{T} is also invertible (its inverse is $(P^{-1})^{\mathrm{T}}$), and so the matrices $P^{\mathrm{T}}AP$ and A are "equivalent matrices" in the sense of MA106, and hence have the same rank.

The rank of the bilinear form τ is defined to be the rank of its matrix A. So we have just shown that the rank of τ is a well-defined property of τ, not depending on the choice of basis we've used.

In fact we can say a little more. It's clear that a vector $\underline{\mathbf{v}} \in K^{n,1}$ is zero if and only if $\underline{\mathbf{v}}^{\mathrm{T}}\underline{\mathbf{w}} = 0$ for all vectors $\underline{\mathbf{w}} \in K^{n,1}$. Since

$$\tau(\mathbf{v}, \mathbf{w}) = \underline{\mathbf{v}}^{\mathrm{T}}A\underline{\mathbf{w}},$$

the kernel of A is equal to the space

$$\{\mathbf{v} \in V : \tau(\mathbf{w}, \mathbf{v}) = 0 \ \forall \mathbf{w} \in V\}$$

(the *right radical* of τ) and the kernel of A^{T} is equal to the space

$$\{\mathbf{v} \in V : \tau(\mathbf{v}, \mathbf{w}) = 0 \ \forall \mathbf{w} \in V\}$$

(the *left radical*). Since A^{T} and A have the same rank, the left and right radicals both have dimension $n - r$, where r is the rank of τ. In particular, the rank of τ is n if and only if the left and right radicals are zero. If this occurs, we'll say τ is *nondegenerate*; so τ is nondegenerate if and only if its matrix (in any basis) is nonsingular.

You could be forgiven for expecting that we were about to launch into a long study of how to choose, given a bilinear form τ on V, the "best" basis for V which makes the matrix of τ as nice as possible. We are *not* going to do this, because although it's a very natural question to ask, it's *extremely* hard! Instead, we'll restrict ourselves to a special kind of bilinear form where life is much easier, which covers most of the bilinear forms that come up in "real life".

Definition 4.2.4. We say bilinear form τ on V is *symmetric* if $\tau(\mathbf{w}, \mathbf{v}) = \tau(\mathbf{v}, \mathbf{w})$ for all $\mathbf{v}, \mathbf{w} \in V$.

We say τ is *antisymmetric* (or sometimes *alternating*) if $\tau(\mathbf{w}, \mathbf{v}) = -\tau(\mathbf{v}, \mathbf{w})$.

An $n \times n$ matrix A is called symmetric if $A^{\mathrm{T}} = A$, and anti-symmetric if $A^{\mathrm{T}} = -A$. We then clearly have:

Proposition 4.2.5. *The bilinear form τ is symmetric or anti-symmetric if and only if its matrix (with respect to any basis) is symmetric or anti-symmetric.*

The best known example of a symmetric form is when $V = \mathbb{R}^n$, and τ is defined by

$$\tau\left(\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}\right) = x_1y_1 + x_2y_2 + \cdots + x_ny_n.$$

This form has matrix equal to the identity matrix I_n with respect to the standard basis of $\mathbb{R}^n$. Geometrically, it is equal to the normal scalar product $\tau(\mathbf{v}, \mathbf{w}) = |\mathbf{v}||\mathbf{w}| \cos\theta$, where θ is the angle between the vectors $\mathbf{v}$ and $\mathbf{w}$.

On the other hand, the form on $\mathbb{R}^2$ defined by $\tau(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}) = x_1y_2 - x_2y_1$ is anti-symmetric. This has matrix $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$.

Technical note: For the theory of symmetric and anti-symmetric bilinear forms to work properly, we need to be able to divide by 2 in the field K. This means that we must assume that $1 + 1 \neq 0$ in K. So, for example, K is not allowed to be the field $\mathbb{F}_2$ of order 2. If you prefer to avoid worrying about technicalities like this, then you can safely assume that K is either $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$.

Having made this assumption, we can easily show that

Proposition 4.2.6. *Any bilinear form τ can be written uniquely as $\tau_1 + \tau_2$ where τ_1 is symmetric and τ_2 is antisymmetric.*

Proof. We just put $\tau_1(\mathbf{v}, \mathbf{w}) = \frac{1}{2}(\tau(\mathbf{v}, \mathbf{w}) + \tau(\mathbf{w}, \mathbf{v}))$ and $\tau_2(\mathbf{v}, \mathbf{w}) = \frac{1}{2}(\tau(\mathbf{v}, \mathbf{w}) - \tau(\mathbf{w}, \mathbf{v}))$. It's clear that τ_1 is symmetric and τ_2 is antisymmetric.

Moreover, given any other such expression $\tau = \tau_1' + \tau_2'$, we have

$$\begin{aligned} \tau_1(\mathbf{v}, \mathbf{w}) &= \frac{\tau_1'(\mathbf{v}, \mathbf{w}) + \tau_1'(\mathbf{w}, \mathbf{v}) + \tau_2'(\mathbf{v}, \mathbf{w}) + \tau_2'(\mathbf{w}, \mathbf{v})}{2} \\ &= \frac{\tau_1'(\mathbf{v}, \mathbf{w}) + \tau_1'(\mathbf{v}, \mathbf{w}) + \tau_2'(\mathbf{v}, \mathbf{w}) - \tau_2'(\mathbf{v}, \mathbf{w})}{2} \end{aligned}$$

from the symmetry and antisymmetry of τ_1' and τ_2'. The last two terms cancel each other and we just have

$$= \frac{2\tau_1'(\mathbf{v}, \mathbf{w})}{2} = \tau_1'(\mathbf{v}, \mathbf{w}).$$

So $\tau_1 = \tau_1'$, and so $\tau_2 = \tau - \tau_1 = \tau - \tau_1' = \tau_2'$, so the decomposition is unique. □

(Notice that $\frac{1}{2}$ has to exist in K for all this to make sense!)

4.3 Quadratic forms

If τ is a bilinear form on a vector space V, then we can consider the function $q(\mathbf{v}) = \tau(\mathbf{v}, \mathbf{v})$. If we fix a basis of V, then we can represent $\mathbf{v}$ by its coordinates, and q will be a polynomial function in the coordinates of $\mathbf{v}$ of which every term is of degree 2, like $3x^2 + 2xz + z^2 - 4yz + xy$. This is what we call a *quadratic form*.

These quadratic forms are important objects. For instance, one sees them often in geometry: an equation like

$$3x^2 + 2xz + z^2 - 4yz + xy = 17$$

describes a surface in $\mathbb{R}^3$, and we'll see later that we can use linear algebra to understand the shape of this surface.

Definition 4.3.1. Let V be a vector space over the field K. Then a *quadratic form* on V is a function $q : V \to K$ that is defined by $q(\mathbf{v}) = \tau(\mathbf{v}, \mathbf{v})$, where $\tau : V \times V \to K$ is a bilinear form.

If we write $\tau = \tau_1 + \tau_2$ with τ_1 symmetric and τ_2 antisymmetric, as in 4.2.6 above, then $\tau(\mathbf{v}, \mathbf{v}) = \tau_1(\mathbf{v}, \mathbf{v})$, so any quadratic form can be written in the form $\tau(\mathbf{v}, \mathbf{v})$ for τ symmetric.

On the other hand, if τ is symmetric then for $\mathbf{u}, \mathbf{v} \in V$,

$$q(\mathbf{u}+\mathbf{v}) = \tau(\mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v}) = \tau(\mathbf{u}, \mathbf{u}) + \tau(\mathbf{v}, \mathbf{v}) + \tau(\mathbf{u}, \mathbf{v}) + \tau(\mathbf{v}, \mathbf{u}) = q(\mathbf{u}) + q(\mathbf{v}) + 2\tau(\mathbf{u}, \mathbf{v})$$

and hence $\tau(\mathbf{u}, \mathbf{v}) = (q(\mathbf{u}+\mathbf{v}) - q(\mathbf{u}) - q(\mathbf{v}))/2$, and τ is completely determined by q. Hence there is a one-one correspondence between symmetric bilinear forms on V and quadratic forms on V.

Let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ be a basis of V. Recall that the coordinates of $\mathbf{v}$ with respect to this basis are defined to be the field elements x_i such that $\mathbf{v} = \sum_{i=1}^n x_i \mathbf{e}_i$.

Let $A = (\alpha_{ij})$ be the matrix of τ with respect to this basis. We will also call A the matrix of q with respect to this basis. Then A is symmetric because τ is, and by Equation (2.1) of Subsection 4.1, we have

$$q(\mathbf{v}) = \underline{\mathbf{v}}^{\mathrm{T}} A \underline{\mathbf{v}} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i \alpha_{ij} x_j = \sum_{i=1}^{n} \alpha_{ii} x_i^2 + 2 \sum_{i=1}^{n} \sum_{j=1}^{i-1} \alpha_{ij} x_i x_j. \tag{3.1}$$

Conversely, if we are given a quadratic form as in the right hand side of Equation (3.1), then is easy to write down its matrix A. For example, if $n = 3$ and $q(\mathbf{v}) = 3x^2 + y^2 - 2z^2 + 4xy - xz$, then $A = \begin{pmatrix} 3 & 2 & -1/2 \\ 2 & 1 & 0 \\ -1/2 & 0 & -2 \end{pmatrix}$.

4.4 Nice bases for quadratic forms

We'll now show how to choose a basis for V which makes a given symmetric bilinear form (or, equivalently, quadratic form) "as nice as possible". This will turn out to be much easier than

the corresponding problem for linear operators.

Theorem 4.4.1. *Let V be a vector space of dimension n equipped with a symmetric bilinear form τ (or, equivalently, a quadratic form q).*

Then there is a basis $\mathbf{b}_1, \dots, \mathbf{b}_n$ of V, and constants $\beta_1, \dots, \beta_n$, such that

$$\tau(\mathbf{b}_i, \mathbf{b}_j) = \begin{cases} \beta_i & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases}.$$

Equivalently,

- *given any symmetric matrix A, there is an invertible matrix P such that $P^{\mathrm{T}}AP$ is a diagonal matrix (i.e. A is congruent to a diagonal matrix);*
- *given any quadratic form q on a vector space V, there is a basis $\mathbf{b}_1, \dots, \mathbf{b}_n$ of v and constants $\beta_1, \dots, \beta_n$ such that*

$$q(x_1\mathbf{b}_1 + \cdots + x_n\mathbf{b}_n) = \beta_1 x_1^2 + \cdots + \beta_n x_n^2.$$

Proof. We shall prove this by induction on $n = \dim V$. If $n = 0$ then there is nothing to prove, so let's assume that $n \geq 1$.

If τ is zero, then again there is nothing to prove, so we may assume that $\tau \neq 0$. Then the associated quadratic form q is not zero either, so there is a vector $\mathbf{v} \in V$ such that $q(\mathbf{v}, \mathbf{v}) \neq 0$. Let $\mathbf{b}_1 = \mathbf{v}$ and let $\beta_1 = q(\mathbf{v})$.

Consider the linear map $V \to K$ given by $\mathbf{w} \mapsto \tau(\mathbf{w}, \mathbf{v})$. This is not the zero map, so its image has rank 1; so its kernel W has rank $n - 1$. Moreover, this $(n-1)$-dimensional subspace doesn't contain $\mathbf{b}_1 = \mathbf{v}$.

By the induction hypothesis, we can find a basis $\mathbf{b}_2, \dots, \mathbf{b}_n$ for the kernel such that $\tau(\mathbf{b}_i, \mathbf{b}_j) = 0$ for all $2 \leq i < j \leq n$; and all of these vectors lie in the space W, so we also have $\tau(\mathbf{b}_1, \mathbf{b}_j) = 0$ for all $2 \leq j \leq n$. Since $\mathbf{b}_1 \notin W$, it follows that $\mathbf{b}_1, \dots, \mathbf{b}_n$ is a basis of V, so we're done. □

Finding the good basis: The above proof is quite short and slick, and gives us very little help if we explicitly want to find the diagonalizing basis. So let's unravel what's going on a bit more explicitly. We'll see in a moment that what's going on is very closely related to "completing the square" in school algebra.

So let's say we have a quadratic form q. As usual, let $B = (\beta_{ij})$ be the matrix of q with respect to some random basis $\mathbf{b}_1, \dots, \mathbf{b}_n$. We'll modify the basis $\mathbf{b}_i$ step-by-step in order to eventually get it into the nice form the theorem predicts. Let β_{ij} be the matrix of q with respect to the basis $\mathbf{b}_1, \dots, \mathbf{b}_n$.

Step 1: Arrange that $q(\mathbf{b}_1) \neq 0$. Here there are various cases to consider.

- If $\beta_{11} \neq 0$, then we're done: this means that $q(\mathbf{b}_1) \neq 0$, so we don't need to do anything.
- If $\beta_{11} = 0$, but $\beta_{ii} \neq 0$ for some $i > 1$, then we just interchange $\mathbf{b}_1$ and $\mathbf{b}_i$ in our basis.
- If $\beta_{ii} = 0$ for all i, but there is some i and j such that $\beta_{ij} \neq 0$, then we replace $\mathbf{b}_i$ with $\mathbf{b}_i + \mathbf{b}_j$; since
$$q(\mathbf{b}_i + \mathbf{b}_j) = q(\mathbf{b}_i) + q(\mathbf{b}_j) + 2\tau(\mathbf{b}_i, \mathbf{b}_j) = 2\beta_{ij},$$
after making this change we have $q(\mathbf{b}_i) \neq 0$, so we're reduced to one of the two previous cases.
- If $\beta_{ij} = 0$ for all i and j, we can stop: the matrix of q is zero, so it's certainly diagonal.

Step 2: Modify $\mathbf{b}_2, \ldots, \mathbf{b}_n$ to make them orthogonal to $\mathbf{b}_1$. Suppose we've done Step 1, but we haven't stopped, so β_{11} is now non-zero. We want to arrange that $\tau(\mathbf{b}_1, \mathbf{b}_i)$ is 0 for all $i > 1$. To do this, we just replace $\mathbf{b}_i$ with

$$\mathbf{b}_i - \frac{\beta_{1i}}{\beta_{11}}\mathbf{b}_1.$$

This works because

$$\tau(\mathbf{b}_1, \mathbf{b}_i - \frac{\beta_{1i}}{\beta_{11}}\mathbf{b}_1) = \tau(\mathbf{b}_1, \mathbf{b}_i) - \frac{\beta_{1i}}{\beta_{11}}\tau(\mathbf{b}_1, \mathbf{b}_1) = \beta_{1i} - \frac{\beta_{1i}}{\beta_{11}}\beta_{11} = 0.$$

This is where the relation to "completing the square" comes in. We've changed our basis by the matrix

$$P = \begin{pmatrix} 1 & -\frac{\beta_{12}}{\beta_{11}} & \cdots & -\frac{\beta_{1n}}{\beta_{11}} \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}$$

so the coordinates of a vector $\mathbf{v} \in V$ change by the inverse of this, which is just

$$P^{-1} = \begin{pmatrix} 1 & \frac{\beta_{12}}{\beta_{11}} & \cdots & \frac{\beta_{1n}}{\beta_{11}} \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}$$

This corresponds to writing

$$q(x_1\mathbf{b}_1 + \cdots + x_n\mathbf{b}_n) = \beta_{11}x_1^2 + 2\beta_{12}x_1x_2 + \cdots + 2\beta_{1n}x_1x_n + C$$

where C doesn't involve x_1 at all, and writing this as

$$\beta_{11}\left(x_1 + \frac{\beta_{12}}{\beta_{11}}x_2 + \cdots + \frac{\beta_{1n}}{\beta_{11}}x_n\right)^2 + C'$$

where C' also doesn't involve x_1. Then our change of basis changes the coordinates so the whole bracketed term becomes the first coordinate of $\mathbf{v}$; we've eliminated "cross terms" involving x_1 and one of the other variables.

Step 3: Induct on n. Now we've managed to engineer a basis $\mathbf{b}_1, \dots, \mathbf{b}_n$ such that the matrix $B = \beta_{ij}$ of q looks like

$$\begin{pmatrix} \beta_{11} & 0 & \dots & 0 \\ 0 & ? & \dots & ? \\ \vdots & \vdots & \ddots & \vdots \\ 0 & ? & \dots & ? \end{pmatrix}$$

So we can now repeat the process with V replaced by the $(n-1)$-dimensional vector space W spanned by $\mathbf{b}_2, \dots, \mathbf{b}_n$. We can mess around as much as we like with the vectors $\mathbf{b}_2, \dots, \mathbf{b}_n$ without breaking the fact that they pair to zero with $\mathbf{b}_1$, since this is true of any vector in W. So we go back to step 1 but with a smaller n, and keep going until we either have an 0-dimensional space or a zero form, in which case we can safely stop.

Example. Let $V = \mathbb{R}^3$ and $q(\begin{pmatrix} x \\ y \\ z \end{pmatrix}) = xy + 3yz - 5xz$, so the matrix of q with respect to the standard basis $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ is

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

 This example sucks.

OK, let us redo it slightly. Since we have only got 3 variables, it's much less work to call them x, y, z than x_1, x_2, x_3. When we change the variables, we will write x_1, y_1, z_1 and so on. We still proceed as in the previous proof. You need to read the proof first! We will use $\overset{\heartsuit}{=}$ for the equalities that need no checking (they are for information purposes only).

First change of basis: All the diagonal entries of A are zero, so we're in Case 3 of Step 1 of the proof above. But a_{12} is $1/2$, which isn't zero; so we replace $\mathbf{e}_1$ with $\mathbf{e}_1 + \mathbf{e}_2$. That is, we work in the basis

$$\mathbf{b}_1 := \mathbf{e}_1 + \mathbf{e}_2, \ \mathbf{b}_2 := \mathbf{e}_2, \ \mathbf{b}_3 := \mathbf{e}_3 .$$

Thus the basis change matrix from $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ to $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ is

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ so that } \begin{pmatrix} x \\ y \\ z \end{pmatrix} \overset{\heartsuit}{=} P \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix}$$

where $\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix}$ is the coordinate expression in the new basis. And we have

$$q(x_1\mathbf{b}_1 + y_1\mathbf{b}_2 + z_1\mathbf{b}_3) = q\begin{pmatrix} x_1 \\ x_1 + y_1 \\ z_1 \end{pmatrix} =$$

$$= x_1(x_1 + y_1) + 3(x_1 + y_1)z_1 - 5x_1z_1 = x_1^2 + x_1y_1 - 2x_1z_1 + 3y_1z_1,$$

so the matrix of q in the basis $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ is

$$B = \begin{pmatrix} 1 & 1/2 & -1 \\ 1/2 & 0 & 3/2 \\ -1 & 3/2 & 0 \end{pmatrix} \overset{\heartsuit}{=} P^T AP.$$

Second change of basis: Now we can use Step 2 of the proof to clear the first row and column by modifying $\mathbf{b}_2$ and $\mathbf{b}_3$, "completing the square". As specified in the Step 2 of the proof, we introduce a new basis $\mathbf{b}'$

$$\mathbf{b}'_1 := \mathbf{b}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \ \mathbf{b}'_2 := \mathbf{b}_2 - \frac{1}{2}\mathbf{b}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1/2 \\ 1/2 \\ 0 \end{pmatrix},$$

$$\mathbf{b}'_3 := \mathbf{b}_3 - (-1)\mathbf{b}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

So the basis change matrix from $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ to $\mathbf{b}'_1, \mathbf{b}'_2, \mathbf{b}'_3$ is

$$P' = \begin{pmatrix} 1 & -1/2 & 1 \\ 1 & 1/2 & 1 \\ 0 & 0 & 1 \end{pmatrix} \overset{\heartsuit}{=} PQ \ \text{ where } \ Q = \begin{pmatrix} 1 & -1/2 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

This corresponds to writing

$$\begin{aligned} \left[x_1^2 + x_1y_1 - 2x_1z_1\right] + 3y_1z_1 &= \left[(x_1 + \frac{1}{2}y_1 - z_1)^2 - \frac{1}{4}y_1^2 - z_1^2 + y_1z_1\right] + 3y_1z_1 \\ &= (x_1 + \frac{1}{2}y_1 - z_1)^2 - \frac{1}{4}y_1^2 + 4y_1z_1 - z_1^2. \end{aligned}$$

In the new basis $x_2\mathbf{b}'_1 + y_2\mathbf{b}'_2 + z_2\mathbf{b}'_3 = (x_2 - \frac{1}{2}y_2 + z_2)\mathbf{b}_1 + y_2\mathbf{b}_2 + z_2\mathbf{b}_3$, which tells us that

$$q(x_2\mathbf{b}'_1 + y_2\mathbf{b}'_2 + z_2\mathbf{b}'_3) = x_2^2 - \frac{1}{4}y_2^2 + 4y_2z_2 - z_2^2.$$

so the matrix of q with respect to the $\mathbf{b}'$ basis is

$$B' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1/4 & 2 \\ 0 & 2 & -1 \end{pmatrix} \overset{\heartsuit}{=} Q^T BQ \overset{\heartsuit}{=} (P')^T AP'.$$

Third change of basis: Now we are in Step 3 of the proof, concentrating on the bottom right 2×2 block. It is an induction step. We must change the second and third basis vectors. Any subsequent changes of basis we make will keep the first basis vector unchanged. We have

$$q(y_2\mathbf{b}'_2 + z_2\mathbf{b}'_3) = -\frac{1}{4}y_2^2 + 4y_2z_2 - z_2^2,$$

the "leftover terms" of the bottom right corner. This is a 2-variable quadratic form.

Since $q(\mathbf{b}_2') = -1/4 \neq 0$, we don't need to do anything for Step 1 of the proof. Using Step 2 of the proof, we replace $\mathbf{b}_1', \mathbf{b}_2', \mathbf{b}_3'$ by another new basis $\mathbf{b}''$:

$$\mathbf{b}_1'' := \mathbf{b}_1',\ \mathbf{b}_2'' := \mathbf{b}_2',\ \mathbf{b}_3'' := \mathbf{b}_3' - \frac{2}{-1/4}\mathbf{b}_2' = \begin{pmatrix}1\\1\\1\end{pmatrix} + 8\begin{pmatrix}-1/2\\1/2\\0\end{pmatrix} = \begin{pmatrix}-3\\5\\1\end{pmatrix}.$$

So the change of basis matrix from $\mathbf{e}$ to $\mathbf{b}''$ is

$$P'' = \begin{pmatrix}1 & -1/2 & -3\\1 & 1/2 & 5\\0 & 0 & 1\end{pmatrix} \overset{\heartsuit}{=} P'Q' \text{ where } Q' = \begin{pmatrix}1 & 0 & 0\\0 & 1 & 8\\0 & 0 & 1\end{pmatrix}.$$

This corresponds, of course, to the completing-the-square operation

$$-\frac{1}{4}y_2^2 + 4y_2z_2 - z_2^2 = -\frac{1}{4}(y_2 - 8z_2)^2 + 15z_2^2.$$

So the matrix of q is now

$$B'' = \begin{pmatrix}1 & 0 & 0\\0 & -1/4 & 0\\0 & 0 & 15\end{pmatrix} \overset{\heartsuit}{=} (Q')^T B' Q' \overset{\heartsuit}{=} (P'')^T A P''.$$

This is diagonal, so we're done: the matrix of q in the basis $\mathbf{b}_1'', \mathbf{b}_2'', \mathbf{b}_3''$ is the diagonal matrix B''.

 Notice that the choice of "good" basis, and the resulting "good" matrix, are extremely far from unique. For instance, in the example above we could have replaced $\mathbf{b}_2''$ with $2\mathbf{b}_2''$ to get the (perhaps nicer) matrix

$$\begin{pmatrix}1 & 0 & 0\\0 & -1 & 0\\0 & 0 & 15\end{pmatrix}.$$

In the case $K = \mathbb{C}$, we can do even better. After reducing q to the form $q(\mathbf{v}) = \sum_{i=1}^n \alpha_{ii}x_i^2$, we can permute the coordinates if necessary to get $\alpha_{ii} \neq 0$ for $1 \leq i \leq r$ and $\alpha_{ii} = 0$ for $r+1 \leq i \leq n$, where $r = \text{rank}(q)$. We can then make a further coordinates change $x_i' = \sqrt{\alpha_{ii}}x_i$ $(1 \leq i \leq r)$, giving $q(\mathbf{v}) = \sum_{i=1}^r (x_i')^2$. Hence we have proved:

Proposition 4.4.2. *A quadratic form q over $\mathbb{C}$ has the form $q(\mathbf{v}) = \sum_{i=1}^r x_i^2$ with respect to a suitable basis, where $r = \text{rank}(q)$.*

Equivalently, given a symmetric matrix $A \in \mathbb{C}^{n,n}$, there is an invertible matrix $P \in \mathbb{C}^{n,n}$ such that $P^{\mathrm{T}}AP = B$, where $B = (\beta_{ij})$ is a diagonal matrix with $\beta_{ii} = 1$ for $1 \leq i \leq r$, $\beta_{ii} = 0$ for $r+1 \leq i \leq n$, and $r = \text{rank}(A)$.

In particular, up to changes of basis, a quadratic form on $\mathbb{C}^n$ is uniquely determined by its rank. We say the rank is the only *invariant* of a quadratic form over $\mathbb{C}$.

When $K = \mathbb{R}$, we cannot take square roots of negative numbers, but we can replace each positive α_i by 1 and each negative α_i by -1 to get:

Proposition 4.4.3 (Sylvester's Theorem)**.** *A quadratic form q over $\mathbb{R}$ has the form $q(\mathbf{v}) = \sum_{i=1}^{t} x_i^2 - \sum_{i=1}^{u} x_{t+i}^2$ with respect to a suitable basis, where $t + u = \mathrm{rank}(q)$.*

Equivalently, given a symmetric matrix $A \in \mathbb{R}^{n,n}$, there is an invertible matrix $P \in \mathbb{R}^{n,n}$ such that $P^{\mathrm{T}}AP = B$, where $B = (\beta_{ij})$ is a diagonal matrix with $\beta_{ii} = 1$ for $1 \leq i \leq t$, $\beta_{ii} = -1$ for $t+1 \leq i \leq t+u$, and $\beta_{ii} = 0$ for $t+u+1 \leq i \leq n$, and $t + u = \mathrm{rank}(A)$.

We shall now prove that the numbers t and u of positive and negative terms are invariants of q. The pair of integers (t, u) is called the *signature* of q.

Theorem 4.4.4 (Sylvester's Law of Inertia)**.** *Suppose that q is a quadratic form on the vector space V over $\mathbb{R}$, and that $\mathbf{e}_1, \ldots, \mathbf{e}_n$ and $\mathbf{e}'_1, \ldots, \mathbf{e}'_n$ are two bases of V such that*

$$q(x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n) = \sum_{i=1}^{t} x_i^2 - \sum_{i=1}^{u} x_{t+i}^2$$

and

$$q(x_1\mathbf{e}'_1 + \cdots + x_n\mathbf{e}'_n) = \sum_{i=1}^{t'} x_i^2 - \sum_{i=1}^{u'} x_{t'+i}^2.$$

Then $t = t'$ and $u = u'$.

