CHRIS GODSIL

MORE LINEAR ALGEBRA

Version: 26/11/2021

Preface

These notes are meant to provide a working knowledge of linear algebra, as might be applied to problems in combinatorics. I assume the reader has had a first course in linear algebra, and is familiar with determinants.

To Do

1. Walk modules. Controllable graphs.
2. Rewrite svd.
3. Expand perturbation theory.
4. Channels
5. Groups: orthogonal, unitary,. . . Matrix exponential. Quadrics.
6. Interlacing, via Courant-Fischer and by rational functions. Need equitable partitions for this. (Put ratfun version in Spectral Decomposition; other goes in or after Eigenvalues and Optimization in Eigenthings chapter)
7. Lie algebras, $sl(2)$ repns.
8. Perron-Frobenius, symbolic dynamics.
9. Lattices

Contents

Part I

Modules

1
Spaces and Subspaces

We review the basic results on vector spaces.

1.1 Vector Spaces

We assume familiarity with the basic terminology of vector spaces—linear combinations, subspaces, linear dependence and independence, span, spanning sets, and bases. We present a proof of the existence of bases (in vector spaces with a finite spanning set).

We define a *circuit* in a vector space V to be a minimal dependent set. Thus if C is a circuit and x is any element of C then $C \setminus x$ is linearly independent. Hence $C \setminus x$ and C have the same span.

1.1.1 Lemma. *If the vector v lies in the span of a set S, then there is a circuit in $S \cup v$ that contains v.*

Proof. Suppose that v is a linear combination of the vectors $x_1, \ldots, x_k$ from S, and that v is not a linear combination of any subset of S with fewer than k elements. Then $x_1, \ldots, x_k$ is linearly independent, for otherwise it contains a circuit and by deleting an element of this circuit, we obtain a set of $k-1$ vectors whose span still contains v. It follows that if for some i, the set

$$\{v, x_1, \ldots, x_k\} \setminus x_i$$

is linearly dependent, then v is a linear combination of at most $k-1$ elements of S. Therefore this set is linearly independent for each i, and so we conclude that $\{v, x_1, \ldots, x_k\}$ is a circuit. □

A basis, we recall, is a linearly independent spanning set. We show they exist if V has a finite spanning set S. If S is linearly independent, there is nothing to prove. Otherwise S contains a circuit C; if $x \in C$ then $C \setminus x$ and C have the same span, and consequently $S \setminus x$ and S have the span. Therefore, by deleting a finite number of elements from S, we obtain a linearly independent set S_1 with the same span as S, and so S_1 is our basis.

Now we show that all finite bases have the same size. To do this we prove the following:

1.1.2 Lemma. *Let V be a vector space. If S is a finite linearly independent subset of V and T is a spanning set, then $|S| \le |T|$.*

Proof. We prove the result by induction on $|S \setminus T|$. Set k equal to $|S \setminus T|$; if $k = 0$ the result is immediate, so suppose $k > 0$. Choose a vector u from $S \setminus T$.

Since T is a spanning set, u is a linear combination of elements of T, and therefore by the lemma above there is a circuit C in $T \cup u$ that contains u. Since S is linearly independent, C is not contained in S and therefore there is an element v in C that does not lie in S. Now v lies in the span of $C \setminus v$, and

$$C \setminus v \subseteq (T \setminus v) \cup u.$$

Therefore v belongs to the span of $(T \setminus v) \cup u$. Since this span contains $T \setminus v$, it contains T.

We conclude that $(T \setminus v) \cup u$ is a spanning set in V that meets S in $k+1$ elements. □

It follows from this that any two finite linearly independent spanning subsets of V have the same size, which we define to be the *dimension* of V. A vector space has finite dimension if and only if it has a finite basis. If V has dimension n then any independent set of size n is a basis, as is spanning set of size n. Each independent set is contained in a basis and, as we already knew, each spanning set contains a basis.

If $\alpha = (v_1, \dots, v_n)$ is an ordered basis for the vector space V and $w \in V$ then there are unique scalars $a_1, \dots, a_n$ such that

$$w = \sum_{i=1}^{n} a_i v_i.$$

The *coordinate vector* $[w]_\alpha$ of w with respect to α is the $n \times 1$ matrix with entries $a_1, \dots, a_n$. The function that maps w to $[w]_\alpha$ is an injective map from V to $\mathbb{F}^n$. We can also show that

$$[w + x]_\alpha = [w]_\alpha + [x]_\alpha, \qquad [cw]_\alpha = c[w]_\alpha.$$

This shows that the coordinate map is an example of a linear mapping.

1.2 Subspaces

The intersection of any two subspaces (indeed, of any collection of subspaces) is a subspace. The union of two subspaces is rarely a subspace—in particular no vector space over an infinite field can be expressed as the union of a finite set of proper subspaces. There is a replacement for union though: the sum $U + V$ of two subspaces U and V. We define this by

$$U + V := \{u + v : u \in U, v \in V\}.$$

We see that $U+V$ is the span of $U\cup V$ and therefore it is a subspace and it is contained in any subspace that contains U and V. Consequently it is the intersection of all subspaces that contained U and V and it follows that the subspaces of a vector space, with the operations of intersection and sum, forms a lattice. If $U\cap V=\{0\}$, we say that $U+V$ is the *direct sum* of U and V.

Here we are concerned with the dimension of $U+V$. For this we need some preliminaries. Suppose U is a subspace of W. We say that a subspace V of W is a *complement* to U if $U\cap V=\{0\}$ and $U+V=W$. We construct examples as follows. Suppose S is a basis of W and (S_1,S_2) is a partition of S into two parts. Let U_i denote the span of S_i. Then U_1+U_2 contains S, and hence it is equal to V. It is also not hard to show that $U_1\cap U_2=\{0\}$. Hence U_2 is a complement to U_1 (and vice versa).

1.2.1 Lemma. *Let W be a vector space with finite dimension. Then any subspace of W has a complement.*

Proof. Let U be a subspace of W and let S be a basis for U. Then there is a basis T for W that contains S, let V be the span of $T\setminus S$. □

1.2.2 Theorem. *If U and V are finite-dimensional subspaces of V, then*

$$\dim(U+V)=\dim(U)+\dim(V)-\dim(U\cap V).$$

Proof. We first establish a special case of the theorem: if U_1 and U_2 are subspaces and $U_1\cap U_2=\{0\}$, then

$$\dim(U_1+U_2)=\dim(U_1)+\dim(U_2).$$

To derive this, we note that if S_i is an independent subset of U_i $(i=1,2)$ and $U_1\cap U_2=\{0\}$ then $S_1\cup S_2$ is linearly independent. Hence the union of a basis of U_1 and a basis of U_2 is a basis for U_1+U_2.

Now we consider the general case. Let V_1 be a complement to $U\cap V$ in V. Then by what we have just proved,

$$\dim(V_1)=\dim(V)-\dim(U\cap V).$$

We show that V_1 is a complement to U in $U+V$. First

$$U+V_1=U+((U\cap V)+V_1)=U+V.$$

Second, $U\cap V_1\subseteq U\cap V$ and $U\cap V_1\subseteq V_1$, so

$$U\cap V_1\subseteq (U\cap V)\cap V_1=\{0\}.$$

Therefore V_1 is a complement to U in $U+V$ and consequently

$$\dim(V_1)=\dim(U+V)-\dim(U).$$

The two expressions for $\dim(V_1)$ imply the result. □

1.3 Linear Mappings

Let V and W be vector spaces over the same field. A function T with domain U and codomain V is a *linear mapping* from U to V if, for all vectors u_1 and u_2 in U,

$$T(u_1 + u_2) = T(u_1) + T(u_2)$$

and if, for all scalars c and all vectors u in U,

$$T(cu) = cT(u).$$

To specify a linear mapping, we must explicitly give its codomain. (This matters most when we consider adjoints.)

A bijective linear mapping is called an *isomorphism*. All this should be familiar. The image and kernel of a linear mapping T are subspaces. The dimension of $\mathrm{im}(T)$ is its *rank* and the dimension of $\ker(T)$ is its *corank*. The following important relation between these parameters is sometimes called the "dimension theorem" for linear mappings.

1.3.1 Theorem. *If T is a linear mapping with domain V then*

$$\mathrm{rk}(T) + \mathrm{cork}(T) = \dim(V).$$

Proof. Choose a basis $v_1, \dots, v_n$ for V such that $v_1, \dots, v_k$ is a basis for $\ker(T)$. Let U be the span of $v_{k+1}, \dots, v_n$. If $u \in U$ and $Tu = 0$, then

$$u \in U \cap \ker(T) = \{0\}.$$

Hence the set $T(v_{k+1}), \dots, T(v_n)$ is linearly independent, and consequently it is a basis for $\mathrm{im}(T)$. □

This is perhaps the most useful formula in linear algebra. An important consequence is that, if T maps V to itself, then it is onto if and only if it is one-to-one.

The coordinate map with respect to a basis is an important example of a linear mapping.

If A is an $m \times n$ matrix over $\mathbb{F}$ then the function that sends $x \in \mathbb{F}^m$ to Ax in $\mathbb{F}^m$ is a linear mapping, often denoted T_A. This gives an even more important class of examples. Note that $\ker(T_A)$ is the null space of A and $\mathrm{im}(T_A)$ is the column space of A, so the dimension theorem yields that

$$\mathrm{rk}(A) + \mathrm{cork}(A) = n.$$

As an application, we rederive the formula for the dimension of the sum of two subspaces. If U and V are vector spaces over the same field, their *external direct sum* is the vector space with vectors

$$\{(u, v) : u \in V,\ v \in V\},$$

where

$$(u_1, v_1) + (u_2, v_2) = (u_1 + u_2, v_1 + v_2)$$

and

$$c(u, v) = (cu, cv).$$

We denote this by $U \oplus V$, and claim that

$$\dim(U \oplus V) = \dim U + \dim V.$$

Now suppose that U and V are subspaces of W. Then we can define a linear map S from $U \oplus V$ to W by

$$S : (u, v) \mapsto u - v.$$

Note that S is a linear map from $U \oplus V$ to the subspace $U + V$ of W. It is easy to see that S is onto, and that its kernel consists of the vectors (x, x), where $x \in U \cap V$. Hence

$$\dim(U + V) = \operatorname{rk}(S) = \dim(U) + \dim(V) - \dim(U \cap V).$$

Define

$$(U, 0) := \{(u, 0) : u \in U\}$$

and define $(0, V)$ similarly. Then $(U, 0)$ and $(0, V)$ are subspaces of $U \oplus V$ having zero intersection and

$$U \oplus V = (U, 0) + (0, V).$$

Thus an external direct sum is a direct sum of subspaces, as in the previous section.

The term "external direct sum" is somewhat confusing. It may help to view this as follows. We have a simple construction of a vector space W from two vector spaces U and V over a field $\mathbb{F}$. The space W is the direct sum, in our original sense, of subspaces isomorphic to U and V.

1.4 Duals and Adjoints

Since we can add linear transformations from V to W and multiply them by scalars, the set $\mathscr{L}(V, W)$ of all linear transformations from V to W forms a vector space. Hence:

1.4.1 Theorem. *If V and W are vector spaces over $\mathbb{F}$, then $\mathscr{L}(V, W)$ is a vector space with dimension* $\dim(V)\dim(W)$.

Proof. We present you a set of linear mappings, and invite you to prove they form a basis.

Let $v_1, \dots, v_n$ be a basis for V and $w_1, \dots, w_m$ be a basis for W. Let $E_{i,j}$ be the element of $\mathscr{L}(V, W)$ given by

$$E_{i,j}(v_r) = \begin{cases} w_j, & \text{if } r = i; \\ 0, & \text{otherwise.} \end{cases}$$

(We use the fact that a linear transformation can be defined by specifying its values on a basis.) This set of $\dim(V)\dim(W)$ operators is the subset we promised. □

Here we will be most interested in the *dual space* $\mathscr{L}(V,\mathbb{F})$, which we denote by V^*. We consider some examples.

Suppose V is the space of all polynomials over $\mathbb{F}$. If $\psi \in V^*$, then ψ is determined by its values on a basis, and hence determined by its values on the powers of x. If we denote $\psi(x^n)$ by ψ_n, then we find that

$$\psi : \sum_{i=0}^{m} p_i x^i \to \sum_{i=0}^{m} p_i \psi_i.$$

Thus each sequence $(\psi_n)_{n\geq 0}$ determines an element of V^*. It follows that we can identify V^* with the space of all formal power series in x.

Each element v of V gives rise to a map from V^* to $\mathbb{F}$, that sends ψ in V^* to $\psi(v)$ in $\mathbb{F}$. This map is linear and injective, and allows us to identify V with a subspace of $(V^*)^*$. The previous example shows that this map need not be an isomorphism in general, but it is an isomorphism when $\dim(V)$ is finite. (This follows from the observation that V, V^* and V^{**} all have the same dimension.)

If $V = \mathbb{F}^n$, then the map that sends an element v to its i-th coordinate is linear, and so belongs to V^*. In this case $V^* \cong V$.

If $V = \mathrm{Mat}_{n\times n}(\mathbb{F})$, then the trace function is an element of V^*.

We cannot resist remarking on one special property of V^*. There is a natural product on it: if $f, g \in V^*$ then fg is defined by $(fg)(u) = f(u)g(u)$.

Let T be a linear map from V to W. If $g \in W^*$, then the composition $g \circ T$ is a linear mapping from V to $\mathbb{F}$; hence it is an element of V^*. Thus we have a mapping that takes an element g of W^* to an element $g \circ T$ in V^*. This map is linear (prove it!), and is called the *adjoint* of T. We denote it by T^*.

(1) Prove that T^* is linear.

(2) Prove that T is one-to-one if and only if T^* is onto, and that T is onto if and only if T^* is one-to-one.

(3) Prove that $T^{**} = T$.

(4) Prove that V is isomorphic to a subspace of V^{**}.

1.5 Bilinear Forms

Suppose Φ is a linear mapping from V to V^*. If $u, v \in V$, then the map

$$(u, v) \mapsto \Phi(u)(v)$$

is linear in each variable. Such a map is called a *bilinear form*. The simplest example arises if we take V to be the space of $n \times 1$ matrices over $\mathbb{F}$. Then we can identify V^* with the space of $1 \times n$ matrices. If $v^T \in V^*$ and $u \in V$, then the value of v^T on u is $v^T u$. So we may take Φ to be the transpose map, and then the bilinear form takes (u, v) to $u^T v$. We generally denote the value of a bilinear form by $\langle u, v\rangle$.

If $u \in V$ and $\Phi(u)(v) = 0$ for all v then $\Phi(u)$ must be the zero vector, and so $u \in \ker\Phi$. If $\Phi(u)(v) = 0$ for all u, then $\operatorname{im}\Phi$ lies in the subspace of V^* formed by the elements f such that $f(v) = 0$. If V is finite dimensional, then V and V^* have the same dimension and $\ker\Phi$ is the zero subspace if and only if $\operatorname{im}\Phi = V^*$. We say that a bilinear form is *non-degenerate* if Φ is invertible; in this case Φ is an isomorphism and we have the following description of V^*:

1.5.1 Lemma. *Let V be a finite-dimensional vector space with a non-degenerate bilinear form. If $f \in V^*$, then there is a vector v in V such that $f(x) = \langle v, x\rangle$.* □

A bilinear form is symmetric if

$$\langle u, v\rangle = \langle v, u\rangle$$

for all u and v. It is *alternating* if

$$\langle u, v\rangle = -\langle v, u\rangle$$

and $\langle u, u\rangle = 0$ for all u. (The first condition implies the second unless we are working over a field of characteristic two.)

We describe one simple construction of bilinear forms. Let A be an $n \times n$ matrix over $\mathbb{F}$. If u and v belong to $\mathbb{F}^n$, define

$$\langle u, v\rangle := u^T A v.$$

It is easy to verify this is bilinear. It is non-degenerate if and only if A is invertible. It is symmetric if and only if $A = A^T$ and alternating if and only if both $A^T = -A$ and all diagonal entries of A are zero.

If S is a subset of V then we define $S^\perp$ to be the set of vectors v such that $\langle v, x\rangle = 0$ for all x in S. (In practice, S will usually be a subspace or a vector.) It is true that if U is a subspace of V, then

$$\dim U^\perp = \dim V - \dim U;$$

but we leave you to prove this. (See the exercises at the end of this section.)

(1) If U is a subspace of V, show that $V = U + U^\perp$ if and only if $U \cap U^\perp = \{0\}$.

(2) Given that $\dim(U^\perp) = \dim(V) - \dim(U)$, prove that $U^{\perp\perp} = U$.

1.6 Counting

We count bases and subspaces in vector spaces over $GF(q)$. Throughout this section we assume that $\mathbb{F}$ has order q. Let $V = \mathbb{F}^n$. Then V contains exactly q^n elements.

We begin by counting the number of subspaces of dimension 1. First we note that two distinct subspaces of dimension 1 have only the zero vector in common, and that a subspace of dimension 1 contains exactly $q-1$ nonzero vectors. It follows that there are exactly $(q^n-1)/(q-1)$ 1-dimensional subspaces of V. This number plays quite a role in what follows, so we define

$$[n] := \frac{q^n-1}{q-1}.$$

(We will write $[n]_q$ if we need to make the order of $\mathbb{F}$ explicit.) Note that $[1] = 1$ and $[2] = q+1$.

We next determine the number of ordered k-tuples $(v_1,\dots,v_k)$ of vectors from V such that $v_1,\dots,v_k$ is linearly independent. Suppose we have such a $(k-1)$-tuple. We can extend it to a k-tuple by choosing vector not in the $(k-1)$-dimensional subspace spanned by the $(k-1)$-tuple. There are $q^n - q^{k-1}$ such factors, and now a simple induction argument yields that the number of ordered k-tuples of linearly independent vectors is

$$(q^n-1)\cdots(q^n-q^{k-1}) = q^{\binom{k}{2}}(q-1)^k[n][n-1]\cdots[n-k+1].$$

Since each k-tuple of linearly independent vectors spans a unique subspace of dimension k, and since each subspace of dimension k gives rise to exactly

$$q^{\binom{k}{2}}(q-1)^k k[k-1]\cdots[1]$$

k-tuples of linearly independent vectors, we find that the number of subspaces of dimension k is

$$q^{\binom{k}{2}}(q-1)^k[n][n-1]\cdots\frac{[n-k+1]}{q^{\binom{k}{2}}}(q-1)^k k[k-1]\cdots[1]$$
$$= \frac{[n][n-1]\cdots[n-k+1]}{[k][k-1]\cdots[1]}. \qquad (1.6.1)$$

This suggests the use of the following notation. We define

$$[n]! := [n][n-1]\cdots[1]$$

and

$$\begin{bmatrix} n \\ k \end{bmatrix} := \frac{[n]!}{[k]![n-k]!}. \qquad (1.6.2)$$

The right side of (1.6.2) is known as the *Gaussian binomial coefficient*. Using it, we have:

1.6.1 Theorem. *The number of subspaces of dimension k in a vector space of dimension n over a field of order q is* ł$\left[\begin{smallmatrix} n \\ k \end{smallmatrix}\right]$. □

We note another consequence. An ordered basis for $\mathbb{F}^n$ is the same thing as an invertible $n \times n$ matrix. Hence:

1.6.2 Lemma. *The number of invertible $n \times n$ matrices over a field of order q is $q^{\binom{n}{2}}(q-1)^n[n]!$.* □

Although it may not be immediately apparent, the Gaussian binomial coefficient is a polynomial in q.

(1) Derive two recurrences for the Gaussian binomial, analagous to

$$\binom{n}{k} = \binom{n-1}{k} = \binom{n-1}{k-1}.$$

(2) Show that the coefficent of t^m in $\left[{k+\ell \atop k}\right]$ is the number of partitions of n with at most k parts, each of size at most ℓ.

(3) Let U be a fixed subspace of $\mathbb{F}^n$ with dimension k. Compute the number of ℓ-dimensional subspaces V of $\mathbb{F}^n$ such that $V \cap U = \{0\}$.

(4) Let U and V be subspaces of $\mathbb{F}^n$ such that $\dim(U) = k$, $\dim(V) = n-k$ and $U \cap V = \{0\}$, where $2k \leq n$. Compute the number of subspaces W with dimension k such that

$$W \cap U = W \cap V = \{0\}.$$

2
Incidence Matrices and Rank

We consider examples of rank arguments in Combinatorics. The basic problem is to derive an upper bound on the size of some set of objects. The solution strategy is to encode the objects as vectors, and then argue that the resulting set of vectors is linearly independent and hence its size is bounded by the dimension of the ambient vector space. It is often natural to present the set of vectors we get as the rows of a matrix, in which case our conceren is the rank of the matrix.

2.1 Fisher's Inequality

A design is a collection of k-subsets (called blocks) of a point set V of size v; it is a t-design if there is a constant λ_t such that any subset T of V with size t lies in exactly λ_t. The number of blocks is denote by b. Lacking warnings to the contrary, we assume $t \geq 2$. (A 1-design is a semi-regular bipartite graph.) If $\mathscr{D}$ is a t-design, and $s \leq t$, then a simple counting argument yields that

$$b\binom{k}{s} = \binom{v}{s}\lambda_s.$$

Hence a t-design is an s-design if $s \leq t$. We note that $\lambda_0 = b$ and that λ_1 (the number of blocks on a point) is denoted by r.

The incidence matrix N of a t-design is the $v \times b$ 01-matrix with rows indexed by points, columns by blocks and with $N_{v,\beta} = 1$ if $v \in \beta$.

2.1.1 Lemma. *If N is the incidence matrix of a 2-design,*

$$NN^T = (r - \lambda_2)I + \lambda_2 J. \qquad \square$$

Since $NJ_{b,v} = rJ_{b,v}$, the lemma yields that

$$N\left(N^T - \frac{\lambda_2}{r}J_{b,v}\right) = (r - \lambda_2)I.$$

Therefore N has a right inverse and so its rows are linearly independent. This implies the number of columns of N is at least v and this yields Fisher's inequality: for any 2-design, $b \geq v$.

One view of what we have done is that we have encoded each point of the design as the characteristic vector of the blocks that contain it, and then verified that these vectors are linearly independent.

2.2 Subset Incidence Matrices

We define $W_{t,k}(v)$ to be the incidence matrix of t-subsets versus k-subsets of a set V of size v. Thus $(W_{t,k}(v))_{\alpha,be} = 1$ if α is a t-subset, β a k-subset of V and $\alpha \subseteq \beta$ (and $(W_{t,k}(v))_{\alpha,be} = 0$ otherwise). When v is clear from the context, we will write $W_{t,k}$.

Although there is no apparent need, we also define a second 01-matrix $\overline{W}_{t,k}(v)$, with rows indexed by t-subsets, columns by k-subsets of V such that

$$\left(\overline{W}_{t,k}(v)\right)_{\alpha,\beta} = 1$$

if and only if $\alpha \cap \beta = \emptyset$. Again we will abbreviate this to $\overline{W}_{t,k}$ when convenient.

To completely define these matrices we should specify an ordering of t-sets and k-sets, but we leave this choice to the reader. As an exercise you might prove that there is a permutation matrix P such that

$$\overline{W}_{t,k} = W_{t,v-k}P.$$

The fundamental result is that if $t \le k$, then $W_{t,k}$ has full rank. In particular if $t \le k$ and $2k \le v$, the rows of $W_{t,k}$ are linearly independent. Note that

$$W_{1,k}W_{1,k}^T = \left(\binom{v-1}{k-1} - \binom{v-2}{k-2}\right) I + \binom{v-2}{k-2} J,$$

Since the right side here is the sum of a positive definite matrix and positive semidefinite matrix, it is positive definite and therefore it is invertible, from which we deduce that the rows of $W_{1,k}$ are linearly independent. As $W_{2,k}$ is the incidence matrix of a 2-design, its rows are linearly independent too. Proving that, in general, $W_{t,k}$ has full rank requires more preparation.

2.2.1 Lemma. *We have:*

(a) $W_{i,t}W_{t,k} = \binom{k-i}{t-i} W_{i,k}$.

(b) $\overline{W}_{i,k}W_{t,k}^T = \binom{v-t-i}{k-t} \overline{W}_{i,t}$.

(c) $W_{i,k}\overline{W}_{t,k}^T = \binom{v-t-i}{k-i} \overline{W}_{i,t}$. □

We will make a lot of use of the first of these three identities.

2.2.2 Lemma. *We have:*

(a) $\overline{W}_{t,k} = \sum_i (-1)^i W_{i,t}^T W_{i,k}$

(b) $W_{t,k} = \sum_i (-1)^i W_{i,t}^T \overline{W}_{i,k}$

Proof. We prove (a) and leave (b) as an exercise. Suppose α is a t-subset of V snd β is a k-subset. Then $\left(\overline{W}_{t,k}\right)_{\alpha,\beta} = 1$ if β lies in the complement of α, and is 0 otherwise.

Now

$$\left(W_{i,t}^T W_{i,k}\right)_{\alpha,\beta} = \binom{|\alpha \cap \beta|}{i}$$

and therefore the (α, β) entry of the sum in (a) is equal to

$$\sum_i (-1)^i \binom{|\alpha \cap \beta|}{i} = \begin{cases} 1, & \text{if } \alpha \cap \beta = \emptyset; \\ 0, & \text{otherwise.} \end{cases}$$

The lemma follows. □

The next lemma provides the main step in proving that $W_{t,k}$ has full rank.

2.2.3 Lemma. *If $t \le k \le v - t$, the matrices $W_{t,k}$ and $\overline{W}_{t,k}$ have the same row space.*

Proof. From Lemma 2.2.1(a) we have

$$W_{i,k} = \binom{k-i}{t-i}^{-1} W_{i,t} W_{t,k}$$

and hence Lemma 2.2.2(a) implies that

$$\overline{W}_{t,k} = \left(\sum_i (-1)^i \binom{k-i}{t-i}^{-1} W_{i,t}^T W_{i,t}\right) W_{t,k}.$$

Therefore each row of $\overline{W}_{t,k}$ is a linear combination of rows of $W_{t,k}$.

It is easy to verify that

$$W_{i,t} \overline{W}_{t,k} = \binom{v-k-i}{t-i} \overline{W}_{i,k}$$

whence Lemma 2.2.2(b) implies that

$$W_{t,k} = \left(\sum_i (-1)^i \binom{v-k-i}{t-i}^{-1} W_{i,t}^T W_{i,t}\right) \overline{W}_{t,k}.$$

and therefire each row of $W_{t,k}$ is a linear combination of rows of $\overline{W}_{t,k}$. □

Now the main result of this section.

2.2.4 Theorem. *If $t \le k \le v - t$, the rows of $W_{t,k}$ are linearly independent.*

Proof. We first consider the case where $v = t + k$. Then $W_{t,v-t}$ and $\overline{W}_{t,v-t}$ are square of the same order. As $\overline{W}_{t,v-t}$ is a permutation matrix, it is invertible and since $\overline{W}_{t,v-t}$ and $W_{t,v-t}$ have the same row space, they have the same rank and therefore $W_{t,v-t}$ is invertible.

Now if $t \le h \le v - t$ then

$$W_{t,h} W_{h,v-t} = \binom{v-2t}{h-t} W_{t,v-t}.$$

Since the matrix on the right of this equation is invertible, it follows that the rows of $W_{t,h}$ are linearly independent. □

The observation that this theorem follows from the fact that $W_{t,v-t}$ is invertible seems to have appeared first in Graver and Jurkat [1].

2.3 *Equitable Partitions*

We develop a version of equitable partitions for $m \times n$ matrices.

Let ρ be a partition of the rows of the matrix[2] N and let σ be a partition of its columns. Let R and S respectively be the characteristic matrices of ρ and σ. If π is partition, let $]pi_i$ denote its i-th cell We say that the pair of partitions (ρ,σ) is *equitable* if for each i and j the submatrix of N with rows indexed by ρ_i and columns indexed by the σ_j has constant row and column sums.

Examples. If N is the incidence matrix of an incidence structure $\mathscr{I}$, the automorphisms of $\mathscr{I}$ are given by permutation matrices P and Q such that $PNQ^T = N$. (Note that Q may not be determined by P, for example, if there are blocks with the same point set.) If ρ is the orbit partition of the automorphism group on points and σ is the orbit partition on blocks, then (ρ,σ) is equitable. To make this more concrete, consider $W_{t,k}$, and let S be a subset of V with size s. We take ρ to be the partition of the t-sets a according to the size of $\alpha \cap S$. Similarly we take σ to be the partition of the k-subsets β according to the value of $\beta \cap S$. Then (ρ,σ) is equitable, with

$$|\rho| = |\sigma| = s+1.$$

You might verify that ρ and σ are orbit partitions, corresponding the subgroup of $\mathrm{Sym}(V)$ that fixes the set S.

The matrix N gives rise to a bipartite graph X with weighted adjacency matrix

$$A = \begin{pmatrix} 0 & N \\ N^T & 0 \end{pmatrix}.$$

If N is an incidence matrix, the cells of ρ and σ form a partition of the vertices of this graph that refines the partition into colour classes, denote it by $\rho \cup \sigma$. Then (ρ, sg) is equitable (as defined above) if and only if $\rho \cup \sigma$ is an equitable partition in our usual sense.

We may refer to the bipartite graph with weighted adjacency matrix

$$\begin{pmatrix} 0 & N^T \\ N & 0 \end{pmatrix}$$

as the *dual* of X. It might be isomorphic to X.[3]

[1]

[2] usually an incidence matrix

[3] and it might not

2.3.1 Theorem. *Let W be a matrix and let ρ and σ respectively be partitions of the rows and columns of W, with characteristic matrices R and S. Then (ρ,σ) is equitable if and only if there are matrices Φ and Ψ such that $WS = R\Phi$ and $R^TW = \Psi S^T$.*

Proof. Let $W_{i,j}$ denote the submatrix of W with rows indexed by the entries of ρ_i and columns indexed by the entries of σ_j.

We have

$$\begin{pmatrix}0 & W\\ W^T & 0\end{pmatrix}\begin{pmatrix}R & 0\\ 0 & S\end{pmatrix} = \begin{pmatrix}0 & WS\\ W^TR & 0\end{pmatrix}$$

and

$$\begin{pmatrix}0 & R\Phi\\ S\Psi & 0\end{pmatrix} = \begin{pmatrix}R & 0\\ 0 & S\end{pmatrix}\begin{pmatrix}0 & \Phi\\ \Psi & 0\end{pmatrix}.$$

Therefore

$$\begin{pmatrix}0 & WS\\ W^TR & 0\end{pmatrix} = \begin{pmatrix}0 & R\Phi\\ S\Psi & 0\end{pmatrix}$$

if and only if

$$\begin{pmatrix}0 & W\\ W^T & 0\end{pmatrix}\begin{pmatrix}R & 0\\ 0 & S\end{pmatrix} = \begin{pmatrix}R & 0\\ 0 & S\end{pmatrix}\begin{pmatrix}0 & \Phi\\ \Psi & 0\end{pmatrix}.$$

We see that

$$A = \begin{pmatrix}0 & W\\ W^T & 0\end{pmatrix}$$

is symmetric, $n \times n$ say, and that

$$Q = \begin{pmatrix}R & 0\\ 0 & S\end{pmatrix}$$

is the characteristic matrix of the partition $\rho \cup \sigma$ of $\{1,\dots,n\}$. Let $\hat{Q}$ be the normalized characteristic matrix of this partition. Then A fixes $\operatorname{col}(Q)$ if and only if A commutes with the projection

$$\hat{Q}(\hat{Q}^T\hat{Q})^{-1}\hat{Q}^T,$$

But this projection is block-diagonal with diagonal blocks of the form $\frac{1}{k}J_k$, and the projection commutes with A if and only if (ρ,σ) is equitable. □

Our equations yield that

$$\Psi S^TS = R^TWS = R^TR\Phi$$

where S^TS and R^TR are diagonal with entries equal to the sizes of the corresponding cells and therefore

$$\Phi = (R^TR)^{-1}R^TWS, \qquad \Psi = R^TWS(S^TS)^{-1}.$$

This shows that Φ and Ψ are determined by R, S and W.

Now $R(R^TR)^{-1}R^T$ represents orthogonal projection onto the space of functions on points that are constant on the cells of ρ, and hence $R\Phi = WS$

if and only if $R(R^TR)^{-1}R^TWS = WS$, that is, if and only if the columns of WS are constant on the cells of ρ. Similarly $R^TW = \Psi S^T$ if and only if the rows of R^TW are constant on the cells of σ.

We also note that if $R^TW = \Psi S^T$ and $z^T\Psi = 0$, then $(Rz)^TW = 0$. Therefore if the rows of W are linearly independent, we have that $z = 0$. This shows that, if the rows of W are linearly independent, so are the rows of Ψ.

2.4 Induced Partitions

Let ρ be a partition of the rows of W with characteristic matrix R. Equality is an equivalence relation on the columns of R^TW we define this partition of the columns to be the partition *induced* by ρ. We denote it by ρ^*.

2.4.1 Theorem. *Let ρ be a partition of the rows of the matrix W. If the rows of W are linearly independent, $|\rho| \le \rho^*$.*

Proof. Let R be the characteristic matrix of ρ. If $z^TR^TW = 0$, then $z^TR^T = 0$. As columns of R are linearly independent, this implies that the rows of R^TW are linearly independent.

Since the rows of R^TW lie in row(W), we have $|\rho| \le rk(W)$.

The number of distinct columns of R^TW is an upper bound on rk(R^TW), and the number of distinct columns is $|\rho^*|$. Our inequality follows. □

This theorem has a very wide range of applications.

Let Γ be a group of automorphisms of an incidence structure of points and blocks, let ρ be the orbit partition on points and let σ be the orbit partition on blocks. We claim[4] that σ is a refinement of the induced partition ρ^*. Hence

2.4.2 Lemma. *Let $\mathscr{I}$ be an incidence structure of points and blocks with incidence matrix W and let Γ be a group of automorphisms of $\mathscr{I}$. If the rows of W are linearly independent, then the number of orbits of Γ on blocks is at least as large as the number of orbits on points.* □

A simple corollary is that if $\mathscr{I}$ is a 2-design and Γ is transitive on its blocks, it is transitive on points.

Let $\mathscr{G}(n,e)$ denote the set of isomorphism classes of graphs on n vertices with e edges.

2.4.3 Lemma. *If $2 \le \binom{n}{2}$, then $|\mathscr{G}(n,e-1)| \le |\mathscr{G}(n,e)|$.*

Proof. Set $N = \binom{n}{2}$. We view a graph on n vertices with e edges as an e-subset of some fixed set of size N (e.g., $E(K_n)$). Let W denote $W_{e,e-1}(N)$. Then Sym(n) acts as a group of automorphisms of the incidence structure of $(e-1)$- versus e-subsets of $\{1,\ldots,N\}$; the orbits on $(e-1)$-subsets are the elements of $\mathscr{G}(n,e-1)$ while the orbits on e-subsets are the elements of $\mathscr{G}(n,e)$. Since the rows of W are linearly independent, the lemma follows. □

[4] and you should prove

We offer a more technical application of Theorem 2.4.1 due to Cameron and Liebler [5].

[5]

2.4.4 Lemma. *Let ρ be partition of the rows of W. If $WW^T = aI + bJ$ for some $a \neq 0$ and b and $|\rho^*| = |\rho|$, then (ρ, ρ^*) is equitable.*

Proof. If $WW^T = aI + bJ$ with $a \neq 0$, then WW^T is invertible and the rows of W are linearly independent. Let R and S be the respective characteristic matrices of ρ and ρ^*. Then there is a matrix Ψ of order $|\rho| \times |\rho *|$ such that

$$R^T W = \Psi S^T$$

Since $|\rho| = |\rho^*|$, we see that Ψ is square. If $z^T\Psi = 0$, then $z^T R^T W = 0$ and it follows that $z = 0$. Therefore Ψ is invertible.

From equation (2.4) we find that

$$WW^T R = WS\Psi^T$$

and consequently

$$(aI + bJ)R\Psi^{-T} = WS.$$

Assume W is $m \times n$ and $|\rho| = r$. We write $J_{(k)}$ to denote the all-ones matrix of order $k \times k$. As $R\mathbf{1} = \mathbf{1}$, find that

$$J_{(m)}R = RJ_{(r)}R^T R$$

and therefore

$$WS = (aR + bJ_{(m)}R)\Psi^{-T} = (aR + bRJ_{(r)}R^T R)\Psi^{-T} = R(aI + J_{(r)}R^T R)\Psi^{-T}.$$

Combined with Equation (2.4), this implies that (ρ, ρ^*) is equitable. □

We present two more applications of this theory as exercises.

We recall that the Gaussian binomial coefficient $\begin{bmatrix} k+\ell \\ k \end{bmatrix}$ is a polynomial in q, and can regarded as the generating function for the number of integer partitions of n with at most k parts, and with each part of size at most ℓ. It is easy to see (given this description) that the coefficients of this polynomial form a symmetric sequence. Your problem is to prove that it is unimodal.

Let Γ be the wreath product $\mathrm{Sym}(\ell) \wr \mathrm{Sym}(k)$, acting on a $k \times \ell$ array of squares by permuting the ℓ squares in a row independently, and by permuting the k rows without changing the orders of the squares in a row. (So $|\Gamma| = (\ell!)^k k!$, and we might also view it as the automorphism group of k vertex-disjoint copies of K_ℓ). Show that the number of orbits of Γ on the sets of n squares from the array is the coefficient of q^n in $\begin{bmatrix} k+\ell \\ k \end{bmatrix}$.

The second application is another proof of Theorem 2.2.4. We work with the incidence structure formed by the t-subsets and k-subsets of a v-set, with incidence matrix $W_{t,k}$. Let τ be a t-subset and let Γ be the subgroup of $\mathrm{Sym}(v)$ formed by the permutations that map τ into itself. (Its order is

$t!(n-t)!$.) Then Γ has $t+1$ orbits on t-subsets, and $t+1$ orbits on k-subsets. Show that matrices Φ andΨ provided by Theorem 2.3.1 are triangular with non-zero diagonals and hence are invertible. Also show that if there is a non-zero vector zE such that $zA^T W_{t,k} = 0$, there is a non-zero vector $\hat{z}$ constant of the cells of the row partition such that $\hat{z}^T W_{t,k} = 0$. Hence derive a contradiction.

2.5 *Null Designs*

Let Ω denote the set of all k-subsets of our v-element set V. We may define a t-design to be a non-negative integer-valued function f on Ω such that, for each t-subset T of V the sum of f on the k-subsets that contain T is equal to λ.[6] Equivalently, viewing f as a vector,

$$W_{t,k} f = \lambda_t \mathbf{1}.$$

If f and g are two t-designs, then $W_{t,k}(f-g) = 0$. We say that h is a *null design* if $W_{t,k} h = 0$. The characteristic function of Ω itself is a t-design[7], whence it follows that each t-design is the sum of a null-design with the complete design.

We give a concrete example, related to Steiner triple systems. Consider the following two sets of triples.

	A			B	
1	2	3	6	2	3
1	4	5	6	4	5
2	4	6	2	4	1
3	5	6	3	5	1

Let V be the set of integers $\{1,\ldots,v\}$, where $v \geq 6$ and let f be the function on triples from V which is 1 on the four triples from A, -1 on the four triples from B and 0 on all other triples. Then f is a null design.

If we are given a Steiner triple system which contains four triples forming a configuration isomorphic to A then we may replace these by the four triples in B. The result is a different Steiner triple system, which may or may not be isomorphic to the original one. As an exercise, construct a Steiner triple system on 13 points that contains a copy of A.

2.6 *Supports*

We derive a lower bound on the size of the support of a null (t,k)-design. The *foundation* of a null (t,k)-design is the union of the blocks in its support. Suppose f is a null (t,k)-design on the point set V and $1 \in V$. By ??? we have

$$\begin{pmatrix} W_{t-1,k-1}(v-1) & 0 \\ W_{t,k-1}(v-1) & W_{t,k}(v-1) \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = 0. \tag{2.6.1}$$

[6] A design is *simple* if f is 01-valued.

[7] the *complete design*

From this it follows that f_1 is a null $(t-1, k-1)$-design on the point set $V \setminus 1$. We call it the *derived design* with respect to the point 1. We call f_2 the *residual design* of f with respect to 1. The next result shows it is a null design too. Either the derived or residual design of f could be identically zero.

2.6.1 Lemma. *Let f be a null (t, k)-design on the point set V and suppose $1 \in V$. Then the derived design of f relative to 1 is a null $(t-1, k-1)$-design and the residual design is a null $(t-1, k)$-design.*

Proof. We have already proved the first claim. As f is also a null $(t-1, k)$-design, we may apply Equation (2.6.1) with $t-1$ in place of t to get

$$W_{t-1,k-1} f_1 + W_{t-1,k} f_2 = 0.$$

As $W_{t-1,k-1} f_1 = 0$, it follows that f_2 is a null $(t-1, k)$-design. □

2.6.2 Theorem. *The support of a null (t, k)-design on V contains at least 2^{t+1} blocks; its foundation contains at least $k+t+1$ points.*

Proof. We consider supports first. Our assertion is easily verified when $t = 0$, so we assume that $t > 0$. Suppose $1 \in V$ and let f_1 and f_2 respectively be the derived and residual designs relative to 1. By the previous lemma, both f_1 and f_2 are null $(t-1)$-designs and, if they are both non-zero, we may assume inductively that each is supported by at least 2^t blocks. Since the supports of f_1 and f_2 are disjoint, the claim follows. If $f_1 = 0$ then ??? yields that $W_{t,k} v - 1 f_2 = 0$. Thus f_2 is a null t-design and we may assume that its support has size at least 2^{t+1}. Similarly if $f_2 = 0$ then $W_{t,k-1} v - 1 f_1 = 0$ and f_1 is a null t-design, therefore its support has size at least 2^{t+1}.

For the foundation, let f be a null t-design on V and choose a point i in V which does not lie in all the blocks in the support of f. Then the residual design g of f relative to i is non-zero and so, by Lemma 2.6.1, it is a null $(t-1, k)$-design on $V \setminus i$. By induction it follows that the foundation of g contains at least $k+t$ points, whence the foundation of f contains at least $k+t+1$. □

One corollary of this theorem is that if $v = k+t$ then the columns of the incidence matrix $W_{t,k}$ are linearly independent, but when $v = k+t$ this matrix is square. Hence we have shown (again) that $W_{t,v-t}$ is invertible over the rationals and it follows (again) that the rows of $W_{t,k}$ are linearly independent over the rationals when $t \le k \le v-k$.

2.7 Null-Designs on Subsets

This section is based on work of Frankl and Pach.

We have defined null (t, k)-designs as functions f on k-subsets of $V = \{1, \dots, n\}$ such that if τ is a t-subset of V, then f sums to zero on the set

$$\{\beta \subseteq V : |\beta| = k,\ \tau \subseteq \beta\}.$$

As

$$W_{t-1,t}W_{t,k} = (k-t+1)W_{t-1,k},$$

the row-space of $W_{t-1,k}$ is a subsopace of the row space of $W_{t,k}$ and if so $|sg \le t|$ a null-design sums to zero on the k-subsets that contain σ. We may extend f to a function on subsets of V by defining it to be zero on sets of size not equal to k.

This leads us to redefine a null-design of strength at least t as a function on subsets of V that sums to zero on the subsets that contain a given set of size at most t. (So a null (t,k)-design is a null design of strength at least t whoe support consists of subsets of size k.)

If f is a function on subsets of V, define a new function f^* by

$$f^*(\tau) := \sum_{\beta\supseteq\tau} f(\beta).$$

Then f is a null-design of strength at least t if f^* is zero on all subsets of size at most t.

Because of the way we defined f^*, we can recover f from f^* by Möbius inversion:[8]

$$f(\beta) = \sum_{\alpha:\alpha\supseteq\beta} (-1)^{|\beta\setminus\alpha|} f^*(\beta).$$

With this in hand, we can create a null-design f_U on a subset U of V from a null design f on V by Möbius inversion on subsets of U applied to the restriction $f^*\restriction U$.

2.7.1 Lemma. *If f is a null-design on V and $U \subseteq V$, then*

$$f_U(\alpha) = \sum_{\gamma\cap U=\alpha} f(\gamma).$$

Proof. We have

$$\begin{aligned} f_U(\alpha) &= \sum_{\alpha\subseteq\beta\subseteq U} (-1)^{|\beta\setminus\alpha|} f^*(\beta) \\ &= \sum_{\alpha\subseteq\beta\subseteq U} (-1)^{|\beta\setminus\alpha|} \sum_{\gamma\supseteq\beta} f(\gamma) \\ &= \sum_{\gamma\subseteq V} f(\gamma) \sum_{\alpha\subseteq\beta\subseteq U\cap\gamma} (-1)^{|\beta\setminus\alpha|} \\ &= \sum_{\gamma\cap U=\alpha} f(\gamma). \qquad\square \end{aligned}$$

One consequence of this lemma is that if $f_U(\alpha) \ne 0$, there is a subset γ of V such that $\gamma\cap U = \alpha$ and $f(\gamma) \ne 0$. The next lemma follows immediately from the definition of f_U.

2.7.2 Lemma. *Let f be a null design of strength at least t and V and let U be a minimal subset of V such that $f^*(U) \ne 0$. If $\alpha \subseteq U$, then*

$$f_U(\alpha) = (-1)^{|U\setminus\alpha|} f^*(U). \qquad\square$$

[8] For Möbius inversion see, for example, https://arxiv.org/abs/1803.06664

2.7.3 Corollary. *The support of a null design of strength at least t has size at least 2^{t+1}.*

Proof. Let U be a minimal subset of V such that $F^*(U) \neq 0$. Then $|U| \geq t+1$. By the lemma, f_U is non-zero on each subset α of U, and by our remark above, for each such α there is a subset γ of V with $\gamma \cap U = \alpha$ and $f(\gamma) \neq 0$. □

2.8 Edge Reconstruction

We apply our bound on the support of null design to the edge-reconstruction problem.

The edge-reconstruction conjecture is the claim that a graph is determined by the collection of its edge-deleted subgraphs. This statement is vague, we offer a precise version of it. We say that the graph Y is an *edge reconstruction* of X if there is a bijection $\beta : E(Y) \to E(X)$ such that, for each edge e in Y we have $Y \setminus e \cong X \setminus \beta(e)$. A graph X is *edge reconstructible* if any edge of X is isomorphic to X.

The graph $K_3 \cup K_1$ is an edge reconstruction of $K_{1,3}$. The edge reconstruction conjecture[9] is the assertion that any graph with at least four edges is edge reconstructible.

In the following discussion, we view a graph on n vertices as a subset of the edges of K_n. (So we are fixing our vertex set.) Thus a graph is just a subset of $E(K_n)$ and hence there are $2^{\binom{n}{2}}$ graphs on n vertices. (Some people might say that we are dealing with "labelled graphs".[10]

For a graph X, define $\mu_X(F)$ to be the number of edge-deleted subgraphs of X isomorphic to F. Let X and Y be graphs on n vertices with e edges, with $e \geq 4$. The edge reconstruction conjecture asserts that if $\mu_X = \mu_Y$, then X and Y are isomorphic.

We translate the problem into linear algebra. Let N denote $\binom{n}{2}$ and define

$$W := W_{e-1,e}(N).$$

If X is a graph on e edges, let ν_X be the function on subsets of $E(K_n)$ defined by

$$\nu_X(F) = \begin{cases} 1, & F \cong X; \\ 0, & \text{otherwise.} \end{cases}$$

We view ν_X as a column vector; we can also view it as the characteristic function of the set of graphs isomorphic to X.

2.8.1 Lemma. *Graphs X and Y on e edges and n vertices have the same collection of edge-deleted subgraphs if and only*

$$W(|\mathrm{Aut}(X)|\nu_X - |\mathrm{Aut}(Y)|\nu_Y) = 0.$$

Proof. The entry of $W\nu_X$ indexed by the graph F on $e-1$ edges is equal to the number of graphs isomorphic to X that contain F (as a subgraph).

[9] Harary, 1964

[10] We will return to this below, we do not find this terminology useful.

Hence the F-entry of $|\mathrm{Aut}(X)|Wv_X$ is equal to

$$|\{\alpha \in \mathrm{Sym}(n) : F \subseteq X^{\alpha}\}| = |\{\alpha \in \mathrm{Sym}(n) : F^{\alpha^{-1}} \subseteq X\}| = \mu_X(F).$$

Therefore $|\mathrm{Aut}(X)|Wv_X = \mu_X$ and this yields the lemma. □

2.8.2 Theorem. *Let X be a graph on n vertices with e edges. If $2e \geq \binom{n}{2} + 1$ or $2^{e-1} > n!$, then X is edge reconstructible.*

Proof. Assume $\mu_X = \mu_Y$ and set

$$z = |\mathrm{Aut}(X)|v_X - |\mathrm{Aut}(Y)|v_Y.$$

We show that if the stated conditions on n and e hold, then $z = 0$.

For the first claim, recall that W has full rank and so if the number of rows is greater then the number of columns, the columns of W are linearly independent. Therefore $z = 0$ and this yields the first claim.

For the second claim, we note that if $z \neq 0$ and $Wz = 0$, then z is a null $(e-1, e)$-design and therefore its support has size at least 2^e. The number of non-zero entries in v_X is $n!/|\mathrm{Aut}(X)|$ and the number in v_Y is $n!/|\mathrm{Aut}(Y)|$, so z has at most $2n!$ non-zero entries. This implies that $2^e \leq 2n!$ and accordingly $2^{e-1} \leq n!$. □

The two parts of the theorem are due respectively to Lovász and Müller. Both parts of the above proof can be extended to reconstruction of k-edge-deleted subgraphs (see Godsil,Krasikov, Roditty) and, even more generally, to k-edge reconstruction of hypergraphs.

Labelled graphs: you are invited to try to express the above arguments using the language of labelled and unlabelled graphs. One point of difficulty is that I do not recall seeing a precise definition of "unlabelled graph" in writing.

3
Primary Decomposition

We use the primary decompostion to decompose vector spaces and linear mappings.

3.1 Modules

Let V be a vector space over $\mathbb{F}$ and let T be an endomorphism of V. A subspace U of V is *T-invariant* if $u \in U$ for all elements u of U. If U is T-invariant, it is invariant under all matrices in the ring $\mathbb{F}[T]$ of polynomials in T. Hence it is a module over this ring; we may also refer to it as a T-module.

(1) If $T = I$ then a T-invariant subspace is just another name for a subspace.

(2) The zero subspace and V itself are T-invariant, for any T.

(3) $\ker(T)$ is T-invariant, because if $u \in \ker(T)$ then $Tu = 0$, and certainly $0 \in \ker(T)$.

(4) The range of T is T-invariant. For if u lies in the range of T then Tu is contained in the range of T.

(5) If U is a subspace of V, the *preimage* of U relative to T is the set

$$\{v \in V : Tv \in U\}.$$

If U is T-invariant, then so is its preimage relative to T. (Since $\ker(T)$ is the preimage of $\{0\}$, this shows that $\ker(T)$ is T-invariant.)

(6) The intersection and sum of T-invariant subspaces are T-invariant.

If U is a T-invariant subspace, then $T \restriction U$ denotes the endomorphism of U that is defined by

$$(T \restriction U)(u) = Tu,$$

for all u in U. We call T the *restriction* of T to U. If U is a 1-dimensional T-invariant subspace and u spans U, then Tu must be a scalar multiple of

u. If u is a non-zero vector and $Tu = \theta u$, we say that u is an *eigenvector* of T with *eigenvalue* θ.

If $u \in W$, then the subspace spanned by vectors

$$T^r v, \qquad r = 0, 1, \ldots$$

is easily seen to be T-invariant. We call it the T-invariant subspace *generated* by v, and observe that is the smallest T-invariant subspace of W that contains v. A T-invariant subspace generated by a single vector u is called a *cyclic subspace* for T. Cyclic subspaces are perhaps the most important class of invariant subspaces.

(1) If $T \in \mathrm{End}(V)$ and T is invertible, show that a T-invariant subspace is T^{-1}-invariant.

3.2 Control Theory

Consider a system of $n+1$ bodies arranged in a line. Assume that if the temperature of the i-th body ($1 \le i \le n$) at time r is $t_i(r)$, then its temperature at time $i+1$ is given by

$$t_i(r+1 = \frac{1}{4}(t_{i-1}(r) + 2t_i(r) + t_{i+1}(r))$$

The temperature of the 0-th body is entirely under our control, we denote its value at time r by $u(r)$. The temperature of the $(n+1)$-st is fixed at zero. If $t(r)$ is the vector in $\mathbb{R}^n$ with i-th entry $t_i(r)$ then t is determined by the equation of the form:

$$t(r+1) = At(r) + u(r)b$$

and the temperature vector $t(0)$ at time zero. In particular, if $n = 5$ then

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0.25 & 0.5 & 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0.5 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0.5 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.5 & 0.25 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

By choosing different values for the terms of the sequence

$$u(0), u(1) \ldots u(m)$$

we can reach a variety of different temperature distributions; are there any we cannot reach?

To study this we assume that $t(0) = 0$. Then

$$\begin{aligned} t(1) &= u(0)b \\ t(2) &= u(1)b + u(0)Ab \\ t(3) &= u(2)b + u(1)Ab + u(0)A^2b \end{aligned}$$

and

$$t(r+1) = \sum_{i=0}^{r} u(r-i)A^i b.$$

If W_r is the matrix

$$W_r = \begin{pmatrix} b & Ab & \cdots & A^{r-1}b \end{pmatrix}$$

then we see that

$$t(r+1) = W_r \begin{pmatrix} u(0) \\ \vdots \\ u(r) \end{pmatrix}.$$

The state $t(r+1)$ is therefore reachable if and only if it lies in the column space of W_r. When $r \geq n$, this column space is precisely the A-cyclic subspace generated by b. (As the vectors $A^r b$ lie in $\mathbb{R}^n$ we have that A^n lies in the column space of W_n and, in general, the rank of W_m equal to the rank of W_n, whenever $m \geq n$.)

In our particular example above, W_6 is an upper triangular matrix with diagonal entries 4^{1-r}, for $r = 1, \dots, 6$. Therefore the cyclic subspace generated by b is $\mathbb{R}^6$, and so all states are reachable after at most six steps. If we change b to

$$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

then the space of reachable states has dimension two—in this case all reachable states have $t_1(r) = t_6(r)$, $t_2(r) = t_5(r)$ and $t_3(r) = t_4(r)$.

(1) Show that, even if $t(0) \neq 0$, it is still true that all states are reachable after at most n steps if and only if the A-cyclic subspace generated by b is all of $\mathbb{R}^n$.

(2) A system given by

$$t(r+1) = At(r), \qquad z(r) = c^T t(r)$$

is *observable* if, given $z(0), z(1), \dots z(m)$ (where $m \geq n$) we can compute $t(i)$ for $i = 0, \dots, m$. (Note: the input for the computation is A, c and the values of z.) Show that this system is observable if and only if the cyclic A^T-subspace generated by c is equal to $\mathbb{R}^n$. Show further that, if the system is observable, we need at most n consecutive values of z to determine all previous states of the system.

3.3 Sums

We consider direct sums of subspaces. Suppose $U_1, \ldots, U_k$ are subspaces of V, and define U_i' to be the sum of the subspaces U_j, where $j \neq i$. We say that V is the *direct sum* of the subspaces U_i if V is the sum of the subspaces U_i and

$$U_i \cap U_i' = \{0\}, \quad (i = 1, \ldots, k). \tag{3.3.1}$$

If this condition holds, we write

$$V = U_1 \oplus \cdots \oplus U_k. \tag{3.3.2}$$

There is a condition equivalent to (3.3.1) that is often easier to work with: V is the direct sum of $U_1, \ldots, U_k$ if and only if for $i = 1, \ldots, n-1$,

$$U_i \cap (U_{i+1} + \cdots + U_k) = \{0\}.$$

We leave you to verify that these two conditions are equivalent.

As an easy consequence of the definition of direct sum, we have

$$\dim(V) = \dim(U_1) + \cdots + \dim(U_k).$$

There is a converse to this: if $U_1, \ldots, U_k$ are subspaces of V whose sum is V and

$$\sum_i \dim(U_i) = \dim(V),$$

then V is the direct sum of the U_i's.

If (3.3.2) holds and $v \in V$, then v can be written in exactly one way as a sum

$$v = u_1 + \cdots + u_k,$$

where $u_i \in U_i$. Define a map $E_i : V \to U_i$ by $E_i(v) = u_i$. Then E_i is linear,

$$E_1 + \cdots + E_k = I,$$

and

$$E_i E_j = \begin{cases} E_i, & \text{if} i = j; \\ 0, & \text{otherwise.} \end{cases}$$

Note that the last condition implies that E_i is *idempotent*, that is, $E_i^2 = E_i$. We call E_i the *projection* onto U_i. Conversely, if $E_1, \ldots, E_k$ is a set of idempotents satisfying these conditions and U_i is the range of E_i, then V is the direct sum of the spaces U_i.

1. If $u_1, \ldots, u_n$ are elements of V and $U_i = \operatorname{span}(u_i)$, show that V is the direct sum of $U_1, \ldots, U_n$ if and only if $u_1, \ldots, u_n$ is a basis for V.

3.4 Invariant Sums

If T is an endomorphism of V, we say a direct sum decomposition of V is T*-invariant* if each summand is. If V is the T-invariant direct sum of $U_1, \dots, U_k$ and $v \in V$ then

$$v = u_1 + \cdots + u_k,$$

where $u_i \in U_i$. Hence

$$T(u) = (T \restriction U_1)(u_1) + \cdots + (T \restriction U_k)(u_k),$$

and so we say that T is the *direct sum* of the operators $T \restriction U_i$. It can be extremely useful to be able to decompose V into a T-invariant direct sum.

We develop a characterization of invariant direct sums in terms of projections. We use the following simple tool.

3.4.1 Lemma. *If E is idempotent, then $x \in \text{im}(E)$ if and only if $x = Ex$.*

Proof. If $x \in \text{im}(E)$ then $x = Ey$ for some y and therefore

$$Ex = E^2 y = Ey = x.$$

If $x = Ex$ then clearly $x \in \text{im}(E)$. □

3.4.2 Theorem. *Suppose $V = V_1 \oplus \cdots \oplus V_k$ and let $E_1, \dots, E_k$ be the set of projections corresponding to the subspaces V_i. Let T be a linear operator on V. Then the direct sum decomposition of V is T-invariant if and only if $TE_i = E_i T$ for each i.*

Proof. We first claim that if E is an idempotent then $\text{im}(E)$ is T-invariant if and only if $(I - E)TE = 0$.

Now $(I - E)TE = 0$ if and only if T maps $\text{im}(E)$ into $\ker(I - E)$. But $(I - E)x = 0$ if and only if $x = Ex$ and so the previous lemma implies that $\ker(I - E) =\in (E)$. This proves our claim.

It follows from this claim that $\text{im}(I - E)$ is T-invariant if and only if $ET(I - E) = 0$.

If $TE = ET$, then both $ET(I - E)$ and $(I - E)TE$ are zero. Conversely, if

$$ET(I - E) = (I - E)TE = 0$$

then

$$0 = ET(I - E) - (I - E)TE = ET - TE$$

and so T and E commute. Hence we have shown that $\text{im}(E)$ and $\text{im}(I - E)$ are T-invariant if and only if $ET = TE$.

Let V_i' be the sum of the subspaces V_j for $j \neq i$. Then $V_i' = \text{im}(I - E_i)$, and so V_i and V_i' are both T-invariant if and only if E_i commutes with T. The theorem follows directly from this. □

Our next result identifies one case where we can express V as a sum of T-invariant subspaces.

3.4.3 Lemma. *Let T be an endomorphism of V. Then $V = \mathrm{im}(T) + \ker(T)$ if and only if* $\mathrm{im}(T) \cap \ker(T) = 0$.

Proof. Suppose $\mathrm{rk}(T) = k$ and $\dim(W) = n$. Then $\dim(\ker(T)) = n - k$ and so $\mathrm{im}(T) + \ker(T) = n$ if and only if $\mathrm{im}(T) \cap \ker(T) = \{0\}$. □

The constraint on T here may also be expressed thus: if $w \in W$ and $T^2 w = 0$ then $Tw = 0$.

As an application of this lemma, suppose that T is idempotent. If $T^2 v = 0$, then $Tv = 0$ and so no non-zero vector Tv lies in $\ker(T)$. Hence V is the direct sum of $\mathrm{im}(T)$ and $\ker(T)$. Note that $T \restriction \ker(T)$ is the zero map.

1. Show that a square matrix of the form
$$P := \begin{pmatrix} 0 & X \\ 0 & I \end{pmatrix}$$
is idempotent. If T is represented by the matrix
$$\begin{pmatrix} A & B \\ C & D \end{pmatrix},$$
show that T fixes $\ker P$ if and only if $C = 0$ and that T fixes $\mathrm{col}(P)$ if and only if
$$XCX - AX + XD - B = 0.$$

3.5 Minimal Polynomials

Let T be an endomorphism of the n-dimensional vector space V. If $v \in V$, then there is a least positive integer r such that $T^r v$ lies in the span of v, $Tv, \ldots, T^{r-1}$. Hence there are scalars $a_1, \ldots, a_r$ such that
$$T^r v + a_1 T^{r-1} v + \cdots + a_0 v = 0.$$

It follows that there is a monic polynomial $\varphi(t)$ such that $\varphi(T)v = 0$. If φ_1 and φ_2 are two polynomials such that $\varphi_i(T)v = 0$, then for all polynomials $a_1(t)$ and $a_2(t)$,
$$(a_1(T)\varphi_1(T) + a_2(T)\varphi_2(T)v = 0,$$
from which it follows that if $\varphi(t)$ is the gcd of $\varphi_1(t)$ and $\varphi_2(t)$, then $\varphi(T)v = 0$. Consequently:

3.5.1 Lemma. *Suppose T is an endomorphism of the finite-dimensional vector space V and $v \in V$. There is a unique monic polynomial of least degree $\psi_v(t)$ such that $\psi_v(T)v = 0$. The degree of ψ_v is equal to the dimension of the subspace generated by v.* □

We call $\psi_v(t)$ the *minimal polynomial* of T relative to v. Since $\dim V = n$, the degree of $\psi_v(t)$ is at most n.

Next we observe that space of endomorphisms of V has dimension n^2, and therefore there is a least integer r, at most n^2, such that I, T,$\dots$,T^r are linearly dependent. It follows that there is a unique monic polynomial ψ of least degree such that $\psi(T) = 0$. It is called the *minimal polynomial* of T. (If L_T denotes the linear operator on $\operatorname{End}(V)$ given by $L_T(M) = TM$, then the minimal polynomial of T is the minimal polynomial of L_T relative to T itself.)

If $v \in V$, then certainly $\psi(T)v = 0$, and it follows that $\psi_v(t)$ must divide $\psi(t)$. Hence $\psi(t)$ is the least common multiple of the polynomials $\psi_v(t)$, as v runs over a basis of V.

3.5.2 Lemma. *Suppose T is an endomorphism of the finite-dimensional vector space V and ψ is the minimal polynomial of T. Then each zero of ψ is an eigenvalue of T.*

Proof. Suppose $\psi_v(\theta) = 0$. Then

$$\psi(t) = (t-\theta)\varphi(t)$$

and therefore

$$(T-\theta I)\varphi(T) = 0.$$

Since φ is a proper factor of ψ, we see that $\varphi(T) \neq 0$. Let w be a non-zero column of $\varphi(T)$. Then $(T-\theta I)w = 0$, and so w is an eigenvector for T with eigenvalue θ. □

If ψ_v is the minimal polynomial of T relative to the vector v and $\psi_v(t) = (t-\theta)\varphi(t)$, then $\varphi(T)v$ is an eigenvector for T with eigenvalue θ. When $\dim V$ is small, this provides an effective way of finding eigenvalues.

For example, suppose $\dim V = 2$, and choose a non-zero vector v. If we are very lucky, v is an eigenvector for T. If not, then T^2v is a linear combination of v and Tv, and ψ_v is quadratic. If θ and τ are the zeros of $\psi_v(t)$, then $(T-\theta I)v$ is an eigenvector for T with eigenvalue τ.

3.6 Primary Decomposition

We use the minimal polynomial of an endomorphism to derive a direct sum decomposition of the space on which it acts. We use the following fact: the greatest common divisor of the polynomials $\varphi_1,\dots,\varphi_k$ is 1 if and only if there are polynomials $a_1,\dots,a_k$ such that

$$a_1\varphi_1 + \cdots + a_k\varphi_k = 1.$$

3.6.1 Theorem. *Let T be an endomorphism of V with minimal polynomial $\psi(t)$. Suppose that $\psi(t) = \prod_{i=1}^r \psi_i(t)$, where the factors ψ_i are pairwise coprime. Set $\varphi_r = \psi/\psi_r$ and let $a_1(t),\dots,a_r(t)$ be polynomials such that*

$\sum_i a_i(t)\varphi_i(t) = 1$. *Then V is the direct sum of T-invariant subspaces U_i, where U_i is the range of the idempotent $a_i(T)\phi_i(T)$. The minimal polynomial of $T\restriction U_i$ is $\psi_i(t)$, and $U_i = \ker\psi_i(T)$.*

Proof. The greatest common divisor of the polynomials φ_i is 1, and so there are polynomials a_i such that

$$a_1\varphi_1 + \cdots + a_r\varphi_r = 1. \tag{3.6.1}$$

Define

$$E_i := a_i(T)\varphi_i(T).$$

Then

$$\sum_{i=1}^r E_i = I.$$

If $i \neq j$ then ψ divides $\varphi_i\varphi_j$, whence

$$E_iE_j = 0.$$

Together the last two equations imply that $E_i^2 = E_i$; thus E_i is an idempotent.

Let U_i denote the range of E_i. If $u \in U_i$ then $E_iu = u$ and so

$$Tu = TE_iu = E_iTu.$$

Therefore Tu lies in the range of E_i, and therefore U_i is T-invariant. Hence V is a direct sum as described.

Next we show that the minimal polynomial of $T\restriction U_1$ is ψ_1. Suppose p is a polynomial such that $p(T)U_i = 0$. Then

$$0 = p(T)E_1 = p(T)a_1(T)\varphi_1(T)$$

which implies that $pa_1\varphi_1$ is divisible by ψ and consequently that ψ_1 divides pa_1. Since ψ_1 divides each of $a_2,\dots,a_r$, it follows from (3.6.1) that a_1 and ψ_1 are coprime. Hence ψ_1 divides p, and we conclude that ψ_1 is the minimal polynomial of $T\restriction U_1$. Setting 1 equal to i, the general result follows. □

Remark: If T has minimal polynomial $\psi(t)$, the ring of all polynomials in T is isomorphic to the quotient ring $\mathbb{F}[t]/(\psi(t))$. The preceding theory is a reflection of the structure theory of this ring.

We use the primary decomposition theorem to prove the following fundamental result.[1]

3.6.2 Theorem. *Let T be an endomorphism of the vector space V over the field $\mathbb{F}$, where $\mathbb{F}$ is algebraically closed. Then there is a diagonalizable endomorphism S and a nilpotent endomorphism N such that S and N are both polynomials in T and $T = S + N$.*

[1] The Jordan-Chevalley decomposition

Proof. Let ψ be the minimal polynomial of T. Since $\mathbb{F}$ is algebraically closed, we may write ψ as

$$\psi(t) = \prod_i (t-\theta_i)^{m_i}.$$

Define ψ_i by

$$\psi_i(t) = \frac{\psi(t)}{(t-\theta_i)^{m_i}}.$$

Let E_i denote $\ker \psi_i(T)$. The polynomials ψ_i are coprime (as a set) and so by the primary decomposition theorem, the E_i are pairwise orthogonal idempotents summing to I. Further each E_i is polynomial in T.

Define S by

$$S = \sum_i \theta_i E_i.$$

If $x \in \ker(T-\theta_i)^{m_i}$ then

$$(T-S)^{m_i} x = (T-\theta_i I)^{m_i} x = 0,$$

from which it follows that $T-S$ is nilpotent. As E_i is a polynomial in T, we see that S is too. □

3.7 The Degree of the Minimal Polynomial

We have seen that the minimal polynomial of an endomorphism T of V is equal to the least common multiple of the minimal polynomials ψ_v, where v runs over the vectors of a basis V. Fortunately something more concrete is true.

3.7.1 Theorem. *If T is an endomorphism of $\mathbb{F}^n$, then there is a vector x such that the minimal polynomial of T relative to x is the minimal polynomial of T.*

Proof. Assume first that the minimal polynomial ψ of T equals $p(t)^m$, where p is irreducible. Then $p(T)^m = 0$ but $p(T)^{m-1} \neq 0$. Choose a vector x such that $p(T)^{m-1}x \neq 0$. If ϕ is monic and $\phi(T)x = 0$ then ϕ must divide ψ. If ϕ divides p^{m-1} then $\phi(T)x \neq 0$. Consequently $\phi = p^m$.

Now suppose that the minimal polynomial of T has the coprime factorization $\psi_1\psi_2$ and that U_1 and U_2 are the summands of the corresponding direct sum decomposition of $\mathbb{F}^n$. Let E_1 and E_2 be the associated idempotents. Suppose that x_i is a vector in U_i such that the minimal polynomial of T relative to x_i is ψ_i. If ϕ is monic and

$$\phi(T)(x_1 + x_2) = 0$$

then

$$0 = E_1\phi(T)(x_1+x_2) = \phi(T)E_1(x_1+x_2) = \phi(T)x_1.$$

This implies that ψ_2 divides ϕ and a similar argument shows that ψ_1 divides it. So ψ divides ϕ and $x_1 + x_2$ is the vector we need.

An easy induction argument based on the last two paragraphs yields that there is always a vector x such that the minimal polynomial of T is the minimal polynomial of T relative to x.

If the field we are working with is infinite, there is an alternative proof. First, the set of relative minimal polynomials ψ_v is finite, since they are all monic divisors of ψ. Suppose $\psi_1,\dots,\psi_r$ is a list of all the possibilities, and let V_i be the set of vectors v such that $\psi_i(T)v = 0$. Then V_i is a subspace of V and the union of the spaces V_i is V itself. But a vector space over an infinite field cannot be the union of a finite number of proper subspaces, hence $V_i = V$ for some i and ψ_i is the minimal polynomial of T.

3.7.2 Corollary. *If* $\dim V = n$ *and* $T \in \operatorname{End}(V)$*, then the degree of the minimal polynomial of* T *is at most* n*.*

Proof. If $T \in \operatorname{End}(V)$ has minimal polynomial $\psi(t)$, then there is a vector v in V such that $\psi(t)$ is the minimal polynomial of T relative to v. Hence, if ψ has degree d, the vectors $v, Av_1, \dots, Av_{d-1}$ are linearly independent. Therefore $\dim(V) \ge d$. □

(1) Let T be an endomorphism of $\mathbb{F}^n$ and let $x_1,\dots,x_n$ be a basis for $\mathbb{F}^n$. If ψ_i denotes the minimal polynomial of T relative to x_i, show that the minimal polynomial of T is the least common multiple of $\psi_1,\dots,\psi_n$.

(2) Prove that a vector space over an infinite field cannot be the union of a finite number of proper subspaces.

3.8 Root Spaces

We consider primary decomposition when the field of scalars is algebraically closed. In this case, if T is a linear operator on V with minimal polynomial $\psi(t)$, then $\psi(t)$ has the coprime factorization

$$\psi(t) = \prod_{i=1}^{k} (t-\theta_i)^{m_i},$$

where $\theta_1,\dots,\theta_k$ are the distinct zeros of ψ. It follows that V is the direct sum of the subspaces

$$\ker(T-\theta_i)^{m_i}.$$

We call these subspaces the *root spaces* of T.

If $v \in V$ and $(T-\theta I)^m v = 0$, then the minimal polynomial of T relative to v divides $(t-\theta)^m$. We say that v is a *root vector* for T if its minimal polynomial relative to T has the form $(t-\theta)^r$, for some integer r. If $(T-\theta I)^r v = 0$ and $v \neq 0$, then θ is an eigenvalue of T.

Since V is the direct sum of the root spaces of T, we have the following fundamental result.

3.8.1 Theorem. *Let V be a finite-dimensional vector space over an algebraically closed field. If T is a linear operator on V, then V has a basis consisting of root vectors of T.* □

The dimension of the root space of an eigenvalue θ of T is called its *algebraic multiplicity*. (The dimension of $\ker(T-\theta I)$ is the *geometric multiplicity* of the eigenvalue.)

3.8.2 Lemma. *Let T be a linear operator on V and let $v_1,\dots,v_n$ be non-zero root vectors. If the respective eigenvalues of these vectors are distinct, then they are linearly independent.*

Proof. Assume $\dim(V)=n$. Suppose that we have scalars $a_1,\dots,a_k$, not all zero, such that

$$\sum_{i=1}^{k} a_i v_i = 0. \tag{3.8.1}$$

We prove by induction on k that $a_1=\cdots=a_k=0$. When $k=1$, this claim is trivial. Assume $k>1$. If we apply $(T-\theta_k I)^n$ to both sides of the above expression we get

$$a_1(T-\theta_k I)^n v_1+\cdots+a_{k-1}(T-\theta_k I)^n v_{k-1}=0. \tag{3.8.2}$$

Since none of $v_1,\dots,v_{k-1}$ lie in the root space belonging to θ_k, none of the $k-1$ terms in this sum is zero. Since each root space is T-invariant, $(T-\theta_i I)^n v_i$ is therefore a non-zero root vector in the root space containing v_i. So by induction, (3.8.2) implies that $a_1=\cdots a_{k-1}=0$. From (3.8.1) it follows that $a_k=0$ too, and we conclude that $v_1,\dots,v_k$ are linearly independent. □

(1) Let T be a linear operator on V with an eigenvalue θ. Show that all root vectors belonging to θ are eigenvectors if and only if

$$\ker(T-\theta I)\cap\operatorname{range}(T-\theta I)=\{0\}.$$

3.9 *Examples of Root Spaces*

We give three examples of root spaces.

Suppose $\dim V=n$ and $e_1,\dots,e_n$ is a basis for V. Thene there is a linear operator T on V such that

$$T(e_i)=\begin{cases} e_{i+1}, & \text{if } i<n;\\ 0, & \text{if } i=n.\end{cases}$$

Thus, if $r<n$ then $T^r(e_1)=e_{i+r}$ and $T^n e_1=0$. In this case V is the root space belonging to the eigenvalue 0.

Let V be $C^\infty(\mathbb{R})$ and let D be differentiation. Then $\ker D^r$ is the space of polynomials of degree less than r. With some work, we can determine $\ker(D-\lambda I)^r$. First we define a linear operator M_λ on V by

$$M_\lambda(f) := e^{\lambda t} f.$$

We claim that $D - \lambda I = M_\lambda D M_{-\lambda}$. (So D and $D - \lambda I$ are similar.)

To verify this we compute

$$\begin{aligned} DM_{-\lambda}(f) &= \frac{d}{dt} e^{-\lambda t} f(t) \\ &= -\lambda e^{-\lambda t} f(t) + e^{-\lambda t} f'(t) \\ &= e^{-\lambda t}(-\lambda f(t) + D(f(t))) \\ &= M_{-\lambda}(D - \lambda I) f. \end{aligned}$$

Since $M_\lambda^{-1} = M_{-\lambda}$, it follows that for all f in V,

$$(M_\lambda D M_{-\lambda})(f) = (D - \lambda I)(f),$$

which is what we claimed.

Now we determine $\ker(D - \lambda I)^r$. We have

$$(D - \lambda I)^r = M_\lambda D^r M_{-\lambda}$$

and therefore $(D - \lambda I)^r(g) = 0$ if and only if

$$M_\lambda D^r M_{-\lambda}(g) = 0.$$

Since M_λ is invertible this holds if and only if

$$D^r M_{-\lambda}(g) = 0.$$

Accordingly $\ker(D - \alpha I)^r$ consists of the functions $g(t)$ such that $e^{-\lambda t} g(t)$ is a polynomial of degree less than r. Therefore $\ker(D - \lambda I)^r$ consists of the functions $e^{\lambda t} p(t)$ where $p(t)$ is a polynomial of degree less than r.

Let $V = \mathbb{C}^{\mathbb{N}}$ and let S be the left shift on V. Define a linear operator M_λ by

$$M_\lambda(a_0, a_1, a_2, \ldots) := (a_0, \lambda a_1, \lambda^2 a_2, \ldots).$$

If $\lambda \neq 0$, then $M_\lambda^{-1} = M_{\lambda^{-1}}$ and

$$S - \lambda I = M_\lambda (S - I) M_\lambda^{-1}.$$

We can show that $\ker(S - I)^r$ consists of the sequences

$$(p(0), p(1), p(2), \ldots)$$

where p is a polynomial of degree less than r, and hence we can show that $\ker(S - \lambda I)^r$ consists of the sequences

$$(p(0), \lambda p(1), \lambda^2 p(2), \ldots)$$

where p is again a polynomial of degree less than r.

The kernel of S^r consists of the sequences $(a_i)_{i \geq 0}$ such that $a_i = 0$ if $i > r$.

3.10 Differential Equations

We begin with two technical results. In this section V is a vector space over $\mathbb{C}$.

3.10.1 Lemma. *Let $T : V \to V$ be linear and suppose that if $\lambda \in \mathbb{C}$, then* $\dim(\ker(T-\lambda I) \le 1$. *If $p(t)$ is a polynomial of degree n, then* $\dim(\ker p(T)) \le n$.

Proof. We prove the result by induction on the degree of $p(t)$. If $n = 1$, there is nothing to prove. Assume $n > 1$.

Suppose θ is a zero of $p(t)$. Then

$$p(t) = (t-\theta)q(t),$$

where q is a polynomial of degree $n-1$. By induction on n, we see that $U = \ker q(T)$ has dimension at most $n-1$.

Now $\ker p(T)$ consists of all vectors v such that $q(T)v$ lies in $\ker(T-\theta I)$. Hence $q(T)$ maps $\ker p(T)$ into $\ker(T-\theta I)$. Let S denote the restriction of $q(T)$ to $\ker p(T)$. Then by the dimension theorem,

$$\dim(\ker p(T)) = \dim\ker(S) + \mathrm{rk}(S) \le \dim(\ker(q(T))) + 1 \le n. \qquad \square$$

The hypotheses of this lemma hold when $V = C^\infty(\mathbb{R})$ and T is differentiation, or when $V = \mathbb{C}^\mathbb{N}$ and T is the left shift.

3.10.2 Theorem. *Let T be a linear operator on V and let $p(t)$ be a polynomial whose zeroes are $\theta_1,\dots,\theta_k$, with respective multiplicities $\nu_1,\dots,\nu_k$. If* $\ker p(T)$ *has finite dimension, it has a basis consisting of root vectors of T; the eigenvalues of these root vectors are the zeros of $p(t)$ and the index of the root vectors with eigenvalue θ_i is at most ν_i.*

Proof. Suppose $K := \ker p(T)$. If $u \in K$, then

$$p(T)Tu = Tp(T)u = 0$$

and therefore K is T-invariant. Hence K is spanned by root vectors of the restriction of T to K, and these are root vectors of T. Suppose z is a root vector of T with eigenvalue θ and index m. Then

$$(T-\theta I)^m z = 0, \quad p(T)z = 0.$$

Therefore the minimal polynomial of T relative to z divides $(t-\theta)^m$ and $p(t)$, and thus it divides $(t-\theta)^\nu$, where ν is the multiplicity of θ as a zero of $p(t)$. $\square$

Let $V = C^\infty(\mathbb{R})$ and let D be differentiation. if

$$p(t) := t^n + a_1 t^{n-1} + \cdots + a_n,$$

then the set of solutions to the differential equation

$$D^n f + a_1 D^{n-1} f + \cdots + a_n f = 0$$

is the kernel of $p(D)$. By Lemma 3.10.1 we see that $\ker p(D)$ has finite dimension and so by Theorem 3.10.2, it follows that $\ker p(D)$ is spanned by root vectors of D whose eigenvalues are zeros of $p(t)$.