Proof. We know that $t + u = t' + u' = \mathrm{rank}(q)$, so it is enough to prove that $t = t'$. Suppose not; by symmetry we may suppose that $t > t'$.

Let V_1 be the span of $\mathbf{e}_1, \ldots, \mathbf{e}_t$, and let V_2 be the span of $\mathbf{e}'_{t'+1}, \ldots, \mathbf{e}'_n$. Then for any non-zero $\mathbf{v} \in V_1$ we have $q(\mathbf{v}) > 0$; while for any $\mathbf{w} \in V_2$ we have $q(\mathbf{w}) \leq 0$. So there cannot be any non-zero $\mathbf{v} \in V_1 \cap V_2$.

On the other hand, we have $\dim(V_1) = t$ and $\dim(V_2) = n - t'$. It was proved in MA106 that

$$\dim(V_1 + V_2) = \dim(V_1) + \dim(V_2) - \dim(V_1 \cap V_2),$$

so

$$\dim(V_1 \cap V_2) = t + (n - t') - \dim(V_1 + V_2) = (t - t') + n - \dim(V_1 + V_2) > 0.$$

The last inequality follows from our assumption on $t - t'$ and the fact $V_1 + V_2$ is a subspace of V and thus had dimensions at most n. Since we have shown that $V_1 \cap V_2 = \{0\}$, this is a contradiction, which completes the proof. □

Remark. *Notice that any non-zero $x \in \mathbb{R}$ is either equal to a square, or -1 times a square, but not both. This property is shared by the finite field $\mathbb{F}_7$ of integers mod 7, so any quadratic form over $\mathbb{F}_7$ can be*

written as a diagonal matrix with only 0's, 1's and -1's down the diagonal (i.e. Sylvester's Theorem holds over $\mathbb{F}_7$). But Sylvester's law of inertia isn't valid in $\mathbb{F}_7$: in fact, we have

$$\begin{pmatrix} 2 & 3 \\ 4 & 2 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 3 \\ 4 & 2 \end{pmatrix} = \begin{pmatrix} 20 & 14 \\ 14 & 20 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix},$$

so the same form has signature $(2,0)$ and $(0,2)$! The proof breaks down because there's no good notion of a "positive" element of $\mathbb{F}_7$, so a sum of non-zero squares can be zero (the easiest example is $1^2+2^2+3^2=0$). So Sylvester's law of inertia is really using something quite special about $\mathbb{R}$.

4.5 Euclidean spaces, orthonormal bases and the Gram–Schmidt process

In this section, we're going to suppose $K = \mathbb{R}$. As usual, we let V be an n-dimensional vector space over K, and we let q be a quadratic form on V, with associated bilinear form τ.

Definition 4.5.1. The quadratic form q is said to be *positive definite* if $q(\mathbf{v}) > 0$ for all $0 \neq \mathbf{v} \in V$.

It is clear that this is the case if and only if $t = n$ and $u = 0$ in Proposition 4.4.3; that is, if q has signature $(n, 0)$.

The associated symmetric bilinear form τ is also called positive definite when q is.

Definition 4.5.2. A vector space V over $\mathbb{R}$ together with a positive definite symmetric bilinear form τ is called a *euclidean space*.

In this case, Proposition 4.4.3 says that there is a basis $\{\mathbf{e}_i\}$ of V with respect to which $\tau(\mathbf{e}_i, \mathbf{e}_j) = \delta_{ij}$, where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

(so the matrix A of q is the identity matrix I_n.) We call a basis of a euclidean space V with this property an *orthonormal basis* of V.

(More generally, any set $\mathbf{v}_1, \ldots, \mathbf{v}_r$ of vectors in V, not necessarily a basis, will be said to be *orthonormal* if $\tau(\mathbf{v}_i, \mathbf{v}_j) = \delta_{ij}$ for $1 \leq i, j \leq r$.)

We shall assume from now on that V is a euclidean space, and that we have chosen an orthonormal basis basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$. Then τ corresponds to the standard scalar product, we shall write $\mathbf{v} \cdot \mathbf{w}$ instead of $\tau(\mathbf{v}, \mathbf{w})$.

Note that $\mathbf{v} \cdot \mathbf{w} = \underline{\mathbf{v}}^{\mathrm{T}} \underline{\mathbf{w}}$ where, as usual, $\underline{\mathbf{v}}$ and $\underline{\mathbf{w}}$ are the column vectors associated with $\mathbf{v}$ and $\mathbf{w}$.

For $\mathbf{v} \in V$, define $|\mathbf{v}| = \sqrt{\mathbf{v} \cdot \mathbf{v}}$. Then $|\mathbf{v}|$ is the length of $\mathbf{v}$. Hence the length, and also the cosine $\mathbf{v} \cdot \mathbf{w} / (|\mathbf{v}||\mathbf{w}|)$ of the angle between two vectors can be defined in terms of the scalar product. Thus a set of vectors is orthonormal if the vectors all have length 1 and are at right angles to each other.

The following theorem tells us that we can always complete a set of orthonormal vectors to an orthonormal basis.

Theorem 4.5.3 (Gram-Schmidt process)**.** *Let V be a euclidean space of dimension n, and suppose that, for some r with $0 \leq r \leq n$, $\mathbf{f}_1, \ldots, \mathbf{f}_r$ are vectors in V such that*

$$\mathbf{f}_i \cdot \mathbf{f}_j = \delta_{ij} \quad \textit{for} \quad 1 \leq i, j \leq r. \tag{$*$}$$

Then $\mathbf{f}_1, \ldots, \mathbf{f}_r$ can be extended to an orthonormal basis $\mathbf{f}_1, \ldots, \mathbf{f}_n$ of V.

Proof. We prove first that $\mathbf{f}_1, \ldots, \mathbf{f}_r$ are linearly independent. Suppose that $\sum_{i=1}^{r} x_i \mathbf{f}_i = \mathbf{0}$ for some $x_1, \ldots, x_r \in \mathbb{R}$. Then, for each j with $1 \leq j \leq r$, the scalar product of the left hand side of this equation with $\mathbf{f}_j$ is $\sum_{i=1}^{r} x_i \mathbf{f}_j \cdot \mathbf{f}_i = x_j$, by $(*)$. Since $\mathbf{f}_j \cdot \mathbf{0} = 0$, this implies that $x_j = 0$ for all j, so the $\mathbf{f}_i$ are linearly independent.

The proof of the theorem will be by induction on $n - r$. We can start the induction with the case $n - r = 0$, when $r = n$, and there is nothing to prove. So assume that $n - r > 0$; i.e. that $r < n$. By a result from MA106, we can extend any linearly independent set of vectors to a basis of V, so there is a basis $\mathbf{f}_1, \ldots, \mathbf{f}_r, \mathbf{g}_{r+1}, \ldots, \mathbf{g}_n$ of V containing the $\mathbf{f}_i$. The trick is to define

$$\mathbf{f}'_{r+1} = \mathbf{g}_{r+1} - \sum_{i=1}^{r} (\mathbf{f}_i \cdot \mathbf{g}_{r+1}) \mathbf{f}_i.$$

If we take the scalar product of this equation by $\mathbf{f}_j$ for some $1 \leq j \leq r$, then we get

$$\mathbf{f}_j \cdot \mathbf{f}'_{r+1} = \mathbf{f}_j \cdot \mathbf{g}_{r+1} - \sum_{i=1}^{r} (\mathbf{f}_i \cdot \mathbf{g}_{r+1})(\mathbf{f}_j \cdot \mathbf{f}_i)$$

and then, by $(*)$, $\mathbf{f}_j \cdot \mathbf{f}_i$ is non-zero only when $j = i$, so the sum on the right hand side simplifies to $\mathbf{f}_j \cdot \mathbf{g}_{r+1}$, and the whole equation simplifies to $\mathbf{f}_j \cdot \mathbf{f}'_{r+1} = \mathbf{f}_j \cdot \mathbf{g}_{r+1} - \mathbf{f}_j \cdot \mathbf{g}_{r+1} = 0$.

The vector $\mathbf{f}'_{r+1}$ is non-zero by linear independence of the basis, and if we define $\mathbf{f}_{r+1} = \mathbf{f}'_{r+1} / |\mathbf{f}'_{r+1}|$, then we still have $\mathbf{f}_j \cdot \mathbf{f}_{r+1} = 0$ for $1 \leq j \leq r$, and we also have $\mathbf{f}_{r+1} . \mathbf{f}_{r+1} = 1$. Hence $\mathbf{f}_1, \ldots, \mathbf{f}_{r+1}$ satisfy the equations $(*)$, and the result follows by inductive hypothesis. □

4.6 Orthogonal transformations

Week 6

If we're working with a euclidean space V, we know what the "length" of a vector in V means, and what the "angle" between vectors is; so we might want to consider transformations from V to itself that preserve lengths and angles – they play nicely with the geometry of the space.

Definition 4.6.1. A linear map $T: V \to V$ is said to be *orthogonal* if it preserves the scalar product on V. That is, if $T(\mathbf{v}) \cdot T(\mathbf{w}) = \mathbf{v} \cdot \mathbf{w}$ for all $\mathbf{v}, \mathbf{w} \in V$.

Since length and angle can be defined in terms of the scalar product, an orthogonal linear map preserves distance and angle, so geometrically it is a rigid map. In $\mathbb{R}^2$, for example, an orthogonal map is either rotation about the origin, or a reflection about a line through the origin.

If A is the matrix of T, then $T(\mathbf{v}) = A\underline{\mathbf{v}}$, so $T(\mathbf{v}) \cdot T(\mathbf{w}) = \underline{\mathbf{v}}^{\mathrm{T}} A^{\mathrm{T}} A \underline{\mathbf{w}}$, and hence T is orthogonal if and only if $A^{\mathrm{T}}A = I_n$, or equivalently if $A^{\mathrm{T}} = A^{-1}$.

Definition 4.6.2. An $n \times n$ matrix is called *orthogonal* if $A^{\mathrm{T}}A = I_n$.

So we have proved:

Proposition 4.6.3. *A linear map $T : V \to V$ is orthogonal if and only if its matrix A (with respect to an orthonormal basis of V) is orthogonal.*

Incidentally, the fact that $A^{\mathrm{T}}A = I_n$ tells us that A (and hence T) is invertible, so $\det(A)$ is non-zero. In fact we can do a little better than that:

Proposition 4.6.4. *An orthogonal matrix has determinant ± 1.*

Proof. We have $A^{\mathrm{T}}A = I_n$, so $\det(A^{\mathrm{T}}A) = \det(I_n) = 1$.

On the other hand, $\det(A^{\mathrm{T}}A) = \det(A^{\mathrm{T}}) \det(A) = (\det A)^2$. So $(\det A)^2 = 1$, implying that $\det A = \pm 1$. □

Example. For any $\theta \in \mathbb{R}$, let $A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$. (This represents a counter-clockwise rotation through an angle θ.) Then it is easily checked that $A^{\mathrm{T}}A = AA^{\mathrm{T}} = I_2$.

One can check that every orthogonal 2×2 matrix with determinant $+1$ is a rotation by some angle θ, and similarly that any orthogonal 2×2 matrix of det -1 is a reflection in some line through the origin. In higher dimensions the taxonomy of orthogonal matrices is a bit more complicated – we'll revisit this in a later section of the course.

Notice that the columns of A are mutually orthogonal vectors of length 1, and the same applies to the rows of A. Let $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n$ be the columns of the matrix A. As we observed in §1, $\mathbf{c}_i$ is equal to the column vector representing $T(\mathbf{e}_i)$. In other words, if $T(\mathbf{e}_i) = \mathbf{f}_i$, say, then $\underline{\mathbf{f}_i} = \mathbf{c}_i$.

Since the (i, j)-th entry of $A^{\mathrm{T}}A$ is $\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_j = \mathbf{f}_i \cdot \mathbf{f}_j$, we see that T and A are orthogonal if and only if

$$\mathbf{f}_i \cdot \mathbf{f}_i = 1 \text{ and } \mathbf{f}_i \cdot \mathbf{f}_j = 0 \ (i \neq j), \quad 1 \leq i, j \leq n. \tag{$*$}$$

By Proposition 4.6.4, an orthogonal linear map is invertible, so $T(\mathbf{e}_i)$ $(1 \leq i \leq n)$ form a basis of V, and we have:

Proposition 4.6.5. *A linear map T is orthogonal if and only if $T(\mathbf{e}_1), \ldots, T(\mathbf{e}_n)$ is an orthonormal basis of V.*

Here's a pretty application of the Gram-Schmidt process and orthogonal matrices. Notice that our proof of Gram-Schmidt actually proved a little more: we showed that if $\mathbf{f}_1, \ldots, \mathbf{f}_r$ is an orthonormal set, and $\mathbf{g}_{r+1}, \ldots, \mathbf{g}_n$ is any way of completing the $\mathbf{f}$'s to a basis of V, then we can find $\mathbf{f}_{r+1}, \ldots, \mathbf{f}_n$ such that

- $\mathbf{f}_1, \ldots, \mathbf{f}_n$ is an orthonormal basis,
- for each $r+1 \leq i \leq n$, f_i is in the linear span of $\mathbf{f}_1, \ldots, \mathbf{f}_r, \mathbf{g}_{r+1}, \ldots, \mathbf{g}_i$.

That is, we've arranged that the basis change matrix from $\mathbf{f}_1, \ldots, \mathbf{f}_r, \mathbf{g}_{r+1}, \ldots, \mathbf{g}_n$ to $\mathbf{f}_1, \ldots, \mathbf{f}_n$ looks like

$$\left(\begin{array}{c|c} I_r & A \\ \hline 0 & B \end{array}\right)$$

with B *upper-triangular*. This is most useful when $r = 0$, when it says that any basis may be modified by an upper-triangular matrix to make it orthonormal.

This slight modification of Gram-Schmidt has a very nice interpretation in terms of matrices:

Proposition 4.6.6 (QR decomposition)**.** *Let A be any $n \times n$ real matrix. Then we can write $A = QR$ where Q is orthogonal and R is upper-triangular.*

Proof. We'll make a simplifying assumption: let's suppose A is invertible.

Let $\mathbf{g}_1, \ldots, \mathbf{g}_n$ be the columns of A, regarded as vectors in $\mathbb{R}^n$. Then since A is invertible, $\mathbf{g}_1, \ldots, \mathbf{g}_n$ is a basis of $\mathbb{R}^n$. We apply Gram-Schmidt orthonormalization to construct an orthonormal basis $\mathbf{f}_1, \ldots, \mathbf{f}_n$ such that $\mathbf{f}_i$ is in the linear span of $\mathbf{g}_1, \ldots, \mathbf{g}_i$ for each i.

Let Q be the matrix whose columns are $\mathbf{f}_1, \ldots, \mathbf{f}_n$. Then Q is an orthogonal matrix, since its columns are orthonormal vectors; and there are real numbers r_{ij} such that

$$\begin{aligned} \mathbf{g}_1 &= r_{11}\mathbf{f}_1 \\ \mathbf{g}_2 &= r_{12}\mathbf{f}_1 + r_{22}\mathbf{f}_2 \\ \mathbf{g}_3 &= r_{13}\mathbf{f}_1 + r_{23}\mathbf{f}_2 + r_{33}\mathbf{f}_3 \\ &\vdots \end{aligned}$$

In other words we have

$$A = QR$$

where R is the upper-triangular matrix with entries r_{ij}. □

(When A isn't invertible, then the columns of A won't be a basis. But it's not hard to show that any matrix A can be written as $A = BR'$ where B is invertible and R' is upper triangular; then writing $B = QR$ we have $A = QRR'$, and RR' is also upper-triangular.)

Example. Consider the matrix

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

We have $\det(A) = 10$, so A is non-singular. Let $\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3$ be the columns of A, as above. Then $|\mathbf{g}_1| = \sqrt{5}$, so

$$\mathbf{f}_1 = \frac{\mathbf{g}_1}{\sqrt{5}} = \begin{pmatrix} -1/\sqrt{5} \\ 2/\sqrt{5} \\ 0 \end{pmatrix}.$$

For the next step, we take $\mathbf{f}_2' = \mathbf{g}_2 - (\mathbf{f}_1 \cdot \mathbf{g}_2)\mathbf{f}_1 = \mathbf{g}_2$, since $\mathbf{f}_1 \cdot \mathbf{g}_2 = 0$. So

$$\mathbf{f}_2 = \frac{\mathbf{g}_2}{|\mathbf{g}_2|} = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}.$$

For the final step, we take the vector

$$\mathbf{f}_3' = \mathbf{g}_3 - (\mathbf{f}_1 \cdot \mathbf{g}_3)\mathbf{f}_1 - (\mathbf{f}_2 \cdot \mathbf{g}_3)\mathbf{f}_2.$$

We have

$$\mathbf{f}_1 \cdot \mathbf{g}_3 = \begin{pmatrix} -1/\sqrt{5} \\ 2/\sqrt{5} \\ 0 \end{pmatrix} \cdot \begin{pmatrix} -2 \\ -1 \\ -2 \end{pmatrix} = 0, \quad \mathbf{f}_2 \cdot \mathbf{g}_3 = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} \cdot \begin{pmatrix} -2 \\ -1 \\ -2 \end{pmatrix} = 2.$$

So $\mathbf{f}_3' = \mathbf{g}_3 - 2\mathbf{f}_2 = \begin{pmatrix} -2 \\ -1 \\ 0 \end{pmatrix}$. We have $|\mathbf{f}_3'| = \sqrt{5}$ again, so

$$\mathbf{f}_3 = \frac{\mathbf{f}_3'}{\sqrt{5}} = \begin{pmatrix} -2/\sqrt{5} \\ -1/\sqrt{5} \\ 0 \end{pmatrix}.$$

Thus Q is the matrix whose columns are $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$, that is

$$Q = \begin{pmatrix} -1/\sqrt{5} & 0 & -2/\sqrt{5} \\ 2/\sqrt{5} & 0 & -1/\sqrt{5} \\ 0 & -1 & 0 \end{pmatrix}.$$

and we have

$$\mathbf{g}_1 = \sqrt{5}\mathbf{f}_1, \quad \mathbf{g}_2 = 2\mathbf{f}_2, \quad \mathbf{g}_3 = 2\mathbf{f}_2 + \sqrt{5}\mathbf{f}_3$$

so $A = QR$ where

$$R = \begin{pmatrix} \sqrt{5} & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & \sqrt{5} \end{pmatrix}.$$

The QR decomposition theorem is a very important technique in numerical calculations with matrices. For instance, QR decomposition gives a quick way of inverting matrices. If $A = QR$, then $A^{-1} = R^{-1}Q^{-1}$. Inverting orthogonal matrices is trivial, as the inverse is just the transpose; inverting upper-triangular matrices is also pretty easy, so we can compute the inverse of A this way, without having to compute the determinant.

4.7 Nice orthonormal bases

Now what if we have a euclidean space V, and a linear operator $T : V \to V$, or a quadratic form q on V (not necessarily the same as the "native" length form on V)? Can we always find an orthonormal basis of V making the matrix of q look reasonably nice? Notice that we're juggling two quadratic forms here – we're trying to make the matrix of q look nice while simultaneously keeping the matrix of the length form being the identity.

Oddly enough, this is *also* a question about linear operators. Given any bilinear form τ on V (not necessarily symmetric), there's a uniquely determined linear operator T on V such that

$$\tau(\mathbf{v}, \mathbf{w}) = \mathbf{v} \cdot (T\mathbf{w}).$$

Conversely, any T determines a τ by the same formula. So once we've fixed a "starting" bilinear form (the inner product), we can get any other bilinear form τ from this via a linear operator, and this gives us a bijection between bilinear forms and linear operators once we've fixed our "starting point". Moreover, the matrix of T, in an orthonormal basis of V, is clearly just the matrix of τ. When we change basis by an orthogonal matrix (to get a new orthonormal basis), the matrix of T changes by $A \mapsto P^{-1}AP$, and the matrix of τ changes by $A \mapsto P^{\mathrm{T}}AP$, but this is OK since $P^{\mathrm{T}} = P^{-1}$ for orthogonal matrices!

In particular, if T is any linear operator, then $(\mathbf{v}, \mathbf{w}) \mapsto (T\mathbf{v}) \cdot \mathbf{w}$ is certainly a bilinear form; so there must be some linear operator S such that

$$(T\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot (S\mathbf{w}) \qquad (*)$$

for all $\mathbf{v}$ and $\mathbf{w}$.

Definition 4.7.1. If $T : V \to V$ is a linear operator on a euclidean space V, then the unique linear map S such that $(*)$ holds is called the *adjoint* of T. We write this as T^*.

It's easy to see that in an orthonormal basis, the matrix of T^* is just the transpose of the matrix of T. It follows from this that a linear operator is orthogonal if and only if $T^* = T^{-1}$; one can also prove this directly from the definition.

We say T is *selfadjoint* if $T^* = T$, or equivalently if the bilinear form $\tau(\mathbf{v}, \mathbf{w}) = \mathbf{v} \cdot (T\mathbf{w})$ is symmetric. Notice that selfadjointness, like orthogonal-ness, is something that only makes sense for linear operators on euclidean spaces; it doesn't make sense to ask if a linear operator on a general vector space is selfadjoint. It should be clear that T is selfadjoint if and only if its matrix in an orthonormal basis of V is a symmetric matrix.

So if V is a euclidean space of dimension n, the following problems are all actually the same:

- given a quadratic form q on V, find an orthonormal basis of V making the matrix of q as nice as possible;
- given a selfadjoint linear operator T on V, find an orthonormal basis of V making the matrix of T as nice as possible;

- given an $n \times n$ symmetric real matrix A, find an orthogonal matrix P such that $P^{\mathrm{T}}AP$ is as nice as possible.

First, we'll warm up by proving a proposition which we'll need in the main proof:

Proposition 4.7.2. *Let A be an $n \times n$ real symmetric matrix. Then A has an eigenvalue in $\mathbb{R}$, and all complex eigenvalues of A lie in $\mathbb{R}$.*

Proof. (To simplify the notation, we will write just $\mathbf{v}$ for a column vector $\underline{\mathbf{v}}$ in this proof.)

The characteristic equation $\det(A - xI_n) = 0$ is a polynomial equation of degree n in x, and since $\mathbb{C}$ is an algebraically closed field, it certainly has a root $\lambda \in \mathbb{C}$, which is an eigenvalue for A if we regard A as a matrix over $\mathbb{C}$. We shall prove that any such λ lies in $\mathbb{R}$, which will prove the proposition.

For a column vector $\mathbf{v}$ or matrix B over $\mathbb{C}$, we denote by $\overline{\mathbf{v}}$ or $\overline{B}$ the result of replacing all entries of $\mathbf{v}$ or B by their complex conjugates. Since the entries of A lie in R, we have $\overline{A} = A$.

Let $\mathbf{v}$ be a complex eigenvector associated with λ. Then

$$A\mathbf{v} = \lambda\mathbf{v} \tag{1}$$

so,taking complex conjugates and using $\overline{A} = A$, we get

$$A\overline{\mathbf{v}} = \overline{\lambda}\overline{\mathbf{v}}. \tag{2}$$

Transposing (1) and using $A^{\mathrm{T}} = A$ gives

$$\mathbf{v}^{\mathrm{T}}A = \lambda\mathbf{v}^{\mathrm{T}}, \tag{3}$$

so by (2) and (3) we have

$$\lambda\mathbf{v}^{\mathrm{T}}\overline{\mathbf{v}} = \mathbf{v}^{\mathrm{T}}A\overline{\mathbf{v}} = \overline{\lambda}\mathbf{v}^{\mathrm{T}}\overline{\mathbf{v}}.$$

But if $\mathbf{v} = (\alpha_1, \alpha_2, \ldots, \alpha_n)^{\mathrm{T}}$, then $\mathbf{v}^{\mathrm{T}}\overline{\mathbf{v}} = \alpha_1\overline{\alpha_1} + \cdots + \alpha_n\overline{\alpha_n}$, which is a non-zero real number (eigenvectors are non-zero by definition). Thus $\lambda = \overline{\lambda}$, so $\lambda \in \mathbb{R}$. □

Now let's prove the main theorem.

Theorem 4.7.3. *Let V be a euclidean space of dimension n. Then:*

- *Given any quadratic form q on V, there is an orthonormal basis $\mathbf{f}_1, \ldots, \mathbf{f}_n$ of V and constants $\alpha_1, \ldots, \alpha_n$, uniquely determined up to reordering, such that*

$$q(x_1\mathbf{f}_1 + \cdots + x_n\mathbf{f}_n) = \sum_{i=1}^{n} \alpha_i (x_i)^2$$

  for all $x_1, \ldots, x_n \in \mathbb{R}$.

- *Given any linear operator $T : V \to V$ which is selfadjoint, there is an orthonormal basis $\mathbf{f}_1, \ldots, \mathbf{f}_n$ of V consisting of eigenvectors of T.*

- *Given any $n \times n$ real symmetric matrix A, there is an orthogonal matrix P such that $P^{\mathrm{T}}AP = P^{-1}AP$ is a diagonal matrix.*

Proof. We've already seen that these three statements are equivalent to each other, so we can prove whichever one of them we like. Notice that in the second and third forms of the statement, it's clear that the diagonal matrix we obtain is similar to the original one; that tells us that in the first statement the constants $\alpha_1, \ldots, \alpha_n$ are uniquely determined (possibly up to re-ordering).

We'll go for proving the second statement. We'll prove this by induction on $n = \dim V$. If $n = 0$ there is nothing to prove, so let's assume the proposition holds for $n - 1$.

Let T be our linear operator. By proposition 4.7.2, T has an eigenvalue in $\mathbb{R}$. Let $\mathbf{v}$ be a corresponding eigenvector in V. Then $\mathbf{f}_1 = \mathbf{v}/|\mathbf{v}|$ is also an eigenvector, and $|\mathbf{f}_1| = 1$. Let α_1 be the corresponding eigenvalue.

We consider the space $W = \{\mathbf{w} \in V : \mathbf{w} \cdot \mathbf{f}_1 = 0\}$. This is clearly a subspace of W of dimension $n - 1$. I claim that T maps W into itself. So suppose $\mathbf{w} \in W$; we want to show that $T(\mathbf{w}) \in W$ also.

We have

$$T(\mathbf{w}) \cdot \mathbf{f}_1 = \mathbf{w} \cdot T(\mathbf{f}_1)$$

since T is selfadjoint. But we know that $T(\mathbf{f}_1) = \alpha_1 \mathbf{f}_1$, so it follows that

$$T(\mathbf{w}) \cdot \mathbf{f}_1 = \alpha_1(\mathbf{w} \cdot \mathbf{f}_1) = 0,$$

since $\mathbf{w} \in W$ so $\mathbf{w} \cdot \mathbf{f}_1 = 0$ by hypothesis.