We want to find all solutions to

$$D^2 f + 3Df + 2f = 0.$$

The solution set of this equation is $\ker p(D)$, where

$$p(t) := t^2 + 3t + 2 = (t+1)(t+2)$$

From our work above, this subspace has a basis consisting of root vectors for D. Since the zeros of $p(t)$ are simple we only need root vectors of index one, that is, we only need eigenvectors. Hence the functions

$$e^{-t},\ e^{-2t}$$

form a basis for the solution space of this differential equation and therefore every solution can be written as

$$Ae^{-t} + Be^{-2t},$$

for some scalars A and B.

Suppose we want all solutions of

$$D^2 f + 2Df + f = 0$$

Here

$$p(t) = (t+1)^2,$$

whence we see that $\ker p(D)$ is spanned by root vectors with eigenvalue -1 and index at most two. Therefore it is spanned by

$$e^{-t},\ te^{-t};$$

the solutions all have the form

$$(A + Bt)e^{-t}$$

for some scalars A and B.

3.11 Linear Recurrence Equations

The Fibonacci sequence $\varphi = (f_n)_{n\ge 0}$ is defined by the recurrence

$$f_{n+1} = f_n + f_{n-1} \tag{3.11.1}$$

and the initial conditions $f_0 = f_1 = 1$. We want to find an explicict expression for the terms of this sequence.

Let S denote the left shift on $\mathbb{C}^{\mathbb{N}}$. Then we may rewrite (3.11.1) as

$$S^2\varphi = S\varphi + \varphi;$$

this suggests we should study $\ker p(S)$, where

$$p(t) = t^2 - t - 1.$$

The zeros of $p(t)$ are

$$\frac{1 \pm \sqrt{5}}{2};$$

denote these by θ and τ, where $\theta > \tau$. It follows from Theorem 3.10.2 that $\ker p(S)$ is spanned by root vectors for θ and τ with index at most one, hence by eigenvectors.

The eigenvector for S with eigenvalue a is the geometric series

$$(1, a, a^2, \dots)$$

and therefore there are constants a and b such that

$$f_n = a\theta^n + b\tau^n.$$

Setting $n = 0$ and $n = 1$ here gives two equations in the unknowns a and b:

$$1 = a + b, \qquad 1 = a\theta + b\tau.$$

We can rewrite the second equation as

$$1 = \frac{a+b}{2} + \frac{a-b}{2}\sqrt{5};$$

since $a + b = 1$ this implies that

$$a - b = \frac{1}{\sqrt{5}}.$$

Therefore

$$a = \frac{1+\sqrt{5}}{2\sqrt{5}} = \frac{\theta}{\sqrt{5}}$$

and

$$b = \frac{-1+\sqrt{5}}{2\sqrt{5}} = -\frac{\tau}{\sqrt{5}}$$

We conclude that

$$f_n = \frac{1}{\sqrt{5}}(\theta^{n+1} - \tau^{n+1}).$$

3.12 Diagonalizability

A matrix A is *diagonalizable* if there is a diagonal matrix D and an invertible matrix L such that $A = LDL^{-1}$, that is, A is similar to a diagonal matrix. If $A = LDL^{-1}$ then $A^k = LD^kL^{-1}$, and so computing ^{k} can be reduced to the simpler task of computing D^k. More generally, it is often possible to reduce questions about diagonalizable matrices to questions are diagonal matrices (which are often trivial).

3.12.1 Theorem. *For an $n \times n$ matrix A over an algebraically closed field $\mathbb{F}$, the following are equivalent:*

(a) A is diagonalizable.

(b) $\mathbb{F}^n$ has a basis that consists of eigenvectors of A.

(c) The minimal polynomial of A has no repeated factors.

Proof. If two matrices are similar, their minimal polynomials are equal, and so (a) implies (c).

If the minimal polynomial has no repeated factors then there are no root vectors of index greater than one, and thus it follows that $\mathbb{F}^n$ has a basis formed from eigenvectors of A.

Finally, suppose that the columns of L are a basis consisting of eigenvectors. Then each column of AL is a scalar multiple of the corresponding column of L, and therefore there is a diagonal matrix D such that $AL = LD$. Since L must be invertible, (a) follows. □

If $\mathbb{F}$ is not algebraically closed (or close to it, like $\mathbb{R}$), then diagonalizability is not usually a useful concept.

4
Frobenius Normal Form

We derive some properties of matrices from the theory we have established, and then develop the theory of the Frobenius normal form.

4.1 Companion Matrices

Let T be an endomorphism of the finite-dimensional vector space V. One of the best ways to study T is to find T-invariant subspaces of V, and cyclic subspaces are the most accessible of these.

The dimension of the subspace U generated by a vector v is the least integer k such that $T^k v$ lies in the span of the vectors

$$v, Tv, \dots, T^{d-1}v,$$

and this set of vectors forms a natural basis for U. Let v_i denote $T^i v$. Then there are scalars $a_1, \dots, a_k$ such that

$$Tv_{d-1} = -a_d v_0 - \cdots - a_1 v_{d-1}. \tag{4.1.1}$$

If $i < d-1$, then $Tv_i = v_{i+1}$ and therefore the matrix representing the action of T on U, relative to the ordered basis $v_0, \dots, v_{d-1}$, has the form

$$\begin{pmatrix} 0 & 0 & \cdots & 0 & -a_d \\ 1 & 0 & & 0 & -a_{d-1} \\ 0 & 1 & & 0 & -a_{d-2} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_1 \end{pmatrix} \tag{4.1.2}$$

We call this matrix the *companion matrix* of the polynomial

$$p(t) = t^d + a_1 t^{d-1} + \cdots + a_d.$$

(We will also refer to this as the *right* companion matrix of p; we will meet other flavours as we proceed.) Since $v_i = T^i v_0$, from (4.1.1) we find that

$$p(T)v_0 = (T^d + a_1 T^{d-1} + \cdots + a_d I)v_0 = 0.$$

Thus $p(t)$ is the minimal polynomial of T relative to v.

We now consider a matrix view of the previous material. Suppose $v \in \mathbb{F}^n$ and $A \in \mathrm{Mat}_{n\times n}(\mathbb{F})$. Assume that the A-cyclic subspace generated by u has dimension d and let the matrix R be given by

$$R := \begin{pmatrix} u & Au & \cdots & A^{d-1}u \end{pmatrix}.$$

Thus $\mathrm{col}(R)$ is the A-cyclic subspace generated by u. If $\psi(t)$ is the minimal polynomial of A relative to u and C is the companion matrix of ψ, then

$$AR = \begin{pmatrix} Au & A^2u & \cdots & A^d u \end{pmatrix} = RC.$$

There is a third view, which is also quite important. Suppose ψ is a polynomial of degree d over $\mathbb{F}$ and let V_ψ be the vector space of polynomials over $\mathbb{F}$ modulo ψ. This vector space is usually denoted by $\mathbb{F}[z]/(\psi(z))$; its elements are equivalence class of polynomials, where polynomials f and g are equivalent if and only if $f-g$ is divisible by ψ. Each equivalence class contains a unique polynomial of degree less than d, and these are the natural representatives of the equivalence classes.

The powers

$$1, z, \ldots, z^{d-1}$$

provide one basis for V_ψ. Multiplication by z is an endomorphism of V_ψ, and the matrix respresenting multiplication by z relative to this basis is easily seen to be the companion matrix of ψ.

1. Let $p(z)$ be a polynomial of degree k as above and let C_p denote its companion matrix. If f is a polynomial of degree less than k, let $\hat{f}$ be the coordinate vector of f relative to the standard basis $1, x, \ldots, x^{k-1}$. Use the fact that $f(z)z^i$ and $z^i f(z)$ have the same remainder modulo p to prove that
$$f(C_p) = \begin{pmatrix} \hat{f} & C_p\hat{f} & \ldots & (C_p)^{k-1}\hat{f} \end{pmatrix}.$$
Deduce that $f(C_p)\hat{g} = g(C_p)\hat{f}$.

2. If C_p is a companion matrix of order $n \times n$, show that $\mathrm{rk}(C_p - \theta I) \ge n-1$, for any element θ of $\mathbb{F}$. Deduce that the geometric multiplicity of any eigenvalue is at most 1. (This implies that C is diagonalizable if and only if the zeros of p are all simple.)

3. Let U be the subspace spanned by the vectors $T^r u$, where $r \ge 0$. If $Su \in U$, show that there is a polynomial p such that $Su = p(T)u$. Hence deduce that if U is S-invariant and $ST = TS$, then $S\restriction U$ is a polynomial in $T\restriction U$.

4.2 Transposes

We introduce a second basis for V_ψ. If

$$\psi(z) = t^d + a_1 t^{d-1} + \cdots + a_d,$$

define polynomials $\psi_1, \dots, \psi_d$ by

$$\psi_i(z) := t^{d-i} + a_1 t^{d-i-1} + \cdots + a_{d-i}.$$

These polynomials can also be defined by the initial condition $\psi_d(z) = 1$ and the backwards recurrence

$$\psi_{i-1}(z) = z\psi_i(z) + a_{d-i+1}. \tag{4.2.1}$$

As a third alternative, we can view $\psi_i(z)$ as the polynomial part of the rational function $z^{-i}\psi(z)$. Since $\psi_i(z)$ is monic of degree $d-i$, we see that these polynomials form a basis for V_ψ, sometimes called the *control basis*.

Suppose $v \in V$ and T is an endomorphism of V with minimal polynomial $\psi(z)$ relative to v. Then the vectors

$$\psi_1(T)v, \dots, \psi_d(T)v$$

form a basis for the T-cyclic subspace U generated by v. It follows from (4.2.1) that

$$T\psi_i(T)v = \begin{cases} -a_d v, & \text{if } i = 1; \\ \psi_{i-1}(T)v - a_{d+1-i}v, & \text{if } 2 \le i \le d. \end{cases}$$

Let $\mathscr{Y}$ be the matrix with columns

$$\psi_0(T)v, \dots, \psi_{d-1}(T)v.$$

Then you may check that $T\mathscr{Y} = \mathscr{Y}C_\psi^T$, from which we see that C_ψ and C_ψ^T are similar. We can say more about this.

4.2.1 Lemma. *Let T be an endomorphism of V with minimal polynomial*

$$\psi(t) = t^d + a_1 t^{d-1} + \cdots + a_d.$$

If Q is the $d \times d$ matrix

$$Q = \begin{pmatrix} a_{d-1} & a_{d-2} & \dots & a_1 & 1 \\ a_{d-2} & a_{d-3} & & 1 & 0 \\ \vdots & & & & \vdots \\ a_1 & 1 & & & 0 \\ 1 & 0 & \dots & & 0 \end{pmatrix}$$

then $C_\psi Q = QC_\psi^T$.

Proof. If $\mathscr{W}$ is the matrix with columns

$$v, Tv, \dots, T^{d-1}v,$$

then $\mathscr{Y} = \mathscr{W}Q$. Now $T\mathscr{Y} = \mathscr{Y}C_\psi^T$ and therefore

$$T\mathscr{W}Q = \mathscr{W}QC_\psi^T.$$

As

$$T\mathscr{W} = \mathscr{W}C_\psi$$

we deduce that $QC_\psi^T Q^{-1} = C_\psi$. □

Note that Q is symmetric and

$$(C_\psi Q)^T = QC_\psi^T = C_\psi Q,$$

therefore $C_\psi Q$ is symmetric. Consequently $C = (QC)Q^{-1}$ is the product of two symmetric matrices. Using this you may prove that any real square matrix is the prpoduct of two symmetric matrices.

4.3 *Eigenvectors for Companion Matrices*

We give explicit formulas for the left and right eigenvectors of a companion matrix. We use $e_1, \dots, e_d$ to denote the standard basis vectors of $\mathbb{F}^d$, as customary.

4.3.1 Lemma. *Let $\psi(z)$ be a polynomial of degree d and let C be its companion matrix. Then*

$$\begin{pmatrix} 1 & z & \cdots & z^{d-1} \end{pmatrix} C = z \begin{pmatrix} 1 & z & \cdots & z^{d-1} \end{pmatrix} - \psi(z) e_d^T.$$

Proof. Suppose

$$\psi(z) = t^d + a_1 t^{d-1} + \cdots + a_d.$$

If $i < d$, the i-th entry of

$$\begin{pmatrix} 1 & z & \cdots & z^{d-1} \end{pmatrix} C$$

is z^{i+1}; while the d-th entry is

$$-(a_1 + a_2 z + \cdots + a_d z^{d-1} = z^d - \psi(z).$$

The lemma follows at once from this. □

If, in the above lemma, we take z to be a zero θ of ψ, then it follows that

$$\begin{pmatrix} 1 & \theta & \cdots & \theta^{d-1} \end{pmatrix}$$

is a left eigenvector of C with eigenvalue θ.

Our next lemma will provide right eigenvectors. Let $\psi_1, \dots, \psi_d$ denote the control basis for V_ψ.

4.3.2 Lemma. *Let $\psi(z)$ be a polynomial of degree d and let C be its companion matrix. Then*

$$C \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_d \end{pmatrix} = z \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_d \end{pmatrix} - \psi(z) e_1.$$

Proof. This is again routine; we leave it as an exercise. □

These two lemmas provides right and left eigenvectors for C, one for each zero θ of $\psi(z)$. If $\psi(z)$ has d distinct zeros, we obtain d distinct left eigenvectors for C. Since the eigenvalues are distinct, these eigenvectors are linearly independent.

If we are working over $\mathbb{R}$ or $\mathbb{C}$, we can say something useful when $\psi(z)$ has zeros with multiplicity greater than 1. The idea is to differentiate both sides of the identity in Lemma 4.3.2. Define

$$\Psi(z) := \begin{pmatrix}\psi_1\\ \vdots \\ \psi_d\end{pmatrix}$$

and let $\Psi^{(r)}(z)$ denote the r-th derivative of $\Psi(z)$. Then

$$C\Psi^{(r)}(z) = z\Psi^{(r)}(z) + r\Psi^{(r-1)} - \psi^{(r)}(z)e_1.$$

If θ is a zero of ψ with multiplicity m and $r < m$, then $\psi^{(r)}(\theta) = 0$ and therefore

$$(C - \theta I)^r \Psi^{(r)}(\theta) = r!\Psi(\theta).$$

Note the since the polynomials ψ_i form a basis for the polynomials of degree less than d, they cannot all be zero at θ; therefore $\Psi(\theta) \neq 0$. It follows that the vectors $\Psi^{(r)}(\theta)$ are a basis for the root space associated with θ. (Exercise: show that these vectors are linearly independent.)

1. By expanding the expression

$$\begin{pmatrix}1 & w & \cdots & w^{d-1}\end{pmatrix} C \begin{pmatrix}\psi_1(z)\\ \vdots \\ \psi_d(z)\end{pmatrix}$$

in two different ways, derive the identity

$$(w - z)\sum_i w^i \psi_i(z) = \psi(w) - \psi(z).$$

(If we take w and z to be zeros of ψ, this gives the orthogonality relation between the right and left eigenvectors of C.)

2. Let Q be the symmetric matrix from **??**. Show that

$$\Psi(z) = Q\begin{pmatrix}1\\ z\\ \vdots \\ z^{d-1}\end{pmatrix},$$

and hence deduce that $C^T = Q^{-1}CQ$.

4.4 Inverses of Companion Matrices

Suppose $A \in \mathrm{Mat}_{n\times n}(\mathbb{F})$ and that $u \in \mathbb{F}^n$ that generates an A-cyclic subspace of dimension d. Let ψ be the minimal polynomial of A relative to u. If

$$\psi(t) := t^k + a_1 t^{k-1} + \cdots + a_k,$$

then C is invertible if and only if $a_k \neq 0$. (There are a number of ways to see this. Perhaps the easiest is to note that if we move the last column of C to the first position, the resulting matrix C' is lower triangular with $(C')_{1,1} = -a_k$ and all other diagonal entries equal to 1.) If C is invertible, there is a simple expression for C^{-1}. To describe this, we need a new operation on polynomials.

If q is a polynomial with degree k, let $\tilde{q}$ denote the polynomial $t^k q(t^{-1})$. (This is q with its coefficients reversed.) Note that if A is invertible, then $p(A) = 0$ if and only if

$$A^k \tilde{p}(A^{-1}) = 0.$$

It follows that if p is the minimal polynomial of A, then $a_k^{-1}\tilde{p}$ is the minimal polynomial of A^{-1}.

Let R be the matrix given by

$$R = \begin{pmatrix} u & Au & \cdots & A^{d-1}u \end{pmatrix}.$$

Then $AR = RC$ and col(R) is the A-cyclic subspace generated by u. If A is invertible, then A^{-1} is a polynomial in A and therefore col(R) is A^{-1}-invariant. Hence there is a matrix D such that

$$A^{-1}R = RD$$

and $D = C^{-1}$. Now

$$A^{-1}\begin{pmatrix} u & Au & \cdots & A^{d-1}u \end{pmatrix} = \begin{pmatrix} A^{-1}u & u & \cdots & A^{d-2} \end{pmatrix},$$

whence D is a $d \times d$ matrix of the form

$$\begin{pmatrix} \gamma & I_{d-1} \\ c_d & 0 \end{pmatrix}.$$

If we write C in the form

$$C = \begin{pmatrix} 0 & a_d \\ I_{d-1} & \alpha \end{pmatrix},$$

then the equation $DC = I$ implies that

$$I_d = \begin{pmatrix} I_{d-1} & a_d\gamma + \alpha \\ 0 & a_d x_d \end{pmatrix}$$

Consequently we must have

$$c_d = a_d^{-1}, \qquad \gamma = -a_d^{-1}\alpha$$

and therefore

$$D = \begin{pmatrix} -a_d^{-1}\alpha & I_{d-1} \\ a_d^{-1} & 0 \end{pmatrix}.$$

This expression for D makes sense if and only if $a_d \neq 0$, because C can be invertible even when A is not. Hence we have proved the following:

4.4.1 Theorem. *Let p be a polynomial with degree k and let C be the companion matrix of p. Then C is invertible if and only if $p(0) \neq 0$. If $p(0) \neq 0$, then*

$$C^{-1} = TDT,$$

where D is the companion matrix of $a_k^{-1}\tilde{p}$ and T is the matrix whose columns are the standard basis vectors in reverse order. □

By way of example, we have

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

If C is a companion matrix and T is the permutation matrix in the previous theorem, we say that TCT is a *left companion matrix*. Analogously we will call C^T a *bottom* companion matrix. And to round off the list, $TC^T T$ is a *top* companion matrix. All four flavours occur in practice.

4.5 Cycles

Let P be the $n \times n$ matrix such that $Pe_1 = e_n$ and, if $2 \leq i \leq n$ then $Pe_i = e_{i-1}$ and $Pe_n = e_1$. Thus if $n = 5$,

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

We see that $P^n = I$ and P is a companion matrix for the polynomial $t^n - 1$. Further $P^{-1} = P^T$, and therefore P is orthogonal.

Let v_θ be the vector of length n with i-th entry θ^{i-1}. Thus

$$v_\theta = \sum_i \theta^{i-1} e_i$$

and consequently, if $\theta^n = 1$, then

$$Pv_\theta = \sum_i \theta^{i-1} Pe_i = \sum_i \theta^{i-1} e_{i-1} = \theta v_\theta.$$

Therefore the vectors v_θ, as θ runs over the distinct n-th roots of unity, are eigenvectors for P. It is not hard to show that, if n is odd, any real eigenvector of P is a scalar multiple of v_1.

Now let $A = P + P^T$. Then A is symmetric and

$$Av_\theta = (\theta + \theta^{-1})v_\theta.$$

Therefore the vectors v_θ, as θ runs over the distinct n-th roots of unity, are eigenvectors for A. Note that here the eigenvalues

$$\theta + \theta^{-1} = \theta + \overline{\theta}$$

are real, even though the eigenvectors themselves are complex (unless θ is real).

The eigenvalues of P are roots of unity. Suppose Q is orthogonal and v is an eigenvector for it with eigenvalue θ. Then $Qv = \theta v$, but

$$\|v\| = \|Qv\| = \|\theta v\| = |\theta|\,\|v\|.$$

It follows that all eigenvalues of an orthogonal matrix lie on the unit circle in the complex plane.

4.6 Circulants and Cyclic Codes

Let P_n be the companion matrix for the polynomial $t^n - 1$. Thus if $n = 5$ then

$$P_5 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

We see that $P_n e_i = e_{i+1}$ if $i < n$ and $P_n e_n = e_1$. A *circulant matrix* is a matrix which is a polynomial in P_n. This is equivalent to stating that a matrix is a circulant if it is square and each row is a cyclic right shift of the row above it. If the first column of the circulant A is

$$\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

then

$$A = \sum_{i=1}^{n} a_i P^{i-1}.$$

It follows that there is an isomorphism between the vector space of $n \times n$ circulant matrices and the space of polynomials with degree less than n. But this is misleading. Suppose a and b are polynomials with degree less than n, and associated circulants A and B respectively. Then the product

AB is a circulant, but the polynomial belonging to it cannot be ab unless the degree of this product is less than n. In fact the polynomial is the remainder of the product $a(t)b(t)$ on division by $t^n - 1$. Thus the space of $n \times n$ circulants is isomorphic to the quotient ring $\mathbb{F}[x]/(x^n - 1)$. This isomorphism is an algebra isomorphism. If $\deg(a) < n$, we use C_g to denote the circulant associated with g.

The row space of an $n \times n$ circulant over $\mathbb{F}$ is a *cyclic code* of length n. Suppose f is a polynomial and that g is its greatest common divisor with $t^n - 1$. Then there are polynomials a and b such that

$$a(t)f(t) + b(t)(t^n - 1) = g(t).$$

Hence

$$C_g = C_{af} = C_a C_f$$

and therefore $\operatorname{row}(C_g) \subseteq \operatorname{row}(C_f)$. On the other hand, $f = f_1 g$ and so

$$C_f = C_{f_1} C_g,$$

which implies that $\operatorname{row}(C_f) \subseteq \operatorname{row}(C_g)$ and hence that $\operatorname{row}(C_f) = \operatorname{row}(C_g)$. This proves that a cyclic code of length n over $\mathbb{F}$ is equal to C_g, for some divisor g of $t^n - 1$.

One of the most important parameters of a code is its dimension. Thus we would like to determine $\operatorname{rk}(C_g)$. If g has degree d, then the submatrix formed by the intersection of the first $n - d$ columns and last n-d rows of C_g is the identity matrix I_{n-d}. Therefore

$$\operatorname{rk}(C_g) \geq n - d.$$

Suppose $a(t)$ is a polynomial of degree less than n, and let $[a]$ denote its coordinate vector with respect to the ordered basis $1, t, \dots, t^{n-1}$. If $C_g[a] =$ 0, then

$$0 = C_g[a] = P^r C_g[a] = C_g P^r [a]$$

for all r and consequently

$$C_g C_a = 0.$$

Equivalently, $C_g C_a = 0$ if and only if $C_g C_a e_1 = 0$. Now $C_g C_a = 0$ if and only if $t^n - 1$ divides $g(t)a(t)$, and accordingly the null space of C_g consists of the vectors $[a]$ such that $(t^n - 1)/g(t)$ divides $a(t)$. If we set $h(t)$ equal to $(t^n - 1)/g(t)$, then the null space of C_g is the column space of C_h. The dimensions of the row and column spaces of C_h are equal, and therefore

$$\operatorname{rk}(C_h) \geq n - (n - d) = d.$$

So by the rank+nullity theorem,

$$\operatorname{rk}(C_g) + \operatorname{rk}(C_h) = n,$$

which forces us to conclude that $\operatorname{rk}(C_g) = n - d$.

If $C_{g^2}[a] = 0$ then $C_{g^2}C_a = 0$ and so $t^n - 1$ divides $g(t)^2 a(t)$. If $t^n - 1$ has no repeated factors, then $t^n - 1$ divides $g(t)^2 a(t)$ if and only if it divides $g(t)a(t)$. In this case it follows that C_g is diagonalizable. If $x^n - 1 = p(t)^2 q(t)$, then $C_{pq}^2 = 0$ and C_{pq} is nilpotent, and not diagonalizable.

However $x^n - 1$ has a repeated factor if and only if its gcd with its derivative nx^{n-1} is not constant, in other words, if and only if n is not divisible by the characteristic of $\mathbb{F}$. In particular, if the characteristic of $\mathbb{F}$ does not divide n, then $\mathbb{F}^n$ is the direct sum of $\ker(C_g)$ and $\operatorname{col}(C_g)$.

Let $\mathbb{E}$ be an extension field of $\mathbb{F}$ in which $t^n - 1$ splits into linear factors. If n and the characteristic of $\mathbb{F}$ are coprime, these factors are all distinct. It follows that each divisor g of $t^n - 1$ is determined by the set of n-th roots of 1 on which it vanishes. Let v_θ be the vector of length n with i-th entry equal to θ^{i-1}. Then if $\theta^n = 1$,

$$C_g v_\theta = g(\theta^{-1})$$

and so $\operatorname{row}(C_g)$ consists of the vectors x^T such that

$$x^T v_\theta = 0$$

whenever θ^{-1} is a zero of g.

4.7 Frobenius Normal Form

A square matrix C is in *Frobenius normal form* if

(a) It is block-diagonal, with diagonal blocks $C_1, \ldots, C_m$.

(b) Each diagonal block is the companion matrix of a polynomial $\psi_i(t)$.

(c) For $i = 1, \ldots, m-1$, the polynomial ψ_{i+1} divides ψ_i.

Thus the Frobenius normal form can be specified by giving the sequence of polynomials ψ_i.

We want to prove that two matrices over a field are similar if and only if they have the same Frobenius normal form. he next lemma is the key.

4.7.1 Lemma. *Let A be an $n \times n$ matrix over $\mathbb{F}$. If U is a non-zero cyclic A-module, then:*

(a) There is a cyclic A-module that contains U nad has a complement.

(b) If $\dim(U)$ equals the degree of the minimal polynomial of A, then U has an A-invariant complement.

Proof. Let u be a non-zero vector and suppose that the A-invariant subspace it generates has dimension k. Let U be the $n \times k$ matrix with the vectors

$$u, Au, \ldots, A^{k-1}u$$

as its columns. Then the columns of U are linearly independent and therefore there is a $n \times k$ matrix V such that $V^T U = I$. Let w denote the last column of V. (Now that we have w, we will ignore V.)

We have

$$w^T A^r A^s u = w^T A^{r+s} u = \begin{cases} 1, & \text{if } r = k-1-s; \\ 0, & \text{if} r < k-1-s. \end{cases}$$

If W is the matrix with columns

$$(A^{k-1})^T w, (A^{k-2})^T w, \dots, A^T w, w$$

then $W^T U$ is a lower triangular matrix with diagonal entries equal to 1. Therefore it is invertible, and therefore $\text{rk}(W) = k$. Let ℓ be the dimension of the A^T-invariant subspace generated by w. Since $\text{rk}(W) = k$, we see that $k \le \ell$.

If $k \ne \ell$, then repeating the above argument with A^T in place of A and w in place of U, we obtain a cyclic subspace for A with dimension at least ℓ. By repeating both these steps a finite number of times, we bring ourselves to the case where $k = \ell$.

Therefore we may assume that $\text{col}(W)$ is A^T-invariant, and so there is a matrix L such that $A^T W = W L^T$. If $W^T x = 0$ then

$$0 = L W^T x = W^T A x;$$

accordingly the null-space K of W^T is A-invariant. Since $W^T U$ is invertible, no non-zero element of $\text{col}(U)$ lies in K. Since $\text{rk}(W) = k$, we see that $\dim K = n - k$ and therefore K is an A-invariant complement to $\text{col}(U)$.

To obtain the last statement of the proof, note that A and A^T have the same minimal polynomial. So if k equals the degree of this polynomial, then $\text{rk}(U) = \text{rk}(W)$. □

It follows readily from this lemma that every square matrix is similar to a block diagonal matrix, where each block is a companion matrix. We can also use it as follows to verify the existence of the Frobenius normal form.

4.7.2 Theorem. *Every square matrix is similar to a matrix in Frobenius normal form.*

Proof. Let A be an $n \times n$ matrix with minimal polynomial $\psi(t)$ of degree k. By **??**, there is a vector u such that ψ is the minimal polynomial of A with respect to u, and therefore u generates a cyclic subspace V of dimension k. By the previous lemma, it follows that this subspace has an A-invariant complement K.

Choose a basis for $\mathbb{F}^n$ consisting of the columns of V followed by a basis for K. Relative to this basis, A is represented by a block-diagonal matrix

$$\begin{pmatrix} L & 0 \\ 0 & B \end{pmatrix},$$

where L is the companion matrix of the minimal polynomial of A. The minimal polynomial of B divides the minimal polynomial of A. By induction on n we see that B is similar to a matrix in Frobenius normal form; stacking L on top of this produces a matrix in Frobenius normal form that is similar to A. □

4.7.3 Lemma. *Assume the matrices N_1 and N_2 are in Frobenius normal form. If N_1 and N_2 are similar, they are equal.*

Proof. Suppose that

$$N_1 := \begin{pmatrix} L_1 & 0 \\ 0 & D_1 \end{pmatrix}, \qquad N_2 := \begin{pmatrix} L_2 & 0 \\ 0 & D_2 \end{pmatrix}$$

are both in Frobenius normal form and that L_1 and L_2 are companion matrices. Then $p(N) = 0$ if and only if $p(L_1) = 0$ and $p(D_1) = 0$ and hence the minimal polynomial of N_1 is the minimal polynomial of L_1. Since N_1 and N_2 are similar, they have the same minimal polynomial, and this is also the minimal polynomial of L_2. Thus L_1 and L_2 have the same minimal polynomial, and as they are companion matrices this implies that they are equal.

Assume now that L_1 is in Frobenius normal form (not necessarily a companion matrix) and that $L_1 = L_2$. Let ψ_1 be the minimal polynomial of D_1. Then $\psi_1(N_1)$ and $\psi_1(N_2)$ are similar and thus

$$\begin{pmatrix} \psi_1(L) & 0 \\ 0 & 0 \end{pmatrix} \sim \begin{pmatrix} \psi_1(L) & 0 \\ 0 & \psi_1(D_2) \end{pmatrix}.$$

This implies that $\psi_1(D_2) = 0$ (prove it!) and we conclude that D_1 and D_2 have the same minimal polynomial. An easy induction argument now yields that $D_1 = D_2$. □

We can use Lemma 4.7.1 to compute the minimal polynomial of a square matrix. First compute a block-diagonal matrix similar to A, with companion matrices as its blocks. The least common multiple of the polynomials associated to these companion matrices is the minimal polynomial of A.

4.8 Applications

We use $\mathscr{C}(M)$ to denote the *commutant* of M, that is, the set of matrices that commute with M. This a subspace that contains all polynomials in M.

4.8.1 Theorem. *Let A and B be square matrices. If $\mathscr{C}(A) \subseteq \mathscr{C}(B)$, then B is a polynomial in A.*

Proof. Assume A is $n \times n$. We can decompose $\mathbb{F}^n$ as the direct sum of A-invariant subspaces $V_1, \ldots, V_k$. For each subspace there is a cyclic vector v_i

such that the powers $A^r v_i$ span V_i. If ψ_i is the minimal polynomial of $A\restriction U_i$, then its degree equals $\dim V_i$, and ψ_{i+1} divides ψ_i.

Let P_i denote the projection on V_i. From **??**, the projections P_i commute with A. Hence they commute with B and, again by **??**, it follows that the subspaces V_i are B-invariant. Therefore $Bv_i \in V_i$ and therefore there is a polynomial g_i such that $Bv_i = g_i(A)v_i$. As $AB = BA$, we have

$$BA^r v_i = A^r Bv_i = A^r g_i(A)v_i = g_i(A)A^r v_i,$$

and therefore $Bv = g_i(A)v$ for all v in V_i.

To complete the proof, we show that $g_i(A)v_i = g_1(A)v_i$. This implies that $B = g_1(A)$.

Let $q_i := \psi_1/\psi_i$. Consider the map that sends $f(A)v_i$ to $q_i(A)f(A)v_1$. If $f(A)v_i = 0$, then ψ_i divides f and so $\psi_1 = \psi_i q_i$ divides $q_i(A)f(A)v_1$. It follows that this is a well-defined linear map from V_i to V_1. We extend it a linear map X_i from V to V by defining $X_i(v) = 0$ if $v \in V_j$ and $j \neq i$; if $v = f(A)v_i$ then $X_i v = q_i(A)f(A)v_1$.

If $i \neq j$ and $v \in V_j$, then $AX_i v = X_i Av = 0$. Further

$$AX_i f(A)v_i = Aq_i(A)f(A)v_1$$

and

$$X_i Af(A)v_i = q_i(A)Af(A)v_1.$$

Therefore X_i commutes with A, and therefore it commutes with B. Now

$$X_i Bv_i = X_i g_i(A)v_1 = g_i(A)q_i(A)v_1$$

and

$$BX_i v_i = Bq_i(A)v_1 = q_i(A)Bv_1 = q_i(A)g_1(A)v_1.$$

Since X_i and B commute, this implies that

$$(g_i(A) - g_1(A))q_i(A)v_1 = 0,$$

whence $(g_i - g_1)q_1$ is divisible by $p_1 = q_i p_i$, and so p_i divides $g_i - g_1$. Consequently

$$g_i(A)v_i = g_1(A)v_i,$$

for all i. □

The above proof follows Prasalov.

4.9 Nilpotent Matrices

A linear mapping or a matrix is *nilpotent* if some power of it is zero. The canonical example is

$$N_2 := \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

whose square is zero. If T is nilpotent then its minimal polynomial is t^k for some k, sometimes called the *index of nilpotency* of T (but not very often, if we can help it). A nilpotent matrix of index 1 is the zero matrix. We note that N_2 is the companion matrix of t^2. More generally the companion matrix of t^k is a nilpotent matrix with index k, which we will denote by N_k. Note that $N_k e_1 = 0$ and $N_k e_{i+1} = e_i$ when $i \geq 1$.

Nilpotent matrices are interesting and useful, but also a source of difficulties. Since $N_k e_1 = 0$, we see that e_1 is an eigenvector of N_k with eigenvalue 0. Since the minimal polynomial of N_k is t^k, we see that 0 is the only eigenvalue of N_k. Further, since $\operatorname{rk}(N_k) = k-1$, the eigenspace associated with 0 has dimension 1, and therefore equals the span of e_1. Consequently eigenvalues and eigenvectors provide very little information about nilpotent matrices.

We have the following structure theorem.

4.9.1 Theorem. *If M is a nilpotent matrix, then it is similar to a block diagonal matrix, where each diagonal block is equal to N_k for some k.*

Proof. The required block diagonal matrix is the Frobenius normal form of M. □

One corollary of this is that the number of similarity classes of $n \times n$ nilpotent matrices over a field equals the number of vectors of non-negative integers

$$(k_1, \ldots, k_n)$$

such that $k_1 \geq k_2 \geq \cdots \geq k_1$ and $\sum_i k_i = n$.

4.9.2 Lemma. *Let A be an $n \times n$ matrix over an algebraically closed field with minimal polynomial $\psi(t)$. Then A is similar to a block diagonal matrix with diagonal blocks of the form $\theta I + N_\theta$, where θ runs over the zeros of ψ, and N_θ is nilpotent with index equal to the multiplicity of θ as a zero of $\psi(t)$.*

Proof. By the primary decomposition theorem Theorem 3.6.1, we know that A is similar to a diagonal matrix with diagonal blocks A_θ indexed by the zeros of ψ, such that the minimal polynomial of A_θ is $(t-\theta)^{m_\theta}$, where m_θ is the multiplicity of θ as a zero of $\psi(t)$. Hence $A - \theta I$ is nilpotent, with index m_θ. Thus we may write

$$A_\theta = \theta I + N_\theta,$$

where N_θ is nilpotent, of index m_θ. □

The corank of $(A-\theta I)^{m_\theta}$ is known as the *algebraic multiplicity* of the eigenvalue θ. This distinguishes it from the *geometric multiplicity*, which is the corank of $A - \theta I$.

We present one application. We wish to determine when a matrix A has a square root, that is, when there is a matrix X such that $X^2 = A$. If $A = LBL^{-1}$ and B has a square root Y, then

$$(LYL^{-1})^2 = LY^2L^{-1} = LBL^{-1} = A.$$

This allows us to use the primary decomposition theory; more precisely, we assume that A is block diagonal with blocks of the form

$$A_\theta = \theta I + N_\theta.$$

It follows that A has a square root if and only if each of its blocks does.

Suppose M is nilpotent. Then $I + M$ has a square root

$$(I+M)^{1/2} = \sum_{r\ge 0} \binom{\frac{1}{2}}{r} M^r.$$

Note that this is a finite sum, since $M^r = 0$ for all but finitely many values of r. It follows that, if $\theta \neq 0$, then

$$\theta I + N_\theta = \theta(I + \theta^{-1}N_\theta)$$

has a square root. Hence we are left with the case where $\theta = 0$, and this our questions reduces to deciding which nilpotent matrices have a square root. If N is nilpotent with index k and $X^2 = N$, then $X^{2k} = 0$ and so X is nilpotent with index $2k$. (This implies that the matrices N_k are not squares.)

Assume now that N is in Frobenius normal form. We claim that the $(k+1) \times (k+1)$ matrix

$$N_k' := \begin{pmatrix} N_k & 0 \\ 0 & 0 \end{pmatrix}$$

has a square root. (You do it!) It follows that N has a square root if and only if its corank is at least as large as the number of non-zero blocks.

4.10 A Similarity Condition

We are given the following two $n \times n$ matrices:

$$\begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}, \qquad \begin{pmatrix} A & B \\ 0 & D \end{pmatrix},$$

where A and D are square. We ask for which matrices B are they similar.

We note that

$$\begin{pmatrix} I & -X \\ 0 & I \end{pmatrix} \begin{pmatrix} A & B \\ 0 & D \end{pmatrix} \begin{pmatrix} I & X \\ 0 & I \end{pmatrix} = \begin{pmatrix} A & AX - XD + B \\ 0 & D \end{pmatrix},$$

and deduce that they are similar if there is a matrix X such that

$$AX - XD = B.$$

We show that this condition is necessary.

Suppose that our two matrices are similar. Then there is an invertible matrix S such that

$$S\begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix} = \begin{pmatrix} A & B \\ 0 & D \end{pmatrix} S.$$

We define linear mappings T_1 and T_2 on the space of $n \times n$ matrices by

$$\begin{aligned} T_1(Y) &:= \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix} Y - Y \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}, \\ T_2(Y) &:= \begin{pmatrix} A & B \\ 0 & D \end{pmatrix} Y - Y \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}. \end{aligned}$$

We have

$$\begin{aligned} S(T_1(S^{-1}Y)) &= S\begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix} S^{-1}Y - Y\begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix} \\ &= \begin{pmatrix} A & B \\ 0 & D \end{pmatrix} Y - Y \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix} \\ &= T_2(Y), \end{aligned}$$

and therefore $\ker(T_1)$ and $\ker(T_2)$ have the same dimension.