So T maps W into itself. Moreover, W is a euclidean space of dimension $n - 1$, so we may apply the induction hypothesis to the restriction of T to W. This gives us an orthonormal basis $\mathbf{f}_2, \ldots, \mathbf{f}_n$ of W consisting of eigenvectors of T. Then the set $\mathbf{f}_1, \ldots, \mathbf{f}_n$ is an orthonormal basis of V, which also consists of eigenvectors of T. □

Although it is not used in the proof of the theorem above, the following proposition is useful when calculating examples. It helps us to write down more vectors in the final orthonormal basis immediately, without having to use Theorem 4.5.3 repeatedly.

Proposition 4.7.4. *Let A be a real symmetric matrix, and let λ_1, λ_2 be two distinct eigenvalues of A, with corresponding eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$. Then $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$.*

Proof. (As in Proposition 4.7.2, we will write $\mathbf{v}$ rather than $\underline{\mathbf{v}}$ for a column vector in this proof. So $\mathbf{v}_1 \cdot \mathbf{v}_2$ is the same as $\mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2$.) We have

$$A\mathbf{v}_1 = \lambda_1 \mathbf{v}_1, \tag{1}$$
$$A\mathbf{v}_2 = \lambda_2 \mathbf{v}_2. \tag{2}$$

The trick is now to look at the expression $\mathbf{v}_1^{\mathrm{T}} A \mathbf{v}_2$. On the one hand, by (2) we have

$$\mathbf{v}_1^{\mathrm{T}} A \mathbf{v}_2 = \mathbf{v}_1 \cdot (A\mathbf{v}_2) = \mathbf{v}_1^{\mathrm{T}}(\lambda_2 \mathbf{v}_2) = \lambda_2(\mathbf{v}_1 \cdot \mathbf{v}_2). \tag{3}$$

On the other hand, $A^{\mathrm{T}} = A$, so $\mathbf{v}_1^{\mathrm{T}} A = \mathbf{v}_1^{\mathrm{T}} A^{\mathrm{T}} = (A\mathbf{v}_1)^{\mathrm{T}}$, so using (1) we have

$$\mathbf{v}_1^{\mathrm{T}} A \mathbf{v}_2 = (A\mathbf{v}_1)^{\mathrm{T}} \mathbf{v}_2 = (\lambda_1 \mathbf{v}_1^{\mathrm{T}}) \mathbf{v}_2 = \lambda_1 (\mathbf{v}_1 \cdot \mathbf{v}_2). \tag{4}$$

Comparing (3) and (4), we have $(\lambda_2 - \lambda_1)(\mathbf{v}_1 \cdot \mathbf{v}_2) = 0$. Since $\lambda_2 - \lambda_1 \neq 0$ by assumption, we have $\mathbf{v}_1^{\mathrm{T}} \mathbf{v}_2 = 0$. □

Example 15. Let $n = 2$ and let A be the symmetric matrix $A = \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}$. Then

$$\det(A - xI_2) = (1-x)^2 - 9 = x^2 - 2x - 8 = (x-4)(x+2),$$

so the eigenvalues of A are 4 and -2. Solving $A\mathbf{v} = \lambda\mathbf{v}$ for $\lambda = 4$ and -2, we find corresponding eigenvectors $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Proposition 4.7.4 tells us that these vectors are orthogonal to each other (which we can of course check directly!), so if we divide them by their lengths to give vectors of length 1, giving $\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$ and $\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix}$ then we get an orthonormal basis consisting of eigenvectors of A, which is what we want. The corresponding basis change matrix P has these vectors as columns, so $P = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix}$, and we can check that $P^{\mathrm{T}} P = I_2$ (i.e. P is orthogonal) and that $P^{\mathrm{T}} A P = \begin{pmatrix} 4 & 0 \\ 0 & -2 \end{pmatrix}$.

Example 16. Let's do an example of the "quadratic form" version of the above theorem. Let $n = 3$ and

$$q(\mathbf{v}) = 3x^2 + 6y^2 + 3z^2 - 4xy - 4yz + 2xz,$$
$$\text{so} \quad A = \begin{pmatrix} 3 & -2 & 1 \\ -2 & 6 & -2 \\ 1 & -2 & 3 \end{pmatrix}.$$

Then, expanding by the first row,

$$\begin{aligned} \det(A - xI_3) &= (3-x)(6-x)(3-x) - 4(3-x) - 4(3-x) + 4 + 4 - (6-x) \\ &= -x^3 + 12x^2 - 36x + 32 = (2-x)(x-8)(x-2), \end{aligned}$$

so the eigenvalues are 2 (repeated) and 8. For the eigenvalue 8, if we solve $A\mathbf{v} = 8\mathbf{v}$ then we find a solution $\mathbf{v} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$. Since 2 is a repeated eigenvalue, we need two corresponding eigenvectors, which must be orthogonal to each other. The equations $A\mathbf{v} = 2\mathbf{v}$ all reduce to $a - 2b + c = 0$, and so any vector $\begin{pmatrix} a \\ b \\ c \end{pmatrix}$ satisfying this equation is an eigenvector for $\lambda = 2$. By

Proposition 4.7.4 these eigenvectors will all be orthogonal to the eigenvector for $\lambda = 8$, but we will have to choose them orthogonal to each other. We can choose the first one arbitrarily, so let's choose $\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$. We now need another solution that is orthogonal to this. In other words, we want a, b and c not all zero satisfying $a - 2b + c = 0$ and $a - c = 0$, and $a = b = c = 1$ is a solution. So we now have a basis $\begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$, $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ of three mutually orthogonal eigenvectors. To get an orthonormal basis, we just need to divide by their lengths, which are, respectively, $\sqrt{6}$, $\sqrt{2}$, and $\sqrt{3}$, and then the basis change matrix P has these vectors as columns, so

$$P = \begin{pmatrix} 1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \\ -2/\sqrt{6} & 0 & 1/\sqrt{3} \\ 1/\sqrt{6} & -1/\sqrt{2} & 1/\sqrt{3} \end{pmatrix}.$$

It can then be checked that $P^{\mathrm{T}}P = I_3$ and that $P^{\mathrm{T}}AP$ is the diagonal matrix with entries 8, 2, 2. So if $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$ is this basis, we have

$$q(x\mathbf{f}_1 + y\mathbf{f}_2 + z\mathbf{f}_3) = 8x^2 + 2z^2 + 2y^2.$$

4.8 Applications of quadratic forms to geometry

4.8.1 Reduction of the general second degree equation

The general equation of the second degree in n variables $x_1, \ldots, x_n$ is

$$\sum_{i=1}^{n} \alpha_i x_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{i-1} \alpha_{ij} x_i x_j + \sum_{i=1}^{n} \beta_i x_i + \gamma = 0. \qquad (\dagger)$$

For fixed values of the α's, β's and γ, this defines a *quadric* curve or surface in n-dimensional euclidean space. To study the possible shapes of the curves and surfaces thus defined, we first simplify this equation by applying coordinate changes resulting from isometries (rigid motions) of $\mathbb{R}^n$; that is, transformations that preserve distance and angle.

By Theorem 4.7.3, we can apply an orthogonal basis change (that is, an isometry of $\mathbb{R}^n$ that fixes the origin) which has the effect of eliminating the terms $\alpha_{ij}x_i x_j$ in the above sum.

Now, whenever $\alpha_i \neq 0$, we can replace x_i by $x_i - \beta_i/(2\alpha_i)$, and thereby eliminate the term $\beta_i x_i$ from the equation. This transformation is just a translation, which is also an isometry.

If $\alpha_i = 0$, then we cannot eliminate the term $\beta_i x_i$. Let us permute the coordinates such that $\alpha_i \neq 0$ for $1 \leq i \leq r$, and $\beta_i \neq 0$ for $r+1 \leq i \leq r+s$. Then if $s > 1$, by using Theorem 4.5.3, we can find an orthogonal transformation that leaves x_i unchanged for $1 \leq i \leq r$ and replaces $\sum_{i=1}^{s} \beta_{r+i} x_{r+i}$ by βx_{r+1} (where β is the length of $\sum_{i=1}^{s} \beta_{r+i} x_{r+i}$), and then we have at most one non-zero β_i; either there are no linear terms at all, or there is just β_{r+1}.

Finally, if there is a non-zero β_{r+1}, then we can perform the translation that replaces x_{r+1} by $x_{r+1} - \gamma/\beta_{r+1}$, and thereby eliminate γ. By dividing the equation through by a constant, we can assume that the constant term is 0 or 1.

We have proved the following theorem:

Theorem 4.8.1. *By rigid motions of euclidean space, we can transform the set defined by the general second degree equation (†) into the set defined by an equation having one of the following three forms:*

$$\sum_{i=1}^{r} \alpha_i x_i^2 = 0,$$

$$\sum_{i=1}^{r} \alpha_i x_i^2 + 1 = 0,$$

$$\sum_{i=1}^{r} \alpha_i x_i^2 + x_{r+1} = 0.$$

Here $0 \leq r \leq n$ and $\alpha_1, \ldots, \alpha_r$ are non-zero constants, and in the third case $r < n$.

In both cases, we shall assume that $r \neq 0$, because otherwise we have a linear equation.

The sets defined by the first two types of equation are called *central quadrics* because they have central symmetry; i.e. if a vector $\mathbf{v}$ satisfies the equation, then so does $-\mathbf{v}$.

We shall now consider the types of curves and surfaces that can arise in the familiar cases $n = 2$ and $n = 3$. These different types correspond to whether the α_i are positive, negative or zero, and whether $\gamma = 0$ or 1.

We shall use x, y, z instead of x_1, x_2, x_3, and α, β, γ instead of α_1, α_2, α_3. We shall assume also that α, β, γ are all strictly positive, and write $-\alpha$, etc., for the negative case.

4.8.2 The case $n = 2$

When $n = 2$ we have the following possibilities.

(i) $\alpha x^2 = 0$. This just defines the line $x = 0$ (the y-axis).

(ii) $\alpha x^2 = 1$. This defines the two parallel lines $x = \pm 1/\sqrt{\alpha}$.

(iii) $-\alpha x^2 = 1$. This is the empty set!

(iv) $\alpha x^2 + \beta y^2 = 0$. The single point $(0, 0)$.

(v) $\alpha x^2 - \beta y^2 = 0$. This defines two straight lines $y = \pm\sqrt{\alpha/\beta}\, x$, which intersect at $(0, 0)$.

(vi) $\alpha x^2 + \beta y^2 = 1$. An ellipse.

(vii) $\alpha x^2 - \beta y^2 = 1$. A hyperbola.

(viii) $-\alpha x^2 - \beta y^2 = 1$. The empty set again.

(ix) $\alpha x^2 - y = 0$. A parabola.

4.8.3 The case $n = 3$

When $n = 3$, we still get the nine possibilities (i) – (ix) that we had in the case $n = 2$, but now they must be regarded as equations in the three variables x, y, z that happen not to involve z.

So, in Case (i), we now get the plane $x = 0$, in case (ii) we get two parallel planes $x = \pm 1/\sqrt{\alpha}$, in Case (iv) we get the line $x = y = 0$ (the z-axis), in case (v) two intersecting planes $y = \pm\sqrt{\alpha/\beta}x$, and in Cases (vi), (vii) and (ix), we get, respectively, elliptical, hyperbolic and parabolic cylinders.

The remaining cases involve all of x, y and z. We omit $-\alpha x^2 - \beta y^2 - \gamma z^2 = 1$, which is empty.

(x) $\alpha x^2 + \beta y^2 + \gamma z^2 = 0$. The single point $(0, 0, 0)$.

(xi) $\alpha x^2 + \beta y^2 - \gamma z^2 = 0$. See Fig. 1.

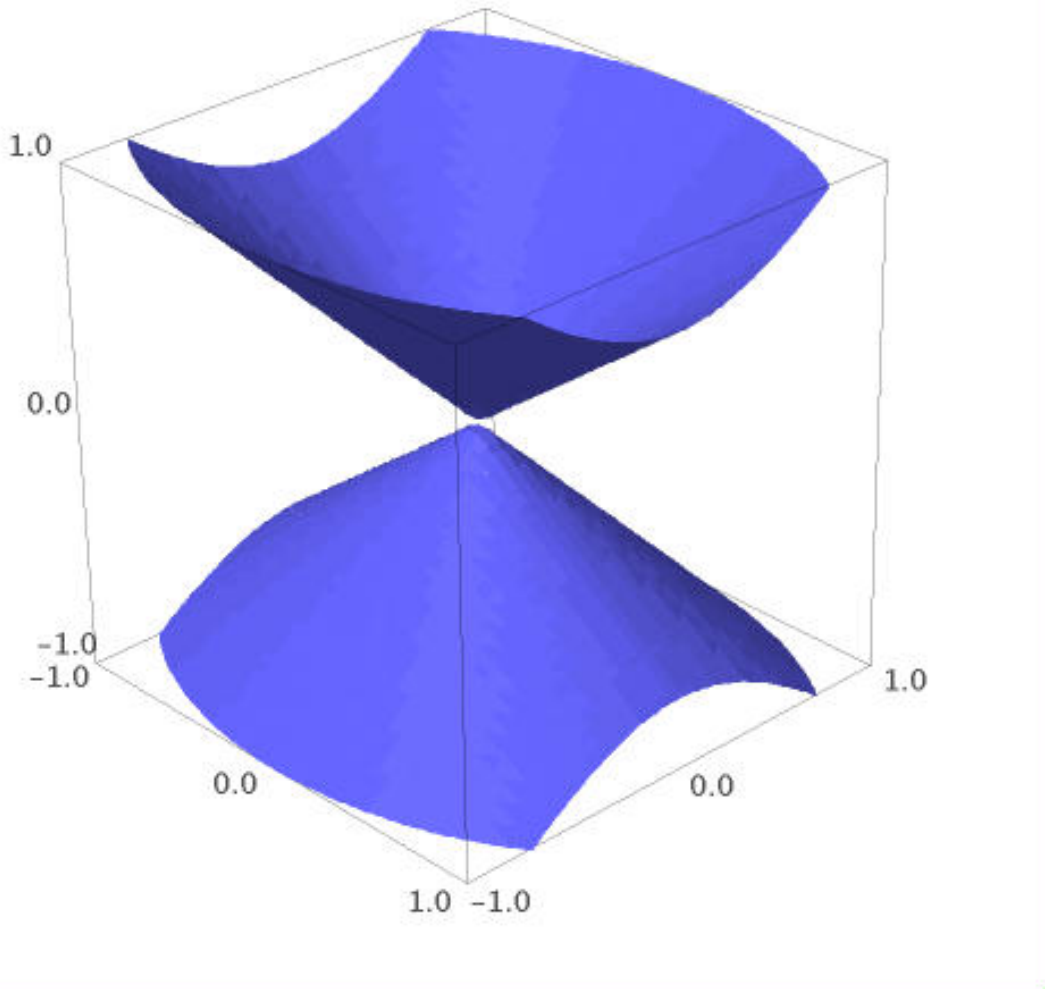


Figure 1: $\frac{1}{2}x^2 + y^2 - z^2 = 0$

This is an elliptical cone. The cross sections parallel to the xy-plane are ellipses of the form $\alpha x^2 + \beta y^2 = c$, whereas the cross sections parallel to the other coordinate planes are generally hyperbolas. Notice also that if a particular point (a, b, c) is on the surface, then so is $t(a, b, c)$ for any $t \in \mathbb{R}$. In other words, the surface contains the straight line through the origin and any of its points. Such lines are called *generators*. When each point of a 3-dimensional surface lies on one or more generators, it is possible to make a model of the surface with straight lengths of wire or string.

(xii) $\alpha x^2 + \beta y^2 + \gamma z^2 = 1$. An ellipsoid. See Fig. 2.

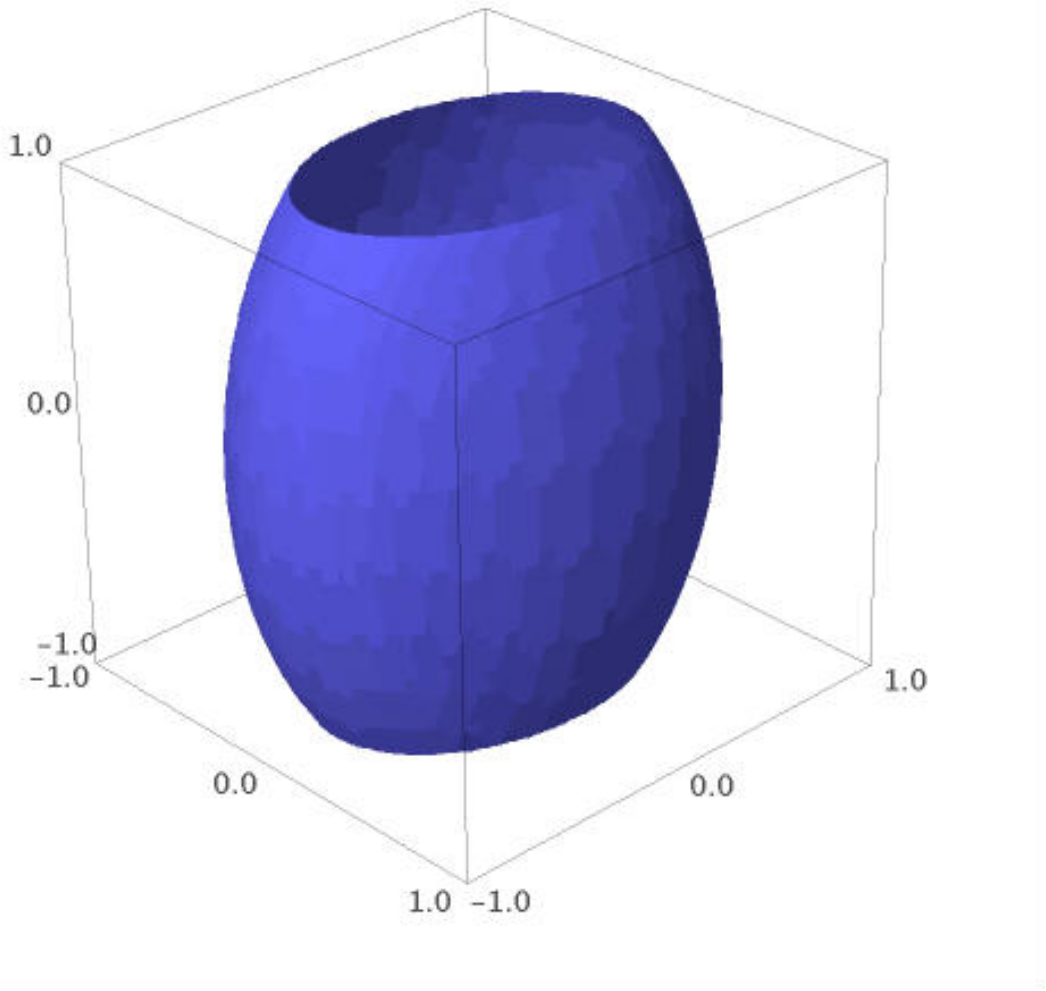


Figure 2: $2x^2 + y^2 + \frac{1}{2}z^2 = 1$

This is a "squashed sphere". It is bounded, and hence clearly has no generators. Notice that if α, β, and γ are distinct, it has only the finite group of symmetries given by reflections in x, y and z, but if some two of the coefficients coincide, it picks up an infinite group of rotation symmetries.

(xiii) $\alpha x^2 + \beta y^2 - \gamma z^2 = 1$. A hyperboloid. See Fig. 3.

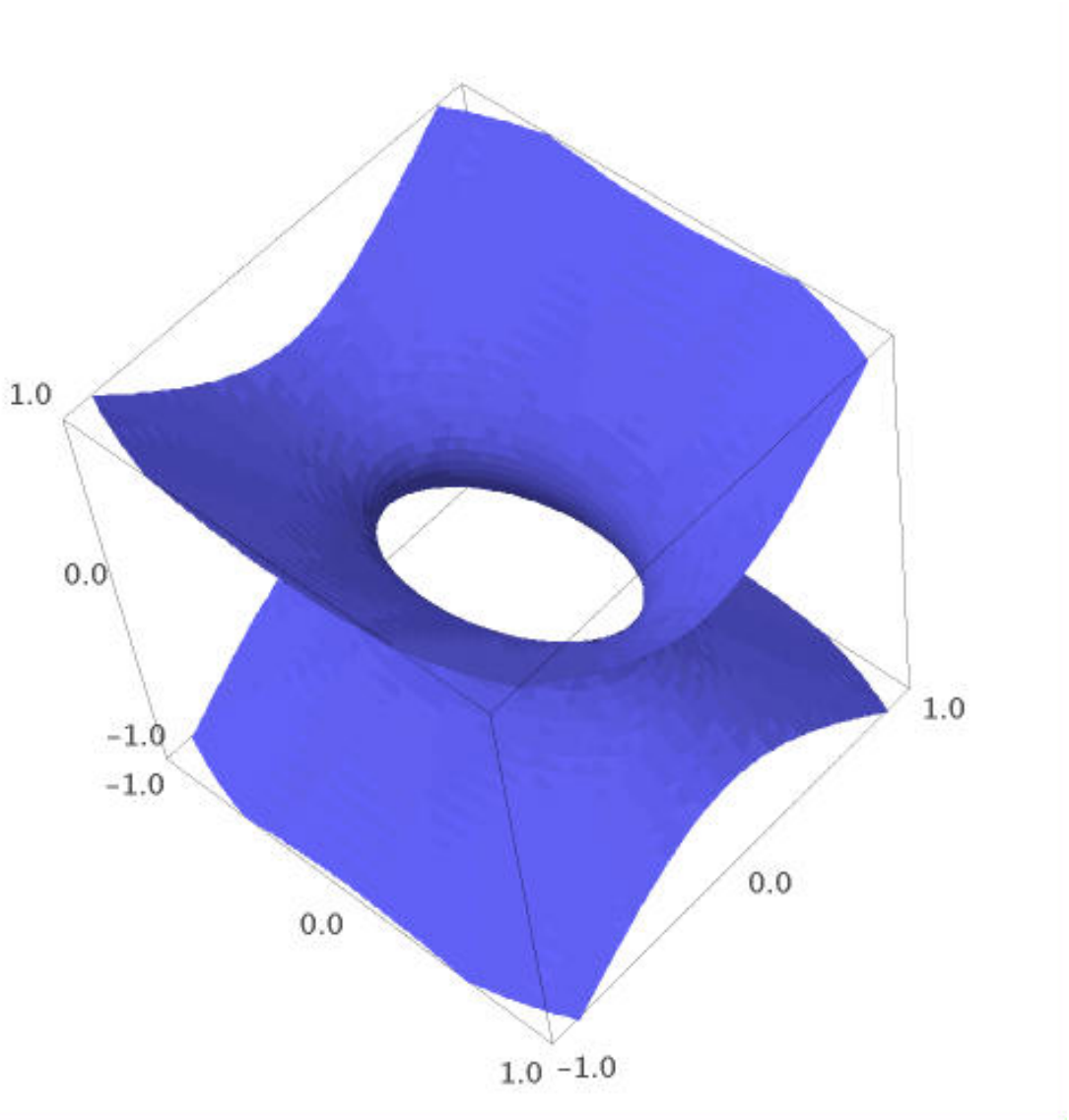


Figure 3: $3x^2 + 8y^2 - 8z^2 = 1$

There are two types of 3-dimensional hyperboloids. This one is connected, and is known as a *hyperboloid of one sheet*. Any cross-section in the xy direction will be an ellipse, and these get larger as z grows (notice the hole in the middle in the picture). Although it is not immediately obvious, each point of this surface lies on exactly two generators; that is, lines that lie entirely on the surface. For each $\lambda \in \mathbb{R}$, the line defined by the pair of equations

$$\sqrt{\alpha}\, x - \sqrt{\gamma}\, z = \lambda(1 - \sqrt{\beta}\, y); \qquad \lambda(\sqrt{\alpha}\, x + \sqrt{\gamma}\, z) = 1 + \sqrt{\beta}\, y.$$

lies entirely on the surface; to see this, just multiply the two equations together. The same applies to the lines defined by the pairs of equations

$$\sqrt{\beta}\, y - \sqrt{\gamma}\, z = \mu(1 - \sqrt{\alpha}\, x); \qquad \mu(\sqrt{\beta}\, y + \sqrt{\gamma}\, z) = 1 + \sqrt{\alpha}\, x.$$

It can be shown that each point on the surface lies on exactly one of the lines in each of these two families.

There is a photo at http://home.cc.umanitoba.ca/~gunderso/model_photos/misc/hyperboloid_of_one_sheet.jpg depicting a rather nice wooden model of a hyperboloid of one sheet, which gives a good idea how these lines sit inside the surface.

(xiv) $\alpha x^2 - \beta y^2 - \gamma z^2 = 1$. Another kind of hyperboloid. See Fig. 4.

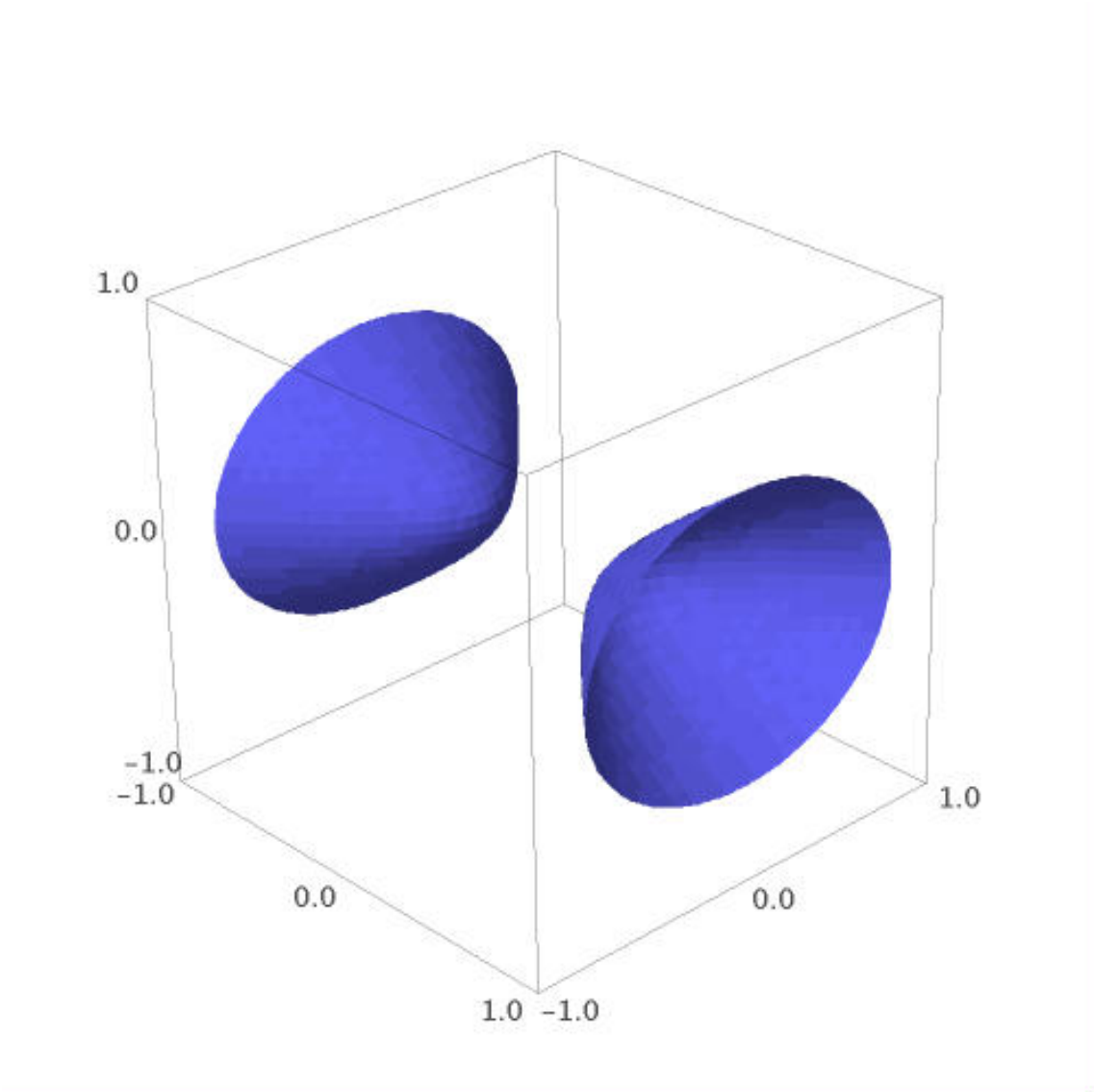


Figure 4: $8x^2 - 12y^2 - 20z^2 = 1$

This one has two connected components and is called a *hyperboloid of two sheets*. It does not have generators.

(xv) $\alpha x^2 + \beta y^2 - z = 0$. An elliptical paraboloid. See Fig. 5.

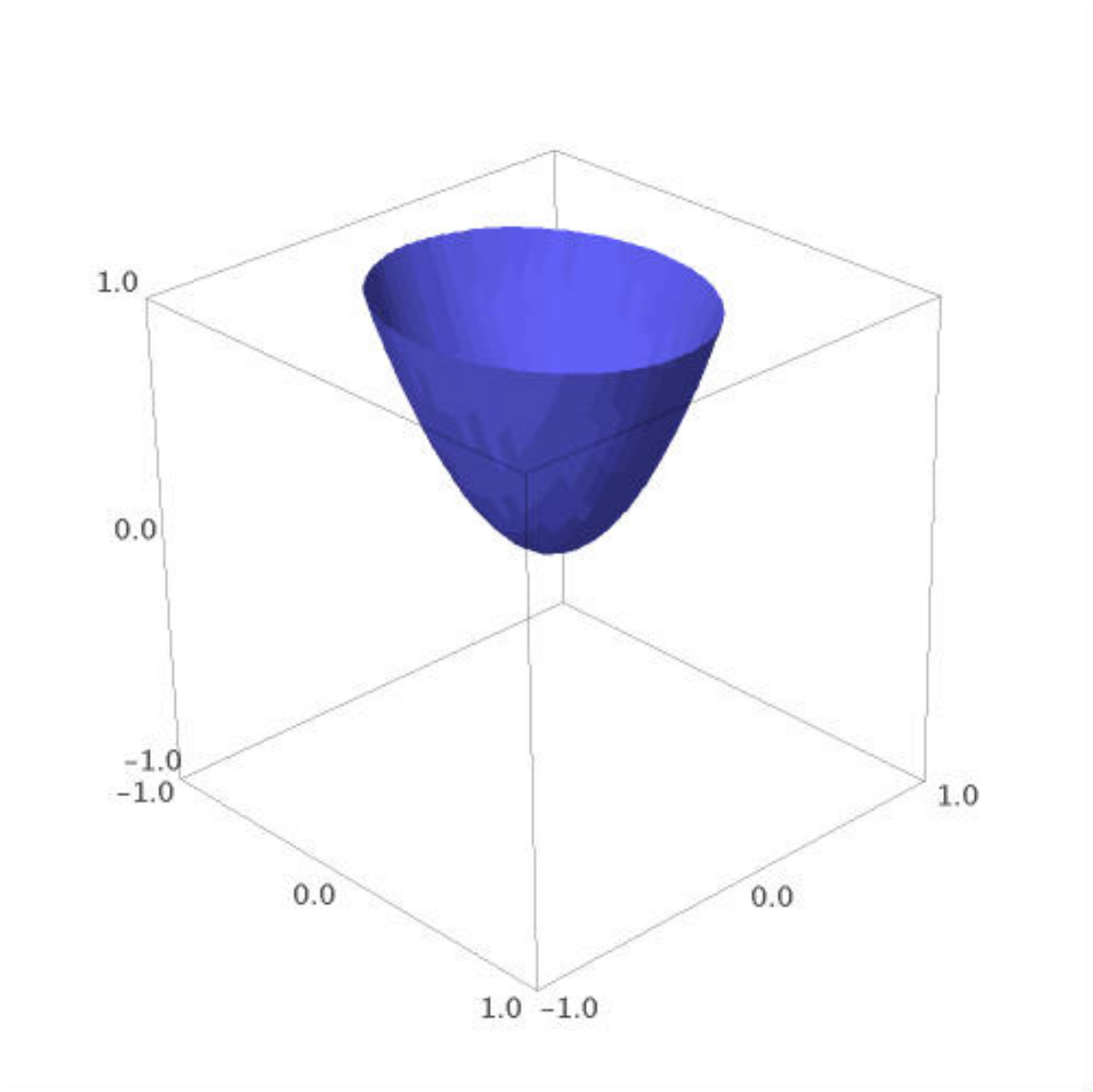


Figure 5: $2x^2 + 3y^2 - z = 0$

Cross-sections of this surface parallel to the xy plane are ellipses, while cross-sections in the yz and xz directions are parabolas. It can be regarded as the limit of a family of hyperboloids of two sheets, where one "cap" remains at the origin and the other recedes to infinity.

(xvi) $\alpha x^2 - \beta y^2 - z = 0$. A hyperbolic paraboloid (a rather elegant saddle shape). See Fig. 6.

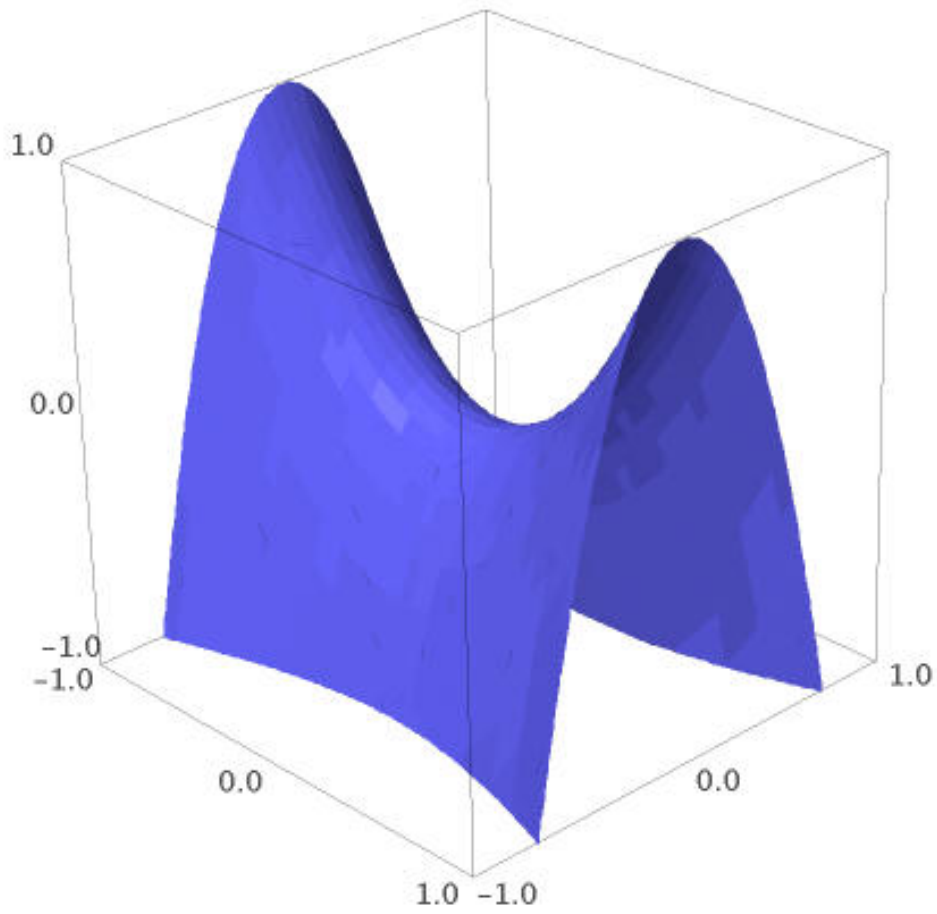


Figure 6: $x^2 - 4y^2 - z = 0$

As in the case of the hyperboloid of one sheet, there are two generators passing through each point of this surface, one from each of the following two families of lines:

$$\lambda(\sqrt{\alpha}\,x - \sqrt{\beta}\,y = z; \qquad \sqrt{\alpha}\,x + \sqrt{\beta}\,y = \lambda.$$
$$\mu(\sqrt{\alpha}\,x + \sqrt{\beta}\,y = z; \qquad \sqrt{\alpha}\,x - \sqrt{\beta}\,y = \mu.$$

Just as the elliptical paraboloid was a limiting case of a hyperboloid of two sheets, so the hyperbolic paraboloid is a limiting case of a hyperboloid of one sheet: you can imagine gradually deforming the hyperboloid of one sheet so the elliptical hole in the middle becomes bigger and bigger, and the result is the hyperbolic paraboloid.

4.9 Singular value decomposition

Week 7

What does a linear map $T : V \to W$ between euclidean spaces look like? This is the question answered by SVD, *singular value decomposition*.

From MA106 Linear Algebra we know that we can choose bases in V and W such that the matrix of T is in the Smith Normal form $\left(\begin{array}{c|c} I_n & 0 \\ \hline 0 & 0 \end{array}\right)$ where n is the rank of T. This answer is unsatisfactory because it does not take the euclidean geometry of V and W into account. In other words, we want to choose orthonormal bases, not just any bases.

Theorem 4.9.1. *(SVD for linear maps) Suppose $T : V \to W$ is a linear map of rank n between euclidean spaces. Then there exist unique positive numbers $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_n > 0$, called the singular values of T, and orthonormal bases of V and W such that the matrix of T with respect to these bases is*

$$\left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right) \quad \textit{where} \quad D = \begin{pmatrix} \gamma_1 & & \\ & \ddots & \\ & & \gamma_n \end{pmatrix}.$$

Proof. Let us consider a new bilinear form on V:

$$\mathbf{u}\clubsuit\mathbf{v} := T(\mathbf{u}) \cdot T(\mathbf{v}) \ .$$

Note that $\mathbf{v}\clubsuit\mathbf{v} = T(\mathbf{v}) \cdot T(\mathbf{v}) \geq 0$ (we call such a bilinear form *positive semidefinite*). By Theorem 4.7.3, there exist unique constants $\alpha_1 \geq \ldots \geq \alpha_m$ (eigenvalues of the matrix of the $\clubsuit$-form) and an orthonormal basis $\mathbf{e}_1, \ldots, \mathbf{e}_m$ of V such that the $\clubsuit$-form is given by $\begin{pmatrix} \alpha_1 & & \\ & \ddots & \\ & & \alpha_m \end{pmatrix}$ in this basis. From the note above ($\clubsuit$-form is positive semidefinite) we easily see that all α_i are non-negative. Suppose $\alpha_k > 0$ is the last positive eigenvalue, that is, $\alpha_{k+1} = \cdots = \alpha_m = 0$.

Since $T(\mathbf{e}_i) \cdot T(\mathbf{e}_j) = \mathbf{e}_i\clubsuit\mathbf{e}_j = \delta_{ij}\alpha_i$, we deduce that $T(\mathbf{e}_{k+1}) = \cdots = T(\mathbf{e}_m) = 0$ and $T(\mathbf{e}_1), \ldots, T(\mathbf{e}_k)$ form an orthogonal set of vectors in W. It follows that k is the rank of T since a set of orthogonal vectors is linearly independent (see the proof of Theorem 4.5.3). Thus, $k = n$. We define $\gamma_i := \sqrt{\alpha_i}$ for all $i \leq k$.

Let us use these image vectors to build an orthonormal basis of W. Since $T(\mathbf{e}_i) \cdot T(\mathbf{e}_i) = \mathbf{e}_i\clubsuit\mathbf{e}_i = \alpha_i$, we know the norms of image vectors: $|T(\mathbf{e}_i)| = \sqrt{\alpha_i} = \gamma_i$. Let $\mathbf{f}_i := \frac{T(\mathbf{e}_i)}{\gamma_i}$ for all $i \leq n$. Extend to an orthonormal basis by the Gram-Schmidt process (Theorem 4.5.3). Since $T(\mathbf{e}_i) = \gamma_i\mathbf{f}_i$ for $i \leq n$ and $T(\mathbf{e}_j) = 0$ for $j > n$, the matrix of T with respect to these bases has the required form.

It remains to prove uniqueness of the singular values. Suppose we have orthonormal bases $\mathbf{e}'_1, \ldots, \mathbf{e}'_m$ of V and $\mathbf{f}'_1, \ldots, \mathbf{f}'_s$ of W, in which T is represented by a matrix $\left(\begin{array}{c|c} B & 0 \\ \hline 0 & 0 \end{array}\right)$ where $B = \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_t \end{pmatrix}$ with $\beta_1 \geq \ldots \geq \ldots \beta_t > 0$. Put $\beta_i = 0$ for $i > t$. Then $\mathbf{e}'_i\clubsuit\mathbf{e}'_j = \beta_i\mathbf{f}'_i \cdot \beta_j\mathbf{f}'_j =$

$\delta_{ij}\beta_i^2$. Thus, $\begin{pmatrix} \beta_1^2 & & \\ & \ddots & \\ & & \beta_m^2 \end{pmatrix}$ is the matrix of the ♣-form in the basis $\mathbf{e}'_1, \ldots, \mathbf{e}'_m$. Uniqueness in Theorem 4.7.3 implies the uniqueness of singular values. □

Before we proceed with some examples, all on the standard euclidean spaces $\mathbb{R}^n$, let us restate the SVD for matrices:

Corollary 4.9.2. *(SVD for matrices) Given any real $k \times m$ matrix A, there exist unique singular values $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_n > 0$ and (non-unique) orthogonal matrices P and Q such that*

$$P^T A Q = P^{-1} A Q = \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right) \quad \textit{where} \quad D = \begin{pmatrix} \gamma_1 & & \\ & \ddots & \\ & & \gamma_n \end{pmatrix}.$$


This corollary is a bit ridiculous. SVD needs to **decompose** A. Thus, the SVD is actually

$$A = P \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right) Q^T .$$

Example. Consider a linear map $\mathbb{R}^2 \to \mathbb{R}^2$, given by the symmetric matrix $A = \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}$, in the example from Section 4.7. There we found the orthogonal matrix $P = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix}$ with $P^{\mathrm{T}} A P = \begin{pmatrix} 4 & 0 \\ 0 & -2 \end{pmatrix}$. This is not the SVD of A because the diagonal matrix contains a negative entry. To get to the SVD we just need to pick different bases for the domain and the range: the columns $\mathbf{c}_1$, $\mathbf{c}_2$ can still be a basis of the domain, while the basis of the range could become $\mathbf{c}_1$, $-\mathbf{c}_2$. This is the SVD:

$$P = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}, \; Q = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix}, \; P^{\mathrm{T}} A Q = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}.$$

The same method works for any symmetric matrix: the SVD is just orthogonal diagonalisation with additional care needed for signs. If the matrix is not symmetric, we need to follow the proof of Theorem 4.9.1 during the calculation.

Example. (from *D. C. Lay, Linear Algebra and Its Applications*) Consider a linear map $\mathbb{R}^3 \to \mathbb{R}^2$, given by $A = \begin{pmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{pmatrix}$. Since $\mathbf{x} \clubsuit \mathbf{y} = A\mathbf{x} \cdot A\mathbf{y} = (A\mathbf{x})^T A\mathbf{y} = \mathbf{x}^T (A^T A)\mathbf{y}$, the matrix of the

♣-form in the standard basis is

$$A^TA = \begin{pmatrix} 4 & 8 \\ 11 & 7 \\ 14 & -2 \end{pmatrix} \begin{pmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{pmatrix} = \begin{pmatrix} 80 & 100 & 40 \\ 100 & 170 & 140 \\ 40 & 140 & 200 \end{pmatrix}.$$

The eigenvalues of this matrix are 360, 90 and 0. Hence the singular values of A are

$$\gamma_1 = \sqrt{360} = 6\sqrt{10} \geq \gamma_2 = \sqrt{90} = 3\sqrt{10}\,.$$

At this stage we are assured of the existence of orthogonal matrices P and Q such that

$$P^{\mathrm{T}}AQ = \begin{pmatrix} 6\sqrt{10} & 0 & 0 \\ 0 & 3\sqrt{10} & 0 \end{pmatrix}.$$

To find the orthogonal matrices we need to find eigenvectors of A^TA:

$$\mathbf{e}_1 = \begin{pmatrix} 1/3 \\ 2/3 \\ 2/3 \end{pmatrix},\ \mathbf{e}_2 = \begin{pmatrix} -2/3 \\ -1/3 \\ 2/3 \end{pmatrix},\ \mathbf{e}_3 = \begin{pmatrix} 2/3 \\ -2/3 \\ 1/3 \end{pmatrix}$$

and then their images under A:

$$\mathbf{f}_1 = \frac{1}{6\sqrt{10}} A\mathbf{e}_1 = \begin{pmatrix} 3/\sqrt{10} \\ 1/\sqrt{10} \end{pmatrix},\ \mathbf{f}_2 = \frac{1}{3\sqrt{10}} A\mathbf{e}_2 = \begin{pmatrix} 1/\sqrt{10} \\ -3/\sqrt{10} \end{pmatrix}.$$

Hence, the orthogonal matrices are

$$P = \begin{pmatrix} 3/\sqrt{10} & 1/\sqrt{10} \\ 1/\sqrt{10} & -3/\sqrt{10} \end{pmatrix},\ Q = \begin{pmatrix} 1/3 & -23 & 2/3 \\ 2/3 & -1/3 & -2/3 \\ 2/3 & 2/3 & 1/3 \end{pmatrix}.$$

The definition of the 2-norm from Section 3.5 easily adopts to a set of linear maps $V \to W$ between euclidean spaces

$$\|T\|_2 := \sup\left\{ \frac{|T(\mathbf{x})|}{|\mathbf{x}|} \mid \mathbf{x} \in V \setminus \{0\} \right\} = \max\{|T(\mathbf{x})| \mid \mathbf{x} \in V, |\mathbf{x}| = 1\}.$$

Notice that linearity allows us to rewrite the definition as $\sup\{|T(\mathbf{x})| \mid \mathbf{x} \in V, |\mathbf{x}| = 1\}$. Then the Extreme Value Theorem holds on a sphere so that the continuous function $|T(\mathbf{x})|$ attains its maximum. The definition is traditionally written as supremum because in that form it works for infinite dimensional spaces as well.

Lemma 4.9.3. *If α_1 is the first (largest) singular value of T, then $\|T\|_2 = \alpha_1$.*

Proof. In the notation of the proof of Theorem 4.9.1, $\alpha_1 = |T(\mathbf{e}_1)| \in \{|T(\mathbf{x})| \mid \mathbf{x} \in V, |\mathbf{x}| = 1\}$. Hence, $\|T\|_2 \geq \alpha_1$.


More Analysis, yeh!?

Indeed, the cleanest proof uses *the Method of Lagrange Multipliers*. Let us consider a point $\mathbf{y} = (y_i)$ on the unit sphere where the function $\mathbf{x} \mapsto |T(\mathbf{x})|^2$ attains its maximum (such point exists by the Extreme Value Theorem). The coordinates are in the orthonormal basis where the SVD is achieved. It follows that all partial derivatives

$$\frac{\partial}{\partial x_i}(|T(\mathbf{x})|^2 - \lambda(|\mathbf{x}| - 1)) = \frac{\partial}{\partial x_i}(\lambda + \sum_j \alpha_j^2 x_j^2 - \sum_j \lambda x_j^2) = 2\alpha_i^2 x_i - 2\lambda x_i = 2(\alpha_i^2 - \lambda)x_i$$

vanish at $\mathbf{y}$ for some Lagrange multiplier $\lambda \in \mathbb{R}$. Since $\mathbf{y} \neq 0$, there exists non-zero y_i. It follows that $\lambda = \alpha_i^2$. Moreover, $\alpha_j \neq \alpha_i$, then $y_j = 0$. Thus, $\|T\|_2 = |T(\mathbf{y})| = \alpha_i \leq \alpha_1$. □

4.10 The complex story

The results in Subsection 4.7 applied only to vector spaces over the real numbers $\mathbb{R}$. There are corresponding results for spaces over the complex numbers $\mathbb{C}$, which we shall summarize here without proofs, although the proofs are similar and analogous to those for spaces over $\mathbb{R}$.

4.10.1 Sesquilinear forms

The key thing that made everything work over $\mathbb{R}$ was the fact that if $x_1, \ldots, x_n$ are real numbers, and $x_1^2 + \cdots + x_n^2 = 0$, then all the x_i are zero. This doesn't work over $\mathbb{C}$, obviously (take $x_1 = 1$ and $x_2 = i$). But we do have something similar if we bring *complex conjugation* into play. As usual, for $z \in \mathbb{C}$, $\overline{z}$ will denote the complex conjugate of z. Then if $z_1\overline{z}_1 + \cdots + z_n\overline{z}_n = 0$, the z's are all zero. So we need to "put bars on half of our formulae". Notice that there was a hint of this in the proof of Proposition 4.7.2.

We'll do this as follows.

Definition 4.10.1. A *sesquilinear form* on a complex vector space V is a function $\tau : V \times V \to \mathbb{C}$ such that

$$\tau(\mathbf{v}, a_1\mathbf{w}_1 + a_2\mathbf{w}_2) = a_1\tau(\mathbf{v}, \mathbf{w}_1) + a_2\tau(\mathbf{v}, \mathbf{w}_2)$$

(as before), but

$$\tau(a_1\mathbf{v}_1 + a_2\mathbf{v}_2, \mathbf{w}) = \overline{a}_1\tau(\mathbf{v}_1, \mathbf{w}) + \overline{a}_2\tau(\mathbf{v}_2, \mathbf{w}),$$

for all vectors v_1, v_2, v, w_1, w_2, w and all $a_1, a_2 \in \mathbb{C}$.

We say such a form is *hermitian symmetric* if

$$\tau(\mathbf{w}, \mathbf{v}) = \overline{\tau(\mathbf{v}, \mathbf{w})}.$$

The word "sesquilinear" literally means "one-and-a-half-times-linear" from its Latin meaning – it's linear in the second argument, but only halfway there in the first argument! We'll often abbreviate "hermitian-symmetric sesquilinear form" to just "hermitian form".

We can represent these by matrices in a similar way to bilinear forms. If τ is a sesquilinear form, and $\mathbf{e}_1, \ldots, \mathbf{e}_n$ is a basis of V, we define the matrix of τ to be the matrix A whose i, j entry is $\tau(\mathbf{e}_i, \mathbf{e}_j)$. Then we have

$$\tau(\mathbf{v}, \mathbf{w}) = \overline{(\underline{\mathbf{v}}^{\mathrm{T}})} A \underline{\mathbf{w}}$$

where $\underline{\mathbf{v}}$ and $\underline{\mathbf{w}}$ are the coordinates of $\mathbf{v}$ and $\mathbf{w}$ as usual. We'll shorten this to $\underline{\mathbf{v}}^* A \underline{\mathbf{w}}$, where the $*$ denotes "conjugate transpose". The condition to be hermitian symmetric translates to the relation $a_{ji} = \overline{a_{ij}}$, so τ is hermitian if and only if A satisfies $A^* = A$.

We have a version here of Sylvester's two theorems (Proposition 4.4.3 and Theorem 4.4.4):

Theorem 4.10.2. *If τ is a hermitian form on a complex vector space V, there is a basis of V in which the matrix of τ is given by*

$$\left(\begin{array}{c|c|c} I_t & & \\ \hline & -I_u & \\ \hline & & 0 \end{array}\right)$$

for some uniquely determined *integers t and u.*

As in the real case, we call the pair (t, u) the *signature* of τ, and we say τ is *positive definite* if its signature is $(n, 0)$ (if V is an n-dimensional space). In this case, the theorem tells us that there is a basis of V in which the matrix of τ is the identity, and in such a basis we have

$$\tau(\mathbf{v}, \mathbf{v}) = \sum_{i=1}^{n} |v_i|^2$$

where $v_1, \ldots, v_n$ are the coordinates of $\mathbf{v}$. Hence $\tau(\mathbf{v}, \mathbf{v}) > 0$ for all non-zero $\mathbf{v} \in V$.

Just as we defined a euclidean space to be a real vector space with a choice of positive definite bilinear form, we have a similar definition here:

Definition 4.10.3. A *Hilbert space* is a finite-dimensional complex vector space endowed with a choice of positive-definite hermitian-symmetric sesquilinear form.

These are the complex analogues of euclidean spaces. If V is a Hilbert space, we write $\mathbf{v} \cdot \mathbf{w}$ for the sesquilinear form on V, and we refer to it as an *inner product*. For any Hilbert space, we can always find a basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of V such that $\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij}$ (an orthonormal basis). Then we can write the inner product matrix-wise as

$$\mathbf{v} \cdot \mathbf{w} = \underline{\mathbf{v}}^* \underline{\mathbf{w}},$$

where $\underline{\mathbf{v}}$ and $\underline{\mathbf{w}}$ are the coordinates of $\mathbf{v}$ and $\mathbf{w}$ and $\underline{\mathbf{v}}^* = \overline{\underline{\mathbf{v}}^{\mathrm{T}}}$ as before.

The canonical example of a Hilbert space is $\mathbb{C}^n$, with the standard inner product given by

$$\mathbf{v} \cdot \mathbf{w} = \sum_{i=1}^{n} \overline{v_i} w_i,$$

for which the standard basis is obviously orthonormal.