Let Y be the matrix

$$Y = \begin{pmatrix} Y_{1,1} & Y_{1,2} \\ Y_{2,1} & Y_{2,2} \end{pmatrix},$$

where the partitioning is compatible with the partitioning of the other matrices above. Then

$$T_1(Y) = \begin{pmatrix} AY_{1,1} - Y_{1,1}A & AY_{1,2} - Y_{1,2}D \\ DY_{2,1} - Y_{2,1}B & DY_{2,2} - Y_{2,2}D \end{pmatrix}$$

and

$$T_2(Y) = \begin{pmatrix} AY_{1,1} - Y_{1,1}A + BY_{2,1} & AY_{1,2} - Y_{1,2}D + BY_{2,2} \\ DY_{2,1} - Y_{2,1}D & DY_{2,2} - Y_{2,2}D \end{pmatrix}.$$

We note that

$$\begin{pmatrix} Y_{1,1} & Y_{1,2} \\ 0 & -I \end{pmatrix}$$

lies in $\ker(T_2)$ if and only if $AY_{1,2} - Y_{1,2}D - B = 0$, and we can prove our claim by showing that there is a matrix of this form in $\ker(T_2)$.

Let $\mathscr{T}_i$ denote the restriction to $\ker T_i$ of the linear map

$$\begin{pmatrix} Y_{1,1} & Y_{1,2} \\ Y_{2,1} & Y_{2,2} \end{pmatrix} \mapsto \begin{pmatrix} Y_{2,1} & Y_{2,2} \end{pmatrix}.$$

We will prove that $\mathscr{T}_1$ and $\mathscr{T}_2$ have the same image.

From the expressions for $T_1(Y)$ and $T_2(Y)$, we see that $\ker \mathscr{T}_1 = \ker \mathscr{T}_2$. Further

$$\operatorname{im} \mathscr{T}_1 = \{\begin{pmatrix} Y_{2,1} & Y_{2,2} \end{pmatrix} : DY_{2,1} - Y_{2,1}A = 0,\ DY_{2,2} - Y_{2,2}D = 0\}$$

and im $\mathscr{T}_2$ consists of the elements of ker$\mathscr{T}_1$ for which there are matrices $Y_{1,1}$ and $Y_{1,2}$ such that

$$CY_{2,1} = Y_{1,1}A - AY_{1,1}, \qquad CY_{2,2} = Y_{1,2}D - AY_{1,2}.$$

It follows that im$\mathscr{T}_2 \subseteq$ im$\mathscr{T}_1$. Now

$$\begin{aligned}\dim((\mathscr{T}_1)) + \mathrm{rk}(\mathscr{T}_1) &= \dim(\ker T_1)\\ \dim((\mathscr{T}_2)) + \mathrm{rk}(\mathscr{T}_2) &= \dim(\ker T_2).\end{aligned}$$

Since $\mathscr{T}_1$ and $\mathscr{T}_2$ have the same corank and since T_1 and T_2 have the same corank, it follows that $\mathscr{T}_1$ and $\mathscr{T}_2$ have the same rank.

Finally, it easy to verify that

$$\begin{pmatrix} 0 & 0 \\ 0 & -I \end{pmatrix} \in \ker T_1$$

whence

$$\begin{pmatrix} 0 & -I \end{pmatrix} \in \operatorname{im} \mathscr{T}_1 = \operatorname{im} \mathscr{T}_2$$

and accordingly there is a matrix in ker T_2 of the form

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

This completes the proof.

4.11 Triangular Maps

A *flag* in V is a sequence $V_0, \dots, V_r$ of distinct subspaces such that

$$V_0 \subset V_1 \subset \cdots \subset V_r.$$

If dim $V = n$, then a flag contains at most $n+1$ subspaces, and a *maximal flag* is a flag with $n+1$ elements. A maximal flag $V_0, \dots, V_n$ has $V_0 = \{0\}$ and $V_n = V$. If $v_1, \dots, v_n$ is a basis for V and we define $V_0 = \{0\}$ and

$$V_i := \operatorname{span}\{v_1, \dots, v_i\}$$

then $V_0, \dots, V_n$ is a maximal flag. There is a converse to this. Suppose that $V_0, \dots, V_n$ is a maximal flag, and for $i = 1, \dots, n$ choose a non-zero vector w_i in $V_i \setminus V_{i-1}$. Then $w_1, \dots, w_n$ is a basis (as you are invited to prove). Let T be an endomorphism of V. A flag $\mathscr{F}$ is T-invariant if each subspace of $\mathscr{F}$ is T-invariant. If $\mathscr{F}$ is T-invariant, we also say that T *fixes* $\mathscr{F}$.

4.11.1 Lemma. *If $\beta = x_1, \dots, x_n$ is a basis for the vector space V and the linear map A fixes the flag associated to β, then the matrix that represents A relative to β is upper triangular.* □

4.11.2 Theorem. *An endomorphism of a finite-dimensional vector space over an algebraically closed field fixes a maximal flag.*

Proof. We prove the result by induction on $\dim V$. We define a hyperplane in V to be a subspace with dimension $\dim(V)-1$. It will be enough to prove that any endomorphism of a vector space fixes a hyperplane H for then, by induction, we may assume that $T\restriction H$ fixes a maximal flag of H.

By Lemma 3.5.2, the adjoint T^* of T has an eigenvector in V^*. Choose such an eigenvector f. Then $T^*f = \lambda f$ for some scalar λ, but T^*f is the composition $f\circ T$ and therefore

$$f(Tv) = \lambda f(v),$$

for all elements v of V. This implies that if $f(v) = 0$, then $f(Tv) = 0$ and therefore $\ker f$ is T-invariant.

Since $f \neq 0$, there is a vector v such that $f(v) \neq 0$. If $f(w) \neq 0$ too, then the vector

$$f(w)v - f(v)w$$

lies in $\ker f$, from which it follows that $\ker f$ is a hyperplane. □

In Section 4.12, we will prove a more concrete version of this result using a variation of the above argument.

(1) Prove that each maximal flag determines a basis, as described above.

(2) Prove that if $f \in V^*$, then $\ker f$ is a hyperplane.

(3) Let S and T be endomorphisms of V that fix the same flag, and suppose $n = \dim V$. Prove that the minimal polynomial of $ST - TS$ divides t^n.

4.12 Triangulations

We prove that if A is a square matrix over $\mathbb{C}$, then there is a unitary matrix L such that $L^{-1}AL$ is triangular. We have already proved a version of this result for linear mappings (see Section **??**) but our argument there did not yield the fact that we could choose L to be unitary.

4.12.1 Theorem. *Let A be an $n\times n$ matrix over $\mathbb{C}$. Then there is a unitary matrix L such that $L^{-1}AL$ is lower triangular.*

Proof. We proceed by induction on n. Let u_1 be an eigenvector for A^* with eigenvalue θ and let U denote the subspace

$$u^\perp = \{x\in\mathbb{C}^n : u^*x = 0\}.$$

Then U is A-invariant: if $v \in U$, then

$$u_1^*Av = (A^*u_1)^Tv = \theta u_1^*v = 0.$$

Let $u_2, \dots, u_n$ be an orthonormal basis for U. Since $u_1 \notin U$, the vectors $u_1, u_2, \dots, u_n$ form an orthonormal basis for $\mathbb{C}^n$. If we define the matrix L_1 by

$$L_1 := \begin{pmatrix} u_1 & u_2 & \cdots & u_n \end{pmatrix}$$

then L_1 is unitary and

$$AL_1 = L_1 \begin{pmatrix} a & 0 \\ b & A_2 \end{pmatrix}.$$

We may assume inductively that there is a unitary matrix M such that $M^{-1} A_2 M$ is lower triangular; then $L = L_1 M$ is the unitary matrix we need. □

Suppose that M is an upper triangular $n \times n$ matrix. If $Mv = \theta v$, then $(M - \theta I)v = 0$ and so $M - \theta I$ is not invertible. The matrix $M - \theta I$ is also upper triangular; it is invertible if and only if its diagonal entries are nonzero. We conclude that the eigenvalues of M are precisely the diagonal entries of M. This generalises the fact the the eigenvalues of a diagonal matrix are its diagonal entries.

4.13 The "Fundamental" "Theorem of Algebra"

The fundamental theorem of algebra is the assertion that any polynomial with coefficients from $\mathbb{C}$ has a root in $\mathbb{C}$. It is equivalent to the claim that every complex matrix has an eigenvector, and we offer a proof of this due to Harm Derkson. The original appears in the American Math. Monthly, and on his web page. (It has been stated that this result is theorem of analysis, not algebra, and is not fundamental. I tend to agree.)

This proof is due to Harm Derksen, American Math. Monthly, 110, (2003), pp. 620- 623.

4.13.1 Theorem. *Every square complex matrix has an eigenvector.*

Before setting out on the proof, some terminology. Let $\mathscr{A}$ be a set with a multiplication defined on it. If $A, B \in \mathscr{A}$, we denote their product by AB. A set $\mathscr{A}$ of endomorphisms of V is an *algebra* if

(a) $\mathscr{A}$ is a vector space over $\mathbb{F}$.

(b) If $A, B \in \mathscr{A}$, then $AB \in \mathscr{A}$.

(c) There is an element I in $\mathscr{A}$ such that $AI = IA = A$ for all A in $\mathscr{A}$.

If V is a vector space over $\mathbb{F}$, then $\operatorname{End}(V)$ is an algebra. If the elements of an algebra $\mathscr{A}$ are endomorphisms of V, it is called an *operator algebra*; if the elements of $\mathscr{A}$ are matrices we call it a *matrix algebra*. The set of all upper triangular matrices is an example of a matrix algebra. The set of strictly upper triangular matrices is not an algebra according to our definition, because it does not contain the identity matrix. An algebra $\mathscr{A}$ is commutative if $AB = BA$ for all A and B in $\mathscr{A}$. If A is a square matrix, the set of all polynomials in A is a commutative algebra.

We note next that if $f(t)$ is a polynomial over $\mathbb{R}$ with odd degree, then f has a real zero. (This is a comparatively simple exercise in calculus.)

We now start the proof of the theorem. We divide it into a number of lemmas.

4.13.2 Lemma. *If A is an $n \times n$ real matrix and n is odd, then A has an eigenvector.*

Proof. The space $\mathbb{R}^n$ is a direct sum of cyclic subspaces for A. Since n is odd, there is a cyclic subspace U for A with odd dimension d. The minimal polynomial ψ of $A{\restriction}U$ has degree d, and therefore there is a real number θ such that $\psi(\theta) = 0$. It follows that A has an eigenvector with eigenvalue θ. □

4.13.3 Lemma. *If $\mathscr{A}$ is a commutative algebra of real $n \times n$ matrices and n is odd, there is a vector z which is an eigenvector for all matrices in A.*

Proof. Let $A_1, \ldots, A_k$ be a basis for $\mathscr{A}$. If $\mathscr{A}$ is generated by I, there is nothing to prove, so we may assume $A_1 \neq I$. By the previous lemma, A_1 has an eigenvector z; let θ be its eigenvalue. The subspaces $\ker(A_1 - \theta I)$ and $\operatorname{im}(A_1 - \theta I)$ are proper non-zero subspaces of $\mathbb{R}^n$ and by the rank theorem,

$$\dim(\ker(A_1 - \theta I)) + \dim(\operatorname{im}(A_1 - \theta I)) = n.$$

Therefore one of these subspaces has odd dimension; we denote it by U.

If $A_1 u = \theta u$, then

$$A_1 A_i u = A_i A_1 u = \theta A_i u$$

and consequently $A_i u \in \ker(A_1 - \theta I)$ if $u \in \ker(A_1 - \theta I)$. If $v = (A_1 - \theta I)w$, then

$$A_i v = A_i (A_1 - \theta I) w = (A_1 - \theta I) A_i w \in \operatorname{im}(A_1 - \theta I).$$

Hence U is invariant under each matrix $A_1, \ldots, A_k$, and so it is invariant under all matrices in $\mathscr{A}$.

Since U is a proper non-zero subspace of $\mathbb{R}^n$ with odd dimension, it follows by induction that there is a vector in U which is an eigenvector for each matrix in $\mathscr{A}$. □

4.13.4 Lemma. *If A is an $n \times n$ complex matrix and n is odd, then A has an eigenvector.*

Proof. Let W denote the vector space of all $n \times n$ Hermitian matrices (which is not an algebra if $n > 1$). We define linear operators L_1 and L_2 by

$$\begin{aligned} L_1(M) &= \frac{1}{2}(AM + MA^*), \\ L_2(M) &= \frac{1}{2i}(AM - MA^*). \end{aligned}$$

If $M = M^*$, then

$$(L_1(M))^* = \frac{1}{2}(AM + MA^*)^* = \frac{1}{2}(MA^* + AM) = L_1(M)$$

and

$$(L_2(M))^* = \frac{1}{-2i}(AM - MA^*)^* = \frac{1}{-2i}(MA^* - AM) = L_2(M).$$

Therefore $L_1, L_2 \in \operatorname{End}(W)$. Also

$$L_1L_2(M) = \frac{1}{2}\frac{1}{2i}[A(AM - MA^*) + (AM - MA^*)A^*] = \frac{1}{2}\frac{1}{2i}[A^2M - M(A^*)^2]$$

and

$$L_2L_1(M) = \frac{1}{2}\frac{1}{2i}[A(AM + MA^*) - (AM + MA^*)A^*] = \frac{1}{2}\frac{1}{2i}[A^2M - M(A^*)^2].$$

Therefore L_1 and L_2 commute.

Now W is a vector space of dimension n^2 over $\mathbb{R}$, and n^2 is odd. If we choose a basis for W, the matrices representing L_1 and L_2 relative to this basis have order $n^2 \times n^2$ and they commute. Consequently they have a common eigenvector, and this is an eigenvector for L_1 and L_2 This eigenvector is a non-zero matrix M such that

$$L_1(M) = \lambda M, \quad L_2(M) = \mu M.$$

Then

$$AM = L_1(M) + iL_2(M) = (\lambda + i\mu)M$$

and this shows that each non-zero column of M is an eigenvector for A. □

4.13.5 Lemma. *If $\mathscr{A}$ is a commutative algebra of complex $n \times n$ matrices and n is odd, there is a vector z which is an eigenvector for all matrices in A.*

Proof. We simply apply the proof of Lemma 4.13.3. If $A_1, \ldots, A_k$ is a basis for $\mathscr{A}$ and A_1 has an eigenvector, then there is a non-zero proper subspace of $\mathbb{C}^n$ of odd dimension over $\mathbb{C}$ which is invariant under $\mathscr{A}$. By induction this contains an eigenvector for $\mathscr{A}$. □

4.13.6 Lemma. *A square complex matrix has an eigenvector.*

Proof. Assume $n = 2^k n_1$, where n_1 is odd. We prove the lemma by induction on k. Let W denote the space of all matrices M in $\operatorname{Mat}_{n\times n}(\mathbb{C})$ such that $M^T = -M$. We note that

$$\dim(W) = \binom{n}{2}$$

and therefore 2^k does not divide $\dim(W)$. We define two mappings L_1 and L_2 as follows:

$$L_1(M) = AM + MA^T,$$
$$L_2(M) = AMA^T.$$

Then $L_1, L_2 \in \mathrm{End}(W)$ and $L_1L_2 = L_2L_1$. Choose a basis for W. The matrices representing L_1 and L_2 relative to this basis commute and have order $\binom{n}{2} \times \binom{n}{2}$. By induction on k, the algebra generated by these matrices has an eigenvector M; this is an eigenvector for L_1 and L_2 and we may assume that its eigenvalues are λ and μ respectively. Hence

$$\mu M = AMA^T = A(L_1(M) - AM) = (\lambda A - A^2)M$$

and so

$$(A^2 - \lambda A + \mu I)M = 0.$$

Let z be a non-zero column of M. Then the minimal polynomial of A relative to z is quadratic, and so the A-cyclic subspace generated by z has dimension at most two. Assume that the minimal polynomial ψ of A relative to z is quadratic, and is equal to

$$t^2 - \lambda t - \mu.$$

This quadratic has two roots in $\mathbb{C}$, and so there are complex numbers θ and τ such that

$$(A - \theta I)(A - \tau I)z = 0.$$

If $(A - \tau I)z = 0$, then z is an eigenvector for A with eigenvalue τ; if $(A - \tau I)z \neq 0$ then $(A - \tau I)z$ is an eigenvector for A with eigenvalue θ. Thus we have shown that A has an eigenvector. □

5

Tensors

5.1 The Kronecker Product

If A and B are matrices over $\mathbb{F}$, we construct their *Kronecker product* $A \otimes B$ by replacing the ij-entry of A with

$$A_{i,j}B,$$

for all i and j. We find that

$$(A \otimes B)(u \otimes v) = Au \otimes Bv$$

and, more generally that

$$(A \otimes B)(C \otimes D) = AC \otimes BD,$$

provided only that the products AC and BD are defined. It follows that if x is an eigenvector for A and y is an eigenvector for B, then $x \otimes y$ is an eigenvector for $A \otimes B$. Consequently the eigenvectors of $A \otimes B$ are just the products $\lambda\mu$, where λ is an eigenvalue of A and μ is an eigenvalue of B. We also have

$$(A \otimes B)^T = A^T \otimes B^T.$$

If X is an $m \times n$ matrix, then $\operatorname{vec}(X)$ is the $mn \times 1$ matrix we get by stacking the columns of X one above the other. In other terms

$$\operatorname{vec}(X) = \sum X_{i,j} e_i \otimes e_j.$$

We have

$$\operatorname{vec}(AX) = (I \otimes A)\operatorname{vec}(X), \qquad \operatorname{vec}(XB) = (B^T \otimes I)\operatorname{vec}(X).$$

It follows for example, that there is a matrix X such that

$$AX - XB = C$$

if and only if

$$(I \otimes A - B^T \otimes I)\operatorname{vec}(X) = \operatorname{vec}(C).$$

The eigenvalues of the matrix $I \otimes A - B^T \otimes I$ are the differences $\mu - \lambda$, where λ is an eigenvalue of A and μ is an eigenvalue of B, and therefore it is invertible if and only if A and B have no eigenvalues in common.

Let P be the matrix such that

$$P(x\otimes y) = y\otimes x.$$

Then P maps $U\otimes V$ to $V\otimes U$. If $V=U$, then $P^2=P$. We say an element u of $V\otimes V$ is *symmetric* if $Pu=u$ and *antisymmetric* if $Pu=-u$. Thus $u\otimes u$ and

$$u\otimes v+v\otimes u$$

are symmetric and

$$u\otimes v-v\otimes u$$

is antisymmetric. (Thus symmetric and antisymmetric elements of $V\otimes V$ are eigenvectors for P, with eigenvalues 1 and -1 respectively.) If A and B belong to $\mathrm{End}(V)$, then

$$P(A\otimes B)P(u\otimes v) = P(A\otimes B)(v\otimes u) = (B\otimes A)(u\otimes v).$$

We also have

$$P\,\mathrm{vec}(X) = \mathrm{vec}(X^T).$$

(1) Show that the matrix $P(A\otimes A^T)$ is symmetric.

(2) Let V be $\mathrm{Mat}_{n\times n}(\mathbb{F})$ and let A be a fixed matrix. If $X\in V$, define the map Ad_A in $\mathrm{End}(V)$ by

$$\mathrm{Ad}_A(X) := AX-XA.$$

If $A^n=0$, prove that $\mathrm{Ad}_A^{2n}=0$.

5.2 Tensor Products

The *tensor product* $U\otimes V$ of two vector spaces U and V over $\mathbb{F}$ is defined as a quotient space. We start with the space of all finitely supported functions $\mathbb{F}^{U\times V}$, the tensor product is the quotient of this subspace modulo the subspace spanned $\mathscr{R}$ by vectors of the following forms:

(a) $a(u,v)-(au,v)$, $a(u,v)-(u,av)$ for $a\in\mathbb{F}$ and (u,v) in $U\times V$.

(b) $(u_1+u_2,v)-(u_1,v)-(u_2,v)$ for $u_1,u_2\in U$, $v\in V$.

(c) $(u,v_1+v_2)-(u,v_1)-(u,v_2)$ for $u\in U$, $v_1,v_2\in V$.

(Here we are using formal sum of finitely many terms to represent elements of $\mathbb{F}^{U\otimes V}$.) We denote the image of (u,v) in $U\otimes V$ by $u\otimes v$. The map that sends (u,v) to $u\otimes v$ is bilinear.

For finite-dimensional vector spaces, there is no harm in identifying the tensor product with Kronecker product.

The tensor product is not commutative, the spaces $U\otimes V$ and $V\otimes U$ are isomorphic but not equal. The tensor product is associative, in that

$$(U\otimes V)\otimes W\cong U\otimes(V\otimes W).$$

The vectors of the form $u \otimes v$ are known as *pure tensors*; they span $U \otimes V$ but do not form a basis. We note that a scalar times a pure tensor is a pure tensor, and so any element of $U \otimes V$ can be expressed as a sum of pure tensors. If $\alpha \in U \otimes V$, we define the *tensor rank* of α to be the least integer r such that α can be expressed as the sum of r pure tensors, that is, the least integer r such that

$$\alpha = \sum_{i=1}^{r} u_i \otimes v_i.$$

The key property of the tensor product is that it allows us to deal with linear maps in place of multilinear maps (at the cost of increasing dimensions). Thus if we have a bilinear map

$$\beta : U \times V \to W,$$

then there is a linear map $\hat{\beta}$ from $U \otimes V$ to W such that

$$\hat{\beta}(u \otimes v) = \beta(u, v).$$

If A and B are linear maps defined on U and V respectively, we define their tensor product $A \otimes B$ by

$$(A \otimes B)(u \otimes v) = Au \otimes Bv.$$

If we have inner products defined on U and V, we can define

$$\langle (u_1 \otimes v_1), (u_2 \otimes v_2) \rangle = \langle u_1, u_2 \rangle \langle v_1, v_2 \rangle.$$

This is a consequence of our definition of the tensor products of maps, because the maps $\langle u_1, ? \rangle$ and $\langle u_2, ? \rangle$ are elements of U^*.

The field $\mathbb{F}$ is a 1-dimensional vector space and so the tensor product $\mathbb{F} \otimes V$ is defined. The map that sends $1 \otimes v$ to v is an isomorphism. If $\psi \in U^*$, then $\psi \times I$ is a linear map from $U \otimes V$ to $\mathbb{F} \otimes V$, and hence it determines a linear map from $U \otimes V$ to V. We will usually identify these two maps.

5.3 *Quadratic Tensors*

We investigate properties of elements of the tensor product $U \otimes V$.

5.3.1 Lemma. *If $\alpha \in U \otimes V$ has tensor rank r and $\alpha = \sum_{i=1}^{r} u_i \otimes v_i$ for some vectors $u_1, \dots, u_r$ and $v_1, \dots, v_r$, then both of these sets of vectors are linearly independent.*

We leave the proof of this as an exercise.

5.3.2 Lemma. *If $\alpha \in U \otimes V$ has tensor rank r and*

$$\alpha = \sum_{i=1}^{r} u_i \otimes v_i = \sum_{i=1}^{r} x_i \otimes y_i$$

then

$$\operatorname{span}\{u_1, \dots, u_r\} = \operatorname{span}\{x_1, \dots, x_r\}, \qquad \operatorname{span}\{v_1, \dots, v_r\} = \operatorname{span}\{y_1, \dots, y_r\}.$$

Proof. There are vectors $\psi_1,\dots,\psi_r$ in U^* such that $\psi_i(u_j) = \delta_{i,j}$. So the image of α under the map $\psi_k \otimes 1$ is v_k, according to the first expression for α, and its image is

$$\sum_{i=1}^{r} \psi_i(x_i) y_i.$$

This shows that $v_k \in \operatorname{span}\{y_1,\dots,y_r\}$, and now everything follows. □

The previous results are analagous to properties of the usual rank of a matrix. This is no accident:

5.3.3 Theorem. *For any two vector spaces U and V, the spaces* $\operatorname{Lin}(U,V)$ *and $U^* \otimes V$ are isomorphic. Under this isomorphism elements of $U^* \otimes V$ with tensor rank r map to operators with rank r.*

Proof. If $\psi \in U^*$ and $v \in V$, let the map $\lambda_{\psi,v}$ be given by

$$\lambda_{\psi,v}(u) = \psi(u)v.$$

This assigns a linear map to each pure tensor in $U^* \otimes V$ and hence gives us a linear map from $U^* \otimes V$ to $\operatorname{Lin}(U,V)$. Denote this map by Λ.

We show that Λ is onto. The first step is to show that each linear map in $\operatorname{Lin}(U,V)$ with rank one is the image of a pure tensor. We leave this as an exercise.

The second step is to show that any $m \times n$ matrix can be written as a sum of rank-one matrices. Suppose A is $m \times n$. If $A \neq 0$, there are vectors x and y such that $xTAy \neq 0$, and so we may assume that we have vectors x and y such that $x^TAy = 1$. Define

$$B = A - Axy^TA^T.$$

Each column of Axy^TA^T is a scalar multiple of Ax, and it follows that the column space of B is contained in the column space of A. Next, $Ax \neq 0$ but

$$Bx = Ax - Axy^TA^Tx = Ax - (x^TAy)Ax = 0.$$

Therefore the column space of B is a proper subspace of the column space of A and so $\operatorname{rk}(B) < \operatorname{rk}(A)$. On the other hand

$$\operatorname{rk}(A) = \operatorname{rk}(B + Axy^TA) \le \operatorname{rk}(B) + \operatorname{rk}(Axy^TA^T) \le \operatorname{rk}(B) + 1$$

and we conclude that $\operatorname{rk}(B) = \operatorname{rk}(A) - 1$. It follows by induction that A can be expressed as the sum of r rank-one matrices. □

Note that it is imediate that a matrix with m rows is the sum of m rank-one matrices, and we can use this to provide a simple proof of the isomorphism in the above theorem. However the relation between tensor rank and the usual rank is inmportant.

5.3.4 Theorem. *We have*

$$\dim(U \otimes V) = \dim(U)\dim(V).$$

Proof. If $u_1, \dots, u_m$ and $v_1, \dots, v_n$ are basis for U and V respectively, then the pure tensors $u_i \otimes v_j$ span $U \otimes V$.

Now suppose that the pure tensors $u_i \otimes v_j$ (for all i and j) are linearly dependent. Then there are linearly independent vectors $u_1, \dots, u_r$ in U and vectors $w_1, \dots, w_r$ in V such that

$$0 = \sum_{i=1}^{r} u_i \otimes w_i.$$

As before, choose elements $f_1, \dots, f_r$ in U^* such that $\psi_i(u_j) = \delta_{i,j}$. If we apply $\psi_k \otimes 1$ to each side of the above expression, we get

$$0 = w_k. \qquad \square$$

5.4 Cubic Tensors

Consider a tensor α in $U \otimes V \otimes W$ given by

$$\alpha = \sum_{i=1}^{r} u_i \otimes \beta_i$$

where $\beta_1, \dots, \beta_r \in V \otimes W$. The subspace $\mathscr{C}(\alpha)$ of $V \otimes W$ spanned by the tensors $\beta_1, \dots, \beta_r$ is an invariant of α. Define the *order* of a subspace of $V \otimes W$ to be the least integer s such that it is contained in the span of s pure tensors. If $\mathscr{C}(\alpha)$ has order s, then there are vectors $v_1, \dots, v_s$ in V and pure tensors $\gamma_1, \dots, \gamma_s$ in $V \otimes W$ such that

$$\alpha = \sum_{i=1}^{s} x_i \otimes \gamma_i$$

Hence the tensor rank of α is at most s. Since no proper subset of $\gamma_1, \dots, \gamma_s$ spans $\mathscr{C}(\alpha)$, it follows that s is the tensor rank of α.

We give one example of the order of a subspace. Identify $V \otimes W$ with the vector space of matrices of order $\dim(V)$ times $\dim(W)$. If $\mathscr{C}$ is the space of upper-triangular 2×2 matrices then $\mathscr{C}$ has dimension three and order four.

For quadratic tensors, we have the following theorem.

5.4.1 Theorem. *For vector spaces V and W, the set*

$$S_k := \{T \in V \otimes W \mid rk(T) \le k\}$$

is closed (i.e., if $\lim_{i \to \infty} T_i = T$ and $rk(T_i) \le k$, then $rk(T) \le k$).

Proof. Each $T \in V \otimes W$ is associated with a matrix A whose rank is equal to the tensor rank of T. Hence the sets S_k are determined by algebraic equations and are closed. $\square$

The set of matrices of rank at most r is a closed set, and so the limit of any sequence of matrices with rank at most r is a matrix with rank at most r. Tensor rank is in general less well behaved. Let V be $\mathbb{R}^2$ with the standard basis e_1, e_2.

5.4.2 Lemma. *The element*

$$T := e_1 \otimes e_1 \otimes e_1 + e_1 \otimes e_2 \otimes e_2 + e_2 \otimes e_1 \otimes e_2$$

of $\mathbb{R}^2 \otimes \mathbb{R}^2 \otimes \mathbb{R}^2$ *has tensor rank three, but is the limit of a sequence of tensors with rank at most two.*

Proof. Define

$$T_\lambda := \lambda^{-1}[e_1 \otimes e_1 \otimes (-e_2 + \lambda e_1) + (e_1 + \lambda e_2) \otimes (e_1 + \lambda e_2) \otimes e_2)].$$

Then

$$T_\lambda - T = \lambda e_2 \otimes e_2 \otimes e_2,$$

whence T_λ converges to T as $\lambda \to 0$.

The only difficulty is to verify that T has tensor rank three. Suppose by way of contradiction that

$$T = (\alpha_1 e_1 + \alpha_2 e_2) \otimes b \otimes c + (\mu_1 e_1 + \mu_2 e_2) \otimes v \otimes w.$$

Then

$$T = e_1 \otimes (\alpha_1 b \otimes c + \mu_1 v \otimes w) + e_2 \otimes (\alpha_2 b \otimes c + \mu_2 v \otimes w).$$

Comparing this with the definition of T, we deduce that

$$\begin{aligned} e_1 \otimes e_1 + e_2 \otimes e_2 &= \alpha_1 b \otimes c + \mu_1 v \otimes w \\ e_1 \otimes e_2 &= \alpha_2 b \otimes c + \mu_2 v \otimes w. \end{aligned}$$

The two vectors on the left in these expressions are linearly independent, and therefore these equations imply that $b \otimes c$ and $v \otimes w$ are linearly independent and that they are linear combinations of the vectors on the left.

Now we use the isomorphism between $\mathbb{R}^2 \otimes \mathbb{R}^2$ and $\mathrm{Mat}_{2\times 2}(\mathbb{R})$. The image of the span of the vectors on the left consists of all matrices of the form

$$\begin{pmatrix} x & y \\ 0 & x \end{pmatrix}$$

All rank-one matrices of this form must have x equal to 0, and so the rank-one matrices of this form span a 1-dimensional space. It follows that $b \otimes c$ and $v \otimes w$ are linearly dependent. This the contradiction we wanted—we conclude that the tensor rank of T is three. □

5.5 Multiplication

Let M be the space of $n \times n$ matrices over some field. Matrix multiplication defines a linear map from $V \otimes V$ to V. By Theorem 5.3.3 we have

$$\mathscr{L}(V \otimes V, V) \cong M^* \otimes M^* \otimes M,$$

and so matrix multiplication can be viewed as a particular element of this space. More concretely, if the elements $E_{i,j}$ form a basis for M and $\epsilon_{i,j}$ denotes the element of M^* that sends a matrix to its ij-entry, then

$$AB = \sum_{i,j,k} \epsilon_{i,j}(A)\epsilon_{j,k}(B)E_{i,k}$$

and so matrix multiplication corresponds to the tensor

$$\sum_{i,j,k} \epsilon_{i,j} \otimes \epsilon_{j,k} \otimes E_{i,k}.$$

This is a sum of n^3 terms, which reflects the fact that the implies algorithm for the product of two $n \times n$ matrices requires n^3 multiplications of scalars. It is surprising and significant that the rank of this tensor is less than n^3. Strassen proved that when $n = 2$, its rank is at most seven, and this has lead to algorithms for matrix multiplication that, for large values of n, are substantially faster than the natural one.

For further information, start with Prasolov.

In the most general sense, an algebra is a vector space V with a bilinear multiplication μ defined on it. As above we can identify μ with a cubic tensor. For is $v_1, \ldots, v_d$ is a basis for V and γ_i is the element of V^* that maps a vector v to its i-th coordinate, then for x and y in V, we have

$$\mu(x, y) = \sum_{i,j=1}^{d} \gamma_i(x)\gamma_j(y)\mu(v_i, v_j).$$

and so we can identify μ with the element

$$\sum_{i,j} \gamma_i \otimes \gamma_j \otimes \mu(v_i, v_j)$$

of $V^* \otimes V^* \otimes V$ or, if we willing to be flexible, with an element of $V^{\otimes 3}$.

5.6 *Semifields*

If

$$V := V_1 \otimes \cdots \otimes V_d$$

and φ lies in the dual space V_j^*, we define $\varphi^{(j)}$ to be the linear map such that

$$\varphi^{(j)}(v_1 \otimes \cdots \otimes v_d) = \varphi(v_j)(v_1 \otimes \cdots \otimes v_{j-1}) \otimes (v_{j+1} \otimes \cdots \otimes v_d)$$

We refer to $\varphi^{(j)}$ as a *contraction*. Following Liebler [1] we define, an element of a tensor product $V_1 \otimes \cdots \otimes V_d$ to be *non-singular* if

(a) $d = 1$, any non-zero element of V is non-singular.

(b) $d > 1$, and any non-zero contraction is non-singular.

[1]

The second-simplest case is when $d = 2$, in this case an element of $V_1 \otimes V_2$ is non-singular if the corresponding matrix is non-singular (in the usual meaning of the term).

(1) Suppose $T \in \operatorname{End}(V)$ and $\operatorname{rk}(T) = 1$. Prove that there is f in V^* and v in V such that $Tx = f(x)v$.

6
Type-II Matrices

6.1 Definitions

If M and N are $m \times m$ matrices, their *Schur product* is the $m \times n$ matrix $M \circ N$ defined by

$$(M \circ N)_{i,j} = M_{i,j} N_{i,j}$$

This is a commutative associative product with the all-ones matrix J as multiplicative identity. If no entry of M is zero, we define the matrix $M^{(-)}$ by

$$(M^{(-)})_{i,j} := M_{i,j}^{-1}$$

and call it the *Schur inverse* of M; clearly $M \circ M^{(-)} = J$. If M is a Schur invertiblematrix we define

$$M_{i/j} := (Me_i) \circ (Me_j)^{(-)}.$$

Thus $M_{i/j}$ is the ratio of the i-th and j-th columns of M.

An $n \times n$ complex matrix w is a *type-II matrix* if it is Schur invertible and

$$WW^{(-)T} = nI$$

Any Hadamard matrix is a type-II matrix, as is the character table of an abelian group. For any nonzero complex number t, the matrix

$$W = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & t & -t \\ 1 & -1 & -t & t \end{pmatrix}$$

is type-II.

If W_1 and W_2 are type-II, so is the Kronecker product $W_1 \otimes W_2$.

We define a *monomial matrix* to be any product of permutation matrices and invertible diagonal matrices. Matrices W_1 and W_2 are *monomially equivalent* if there are monomial matrices M and N such that $W_2 = MW_1N$. If W is monomially equivalent to a type-II matrix, it is a type-II matrix.[1]

[1] as you may show

If W is type-II, so is its transpose W^T, but in general W and W^T are not monomially equivalent.

Our next result introduces an important class of type-II matrices. We say that a complex matrix is *flat* if its entries all have the same absolute value.

6.1.1 Lemma. *Suppose W is a square Schur-invertible matrix. Then any two of the following statements imply the third:*

(a) W is type-II.

(b) W is flat.

(c) W is unitary.

A flat unitary matrix is commonly referred to as a *complex Hadamard matrix*.

6.2 Traces and Type-II Matrices

6.2.1 Lemma. *An $n \times n$ matrix W is type-II if and only if for any two diagonal matrices D_1 and D_2,*

$$\langle D_1, W^{-1} D_2 W\rangle = \frac{1}{n} \operatorname{tr}(D_1) \operatorname{tr}(D_2).$$

Proof. We have

$$\langle e_i e_i^T, W^{-1} e_j e_j^T W\rangle = \operatorname{tr}(e_i e_i^T W^{-1} e_j e_j^T W) = e_i^T W^{-1} e_j e_j^T W e_i = (W^{-1})_{i,j} W_{j,i},$$

and so our claim holds for $D_1 = e_i e_i^T$ and $D_2 = e_j e_j^T$ if and only if

$$(W^{-1})_{i,j} W_{j,i} = \frac{1}{n}.$$

It holds for all i and j if and only if $W^{-1} = \frac{1}{n} W^{(-1)T}$, i.e., if W is type-II. The result now follows by linearity. □

6.2.2 Corollary. *If W is type-II of order $n \times n$ and D is diagonal,*

$$(W^{-1} D W)_{i,i} = \frac{1}{n} \operatorname{tr}(D). \qquad \square$$

6.3 Compressions and Projections

A *resolution of the identity* is a sequence of projections $Q_1, \dots, Q_m$ such that $\sum_i Q_i = I$. (So each direct sum decomposition of a vector space provides a resolution of the identity.[2])

6.3.1 Lemma. *If $P_1, \dots, P_m$ are projections and their sum is a projection, then $P_i P_j = 0$ when $i \neq j$.*

[2] The converse is also true

Proof. Assume $Q = \sum_i P_i$. Then

$$Q = Q^2 = \sum_i P_i + \sum_{i \neq j} P_i P_j = Q + \sum_{i \neq j} P_i P_j$$

and therefore

$$0 = \operatorname{tr}\left(\sum_{i \neq j} P_i P_j\right) = \sum_i \operatorname{tr}(P_i P_j).$$

Since projections are positive semidefinite, if $\operatorname{tr}(P_i P_j) = 0$, then $P_i P_j = 0$. □

If follows that if $\sum_i P_i = I$, then the matrices P_i commute. Since these matrices are normal, there is a change of basis that makes them diagonal. Hence if $P_1, \ldots, P_m$ is a resolution of the identity, by making a change of basis we may assume each P_i is diagonal, with all diagonal entries 0 or 1.

If $\mathscr{A}$ is an algebra and $P_1, \ldots, P_m$ is a resolution of the identity formed by projections in $\mathscr{A}$, we define a map Ψ from $\mathscr{A}$ to itself by setting

$$\Psi(M) := \sum_i P_i M P_i.$$

You may verify that the image of Ψ is a subalgebra of $\operatorname{End}(\mathscr{A})$. The map Ψ is referred to as a *compression*. (If we have chosen a basis so the P_i's are diagonal, then $\Psi(M)$ will be block diagonal—equivalently Ψ just sets the off-diagonal blocks to zero.)

6.3.2 Lemma. *Let Ψ be the element of* $\operatorname{End}(\mathscr{A})$ *arising from the resolution of the identity $P_1, \ldots, P_m$. Then*

(a) Ψ is self-adjoint.

(b) Ψ is a projection.

(c) The image of Ψ is set of matrices in $\mathscr{A}$ that commute with $P_1, \ldots, P_m$.

Proof. We calculate:

$$\langle M, \Psi(N)\rangle = \sum_i \operatorname{tr}(M^* P_i N P_i) = \sum_i \operatorname{tr}((P_i M^* P) N) - \langle (, \Psi\rangle(M), N).$$

This gives (a) and, since $\Psi^2 = \Psi$, we get (b).

Since $\Psi(M)$ is a sum of matrices that commute with $P_1, \ldots, P_m$, it commutes with each element of $P_1, \ldots, P_m$. Since Ψ acts as the identity on the commutant of $P_1, \ldots, P_m$, we see that Ψ is onto. □

6.3.3 Lemma. *Suppose $P_1, \ldots, P_k$ are pairwise orthogonal projections summing to I. If W is a $k \times k$ type-II matrix and we define*

$$U_i = \frac{1}{\sqrt{n}} \sum_j W_{i,j} P_j \qquad (i = 1, \ldots, k),$$

then $U_1, \ldots, U_k$ are invertible and

$$\sum_i P_i \otimes P_i = \sum_i U_i \otimes U_i^{-1}.$$

If W is unitary, so are $U_1, \ldots, U_k$.