Remark. *Technically, I should say "finite-dimensional Hilbert space". There are lots of interesting infinite-dimensional Hilbert spaces, but we won't say anything about them in this course. (Curiously, one never seems to come across infinite-dimensional euclidean spaces.)*

4.10.2 Operators on Hilbert spaces

In our study of linear operators on euclidean spaces, the idea of the *adjoint* of an operator was important. There's an analogue of it here:

Definition 4.10.4. Let $T : V \to V$ be a linear operator on a Hilbert space V. Then there is a unique linear operator $T^* : V \to V$ (the *hermitian adjoint* of V) such that

$$T(\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot T^*(\mathbf{w}).$$

It's clear that if A is the matrix of T in an orthonormal basis, then the matrix of T^* is A^*.

Definition 4.10.5. We say that T is

- *selfadjoint* if $T^* = T$,
- *unitary* if $T^* = T^{-1}$,
- *normal* if $T^*T = TT^*$.

Exercise. If T is unitary, then $T(\mathbf{u}) \cdot T(\mathbf{v}) = \mathbf{u} \cdot \mathbf{v}$ for all $\mathbf{u}, \mathbf{v}$ in V.

If A is the matrix of T in an orthonormal basis, then it's clear that T is selfadjoint if and only if $A^* = A$ (a hermitian-symmetric matrix), unitary if and only if $A^* = A^{-1}$ (a *unitary matrix*), and normal if and only if $A^*A = AA^*$ (a *normal matrix*). In other words, these properties are preserved under unitary bases changes:

Lemma 4.10.6. *If $A \in \mathbb{C}^{n,n}$ is normal (selfadjoint, unitary) and $P \in \mathbb{C}^{n,n}$ is unitary, then P^*AP is normal (selfadjoint, unitary).*

Proof. Let $B = P^*AP$. Using the property $(MN)^* = N^*M^*$, we compute that in the first (normal) case,

$$BB^* = (P^*AP)(P^*AP)^* = P^*APP^*A^*P = P^*AA^*P = P^*A^*AP = (P^*A^*P)(P^*AP) = B^*B.$$

In the second (selfadjoint) case, $B^* = P^*A^*P = P^*AP = B$. In the third (unitary) case, $BB^* = P^*APP^*A^*P = P^*AA^*P = P^*P = I$. □

Notice that if A is unitary and the entries of A are real, then A must be orthogonal, but the definition also includes things like

$$\begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix}.$$

Similarly, a matrix with real entries is hermitian-symmetric if and only if it's symmetric, but

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

is a hermitian-symmetric matrix that's not symmetric.

Both selfadjoint and unitary operators are normal. The generalisation of Theorem 4.7.3 applies to all three types of operators.

Theorem 4.10.7. *The following statements hold for a linear operator $T : V \to V$ on a Hilbert space.*

(i) T is normal if and only if there exists an orthonormal basis of V consisting of eigenvectors of T.

(ii) T is selfadjoint if and only if there exists an orthonormal basis of V consisting of eigenvectors of T with real eigenvalues.

(iii) T is unitary if and only if there exists an orthonormal basis of V consisting of eigenvectors of T with eigenvalues of absolute value 1.

Proof. Let us choose an orthonormal basis of V and replace V with $\mathbb{C}^n$. Let A be the matrix of T.

The "if part" of (i). Then $A = P^*DP$ for a diagonal matrix D and a unitary matrix P. Clearly, D is normal. By Lemma 4.10.6, A is normal.

The "only if part" of (i). We know that A is normal. We proceed by induction on n. If $n = 1$, there is nothing to prove. Let us assume we have proved the statement for all dimensions less than n. The matrix A will have an eigenvector $v \in \mathbb{C}^n$ corresponding to some eigenvalue λ. Let W be the vector subspace of all vectors x satisfying $Ax = \lambda x$. If $W = \mathbb{C}^n$ then A is a scalar matrix and we are done. Otherwise, we have a nontrivial[9] decomposition $\mathbb{C}^n = W \oplus W^\perp$ where $W^\perp = \{\mathbf{v} \in \mathbb{C}^n \mid \forall \mathbf{w} \in W,\ \mathbf{v}^* \cdot \mathbf{w} = 0\}$

Let us notice that $A^*W \subseteq W$ because $AA^*x = A^*Ax = A^*\lambda x = \lambda(A^*x)$ for any $x \in W$. It follows that $AW^\perp \subseteq W^\perp$ since $(Ay)^*x = y^*(A^*x) \in y^*W = 0$ so $(Ay)^*x = 0$ for all $x \in W$, $y \in W^\perp$. Similarly, $A^*W^\perp \subseteq W^\perp$.

Now choose orthonormal bases of W and $W^\perp$. Together they form a new orthonormal basis of $\mathbb{C}^n$. The change of basis matrix P is unitary. Hence, by Lemma 4.10.6, the matrix $P^*AP = \begin{pmatrix} B & 0 \\ 0 & C \end{pmatrix}$ is normal. It follows that the matrices B and C are normal of smaller size and we can use our induction hypothesis to complete the proof.

The "only if part" of (ii) and (iii)). The matrix A is also normal. By part **(i))**, $A = P^*DP$ for a diagonal matrix D and a unitary matrix P. By Lemma 4.10.6, D is selfadjoint (or unitary). Observe that D is selfadjoint if and only if all its diagonal entries are real. Similarly, D is unitary if and only if all its diagonal entries have absolute value 1.

The "if part" of (ii). Then $A = P^*DP$ for a real diagonal matrix D and a unitary matrix P. Clearly, D is selfadjoint. By Lemma 4.10.6, A is selfadjoint.

[9] i.e., neither W nor $W^\perp$ is zero.

The "if part" of (iii). Then $A = P^*DP$ for a diagonal matrix D with diagonal entries of absolute value 1 and a unitary matrix P. Clearly, D is unitary. By Lemma 4.10.6, A is unitary. □

Example. Let $A = \begin{pmatrix} 6 & 2+2i \\ 2-2i & 4 \end{pmatrix}$. Then

$$c_A(x) = (6-x)(4-x) - (2+2i)(2-2i) = x^2 - 10x + 16 = (x-2)(x-8),$$

so the eigenvalues are 2 and 8. Corresponding eigenvectors are $\mathbf{v}_1 = (1+i, -2)^{\mathrm{T}}$ and $\mathbf{v}_2 = (1+i, 1)^{\mathrm{T}}$. We find that $|\mathbf{v}_1|^2 = \mathbf{v}_1^*\mathbf{v}_1 = 6$ and $|\mathbf{v}_2|^2 = 3$, so we divide by their lengths to get an orthonormal basis $\mathbf{v}_1/|\mathbf{v}_1|$, $\mathbf{v}_2/|\mathbf{v}_2|$ of $\mathbb{C}^2$. Then the matrix

$$P = \begin{pmatrix} \frac{1+i}{\sqrt{6}} & \frac{1+i}{\sqrt{3}} \\ \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix}$$

having this basis as columns is unitary and satisfies $P^*AP = \begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix}$.

4.10.3 Singular value decomposition

For complex matrices, the singular value decomposition works exactly the same as for real matrices. The only difference is that the orthonormal bases are accounted for now by unitary matrices. We restate these results for completeness of expositions but we will not prove them.

Theorem 4.10.8. *(SVD for linear maps) Suppose $T : V \to W$ is a linear map of rank n between Hilbert spaces. Then there exist unique positive numbers $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_n > 0$, called the singular values of T, and orthonormal bases in V and W such that the matrix of T with respect to these bases is*

$$\left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right) \quad \textit{where} \quad D = \begin{pmatrix} \gamma_1 & & \\ & \ddots & \\ & & \gamma_n \end{pmatrix}.$$

Corollary 4.10.9. *(SVD for complex matrices) Given any complex $k \times m$ matrix A, there exist unique (real) singular values $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_n > 0$ and (non-unique) unitary matrices P and Q such that*

$$P^*AQ = P^{-1}AQ = \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right) \quad \textit{where} \quad D = \begin{pmatrix} \gamma_1 & & \\ & \ddots & \\ & & \gamma_n \end{pmatrix}.$$

Lemma 4.9.3 works as stated as well: $\|T\|_2 = \alpha_1$.

5 Finitely Generated Abelian Groups

Week 8

In the first four sections of the course, we've always been thinking about vector spaces over *fields*. The idea of this section is to show that some of the same ideas work with the field K replaced by the integers $\mathbb{Z}$, even though $\mathbb{Z}$ isn't a field; and that this is strongly related to the *group theory* which most of you will have seen in MA136 Introduction to Abstract Algebra last year. Do not worry if you did not take that module, we will cover all of the group theory we need in the following sections.

5.1 Definitions

Definition 5.1.1. An abelian group is a set G together with a binary operation, which we write as addition, and which satisfies the following properties:

(i) (*Closure*) for all $g, h \in G$, $g + h \in G$;

(ii) (*Associativity*) for all $g, h, k \in G$, $(g + h) + k = g + (h + k)$;

(iii) there exists an element $0_G \in G$ such that:

  (a) (*Identity*) for all $g \in G$, $g + 0_G = g$; and

  (b) (*Inverse*) for all $g \in G$ there exists $-g \in G$ such that $g + (-g) = 0_G$;

(iv) (*Commutativity*) for all $g, h \in G$, $g + h = h + g$.

Usually we just write 0 rather than 0_G. We only write 0_G if we need to distinguish between the zero elements of different groups.

The commutativity axiom (iv) is not part of the definition of a general group, and for general (non-abelian) groups, it is more usual to use multiplicative rather than additive notation. All groups in this course should be assumed to be abelian, although many of the definitions in this section apply equally well to general groups.

Examples. 1. The integers $\mathbb{Z}$.

2. Fix a positive integer $n > 0$ and let

$$\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\} = \{\, x \in \mathbb{Z} \mid 0 \leq x < n \,\}.$$

   where addition is computed modulo n. So, for example, when $n = 9$, we have $2 + 5 = 7$, $3 + 8 = 2$, $6 + 7 = 4$, etc. Note that the inverse $-x$ of $x \in \mathbb{Z}_n$ is equal to $n - x$ in this example.

3. Examples from linear algebra. Let K be a field.
   (i) The elements of K form an abelian group under addition.
   (ii) The non-zero elements of K form an abelian group $K^\times$ under multiplication.
   (iii) The vectors in any vector space form an abelian group under addition.

Proposition 5.1.2 (The cancellation law)**.** *Let G be any group, and let $g, h, k \in G$. Then $g + h = g + k \Rightarrow h = k$.*

Proof. Add $-g$ to both sides of the equation and use the Associativity and Identity axioms. □

For any group G, $g \in G$, and integer $n > 0$, we define ng to be $g + g + \cdots + g$, with n occurrences of g in the sum. So, for example, $1g = g$, $2g = g + g$, $3g = g + g + g$, etc. We extend this notation to all $n \in \mathbb{Z}$ by defining $0g = 0$ and $(-n)g = -(ng)$ for $-n < 0$. Overall, this defines a scalar action $\mathbb{Z} \times G \to G$ which allows as to think of abelian groups as "vector spaces over $\mathbb{Z}$" (or using precise terminology $\mathbb{Z}$-modules - algebraic modules will play a significant role in *Rings and Modules* in year 3).

Definition 5.1.3. A group G is called *cyclic* if there exists an element $x \in G$ such that every element of G is of the form mx for some $m \in \mathbb{Z}$.

The element x in the definition is called a *generator* of G. Note that $\mathbb{Z}$ and $\mathbb{Z}_n$ are cyclic with generator $x = 1$.

Definition 5.1.4. A bijection $\phi : G \to H$ between two (abelian) groups is called an *isomorphism* if $\phi(g + h) = \phi(g) + \phi(h)$ for all $g, h \in G$, and the groups G and H are called *isomorphic* is there is an isomorphism between them.

The notation $G \cong H$ means that G is isomorphic to H; isomorphic groups are often thought of as being essentially the same group, but with elements having different names.

Note (exercise) that any isomorphism must satisfy $\phi(0_G) = 0_H$ and $\phi(-g) = -\phi(g)$ for all $g \in G$.

Proposition 5.1.5. *Any cyclic group G is isomorphic either to $\mathbb{Z}$ or to $\mathbb{Z}_n$ for some $n > 0$.*

Proof. Let G be cyclic with generator x. So $G = \{ mx \mid m \in \mathbb{Z} \}$. Suppose first that the elements mx for $m \in \mathbb{Z}$ are all distinct. Then the map $\phi : \mathbb{Z} \to G$ defined by $\phi(m) = mx$ is a bijection, and it is straightforward to check that it is an isomorphism.

Otherwise, we have $lx = mx$ for some $l < m$, and so $(m - l)x = 0$ with $m - l > 0$. Let n be the least integer with $n > 0$ and $nx = 0$. Then the elements $0x = 0, 1x, 2x, \ldots, (n-1)x$ of G are all distinct, because otherwise we could find a smaller n. Furthermore, for any $mx \in G$, we can write $m = rn + s$ for some $r, s \in \mathbb{Z}$ with $0 \leq s < n$. Then $mx = (rn + s)x = sx$, so $G = \{ 0, 1x, 2x, \ldots, (n-1)x \}$, and the map $\phi : \mathbb{Z}_n \to G$ defined by $\phi(m) = mx$ for $0 \leq m < n$ is a bijection, and we check that it is an isomorphism. □

Definition 5.1.6. For an element $g \in G$, the least integer $n > 0$ with $ng = 0$, if it exists, is called the *order* of g and we denote the order of g by $|g|$. If there is no such n, then g has infinite order and we write $|g| = \infty$.

Exercise. If $\phi : G \to H$ is an isomorphism then $|g| = |\phi(g)|$ for all $g \in G$.

Definition 5.1.7. A group G is *generated* or *spanned* by a subset X of G if every $g \in G$ can be written as a finite sum $\sum_{i=1}^{k} m_i x_i$, with $m_i \in \mathbb{Z}$ and $x_i \in X$. It is finitely generated if it has a finite generating set $X = \{x_1, \ldots, x_n\}$.

So a group is cyclic if and only if it has a generating set X with $|X| = 1$.

In general, if G is generated by X, then we write $G = \langle X \rangle$ or $G = \langle x_1, \ldots, x_n \rangle$ when $X = \{x_1, \ldots, x_n\}$ is finite.

Definition 5.1.8. The direct sum of groups $G_1, \ldots, G_n$ is defined to be the set $\{ (g_1, g_2, \ldots, g_n) \mid g_i \in G_i \}$ with component-wise addition

$$(g_1, g_2, \ldots, g_n) + (h_1, h_2, \ldots, h_n) = (g_1 + h_1, g_2 + h_2, \ldots, g_n + h_n).$$

This is a group with identity element $(0, 0, \ldots, 0)$ and $-(g_1, g_2, \ldots, g_n) = (-g_1, -g_2, \ldots, -g_n)$.

In general (non-abelian) group theory this is more often known as the direct product of groups.

The main result of this section, known as the *fundamental theorem of finitely generated abelian groups*, is that every finitely generated abelian group is isomorphic to a direct sum of cyclic groups. (This is not true in general for abelian groups, such as the additive group $\mathbb{Q}$ of rational numbers, which are not finitely generated.)

5.2 Subgroups, cosets and quotient groups

Definition 5.2.1. A subset H of a group G is called a *subgroup* of G if it forms a group under the same operation as that of G.

Lemma 5.2.2. *If H is a subgroup of G, then the identity element 0_H of H is equal to the identity element 0_G of G.*

Proof. Using the identity axioms for H and G, $0_H + 0_H = 0_H = 0_H + 0_G$. Now by the cancellation law, $0_H = 0_G$. □

The definition of a subgroup is *semantic* in its nature. While it precisely pinpoints what a subgroup is, it is quite cumbersome to use. The following proposition gives a usable criterion.

Proposition 5.2.3. *Let H be a subset of a group G. The following statements are equivalent.*

(i) *H is a subgroup of G.*

(ii) *(a) H is nonempty; and*

(b) $h_1, h_2 \in H \Rightarrow h_1 + h_2 \in H$; and

(c) $h \in H \Rightarrow -h \in H$.

(iii) *(a) H is nonempty; and*

(b) $h_1, h_2 \in H \Rightarrow h_1 - h_2 \in H$.

Proof. If H is a subgroup of G then it is nonempty as it contains 0. Moreover, $h_1 - h_2 = h_1 + (-h_2) \in H$ if h_1 and h_2 are in H. Thus, (i) implies (iii).

To show that (iii) implies (ii) we pick $x \in H$. Then $0 = x - x \in H$. Now $-h = 0 - h \in H$ for any $h \in H$. Finally, $h_1 + h_2 = h_1 - (-h_2) \in H$ for all $h_1, h_2 \in H$.

To show that (ii) implies (i) we need to verify the four group axioms in H. Two of these, 'Closure', and 'Inverse', are the conditions (b) and (c). The other two axioms are 'Associativity' and 'Identity'. Associativity holds because it holds in G, and H is a subset of G. Since we are assuming that H is nonempty, there exists $h \in H$, and then $-h \in H$ by (c), and $h + (-h) = 0 \in H$ by (b), and so 'Identity' holds, and H is a subgroup. □

Examples. 1. There are two standard subgroups of any group G: the whole group G itself, and the *trivial* subgroup $\{0\}$ consisting of the identity alone. Subgroups other than G are called *proper* subgroups, and subgroups other than $\{0\}$ are called *non-trivial* subgroups.

2. If g is any element of any group G, then the set of all integer multiples $\{ mg \mid m \in \mathbb{Z} \}$ forms a subgroup of G called the cyclic subgroup generated by g.

Let us look at a few specific examples. If $G = \mathbb{Z}$, then $5\mathbb{Z}$, which consists of all multiples of 5, is the cyclic subgroup generated by 5. Of course, we can replace 5 by any integer here, but note that the cyclic groups generated by 5 and -5 are the same.

If $G = \langle g \rangle$ is a finite cyclic group of order n and m is a positive integer dividing n, then the cyclic subgroup generated by mg has order n/m and consists of the elements kmg for $0 \leq k < n/m$.

Exercise. What is the order or the cyclic subgroup generated by mg for general m (where we drop the assumption that $m|n$)?

Exercise. Show that the group of non-zero complex numbers $\mathbb{C}^\times$ under the operation of multiplication has finite cyclic subgroups of all possible orders.

Definition 5.2.4. Let $g \in G$. Then the *coset* $H + g$ is the subset $\{ h + g \mid h \in H \}$ of G.

(*Note*: Since our groups are abelian, we have $H + g = g + H$, but in general group theory the right and left cosets $H + g$ and $g + H$ can be different.)

Examples. 1. $G = \mathbb{Z}$, $H = 5\mathbb{Z}$. There are just 5 distinct cosets $H = H + 0 = \{ 5n \mid n \in \mathbb{Z} \}$, $H + 1 = \{ 5n + 1 \mid n \in \mathbb{Z} \}$, $H + 2$, $H + 3$, $H + 4$. Note that $H + i = H + j$ whenever $i \equiv j \pmod 5$.

2. $G = \mathbb{Z}_6$, $H = \{0, 3\}$. There are 3 distinct cosets, $H = H + 3 = \{0, 3\}$, $H + 1 = H + 4 = \{1, 4\}$, and $H + 2 = H + 5 = \{2, 5\}$,

3. $G = \mathbb{C}^\times$, the group of non-zero complex numbers under multiplication and the subgroup $S^1 = \{z,\ |z| = 1\}$, which is the unit circle. The cosets are circles. There are uncountably many distinct cosets, one for each positive real number (radius of a circle).

Proposition 5.2.5. *The following are equivalent for $g, k \in G$:*

(i) $k \in H + g$.

(ii) $H + g = H + k$.

(iii) $k - g \in H$.

Proof. Clearly $H + g = H + k \Rightarrow k \in H + g$, so $(ii) \Rightarrow (i)$.

If $k \in H + g$, then $k = h + g$ for some fixed $h \in H$, so $g = k - h$. Let $f \in H + g$. Then, for some $h_1 \in H$, we have $f = h_1 + g = h_1 + k - h \in H + k$, so $H + g \subseteq H + k$. Similarly, if $f \in H + k$, then for some $h_1 \in H$, we have $f = h_1 + k = h_1 + h + g \in H + g$, so $H + k \subseteq H + g$. Thus $H + g = H + k$, and we have proved that $(i) \Rightarrow (ii)$.

If $k \in H + g$, then, as above, $k = h + g$, so $k - g = h \in H$ and $(i) \Rightarrow (iii)$.

Finally, if $k - g \in H$, then putting $h = k - g$, we have $h + g = k$, so $k \in H + g$, proving $(iii) \Rightarrow (i)$. □

Corollary 5.2.6. *Two cosets $H + g_1$ and $H + g_2$ of H in G are either equal or disjoint.*

Proof. If $H + g_1$ and $H + g_2$ are not disjoint, then there exists an element $k \in (H + g_1) \cap (H + g_2)$, but then $H + g_1 = H + k = H + g_2$ by Proposition 5.2.5. □

Corollary 5.2.7. *The cosets of H in G partition G.*

Proposition 5.2.8. *If H is finite, then all cosets have exactly $|H|$ elements.*

Proof. Since $h_1 + g = h_2 + g \Rightarrow h_1 = h_2$ by the cancellation law, it follows that the map $\phi : H \to H + g$ defined by $\phi(h) = h + g$ is a bijection, and the result follows. □

Corollary 5.2.7 and Proposition 5.2.8 together imply:

Theorem 5.2.9 (Lagrange's Theorem)**.** *Let G be a finite (abelian) group and H a subgroup of G. Then the order of H divides the order of G.*

Definition 5.2.10. The number of distinct right cosets of H in G is called the *index* of H in G and is written as $|G : H|$.

If G is finite, then we clearly have $|G : H| = |G|/|H|$. But, from the example $G = \mathbb{Z}$, $H = 5\mathbb{Z}$ above, we see that $|G : H|$ can be finite even when G and H are infinite.

Proposition 5.2.11. *Let G be a finite (abelian) group. Then for any $g \in G$, the order $|g|$ of g divides the order $|G|$ of G.*

Proof. Let $|g| = n$. We saw in Example 2 above that the integer multiples $\{ mg \mid m \in \mathbb{Z} \}$ of g form a subgroup H of G. By minimality of n, the distinct elements of H are $\{0, g, 2g, \ldots, (n-1)g\}$, so $|H| = n$ and the result follows from Lagrange's Theorem. □

As an application, we can now immediately classify all finite (abelian) groups whose order is prime.

Proposition 5.2.12. *Let G be a (abelian) group having prime order p. Then G is cyclic; that is, $G \cong \mathbb{Z}_p$.*

Proof. Let $g \in G$ with $0 \neq g$. Then $|g| > 1$, but $|g|$ divides p by Proposition 5.2.11, so $|g| = p$. But then G must consist entirely of the integer multiples mg $(0 \leq m < p)$ of g, so G is cyclic. □

Definition 5.2.13. If A and B are subsets of a group G, then we define their sum $A + B = \{ a + b \mid a \in A, b \in B \}$.

Lemma 5.2.14. *If H is a subgroup of the abelian group G and $H + g$, $H + h$ are cosets of H in G, then $(H + g) + (H + h) = H + (g + h)$.*

Proof. Since G is abelian, this follows directly from commutativity and associativity. □

Theorem 5.2.15. *Let H be a subgroup of an abelian group G. Then the set G/H of cosets $H + g$ of H in G forms a group under addition of subsets.*

Proof. We have just seen that $(H + g) + (H + h) = H + (g + h)$, so we have closure, and associativity follows easily from associativity of G. Since $(H + 0) + (H + g) = H + g$ for all $g \in G$, $H = H + 0$ is an identity element, and since $(H - g) + (H + g) = H - g + g = H$, $H - g$ is an inverse to $H + g$ for all cosets $H + g$. Thus the four group axioms are satisfied and G/H is a group. □

Definition 5.2.16. The group G/H is called the *quotient group* (or the *factor group*) of G by H.

Notice that if G is finite, then $|G/H| = |G : H| = |G|/|H|$. So, although the quotient group seems a rather complicated object at first sight, it is actually a smaller group than G.

Examples.

1. Let $G = \mathbb{Z}$ and $H = m\mathbb{Z}$ for some $m > 0$. Then there are exactly m distinct cosets, H, $H + 1$, …, $H + (m - 1)$. If we add together k copies of $H + 1$, then we get $H + k$. So G/H is cyclic of order m and with generator $H + 1$. So by Proposition 5.1.5, $\mathbb{Z}/m\mathbb{Z} \cong \mathbb{Z}_m$.
2. $G = \mathbb{R}$ and $H = \mathbb{Z}$. The quotient group G/H is isomorphic to the circle subgroup S^1 of the multiplicative group $\mathbb{C}^\times$. One writes an explicit isomorphism $\phi : G/H \to S^1$ by $\phi(x + \mathbb{Z}) = e^{2\pi x i}$.

5.3 Homomorphisms and the first isomorphism theorem

Definition 5.3.1. Let G and H be groups. A *homomorphism* ϕ from G to H is a map $\phi : G \to H$ such that $\phi(g_1 + g_2) = \phi(g_1) + \phi(g_2)$ for all $g_1, g_2 \in G$.

Homomorphisms correspond to linear transformations between vector spaces.

Note that an isomorphism is just a bijective homomorphism. There are two other types of 'morphism' that are worth mentioning at this stage.

A homomorphism ϕ is injective if it is an injection; that is, if $\phi(g_1) = \phi(g_2) \Rightarrow g_1 = g_2$. A homomorphism ϕ is surjective if it is a surjection; that is, if $\mathrm{im}(\phi) = H$. Sometimes, a surjective homomorphism is called *epimorphism* while an injective homomorphism is called *monomorphism* but we will not use this terminology in these lectures.

Lemma 5.3.2. *Let $\phi : G \to H$ be a homomorphism. Then $\phi(0_G) = 0_H$ and $\phi(-g) = -\phi(g)$ for all $g \in G$.*

Proof. Exercise. (Similar to results for linear transformations.) □

Example. Let G be any group, and let $n \in \mathbb{Z}$. Then $\phi : G \to G$ defined by $\phi(g) = ng$ for all $g \in G$ is a homomorphism.

Kernels and images are defined as for linear transformations of vector spaces.

Definition 5.3.3. Let $\phi : G \to H$ be a homomorphism. Then the *kernel* $\ker(\phi)$ of ϕ is defined to be the set of elements of G that map onto 0_H; that is,

$$\ker(\phi) = \{\, g \mid g \in G,\ \phi(g) = 0_H \,\}.$$

Note that by Lemma 5.3.2 above, $\ker(\phi)$ always contains 0_G.

Proposition 5.3.4. *Let $\phi : G \to H$ be a homomorphism. Then ϕ is injective if and only if* $\ker(\phi) = \{0_G\}$.

Proof. Since $0_G \in \ker(\phi)$, if ϕ is injective then we must have $\ker(\phi) = \{0_G\}$. Conversely, suppose that $\ker(\phi) = \{0_G\}$, and let $g_1, g_2 \in G$ with $\phi(g_1) = \phi(g_2)$. Then $0_H = \phi(g_1) - \phi(g_2) = \phi(g_1 - g_2)$ (by Lemma 5.3.2), so $g_1 - g_2 \in \ker(\phi)$ and hence $g_1 - g_2 = 0_G$ and $g_1 = g_2$. So ϕ is injective. □

Theorem 5.3.5. *(i) Let $\phi : G \to H$ be a homomorphism. Then* $\ker(\phi)$ *is a subgroup of G and* $\mathrm{im}(\phi)$ *is a subgroup of H.*

(ii) Let H be a subgroup of a group G. Then the map $\phi : G \to G/H$ defined by $\phi(g) = H + g$ is a surjective homomorphism with kernel H.

Proof. (i) is straightforward using Proposition 5.2.3. For (ii), it is easy to check that ϕ is surjective, and $\phi(g) = 0_{G/H} \Leftrightarrow H + g = H + 0_G \Leftrightarrow g \in H$, so $\ker(\phi) = H$. □

The following lemma explains a connection between quotients and homomorphisms. It clarifies the trickiest point in the proof of the forthcoming *First Isomorphism Theorem*.

Lemma 5.3.6. *Let $\phi : G \to H$ be a homomorphism with a kernel K, A a subgroup of G. The homomorphism ϕ determines a homomorphism $\overline{\phi} : G/A \to H$ via $\overline{\phi}(A + g) = \phi(g)$ for all $g \in G$ if and only if $A \subseteq K$.*

Proof. We need to show that $\overline{\phi}(A + g) = \phi(g)$ does actually define a map $\overline{\phi} : G/A \to H$. This is not immeditaley obvious because cosets have different representatives, i.e. we can have $A + g = A + h$ with $g \neq h$. So for $\overline{\phi}$ to make sense we need to ensure $\phi(g) = \phi(h)$ whenever $A + g = A + h$. This is called checking that $\overline{\phi}$ is *well-defined*. Now we know what to check, it is not too difficult. Suppose that $A + g = A + h$. Then $g = a + h$ for some $a \in A$. Hence, $\overline{\phi}$ is well-defined if and only if $\phi(g) = \phi(a) + \phi(h) = \phi(h)$ for all $g, h \in G$, $a \in A$ if and only if $\phi(a) = 0$ for all $a \in A$ if and only if $A \subseteq K$.

Observe also that, once $\overline{\phi}$ is well-defined, it is trivially a homomorphism, since so is ϕ:

$$\overline{\phi}(A + h) + \overline{\phi}(A + g) = \phi(h) + \phi(g) = \phi(h + g) = \overline{\phi}(A + h + g) = \overline{\phi}((A + h) + (A + g)).$$

□

If one denotes the set of all homomorphism from G to H by $\hom(G, H)$ there is an elegant way to reformulate Lemma 5.3.6. The composition with the quotient map $\psi : G \to G/A$ defines a bijection

$$\hom(G/A, H) \to \{\alpha \in \hom(G, H) \mid \alpha(A) = \{0\}\}, \quad \phi \mapsto \phi \circ \psi.$$

Theorem 5.3.7 (The First Isomorphism Theorem)**.** *Let $\phi : G \to H$ be a homomorphism with kernel K. Then $G/K \cong \mathrm{im}(\phi)$. More precisely, there is an isomorphism $\overline{\phi} : G/K \to \mathrm{im}(\phi)$ defined by $\overline{\phi}(K + g) = \phi(g)$ for all $g \in G$.*

Proof. The map $\overline{\phi}$ is a well-defined homomorphism by Lemma 5.3.6. Clearly, $\mathrm{im}(\overline{\phi}) = \mathrm{im}(\phi)$. Finally,

$$\overline{\phi}(K + g) = 0_H \iff \phi(g) = 0_H \iff g \in K \iff K + g = K + 0_G = 0_{G/K}.$$

By Proposition 5.3.4, $\overline{\phi}$ is injective. Thus $\overline{\phi} : G/K \to \mathrm{im}(\phi)$ is an isomorphism. □

We shall be using this theorem later, when we prove the main theorem on finitely generated abelian group in Subsection 5.7. The crucial observation is that any finitely generated abelian group is the cokernel of a homomorphism between two finitely generated free abelian groups, which we will discuss in the next section.

5.4 Free abelian groups

Definition 5.4.1. The direct sum $\mathbb{Z}^n$ of n copies of $\mathbb{Z}$ is known as a (finitely generated) *free abelian group* of rank n.

More generally, a finitely generated abelian group is called free abelian if it is isomorphic to $\mathbb{Z}^n$ for some $n \geq 0$.

(The free abelian group $\mathbb{Z}^0$ of rank 0 is defined to be the trivial group $\{0\}$ containing the single element 0.)

The groups $\mathbb{Z}^n$ have many properties in common with vector spaces such as $\mathbb{R}^n$, but we must expect some differences, because $\mathbb{Z}$ is not a field.


Given we wish to exploit this connection we choose to write elements of $\mathbb{Z}^n$ as columns, rather than rows. So $\begin{pmatrix}1\\0\end{pmatrix} \in \mathbb{Z}^2$ etc.

We then define the standard basis of $\mathbb{Z}^n$ exactly as for $\mathbb{R}^n$; that is, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i$ has a 1 in its i-th component and 0 in the other components. This has the same properties as a basis of a vector space; i.e. it is linearly independent and spans $\mathbb{Z}^n$.

Definition 5.4.2. Elements $x_1, \ldots, x_n$ of an abelian group G are called *linearly independent* if $\alpha_1 x_1 + \cdots + \alpha_n x_n = 0_G$ for $\alpha_1, \ldots, \alpha_n \in \mathbb{Z}$ implies that $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0_{\mathbb{Z}}$.

Definition 5.4.3. Elements $x_1, \ldots, x_n$ form a *free basis* of the abelian group G if and only if they are linearly independent and generate (span) G.

Example. It's clear that the standard basis $\mathbf{x}_1 = (1, 0, \ldots, 0)^T, \mathbf{x}_2 = (0, 1, \ldots, 0)^T, \ldots, \mathbf{x}_n = (0, 0, \ldots, 1)^T$ is indeed a free basis of $\mathbb{Z}^n$ but there are others; for instance, $\left\{ \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix} \right\}$ is a free basis of $\mathbb{Z}^2$.

It's important to notice, though, that a subset of $\mathbb{Z}^n$ which is a basis of $\mathbb{Q}^n$ need not be a free basis of $\mathbb{Z}^n$. For instance, $\left\{ \begin{pmatrix}2\\0\end{pmatrix}, \begin{pmatrix}0\\2\end{pmatrix} \right\}$ is not a free basis of $\mathbb{Z}^2$, since we can't write all elements of $\mathbb{Z}^2$ as linear combinations of these vectors with *integer* coefficients – we'll need to divide by 2 at some point. This also shows that a set of n linearly independent elements of $\mathbb{Z}^n$ needn't be a free basis.

Now consider elements $g_1, \ldots, g_n$ of an abelian group G. It is possible to extend the assignment $\phi(\mathbf{x}_i) = g_i$ to a group homomorphism $\phi : \mathbb{Z}^n \to G$. As a function we define $\phi((a_1, a_2, \ldots, a_n)^{\mathrm{T}}) = \sum_{i=1}^n a_i g_i$. We leave the proof of the following result as an exercise.

Proposition 5.4.4. *(i) The function ϕ is a group homomorphism.*

(ii) The set of elements $\{g_i\}$ are linearly independent if and only if ϕ is injective.

(iii) The set of elements $\{g_i\}$ span G if and only if ϕ is surjective.

(iv) The set of elements $\{g_i\}$ form a free basis of G if and only if ϕ is an isomorphism.

Note that this proposition makes perfect sense for vector spaces. Also note that the last statement implies that $g_1, \ldots, g_n$ is a free basis of G if and only if every element $g \in G$ has a unique expression $g = \alpha_1 g_1 + \cdots + \alpha_n g_n$ with $\alpha_i \in \mathbb{Z}$, very much like for vector spaces.

Before Proposition 5.4.4 we were trying to extend the assignment $\phi(\mathbf{x}_i) = g_i$ to a group homomorphism $\phi : \mathbb{Z}^n \to G$. Note that the extension we wrote is unique. This is the key to the next corollary. The details of the proof are left to the reader.

Corollary 5.4.5 (Universal property of the free abelian group)**.** *Let G be a free abelian group with a free basis $g_1, \ldots, g_n$. Let H be an abelian group and $a_1, \ldots a_n \in H$. Then there exists a unique group homomorphism $\phi : G \to H$ such that $\phi(g_i) = a_i$ for all i.*

As for finite dimensional vector spaces, it turns out that any two free bases of a free abelian group have the same size, but this has to be proved. It will follow directly from the next theorem.

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be the standard free basis of $\mathbb{Z}^n$, and let $\mathbf{y}_1, \ldots \mathbf{y}_m$ be another free basis. As in Linear Algebra, we can define the associated change of basis matrix P (with original basis $\{\mathbf{x}_i\}$ and new basis $\{\mathbf{y}_i\}$) , where the columns of P are $\mathbf{y}_i$; that is, they express $\mathbf{y}_i$ in terms of $\mathbf{x}_i$. For example, if $n = m = 2$, $\mathbf{y}_1 = \begin{pmatrix} 2 \\ 7 \end{pmatrix}$, $\mathbf{y}_2 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$, then $P = \begin{pmatrix} 2 & 1 \\ 7 & 4 \end{pmatrix}$. In general, $P = (\rho_{ij})$ is an $n \times m$ matrix with $\mathbf{y}_j = \sum_{i=1}^{n} \rho_{ij} \mathbf{x}_i$ for $1 \leq j \leq m$.

Theorem 5.4.6. *Let $\mathbf{y}_1, \ldots, \mathbf{y}_m \in \mathbb{Z}^n$ with $\mathbf{y}_j = \sum_{i=1}^{n} \rho_{ij} \mathbf{x}_i$ for $1 \leq j \leq m$. Then the following are equivalent:*

(i) $\mathbf{y}_1, \ldots, \mathbf{y}_m$ is a free basis of $\mathbb{Z}^n$;

(ii) $n = m$ and P is an invertible matrix such that P^{-1} has entries in $\mathbb{Z}$;

(iii) $n = m$ and $\det(P) = \pm 1$.

(A matrix $P \in \mathbb{Z}^{n,n}$ with $\det(P) = \pm 1$ is called *unimodular*.)

Proof. (i) $\Rightarrow$ (ii). If $\mathbf{y}_1, \ldots, \mathbf{y}_m$ is a free basis of $\mathbb{Z}^n$ then it spans $\mathbb{Z}^n$, so there is an $m \times n$ matrix $T = (\tau_{ij})$ with $\mathbf{x}_k = \sum_{j=1}^{m} \tau_{jk} \mathbf{y}_j$ for $1 \leq k \leq n$. Hence

$$\mathbf{x}_k = \sum_{j=1}^{m} \tau_{jk} \mathbf{y}_j = \sum_{j=1}^{m} \tau_{jk} \sum_{i=1}^{n} \rho_{ij} \mathbf{x}_i = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} \rho_{ij} \tau_{jk} \right) \mathbf{x}_i,$$

and, since $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is a free basis, this implies that $\sum_{j=1}^{m} \rho_{ij} \tau_{jk} = 1$ when $i = k$ and 0 when $i \neq k$. In other words $PT = I_n$, and similarly $TP = I_m$, so P and T are inverse matrices. But we

can think of P and T as inverse matrices over the field Q, so it follows from First Year Linear Algebra that $m = n$, and $T = P^{-1}$ has entries in $\mathbb{Z}$.

(ii) $\Rightarrow$ (i). If $T = P^{-1}$ has entries in $\mathbb{Z}$ then, again thinking of them as matrices over the field $\mathbb{Q}$, $\text{rank}(P) = n$, so the columns of P are linearly independent over Q and hence also over $\mathbb{Z}$. Since the columns of P are just the column vectors representing $\mathbf{y}_1, \dots \mathbf{y}_m$, this tells us that $\mathbf{y}_1, \dots \mathbf{y}_m$ are linearly independent.

Using $PT = I_n$, for $1 \leq k \leq n$ we have

$$\sum_{j=1}^{m} \tau_{jk}\mathbf{y}_j = \sum_{j=1}^{m} \tau_{jk} \sum_{i=1}^{n} \rho_{ij}\mathbf{x}_i = \sum_{i=1}^{n} (\sum_{j=1}^{m} \rho_{ij}\tau_{jk})\mathbf{x}_i = \mathbf{x}_k,$$

because $\sum_{j=1}^{m} \rho_{ij}\tau_{jk}$ is equal to 1 when $i = k$ and 0 when $i \neq k$. Since $\mathbf{x}_1, \dots \mathbf{x}_n$ spans $\mathbb{Z}^n$, and we can express each $\mathbf{x}_k$ as a linear combination of $\mathbf{y}_1, \dots \mathbf{y}_m$, it follows that $\mathbf{y}_1, \dots \mathbf{y}_m$ span $\mathbb{Z}^n$ and hence form a free basis of $\mathbb{Z}^n$.

(ii) $\Rightarrow$ (iii). If $T = P^{-1}$ has entries in $\mathbb{Z}$, then $\det(PT) = \det(P)\det(T) = \det(I_n) = 1$, and since $\det(P), \det(T) \in \mathbb{Z}$, this implies $\det(P) = \pm 1$.

(iii) $\Rightarrow$ (ii). From First year Linear Algebra, $P^{-1} = \frac{1}{\det(P)} \text{adj}(P)$, so $\det(P) = \pm 1$ implies that P^{-1} has entries in $\mathbb{Z}$. □

Example. If $n = 2$ and $\mathbf{y}_1 = \begin{pmatrix} 2 \\ 7 \end{pmatrix}$, $\mathbf{y}_2 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$, then $\det(P) = 8 - 7 = 1$, so $\mathbf{y}_1, \mathbf{y}_2$ is a free basis of $\mathbb{Z}^2$.

But, if $\mathbf{y}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\mathbf{y}_2 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$, then $\det(P) = 2$, so $\mathbf{y}_1, \mathbf{y}_2$ is not a free basis of $\mathbb{Z}^2$.

Recall that in Linear Algebra over a field, any set of n linearly independent vectors in a vector space V of dimension n form a basis of V. This example shows that this result is not true in $\mathbb{Z}^n$, because $\mathbf{y}_1$ and $\mathbf{y}_2$ are linearly independent but do not span $\mathbb{Z}^2$.

But as in Linear Algebra, for $\mathbf{v} \in \mathbb{Z}^n$, if $\mathbf{x}$ and $\mathbf{y}$ are the column vectors representing $\mathbf{v}$ using free bases $\mathbf{x}_1, \dots \mathbf{x}_n$ and $\mathbf{y}_1, \dots \mathbf{y}_n$, respectively, then we have $\mathbf{x} = P\mathbf{y}$, so $\mathbf{y} = P^{-1}\mathbf{x}$.

5.5 Unimodular elementary row and column operations and the Smith normal form for integral matrices

We interrupt our discussion of finitely generated abelian groups at this stage to investigate how the row and column reduction process of Linear Algebra can be adapted to matrices over $\mathbb{Z}$. Recall from MA106 that we can use elementary row and column operations to reduce an $m \times n$ matrix of rank r over a field K to a matrix $B = (\beta_{ij})$ with $\beta_{ii} = 1$ for $1 \leq i \leq r$ and $\beta_{ij} = 0$ otherwise. We called this the *Smith Normal Form* of the matrix. We can do something similar over $\mathbb{Z}$, but the non-zero elements β_{ii} will not necessarily all be equal to 1.

The reason that we disallowed $\lambda = 0$ for the row and column operations (R3) and (C3) (multiply a row or column by a scalar λ) was that we wanted all of our elementary operations to be reversible. When performed over $\mathbb{Z}$, (R1), (C1), (R2) and (C2) are reversible, but (R3) and (C3) are reversible only when $\lambda = \pm 1$. So, if A is an $m \times n$ matrix over $\mathbb{Z}$, then we define the three types of *unimodular* elementary row and column operations as follows:

(UR1): Replace some row $\mathbf{r}_i$ of A by $\mathbf{r}_i + t\mathbf{r}_j$, where $j \neq i$ and $t \in \mathbb{Z}$;

(UR2): Interchange two rows $\mathbf{r}_i$ and $\mathbf{r}_j$ of A;

(UR3): Replace some row $\mathbf{r}_i$ of A by $-\mathbf{r}_i$.

(UC1): Replace some column $\mathbf{c}_i$ of A by $\mathbf{c}_i + t\mathbf{c}_j$, where $j \neq i$ and $t \in \mathbb{Z}$;

(UC2): Interchange two columns $\mathbf{c}_i$ and $\mathbf{c}_j$ of A;

(UC3): Replace some column $\mathbf{c}_i$ of A by $-\mathbf{c}_i$.

Recall from MA106 that performing elementary row or column operations on a matrix A corresponds to multiplying A on the left or right, respectively, by an elementary matrix. These elementary matrices all have determinant ± 1 (1 for (UR1) and -1 for (UR2) and (UR3)), so are unimodular matrices over $\mathbb{Z}$.

By checking what left and right multiplying by these elementary matrices do we can check that the unimodular row and column operations correspond to the following change of bases, where $\mathbf{e}_1, \ldots, \mathbf{e}_n$ is a basis for $\mathbb{Z}^n$ (the domain of the linear map which A represents) and $\mathbf{f}_1, \ldots, \mathbf{f}_m$ is a basis for $\mathbb{Z}^m$ (the target of the linear map). Notice that column operations correspond to changing the basis of the domain, whereas row operations correspond to changing the basis of the target.

(UC1): $\mathbf{e}_i \to \mathbf{e}_i + t\mathbf{e}_j$; (UC2): $\mathbf{e}_i \leftrightarrow \mathbf{e}_j$; (UC3): $\mathbf{e}_i \to -\mathbf{e}_i$.

(UR1): $\mathbf{f}_j \to \mathbf{f}_j - t\mathbf{f}_i$; (UR2): $\mathbf{f}_i \leftrightarrow \mathbf{f}_j$; (UR3): $\mathbf{f}_i \to -\mathbf{f}_i$.

Theorem 5.5.1. *Let A be an $m \times n$ matrix over $\mathbb{Z}$ with rank r. Then, by using a sequence of unimodular elementary row and column operations, we can reduce A to a matrix $B = (\beta_{ij})$ with $\beta_{ii} = d_i$ for $1 \leq i \leq r$ and $\beta_{ij} = 0$ otherwise, and where the integers d_i satisfy $d_i > 0$ for $1 \leq i \leq r$, and $d_i | d_{i+1}$ for $1 \leq i < r$. Subject to these conditions, the d_i are uniquely determined by the matrix A.*

Proof. We shall not prove the uniqueness part here. The fact that the number of non-zero β_{ii} is the rank of A follows from the fact that unimodular row and column operations do not change the rank. We use induction on $m + n$. The base case is $m = n = 1$, where there is nothing to prove. Also if A is the zero matrix then there is nothing to prove, so assume not.

Let d be the smallest entry with $d > 0$ in any matrix $C = (\gamma_{ij})$ that we can obtain from A by using unimodular elementary row and column operations. By using (R2) and (C2), we can move d to position $(1, 1)$ and hence assume that $\gamma_{11} = d$. If d does not divide γ_{1j} for some $j > 0$, then we can write $\gamma_{1j} = qd + r$ with $q, r \in \mathbb{Z}$ and $0 < r < d$, and then replacing the j-th column $\mathbf{c}_j$ of C by $\mathbf{c}_j - q\mathbf{c}_1$ results in the entry r in position $(1, j)$, contrary to the choice of d. Hence $d | \gamma_{1j}$ for $2 \leq j \leq n$ and similarly $d | \gamma_{i1}$ for $2 \leq i \leq m$.

Now, if $\gamma_{1j} = qd$, then replacing $\mathbf{c}_j$ of C by $\mathbf{c}_j - q\mathbf{c}_1$ results in entry 0 position $(1, j)$. So we can assume that $\gamma_{1j} = 0$ for $2 \leq j \leq n$ and $\gamma_{i1} = 0$ for $2 \leq i \leq m$. If $m = 1$ or $n = 1$, then we are done. Otherwise, we have $C = (d) \oplus C'$ for some $(m-1) \times (n-1)$ matrix C'. By inductive hypothesis, the result of the theorem applies to C', so by applying unimodular row and column operations to C which do not involve the first row or column, we can reduce C to $D = (\delta_{ij})$, which satisfies $\delta_{11} = d$, $\delta_{ii} = d_i > 0$ for $2 \leq i \leq r$, and $\delta_{ij} = 0$ otherwise, where $d_i | d_{i+1}$ for $2 \leq i < r$. To complete the proof, we still have to show that $d | d_2$. If not, then adding row 2 to row 1 results in d_2 in position (1,2) not divisible by d, and we obtain a contradiction as before. □

So how do we find the Smith Normal Form of a matrix then? The general strategy is to reduce the size of entries in the first row and column, until the (1,1)-entry divides all other entries in the first row and column. Then we can clear all of these other entries with repeated use of (UR1) and (UC1). Let us elaborate on this strategy. First we state a useful result in finding the SNF. We leave the proof as an exercise.

Lemma 5.5.2. *Let $A \in \mathbb{Z}^{m,n}$ with SNF having non-zero diagonal entries $d_1, \ldots, d_r$. Then the greatest common divisor of all of the entires of A is equal to d_1.*

One can generalise this to the greatest common divisor of $k \times k$ minors. You may want to think about what they tell you about the SNF. This is one way to prove uniqueness of the SNF (see Assignment 5).

Strategy for finding the SNF of $A \in \mathbb{Z}^{m,n}$

Step 1: Find d_1, using Lemma 5.5.2.

Step 2: If d_1 occurs as an entry in A move it to the $(1,1)$ entry using (UC2) and (UR2). If $-d_1$ occurs then use (UR3) or (UC3) and then move it into the $(1,1)$ entry. This is the easier scenario.

If d_1 does not occur then we need to do some (or lots) of division with remainder. Let x be the smallest entry (with respect to absolute value) occurring in A, say in position (i, j).

Now, we again have a dichotomy into an easier case and a harder case. Does x divide everything else in the ith row and jth column? If not, let y be an entry that x does not divide, in position (k, l) with $k = i$ or $l = j$. Then $y = sx + r$ with $r < x$ and using (UC1) or (UR1) we can obtain $r = y - sx$ as an entry of A (add $-sx$ of row i to row k or $-sx$ of column j to column l). At this point we have reduced the smallest entry in $\tilde{A}$ (the matrix obtained from A by doing the unimodular row or column operation required) and so we can start Step 2 again.

Now it remains to deal with the case where x divides everything else in row i and column j. We start by clearing all of these entries to 0. This is straightforward using (UC1) and (UR1), and is possible since x divides each entry:

$$r_d \to r_d - \frac{a_{d,j}}{x} r_i, \quad d \neq i$$

and

$$c_d \to c_d - \frac{a_{i,d}}{x} c_j, \quad d \neq j$$

do the job. We know there still exists an element in $\tilde{A}$ which is not divisible by x, again let y be such an entry, in position (k,l). Then looking at the intersection of row i and k with column j and l we find a 2×2-matrix which looks as follows (assuming $i<k$ and $j<l$, otherwise the picture is similar but x and y are in different positions, still opposite each other diagonally):

$$\begin{pmatrix} x & 0 \\ 0 & y \end{pmatrix}.$$

We want $r = y - sx$ with $r < x$ to appear in $\tilde{A}$. We can do this in two moves. First use (UR1) to add $-s$ times row i to row k. We then use (UC1) to add column j to column l. This will result in r appearing where y was before:

$$\begin{pmatrix} x & -sx \\ -sx & r = y - sx \end{pmatrix}.$$

(There are other ways to do this, feel free to experiment!) Again, we have reduced the smallest entry in $\tilde{A}$ and so we can start Step 2 again.

Step 3: With d_1 in the $(1,1)$ entry, we can use it to clear everything else in the first row and column since d_1 divides all entries in the matrix (just as we did in Step 2).

Step 4: The matrix $\tilde{A}$ now has entry d_1 in the $(1,1)$ entry and 0s elsewhere in row and column 1. We now go back to Step 1, working on the $m-1\times n-1$ matrix in the bottom right hand corner of $\tilde{A}$. We can repeat Steps 1 to 3 without changing the entries in row 1 and column 1. Therefore, repeating this process will terminate and will yield the SNF of A.

Example 17. $A = \begin{pmatrix} 42 & 21 \\ -35 & -14 \end{pmatrix}$.

We calculate that $d_1 = \gcd(42,21,-35,-14) = 7$. It does not appear in A so we need to use unimodular row and column operations to make that happen. It is often easy in practice to see how to do this and the general strategy should be seen as a guide rather than an algorithm you must follow. On this occasion we notice that dividing -35 by -14 we get remainder $-7 = -d_1$. We can then negate row 2 and we see 7 occurring in the matrix. Everything is straightforward from that point onwards.

Matrix	Operation	Matrix	Operation
$\begin{pmatrix} 42 & 21 \\ -35 & -14 \end{pmatrix}$	$\mathbf{c}_1 \to \mathbf{c}_1 - 2\mathbf{c}_2$	$\begin{pmatrix} 0 & 21 \\ -7 & -14 \end{pmatrix}$	$\mathbf{r}_2 \to -\mathbf{r}_2$
			$\mathbf{r}_1 \leftrightarrow \mathbf{r}_2$
$\begin{pmatrix} 7 & 14 \\ 0 & 21 \end{pmatrix}$	$\mathbf{c}_2 \to \mathbf{c}_2 - 2\mathbf{c}_1$	$\begin{pmatrix} 7 & 0 \\ 0 & 21 \end{pmatrix}$	

Example 18. $A = \begin{pmatrix} -18 & -18 & -18 & 90 \\ 54 & 12 & 45 & 48 \\ 9 & -6 & 6 & 63 \\ 18 & 6 & 15 & 12 \end{pmatrix}$.

This time we want to notice what the greatest common divisor of the entries is without doing too many calculations. Firstly, we spot 9 and 6, so d_1 must be 1 or 3. Looking at all the other entries we see they are all divisible by 3. So $d_1 = 3$. Again, this does not appear in A so we need to use unimodular row and column operations to make that happen. We spot 9 and 6 in row 3, so we can get 3 to appear just by adding the negative of column 3 to column 1 (plenty of other choices here!). We carry on in this way to obtain the SNF.