Proof. We have $U_i^{-1} = n^{-1/2} \sum_j W_{i,j}^{-1} P_j$ and consequently

$$\begin{aligned} nU_i \otimes U_i^{-1} &= \left(\sum_j W_{i,j} P_j \right) \otimes \left(\sum_k W_{i,k}^{-1} P_k \right) \\ &= \left(\sum_{j,k} W_{i,j} W_{i,k}^{-1} \right) P_j \otimes P_k. \end{aligned}$$

Therefore

$$\begin{aligned} n \sum_i U_i \otimes U_i^{-1} &= \sum_{j,k} \left(\sum_i W_{i,j} W_{i,k}^{-1} \right) P_j \otimes P_k \\ &= \sum_{j,k} (W^{(-)T} W)_{j,k} P_j \otimes P_k \\ &= n \sum_j P_j \otimes P_j. \end{aligned}$$

If W is flat, the eigenvalues of the matrices U_i are complex numbers of norm one, since U_i is Hermitian it follows that U_i is unitary. □

6.4 *Classical Colourings*

We give a description of classical colourings of graphs in linear algebraic terms, with a view to presenting a quantum analog later.

A colouring of the graph X is a partition π of $V(X)$, such that the subgraphs induced by the cells of π are cocliques. We can represent a partition by its *characteristic matrix*. If $|\pi|$ denotes the number of cells of π this is the $|V(X)| \times |\pi|$ matrix whose i-th column is the characteristic vector of the i-th cell of π. So the characteristic matrix is a 01-matrix with column sum equal to $\mathbf{1}$.

If N is the characteristic matrix of a partition with m classes we can, for each i, convert the i-th column to a diagonal matrix P_i. (Thus $(P_i)_{u,u} = (Ne_i)_u$.) Then $P_i^2 = P_i$ and $\sum_i P_i = I$, and we have a resolution of the identity. We refer to the matrices P_i as the projections associated to the cells of π.

6.4.1 Lemma. *Assume $P_1, \dots, P_m$ are the projections associated to a partition π of $V(X)$. Then π is a colouring if and only if $\sum_i P_i A(X) P_i = 0$.* □

We use this lemma, along with earlier work, to derive an eigenvalue bound on the chromatic number.

6.4.2 Theorem. *Let X be a graph on n vertices with eigenvalues*

$$\theta_1 \geq \cdots \geq \theta_n.$$

Then

$$\chi(X) \geq 1 - \frac{\theta_1}{\theta_n}.$$

Proof. Let π be the partition of a c-colouring of X. Let $P_1, \dots, P_c$ be the corresponding projections and let $U_1, \dots, U_c$ be the unitary matrices derived from the projections using some flat type-II matrix. Then

$$0 = \sum_{i=1}^{c} P_i A P_i = \sum_i U_i A U_i^*$$

and therefore

$$U_1 A U_1^* = -\sum_{i=2}^{c} U_i A U_i^*$$

whence

$$A = -\sum_{i=2}^{c} U_1^* U_i A U_i^* U_1.$$

This implies that

$$\begin{aligned} \theta_1 &\le (c-1)\max\{\theta_1(U_1^* U_i(-A)U_i^* U_1) : 2 \le i \le c\} \\ &\le (c-1)\max\{\theta_1(-A) : 2 \le i \le c\} \\ &= (c-1)(-\theta_n) \end{aligned}$$

and the bound follows. □

6.5 *Quantum Permutations*

A *quantum permutation* P is an $n \times n$ matrix such that each entry is a $d \times d$ projection, and the projections in each row and column sum to I_d. We prefer to view P as a matrix over the ring of $d \times d$ matrices but, occasionally it is convenient to view it as an $nd \times nd$ matrix with blocks of size $d \times d$. In this case we will write $\tilde{P}$ to warn the reader of the change of viewpoint.

Note that if $Q_1, \dots, Q_k$ are projections and $\sum_i Q_i = I$, then $Q_i Q_j = 0$ when $i \neq j$. If the entries in a quantum permutation P all have rank one, then P is also known as a *quantum Latin square*.

6.5.1 Lemma. *Suppose P is an $n \times n$ quantum permutation with $d \times d$ projections as entries. Then $\tilde{P}$ is unitary.*

Proof. Easy exercise. □

An important consequence of this result is that P and $\tilde{P}$ are invertible.

Following Roberson et al [3], we define two graphs X and Y on n vertices to be *quantum isomorphic* if there is a quantum permutation P of order $n \times n$, with entries projections of order $d \times d$, such that

$$(A(X) \otimes I_d)\tilde{P} = \tilde{P}(A(Y) \otimes I_d).$$

If $X = Y$, we have a *quantum automorphism* of X. Since P is unitary, the matrices $A(X) \otimes I$ and $A(Y) \otimes I$ are similar, and so we see that quantum isomorphic graphs are cospectral. We'll see that more is true, but there are graphs that are quantum isomorphic but not isomorphic. (See [4].)

[3]

[4]

An automorphism of the graph X on n vertices can be specified by an $n \times n$ permutation matrix Q such that $QA = AQ$. Then $Q \otimes I$ and $A \otimes I$ commute, and we see that any automorphism of a graph gives rise to quantum automorphism,

6.5.2 Lemma. *If P is a quantum permutation, $\tilde{P}$ commutes with $J \otimes I_d$.* □

This result is easy to prove, and is left to the reader. One consequence of it is that quantum isomorphic graphs are cospectral with cospectral complements.

Our next results holds provided the entries in any row of P satisfy

$$P_{i,j}P_{i,k} = \delta_{j,k}P_{i,j};$$

they do not need to be projections.

6.5.3 Lemma. *If P is a quantum permutation and $\tilde{P}$ commutes with $M \otimes I$ and $N \otimes I$, it commutes with $(M \circ N) \otimes I$.*

Proof. The ij-block of $(M \otimes I)\tilde{P}$ is

$$\sum_r M_{i,r}P_{r,j}$$

and, by hypothesis, this is equal to the ij-block of $\tilde{P}(M \otimes I)$:

$$\sum_s M_{s,j}P_{i,s}.$$

We have

$$\sum_r M_{i,r}P_{r,j} \sum_s N_{i,s}P_{s,j} = \sum_r (M_{i,r}N_{i,r})P_{r,j}$$

where the right side is the ij-block of $((M \circ N) \otimes I)\tilde{P}$. Similarly

$$\sum_r M_{r,j}P_{i,r} \sum_r N_{r,j}P_{i,r} = \sum_r (M_{r,j}N_{r,j})P_{i,r}$$

where the right side is the ij-block of $\tilde{P}((M \circ N) \otimes I)$. Since the left sides of the previous pair of equations are equal, our result follows. □

6.5.4 Lemma. *Let P be a quantum permutation. The set of matrices M such that $M \otimes I$ commutes with $\tilde{P}$ is $*$-closed.*

Proof. Since the entries of P are Hermitian, we have

$$(\tilde{P}(M^* \otimes I))_{x,y} = \sum_r P_{x,r}M^*_{r,y} = \sum_r \left(P_{x,r}M_{r,y}\right)^* = \left((\tilde{P}(M \otimes I))_{x,y}\right)^*$$

and, if P and $M \otimes I$ commute, then

$$\left((\tilde{P}(M \otimes I))_{x,y}\right)^* = ((M^* \otimes I)\tilde{P})_{x,y}.$$

It follows that if $M \otimes I$ commutes with $\tilde{P}$, so does $M^* \otimes I$. □

From the previous two lemmas we see that the set of matrices M such that $\tilde{P}$ commutes with $(M \otimes I)$ is a coherent algebra.

6.6 Quantum Colourings

We introduce a quantum version of colouring. The basic idea is that a quantum c-colouring of X is given by an $n \times c$ matrix whose entries are $d \times d$ projections, with the projections in each row sum to the identity matrix I_d. (If $d = 1$, we will get a classical colouring.)

6.6.1 Lemma. *If $Q_1, \ldots, Q_m$ are projections with sum I_d, then $Q_i Q_j = 0$ if $i = j$. If $R_1, \ldots, R_m$ is a second sequence of projections summing to I_d, then $\sum_i Q_i R_i = 0$ if and only if $Q_i R_i = 0$ for all i.*

Proof. Set $S = \sum_i Q_i$. Then

$$S^2 = \sum_i Q_i + \sum_{i \neq j} Q_i Q_j.$$

As $S = I$, this implies that $\sum_{i \neq j} Q_i Q_j = 0$ and consequently

$$0 = \sum_{i \neq j} \operatorname{tr}(Q_i Q_j).$$

Since projections are positive semidefinite, $\operatorname{tr}(Q_i Q_j) \geq 0$ and equality holds if and only if $Q_i Q_j = 0$. □

If $Q_1, \ldots, Q_m$ and $R_1, \ldots, R_m$ are resolutions of the identity such that $Q_i R_i = 0$ for all i, we say the resolutions are *orthogonal*.[5]

We now define a *quantum m-colouring* of X to be a $|V(X)| \times m$ matrix N of $d \times d$ projections such that:

(a) Each row is a resolution of I_d.

(b) If vertices i and j are not adjacent in X, the partitions in the rows $e_i^t N$ and $e_j^T N$ are orthogonal.

We refer to d as the *index* of the quantum colouring. The quantum colourings of index one are precisely the clasical colourings.

6.6.2 Lemma. *A quantum m-colouring of K_m is a quantum permutation.*

Proof. if N is a quantum m-colouring of K_m, then the projections in any column of N are pairwise orthogonal and hence each column sum is an idempotent of order $d \times d$. Therefore

$$D_j := I - \sum_{i=1}^{m} N_{i,j} \succcurlyeq 0$$

and so $\sum_j D_j \succcurlyeq 0$. But

$$\sum_j D_j = \sum_{j=1}^{m} \sum_{i=1}^{m} (I - \sum_j N_{i,j}) = \sum_{i=1}^{m} \sum_{j=1}^{m} (I - \sum_j N_{i,j}) = 0$$

and therefore $D_j = 0$ for all j, that is, $\sum_i N_{i,j} = I_d$. □

[5] just what we need, another meaning assigned to ‘orthogonal’

6.7 *The Nomura Algebra of a Type-II Matrix*

If W is Schur-invertible, we define $W_{i/j}$ to be the vector $We_i \circ (W^{(-)e_j})$.

To each $m \times n$ Schur-invertible matrix W we associate its *Nomura algebra*, defined to be the set of $m \times m$ matrices M such that each ratio $W_{i/j}$ is an eigenvector. Hence M lies in the Nomura algebra of W if and only if there are scalars $\Theta_{i,j}(M)$ such that

$$MW_{i/j} = \Theta_{i,j}(M) W_{i/j}.$$

We denote this algebra by $\mathcal{N}_W$. It contains the identity matrix, so it is at least not empty.

6.7.1 Lemma. *A square Schur-invertible matrix W is type-II if and only if $J \in \mathcal{N}_W$.* □

So if W is type-II, the dimension of its Nomura algebra is at least two.

There is a non-trivial class of examples based on finite abelian groups. Assume G is an abelian group of order n, given by $n \times n$ permutation matrices, and let W be its character table, with rows indexed by group elements and columns by characters. Then $W_{i/j}$ is a character of G, and therefore $\mathcal{N}_W$ consists of the matrices M for which there is a diagonal matrix D such that $MW = WD$. It is not hard to verify that all permutation matrices in G belong to $\mathcal{N}_W$.

It is surprisingly difficult to provide examples of type-II matrices where the dimension of the Nomura algebra is greater than two. We can use products to get examples which we deem trivial: It can be proved that if W_1 and W_2 are type-II matrices, then

$$\mathcal{N}_{W_1 \otimes W_2} \cong \mathcal{N}_{W_1} \otimes \mathcal{N}_{W_2}.$$

Hence if W is the Kronecker product of k type-II matrices,

$$\dim(\mathcal{N}_W) \geq 2^k.$$

A type-II matrix W is a *spin model* if $W \in \mathcal{N}_W$ (or, more precisely, if W is monomially equivalent to an element of $\mathcal{N}_W$). Spin models are important because they give rise to link invariants. The character tables of abelian groups provide examples of spin models where the type-II matrices are flat; the only known examples where the type-II matrix is not flat is one based on the Higman-Sims graph (due to Jaeger [6]) and a family due to Nomura [7] based on Hadamard matrices.

[6]

[7]

6.8 *The Matrix of Idempotents of a Type-II Matrix*

We describe an operation on type-II matrices which we can use to construct quantum permutations. Assume W is an $n \times n$ type-II matrix and, for each i and j, define a rank-1 matrix $\mathscr{Y}_{i,j}$ by

$$Y_{i,j} := \frac{1}{n} W_{i/j} (W_{j/i})^T.$$

Let $\mathscr{Y}_W$ denote the $n \times n$ matrix with ij-entry equal to $Y_{i,j}$ (for all i and j). We call it the *matrix of idempotents* of W.

We observe that

$$Y_{i,i} = \frac{1}{n} J$$

and

$$Y_{i,j}^T = Y_{j,i}.$$

The latter implies that $\mathscr{Y}_W$ is symmetric. Further

$$Y_{i,j}^{(-)} = n W_{j/i} (W_{i/j})^T = n^2 Y_{j,i}.$$

If $\mathscr{Y}^\tau$ denotes the matrix we get by replacing each entry of $\mathscr{Y}$ by its transpose (i.e., the partial transpose of $\mathscr{Y}$), then

$$\mathscr{Y}^\tau = \frac{1}{n} \mathscr{Y}^{(-)}.$$

Finally, if W is flat, then $Y_{i,j}$ is Hermitian.

6.8.1 Theorem. *Let $\mathscr{Y}$ be the matrix of idempotents of the $n \times n$ type-II matrix W. Then each row and column of $\mathscr{Y}$ sums to I.*

Proof. Let $\partial_i(M)$ denote the diagonal matrix such that $(\partial_i(M)_{r,r} = (Me_i)_r$. We have

$$n \sum_j Y_{i,j} = \sum_j W_{i/j} (W_{j/i})^T = \partial_i(W) \left(\sum_j (We_j)^{(-)} (We_j)^T \right) \partial_i(W)^{-1}.$$

Here the inner sum is equal to

$$W^{(-)} W^T = (W W^{(-)T})^T = nI.$$

Since $\mathscr{Y}$ is symmetric, the result follows. □

Let S be the endomorphism of $\mathbb{C}^n \otimes \mathbb{C}^n$ that sends $u \otimes v$ to $v \otimes u$. Note that $S^2 = I$ and S is a permutation matrix.

6.8.2 Theorem. *If W is a type-II matrix, its matrix of idempotents $\mathscr{Y}$ is a type-II matrix. If W is flat, then $\mathscr{Y}$ is flat and is a quantum permutation.*

Proof. For fixed i, the vectors We_j form a basis of $\mathbb{C}^n$ and the vectors $n^{-1}(We_j)^{(-)}$ form a basis dual to this. Hence the matrices

$$\frac{1}{n} (We_j)^{(-1)} (We_j)^T$$

are pairwise orthogonal idempotents and sum to I. Therefore for fixed i the matrices $Y_{i,j}$ are pairwise orthogonal idempotents that sum to I.

Since $\mathscr{Y}^T = \mathscr{Y}$, it also follows that each column of $\mathscr{Y}$ consists of pairwise orthogonal idempotents that sum to I. If W is flat, then $Y_{i,j}$ is Hermitian. □

6.8.3 Corollary. *If W is a Hadamard matrix, $\mathscr{Y}_W$ is a Hadamard matrix of Bush type.* □

6.8.4 Lemma. *If W is type-II, then $\mathscr{Y}_{W^T} = S\mathscr{Y}_W S$.*

Proof. We have

$$n(Y_{i,j})_{r,s} = \frac{W_{r,i}}{W_{r,j}}\frac{W_{s,j}}{W_{s,i}} = \frac{W_{r,i}}{W_{s,i}}\frac{W_{s,j}}{W_{r,j}} = \frac{W^T_{i,r}}{W^T_{i,s}}\frac{W^T_{j,s}}{W^T_{j,r}} = n(Y_{r,s}(W^T))_{i,j}.$$

Here the left hand and right hand terms are equal respectively to

$$(e_i \otimes e_r)^T \mathscr{Y}_W (e_j \otimes e_s), \qquad (e_r \otimes e_i)^T \mathscr{Y} (e_s \otimes e_j)$$

and the result follows. □

6.9 Commutants

After one preparatory lemma, we derive a useful second characterization of $\mathscr{N}_W$.

6.9.1 Lemma. *If W is an $n \times n$ type-II matrix, then $(W_{k/i})^T W_{i/j} = n\delta_{j,k}$.*

Proof. We have

$$\begin{aligned}(W_{k/i})^T W_{i/j} = \sum_r \frac{W_{r,k}}{W_{r,i}}\frac{W_{r,i}}{W_{r,j}} = \sum_r \frac{W_{r,k}}{W_{r,j}} &= (W^{(-)}e_j)^T W e_k \\ &= (W^{(-)T}W)_{j,k} \\ &= n\delta_{j,k}.\end{aligned}$$

6.9.2 Theorem. *Let W be a type-II matrix. For a matrix M, the following are equivalent:*

(a) $M \in \mathscr{N}_W$.

(b) Each vector $W_{i/j}$ is an eigenvector for M.

(c) Each vector $(W_{j/i})^T$ is a left eigenvector for M.

(d) M commutes with $Y_{i,j}$ for all i and j.

Proof. From the definition of $\mathscr{N}_W$, (a) implies (b).

If L is a matrix of eigenvectors for A and D is the corresponding diagonal matrix of eigenvalues, then $AL = LD$, so $L^{-1}A = DL^{-1}$ and the rows of L^{-1} are left eigenvectors for A. If

$$L = \begin{pmatrix} W_{1/j} & \dots & W_{n/j}\end{pmatrix}$$

then, by the previous lemma,

$$\begin{pmatrix} (W_{j/1})^T \\ \vdots \\ (W_{j/n})^T \end{pmatrix} L = nI.$$

Therefore the vectors $(W_{j/i})^T$ are left eigenvectors for M.

Finally, suppose $Mxy^T = xy^T$. As the column space of Mxy^T is spanned by Mx and the column space of $xy^T M$ is spanned by x, we see that x is an eigenvector for M. Similarly y^T is a left eigenvector. Conversely, if x is a right eigenvector for M and y a left eigenvector for M, then xy^T and M commute. Hence (b) and (c) are equivalent to (d). □

6.9.3 Corollary. *If W is type-II, then $\mathcal{N}_W$ is transpose-closed.*

Proof. The matrix M commutes with $Y_{i,j}$, if and only if M^T commutes with $(Y_{i,j])^T})$. Since $(Y_{i,j])^T}) = Y_{j,i}$, the corollary follows. □

6.10 Coherent Algebras

If P is a permutation matrix, then $P(A \circ B) = (PA) \circ (PB)$. We derive an analogous result for quantum permutations.

6.10.1 Lemma. *Let $\mathcal{Z}$ be an $n \times n$ matrix with entries $d \times d$ idempotents. If the entries in each row and in each column are pairwise orthogonal and sum to I_d, then for any $n \times n$ matrices M and N,*

$$\begin{aligned} ((M \circ N) \otimes I)\mathcal{Z} &= ((M \otimes I)\mathcal{Z}) \circ ((N \otimes I)\mathcal{Z}) \\ \mathcal{Z}((M \circ N) \otimes I) &= (\mathcal{Z}(M \otimes I)) \circ (\mathcal{Z}(N \otimes I)) \end{aligned}$$

Proof. We have

$$((M \otimes I)\mathcal{Z})_{i,j} = \sum_r M_{i,r}\mathcal{Z}_{r,j}, \quad ((N \otimes I)\mathcal{Z})_{i,j} = \sum_r N_{i,r}\mathcal{Z}_{r,j}.$$

Since the entries in a column of $\mathcal{Z}$ are pairwise orthogonal,

$$((M \otimes I)\mathcal{Z})_{i,j}((N \otimes I)\mathcal{Z})_{i,j} = \sum M_{i,r}N_{i,r}\mathcal{Z}_{r,j} = (((M \circ N) \otimes I)\mathcal{Z})_{i,j}.$$

This proves the first equality, the second follows similarly. □

6.11 A Nomura Algebra is Schur-Closed

Let W be a type-II matrix. Recall that if $M \in \mathcal{N}_W$, then $\Theta_{i,j}(M)$ is the eigenvalue of M on the eigenvector $W_{i/j}$. Accordingly we define $\Theta_W(M)$ to be the matrix with

$$(\Theta_W(M))_{i,j} = \Theta_{i,j}(M);$$

it is the *matrix of eigenvalues* of M. Clearly, if $M, N \in \mathcal{N}_W$, then

$$\Theta_W(MN) = \Theta_W(M) \circ \Theta_W(N).$$

If M and N are square matrices of the same order, their Lie bracket is

$$[M,N] := MN - NM.$$

Obviously $[M,N] = 0$ if and only if M and N commute (and this is the only property of the Lie bracket that we will use.)

6.11.1 Theorem. *Let W by type-II and let $\mathcal{Y}$ be its matrix of idempotents. Then*

$$\mathcal{N}_W = \{M : [I \otimes M, \mathcal{Y}_W] = 0\},$$

and

$$\mathcal{N}_{W^T} = \{N : [N \otimes I, \mathcal{Y}_W] = 0\}.$$

Proof. We have that $[I \otimes M, \mathcal{Y}] = 0$ if and only if $[M, Y_{i,j}] = 0$ for all i and j. Now M commutes with a rank-1 matrix uv^* if and only if u is a right eigenvector for M and v^* is a left eigenvector. Hence $[M, Y_{i,j}] = 0$ for fixed i and all j if and only if $M \in \mathcal{N}_W$.

For the second claim,

$$S((N \otimes I)\mathcal{Y}_W)S = (I \otimes N)\mathcal{Y}_{W^T},$$

from which the assertion follows. □

6.11.2 Corollary. *If W is a type-II matrix, then $\mathcal{N}_W$ is Schur-closed.* □

Since $\mathcal{N}_W$ is closed under transpose, it follows that it is the Bose-Mesner algebra of an association scheme. Similarly $\mathcal{N}_{W^T}$ will be a Bose-Mesner algebra; the relation between these two algebras is described in the following theorem, which we would like to be able to prove using the machinery at hand.

We have

$$(M \otimes I)\mathcal{Y} = (\Theta(M) \otimes J) \circ \mathcal{Y}.$$

6.11.3 Theorem (Nomura)**.** *If W is a type-II matrix of order $n \times n$, then*

$$\Theta_W(M) \in \mathcal{N}_{W^T}$$

and

$$\Theta_W(M \circ N) = \frac{1}{n}\Theta_W(M)\Theta_W(N).$$ □

7
The Smith Normal Form

In this chapter we study some linear algebra over rings. The most important rings we use are $\mathbb{Z}$ and $\mathbb{F}[x]$.

7.1 Domains

Let R be a commutative ring. We say that an element a of R divides an element b if $b = ax$ for some x. We call R a *domain* if it has no divisors of zero, that is, if $a, b \in R$ and $ab = 0$ then $a = 0$ or $b = 0$. Clearly any field is a domain. Further examples are provided by the integers $\mathbb{Z}$ and $\mathbb{F}[x]$, the ring of polynomials in x with coefficients from $\mathbb{F}$.

An ideal of R is a non-empty subset I such that if $a \in I$ and $r \in R$, then $ra \in I$. The even integers form an ideal in $\mathbb{Z}$. The polynomials p in $\mathbb{F}[x]$ such that $p(1) = 0$ provide a second example. If I and J are subsets of R, then IJ is given by

$$IJ := \{ab : a \in I,\ b \in J\}.$$

Thus the subset I of R is an ideal if $RI \subseteq I$. The only ideal of R that contains 1 is R itself. It follows that a proper ideal cannot contain an invertible element of R. If $S \subseteq R$, then the set SR is an ideal; we call it the ideal generated by S. It consists of all R-linear combinations of the elements of S. An ideal generated by a single element is called a *principal ideal*. For example, the even integers $2\mathbb{Z}$ form a principal ideal in $\mathbb{Z}$. If I is the principal ideal generated by d, then I consists of the elements of R that are divisible by d. A *principal ideal domain* is a ring in which every ideal is principal. Both $\mathbb{Z}$ and $\mathbb{F}[x]$ are examples.

An ideal I is *prime* if it is a proper ideal and, whenever $ab \in I$, either a or b lies in I. If $m \in \mathbb{Z}$, then $m\mathbb{Z}$ is a prime ideal if and only if m is a prime. A ring is a domain if and only if the sero ideal is prime.

Suppose R is a principal ideal domain and $a, b \in R$. The ideal generated by a and b is generated by some element d, which divides both a and b. Since this ideal consists of the R-linear combinations of a and b, there are elements r and s of R such that

$$d = ra + sb.$$

It follows that if c divides a and b, then c divides d and therefore d is a greatest common divisor of a and b.

If d divides e and e divides d, we have

$$d = d_1 e, \quad e = e_1 d$$

whence $d = d_1 e_1 d$. Therefore

$$(1 - d_1 e_1)d = 0$$

and so $d_1 e_1 = 1$; hence both d_1 and e_1 are units of R. It follows that, in a principal ideal domain, any two non-zero elements have a greatest common divisor, which is unique up to multiplication by a unit.

It can be difficult to verify that a given ring is a principal ideal domain. There is one case where it is easy. We say R is a *Euclidean domain* if there is a function ρ from $R \setminus 0$ to $\mathbb{N}$ such that

(a) If $a, b \in R$ then $\rho(ab) \geq \rho(a)$.

(b) If $a, b \in R$, there are elements q and r such that $b = qa + r$ and $\rho(r) < \rho(a)$.

The advantage of Euclidean domains is that we can compute the greatest common divisor of any two elements using the usual Euclidean algorithm. Also, a Euclidean domain is a principal ideal domain.

We consider examples. If $R = \mathbb{Z}$, take $\rho(a)$ to be $|x|$. If $R = \mathbb{F}[x]$, use $\rho(p) = \deg(p)$. If $p, q \in \mathbb{F}[x]$, we say the rational function p/q is *proper* if $\deg(p) < \deg(q)$. If we define ρ by

$$\rho\left(\frac{p}{q}\right) := \deg(q) - \deg(p).$$

The set of proper rational functions over $\mathbb{F}$, with this function ρ, forms a Euclidean domain. If $\mathbb{F} = \mathbb{C}$, the proper rational functions are the rational functions that are bounded at infinity.

1. Prove that a finite domain is a field.

2. If (R, ρ) is a Euclidean domain and x is a unit in R, show that $\rho(ax) = \rho(a)$ for all a in R.

7.2 Localization

Let R be a domain. A subset S of R is *multiplicatively closed* if

(a) $0 \notin S$ and $1 \in S$,

(b) If $a, b \in S$, then $ab \in S$.

In $\mathbb{Z}$, the set of integers not divisible by a given prime is multiplicatively closed. The set of non-zero elements of R is also multiplicatively closed.

Using S, we can construct a new ring, denoted $R[S^{-1}]$. It elements are equivalence classes of ordered pairs from $R \times S$. We define (a, s) and (b, t) to be equivalent if there is an element x of R such that $b = ax$ and $t = sx$. The product of the pairs (a, s) and (b, t) is (ab, st); their sum is

$$(at + bs, st).$$

These definitions will seem more familiar if we write our pairs as ratios a/s. We then see that if $R = \mathbb{Z}$ and S consists of the non-zero integers, $R[S^{-1}] = \mathbb{Q}$. If $R = \mathbb{F}[x]$ and S consists of all powers of x, then $R[S^{-1}]$ is known as the ring of *Laurent polynomials*. It consists of the rational functions of the form $x^k p(x)$, where $p \in \mathbb{F}[x]$ and $k \in \mathbb{Z}$.

If $S = R \setminus 0$, then the ring $R[S^{-1}]$ is called the *quotient field* of R. The quotient field of $\mathbb{Z}$ is $\mathbb{Q}$, as we have just noted. The quotient field of $\mathbb{F}[x]$ is ring of rational functions in x, denoted $\mathbb{F}(x)$.

We can view $R[S^{-1}]$ as being constructed by adjoining the multiplicative inverse of each element of S to R. The element of $R[S^{-1}]$ of the form $a/1$ form a subring isomorphic to R. If $a \in R$ and $s \in S$, then $a/1$ and a/s generate the same ideal. It follows from this that ideals of $R[S^{-1}]$ correspond to the ideals I of R such that $I \cap S = \emptyset$. An important consequence is that $R[S^{-1}]$ is a principal ideal domain if R is.

We consider some examples. Let $R = \mathbb{C}[x]$ and let C be a subset of $\mathbb{C}$, for example, the unit disc. Then the polynomials $p(x)$ with no zeros in C form a multiplicatively closed subset S. The ring $R[S^{-1}]$ consists of the rational functions with no pole in the unit disc.

1. Prove that if I is an ideal of R, then $R \setminus I$ is multiplicatively closed if and only if I is prime.

2. Let S be a multiplicatively closed subset of the domain R. Prove that each ideal of $R[S^{-1}]$ consists of the elements a/s, where a comes from a given ideal I of R, and $s \in S$.

7.3 Binet-Cauchy

We prove a useful determinental identity. If C is a $k \times n$ matrix and $S \subseteq \{1, \ldots, n\}$ of size k, define $p_S(C)$ to be the determinant of the submatrix of C formed by the columns of C indexed by entries from S.

The following in the Binet-Cauchy identity. The cases $k = 1$ and $k = n$ should be familiar.[1]

7.3.1 Theorem. *If A and B are $k \times n$ matrices over a commutative ring, then*

$$\det(AB^T) = \sum_{|S|=k} p_S(A) p_S(B).$$

[1] proof by Tao, from wikipedia

Proof. Recall that

$$\det(tI + B^T A) = t^{n-k} \det(tI + AB^T).$$

On the right, the coefficient of t^{n-k} is $\det(AB^T)$. On the left, this coefficient is the sum of the principal $k \times k$ minors of $B^T A$.[2] If $|S|$ is a k-subset of $\{1, ldots, n\}$, the (S, S) minor of $B^T A$ is $p_S(B) p_S(A)$. □

If Q is $k \times k$ and A is $k \times n$ and S is a k-subset of $\{1, ldots, n\}$, we have

$$p_S(QA) = \det(Q) p_S(A)$$

The quantities $p_S(A)$ are the *Plücker coordinates* of the row space of A. You may show that A and B have the same row space if and only there is a non-zero scalar γ such that $p_S(A) = \gamma p_S(B)$ for all k-subsets S.

7.4 Fitting Invariants

Let A be an $m \times n$ matrix over a ring R. We define the *Fitting invariant* $F_k(A)$ to be the ideal generated by the $k \times k$ minors of A, where $1 \le k \le \min\{m, n\}$. If R is a principal ideal domain, then the ideal $F_k(A)$ is generated by an element f_k, and so we may use the sequence of elements $f_1, \dots, f_{m \wedge n}$, rather than the ideals $F_k(A)$.

7.4.1 Lemma. *Let A be an $m \times n$ matrix over R, where $m \le n$. Then the following are equivalent:*

(a) A has a right inverse.

(b) $F_k(A) = R$ for $k = 1, \dots, m$.

(c) $F_m(A) = R$.

Proof. First suppose that B is a right inverse for A over R. Then B is $m \times n$ and, since $AB = I_m$, by the Binet-Cauchy identity,

$$1 = \det(AB) = \sum_S p_S(A) p_S(B)$$

There the ideal generated by the $k \times k$ minors of A is R, equivalently $F_m(A) = R$.

Now suppose that $F_m(A) = R$. Let S be a set of m columns of A and set d_S equal to $\det A_S$. Let M be the $n \times m$ matrix such that $M_S = \operatorname{adj}(A_S)$ and $Me_i = 0$ if $i \notin S$. Then

$$AM = d_S I.$$

If T is a second subset of m columns of A and N is constructed analogously to M, then $AN = d_T I$ and therefore

$$A(xM + yN) = (xd_S + yd_T)I.$$

It follows that if the minors d_S generate R, then there is a right inverse for A. □

[2] using a result of Laplace

One consequence of this lemma is that if A is $m \times n$ and $F_m(A) = R$, then $F_k(A) = R$ for $k = 1, \dots, m$. We recall that two matrices A and B are equivalent over R if there are invertible matrices P and Q over R such that $B = PAQ$.

7.4.2 Lemma. *Let A and B be two $m \times n$ matrices over R. If A and B are equivalent, they have the same Fitting invariants.*

Proof. Assume Q is invertible. By Binet-Cauchy, each $r \times r$ minor of QB lies in the ideal generated by the $r \times r$ minors of B and hence $F_r(QB) \le F_r(B)$.

By the same argument, since the entries of Q^{-1} lie in R, we have

$$F_r(B) = F_r(Q^{-1}(QB)) \le F_r(QB).$$

The lemma follows. □

In Section 7.5, we will see that if R is a principal ideal domain, then two matrices of the same order are equivalent if and only if they have the same Fitting invariants. Note that A and A^T have the same Fitting invariants.

1. Let A be a $m \times n$ matrix over $\mathbb{Z}$. Show that if, for each prime p, the rank of A modulo p equals its rank over $\mathbb{Q}$, then the greatest common divisor of the $m \times m$ minors of A is 1.

7.5 Smith Normal Form

Let A and B be two $m \times n$ matrices over a commutative ring R. (Think $\mathbb{Z}[x]$.) We say that A and B are *equivalent* over R if there are invertible matrices P and Q such that $PAQ = B$. We want to decide if two given matrices are equivalent.

7.5.1 Theorem. *Let A be a matrix over a principal ideal domain R. Then there is a unique matrix D over R which is equivalent to A such that $D_{ij} = 0$ if $i \neq j$ and $D_{i,i}$ divides $D_{i+1,i+1}$ for $i = 1, \dots, n-1$.*

Proof. We first show that A is equivalent to a matrix D such that $D_{ij} = 0$ if $i \neq j$, and only then show that D can be arranged to have the form stated.

Suppose a and b are two elements of R, and suppose that the ideal they generate is generated by d. Then there must be elements s and t of R such that $sa + tb = d$. Further, there are elements a_1 and b_1 such that $a = a_1 d$ and $b = b_1 d$. Hence

$$\begin{pmatrix} s & t \\ -b_1 & a_1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} d \\ 0 \end{pmatrix}.$$

As $sa_1 + tb_1 = 1$ the determinant of

$$\begin{pmatrix} s & t \\ -b_1 & a_1 \end{pmatrix}$$

is 1 and therefore this matrix is invertible.

If the i-th of A is x and the j-th row is y and we may replace x be $sx+ty$ and y by $-b_1x+a_1y$, the resulting matrix is equivalent to A.

We may permute the columns of A so that any zero columns are last. Having done this, we may convert A to an equivalent matrix where $A_{1,1} = a \neq 0$ and $A_{i,1} = 0$ if $i > 1$. If a divides each entry of the first row of A then A is equivalent to a matrix of the form

$$\begin{pmatrix} a & 0 \\ 0 & A_1 \end{pmatrix}$$

and we can prove our claim by induction.

If a does not divide each entry in the first row, then we may operate on the columns of A, converting it to an equivalent matrix with $A_{1,1} = a'$ and $A_{1,j} = 0$ if $j > 1$. Further the ideal generated by a is properly contained in the ideal generated by a'. We hope now that a' divides each entry in the first column of A. If so then we reduce to the previous induction. If not, we operate on the rows again. Since the ideal generated by $A_{1,1}$ form a strictly increasing sequence, and since R does not contain an infinite increasing sequence of ideals, we conclude that A is equivalent to a matrix with $A_{i,1} = 0$ when $i > 1$ and $A_{1,j} = 0$ when $j > 1$. This proves our claim.

To reduce R to the required form, we observe that the two matrices

$$\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, \qquad \begin{pmatrix} a & 0 \\ sa+tb & b \end{pmatrix}$$

are equivalent; given this it is easy to see R is equivalent to a matrix satisfying the divisibility condition we gave.

The problem left is to prove that R is unique. This follows because A and D have the same Fitting ideals, and because a diagonal matrix which satisfies our divisibility condition is determined by its Fitting ideals. □

If R is a Euclidean domain, then we can use elementary row and column operations rather than the 2 × 2 matrices we described. Note that over a principal ideal domain there may be invertible 2 × 2 matrices that are not products of elementary matrices. Cohn [3] gives an example over $\mathbb{Q}(\sqrt{-19})$. (No problems over Euclidean domains.) [3]

The matrix R whose existence is guaranteed by the theorem is called the *Smith normal form* of A. If A is square then $\det(A)$ is a unit times $\det(R)$. Computing the Smith normal form, even over $\mathbb{Z}$, is one of the more difficult problems in linear algebra. If implemented as described then the number of digits in an entry can double at each step.

Generally one only meets the Smith normal form for matrices over $\mathbb{Z}$ and over $\mathbb{F}[z]$; there are a number of interesting Euclidean domains that arise in control theory, related to rational functions. Call a rational function p/q in $\mathbb{F}(z)$ *bounded* if $\deg p \leq \deg q$. If we define

$$\rho_1(p/q) = \deg q - \deg p,$$

the bounded rational functions form a Euclidean domain relative to the function ρ_1.

For a second example, let S be a subset of the complex plane, and call a polynomial *stable* if its zeros all lie in S. The set of stable polynomials is multiplicatively closed, and so the rational functions p/q where q is stable form a ring. If we define $\rho_2(p/q)$ to be the number of zeros of p not in S then this ring is a Euclidean domain relative to ρ_2.

The intersection of these two rings has the baroque denotation RH_+^∞. If, as is standard, S is the open left half-plane, this ring consists of the rational functions that are uniformly bounded on the closed right half-plane. It is a Euclidean domain relative to the function $\rho_1 + \rho_2$.

8
Polynomial and Rational Matrices

A *polynomial matrix* is a matrix whose entries come from the ring $\mathbb{F}[z]$. A *rational matrix* is a matrix whose entries come from the field of rational functions $\mathbb{F}(z)$. We will also have occasion to consider matrices whose entries are formal power series or Laurent series, but we will not assign names to these. Any matrix polynomial $A(z)$ can be written as a polynomial in z with coefficients A_i from $\mathrm{Mat}_{m\times n}(\mathbb{F})$:

$$A(z) = \sum_i A_i z^i.$$

(This encodes an isomorphism between the ring of polynomial matrices, and the ring of polynomials with matrix coefficients, but this distinction goes beyond the level of sophistication to which we aspire.) The *degree* is the maximum degree of an entry. We will also be concerned with the degrees of the rows and/or columns of polynomial matrices. The key here is to note that each column of a polynomial matrix is a polynomial matrix, and so has a well-defined degree.

We consider one pertinent example. If A is $n \times n$, then $tI - A$ is a polynomial matrix with degree one. We have

$$(zI - A)\,\mathrm{adj}(zI - A) = \det(zI - A)I.$$

Here $\mathrm{adj}(zI - A)$ is also a matrix polynomial, with degree $n - 1$, and

$$(zI - A)^{-1} = \frac{1}{\det(zI - A)}\,\mathrm{adj}(A).$$

In this chapter, we will study the basic properties of polynomial and rational matrices.

8.1 Series

A rational function is *proper* if the degree of its numerator is less than the degree of its denominator; if the degree of its numerator equals that of its denominator we say it is *bounded*[1] A rational matrix is proper if

[1] bounded = bounded at infinity; in control theory the corresponding terms are “strictly proper” and “proper”.

its elements are proper and bounded if they are bounded. The bounded rational matrices form a ring, and the strictly proper rational matrices form a proper ideal in this ring.[2]

We can view the ring of polynomials $\mathbb{F}[z]$ as a subring of the ring of formal power series $\mathbb{F}[[z]]$. This has some use, for example if $p(z)$ is a polynomial and $p(0) \neq 0$, then $p(z)$ has a multiplicative inverse in $\mathbb{F}[[z]]$. In a similar way, we can represent rational functions by formal Laurent series.