Matrix	**Operation**	**Matrix**	**Operation**
$\begin{pmatrix} -18 & -18 & -18 & 90 \\ 54 & 12 & 45 & 48 \\ 9 & -6 & 6 & 63 \\ 18 & 6 & 15 & 12 \end{pmatrix}$	$\mathbf{c}_1 \to \mathbf{c}_1 - \mathbf{c}_3$	$\begin{pmatrix} 0 & -18 & -18 & 90 \\ 9 & 12 & 45 & 48 \\ 3 & -6 & 6 & 63 \\ 3 & 6 & 15 & 12 \end{pmatrix}$	$\mathbf{r}_1 \leftrightarrow \mathbf{r}_4$
$\begin{pmatrix} 3 & 6 & 15 & 12 \\ 9 & 12 & 45 & 48 \\ 3 & -6 & 6 & 63 \\ 0 & -18 & -18 & 90 \end{pmatrix}$	$\mathbf{r}_2 \to \mathbf{r}_2 - 3\mathbf{r}_1$ $\mathbf{r}_3 \to \mathbf{r}_3 - \mathbf{r}_1$	$\begin{pmatrix} 3 & 6 & 15 & 12 \\ 0 & -6 & 0 & 12 \\ 0 & -12 & -9 & 51 \\ 0 & -18 & -18 & 90 \end{pmatrix}$	$\mathbf{c}_2 \to \mathbf{c}_2 - 2\mathbf{c}_1$ $\mathbf{c}_3 \to \mathbf{c}_3 - 5\mathbf{c}_1$ $\mathbf{c}_4 \to \mathbf{c}_4 - 4\mathbf{c}_1$
$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & -6 & 0 & 12 \\ 0 & -12 & -9 & 51 \\ 0 & -18 & -18 & 90 \end{pmatrix}$	$\mathbf{c}_2 \to -\mathbf{c}_2$ $\mathbf{c}_2 \to \mathbf{c}_2 + \mathbf{c}_3$	$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 6 & 0 & 12 \\ 0 & 3 & -9 & 51 \\ 0 & 0 & -18 & 90 \end{pmatrix}$	$\mathbf{r}_2 \leftrightarrow \mathbf{r}_3$
$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & -9 & 51 \\ 0 & 6 & 0 & 12 \\ 0 & 0 & -18 & 90 \end{pmatrix}$	$\mathbf{r}_3 \to \mathbf{r}_3 - 2\mathbf{r}_2$	$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & -9 & 51 \\ 0 & 0 & 18 & -90 \\ 0 & 0 & -18 & 90 \end{pmatrix}$	$\mathbf{c}_3 \to \mathbf{c}_3 + 3\mathbf{c}_2$ $\mathbf{c}_4 \to \mathbf{c}_4 - 17\mathbf{c}_2$
$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 18 & -90 \\ 0 & 0 & -18 & 90 \end{pmatrix}$	$\mathbf{c}_4 \to \mathbf{c}_4 + 5\mathbf{c}_3$ $\mathbf{r}_4 \to \mathbf{r}_4 + \mathbf{r}_3$	$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 18 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	

Note: There is also a generalisation to integer matrices of the the row reduced normal form from Linear Algebra, where only row operations are allowed. This is known as the *Hermite Normal Form* and is more complicated.

5.6 Subgroups of free abelian groups

Proposition 5.6.1. *Any subgroup of a finitely generated abelian group is finitely generated.*

Proof. Let $K < G$ with G an abelian group generated by $x_1, \ldots, x_n$. We shall prove by induction on n that K can be generated by at most n elements. If $n = 1$ then G is cyclic. Write $G = \{nx | n \in \mathbb{Z}\}$. Let m be the smallest positive number such that $mx \in K$. If such a number does

not exist then $K = \{0\}$. Otherwise, $K \supseteq \{nmx | n \in \mathbb{Z}\}$. The opposite inclusion follows using division with a remainder: write $t = qm + r$ with $0 \leq r < m$. Then $tx \in K$ if and only if $rx = (t - mq)x \in K$ if and only if $r = 0$ due to minimality of m. In both cases K is cyclic.

Suppose $n > 1$, and let H be the subgroup of G generated by $x_1, \ldots, x_{n-1}$. By induction, $K \cap H$ is generated by $y_1, \ldots, y_{m-1}$, say, with $m \leq n$. If $K \leq H$, then $K = K \cap H$ and we are done, so suppose not.

Then there exist elements of the form $h + tx_n \in K$ with $h \in H$ and $t \neq 0$. Since $-(h + tx_n) \in K$, we can assume that $t > 0$. Choose such an element $y_m = h + tx_n \in K$ with t minimal subject to $t > 0$. We claim that K is generated by $y_1, \ldots, y_m$, which will complete the proof. Let $k \in K$. Then $k = h' + ux_n$ with $h' \in H$ and $u \in \mathbb{Z}$. If t does not divide u then we can write $u = tq + r$ with $q, r \in \mathbb{Z}$ and $0 < r < t$, and then $k - qy_m = (h' - qh) + rx_n \in K$, contrary to the choice of t. So $t|u$ and hence $u = tq$ and $k - qy_m \in K \cap H$. But $K \cap H$ is generated by $y_1, \ldots, y_{m-1}$, so we are done. □

Now let H be a subgroup of the free abelian group $\mathbb{Z}^n$, and suppose that H is generated by $\mathbf{v}_1, \ldots, \mathbf{v}_m$. Then H can be represented by an $n \times m$ matrix A in which the columns are $\mathbf{v}_1, \ldots, \mathbf{v}_m$.

Example 19. If $n = 3$ and H is generated by $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 3 \\ -1 \end{pmatrix}$ and $\mathbf{v}_2 = \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}$, then

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

As we saw above, if we use a different free basis $\mathbf{y}_1, \ldots, \mathbf{y}_n$ of $\mathbb{Z}^n$ with basis change matrix P, then each column $\mathbf{v}_j$ of A is replaced by $P^{-1}\mathbf{v}_j$, and hence A itself is replaced by $P^{-1}A$.

So in Example 19, if we use the basis $\mathbf{y}_1 = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}$, $\mathbf{y}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$, $\mathbf{y}_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ of $\mathbb{Z}^3$, then

$$P = \begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad P^{-1} = \begin{pmatrix} 1 & -1 & -1 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{pmatrix}, \quad P^{-1}A = \begin{pmatrix} -1 & 1 \\ -1 & 1 \\ 2 & 1 \end{pmatrix}.$$

For example, the first column $\begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}$ of $P^{-1}A$ represents $-\mathbf{y}_1 - \mathbf{y}_2 + 2\mathbf{y}_3 = \begin{pmatrix} 1 \\ 3 \\ -1 \end{pmatrix} = \mathbf{v}_1$.

In particular, if we perform a unimodular elementary row operation on A, then the resulting matrix represents the same subgroup H of $\mathbb{Z}^n$ but using a different free basis of $\mathbb{Z}^n$.

We can clearly replace a generator $\mathbf{v}_i$ of H by $\mathbf{v}_i + r\mathbf{v}_j$ for $r \in \mathbb{Z}$ without changing the subgroup H that is generated. We can also interchange two of the generators or replace one of the generators $\mathbf{v}_i$ by $-\mathbf{v}_i$ without changing H. In other words, performing a unimodular elementary

column operation on A amounts to changing the generating set for H, so again the resulting matrix still represents the same subgroup H of $\mathbb{Z}^n$.

Summing up, we have:

Proposition 5.6.2. *Suppose that the subgroup H of $\mathbb{Z}^n$ is represented by the matrix $A \in \mathbb{Z}^{n,m}$. Then if the matrix $B \in \mathbb{Z}^{n,m}$ is obtained by performing a sequence of unimodular row and column operations on A, then B represents the same subgroup H of $\mathbb{Z}^n$ using a (possibly) different free basis of $\mathbb{Z}^n$.*

In particular, by Theorem 5.5.1, we can transform A to a matrix B in Smith Normal Form. So, then if B represents H with the free basis $\mathbf{y}_1, \ldots, \mathbf{y}_n$ of $\mathbb{Z}^n$, then the r non-zero columns of B correspond to the elements $d_1\mathbf{y}_1, d_2\mathbf{y}_2, \ldots, d_r\mathbf{y}_r$ of $\mathbb{Z}^n$. So we have:

Theorem 5.6.3. *Let H be a subgroup of $\mathbb{Z}^n$. Then there exists a free basis $\mathbf{y}_1, \ldots, \mathbf{y}_n$ of $\mathbb{Z}^n$ such that $H = \langle\, d_1\mathbf{y}_1, d_2\mathbf{y}_2, \ldots, d_r\mathbf{y}_r \,\rangle$, where each $d_i > 0$ and $d_i | d_{i+1}$ for $1 \leq i < r$.*

In Example 19, it is straightforward to calculate the Smith Normal Form of A, which is $\begin{pmatrix} 1 & 0 \\ 0 & 3 \\ 0 & 0 \end{pmatrix}$, so $H = \langle \mathbf{y}_1, 3\mathbf{y}_2 \rangle$.

By keeping track of the unimodular row operations carried out, we can, if we need to, find the free basis $\mathbf{y}_1, \ldots, \mathbf{y}_n$ of $\mathbb{Z}^n$ such that H has this nice form. Using the formulae in Section 5.5, noting that we start from the standard free basis, we can do this in Example 19.

Matrix	Operation	New free basis
$\begin{pmatrix} 1 & 2 \\ 3 & 0 \\ -1 & 1 \end{pmatrix}$	$\mathbf{r}_2 \to \mathbf{r}_2 - 3\mathbf{r}_1$	$\mathbf{y}_1 = \begin{pmatrix} 1 \\ 3 \\ 0 \end{pmatrix}, \mathbf{y}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{y}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$
$\begin{pmatrix} 1 & 2 \\ 3 & 0 \\ -1 & 1 \end{pmatrix}$	$\mathbf{r}_3 \to \mathbf{r}_3 + \mathbf{r}_1$	$\mathbf{y}_1 = \begin{pmatrix} 1 \\ 3 \\ -1 \end{pmatrix}, \mathbf{y}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{y}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$
$\begin{pmatrix} 1 & 2 \\ 0 & -6 \\ 0 & 3 \end{pmatrix}$	$\mathbf{c}_2 \to \mathbf{c}_2 - 2\mathbf{c}_1$	$\mathbf{y}_1 = \begin{pmatrix} 1 \\ 3 \\ -1 \end{pmatrix}, \mathbf{y}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{y}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$
$\begin{pmatrix} 1 & 0 \\ 0 & -6 \\ 0 & 3 \end{pmatrix}$	$\mathbf{r}_2 \to \mathbf{r}_3$	$\mathbf{y}_1 = \begin{pmatrix} 1 \\ 3 \\ -1 \end{pmatrix}, \mathbf{y}_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \mathbf{y}_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$
$\begin{pmatrix} 1 & 0 \\ 0 & 3 \\ 0 & -6 \end{pmatrix}$	$\mathbf{r}_3 \to \mathbf{r}_3 + 2r_2$	$\mathbf{y}_1 = \begin{pmatrix} 1 \\ 3 \\ -1 \end{pmatrix}, \mathbf{y}_2 = \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}, \mathbf{y}_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$
$\begin{pmatrix} 1 & 0 \\ 0 & 3 \\ 0 & 0 \end{pmatrix}$		

5.7 General finitely generated abelian groups

Let G be a finitely generated abelian group. If G has n generators, Proposition 5.4.4 gives a surjective homomorphism $\phi : \mathbb{Z}^n \to G$. From the First isomorphism Theorem (Theorem 5.3.7) we deduce that $G \cong \mathbb{Z}^n / K$, where $K = \ker(\phi)$. So we have proved that every finitely generated abelian group is isomorphic to a quotient group of a free abelian group.

From the definition of ϕ, we see that

$$K = \{ (\alpha_1, \alpha_2, \ldots, \alpha_n)^{\mathrm{T}} \in \mathbb{Z}^n \mid \alpha_1 x_1 + \cdots + \alpha_n x_n = 0_G \}.$$

By Theorem 5.6.1, this subgroup K is generated by finitely many elements $\mathbf{v}_1, \ldots, \mathbf{v}_m$ of $\mathbb{Z}^n$. The notation

$$\langle \mathbf{x}_1, \ldots, \mathbf{x}_n \mid \mathbf{v}_1, \ldots, \mathbf{v}_m \rangle$$

is often used to denote the quotient group $\mathbb{Z}^n / K$, so we have

$$G \cong \langle \mathbf{x}_1, \ldots, \mathbf{x}_n \mid \mathbf{v}_1, \ldots, \mathbf{v}_m \rangle.$$

Now we can apply Theorem 5.6.3 to this subgroup K, and deduce that there is a free basis $\mathbf{y}_1, \ldots, \mathbf{y}_n$ of $\mathbb{Z}^n$ such that $K = \langle\, d_1\mathbf{y}_1, \ldots, d_r\mathbf{y}_r \,\rangle$ for some $r \leq n$, where each $d_i > 0$ and $d_i | d_{i+1}$ for $1 \leq i < r$.

So we also have

$$G \cong \langle\, \mathbf{y}_1, \ldots, \mathbf{y}_n \mid d_1\mathbf{y}_1, \ldots, d_r\mathbf{y}_r \,\rangle,$$

and G has generators $y_1, \ldots, y_n$ with $d_i y_i = 0$ for $1 \leq i \leq r$.

Proposition 5.7.1. *The group*

$$\langle\, \mathbf{y}_1, \ldots, \mathbf{y}_n \mid d_1\mathbf{y}_1, \ldots, d_r\mathbf{y}_r \,\rangle$$

is isomorphic to the direct sum of cyclic groups

$$\mathbb{Z}_{d_1} \oplus \mathbb{Z}_{d_2} \oplus \ldots \oplus \mathbb{Z}_{d_r} \oplus \mathbb{Z}^{n-r}.$$

Proof. This is another application of the First Isomorphism Theorem. Let $H = \mathbb{Z}_{d_1} \oplus \mathbb{Z}_{d_2} \oplus \mathbb{Z}_{d_r} \oplus \mathbb{Z}^{n-r}$, so H is generated by $y_1, \ldots, y_n$, with $y_1 = (1, 0, \ldots, 0), \ldots, y_n = (0, 0, \ldots, 1)$. Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be the standard free basis of $\mathbb{Z}^n$. Then, by Proposition 5.4.4, there is a surjective homomorphism ϕ from $\mathbb{Z}^n$ to H for which

$$\phi(\alpha_1\mathbf{y}_1 + \cdots + \alpha_n\mathbf{y}_n) = \alpha_1 y_1 + \cdots + \alpha_n y_n$$

for all $\alpha_1, \ldots, \alpha_n \in \mathbb{Z}$. Then, by Theorem 5.3.7, we have $H \cong \mathbb{Z}^n / K$, with

$$K = \{\, (\alpha_1, \alpha_2, \ldots, \alpha_n)^{\mathrm{T}} \in \mathbb{Z}^n \mid \alpha_1 y_1 + \cdots + \alpha_n y_n = 0_H \,\}.$$

Now $\alpha_1 y_1 + \cdots + \alpha_n y_n$ is the element $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ of H, which is the zero element if and only if α_i is the zero element of $\mathbb{Z}_{d_i}$ for $1 \leq i \leq r$ and $\alpha_i = 0$ for $r + 1 \leq i \leq n$.

But α_i is the zero element of $\mathbb{Z}_{d_i}$ if and only if $d_i | \alpha_i$, so we have

$$K = \{\, (\alpha_1, \alpha_2, \ldots, \alpha_r, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{Z}^n \mid d_i | \alpha_i \text{ for } 1 \leq i \leq r \,\}$$

which is generated by the elements $d_1\mathbf{y}_1, \ldots, d_r\mathbf{y}_r$. So

$$H \cong \mathbb{Z}^n / K = \langle\, \mathbf{y}_1, \ldots, \mathbf{y}_n \mid d_1\mathbf{y}_1, \ldots, d_r\mathbf{y}_r \,\rangle.$$

□

Putting all of these results together, we get the main theorem:

Theorem 5.7.2 (The fundamental theorem of finitely generated abelian groups)**.** *If G is a finitely generated abelian group, then G is isomorphic to a direct sum of cyclic groups. More precisely, if G is generated by n elements then, for some r with $0 \leq r \leq n$, there are integers $d_1, \ldots, d_r$ with $d_i > 0$ and $d_i | d_{i+1}$ such that*

$$G \cong \mathbb{Z}_{d_1} \oplus \mathbb{Z}_{d_2} \oplus \ldots \oplus \mathbb{Z}_{d_r} \oplus \mathbb{Z}^{n-r}.$$

So G is isomorphic to a direct sum of r finite cyclic groups of orders $d_1, \ldots, d_r$, and $n - r$ infinite cyclic groups.

There may be some factors $\mathbb{Z}_1$, the trivial group of order 1. These can be omitted from the direct sum (except in the case when $G \cong \mathbb{Z}_1$ is trivial). It can be deduced from the uniqueness part of Theorem 5.5.1, which we did not prove, that the numbers in the sequence $d_1, d_2, \ldots, d_r$ that are greater than 1 are uniquely determined by G.

Note that, $n - r$ may be 0, which is the case if and only if G is finite. At the other extreme, if all $d_i = 1$, then G is free abelian.

The group G corresponding to Example 17 in Section 5.5 is

$$\langle\, \mathbf{x}_1, \mathbf{x}_2 \mid 42\mathbf{x}_1 - 35\mathbf{x}_2,\ 21\mathbf{x}_1 - 14\mathbf{x}_2 \,\rangle$$

and we have $G \cong \mathbb{Z}_7 \oplus \mathbb{Z}_{21}$, a group of order $7 \times 21 = 147$.

The group defined by Example 18 in Section 5.5 is

$$\begin{aligned}\langle\, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4 \mid\ & -18\mathbf{x}_1 + 54\mathbf{x}_2 + 9\mathbf{x}_3 + 18\mathbf{x}_4,\quad -18\mathbf{x}_1 + 12\mathbf{x}_2 - 6\mathbf{x}_3 + 6\mathbf{x}_4,\\ & -18\mathbf{x}_1 + 45\mathbf{x}_2 + 6\mathbf{x}_3 + 15\mathbf{x}_4,\quad 90\mathbf{x}_1 + 48\mathbf{x}_2 + 63\mathbf{x}_3 + 12\mathbf{x}_4 \,\rangle,\end{aligned}$$

which is isomorphic to $\mathbb{Z}_3 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_{18} \oplus \mathbb{Z}$, and is an infinite group with a (maximal) finite subgroup of order $3 \times 3 \times 18 = 162$,

The group defined by Example 19 in Section 5.6 is

$$\langle\, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \mid \mathbf{x}_1 + 3\mathbf{x}_2 - \mathbf{x}_3,\ 2\mathbf{x}_1 + \mathbf{x}_3 \,\rangle,$$

and is isomorphic to $\mathbb{Z}_1 \oplus \mathbb{Z}_3 \oplus \mathbb{Z} \cong \mathbb{Z}_3 \oplus \mathbb{Z}$, so it is infinite, with a finite subgroup of order 3.

5.8 Finite abelian groups

In particular, for any finite abelian group G, we have $G \cong \mathbb{Z}_{d_1} \oplus \mathbb{Z}_{d_2} \oplus \cdots \oplus \mathbb{Z}_{d_r}$, where $d_i | d_{i+1}$ for $1 \leq i < r$, and $|G| = d_1 d_2 \cdots d_r$.

From the uniqueness part of Theorem 5.5.1 (which we did not prove), it follows that, if $d_i | d_{i+1}$ for $1 \leq i < r$ and $e_i | e_{i+1}$ for $1 \leq i < s$. then $\mathbb{Z}_{d_1} \oplus \mathbb{Z}_{d_2} \oplus \mathbb{Z}_{d_r} \cong \mathbb{Z}_{e_1} \oplus \mathbb{Z}_{e_2} \oplus \mathbb{Z}_{e_s}$ if and only if $r = s$ and $d_i = e_i$ for $1 \leq i \leq r$.

So the isomorphism classes of finite abelian groups of order $n > 0$ are in one-one correspondence with expressions $n = d_1 d_2 \cdots d_r$ for which $d_i | d_{i+1}$ for $1 \leq i < r$. This enables us to classify isomorphism classes of finite abelian groups.

Examples.

1. $n = 4$. The decompositions are 4 and 2×2, so $G \cong \mathbb{Z}_4$ or $\mathbb{Z}_2 \oplus \mathbb{Z}_2$.
2. $n = 15$. The only decomposition is 15, so $G \cong \mathbb{Z}_{15}$ is necessarily cyclic.
3. $n = 36$. Decompositions are 36, 2×18, 3×12 and 6×6, so $G \cong \mathbb{Z}_{36}$, $\mathbb{Z}_2 \oplus \mathbb{Z}_{18}$, $\mathbb{Z}_3 \oplus \mathbb{Z}_{12}$ and $\mathbb{Z}_6 \oplus \mathbb{Z}_6$.

Although we have not proved in general that groups of the same order but with different decompositions of the type above are not isomorphic, this can always be done in specific examples by looking at the orders of elements.

We saw in an exercise above that if $\phi : G \to H$ is an isomorphism then $|g| = |\phi(g)|$ for all $g \in G$. So isomorphic groups have the same number of elements of each order.

Note also that, if $g = (g_1, g_2, \ldots, g_n)$ is an element of a direct sum of n groups, then $|g|$ is the least common multiple of the orders $|g_i|$ of the components of g.

So, in the four groups of order 36, $G_1 = \mathbb{Z}_{36}$, $G_2 = \mathbb{Z}_2 \oplus \mathbb{Z}_{18}$, $G_3 = \mathbb{Z}_3 \oplus \mathbb{Z}_{12}$ and $G_4 = \mathbb{Z}_6 \oplus \mathbb{Z}_6$, we see that only G_1 contains elements of order 36. Hence G_1 cannot be isomorphic to G_2, G_3 or G_4. Of the three groups G_2, G_3 and G_4, only G_2 contains elements of order 18, so G_2 cannot be isomorphic to G_3 or G_4. Finally, G_3 has elements of order 12 but G_4 does not, so G_3 and G_4 are not isomorphic, and we have now shown that no two of the four groups are isomorphic to each other.

As a slightly harder example, $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_4$ is not isomorphic to $\mathbb{Z}_4 \oplus \mathbb{Z}_4$, because the former has 7 elements of order 2, whereas the latter has only 3.

6 Vistas

Week 9

In this section we present four topics of different nature that were part of Algebra-1 in the past. They are not examinable but they can help to broaden your horizons.

6.1 Heisenberg uncertainty

With all the linear algebra we know it is a little step aside to understand basics of quantum mechanics. We discuss Schrödinger's picture[10] of quantum mechanics and derive (mathematically) *Heisenberg's Uncertainty Principle*.

The main ingredient of quantum mechanics is a Hilbert space $(V, <, >)$. There are physical arguments showing that euclidean vector spaces are no good and that V must be infinite-dimensional[11]. Here we just take their conclusions at face value. The states of the system are lines in V. We denote by $[\mathbf{v}]$ the line $\mathbb{C}\mathbf{v}$ spanned by $\mathbf{v} \in V$. We use *normalised* vectors, i.e., $\mathbf{v}$ such that $< \mathbf{v}, \mathbf{v} >= 1$ to present states as this makes formulas slightly easier.

It is impossible to observe the state of the quantum system but we can try to observe some physical quantities such as momentum, energy, spin, etc. Such physical quantities become *observables*, i.e., selfadjoint linear operators $\Phi : V \to V$. Selfadjoint in this context means that $< x, \Phi y >=< \Phi x, y >$ for all $x, y \in V$. Sweeping a subtle mathematical point under the carpet[12], we assume that Φ is diagonalisable with eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots$ and eigenvalues $\phi_1, \phi_2, \ldots$. Proof of Theorem 4.10.7 goes through in the infinite dimensional case, so we conclude that all ϕ_i belong to $\mathbb{R}$. Back to Physics, if we measure Φ on a state $[\mathbf{v}]$ with normalised $\mathbf{v} = \sum_n \alpha_n \mathbf{e}_n$ then the measurement will return ϕ_n as a result with probability $|\alpha_n|^2$

One observable is energy $H : V \to V$, often called hamiltonian. It is central to the theory because it determines the time evolution $[\mathbf{v}(t)]$ of the system by Shrodinger's equation:

$$\frac{d\mathbf{v}(t)}{dt} = \frac{1}{i\hbar} H\mathbf{v}(t)$$

where $\hbar \approx 10^{-34}$ Joule per second[13] is the reduced Planck constant. We know how to solve this equation: $\mathbf{v}(t) = e^{tH/i\hbar}\mathbf{v}(0)$.

As a concrete example, let us look at the quantum oscillator. The full energy of the classical harmonic oscillator mass m and frequency ω is

$$h = \frac{p^2}{2m} + \frac{1}{2} m\omega^2 x^2$$

[10] The alternative is Heisenberg's picture but we have no time to discuss it here.

[11] It is not as in Definition 4.10.3. Another subtle mathematical point is completness: we will overlook that one too.

[12] If V were finite-dimensional we could have used Theorem 4.10.7. But V is infinite dimensional! To ensure diagonalisability V must be complete with respect to the hermitian norm. Such spaces are called Hilbert spaces. Diagonalisability is still subtle as eigenvectors do not span the whole V but only a dense subspace. Furthermore, if V admits no dense countably dimensional subspace, further difficulties arise... Pandora box of functional analysis is wide open, so let us try to keep it shut.

[13] Notice the physical dimensions: H is energy, t is time, i dimensionless, $\hbar$ equalises the dimensions in the both sides irrespectively of what $\mathbf{v}$ is.

where x is the position and $p = mx'$ is the momentum. To quantise it, we have to play with this expression. The vector subspace of the space of all smooth functions $C^\infty(\mathbb{R}, \mathbb{C})$ admits a convenient subspace $V = \{f(x)e^{-x^2/2} \mid f(x) \in \mathbb{C}[x]\}$, which we make Hilbert by defining

$$< \phi(x), \psi(x) >:= \int_{-\infty}^{\infty} \bar{\phi}(x)\psi(x)dx$$

and then completing it with respect to this norm (we will largely disregard this completion). Quantum momentum and quantum position are linear operators (observables) on this space:

$$P(f(x)) = -i\hbar f'(x), \;\; X(f(x)) = f(x) \cdot x.$$

The quantum Hamiltonian is a second order differential operator operator given by the same equation

$$H = \frac{P^2}{2m} + \frac{1}{2}m\omega^2 X^2 = -\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + \frac{1}{2}m\omega^2 x^2.$$

As mathematicians, we can assume that $m = 1$ and $\omega = 1$, so that $H(f) = (fx^2 - f'')/2$. The eigenvectors of H are Hermite functions

$$\Psi_n(x) = (-1)^n e^{x^2/2}(e^{-x^2})^{(n)} \;,\; n = 0, 1, 2 \ldots$$

with eigenvalues $n + 1/2$ which are discrete *energy levels* of the quantum oscillator. Notice that $< \Psi_k, \Psi_n >= \delta_{k,n} 2^n n! \sqrt{\pi}$, so they are orthogonal but not orthonormal. The states $[\Psi_n]$ are *pure* states: they do not change with time and always give $n + 1/2$ as energy. If we take a system in a state $[\mathbf{v}]$ where

$$\mathbf{v} = \sum_n \alpha_n \frac{1}{\pi^4 2^{n/2} n!}\Psi_n$$

is normalised, then the measurement of energy will return $n + 1/2$ with probability $|\alpha_n|^2$ Notice that the measurement breaks the system!! It changes it to the state $[\Psi_n]$ and all future measurements will return the same energy!