Suppose

$$p(z) = z^n + p_1 z^{n-1} + \cdots + p_n.$$

Then

$$p(z)^{-1} = z^{-n}\left(1 + \frac{p_1}{z} + \cdots + \frac{p_n}{z^n}\right)^{-1}$$

Hence $p(z)^{-1}$ has a formal power series expansion in z^{-1}, and it follows that any rational function has an expansion as a formal Laurent series in z^{-1}. If $p(z)/q(z)$ is a rational function then

$$\frac{p(z)}{q(z)} = \sum_{i=-k}^{\infty} a_i z^{-i},$$

where $k = \deg(p) - \deg(k)$. Hence the ring of rational functions in z is isomorphic to a subring of the ring of Laurent series in z^{-1}, and the image of the bounded rational functions under this isomorphism is the ring of formal power series in z. The proper rational functions map to the formal power series with contstant term equal to 0.

Since we have used nothing more than the geometric series expansion, everything goes over to matrix rational functions: these are isomorphic to a subring of the ring of Laurent series in z^{-1} with matrix coefficients, bounded rational matrices correspond to formal power series and proper rational matrices to formal power series with constant term equal to 0. From this we see, for example, that the bounded rational matrices form a ring, and the proper rational functions form an ideal in this ring. We note one other property we will need.

8.1.1 Lemma. *If $M(z)$ is a proper rational matrix, then $I + M(z)$ is invertible, and its inverse is a bounded rational matrix.*

Proof. Since $M(z)$ is strictly proper it has a series expansion

$$M(z) = \sum_{i\geq 0} M_i z^{-i}$$

Hence $I + M(z)$ is a formal power series with constant term I, and therefore it has a multiplicative inverse, which is again a formal power series with constant term I. □

8.2 Polynomial Matrices

We develop some of the basic properties of polynomial matrices.

[2] In fact they form the unique maximal ideal, the ring is *local*.

Every polynomial matrix is a rational matrix. Since

$$A(z)\,\mathrm{adj}(A(z)) = \det(A(z))I$$

we see that if $\det(A(z)) \neq 0$, then

$$\frac{1}{\det(A(z))}\,\mathrm{adj}(A(z))$$

is the inverse of $A(z)$ in the ring of rational matrices. Thus a polynomial matrix $A(z)$ has a rational inverse if and only if its determinant is not zero, although $A(z)$ may not be invertible for certain values of z in $\mathbb{F}$. A polynomial matrix has a polynomial inverse if and only if its determinant is a non-zero constant. We say that a square matrix over a ring is *unimodular* if its determinant is a unit. Since the units in $\mathbb{F}[x]$ are the non-zero constants, a polynomial matrix is unimodular if and only if it has a polynomial inverse. More generally, we recall from Section 7.4 that an $m \times n$ matrix A over a ring R has a right inverse if and only if the ideal generated by the $m \times m$ minors of A is equal to R.

Suppose $A(z)$ is an $m \times n$ polynomial matrix with linearly independent rows that is not right invertible. Then the greatest common divisor of the $m \times m$ minors of $A(z)$ is a polynomial of positive degree. It follows that there are values of z in the algebraic closure of $\mathbb{F}$ such that $\mathrm{rk}(A(z)) < m$.

8.3 Division

The *degree* of a vector over $\mathbb{F}[t]$ is the maximum degree of an entry. If A is a square matrix over $\mathbb{F}[t]$ and d_i is the degree of its i-th column, then

$$\deg(\det(A)) \leq \sum_i d_i.$$

We say that A is *column reduced* if equality holds.

If $a(z)$ and $d(z)$ are polynomials over a field, there are unique polynomials $q(z)$ and $r(z)$ such that $\deg r < \deg d$ and

$$a(z) = q(z)d(z) + r(z).$$

We establish a matrix version of this.

8.3.1 Theorem. *Suppose $D(z)$ and $N(z)$ are polynomial matrices of orders $n \times n$ and $m \times n$ respectively, and $D(z)$ is column reduced. Then $N(z)D(z)^{-1}$ is proper if and only if each column of $N(z)$ has degree less than the degree of the corresponding column of $D(z)$.*

Proof. Suppose first that $G(z) = N(z)D(z)^{-1}$ is proper. We have

$$N(z) = G(z)D(z)$$

and if $N_i(z)$ and $D_i(z)$ denote the i-th columns of $N(z)$ and $D(z)$ respectively,

$$N_i(z) = G(z)D_i(z).$$

Since $G(z)$ is proper, the degree of an element of $N_i(z)$ is less than the degree of the corresponding element of $D_i(z)$. (Note that for this part of the argument we did not need $D(z)$ to be column reduced.)

Assume now that $D(z)$ is column reduced and that the degree of each column of $N(z)$ is less than the degree of the corresponding column of $D(z)$. We may write

$$D(z) = HS(z) + L(z),$$

where H is the leading coefficient matrix of $D(z)$. Then

$$D(z)^{-1} = S(z)^{-1}H^{-1}(I + L(z)S(z)^{-1}H^{-1})^{-1}$$

Therefore

$$N(z)D(z)^{-1} = (N(z)S(z)^{-1})\,H^{-1}(I + L(z)S(z)^{-1}H^{-1})^{-1}$$

is the product of two rational matrices. The factor $N(z)S(z)^{-1}$ is proper by hypothesis. Regarding the second factor, $L(z)S(z)^{-1}H^{-1}$ is proper and so by Lemma 8.1.1, we see that $(I + L(z)S(z)^{-1}H^{-1})^{-1}$ is a proper rational matrix. It follows that $N(z)D(z)^{-1}$ is proper, as required. □

8.3.2 Theorem. *Suppose $D(z)$ and $A(z)$ are polynomial matrices and $D(z)$ is invertible and column-reduced. Then there are unique polynomial matrices $Q(z)$ and $R(z)$ such that for each i, the degree of the i-th column of P_1 is less than the degree of the i-th column of D, and*

$$A(z) = Q(z)D(z) + R(z).$$

Proof. The matrix $A(z)D(z)^{-1}$ is rational and so

$$A(z)D(z)^{-1} = Q(z) + P(z),$$

where $P(z)$ is polynomial and $R(z)$ is a proper rational matrix. Hence

$$A(z) = Q(z)D(z) + P(Z)D(z)$$

and, since $A(z)$ and $Q(z)D(Z)$ are polynomial matrices, so is $P(z)D(z)$. Let $R(z) := P(z)D(z)$. Then $R(Z)D(z)^{-1}$ is proper and so by Theorem 8.3.1 the degree of each column of $R(z)$ has degree less than the degree of the corresponding column of $R(z)$.

Now suppose

$$A(z) = Q_1(z)D(z) + R_1(z)$$

where P_1 and Q_1 are polynomial and for each i, the degree of the i-th column of P_1 is less than the degree of the i-th column of D. Then

$$(Q - Q_1)D + (R - R_1) = 0$$

and therefore

$$Q - Q_1 = (R_1 - R)D^{-1}.$$

Here the left side is a polynomial matrix, while by Theorem 8.3.1, the right side is a proper rational matrix. Therefore both sides are zero, and therefore $Q(z)$ and $R(z)$ are unique. □

Note that we do not get a version of the Euclidean algorithm, because there is no guarantee that the remainder $R(z)$ is not a zero divisor or, if not, that it is reduced. So we cannot expect to be able to divide $Q(z)$ by $R(z)$.

8.4 Cayley-Hamilton

Assume A is square. The matrix $zI-A$ is column reduced and linear so if we divide by it, the remainder must be a constant matrix. We can give an explicit formula for the remainder.

8.4.1 Lemma. *Suppose $F(z)=\sum_{i=0}^{r} F_i z^i$. Then remainder of $F(z)$ on right division by $zI-A$ is $\sum_i F_i A^i$. The remainder on left division is $\sum_i A^i F_i$.*

Proof. We can write

$$(zI-A)^{-1} = z^{-1}\sum_{i\geq 0} A^i z^{-i};$$

the coefficient of z^{-1-j} in $F(z)(zI-A)^{-1}$ is then

$$F_0A^j + F_1A^{j+1} + \cdots + F_rA^{j+r} = (F_0 + F_1A + \cdots + F_rA_r)A^j.$$

Therefore there is polynomial matrix $Q(t)$ such that

$$F(z)(zI-A)^{-1} = Q(z) + (F_0 + F_1A + \cdots + F_rA^r)(zI-A)^{-1}$$

and the remainder on right division of $F(z)$ by $(zI-A)^{-1}$ is $F_0 + F_1A + \cdots + F_rA^r$, as claimed.

We've left the left division as an exercise. □

We write $F(A)$ to denote the remainder of $F(z)$ on right division by $zI-A$.

This last result is an extension of the result that the remainder of $p(z)$ on division by $z-a$ is $p(a)$. It also implies the Cayley-Hamilton theorem. For suppose that $\phi(z)$ is the characteristic polynomial of A, and consider the remainder on left division of $\phi(z)I$ by $zI-A$. By the lemma,

$$\phi(z)I = Q(z)(zI-A) + \phi(A)$$

and since $\phi(z) = (zI-A)\operatorname{adj}(zI-A)$, we find that

$$(zI-A)(\operatorname{adj}(zI-A) - Q(z)) = \phi(A).$$

As $\operatorname{adj}(zI-A)$ is a matrix polynomial, it follows that $\operatorname{adj}(zI-A) - Q(z) = 0$ and consequently $\phi(A)$ is zero.

Let $d(z)$ denote the greatest common divisor of the entries of $\operatorname{adj}(zI-A)$ and let $C(z)$ be the matrix polynomial

$$C(z) = d(z)^{-1}\operatorname{adj}(zI-A).$$

If $p(z)$ is the polynomial $\phi(z)/d(z)$ then, since

$$\phi(z)I = \operatorname{adj}(zI-A)\,(zI-A)$$

we see that $p(A) = 0$. Let $\psi(z)$ be the minimal polynomial of A and let $\Psi(z)$ be the matrix polynomial satisfying

$$\psi(z)I = \Psi(z)(zI - A).$$

If $c(z) := p(z)/\psi(z)$, then

$$C(z)(zI - A) = p(z)I = c(z)\psi(z)I = c(z)\Psi(z)(zI - A).$$

As $zI - A$ is invertible, this implies that $C(z) = c(z)\Psi(z)$. Since the greatest common divisor of the entries if $C(z)$ is 1, it follows that $c(z) = 1$. Thus we have shown that $\psi(z) = \phi(z)/d(z)$.

8.5 Equivalence and Similarity

We derive a characterization of matrix similarity over fields.

8.5.1 Theorem. *Two $n \times n$ matrices A and B over $\mathbb{F}$ are similar if and only if $tI - A$ and $tI - B$ are equivalent over $\mathbb{F}[t]$.*

Proof. If A and B are similar, say $B = L^{-1}AB$, then

$$L^{-1}(tI - A)L = tI - B$$

and thus $tI - A$ and $tI - B$ are equivalent.

So assume that $tI - A$ and $tI - B$ are equivalent over $\mathbb{F}[t]$. Then there are invertible matrices $P(t)$ and $Q(t)$ such that

$$P(t)(tI - A) = (tI - B)Q(t).$$

There are matrices $P_0(t)$ and P_1 such that $\deg(P_0(t)) < \deg(P(t))$ and P_0 is constant and

$$P(t) = (tI - B)P_0(t) + P_1$$

Similarly we have

$$Q(t) = Q_0(t)(tI - A) + Q_1$$

It follows that

$$((tI - B)P_0(t) + P_1)(tI - A) = ((tI - B)Q_0(t) + Q_1)(tI - A)$$

and therefore

$$(tI - B)(P_0(t) - Q_0(t))(tI - A) = -P_1(tI - A) + (tI - B)Q_1.$$

Here the right side has degree at most one but if $P_0(t) \neq Q_0(t)$, the left side has degree at least two. Consequently $P_0(t) = Q_0(t)$ and so

$$tP_1 - P_1A = tQ_1 - BQ_1. \tag{8.5.1}$$

This implies that $P_1 = Q_1$ and we can complete our proof by showing that Q_1 is invertible.

We may assume that

$$Q(t)^{-1} = R_0(t)(tI - B) + R_1.$$

Now

$$\begin{aligned} I = Q(t)^{-1}Q(t) &= (R_0(t)(tI-B)+R_1)(Q_0(t)(tI-A)+Q_1) \\ &= R_0(t)(tI-B)Q_0(t)(tI-A) + R_0(t)(tI-B)Q_1 + R_1Q_0(t)(tI-A) + R_1Q_1 \end{aligned}$$

and since (from Equation (8.5.1))

$$(tI-B)Q_1 = P_1(tI-A)$$

we have

$$I = (R_0(t)(tI-B)Q_0(t) + R_0(t)P_1(tI-A) + R_1Q_0(t))(tI-A) + Q_1R_1.$$

This implies that $I = Q_1R_1$ and therefore Q_1 is invertible. □

It follows that we can decide if two matrices A and B are similar by computing the Smith normal forms of $tI - A$ and $tI - B$.

8.6 *An Identity*

We will need the following result.

8.6.1 Lemma. *Let $C = (c_{i,j})$ be a square matrix. Then*

$$\frac{\partial}{\partial c_{i,j}}(zI-C)^{-1} = (zI-C)^{-1}e_ie_j^T(zI-C)^{-1}.$$

Proof. This is an easy consequence of the following identity, which itself is easily verified.

$$(zI-C)^{-1} - (zI-D)^{-1} = (zI-C)^{-1}(C-D)(zI-D)^{-1}. \qquad \square$$

The matrix Ψ in the proof of the next result is defined in **??**.

8.6.2 Theorem. *Let ψ be a polynomial of degree d, let C_ψ be its companion matrix and let $E_\psi(z)$ denote the $d \times d$ matrix with ij-entry equal to $\psi_i(z)z^{j-1}/\psi(z)$. Let N be the companion matrix of z^d. Then*

$$(zI-C_\psi)^{-1} - E_\psi(z) = \left(N + zN^2 + \cdots + z^{d-2}N^{d-1}\right)^T.$$

Proof. The right side of this identity is independent of ψ (apart from its degree). By **??**,

$$\psi(z)^{-1}\Psi(z) = (zI-C_\psi)^{-1}e_1$$

and therefore

$$E_\psi(z) = (zI-C_\psi)^{-1}e_1\begin{pmatrix}1 & z & \cdots & z^{d-1}\end{pmatrix}.$$

Our strategy is to show that $(zI - C_\psi)^{-1} - E_\psi(z)$ is independent of ψ, and then evaluate it when $\psi = z^d$.

Assume

$$\psi(z) = t^d + a_1 t^{d-1} + \cdots + a_d;$$

then $C_{i,d} = -a_i$ and, by the previous lemma

$$\frac{\partial}{\partial a_i}(zI - C_\psi)^{-1} = -(zI - C_\psi)^{-1} e_i e_d^T (zI - C_\psi)^{-1}.$$

From **??** we have

$$e_d^T (zI - C_\psi)^{-1} = \psi(z)^{-1} e_i \begin{pmatrix} 1 & z & \cdots & z^{d-1} \end{pmatrix}$$

and therefore

$$\frac{\partial}{\partial a_i}(zI - C_\psi)^{-1} = -\psi(z)^{-1}(zI - C_\psi)^{-1} e_i \begin{pmatrix} 1 & z & \cdots & z^{d-1} \end{pmatrix}.$$

We have

$$\frac{\partial}{\partial a_i} E_\psi(z) = \frac{\partial}{\partial a_i}(zI - C_\psi)^{-1} e_1 \begin{pmatrix} 1 & z & \cdots & z^{d-1} \end{pmatrix}.$$

Since

$$e_i \begin{pmatrix} 1 & z & \cdots & z^{d-1} \end{pmatrix} e_1 \begin{pmatrix} 1 & z & \cdots & z^{d-1} \end{pmatrix} = e_i \begin{pmatrix} 1 & z & \cdots & z^{d-1} \end{pmatrix},$$

it follows that

$$\frac{\partial}{\partial a_i}(zI - C_\psi)^{-1} = \frac{\partial}{\partial a_i} E_\psi(z).$$

We conclude that $(zI - C_\psi)^{-1} - E_\psi(z)$ is independent of ψ.

Now suppose $\psi(z) = z^d$ and $N = C_\psi$. Then $N^d = 0$,

$$(zI - N)^{-1} = z^{-1}(I - z^{-1}N)^{-1} = \sum_{i=0}^{d-1} z^{-i} N^i,$$

and

$$\big(E_\psi(z)\big)_{i,j} = z^{-i+j-1}.$$

The theorem follows at once. □

(1) By setting $z = 0$ in Theorem 8.6.2, deduce the expression for the inverse of an invertible companion matrix in Theorem 4.4.1.

8.7 Resolvents

Let A be an $n \times n$ matrix. In this section we generally work over any field that contains all the eigenvalues of A. The *resolvent* $R(z)$ of A is the matrix $(zI - A)^{-1}$. As

$$R(z) = \frac{1}{\det(zI - A)} \operatorname{adj}(zI - A),$$

each entry of $R(z)$ is a rational function. Let θ be an eigenvalue of A with multiplicity m. Then there are matrices A_i such that

$$R(z) = \sum_{r=-m}^{\infty} A_r(z-\theta)^r;$$

we wish to determine these matrices.

The key to this is the following identity.

8.7.1 Theorem. *If $R(z)$ is the resolvent of some matrix then*

$$R(z) - R(w) = -(z-w)R(z)R(w).$$

Proof. Let $R(z)$ be the resolvent of A. Then

$$(zI - A)(R(z) - R(w))(wI - A) = (wI - A) - (zI - A) = (w - z)I,$$

whence the result follows. □

We note a simple consequence of this.

8.7.2 Lemma. *If A is symmetric, then all poles of the entries of $R(z)$ are simple.*

Proof. Suppose that θ is an eigenvalue of A and

$$R(z) = \sum_{r\ge -m} A_r(z-\theta)^r.$$

Here $m \ge 1$ and we may assume without loss that $A_{-m} \ne 0$. Then $R(z)^T R(z)$ is equal to $A_{-m}^T A_{-m}(z-\theta)^{-2m}$, plus terms of higher order and, as A is symmetric, $R(z)^T R(z) = R(z)^2$. On the other hand, from Theorem 8.7.1 we have that

$$\frac{d}{dz}R(z) = -R(z)^2.$$

The term of least order in $R(z)'$ is $-mA_{-m}(z-\theta)^{-m-1}$; consequently we must have $m+1 = 2m$, i.e., $m = 1$. □

8.7.3 Lemma. *Suppose that $R(z)$ is the resolvent of A and that θ is an eigenvalue of A with multiplicity m. If $R(z) = \sum_{-m}^{\infty} A_r(z-\theta)^r$ then*

$$A_rA_s = \begin{cases} -A_{r+s+1}, & r,s \ge 0; \\ A_{r+s+1}, & r+s \ge -m-1,\ r,s \le -1; \\ 0, & \text{otherwise.} \end{cases}$$

Proof. We assume that 0 is an eigenvalue of A, and seek to determine the coefficients A_r in the expansion $R(z) = \sum_{r\ge -m} A_r z^r$. From Theorem 8.7.1 we have

$$-\sum_{r,s\ge -m} A_rA_s z^r w^s = -R(z)R(w) = \frac{R(z)-R(w)}{z-w} = \sum_{r\ge -m} A_r \frac{z^r - w^r}{z-w}.$$

The lemma follows for $\theta = 0$ by comparing coefficients of $z^i w^j$ in the two series above, and the general result is an easy consequence of this. □

From this result we see that the matrices A_i, $i = -m, -m+1, \dots$ commute. We also find that:

$$\begin{aligned} A_r &= (-1)^r A_0^{r+1}, \qquad \text{if } r \geq 0, \\ A_{-r} &= (A_{-2})^{r-1}, \qquad \text{if } r \geq 2, \\ A_{-1}A_{-r} &= A_{-r}, \qquad \text{if } r \geq 0. \end{aligned}$$

Therefore the coefficients in our Laurent series for $R(z)$ are determined by A_0, A_{-1} and A_{-2}, where $(A_{-1})^2 = A_{-1}$ and $(A_2)^m = 0$. Thus A_{-1} is idempotent and A_{-2} is nilpotent, let us denote them respectively by E_θ and N_θ. Now note that

$$(tI - A)R(z) = ((t-z)I + zI - A)R(z) = (t-z)R(z) + I;$$

Putting $t = \theta$ in this yields

$$\begin{aligned} (\theta I - A)A_r &= A_{r-1}, \quad r \neq 0, \\ (\theta I - A)A_0 &= A_{-1} - I. \\ (\theta I - A)A_{-m} &= 0. \end{aligned}$$

Hence

$$N_\theta = (\theta I - A)E_\theta.$$

Define the *principal part* $P_\theta(z)$ of $R(z)$ by

$$P_\theta(z) := \sum_{r=1}^{m} A_{-r}(z-\theta)^{-r}.$$

Thus

$$\begin{aligned} P_\theta(z) &= (z-\theta)^{-1}E_\theta + \sum_{r=1}^{m-1} N_\theta^r (z-\theta)^{-r} \\ &= (z-\theta)^{-1} \sum_{r=0}^{m-1} (\theta I - A)^r E_\theta (z-\theta)^{-r} \end{aligned}$$

We note that, if θ and τ are distinct eigenvalues of A, then $P_\theta P_\tau = 0$ and so $E_\theta E_\tau = 0$. We have the following result, which provides a partial fraction decomposition of the resolvent.

8.7.4 Theorem. *Let $R(z)$ be the resolvent of A and let $P_\theta(z)$ be the principal part of $R(z)$ at θ. Then $R(z) = \sum_\theta P_\theta(z)$.*

Proof. A rational function in z is called *proper* if the degree of its numerator is less than the degree of its denominator. A proper rational function with no poles is constant. The set of proper rational functions is a vector space.

We note that the entries of $R(z)$ and the entries of $P_\theta(z)$ are proper rational functions. Hence each entry of the difference

$$R(z) - \sum_\theta P_\theta(z);$$

is a proper rational function. By the construction of $P_\theta(z)$, these rational functions have no poles. As both $R(z)$ and $P_\theta(z)$ converge to zero as $z \to \infty$, our theorem follows. □

We know that, if m is the multiplicity of θ as an eigenvalue of A then $A_r = 0$ when $r < -m$, equivalently $(\theta I - A)^m E_\theta = 0$. This implies that the order of the pole of $R(z)$ at θ is at most m.

8.7.5 Theorem. *The order of the pole of $R(z)$ at θ is equal to the multiplicity of θ as a zero of the minimal polynomial of A.*

Proof. Let $\psi(z)$ denote the minimal polynomial of A, let $\nu(\theta)$ be the multiplicity of θ as a zero of $\psi(z)$ and suppose

$$\psi_\theta(z) = \frac{\psi(z)}{(z-\theta)^{\nu(\theta)}}.$$

Let $\mathscr{A}_\theta$ denote the space spanned by the matrices $(\theta I - A)^i E_\theta$ and let $d(\theta)$ be its dimension. Thus $d(\theta)$ is the greatest integer such that $(\theta I - A)^{d(\theta)-1} E_\theta \neq 0$.

As $(\theta I - A)^{d(\theta)} P_\theta(z) = 0$, it follows that

$$\prod_\theta (\theta I - A)^{d(\theta)} R(z) = 0.$$

Since $R(z)$ is invertible, this implies that

$$\prod_\theta (\theta I - A)^{d(\theta)} = 0.$$

From the definition of the minimal polynomial we then deduce that $\nu(\theta) \le d(\theta)$, for all eigenvalues θ of A. We show next that $\nu(\theta) = d(\theta)$.

The matrices $(\theta I - A)^i E_\theta$ for $i = 0, 1, \dots, d(\theta) - 1$ form a basis for $\mathscr{A}_\theta$. As $\psi_\theta(\theta) \neq 0$, it follows that the matrix representing the action of $\psi_\theta(A)$ relative to this basis is triangular, with non-zero diagonal entries. In particular, it is invertible. On the other hand, if $M \in \mathscr{A}_\theta$, then

$$0 = (\theta I - A)^{\nu(\theta)} \psi_\theta(A) M = \psi_\theta(A)(\theta I - A)^{\nu(\theta)} M,$$

and this implies that $(\theta I - A)^\nu(\theta)$ acts as the zero operator on $\mathscr{A}_\theta$. It follows that $\nu(\theta) \ge d(\theta)$. □

8.7.6 Corollary. *For each eigenvalue θ, the matrix E_θ is a polynomial in A.*

Proof. Since $zR(z) \to I$ and $zP_\theta(z) \to E_\theta$ as $z \to \infty$, Theorem 8.7.4 implies that

$$I = \sum_\theta E_\theta. \tag{8.7.1}$$

It follows from the proof of Theorem 8.7.5 that $\psi_\theta(A) E_\tau = 0$ if $\tau \neq \theta$, whence (8.7.1) yields that

$$(\theta I - A)^i \psi_\theta(A) = (\theta I - A)^i \psi_\theta(A) E_\theta.$$

Referring to the proof of Theorem 8.7.5 again, we see that $\psi_\theta(A)E_\theta$ lies in $\mathscr{A}_\theta$. It is not hard to show that the matrices $(\theta I - A)^i \psi_\theta(A)E_\theta$ for $i = 0, 1, \dots, \nu(\theta)$ form a basis for $\mathscr{A}_\theta$, and accordingly each matrix in $\mathscr{A}_\theta$ must be a polynomial in A. □

8.7.7 Corollary. *Any square matrix A is the sum of a diagonalizable and a nilpotent matrix, each of which is a polynomial in A.*

Proof. As $E_\theta E_\tau = 0$ when $\theta \neq \tau$ and $E_\theta^2 = E_\theta$, the column space of E_θ is an eigenspace for all the idempotents E_τ. Given this, (8.7.1) implies that $\mathbb{F}^n$ is the direct sum of eigenspaces of E_θ. Hence E_θ is diagonalizable; more generally any linear combination of the matrices E_θ is diagonalizable. It is also a polynomial in A.

As $AE_\theta = E_\theta + N_\theta$, it also follows from (8.7.1) that

$$A = \sum_\theta (\theta E_\theta + N_\theta) = \sum_\theta \theta E_\theta + \sum_\theta N_\theta.$$

Since $N_\theta N_\tau = 0$ when $\theta \neq \tau$, it follows that $\sum_\theta N_\theta$ is nilpotent. Since $N_\theta = (\theta I - A)E_\theta$, we see that N_θ is a polynomial in A and, therefore, $\sum_\theta N_\theta$ is too. □

The last result implies that symmetric matrices are diagonalizable—if A is symmetric, so is any polynomial in A, but the only symmetric nilpotent matrix is the zero matrix. It is slightly more difficult to see that the only normal nilpotent matrix is the zero matrix; from this it follows that normal matrices are diagonalizable.

8.7.8 Corollary. *Let $\varphi(z)$ be the characteristic polynomial of A and let $g(z)$ be the greatest common divisor of the determinants of the $(n-1)\times(n-1)$ submatrices of $zI - A$. Then $\varphi(z)/g(z)$ is the minimal polynomial of A.*

Proof. Let θ be an eigenvalue of A, with multiplicity m, and let ν be its multiplicity as a zero of $\psi(z)$. Let $f_{i,j}(z)$ be the ij-minor $zI - A$. It follows from Theorem 8.7.5 that no entry of $R(z)$ has a pole of order greater than $\nu(\theta)$ at θ, and that some entry has a pole of this order at θ. In other words $(z-\theta)^{m-\nu}$ divides each polynomial $f_{i,j}(z)$, and divides one of these polynomials exactly. The result follows immediately. □

8.8 Paraunitary Matrices

Suppose

$$A(z) = \sum_{r=0}^{m} A_r z^r,$$

where $A_r \in \mathrm{Mat}_{n\times n}(\mathbb{C})$ and let $A^*(z^{-1})$ be given by

$$A^*(z) = \sum_{r=0}^{m} A_r^* z^{-r}.$$

We say that $A(z)$ is paraunitary if

$$A(z)A^*(z^{-1}) = I.$$

One consequence of this definition is that if $A(z)$ is paraunitary, then $A(z)$ is unitary when $\|z\| = 1$. The product of paraunitary matrices is paraunitary.

By way of example if $v \in \mathbb{C}^n$ and $\|v\| = 1$, then an easy computation shows that

$$V(z) := I - vv^* + zvv^*$$

is paraunitary. We also see that $V(1) = I$ and, with some effort, that

$$\det V(z) = 1.$$

Paraunitary matrices of this type are called *primitive*. Note that $V^*(z) = V(z)$, so $V(z)V(z^{-1}) = I$.

8.8.1 Lemma. *3.1 If $A(z)$ is paraunitary, then* $\det A(z) = z^m$ *for some non-negative integer m.*

Proof. Suppose $p(z) := \det A(z)$, and let $\overline{p}(z)$ be the polynomial whose coefficients are the complex conjugates of those of $p(z)$. Then $\overline{p}(z^{-1}) = \det A^*(z^{-1})$ and since $A(z)A^*(z^{-1}) = I$, we have

$$p(z)\overline{p}(z^{-1}) = \det(A(z))\det(A^*(z^{-1})) = 1.$$

Suppose $p(z)$ has degree d and that z^e is the highest power of z that divides $p(z)$ Then

$$p(z)\overline{p}(z^{-1}) = \frac{p(z)q(z)}{z^d},$$

where $q(z)$ is a polynomial of degree $d-e$. Therefore $p(z)q(z)$ has degree $d-e$, and the lemma follows at once. □

8.8.2 Lemma. *Let $A(z)$ be a paraunitary matrix. Then $A(z)$ is constant if and only if* $\det A(z) = 1$.

Proof. Suppose

$$A(z) = \sum_{r=0}^{m} z^r A_r$$

and $A_m \neq 0$. If $m = 0$ then $A(z)$ is constant and $\det A(z) = 1$.

If $m > 0$, then the coefficient of z^{-m} in the product $A(z)A^*(z^{-1})$ is $A_0 A_m^*$, whence $A_0 A_m^* = 0$ and A_0 is singular. Then

$$\det A(z)\det(A_0 + zB(z)),$$

where $B(z)$ is a polynomial matrix. Therefore the constant term of $\det A(z)$ is $\det A(0)$, which is zero. We conclude that $\det A(z)$ is a positive power of z. □

8.8.3 Theorem. *If $A(z)$ is a paraunitary matrix and* $\det A(z) = z^d$*, then $A(z) = A(1)W(z)$, where $W(z)$ is the product of d primitive paraunitary matrices.*

Proof. If $d = 0$ then $A(z) = A(1)$ and there is nothing to prove, so we assume $d > 0$. As in the proof of the previous lemma, it follows that the constant term A_0 in $A(z)$ is singular and therefore there is a unit vector v such that $v^* A_0 = 0$.

Suppose $A(z) = -\sum_{r=0}^{m} A_r z^r$ and

$$V(z) := I - vv^* + zvv^*$$

and consider the product

$$B(z) = V(z^{-1})A(z) = (I - vv^* + z^{-1}vv^*)(A_0 + A_1 z + \cdots + A_m z^m).$$

Since $v^* A_0 = 0$, we see that $B(z)$ is a polynomial matrix and hence that is is paraunitary.

Since $V(z)B(z) = A(z)$ we have

$$z \det B(z) = z^d$$

and consequently $\det B(z) = z^{d-1}$. The theorem follows now by induction on the degree of $\det A(z)$. □

Paraunitary matrices play a significant role in the theory of filter banks and in some treatments of wavelets. For the latter, see Resnikoff and Wells "Wavelet Analysis" (Springer, New York) 1998.

Part II

Eigenthings

9
Orthogonality

We study inner product spaces over $\mathbb{R}$ and $\mathbb{C}$.

9.1 Properties of Projections

Let U be subspace of the inner product space V. Then orthogonal projection onto U is a function P from V to itself such that, for all v in V, we have $v - P(v) \in U^{\perp}$. We establish a number of properties of P, the most important of which is that it is linear.

9.1.1 Lemma. *Let P be the orthogonal projection of V onto U. Then P is linear mapping and*

(a) $\operatorname{im}(P) = U$.

(b) $\ker(P) = U^{\perp}$.

(c) $P^2 = P$.

(d) *If $v, w \in V$, then $\langle v, Pw\rangle = \langle Pv, w\rangle$.*

Proof. Suppose $v, w \in V$. Then $v - P(v)$ and $w - P(w)$ both belong to $U^{\perp}$, whence

$$(v + w) - (P(v) + P(w)) = (v - P(v) + (w - P(w)) \in U^{\perp}.$$

Since $P(v) + P(w) \in U$, this implies that $P(v) + P(w)$ is the orthogonal projection of $v + w$ onto U. Therefore P is linear.

Since $Pv \in U$ for all v in V, we see that $\operatorname{im}(P) \subseteq U$, and since $Pu = u$ for all u in U, we see that $\operatorname{im}(P) = U$ and $P^2 = P$. This proves (a) and (c).

For (b), we note that if $P(v) = 0$ then $v \in U^{\perp}$. On the other hand, $v - P(v) \in U^{\perp}$ and so if $v \in U^{\perp}$ then $P(v) \in U^{\perp}$. Since $P(v) \in U$, this implies that $P(v) = 0$.

Finally, for any vectors v and w we have

$$\langle v - Pv, Pw\rangle = 0, \qquad \langle Pv, Pw - w\rangle = 0.$$

Summing these two expressions yields

$$0 = \langle v, Pw\rangle - \langle P, v\rangle Pw + \langle Pv, Pw\rangle - \langle Pv, w\rangle,$$

whence (d) follows. □

Linear mappings P such that $P^2 = P$ are called *idempotent*. If $\langle v, Pw\rangle = \langle Pv, w\rangle$ for all v and w, we say P is *self adjoint* with respect to the inner product.

9.2 Matrices Representing Projections

If we are working in Euclidean space—$\mathbb{R}^n$ with dot product—then we can give an explicit formula for the matrix representing orthogonal projection.

9.2.1 Lemma. *Let V be $\mathbb{R}^n$ equipped with dot product, and let U be a subspace of V with dimension k. If B is an $n \times k$ matrix whose columns form a basis for U, the matrix representing orthogonal projection on U is*

$$B(B^TB)^{-1}B^T.$$

Proof. We offer two proofs. The first is a simple verification that the quoted formula is correct. First you may easily verify that $B(B^TB)^{-1}B^T$ is symmetric. Then we compute that

$$(I - B(B^TB)^{-1}B)B = B - B(B^TB)^{-1}B^TB = B - B = 0$$

and therefore

$$\begin{aligned}(v - B(B^TB)^{-1}B^Tv)^TB &= v^T(I - B(B^TB)^{-1}B)^TB\\ &= v^T(I - B(B^TB)^{-1}B)B\\ &= 0.\end{aligned}$$

So, if $u := B(B^TB)^{-1}B^Tv$, then $v - u$ is orthogonal to each column of B. Hence it lies in $U^\perp$, and therefore u is the orthogonal projection of v onto U.

A difficulty with the previous argument is that it gives no indication how we found the matrix $B(B^TB)^{-1}B^T$ in the first place. We outline the reasoning. Suppose Q is the matrix representing orthogonal projection on U. Then $\operatorname{rk} Q = k$ and by Theorem **??** we can write

$$Q = AB^T,$$

where A and B are $n \times k$ matrices with rank k. Since $\operatorname{col}(A) = \operatorname{col}(Q) = U$, the columns of A form a basis for U. Since the columns of A are linearly independent, if $Ax = 0$ then $x = 0$. Therefore $\ker B^T = \ker P = U^\perp$ and consequently $\operatorname{col}(B) \subseteq U^{\perp\perp} = U$. As $\operatorname{rk} B = k$, this shows that $\operatorname{col} B = U$. Since each column of A lies in U, we have

$$A = QA = AB^TA$$

and therefore $B^T A = I$. On the other hand, the columns of B form a basis for U, so each column of A is a linear combination of columns of B, and therefore there is a $k \times k$ matrix M such that $A = BM$. If $B^T A = I$, this implies that $B^T BM = I$ and so $M = (B^T B)^{-1}$. Accordingly

$$Q = B(B^T B)^{-1} B^T. \qquad \square$$

(It might be a useful exercise to identify where in this argument we have used that our inner product is the dot product.)

To sum up, we have two ways to compute the orthogonal projection of a vector v onto a subspace U. If we are given an orthogonal basis for U, we can use (**??**). If we are working in $\mathbb{R}^n$ with dot product and given a basis for U, we could construct $Q = B(B^T B)^{-1} B^T$, in which case the answer is Qv.

(1) If $\langle\cdot,\cdot\rangle$ is the dot product, show that (**??**) implies that $P = P^T$.

(2) Suppose B and C are $n \times k$ matrices with rank k and the same column space. Prove that $B(B^T B)^{-1} B = C(C^T C)^{-1} C$.

(3) Let $u_1, \dots, u_k$ be an orthogonal basis for the subspace U. Show that the matrix representing orthogonal projection on U is equal to

$$\sum_{i=1}^{k} \langle u_i, u_i\rangle^{-1} u_i u_i^T.$$

9.3 *Least Squares*

We consider a version of the least squares problem. Let W be an $m \times n$ matrix where $m < n$, and $\text{rk}(W) = m$. Then the system of equations

$$Wx = v \tag{9.3.1}$$

will have infinitely many solutions, but some may suit us better than others. For example, in the control theory setting of Chapter 14, a solution x to an equation of the form $Wx = v$ represented a sequence of inputs that would drive our system to a chosen state. In this case, $x^T x$ corresponds to the power that this sequence would require, and it would be very natural to seek to minimize it. Thus we want to find the solution to (9.3.1) with minimum squared length.

Suppose that x is any solution to (9.3.1), and let $\bar{x}$ be the orthogonal projection of x on $\text{col}(W^T)$. Then $x - \bar{x}$ is orthogonal to $\text{col}(W^T)$, and therefore $W(x - \bar{x}) = 0$. Hence

$$W\bar{x} = Wx = v.$$

If y is another solution to (9.3.1), then $Wy = W\bar{x}$ and so $y - \bar{x}$ is in the null space of W. Consequently $y - \bar{x}$ is orthogonal to $\bar{x}$ and

$$\|y\|^2 = \|y - \bar{x}\|^2 + \|\bar{x}\|^2.$$

Thus $\bar{x}$ is the solution to (9.3.1) with minimum norm.

How can we compute $\bar{x}$? If we can assume that the rows of W are linearly independent then, by Lemma 9.2.1, the matrix representing orthogonal projection onto col W^T is

$$Q := W^T(WW^T)^{-1}W$$

and our solution is Qx. However we do not need to find x; we have

$$Qx = W^T(WW^T)^{-1}Wx = W^T(WW^T)^{-1}v,$$

and we can proceed as follows: given v, solve the system

$$WW^T z = v,$$

the desired solution is $W^T z$. (This approach avoids the need to compute the inverse of WW^T. Computing an inverse explicitly is rarely worth the trouble. It may also pay to avoid computing WW^T, but we digress....)

In **??**, we will develop a general method for least squares problems, which does not require that the rows of W are linearly independent.

9.4 *Orthogonal Polynomials*

Let V be the space of all real polynomials, or the vector space of polynomials with degree at most n. Assume V is equipped with an inner product such that

$$\langle p, xq\rangle = \langle xp, q\rangle,$$

and, if $p(x)$ is non-negative and not zero, then $\langle 1, p\rangle > 0$. All our examples have these properties.