Alternatively, it is possible to model the quantum oscillator on the vector space $W = \mathbb{C}[x]$ of polynomials. One has to use the natural linear bijection

$$\alpha : W \to V, \; \alpha(f(x)) = f(x)e^{-x^2/2}$$

and transfer all the formulas to W. The metric becomes

$$< f, g >=< \alpha(f), \alpha(g) >= \int_{-\infty}^{\infty} \bar{f}(x)g(x)e^{-x^2}dx \;,$$

the formulas for P and X changes accordingly, and at the end one arrives at Hermite polynomials $\alpha^{-1}(\Psi_n(x)) = (-1)^n e^{x^2}(e^{-x^2})^{(n)}$ instead of Hermite functions.

Let us go back to an abstract system with two observables P and Q. It is pointless to measure Q after measuring P as the system is broken. But can we measure them simultaneously? The answer is given by Heisenberg's uncertainty principle. Mathematically, it is a corollary of Schwarz's inequality:

$$||\mathbf{v}||^2 \cdot ||\mathbf{w}||^2 \; = < \mathbf{v}, \mathbf{v} >< \mathbf{w}, \mathbf{w} > \geq \; | < \mathbf{v}, \mathbf{w} > |^2.$$

Let $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots$ be eigenvectors for P with eigenvalues $p_1, p_2, \ldots$. The probability that p_j is returned after measuring on $[\mathbf{v}]$ with $\mathbf{v} = \sum_n \alpha_n \mathbf{e}_n$ depends on the multiplicity of the eigenvalue:

$$\text{Prob}(p_j \text{ is returned}) = \sum_{p_k = p_j} |\alpha_k|^2.$$

Hence, we should have the expected value

$$\mathcal{E}(P, \mathbf{v}) = \sum_k p_k |\alpha_k|^2 = \sum_k < \alpha_k \mathbf{e}_k, p_k \alpha_k \mathbf{e}_k > = < \mathbf{v}, P(\mathbf{v}) > .$$

To compute the expected quadratic error we use the shifted observable $P_{\mathbf{v}} = P - \mathcal{E}(P, \mathbf{v})I$:

$$\mathcal{D}(P, \mathbf{v}) = \sqrt{\mathcal{E}({P_{\mathbf{v}}}^2, \mathbf{v})} = \sqrt{< \mathbf{v}, P_{\mathbf{v}}(P_{\mathbf{v}}(\mathbf{v})) >} = \sqrt{< P_{\mathbf{v}}(\mathbf{v}), P_{\mathbf{v}}(\mathbf{v}) >} = ||P_{\mathbf{v}}(\mathbf{v})||$$

where we use the fact that P and $P_{\mathbf{v}}$ are hermitian. Notice that $\mathcal{D}(P, \mathbf{v})$ has a physical meaning of uncertainty of measurement of P. Notice also that the operator $PQ - QP$ is no longer hermitian in general but we can still talk about its expected value. Here goes Heisenberg's principle.

Theorem 6.1.1.

$$\mathcal{D}(P, \mathbf{v}) \cdot \mathcal{D}(Q, \mathbf{v}) \ \geq \ \frac{1}{2} |\mathcal{E}(PQ - QP, \mathbf{v})|$$

Proof. In the right hand side, $\mathcal{E}(PQ - QP, \mathbf{v}) = \mathcal{E}(P_{\mathbf{v}}Q_{\mathbf{v}} - Q_{\mathbf{v}}P_{\mathbf{v}}, \mathbf{v}) = < \mathbf{v}, P_{\mathbf{v}}Q_{\mathbf{v}}(\mathbf{v}) > - < \mathbf{v}, Q_{\mathbf{v}}P_{\mathbf{v}}(\mathbf{v}) > = < P_{\mathbf{v}}(\mathbf{v}), Q_{\mathbf{v}}(\mathbf{v}) > - < Q_{\mathbf{v}}(\mathbf{v}), P_{\mathbf{v}}(\mathbf{v}) >$. Remembering that the form is hermitian,

$$\mathcal{E}(PQ - QP, \mathbf{v}) = < P_{\mathbf{v}}(\mathbf{v}), Q_{\mathbf{v}}(\mathbf{v}) > - \overline{< P_{\mathbf{v}}(\mathbf{v}), Q_{\mathbf{v}}(\mathbf{v}) >} = 2 \cdot \text{Im}(< P_{\mathbf{v}}(\mathbf{v}), Q_{\mathbf{v}}(\mathbf{v}) >) \, ,$$

twice the imaginary part. So the right hand side is estimated by Schwarz's inequality:

$$\text{Im}(< P_{\mathbf{v}}(\mathbf{v}), Q_{\mathbf{v}}(\mathbf{v}) >) \ \leq \ | < P_{\mathbf{v}}(\mathbf{v}), Q_{\mathbf{v}}(\mathbf{v}) > | \ \leq \ ||P_{\mathbf{v}}(\mathbf{v})|| \cdot ||Q_{\mathbf{v}}(\mathbf{v})|| \, .$$

□

Two cases of particular physical interest are *commuting observables*, i.e. $PQ = QP$ and *conjugate observables*, i.e. $PQ - QP = i\hbar I$. Commuting observable can be measured simultaneously with any degree of certainty. Conjugate observables obey Heisenberg's uncertainty:

$$\mathcal{D}(P, \mathbf{v}) \cdot \mathcal{D}(Q, \mathbf{v}) \ \geq \ \frac{\hbar}{2}.$$

6.2 Modules over PID's

Here we will see why two of the main theorems of this course were really different versions of the same theorem.

Definition 6.2.1. A *ring* is a set R with two binary operations $+$ and $\times$, such that

- $(R, +)$ is an abelian group,
- $(R, \times)$ satisfies group axioms G1 (closure), G2 (associativity), and G3 (identity),
- $(r + s) \times t = r \times t + s \times t$ and $r \times (s + t) = r \times s + r \times t$ (distributive law).

If $r \times s = s \times r$ for all $r, s \in R$, then we say R is a *commutative ring*.

Note that R needn't be a group under multiplication, and the additive and multiplicative identities 0_R and 1_R won't necessarily be the same. (In fact the only ring which is a group under multiplication is the zero ring, and this is also the only ring with $0_R = 1_R$.)

Any field is an example of a commutative ring; but there are many more, e.g. $\mathbb{Z}$ is a commutative ring, as is $\mathbb{Z}_n$ for any integer n, and so is $K[X]$ for any field K. An example of a ring that's not commutative is the ring $K^{2,2}$ of 2×2 matrices over a field K.

If R is a commutative ring, an *R-module* is an abelian group M with a binary operation $R \times M \to M$ (written as $r \cdot m$, or just rm) such that

- $rs \cdot m = r \cdot (s \cdot m)$,
- $(r + s) \cdot m = r \cdot m + s \cdot m$,
- $r \cdot (m + n) = r \cdot m + r \cdot n$.

If R is a field, then an R-module is the same thing as a vector space over R; but the definition makes sense for any commutative ring, and we have two very important examples:

- Any abelian group G is a $\mathbb{Z}$-module (in a unique way), with the module structure given earlier, where we defined what ng meant where $n \in \mathbb{Z}$ and $g \in G$ (it was just g added to itself n times).
- If V is a vector space over K, and $T : V \to V$ is a linear operator, then V becomes a $K[X]$-module via the rule
$$f \cdot v = f(T)(v)$$
for $f \in K[X]$ and $v \in V$.

Now, $\mathbb{Z}$ and $K[X]$ are both rather nice rings. Firstly, their multiplication is well-behaved: if R is either of these rings, and $r, s \in R$ satisfy $rs = 0$, then $r = 0$ or $s = 0$. (Notice that this isn't automatic from the axioms: the integers mod 4 are a ring, but $2 \times 2 = 0$ in this ring). Rings with this nice property are called *integral domains*.

Finally, they satisfy the following nice property: any set (not necessarily finite) of elements of R has a *greatest common divisor* – that is, if X is any subset of R, then there is an element $x \in R$ such that

- each element in X is of the form ux for $u \in R$,
- x can be written as $u_1x_1 + \cdots + u_nx_n$ for some elements $x_1, \ldots, x_n$ of X.

It would be reasonable to call rings like this "GCD rings", but they're conventionally known as "principal ideal rings" for a reason you'll see in Algebra II next term. A principal ideal ring

which is also an integral domain is called a "principal ideal domain" or just a "PID".

Theorem 6.2.2. *Let R be a principal ideal domain, and M a finitely-generated R-module. Then there is an integer $s \geq 0$ and elements $d_1, \ldots, d_r$ of R, uniquely determined up to units in R and satisfying $d_i \mid d_{i+1}$ for each i, such that*

$$M \cong R^s \oplus \frac{R}{d_1 R} \oplus \frac{R}{d_2 R} \oplus \cdots \oplus \frac{R}{d_r R}.$$

The proof is almost exactly the same as that of Smith normal form, with integer matrices replaced by matrices over R. Now let's see what this gives us for our two examples:

If $R = \mathbb{Z}$, then a finitely-generated abelian group is certainly a finitely-generated R-module, and hence this gives us the FTFGAG.

If $R = K[X]$ and M is a vector space with a linear operator on it, then this theorem will give us (after a bit of massaging) the JCF theorem! Out of the box, the theorem tells us that V is a direct sum of sub-R-modules, each of which looks like $R/f_i(X)R$ as a $K[X]$-module. But if $K = \mathbb{C}$, then each f_i will look like a product $(X - \lambda_1)^{a_1} \ldots (X - \lambda_t)^{a_t}$, and it's easy to see that

$$\frac{R}{f_i R} \cong \frac{R}{(X-\lambda_1)^{a_1} R} \oplus \cdots \oplus \frac{R}{(X-\lambda_t)^{a_t} R}.$$

So V has a decomposition as a direct sum of T-stable subspaces, with each subspace isomorphic as an R-module to $\frac{R}{(X-\lambda)^k R}$ for some k and λ. But the image of the k vectors $\{(X-\lambda)^{k-1}, (X-\lambda)^{k-2}, \ldots, (X-\lambda), 1\} \subset \frac{R}{(X-\lambda)^k R}$ is then a Jordan chain forming a basis of the subspace! So without too much work, we've deduced both Theorem 5.7.2 and Theorem 2.7.2 from the same piece of abstract algebraic machinery.

6.3 Tensor products

Given two abelian groups A and G, one can form a new abelian group $A \otimes B$, their *tensor product*, dont' confuse with the direct product. We consider the direct product $X = A \times B$ as a set. Let F be the free abelian group with X as a basis:

$$F = \mathbb{Z}^X = \langle\, A \times G \mid \emptyset \,\rangle.$$

Elements of F are formal finite $\mathbb{Z}$-linear combinations $\sum_i n_i(a_i, g_i)$, $n_i \in \mathbb{Z}$, $a_i \in A$, $g_i \in G$. Let F_0 be the subgroup of F generated by the following elements

$$\begin{aligned} &(a+b, g) - (a, g) - (b, g), \quad n(a, g) - (na, g), \\ &(a, g+h) - (a, g) - (a, h), \quad n(a, g) - (a, ng) \end{aligned}$$

for all possible $\alpha \in \mathbb{Z}$, $a, b \in A$, $g, h \in G$. *The tensor product* is the quotient group

$$A \otimes G = F/F_0 = \langle\, A \times G \mid \text{ relations above } \,\rangle.$$

We have to get used to this definition that seems strange at the first glance. First, it is easy to materialize certain elements of $A \otimes G$. *Elementary tensors* are

$$a \otimes g = (a, g) + F_0$$

for various $a \in A$, $g \in G$. However, it is important to realize that not all tensors are elementary. Generators for F_0 become relations on elementary tensors,

$$\begin{aligned} (a+b) \otimes g = a \otimes g + b \otimes g, \quad n(a \otimes g) = (na) \otimes h, \\ a \otimes (g+h) = a \otimes g + a \otimes h, \quad n(a \otimes g) = a \otimes (nh), \end{aligned}$$

so a general element of $A \otimes G$ is a sum $\sum_i a_i \otimes g_i$.

Exercise. *Show that $a \otimes 0 = 0 \otimes g = 0$ for all $a \in A$, $g \in G$.*

If elements b_i span A, while elements h_j span G, then $b_i \otimes h_j$ span $A \otimes G$. Indeed, given $\sum_k a_k \otimes g_k \in A \otimes G$, we can express all $a_k = \sum_i n_{ki} b_i$, $g_k = \sum_j m_{kj} h_j$. Then

$$\sum_k a_k \otimes g_k = \sum_k (\sum_i n_{ki} b_i) \otimes (\sum_j m_{kj} h_j) = \sum_{k,i,j} n_{ki} m_{kj} b_i \otimes g_j$$

In fact, even a more subtle statement holds.

Exercise. Let b_i, be a basis of A, h_j, a basis of G. Then the elementary tensors $b_i \otimes h_j$ constitute a basis of $A \otimes G$.

In particular, a tensor product of two free groups is free. However, for general groups tensor products could behave in quite an unpredictable way. For instance, $\mathbb{Z}_2 \otimes \mathbb{Z}_3 = 0$. Indeed,

$$1_{\mathbb{Z}_2} \otimes 1_{\mathbb{Z}_3} = 3 \cdot 1_{\mathbb{Z}_2} \otimes 1_{\mathbb{Z}_3} = 1_{\mathbb{Z}_2} \otimes 3 \cdot 1_{\mathbb{Z}_3} = 0.$$

To help sorting out zero from non-zero elements in tensor products we need to understand a connection between tensor products and bilinear maps. Let A, G, and H be abelian groups.

Definition 6.3.1. A function $\omega : A \times G \to H$ is a bilinear map if

$$\begin{aligned} \omega(a+b, g) = \omega(a, g) + \omega(b, g), \quad n\omega(a, g) = \omega(na, g) \\ \omega(a, g+h) = \omega(a, g) + \omega(a, h), \quad n\omega(a, g) = \omega(a, ng) \end{aligned}$$

for all possible $n \in \mathbb{Z}$, $a, b \in A$, $g, h \in G$.

Let $Bil(A \times G, H)$ be the set of all bilinear maps from $A \times G$ to H.

Lemma 6.3.2. ***(Universal property of tensor product.)***. *The function*

$$\theta : A \times G \to A \otimes G, \quad \theta(a, g) = a \otimes g$$

is a bilinear map. This bilinear map is universal, i.e. the composition with θ defines a bijection

$$\hom(A \otimes G, H) \to Bil(A \times G, H), \quad \phi \mapsto \phi \circ \theta.$$

Proof. The function θ is a bilinear map: the four propeties of a bilinear map easily follow from the corresponding generators of F_0. For instance, $\theta(a+b,g) = \theta(a,g) + \theta(b,g)$ because $(a+b,g) - (a,g) - (b,g) \in F_0$.

Let Fun denote the set of functions between two sets. Since F is free with basis $A \times G$ we have a bijection[14]

$$\hom(F,H) \to Fun(A \times G, H).$$

It follows from Theorem 5.3.7 that bilinear maps correspond to functions vanishing on F_0, i.e., to linear maps from F/F_0. □

In the following section we will need a criterion for elements of $\mathbb{R} \otimes S^1$ to be nonzero. The circle group S^1 is a group under multiplication, creating certain confusion for tensor products. To avoid this confusion we identify the multiplicative group S^1 with the additive group $\mathbb{R}/2\pi\mathbb{Z}$ via the natural isomorphism $e^{xi} \mapsto x + 2\pi\mathbb{Z}$.

Proposition 6.3.3. *Let $a \otimes (x + 2\pi\mathbb{Z}) \in \mathbb{R} \otimes \mathbb{R}/2\pi\mathbb{Z}$ where $x \in \mathbb{R}$. Then $a \otimes (x + 2\pi\mathbb{Z}) = 0$ if and only if $a = 0$ or $x/\pi \in \mathbb{Q}$.*

Proof. If $a = 0$, then $a \otimes (x + 2\pi\mathbb{Z}) = 0$. If $a \neq 0$ and $x = n\pi/m$ with $m, n \in \mathbb{Z}$, then

$$a \otimes (x+2\pi\mathbb{Z}) = 2m(a/2m) \otimes (x+2\pi\mathbb{Z}) = a/2m \otimes 2m(x+2\pi\mathbb{Z}) = a \otimes (2n\pi + 2\pi\mathbb{Z}) = a \otimes 0 = 0.$$

In the opposite direction, let us consider $a \otimes (x + 2\pi\mathbb{Z})$ with $a \neq 0$ and $x/\pi \notin \mathbb{Q}$. It suffices to construct a bilinear map $\phi : \mathbb{R} \times \mathbb{R}/2\pi\mathbb{Z} \to A$ to some group A such that $\phi(a, x + 2\pi\mathbb{Z}) \neq 0$. By Lemma 6.3.2 this gives a homomorphism $\widetilde{\phi} : \mathbb{R} \otimes \mathbb{R}/2\pi\mathbb{Z} \to A$ with $\widetilde{\phi}(a \otimes (x + 2\pi\mathbb{Z})) = \phi(a, x + 2\pi\mathbb{Z}) \neq 0$. Hence, $a \otimes (x + 2\pi\mathbb{Z}) \neq 0$.

Let us consider $\mathbb{R}$ as a vector space over $\mathbb{Q}$. The subgroup $\pi\mathbb{Q}$ of $\mathbb{R}$ is a vector subspace, hence the quotient group $A = \mathbb{R}/\pi\mathbb{Q}$ is also a vector space over $\mathbb{Q}$. Since $2\pi\mathbb{Z} \subset \pi\mathbb{Q}$, we have a homomorphism

$$\beta : \mathbb{R}/2\pi\mathbb{Z} \to \mathbb{R}/\pi\mathbb{Q}, \quad \beta(z + 2\pi\mathbb{Z}) = z + \pi\mathbb{Q}.$$

Since $x/\pi \notin \mathbb{Q}$, $\beta(x + 2\pi\mathbb{Z}) \neq 0$. Choose a basis e_i of $\mathbb{R}$ over $\mathbb{Q}$ such that $e_1 = a$. Let $e^i : \mathbb{R} \to \mathbb{Q}$ be the linear function[15] computing the i-th coordinate in this basis:

$$e^i(\sum_j x_j e_j) = x_i.$$

The required bilinear map is defined using multiplication by a scalar in $A = \mathbb{R}/\pi\mathbb{Q}$: $\phi(b, z + 2\pi\mathbb{Z}) = e^1(b)\beta(z + 2\pi\mathbb{Z})$. Clearly, $\phi(a, x + 2\pi\mathbb{Z}) = e^1(e_1)\beta(x + 2\pi\mathbb{Z}) = 1 \cdot \beta(x + 2\pi\mathbb{Z}) = x + \pi\mathbb{Q} \neq 0$. □

Exercise. $\mathbb{Z}_n \otimes \mathbb{Z}_m \cong \mathbb{Z}_{\gcd(n,m)}$.

[14] This is known as the universal property of a free abelian group.
[15] commonly known as a covector

6.4 Third Hilbert's problem

All the hard work we have done is going to pay off now. We will understand a solution of the third Hilbert problem. In 1900 Hilbert formulated 23 problems that, in his view, would influence Mathematics of the 20th century. The third problem was the first solved: the same year 1900 by Dehn, which is quite remarkable as the problem was missing from Hilbert's lecture and appeared in print only in 1902, two years after solution.

In his third problem Hilbert asks whether two 3D polytopes of the same volume are *scissor congruent*. Recall that M and N are congruent if there is a motion that moves M to N. M and N are scissor congruent if one can cut M into pieces M_i (cutting along planes) and N into pieces N_i such that individual pieces M_i and N_i are congruent for each i.

Let us consider a "scissor group" generated by all n-dimensional polytopes P

$$\mathbb{P}_n = \langle\, P \mid M - N,\ A - B - C \,\rangle$$

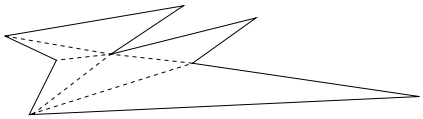
with these relations for each pair M, N of congruent polytopes and each cut $A = B \cup C$ of a polytope by a hyperplane. For a polytope M, let use denote by $[M] \in \mathbb{P}_n$ its class in the scissor group. Clearly, M and N are scissor congruent if and only if $[M] = [N]$. By Theorem 5.3.7, n-dimensional volume is a homomorphism

$$\nu_n : \mathbb{P}_n \to \mathbb{R}, \quad \nu_n([M]) = volume(M).$$

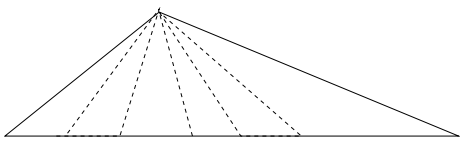
The 3rd Hilbert problem is whether ν_3 is injective.

Theorem 6.4.1. *ν_2 is injective.*

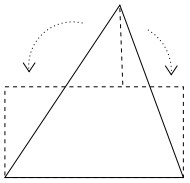
Proof. For a polygon M, there are triangles T_1, $T_2 \ldots T_n$ such that $[M] = [T_1] + [T_2] + \ldots [T_n]$. It follows from triangulation of M illustrated on the next picture.



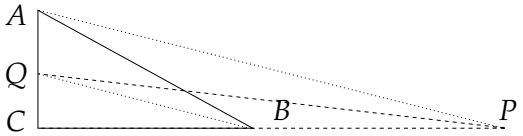
It suffices to show that if two triangles T and T' have the same area, then $[T] = [T']$. Indeed, using it, one can reshape triangles to T'_1, $T'_2 \ldots T'_n$ so that they add up to a triangle T:



Then $[M] = [T_1] + [T_2] + \ldots [T_n] = [T'_1] + [T'_2] + \ldots [T'_n] = [T]$ and we supposedly know that two triangles of the same area are sciccors equivalent. The following picture shows that a triangles with the base b and the height h is equivalent to the rectangle with sides b and $h/2$.



In particular, any polygon is equivalent to a right-angled triangle. The last picture shows who two right-angled triangles of the same area are scissors congurent.



The equal area triangles are CAB and CPQ. This means that $|CA||CB| = |CP||CQ|$. Hence, $|CA|/|CB| = |CP|/|CQ|$ and triangles CAP and CBQ are similar. In particular, AP and BQ are parallel, thus triangles APB and APQ share the same base and height and, consequently, scissors congruent. Finally,

$$[CAB] = [CAP] - [APB] = [CAP] - [APQ] = [CPQ]$$

□

Observe that ν_n is surjective, hence ν_2 is an isomorphism and $\mathbb{P}_2 \cong \mathbb{R}$. However, ν_3 is not injective, disproving the 3rd Hilbert problem.

Theorem 6.4.2. *Let T be a regular tetrahedron, C a cube, both of unit volume. Then $[T] \neq [C] \in \mathbb{P}_3$.*

Proof. Let M be a polytope with the of edges I. For each edge i, let h_i be its length h_i, $\alpha_i \in \mathbb{R}/2\pi\mathbb{Z}$ the angle near this edge. The *Dehn invariant* of M is

$$\delta(M) = \sum_i h_i \otimes \alpha_i \in \mathbb{R} \otimes \mathbb{R}/2\pi\mathbb{Z}.$$

Using Theorem 5.3.7, we conclude that δ is a well-defined homomorphism

$$\delta : \mathbb{P}_3 \to \mathbb{R} \otimes \mathbb{R}/2\pi\mathbb{Z}\,.$$

Indeed, δ defines a homomorphism from the free group F generated by all polytopes. Keeping in mind $\mathbb{P}_3 = F/F_0$, we need to check that δ vanishes on generators of F_0. Clearly, $\delta(M - N) = 0$ if M and N are congruent. It is slightly more subtle to see that $\delta(A - B - C) = 0$ if $A = B \cup C$ is a cut. One can collect the terms in 4 types of groups, with the zero sum in each group.

Survivor: an edge length h with angle α survives completely in B or C. This contributes $h \otimes \alpha - h \otimes \alpha = 0$.

Edge cut: an edge length h with angle α is cut into edges of lengths h_B in B and h_C in C. This contributes $h \otimes \alpha - h_B \otimes \alpha - h_C \otimes \alpha = (h - h_B - h_C) \otimes \alpha = 0 \otimes \alpha = 0$.

Angle cut: an edge length h with angle α has its angle cut into angles α_B in B and α_C in C. This contributes $h \otimes \alpha - h \otimes \alpha_B - h \otimes \alpha_C = h \otimes (\alpha - \alpha_B - \alpha_C) = h \otimes 0 = 0$.

New edge: a new edge of length h is created. If its angle in B is α, then its angle in C is $\pi - \alpha$. This contributes $-h \otimes \alpha - h \otimes (\pi - \alpha) = -h \otimes \pi = 0$, by Proposition 6.3.3.

Finally, using Proposition 6.3.3,

$$\delta([C]) = 12(1 \otimes \frac{\pi}{4}) = 12 \otimes \frac{\pi}{4} = 0, \text{ while } \delta([T]) = 6\frac{\sqrt{2}}{\sqrt[3]{3}} \otimes \arccos\frac{1}{3} \neq 0,$$

by Lemma 6.4.3. Hence, $[C] \neq [T]$. □

Lemma 6.4.3. $\arccos(1/3)/\pi \notin \mathbb{Q}$.

Proof. Let $\arccos(1/3) = q\pi$. We consider a sequence $x_n = \cos(2^n q\pi)$. If q is rational, then this sequence admits only finitely many values. On the other hand, $x_0 = 1/3$ and

$$x_{n+1} = \cos(2 \cdot 2^n q\pi) = 2\cos^2(2^n q\pi) - 1 = 2x_n^2 - 1.$$

Computing several first terms,

$$x_1 = \frac{-7}{9},\ x_2 = \frac{17}{81},\ x_3 = \frac{-5983}{3^8},\ x_4 = \frac{28545857}{3^{16}}, \ldots$$

as can be easily shown the denominators grow indefinitely. Contradiction. □

Now it is natural to ask what exactly the group $\mathbb{P}_3$ is. It was proved later (in 1965) that the joint homomorphism $(\nu_3, \delta) : \mathbb{P}_3 \to \mathbb{R} \times (\mathbb{R} \otimes \mathbb{R}/2\pi\mathbb{Z})$ is injective. It is not surjective: the image can be explicitly described but we won't do it here.