If we apply Gram-Schmidt to the basis of V formed by the powers of x, we obtain a sequence of polynomials $(p_r)_{r\ge0}$, where p_r has degree r. A *sequence of orthogonal polynomials* is an orthogonal set of polynomials $(p_r)_{r\ge0}$, where p_r has degree r (and $p_0 \ne 0$). If we multiply each member of a sequence of orthogonal polynomials by a non-zero scalar, the result is still a sequence of orthogonal polynomials.

9.4.1 Lemma. *The sequence of polynomials $(p_r)_{r\ge0}$ is an orthogonal basis if and only if p_r is non-zero and is orthogonal to all polynomials of degree less than r.* □

9.4.2 Lemma. *Let $(p_r)_{r\ge0}$ be a sequence of orthogonal polynomials. If $p_r(x) = a(x)b(x)$, where a and b are polynomials and $b(x) \ge 0$ for all x, then b is constant.*

Proof. We have

$$\langle p_r, a\rangle = \langle ab, a\rangle = \langle 1, a^2b\rangle.$$

Now a^2b is non-zero and non-negative and therefore $\langle 1, a^2b\rangle > 0$. But, if the degree of b is positive, then the degree of a is less than r and, by the previous lemma, $\langle p_r, a\rangle = 0$. We conclude that b must be constant. □

9.4.3 Theorem. *If p is a member of a sequence of orthogonal polynomials, its zeros are real and simple.*

Proof. Suppose θ is a complex zero of p. Then its complex conjugate $\bar{\theta}$ is also a zero of p and therefore the real quadratic polynomial

$$(x-\theta)(x-\bar{\theta})$$

divides p. Since this quadratic has two complex roots and is monic, it is non-negative. By the previous lemma, it cannot divide p. This proves the first claim.

For the second, note that $(x-\theta)^2$ is non-negative and the same technique yields that this cannot divide p. □

(1) Let $\langle p,q\rangle := \int_0^1 p(x)q(x)\,dx$. Show that if p is a member of the sequence of orthogonal polynomials associated to this inner product, all zeros of p lie in the interval $(0,1)$.

(2) Suppose p_r and p_{r+1} are consecutive members of a sequence of orthogonal polynomials. Show that they cannot have a common zero.

9.5 The Three-Term Recurrence

We provide an easier way to construct families of orthogonal polynomials. The key is to note that

$$\langle xp_r, p_j\rangle = 0$$

if $j \notin \{r-1, r, r+1\}$. For if $j < r-1$ then xp_j has degree less than r, and therefore

$$\langle xp_r, p_j\rangle = \langle xp_r, p_j\rangle = 0.$$

If $j > r+1$ then similarly p_j is orthogonal to xp_r.

9.5.1 Theorem. *Let $(p_r)_{r\geq 0}$ be a sequence of monic orthogonal polynomials. Then*

$$p_{n+1} = (x-a_n)p_n - b_n p_{n-1},$$

where $a_n = \langle xp_n, p_n\rangle / \langle p_n, p_n\rangle$ and $b_n = \langle p_n, p_n\rangle / \langle p_{n-1}, p_{n-1}\rangle$.

Proof. From our remarks just above, xp_n is a linear combination of p_{n-1}, p_n and p_{n+1}. Thus we may write

$$xp_n = \gamma p_{n+1} + \alpha p_n + \beta p_{n-1}.$$

Here

$$\gamma = \frac{\langle xp_n, p_{n+1}\rangle}{\langle p_{n+1}, p_{n+1}\rangle}.$$

Since p_{n+1} is monic, $xp_n = p_{n+1} - q$, where q has degree less than n. So

$$\langle xp_n, p_{n+1}\rangle = \langle p_{n+1}, p_{n+1}\rangle$$

and therefore $\gamma = 1$.

Next we see that $\alpha = \langle xp_n, p_n\rangle / \langle p_n, p_n\rangle$ and

$$\beta = \frac{\langle xp_n, p_{n-1}\rangle}{\langle p_{n-1}, p_{n-1}\rangle}.$$

Arguing as before,

$$\langle xp_n, p_{n-1}\rangle = \langle p_n, p_n\rangle$$

and this leads to the stated expression for b_n. □

One consequence of the formulas for the coefficients in this recurrence is that $b_n > 0$ for all n.

There is another way of stating the last result. Let $(p_r)_{r\ge 0}$ be a monic sequence of orthogonal polynomials. Let M_x denote the linear transformation that maps a polynomial p to xp. Then the matrix representing M_x with respect to the basis $(p_r)_{r\ge 0}$ is

$$\begin{pmatrix} a_0 & b_1 & & & \\ 1 & a_1 & b_2 & & \\ & 1 & a_2 & b_3 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}$$

This is an example of a *tridiagonal matrix*.

9.6 Numerical Integration

We want to compute definite integrals of the form

$$\int_a^b f(t)\, w(t) dt.$$

Here $w(t)$ is a *weight function*. For example if the interval of integration is $[0,\infty]$, then we may use $w(t) = e^t$. But for now we take $w(t)$ to be identically 1, and the interval of integration will be $[0,1]$. So all we want is

$$\int_0^1 f(t)\, dt.$$

The problem is that we do not know the anti-derivative of f, and so we seek a procedure that will produce a reasonably accurate answer in reasonable time.

There are many possible notions of what ‘reasonably accurate’ might mean. Before we discuss this, we specify the sort of procedure we want in more detail. The first thing is to note that the map

$$\mathscr{S} : f \mapsto \int_0^1 f(t)\, dt.$$

is a linear map from the space $C[0,1]$ of continuous functions on $[0,1]$ to $\mathbb{R}$. Hence it is an element of the dual space $C[0,1]^*$. This has the property that if $f \geq 0$, then $\mathscr{S}(f) \geq 0$—it maps non-negative functions to non-negative real numbers. In this context, the elements of $C[0,1]^*$ are known as *linear functionals*, and we say a linear functional is *non-negative* if it maps non-negative functions to non-negative numbers.

There are many other non-negative linear functionals, and amongst the simplest are the *evaluation maps* e_a, for $a \in \mathbb{R}$, given by

$$e_a(f) := f(a).$$

Our aim is to choose an increasing sequence of *nodes* $\theta_1, \dots, \theta_n$ and a sequence of *weights* $w_1, \dots, w_n$, such that the linear functional

$$\mathscr{Q} := \sum_i w_i e_{\theta_i}$$

is a good approximation to $\mathscr{S}$. We call a linear functional of this form a *quadrature scheme*. Define the *degree of precision* of $\mathscr{Q}$ to be the greatest integer k such that

$$\mathscr{Q}(p) = \int_0^1 p(t)\,dt$$

for all polynomials p with degree at most k.

By way of example, if $\mathscr{Q}$ has degree of precision 1, then

$$\mathscr{Q}(1) = 1 \qquad \mathscr{Q}(t) = \frac{1}{2}.$$

These hold if and only if

$$\sum_i w_i = 1, \qquad \sum_i w_i \theta_i = \frac{1}{2}.$$

It is easy to find nodes and weights for which these conditions hold.

We will be more greedy. Suppose we are given nodes $a_1, \dots, a_n$, and that we try to find weights to go with them. Let $p_1, \dots, p_n$ be the Lagrange interpolating polynomials at the given nodes. Thus

$$p_i(a_j) = \delta_{i,j}.$$

Then

$$\mathscr{Q}(p_i) = \sum_j w_j p_i(a_j) = w_i$$

and, if we want degree of precision at least $n-1$, we will need

$$w_i = \int_0^1 p_i(t)\,dt.$$

There is one problem: there are good reasons to require that the weights w_i be non-negative, and it is not clear how to choose the nodes to ensure this in general.

We can go further if we use orthogonal polynomials. Define an inner product on $C[0,1]$ by

$$\langle p,q\rangle = \int_0^1 p(t)q(t)\,dt.$$

Let $p_0,\dots,p_n$ be the first $n+1$ members of the corresponding family of orthogonal polynomials, and let $\theta_1,\dots,\theta_n$ be the zeroes of p_n in increasing order. (We know by Theorem 9.4.3 that these zeroes are real and distinct.) Using the Lagrange interpolating polynomials, we compute the weights w_i for a quadrature scheme with degree of precision at least $n-1$. Then, as Gauss first noted, a miracle occurs: the degree of precision of our scheme is $2n-1$.

We verify this. Suppose f is a polynomial with degree at most $2n-1$. By the Euclidean algorithm, there are polynomials q and r, both with degree at most $n-1$, such that

$$f(t) = q(t)p_n(t) + r(t).$$

Now

$$\int_0^1 f(t)\,dt = \langle 1, f(t)\rangle = \langle 1, q(t)p_n(t)\rangle + \langle 1, r(t)\rangle.$$

Since q has degree less than n,

$$\langle 1, q(t)p_n(t)\rangle = \langle q(t), p_n(t)\rangle = 0,$$

and therefore

$$\int_0^1 f(t)\,dt = \int_0^1 r(t)\,dt.$$

Because the degree of $r(t)$ is at most $n-1$, the integral on the right can be computed exactly using our (well, Gauss's) quadrature scheme. Hence this scheme has degree of precision at least $2n-1$. However it is exactly $2n-1$, because

$$\mathscr{Q}(p_n(t)^2) = 0 < \int_0^1 p_n(t)^2\,dt.$$

10
Eigenvectors and Eigenvalues

In this chapter we undertake a study of questions related to existence of eigenvectors and eigenvalues. Our focus is on self-adjoint operators, because that is where eigenvalues are most useful.

10.1 Self-Adjoint Operators

If S is an operator on an inner product space V, we define the *adjoint* S^* of S to be an operator such that, for all u, v in V we have

$$\langle S^* u, v\rangle = \langle u, Sv\rangle.$$

It is an easy exercise to show that, if it exists, the adjoint is unique. If V is $\mathbb{C}^n$ and the inner product is the usual complex dot product, amd M is a matrix repsenting S on V, then S^8 is represented by the conjugate transpose of M, which we usually denote by M^*.

By way of a second example, if V is the vector space of all real polynomials and

$$\langle p, q\rangle := \int_a^b p(x)q(x)\, w(x) dx$$

then

$$\langle xp, q\rangle = \langle p, xq\rangle.$$

Hence the operation of multiplication by x is a self-adjoint linear mapping of V. (This is why the theory of orthogonal polynomials is so rich.)

We turn to the existence question. There is a notational difficulty that arises because, outside the context of inner product space, the dual of a linear mapping is often referred as the adjoint (with good reason). We will temporarily use S^d to denote the dual of a linear mapping.

There are two steps to the existence proof. Assume V is an inner product space. If $a \in V$ then we have a map θ_a in V^* given by

$$\theta_a(v) = \langle a, v\rangle.$$

This is linear, and is an isomorphism from V to its dual V^*. If $\in \operatorname{End} V$ then, by the definition of the dual,

$$\langle u, Sv\rangle = (\theta_a \circ S)(b) = (S^d \theta_a)(b).$$

Now $S^d \circ \theta_a \in V^8$ and, since the map $a \mapsto \theta_a$ is an isomorphism, there is a vector $S^*(a)$ in V such that

$$S^d \circ \theta_a = \theta_{S^*(a)}.$$

As the notation suggests, and as you should prove, the map $a \mapsto S^*(a)$ is linear. We say that S^* is the *adjoint* of S.

10.2 Diagonalizability

Our chief tool is the following general result.

10.2.1 Theorem. *Let $\mathscr{B}$ be a commutative matrix algebra with identity over an algebraically closed field. Assume that if $N \in \mathscr{B}$ and $N^2 = 0$, then $N = 0$. Then $\mathscr{B}$ has a basis of pairwise orthogonal idempotents.*

Proof. As a first step, we show that each element of $\mathscr{B}$ is a linear combination of idempotents.

Suppose $A \in \mathscr{B}$. Let $\psi(t)$ be the minimal polynomial of A and assume that

$$\psi(t) = \prod_{i=1}^{k} (t-\theta_i)^{m_i}.$$

If

$$\psi_i(t) := \frac{\psi(t)}{(t-\theta_i)^{m_i}},$$

then from the primary decomposition theorem we know that there are polynomials $f_1(t), \dots, f_k(t)$ such that

$$I = \sum_i f_i(A)\psi_i(A). \tag{10.2.1}$$

Each $f_i(A)\psi_i(A)$ is an idempotent for each i, which we denote it by E_i. Further,

$$(A-\theta_i I)^{m_i} E_i = 0$$

and consequently

$$[(A-\theta_i I)E_i]^{m_i} = 0.$$

Given our hypothesis, it follows that $(A-\theta_i I)E_i = 0$ and we may rewrite (10.2.1) as

$$I = E_1 + \cdots + E_k.$$

Therefore

$$A = AE_1 + \cdots + AE_k = \theta_1 E_1 + \cdots + \theta_k E_k,$$

thus expressing A as a linear combination of idempotents.

We have shown that $\mathscr{B}$ is spanned by idempotents. The essential problem that remains is to show that minimal idempotents exist. Suppose E and F are distinct idempotents and $E \le F$. Then

$$F(I-E) = F-E \neq 0$$

but $E(I-E) = 0$. Hence the column space of E must be a proper subspace of the column space of F. Therefore if $E_1, \dots, E_m$ are distinct idempotents and

$$E_1 \le \cdots \le E_m$$

then $m \le n+1$. We conclude that minimal idempotents exist.

Now we prove that each idempotent is a sum of minimal idempotents. Suppose F is an idempotent and E is a minimal idempotent. If $EF \neq 0$, then $EF \le E$ and therefore $EF = E$. This also shows that distinct minimal idempotents are orthogonal. Let F_0 be the sum of the distinct minimal idempotents E such that $E \le F$. Then F_0 is an idempotent. If $F_0 \neq F$ then $F - F_0$ is an idempotent and so there is a minimal idempotent below it, which contradicts our choice of F_0. We conclude that $\mathscr{B}$ is spanned by minimal idempotents. □

10.2.2 Corollary. *Let $\mathscr{B}$ be a commutative matrix algebra of $n \times n$ matrices with identity over an algebraically closed field $\mathbb{F}$. Assume that if $N \in \mathscr{B}$ and $N^2 = 0$, then $N = 0$. Then there is a basis of $\mathbb{F}^n$ relative to which all elements are diagonalizable.*

Proof. Take the union of bases of the column spaces for the idemptoents. □

10.3 Diagonalization of Self-adjoint Operators

We prove that self-adjoint operators are diagonalizable, and more.

10.3.1 Theorem. *Let S be a self-adjoint operator on the inner product space V. Then*

(a) The minimal polynomial of S has only simple zeros.

(b) Eigenvectors of S with distinct eigenvalues are orthogonal.

(c) The eigenvalues of S are real.

(d) S is diagonalizable.

Proof. Let $\mathscr{A}$ denote the algebra generated by S. If $M \in \mathscr{A}$ and $M^2 = 0$, then

$$0 = \operatorname{tr}(M^2) = \|M\|^2$$

and therefore $M = 0$. By Corollary 10.2.2, each element of $\mathscr{A}$ is diagonalizable and, consequently, the minimal polynomial of each element has only simple zeros.

Next, suppose

$$Su = \theta u, \quad Sv = \tau v.$$

Then

$$\tau \langle u, v\rangle = \langle u, \tau v\rangle = \langle u, Sv\rangle = \langle Su, v\rangle = \langle \theta u, v\rangle = \theta \langle u, v\rangle;$$

and we conclude that either $\theta = \tau$, or $\langle u, v\rangle = 0$.

On the other hand, the proof of Corollary 10.5.2 shows that each S^*S-invariant subspace of V contains an eigenvector for S^*S with a non-negative real eigenvalue. Since $S^*S = S^2$, any S-invariant subspace is S^*S-invariant. Suppose the eigenvalue of S^2 on V_i is σ_i. Then $\sigma_i = \theta_i^2$, and therefore θ is real. □

10.3.2 Corollary. *Suppose S is a self-adjoint operator on the inner product space V. Then there is an orthogonal basis for V formed of eigenvectors for S.* □

We offer a second proof that for any self-adjoint operator on a finite dimensional space, there is an orthogonal basis for the space that consists of eigenvectors. The is a consequence of the following result.

10.3.3 Lemma. *Let S be a self-adjoint operator on the inner product space V. If U is an S-invariant subspace of V, then $U^\perp$ is S-invariant.*

Proof. If $v \in U^\perp$, then

$$\langle Sv, u\rangle = \langle v, Su\rangle$$

and therefore Sv lies in $U^\perp$. □

This makes everything easy. Suppose S is self-adjoint and λ is a zero of its minimal polynomial. Then there is an eigenvector z associated with λ. Its span U is certainly S-invariant, and hence $U^\perp$ is S-invariant. The restriction of S to $U^\perp$ is self-adjoint (prove it) and therefore by induction on the dimension, $U^\perp$ has an orthogonal basis formed from eigenvectors. This basis together with z provides the basis of V that we need. □

10.4 Rank-1 Approximation

Let B be a complex $m \times n$ matrix. We want the rank-1 matrix Z such that $\|B - Z\|$ is minimal.

10.4.1 Lemma. *If B is $m \times n$ and y and z are unit vectors in $\mathbb{C}^m$ and $\mathbb{C}^n$ respectively, then the minimum value of $\|A - \lambda yz^*\|^2$ is equal to*

$$\langle A, A\rangle - \langle yz^*, A\rangle\langle A, yz^*\rangle,$$

and occurs when $\lambda = \langle yz^*, A\rangle$.

Proof. We have

$$\begin{aligned}\langle B-\lambda yz^*, B-\lambda yz^*\rangle &= \langle B,B\rangle - \overline{\lambda}\langle yz^*,B\rangle - \lambda\langle B,yz^*\rangle + \lambda\overline{\lambda}\\ &= \langle yz^*,B\rangle\langle B,yz^*\rangle - \overline{\lambda}\langle yz^*,B\rangle - \lambda\langle B,yz^*\rangle + \lambda\overline{\lambda} + \langle B,B\rangle - \langle yz^*,B\rangle\langle B,yz^*\rangle\\ &= (\langle yz^*,B\rangle - \lambda)(\langle B,yz^*\rangle - \overline{\lambda}) + \langle B,B\rangle - \langle yz^*,B\rangle\langle B,yz^*\rangle.\end{aligned}$$

As

$$(\langle yz^*,B\rangle - \lambda)(\langle B,yz^*\rangle - \overline{\lambda}) = \|\langle B,yz^*\rangle - \overline{\lambda}\|^2$$

and the lemma follows. □

Now our task is to determine unit vectors y and z so that $\langle yz^*,B\rangle\langle B,yz^*\rangle$ is maximal. Set $Q(y) = \langle yz^*,B\rangle\langle B,yz^*\rangle$.

Then

$$\begin{aligned}Q(y+h) &= ((y+h)^*Bz)(z^*B(y+h))\\ &= (y^*Bz + h^*Bz)(z^*By + z^*Bh)\\ &= Q(y) + h^*Bz + z^*Bh + (h^*Bz)(z^*Bh).\end{aligned}$$

If $h^*y = 0$ and $\|h\|$ is small and y maximizes $Q(y)$, we see that

$$h^*Bz + z^*Bh = 0.$$

Replacing h by ih, we also find that

$$-h^*Bz + z^*Bh = 0,$$

and therefore we must have $h^*Bz = 0$. So $h \in (Bz)^\perp$ if $h \in y^\perp$, implying that $y^\perp \le (Bz)^\perp$ and hence that Bz lies in $\langle y\rangle$.

A similar argument shows that if z is a unit vector that maximizes $\langle yz^*,B\rangle\langle B,yz^*\rangle$, then $B^Ty \in \langle z\rangle$. Hence we assume that

$$B^Ty = \lambda z, \qquad Bz = \mu y. \tag{10.4.1}$$

Then

$$\lambda\mu y = \lambda Bz = BB^Ty, \qquad \mu\lambda z = \mu B^Ty = B^TBz$$

and so y and z are respectively eigenvectors of BB^T and B^TB. From (10.4.1) we also find that

$$z^TB^Ty = \lambda, \quad y^TBz = \mu$$

and there $\lambda = \mu$.

10.5 Eigenvectors and Optimization

We present a result which may appear to be of limited interest, but it provides an important reason why we should be interested in eigenvectors. It also illustrates how self-adjoint operators can arise in practice.

10.5.1 Lemma. *Let L be a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$, let U be a subspace of $\mathbb{R}^n$, and let u be a unit vector in U such that $\|Lu\|$ is maximal. If $h \in U$ and $h^T u = 0$, then $h^T L^T Lu = 0$.*

Proof. We have

$$\|L(u+th)\|^2 = (u+th)^T L^T L(u+th) = u^T L^T Lu + 2tu^T L^T Lh + t^2 h^T L^T Lh.$$

Since $\langle u, h\rangle = 0$, we have

$$\|u+th\|^2 = \|u\|^2 + t^2\|h\|^2 = 1 + t^2\|h\|^2.$$

Assuming that t is small enough that t^2 is negligible, we find that

$$\|Lu\|^2 - \frac{\|L(u+th)\|^2}{\|u+th\|^2} \approx -2th^T L^T Lu.$$

We may choose t to be positive or negative; as we have chosen the unit vector u in U to maximize $\|Lu\|$ it follows that if h is orthogonal to u, then $h^T L^T Lu = 0$, and therefore Lu and h are orthogonal. □

Now we present the application of this lemma to eigenvectors.

10.5.2 Corollary. *Let L be a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$, let U be a subspace of $\mathbb{R}^n$, and let u be a unit vector in U such that $\|Lu\|$ is maximal. If U is $L^T L$-invariant, then u is an eigenvector of $L^T L$, and its eigenvalue is non-negative and real.*

Proof. Suppose u is as stated. From the previous lemma we see that if $h \in U$ and $h \in u^\perp$, then $h^T L^T Lu = 0$. Therefore

$$U \cap u^\perp \subseteq (L^T L)^\perp,$$

from which we have

$$L^T Lu \in \operatorname{span}(u) + U^\perp.$$

Therefore $L^T Lu = \theta u + v$, where $v \in U^\perp$. But U is $L^T L$-invariant, and therefore $L^T Lu \in U$. Hence

$$L^T Lu - \theta u \in U^\perp \cap U = \{0\}$$

and so u is an eigenvector for $L^T L$. □

Obviously $\mathbb{R}^n$ itself is $L^T L$-invariant, and thus it follows that if u is a unit vector in $\mathbb{R}^n$ that maximizes $u^T L^T Lu$, then u is an eigenvector for $L^T L$ Since the associated eigenvalue is the maximum value of a non-negative real function, the final claim holds. □

We consider one important case where we are interested in maximizing $\|Lu\|$ over unit vectors. Let A be an $n \times n$ invertible matrix and consider the system of linear equations

$$Ax = b. \tag{10.5.1}$$

If z is a vector then the solution to $Ax = b + z$ is $A^{-1}b + A^{-1}z$. Thus we may say that an error z in b leads to an error $A^{-1}z$ in the solution to (10.5.1).

Which vector z leads to the greatest error? It is clear that if, for example, we replace z by $2z$ then the error is doubled, thus it makes sense to consider

$$\max_{\|z\|=1} \|A^{-1}z\|.$$

From our considerations above, the maximum value of this occurs when z is an eigenvector of

$$(A^{-1})^T A^{-1} = (AA^T)^{-1}.$$

The magnitude of the error will be given by the eigenvalue associated with z. We will see that the eigenvalues of AA^T are real and positive. If the matrix M is invertible, then θ is an eigenvalue of M if and only if θ^{-1} is an eigenvalue of M^{-1}. It follows that the solution of (10.5.1) will be most sensitive to errors in b when the least eigenvalue of AA^T is small.

10.6 The Singular Value Decomposition

If the $m \times n$ matrix A has rank k, then it can be shown that there is an $m \times k$ matrix X and a $k \times n$ matrix Y such that $\operatorname{rk}(X) = \operatorname{rk}(Y) = k$ and $A = XY^T$. When we work over $\mathbb{R}$ (or $\mathbb{C}$), we can prove a somewhat stronger version of this, known as the *singular value decomposition*. This is extremely important in practice.

10.6.1 Theorem. *Let A be a non-zero real matrix with rank k. Then $A = Y\Sigma X^T$, where*

(a) $X^T X = I_k$,

(b) *Σ is a $k \times k$ diagonal matrix Σ with positive diagonal entries,*

(c) $Y^T Y = I_k$.

Proof. Assume A is $m \times n$. Using induction on k, we construct an orthonormal subset $x_1, \ldots, x_k$ of $\mathbb{R}^n$ and an orthonormal subset $y_1, \ldots, y_k$ of $\mathbb{R}^m$ such that $y_i = \sigma_i A x_i$ and

$$A = \sum_{i=1}^{k} \sigma_i y_i x_i^T.$$

This is equivalent to the statement of the theorem.

Let U_0 denote $\mathbb{R}^n$ and let x_1 be a unit vector in U_0 such that $\|Ax_1\|$ is maximal. Set σ_1 equal to $\|Ax\|$ and define

$$x_1 := x, \quad y_1 = \sigma_1^{-1} A x_1.$$

Let U_1 denote $x_1^{\perp}$. By Lemma 10.5.1 we see that if $h^T x_1 = 0$, then $h^T A^T A x_1 =$ and consequently $A(U_1^{\perp}) \subseteq A(U_1)^{\perp}$.

Suppose

$$A_1 := A - \sigma_1 y x^T.$$

Since y lies in the column space of A, we see that $\operatorname{col}(A_1) \subseteq \operatorname{col}(A)$. Since $A \neq 0$ we see that $x_1 \neq 0$ and $y_1 \neq 0$. Therefore $Ax \neq 0$, but

$$A_1 x = Ax - \sigma_1 y x^T x = \sigma_1 y - \sigma_1 y = 0.$$

Consequently $\operatorname{rk}(A_1) < \operatorname{rk}(A)$. As $\operatorname{rk}(\sigma_1 y x^T) = 1$ it follows that $\operatorname{rk}(A_1) = k-1$.

Note next that if $x \in U_1$, then $Ax = A_1 x$ and so A and A_1 agree on U_1. Working now with A_1 and U_1, we conclude by induction on k that there are orthogonal unit vectors $x_2, \dots, x_k$ in $\mathbb{R}^n$ and orthogonal unit vectors $y_2, \dots, y_k$ in $\mathbb{R}^m$, such that $y_i = A_1 x_i$ and, if $\sigma_i := \|A_1 x_i\|$, then

$$A_1 = \sum_{i=2}^{k} \sigma_i y_i x_i^T.$$

Our theorem follows immediately. □

In numerical work, the following alternative version of the singular value decomposition may be more useful. (It does not assume we know the rank of A.)

10.6.2 Corollary. *If A is a square matrix, there are orthogonal matrices X and Y, and a non-negative diagonal matrix Σ such that $A = Y\Sigma X^T$.* □

The matrices Y and X in the singular value decomposition $Y\Sigma X^T$ of A are not unique in any useful sense. However Σ is determined up to a permutation. Its entries are known as the *singular values* of A; there are usually denoted by $\sigma_1, \dots, \sigma_n$, with the assumption that they form a non-increasing sequence.

The easiest way to see that the singular values are determined by A is to verify that they are the squares are the eigenvalues of AA^T. To show this, note that

$$AA^T = Y\Sigma X^T X \Sigma Y^T = Y\Sigma^2 Y^T,$$

and therefore

$$AA^T Y = Y\Sigma^2.$$

It follows from this that the columns of Y are eigenvectors for AA^T, and the diagonal entries of Σ^2 are its eigenvalues.

In a similar fashion we can show that the squares of the singular values of A are the eigenvalues of $A^T A$. Hence we see that AA^T and $A^T A$ have the same eigenvalues. (This actually holds over any field, although the proof at hand only works over $\mathbb{R}$ or, with modest extra effort, over $\mathbb{C}$.)

(1) Prove Corollary 10.6.2.

(2) Compute the singular values of a companion matrix. (You may work with either CC^T or $C^T C$, but one is significantly easier. First show that all but two of the singular values are equal to 1.)

(3) Show that the sum of the singular values of a square matrix is a norm.

(4) If $\sigma_1(A)$ denote the largest singular value of A, show that it is a norm.

.

10.7 Least Squares

We consider the system of linear equations

$$Ax = b \tag{10.7.1}$$

where A is $m \times n$. In **??** we considered the case where the rows of A are linearly independent. Then the columns of A span $\mathbb{R}^m$, and we want the vector x with minimum norm such that $Ax = b$. The second, and more commonly met situation, is when the columns of A are linearly independent, and we want the vector x such that $\|b - Ax\|^2$ is minimal.

We draw attention to one difficulty. It is in fact a non-trivial numerical problem to determine the rank of a real matrix, and so it may not be easy to verify that the rows or columns of A are linearly independent. In fact, the best way to determine the rank in finite precision arithmetic is to use the singular value decomposition $A = Y\Sigma X^T$, since $\mathrm{rk}(A) = \mathrm{rk}(\Sigma)$. (Thus determining the rank of A is reduced to determining the rank of a diagonal matrix; in the presence of rounding errors and uncertainties in the data, this still may require thought.) But rather than using the singular value decomposition just to get the rank of A, we can use it to solve the least squares problem.

10.7.1 Lemma. *Let A be an $m \times n$ real matrix with singular value decomposition $A = Y\Sigma X^T$, where Σ is $k \times k$ and invertible. Then the vector z of minimum norm, such that $b - Az$ has minimum norm is given by*

$$z = X\Sigma^{-1}Y^T b.$$

Proof. We note that the columns of Y form an orthonormal basis for $\mathrm{col}(A)$, whence the matrix representing projection onto $\mathrm{col}(A)$ is YY^T. Similarly, the columns of X form an orthonormal basis for $\mathrm{col}(A^T)$, and therefore XX^T is the matrix representing projection onto $\mathrm{col}(A^T)$.

Consequently $y = YY^T b$ is the vector in $\mathrm{col}(A)$ closest to b. Suppose $Ax = y$. Then

$$Y\Sigma X^T x = YY^T b$$

and, multiplying both sides on the left by Y^T, we have

$$X^T x = \Sigma^{-1}Y^T b.$$

Now $XX^T x$ is the projection of x onto $\mathrm{col}(A^T)$, and accordingly

$$z = X\Sigma^{-1}Y^T b$$

is the vector of minimum norm such that Az is closest to b. □

10.8 Legendre Polynomials

Let V be Pol($\mathbb{R}$), the vector space of all real polynomials, with inner product

$$\langle p, q\rangle = \int_{-1}^{1} p(t)q(t)\,dt.$$

Define a linear mapping $L : V \to V$ by

$$L(p) = (1-t^2)p'' - 2tp'.$$

If $n \geq 2$ then

$$L(t^n) = (1-t^2)n(n-1)t^{n-2} - 2nt^n = -n(n+1)t^n + n(n-1)t^{n-2}. \qquad (10.8.1)$$

It follows that

$$\langle t^m, Lt^n\rangle = \int_{-1}^{1} (n(n-1)t^{m+n-2} - n(n+1)t^{m+n})\,dt;$$

when $m+n$ is odd the integral here is zero, if $m+n$ is even then it is

$$\left[\frac{2n(n-1)}{m+n-1} - \frac{2n(n+1)}{m+n+1}\right] = -\frac{4mn}{(m+n)^2-1}.$$

Hence

$$\langle t^m, Lt^n\rangle = \langle Lt^m, t^n\rangle$$

for all m and n. It follows that for any polynomials p and q,

$$\langle p, Lq\rangle = \langle Lp, q\rangle,$$

and therefore L is self-adjoint. (This can also be proved directly using integration by parts.)

It follows that the eigenvalues of L are real, and eigenvectors with distinct eigenvectors are orthogonal with respect to the above inner product. It is not hard to determine the eigenvalues of L. From (10.8.1) we see that Pol$_n(\mathbb{R})$ is L-invariant and further, if L_n denotes the restriction of L to Pol$_n(\mathbb{R})$ and $\beta = \{1, t, \dots, t^n\}$ is the standard basis for Pol$_n(\mathbb{R})$, then

$$[L_n]_\beta = \begin{pmatrix} 0 & 0 & 2 & & & \\ & -2 & 0 & 6 & & \\ & & -6 & 0 & 12 & \\ & & & -12 & 0 & \\ & & & & \ddots & \end{pmatrix}$$

This is a triangular matrix, and reveals that the eigenvalues of L_n are the integers $-m(m-1)$ for $m = 1, \dots, n$.

As the eigenvalues are distinct, each eigenspace is 1-dimensional and is thus spanned by a polynomial. The polynomial with eigenvalue $-m(m-1)$ will have degree m and is a solution of *Legendre's equation*:

$$(1-t^2)p'' - 2tp' + m(m-1)p = 0.$$

We call p_m the *Legendre polynomial* of degree m. The first five Legendre polynomials are as follows:

$$\begin{aligned}
p_0 &= 1\\
p_1 &= t\\
p_2 &= 3t^2 - 1\\
p_3 &= 5t^3 - 3t\\
p_4 &= 35t^4 - 30t^2 + 3.
\end{aligned}$$

It makes no harm if we replace p_i by any non-zero scalar multiple of itself, and it is customary to choose the multiple so that $p_i(1) = 1$. (But we have not done that here.)

There are a number of related examples (of self-adjoint linear operators on $P(\mathbb{R})$). We summarize some of them here. The numbers $\lambda_0, \lambda_1, \dots$ are the eigenvalues of the operator.

(a) **Chebyshev**.

$$Lp = (1-t^2)p'' - tp'; \quad \langle p,q\rangle = \int_{-1}^{1} p(t)q(t)\,\frac{dt}{\sqrt{1-t^2}}; \quad \lambda_n = -n^2.$$

(a) **Laguerre**.

$$Lp = tp'' + (1-t)p'; \quad \langle p,q\rangle = \int_0^\infty p(t)q(t)e^{-t}\,dt; \quad \lambda_n = -n.$$

(a) **Hermite**.

$$Lp = p'' - tp'; \quad \langle p,q\rangle = \int_{-\infty}^{\infty} p(t)q(t)e^{-t^2/2}\,dt; \quad \lambda_n = -n.$$

In general, if

$$Lp = fp'' + gp'$$

then we may write

$$Lp = w^{-1}(wfp')'$$

where

$$w(t) = \frac{1}{f(t)}\exp\int_\alpha^t \frac{g(u)}{f(u)}\,du.$$

(The value of the lower limit α in this integral will be determined by context.) Then L is self-adjoint relative to the inner product

$$\langle p,q\rangle = \int p(t)q(t)\,w(t)dt.$$

To see this, compute in outline as follows:

$$\begin{aligned}
\langle Lp,q\rangle = \int q(wfp')'\,dt = -\int wfp'q'\,dt &= -\int wfq'p'\,dt\\
&= \int p(wfq')'\,dt\\
&= \langle p,Lq\rangle.
\end{aligned}$$

For this computation to be accurate, $f(t)w(t)$ must vanish at the endpoints of the interval over which we integrate.

The eigenvectors of L will be polynomials only if $w(t)$ satisfies further restrictions.

10.9 Computing Eigenvalues

How do people really compute the eigenvalues of symmetric matrices? They do not use the method offered in most introductory linear algebra course—compute the characteristic polynomial, find its zeros—that is probably the fourth best method. Here we outline the second best.

So, suppose A is a real symmetric $n \times n$ matrix. We want to find an orthogonal matrix L such that $L^T AL$ is diagonal. What we will actually do is to describe how to find a sequence of orthogonal matrices $S_1, \dots, S_r$ such that all off-diagonal entries of

$$S_r^T \cdots S_1^T AS_1 \cdots S_r$$

are very small, we can then take the diagonal entries of this matrix to be the eigenvalues of A.

The basic idea is to note that we can diagonalize symmetric 2×2 matrices. Using this we choose the matrix S_{i+1} so that it makes some off-diagonal entry of $S_i^T \cdots S_1^T AS_1 \cdots S_i$ equal to 0. Unfortunately this will usually make some off-diagonal entries non-zero, when they were already zero. This will make us work harder, but will not prevent eventual success.

If M is a symmetric matrix then there is an orthogonal matrix L such that $L^T ML$ is diagonal; if M is 2×2 then we may assume that L has the form

$$\begin{pmatrix} c & -s \\ s & c \end{pmatrix}.$$

where $c^2 + s^2 = 1$. (We could, but do not, assume that $c \geq 0$.) Now suppose that A is a symmetric $n \times n$ matrix, that B is the leading principal 2×2 submatrix of A and that R is an orthogonal matrix such that $R^T BR$ is diagonal. Let S denote the matrix

$$\begin{pmatrix} R & 0 \\ 0 & I_{n-2} \end{pmatrix}.$$

Then

$$B = S^T AS$$

is similar to A and $B_{1,2} = 0$. In general, if $A_{i,j} \neq 0$ then there is an orthogonal matrix S such that the only non-zero off-diagonal entries of S are the ij and ji entries, and $(S^T AS)_{i,j} = 0$. We call S a *Givens rotation*.

How does this help us. If A and B are $n \times n$ matrices, define

$$\langle A, B \rangle = \operatorname{tr} AB^T.$$

Then $\|A\|^2$ is the sum of the squares of the entries of A and, if L is orthogonal,

$$\|L^T AL\|^2 = \operatorname{tr}(L^T A^T LL^T A^T L) = \operatorname{tr}(L^T LAA^T L) = \operatorname{tr}(AA^T) = \|A\|^2.$$

Let $\operatorname{sqo}(A)$ denote the sum of the squares of the off-diagonal entries of A. Note that, in passing from A to $S^T AS$, the only diagonal entries that change are the ii- and jj-entries and that the sum of the squares of these two entries increases by $2(A_{i,j})^2$. It follows that, if S and A are as above,

$$\operatorname{sqo}(S^T AS) = \operatorname{sqo}(A) - 2(A_{i,j})^2.$$

If $\operatorname{sqo}(A) = c$, there are indices i and j such that

$$(A_{i,j})^2 \geq \frac{c}{n(n-1)}$$

and hence there is a Givens rotation S such that

$$\operatorname{sqo}(S^T AS) \leq c\left(1 - \frac{2}{n(n-1)}\right).$$

This implies that, by successively applying Givens rotations, we can form a matrix M, orthogonally similar to A and such that $\operatorname{sqo}(M)$ is as small as we like. The diagonal entries of M will be the eigenvalues of A.

10.10 Jacobi: An Example

By way of example, suppose that

$$A = \begin{pmatrix} 1 & 0.5 & 0.3333 \\ 0.5 & 0.3333 & 0.25 \\ 0.3333 & 0.25 & 0.2 \end{pmatrix}.$$

Then Jacobi's method runs through the following iterations.

$$\begin{pmatrix} c \\ s \end{pmatrix} = \begin{pmatrix} -0.47185 \\ 0.88167 \end{pmatrix}; \quad [1,2] \to \begin{pmatrix} 0.065741 & 0 & 0.063132 \\ 0 & 1.2675 & -0.41185 \\ 0.063132 & -0.41185 & 0.2 \end{pmatrix}$$

$$\begin{pmatrix} c \\ s \end{pmatrix} = \begin{pmatrix} -0.32269 \\ -0.94650 \end{pmatrix}; \quad [2,3] \to \begin{pmatrix} 0.06574 & -0.05975 & -0.02037 \\ -0.05975 & 0.05958 & 0 \\ -0.02037 & 0 & 1.40801 \end{pmatrix}$$

$$\begin{pmatrix} c \\ s \end{pmatrix} = \begin{pmatrix} -0.68867 \\ -0.72507 \end{pmatrix}; \quad [1,2] \to \begin{pmatrix} 0.002829 & 0 & 0.01403 \\ 0 & 0.12250 & -0.01477 \\ 0.01403 & -0.01477 & 1.40801 \end{pmatrix}$$

$$\begin{pmatrix} c \\ s \end{pmatrix} = \begin{pmatrix} -0.99993 \\ -0.01149 \end{pmatrix}; \quad [2,3] \to \begin{pmatrix} 0.00283 & -0.00016 & -0.01403 \\ -0.00016 & 0.12232 & 0 \\ -0.01403 & 0 & 1.40818 \end{pmatrix}$$

$$\begin{pmatrix} c \\ s \end{pmatrix} = \begin{pmatrix} -0.99995 \\ -0.00998) \end{pmatrix}; \quad [1,3] \to \begin{pmatrix} 0.00269 & 0.00016 & 0 \\ 0.00016 & 0.12233 & 0 \\ 0 & 0 & 1.40832 \end{pmatrix}$$

$$\begin{pmatrix} c \\ s \end{pmatrix} = \begin{pmatrix} -1.0 \\ 0.00135 \end{pmatrix}; \quad [1,2] \to \begin{pmatrix} 0.00268 & 0 & 0 \\ 0 & 0.122327 & 0 \\ 0 & 0 & 1.40832 \end{pmatrix}$$

Here the diagonal entries are the eigenvalues of A, and further iterations do not change them.

(1) Suppose that $\langle Av, Aw\rangle = 0$ whenever $\langle v, w\rangle = 0$. Prove, or disprove, that A is a scalar multiple of an orthogonal matrix.

(2) Suppose $Q^2 = Q$ and $Q = Q^T$. Show that $I - 2Q$ is a symmetric orthogonal matrix, and explain the connection to reflections.

(3) Prove that an involution is symmetric if and only if it is orthogonal.

(4) Show that each involution has the form $I - 2P$, for some idempotent P.

(5) Show that, if A and A^{-1} are similar, there is an involution T such that $TAT = A^{-1}$.

11
Spectral Decomposition

11.1 Self-Adjoint Operators

The spectral decomposition of an operator is a more concrete form of diagonalizability. It is most useful when the operator is self-adjoint, so we confine ourselves to that case.

Suppose S is an operator on the inner product space V and that the minimal polynomial ψ of S is given by

$$\psi(t) = \prod_{i=1}^{k} (t-\theta_i),$$

where the zeros θ_i are distinct. (Thus A is diagonalizable.) By the primary decomposition theorem (Theorem 3.6.1), there are polynomials p_i such that

$$I = \sum_{i=1}^{k} p_i(S), \tag{11.1.1}$$

where

(a) $p_i(S)$ is idempotent,

(b) $p_(S)p_j(S) = 0$ if $i \neq j$, and

(c) S acts on $\operatorname{col}(p_i(S))$ as multiplication by θ_i.

Assume $E_i := p_i(S)$. Then $SE_i = \theta_i E_i$ and, by (11.1.1),

$$S = \sum_i \theta_i E_i. \tag{11.1.2}$$

Equation (11.1.2) is known as the *spectral decomposition* of S.

One consequence of the spectral decomposition is that

$$S^n = \sum_i \theta_i^n E_i;$$

this can provide a simple way to compute powers of S.

If S is self-adjoint, then the operators E_i are self-adjoint, because each E_i is a polynomial in S.

We now offer a matrix view of the spectral decomposition. If A is a diagonalizable $n \times n$ matrix, then $\mathbb{F}^n$ has a basis consisting of eigenvectors for A. Let L be the matrix with these eigenvectors as its columns. Then L is an invertible matrix and there is a diagonal matrix D such that

$$AL = LD.$$

It follows that

$$A = LDL^T.$$

We can write D as a sum of 01-diagonal matrices D_i:

$$D = \sum_i \theta_i D_i,$$

where $\theta_1, \dots, \theta_m$ are the distinct eigenvalues of A and $\sum_i D_i = I$. Accordingly

$$A = \sum_i \theta_i LD_i L^{-1}.$$

It is easy to verify that

$$(LD_i L^{-1})^2 = LD_i L^{-1}$$

and, if $i \neq j$, then $D_i D_j = 0$ and

$$LD_i L^{-1} LD_j L^{-1} = LD_i D_j L^{-1} = 0$$

11.2 Commutative Algebras

Two idempotents E and F are *orthogonal* if $EF = 0$. For example, if E is an idempotent, then E and $I - E$ are orthogonal idempotents. We can define a partial ordering on the idempotents of a commutative algebra $\mathscr{A}$ as follows. If E and F are idempotents in $\mathscr{A}$, we declare that $E \leq F$ if $FE = E$. This relation is reflexive, antisymmetric and transitive; therefore it is a partial order. A *minimal idempotent* is a minimal element of the set of nonzero idempotents, relative to this order. If E and F are idempotents, then $EF \leq E, F$. It follows that if E and F are minimal, then they are orthogonal.

11.2.1 Theorem. *Let $\mathscr{B}$ be a commutative matrix algebra with identity over an algebraically closed field. Assume that if $N \in \mathscr{B}$ and $N^2 = 0$, then $N = 0$. Then $\mathscr{B}$ has a basis of pairwise orthogonal idempotents.*

Proof. As a first step, we show that each element of $\mathscr{B}$ is a linear combination of idempotents.

Assme the matrices in $\mathscr{B}$ have order $n \times n$. Suppose $A \in \mathscr{B}$ and let $\psi(t) = \prod_{i=1}^k (t - \theta_i)^{m_i}$ be its minimal polynomial. There are idempotents E_i, summing to i, such that $\operatorname{im}(E_i)$ is the root space associated with θ_i, and $\mathbb{F}^n$ is the direct sum of these root spaces.

Further, the minimal polynomial of A on $\operatorname{im}(E_i)$ is $(t - \theta_i)^{m_i}$, and hence we have

$$0 = (A - \theta_i I)^{m_i} E_i = ((A - \theta_i I)E_i)^{m_i}.$$

If $m_i > 1$, we set $k = \lfloor (m_i + 1)2 \rfloor$ and $N = ((A - \theta_i I)E_i)^k$. Then $N \neq 0$ but $N^2 = 0$. We conclude that zeros of the minimal polynomial of A are simple. We also see that $\operatorname{im}(E_i)$ is an eigenspace for A and as $I = \sum_i E_i$ it follows that

$$A = AI = \sum_i AE_i = \sum_i \theta_i E_i.$$

Therefore A is a linear combination of idempotents belonging to $\mathscr{B}$, and it follows that $\mathscr{B}$ is spanned by idempotents.

The problem that remains is to show that minimal idempotents exist. Suppose E and F are distinct idempotents and $E \leq F$. Then

$$F(I - E) = F - E \neq 0$$

but $E(I - E) = 0$. Hence the column space of E must be a proper subspace of the column space of F. Therefore if $E_1, \dots, E_m$ are distinct idempotents and

$$E_1 \leq \cdots \leq E_m$$

then $m \leq n + 1$. We conclude that minimal idempotents exist.

Now we prove that each idempotent is a sum of minimal idempotents. Suppose F is an idempotent and E is a minimal idempotent. If $EF \neq 0$, then $EF \leq E$ and therefore $EF = E$. This also shows that distinct minimal idempotents are orthogonal. Let F_0 be the sum of the distinct minimal idempotents E such that $E \leq F$. Then F_0 is an idempotent. If $F_0 \neq F$ then $F - F_0$ is an idempotent and so there is a minimal idempotent below it, which contradicts our choice of F_0. We conclude that $\mathscr{B}$ is spanned by minimal idempotents. □

A matrix N is nilpotent if $N^k = 0$ for some k. Theorem 11.2.1 asserts that a commutative matrix algebra with identity has a basis of orthogonal idempotents if there are no non-zero nilpotent matrices in it. Since a non-zero linear combination of pairwise orthogonal idempotents cannot be nilpotent, this condition is necessary too. A commutative algebra is *semisimple* if it contains no non-zero nilpotent elements.

11.3 Normal Operators

An operator A on an inner product space is *normal* if $AA^* = A^*A$. We consider examples. Clearly any self-adjoint operator is normal. Unitary operators are a second important class. If $A = L^*DL$ where D is diagonal and L is unitary, then

$$AA^* = L^*DLL^*\overline{D}L = L^*D\overline{D}L = L^*\overline{D}DL = A^*A$$

and so any matrix that is unitarily similar to a diagonal matrix is normal.

Exercise: determine which complex 2 × 2 matrices are normal.

Exercise: If H is normal, show that we can write it as $H = A + iB$, where A and B are Hermitian and commute.

11.3.1 Theorem. *Suppose $\mathscr{A}$ is a commutative subalgebra of* $\operatorname{Mat}_{v\times v}(\mathbb{C})$ *that is closed under conjugate transpose and contains the identity. Then $\mathscr{A}$ has a basis of matrix idempotents $E_0,\dots,E_d$ such that*

(a) $E_iE_j = \delta_{i,j}E_i$.

(b) The columns of E_i are eigenvectors for each matrix in $\mathscr{A}$.

(c) $\sum_{i=0}^{d} E_i = I$.

(d) $E_i^* = E_i$. □

Proof. Suppose $N \in \mathscr{A}$ and $N^2 = 0$. Then

$$0 = (N^*)^2N^2 = (N^*N)^2$$

and hence

$$0 = \operatorname{tr}((N^*N)^2) = \operatorname{tr}((N^*N)^*(N^*N)).$$

If $H := N^*N$, then $\operatorname{tr}(H^*H) = 0$ if and only if $H = 0$, so we deduce that $N^*N = 0$. But then $\operatorname{tr}(N^*N) = 0$ and therefore $N = 0$. Hence $\mathscr{A}$ satisfies the hypotheses of 11.2.1, and therefore it has a basis that consists of pairwise orthogonal idempotents.

We show that the idempotents E_i are Hermitian. Since $\mathscr{A}$ is closed under transpose and complex conjugation, $E_i^* \in \mathscr{A}$. Therefore there are scalars $a_0,\dots,a_d$ such that

$$E_i^* = \sum_j a_jE_j$$

and so

$$E_i^*E_i = f_iE_i.$$

Since $\operatorname{tr}(E_i^*E_i) > 0$ and $\operatorname{tr}(E_j) > 0$, it follows that $f_i \neq 0$. But E_i^* is a minimal idempotent, and therefore $f_j = 0$ if $j \neq i$. This implies that E_i^* is a scalar multiple of E_i, but $\operatorname{tr}(E_i) = \operatorname{tr}(E_i^*)$, and therefore $E_i^* = E_i$. □

11.3.2 Theorem. *If A is normal, then A is unitarily similar to a diagonal matrix.*

12
Cospectral Graphs

We present some constructions of cospectral graphs, along with the related theory.

12.1 $K_{1,4}$, $C_4 \cup K_1$

The smallest pair of non-isomorphic cospectral graphs are $K_{1,4}$ and $C_4 \cup K_1$.

Note that C_4 is also $K_{2,2}$ and recall that

$$\phi(K_{m,n}, t) = t^{m+n-2}(t^2 - mn).$$

(One way of seeing this is that the functions of the vertices of $K_{m,n}$ that sum to zero on each colour class form an $(m+n-2)$-dimensional subspace of eigenvectors with eigenvalue 0. The remaining eigenvectors are orthogonal to those already in hand, so they are constant on colour classes. Hence the corresponding eigenvalues are the eigenvalues of the quotient relative to the colour partition, which is equitable.)

In particular we have that the graphs

$$K_{1,mn}, \quad K_{m,n} \cup (m-1)(n-1)K_1$$

are cospectral. (You might prove that $K_{m,n}$ is determined by its spectrum.)

12.2 Direct Products

The direct product $P_3 \times P_3$ is the disjoint union of $K_{1,4}$ and C_4. This leads to interesting consequences.

The direct product of two connected graphs is connected if at least one of its components is not bipartite. The direct product of two connected bipartite graphs is the disjoint union of two bipartite graphs. The direct product of graphs X and Y has adjacency matrix $A(X) \otimes A(Y)$. If X and Y are bipartite, we can assume that their adjacency matrices of the respective forms

$$\begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 & C \\ C^T & 0 \end{pmatrix}$$

with tensor product

$$\begin{pmatrix} 0 & 0 & 0 & B\otimes C \\ 0 & 0 & B\otimes C^T & \\ 0 & B^T\otimes C & 0 & 0 \\ B^T\otimes C^T & 0 & 0 & \end{pmatrix}.$$

This the adjacency matrix of the disjoint union of two graphs with adjacency matrices

$$\begin{pmatrix} 0 & B\otimes C \\ B^T\otimes C^T & 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 & B\otimes C^T \\ B^T\otimes C & \end{pmatrix}.$$

The square of these matrices are respectively

$$\begin{pmatrix} BB^T\otimes CC^T & 0 \\ 0 & B^TB\otimes C^TC \end{pmatrix}, \qquad \begin{pmatrix} BB^T\otimes C^TC & 0 \\ 0 & B^TB\otimes CC^T \end{pmatrix}.$$

12.2.1 Lemma. *If M is an $m\times n$ matrix and N is $n\times m$, then MN and NM have the same non-zero eigenvalues with the same multiplcity.*

Proof. First we show that

$$\det(I-MN)=\det(I-NM).$$

We have

$$\begin{pmatrix} I & 0 \\ -N & I \end{pmatrix}\begin{pmatrix} I & M \\ N & I \end{pmatrix}=\begin{pmatrix} I & M \\ 0 & I-NM \end{pmatrix}$$

and

$$\begin{pmatrix} I & M \\ N & I \end{pmatrix}\begin{pmatrix} I & 0 \\ -N & I \end{pmatrix}=\begin{pmatrix} I-MN & M \\ 0 & I \end{pmatrix};$$

taking determinants of both sides of the first of these equations yields

$$\det\begin{pmatrix} I & M \\ N & I \end{pmatrix}=\det(I-NM)$$

and from the second that

$$\det\begin{pmatrix} I & M \\ N & I \end{pmatrix}=\det(I-MN).$$

This shows that $\det(I-tMN)=\det(I-tMN)$.

Now we deduce

$$t^m\phi(MN,t^{-1}=t^m\det(t^{-1}-MN)=t^n\det(t^{-1}-NM)=tn^{\phi}(NM,t^{-1})$$

and therefore $\phi(MN,t)=t^{m-n}\phi(NM,t)$. □

12.3 The Partitioned Tensor Product

12.4 Subdivisions and Line Graphs

The *incidence matrix* of a graph X is the $|V(X)| \times |E(X)|$ 01-matrix with ue-entry equal to one if the vertex u lies on the edge e. We usually denote it by B. If D is the diagonal matrix of valencies of X, then

$$BB^T = D + A(X)$$

and

$$B^T B = 2I + A(L(X)).$$

The matrix $D + A(X)$ is the *unsigned Laplacian* of X, and we see that $D + A(X)$ and $2I + A(L(X))$ have the same non-zero eigenvalues with the same multiplicities.

The bipartite graph with adjacency matrix

$$\begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}$$

is the *subdivision graph* $S(X)$ of X. The square of this matrix is

$$\begin{pmatrix} BB^T & 0 \\ 0 & B^T B \end{pmatrix}$$

and therefore the spectrum of $S(X)$ is determined by the spectrum of $L(X)$ (or by the spectrum of $D + A$).

The graphs $K_{1,3}$ and $K_3 \cup K_1$ both have line graph equal to K_3; it follows that there subdivision graphs are cospectral. These graphs are respectively the subdivision graph of $K_{1,3}$ and $C_6 \cup K_1$. So these graphs are cospectral, what is surprising is that their complements are cospectral as well, and this is the smallest pair of graphs that are cospectral with cospectral complements.

But nothing we have said implies that the complements are cospectral. We address this deficiency.[1]

The incidence matrix of $K_{1,3}$ can be taken to be

$$C_1 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and the incidence matrix of $K_3 \cup K_1$ can be taken to be

$$C_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

[1] in an admittedly roundabout way

12.5 Congruence

If we want to show that the matrices

$$\begin{pmatrix} 0 & C_1 \\ C_1^T & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & C_2 \\ C_2^T & 0 \end{pmatrix}$$

are similar, it is reasonable to look for orthogonal matrices L and M such that

$$\begin{pmatrix} L & 0 \\ 0 & M \end{pmatrix} \begin{pmatrix} 0 & C_1 \\ C_1^T & 0 \end{pmatrix} \begin{pmatrix} L^T & 0 \\ 0 & M^T \end{pmatrix} = \begin{pmatrix} 0 & LC_1M^T \\ MC_1^TL^T & 0 \end{pmatrix} = \begin{pmatrix} 0 & C_2 \\ C_2^T & 0 \end{pmatrix},$$

that is, such that $LC_1M^T = C_2$. We are greedy and observe that if L is orthogonal and $LC_1 = C_2$, then our two subdivision graphs are cospectral.

12.5.1 Theorem. *Let U and V be $d \times m$ matrices. There is an orthogonal matrix Q such that $QU = V$ if and only if $U^TU = V^TV$.*

Proof. Let the columns of U and V be respectively $u_1, \ldots, u_m$ and $v_1, \ldots, v_m$. Assume that $\mathrm{rk}(U) = e$ and the $u_1, \ldots, u_e$ is a basis for the column space of U. Then $v_1, \ldots, v_e$ is a basis for the column space of V.

Since u_1 and v_1 have the same length, the matrix Q_1 representing reflection in the hyperplane $(u_1 - v_1)^\perp$ is an orthogonal matrix swapping u_1 and v_1 and

$$(Q_1U)^TQ_1U = U^TU = V^TV.$$

Now assume inductively that $u_i = v_i$ for $i = 1, \ldots, k$, with $1 \le k \le e$. If y and z are two vectors such that $\langle y, y\rangle = \langle z, z\rangle$ and

$$\langle u_i, y\rangle = \langle u_i, z\rangle, \quad (i = 1, \ldots, k)$$

then $y - z$ is orthogonal to $u_1, \ldots, u_k$ and the reflection in $(y - z)^\perp$ fixes $u_1, \ldots, u_k$ and swaps y and z. This implies that, if $k < e$, there is an orthogonal matrix Q_{k+1} such that the first $k + 1$ columns of $Q_{k+1}U$ and V are equal.

To complete the proof, we observe that if the first e columns of U span $\mathrm{col}(U)$ and are equal to the first e columns of V, then $U = V$. The theorem follows. □

This theorem generalizes that fact that if the sides of two triangles have the same length, the triangles are congruent: that is, there is a composition of a translation, a rotation and (possibly) a reflection that maps one triangle to the other.

From this theorem we deduce that there is an orthogonal matrix Q such that $QC_1 = C_2$. It is easy to verify that we take Q to be

$$\frac{1}{2}J - I$$

Since Q is symmetric it is an involution. We have

$$\begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} 0 & C_1 \\ C_1^T & 0 \end{pmatrix} \begin{pmatrix} Q^T & 0 \\ 0 & I \end{pmatrix} = \begin{pmatrix} QC_1 & 0 \\ 0 & C_1^T Q^T \end{pmatrix} = \begin{pmatrix} 0 & C_2 \\ C_2^T & 0 \end{pmatrix}$$

Next we note that $QJ = J$ and $C_2 = J - C_1$, from which it follows that the complements of $S(K_{1,3})$ and $C_6 \cup K_1$ are cospectral.

12.6 Local Switching

12.7 Extended Adjacency Algebras

It is an experimental fact that if two graphs are cospectral, their complements are likely to be cospectral. Johnson and Newman proved that if graphs X_1 and X_2 are cospectral with cospectral complements, there is an orthogonal matrix L such that

$$A(X_2) = L^T A(X_1) L, \qquad A(\overline{X_2}) = L^T A(\overline{X_1}) L.$$

This implies that $LJ = J$ and that for any scalars α and β,

$$L^T(\alpha A(X_1) + \beta A(\overline{X_1})) = \alpha A(X_2) + \beta A(\overline{X_2}).$$

We will present a proof of the result of Johnson and Newman, and related results concerning cospectral vertices.

Let z be a vector in $\mathbb{R}^n$. Then

$$\begin{aligned} \det(tI - A - zz^T) &= \det(tI - A)\det(I - (tI - A)^{-1} zz^T) \\ &= \det(tI - A)(1 - z^T (tI - A)^{-1} z) \end{aligned}$$

and consequently

$$\frac{\phi(A + zz^T, t)}{\phi(A, t)} = 1 - z^T (tI - A)^{-1} z = 1 - \sum_r \frac{z^T E_r z}{t - \theta_r}.$$

If z is the characteristic vector of a subset S of $V(X)$, then

$$z^T (I - tA)^{-1} z$$

is the generating function for the walks on X that start and finish on vertices in S. So we have a relation between the characteristic polynomial $\phi(A + zz^T, t)$, the generating function for walks starting and ending S, and the sequence of numbers $z^T E_r z$. The latter are the squared lengths of the projections of z into the eigenspaces of A. The set of eigenvalues

$$\{\theta_r : z^T E_r z \neq 0\}$$

is the *eigenvalue support* of z.

The *walk module* $\langle z \rangle_A$ generated by z is the A-invariant subspace spanned by the vectors $A^k z$ for $k \geq 0$. If A has spectral decomposition

$A = \sum_r \theta_r E_r$, the non-zero vectors $E_r z$ form an orthogonal basis for $\langle z \rangle_A$, and so $\dim(\langle z \rangle_A)$ is equal to the size of the eigenvalue support of z. If S is the eigenvalue support of z, then the minimal polynomial of A acting on $\langle z \rangle_A$ is

$$\prod_{\theta_r \in S} (t - \theta_r).$$

We use $\mathscr{A}(z)$ to denote the algebra $\langle A, zz^T \rangle$ generated by A and zz^T.

12.7.1 Lemma. *The walk module $\langle z \rangle_A$ is an irreducible module for $\mathscr{A}(z)$.*

Proof. Assume U is an $\mathscr{A}(z)$-submodule of $\langle z \rangle_A$. Since A and zz^T are symmetric, $U^\perp$ is also a submodule for $\mathscr{A}(z)$.

If $z \in U$, then $U = \langle z \rangle_A$. If $u \in U$ and $z^T U \neq 0$, then $zz^T u \neq 0$ and therefore $z \in U$. So we may assume that $U \leq z^\perp$, and therefore $z \in U^\perp$. But this implies that $U^\perp = \langle z \rangle_A$ and $U = 0$. □

[Johnson and Newman]

13
Perturbation Theory

It seems reasonable that if A and B are square matrices of the same order and ϵ is small, then we should be able to relate the eigenvalues of $A+\epsilon B$ to the eigenvalues of A. This can be done, provided that we restrict to the case where A and B are Hermitian, and provided we are prepared to apply considerable effort.

Sources: Kato. Lancaster and Tismenetsky. Avrachenkov, Filar and Howlett.

13.1 Kato

We want information about the eigenvalues and eigenvectors of the matrices $A + tB$, where A and B are Hermitian and $t \in \mathbb{R}$. We call these *Hermitian pencils*.

The authorative source from the information we want is Kato's book.[1]. Extracting the information we want is not entirely trivial. I am going to to give a summary of the relevant results but, be warned, what I offer will be paraphrases.

13.1.1 Theorem. *Let A and B be $n \times n$ Hermitian matrices. Then there is an integer m and analytic functions $\theta_1(u),\dots,\theta_m(u)$ such that $\theta_i(\gamma)$ is an eigenvalue of $A+\gamma B$ (for each i). There are corresponding orthogonal projections $F_1(\gamma),\dots,F_m(\gamma)$; these are analytic functions of γ and $F_r(\gamma)$ is the projection onto the $\theta_r(\gamma)$-eigenspace of $A+\gamma B$.*

This theorem is a specialization of Theorem 1.8 on page 70 of Kato. It implies the following.

13.1.2 Lemma. *Let A and B be $n \times n$ Hermitian matrices. Then there is an integer m such that the matrix $A+\gamma B$ has at most m distinct eigenvalues, and there are only finitely many values of γ for which the number of eigenvalues is less than m.*

A warning: If B is Hermitian then $B = LDL^*$ for some unitary matrix L

[1] Perturbation Theory for Linear Operators

and some real diagonal matrix D. Hence

$$A+tB = A+tLDL^* = L(L^*AL+tD)^*$$

and we see that restricting B to be diagonal will not make our lives any easier.

13.2 Basics

The first thing to note is that the zeros of a polynomial are continuous functions of its coefficients. (The map from zeros to coefficients is differentiable and invertible, so we can apply the inverse function theorem.)

If L is unitary then

$$L^*(A+\gamma B)L = LAL^* + \gamma L^*BL$$

and therefore image of a Hermitian pencil under unitary conjugation is a Hermitian pencil. This means that we can always assume that one of the matrices A and B in a Hermitian pencil is diagonal. Unfortunately this does not simplify things, the diagonal case is as hard as the general case. If A and B commute, then we can simultaneously diagonalize them, in which case all difficulties vanish.

An issue here is that if we are working with two matrices A and B, then we will likely be concerned with the structure of the algebra $\langle A, B\rangle$, and this algebra has a distressing tendency to be the full matrix algebra:

13.2.1 Theorem. *Let A be Hermitian $d \times d$ and let z be an element of $\mathbb{C}^d$. If no eigenvector of A is orthogonal to z, then $\langle A, zz^*\rangle = \mathrm{Mat}_{d\times d}(\mathbb{C})$.* □

The condition on A and z is equivalent to assuming that the pair (A, z) is controllable. For details and the proof of the theorem, see GS&CQW.

What if (A, z) is not controllable? If A has spectral decomposition

$$A = \sum_r \theta_r E_r$$

then the non-zero vectors $E_r z$ form an orthogonal basis for the A-module $\mathcal{M}$ generated by z. Since this module contains z, it is actually a module for $\langle A, zz^*\rangle$. The orthogonal complement $cM^\perp$ to $\mathcal{M}$ is spanned by eigenvectors y of A such that $z^*y = 0$, and so zz^* acts on $\mathcal{M}^\perp$ as the zero operator. The module $\mathcal{M}$ itself is irreducible (under $\langle A, zz^*\rangle$), and therefore by a theorem due to Burnside, $\langle A, zz^*\rangle$ acts on it as the full matrix algebra.

If A is the adjacency matrix of a strongly regular graph X and $z = e_a$ for some vertex a, then $\dim \mathcal{M} = 3$, and is known as the *standard module* for the Terwilliger algebra of X.

13.3 Rank-1 Updates

We have

$$\begin{aligned}\phi(A+ww^T,t) = \det(tI-A-ww^T) &= \det(tI-A)\det(I-(tI-A)^{-1}ww^T)\\ &= \phi(A,t)(1-w^T(tI-A)^{-1}w)\end{aligned}$$

and from this we deduce:

13.3.1 Lemma. *If A has spectral decomposition $A=\sum_r \theta_r E_r$, then*

$$\frac{\phi(A+ww^T,t)}{\phi(A,t)} = 1-\sum_r \frac{w^TE_rw}{t-\theta_r}.$$

The *eigenvalue support* of a vector w is the set

$$\{\theta_r : E_rw \neq 0\}.$$

Note that, since E_r is a projection, $E_rw = 0$ if and only if $w^*E_rw = 0$. The size of the eigenvalue support is the number of poles of the rational function $\phi(A(w),t)/\phi(A,t)$.

13.3.2 Theorem. *Let A be a Hermitian matrix, let w be a vector in $\mathbb{C}^n$ and let*

$$\theta_1 > \cdots > \theta_k$$

be the eigenvalue support of w. Then

(a) the eigenvalues of $A+\gamma ww^T$ interlace the eigenvalues of A; more precisely $\theta_1(\gamma)\geq\theta_1$ and if $r>1$, then

$$\theta_{r-1}(0) \geq \theta_r(\gamma) \geq \theta_r(0).$$

(b) The function $\theta_r(\gamma)$ is constant if and only if $\theta_r(0)$ is not in the eigenvalue support of x.

Suppose that $z\neq 0$ and

$$(A+ww^T)z = \lambda z.$$

Then

$$\langle w,z\rangle w = (\lambda I-A)z$$

and, if λ is not an eigenvalue of A, we see that $\langle w,z\rangle \neq 0$ and

$$z = \langle w,z\rangle(\lambda I-A)^{-1}w.$$

Therefore

$$z = \langle w,z\rangle \sum_r \frac{1}{\lambda-\theta_r}E_rw.$$

Adding a loop at a vertex is a special case of a rank-1 update. We recall that

$$\frac{\phi(X\setminus i,t)}{\phi(X,t)} = \sum_r \frac{e_i^TE_re_i}{t-\theta_r}$$

and so from Lemma 13.3.1,

$$\frac{\phi(A+\gamma e_ie_i^T,t)}{\phi(A,t)} = 1-\gamma\frac{\phi(X\setminus i,t)}{\phi(X,t)}.$$

13.4 Commutants

Let A be Hermitian with spectral decomposition $A = \sum_r \theta_r E_r$. Define a map Ψ_A on $\mathrm{Mat}_{n\times n}(\mathbb{C})$ by

$$\Psi(M) := \sum_r E_r M E_r.$$

13.4.1 Theorem. *If A is Hermitian, Ψ is orthogonal projection onto the commutant of A.*

Proof. As $E_r M E_r$ commutes with A, it is immediate that the image of Ψ lies in the commutant of A. If M commutes with A, it commutes with each idempotent E_r and accordingly

$$M = IMI = \sum_{r,s} E_r M E_s.$$

If $r \neq s$, then $E_r M E_s = M E_r E_s = 0$, and therefore the commutant of A is the image of Ψ.

It is also clear that $\Psi^2 = \Psi$, so Ψ is idempotent. Now if $M, N \in \mathrm{Mat}_{n\times n}(\mathbb{C})$, then

$$\begin{aligned}\langle N, \Psi(M)\rangle = \operatorname{tr} N^T \Psi(M) &= \sum_r \operatorname{tr}(N^T E_r M E_r)\\ &= \sum_r \operatorname{tr}(E_r N^T E_r M)\\ &= \langle \Psi(N), M\rangle\end{aligned}$$

and so Ψ is self-adjoint. □

13.5 The Eigenvalues of a Hermitian Pencil

If A is diagonal, then its Schur idempotents are diagonal 01-matrices. If the i-th eigenvalue of A is θ_i and has mutiplicity m_i (for $i = 1, \dots, k$), then the commutant of A consists of the block-diagonal matrices with k blocks, where the i-th block is $m_i \times m_i$. (Hence the dimension of the commutant is $\sum_i m_i^2$.) The orthogonal complement to the commutant consists of the matrices Schur-orthogonal to the block-diagonal matrix

$$J_{m_1} \oplus \cdots \oplus J_{m_k}.$$

If B commutes with A, we can express the eigenvalues of $A + tB$ in terms of the eigenvalues of A and B. To help with determining the eigenvalues of the pencil when A and B do not commute, we describe a more complicated way of getting at the eigenvalues in the commutative case.

Assume A has spectral decomposition

$$A = \sum_r \theta_r E_r$$

If B commutes with A, then each eigenspace of A is B-invariant and therefore has an orthogonal basis formed from eigenvectors of B. Let E be a spectral idempotent of A and assume its rank is m and that the corresponding eigenvalue is θ. There is an $n \times m$ matrix U such that $U^*U = I_m$ and $UU^* = E$; its column space is the eigenspace associated with E. The matrix that represents the restriction of B to $\mathrm{col}(U)$ is U^*BU and, if its eigenvalues are

$$\nu_1, \ldots, \nu_r$$

with respective multiplicities

$$\mu_1, \ldots, \mu_r,$$

the eigenvalues of the restriction of $A+\gamma B$ to $\mathrm{col}(U)$ are

$$\theta + \gamma\nu_1, \cdots, \theta + \gamma\nu_r$$

with multiplicities as above. We will establish very similar expressions in the case where A and B need not commute.

We assume A has spectral decomposition $A = \sum_i \theta_i E_r$, and that θ_i has multiplicity m_i. Assume $E_r = U_r U_r^*$, as before. Let B_0 be the orthogonal projection of B onto the commutant of A and set $B_1 = B - B_0$. Then

$$B_0 = \sum_i E_i B E_i.$$

and

$$E_i B_1 E_i = 0$$

for all i.

We now appeal to Theorem 11.7.1 of Lancaster & Tismenetsky “The Theory of Matrices”, which tells us that if $\sum_i E_i B E_i = 0$, then the linear terms in the series expansions of the eigenvalues of $A + tB$ are zero. Equivalently, the linear terms depend only on the eigenvalues of $\sum_i E_i B E_i$.

13.6 Adding Loops to Strongly Regular Graphs

Let X be a strongly regular graph with parameters $(n, k; a, c)$. We consider the pencil

$$A + \gamma e_i e_i^T.$$

We assume X is primitive[2] and that its eigenvalues are

$$k > \theta > \tau.$$

As X is primitive, k is a simple eigenvalue. We denote the respective multiplicities of θ and τ by m_θ and m_τ. We know that

$$m_\theta = \frac{(n-1)\tau + k}{\tau - \theta}, \quad m_\tau = \frac{(n-1)\theta + k}{\theta - \tau}.$$

[2] X and $\overline{X}$ are connected

We will use ℓ to denote $n-1-k$; this is the valency of the complement of X. We also have explicit formulas for the spectral idempotents of A:

$$E_k = \frac{1}{n}J$$

and

$$E_\theta = \frac{m_\theta}{n}\left(1+\frac{\theta}{k}A-\frac{\theta+1}{\ell}\overline{A}\right), \quad E_\tau = \frac{m_\tau}{n}\left(1+\frac{\tau}{k}A-\frac{\tau+1}{\ell}\overline{A}\right).$$

Hence the eigenvalues of $A+\gamma e_1 e_1^T$ are θ and τ (with multiplicities $m_\theta - 1$ and $m_\theta - 1$ respectively), and the three zeros of the rational function

$$1-\gamma\left(\frac{(E_k)_{1,1}}{t-k}+\frac{(E_\theta)_{1,1}}{t-\theta}+\frac{(E_\tau)_{1,1}}{t-\tau}\right) = 1-\frac{\gamma}{n}\left(\frac{1}{t-k}+\frac{m_\theta}{t-\theta}+\frac{m_\tau}{t-\tau}\right).$$

The walk module generated by e_i has the vectors

$$e_i,\ Ae_i, \overline{A}e_i$$

as an orthogonal basis and is invariant under $e_1 e_1^T$. The matrix representing the action of A on this module is

$$\begin{pmatrix} 0 & k & 0 \\ 1 & a & k-1-a \\ 0 & c & k-c \end{pmatrix}$$

and it follows that $A+\gamma e_1 e_1^T$ is represented[3] by

$$\begin{pmatrix} \gamma & k & 0 \\ 1 & a & k-1-a \\ 0 & c & k-c \end{pmatrix}.$$

Note that the walk-module is a module for $\langle A, e_1 e_1^T\rangle$, and it is irreducible (by Theorem 13.2.1).

13.7 Adding Edges

We work with the pencil

$$A+\gamma B$$

where $B = e_i e_j^T + e_j e_i^T$ (with $i \neq j$). Thus if $i=1$ and $j=2$,

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

13.7.1 Lemma. *If B is a Hermitian matrix and E is a projection, then*

$$n^-(EBE) \leq n^-(B),\ n^+(EBE) \leq n^+(B).$$

[3] as we might expect

Proof. Write $B = B_0 - B_1$ where B_0 and B_1 are positive semidefinite. Then

$$EBE = EB_0E - EB_1E$$

and so

$$n^+(EBE) \leq \mathrm{rk}(EB_0E) \leq \mathrm{rk}(B_0) = n^+(B)$$

and similarly $n^-(EBE) \leq n^-(B)$. □

We note that EBE and $BEE = BE$ have the same non-zero eigenvalues with the same multiplicities. As only the first i-th and j-th rows of BE are not zero, the non-zero eigenvalues of EBE are the eigenvalues of

$$C = \begin{pmatrix} E_{j,i} & E_{j,j} \\ E_{i,i} & E_{i,j} \end{pmatrix}.$$

Since $E = E^T$, we have $E_{i,j} = E_{j,i}$ and the eigenvalues of C are

$$E_{i,j} \pm \sqrt{E_{i,i}E_{j,j}}.$$

If $A = A(X)$ and i and j are cospectral, $E_{i,i} = E_{j,j}$; if i and j are strongly cospectral, then $E_{i,j} = \pm E_{i,i}$ and $\mathrm{rk}(C) \leq 1$. The matrix

$$\begin{pmatrix} E_{i,i} & E_{i,j} \\ E_{j,i} & E_{j,j} \end{pmatrix}$$

is a principal submatrix of E and, with that, is positive semidefinite. Consequently

$$|E_{i,j}|^2 \leq E_{i,i}E_{j,j}$$

and this implies that $\det(C) \leq 0$.

14
Control

We think a linear system as a kind of 'black box'. At time intervals $t = 1, \dots$ it receives an input, returns an output and moves to a new state. The states are elements of its *state space*, the inputs come from an *input space* and the outputs belong to the *output space*. If these elements are represented by vectors $x(i)$, $u(i)$ and $y(i)$ respectively, then they are related by the system of equations

$$\begin{aligned} x(n+1) &= Ax(n) + Bu(n), \\ y(n) &= Cx(n) + Du(n), \end{aligned}$$

for all non-negative integers n. Thus the behaviour of the system is governed by the four matrices A, B, C and D, which we often write as a 2×2 matrix:

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

We will call this the *state-space* description of our system. The state-space matrix need not be square, but A must be.

What we have just described is more usually known as a discrete linear system. Since we will not consider continuous systems at any length, dropping the adjective should not cause problems.

14.1 Buffalos

By way of a first example, we consider a model for the US buffalo population, from J. J. Truxal "Introductory System Engineering", (McGraw-Hill, New York) 1972. In this section we describe the underlying uncontrolled system; in the next section we consider the controlled version.

Let c_i and b_i respectively denote the number of female and male buffalo at the start of year i. We assume that buffalo are mature at age 2, and that each year five percent of the adults die. Female buffalo start breeding at age 2; the number of female calves born in year i is $0.12c_{i-2}$, the number of males is $0.14c_{i-2}$. Thus the population is governed by the two recurrences:

$$\begin{aligned} c_n &= 0.95c_{n-1} + 0.12c_{n-2} \\ b_n &= 0.95b_{n-1} + 0.14c_{n-2} \end{aligned}$$

We analyse the female population. If we define

$$C_n := \begin{pmatrix} c_n \\ c_{n-1} \end{pmatrix},$$

then

$$C_{n+1} = \begin{pmatrix} 0.95 & 0.12 \\ 1 & 0 \end{pmatrix} C_n.$$

Suppose

$$A := \begin{pmatrix} 0.95 & 0.12 \\ 1 & 0 \end{pmatrix}.$$

Then the minimal polynomial of A is

$$t^2 - 0.95t - 0.12,$$

which has distinct roots. Hence we can compute the spectral decomposition of A, with the result that

$$A^n = (1.0629)^n E_1 + (-0.1122)^n E_2,$$

where

$$E_1 = \begin{pmatrix} 0.9040 & 0.1021 \\ 0.8505 & 0.0960 \end{pmatrix}, \qquad E_2 = \begin{pmatrix} 0.0960 & -0.1021 \\ -0.8505 & 0.9040 \end{pmatrix}.$$

(The matrices E_1 and E_2 are idempotent and $E_1E_2 = E_2E_1 = 0$.) From this we learn that, in the long term, the number of female buffalo will increase annually by 6.29%. The actual numbers at the end of year n will be closely approximated by the vector

$$(1.0629)^{n-1}(0.9040c_2 + 0.1021c_1).$$

This shows that the size of the population is sensitive to the initial conditions, even though the growth rate is not.

We now consider the males too. Suppose

$$D_n := \begin{pmatrix} c_n \\ c_{n-1} \\ b_n \end{pmatrix}.$$

Then

$$D_{n+1} = \begin{pmatrix} 0.95 & 0.12 & 0 \\ 1 & 0 & 0 \\ 0 & 0.14 & 0.95 \end{pmatrix} D_n.$$

Here the coefficient matrix is block-triangular, and its minimal polynomial is

$$(t - 0.95)(t^2 - 0.95t - 0.12).$$

(1) Show that the male population grows as a power of 1.0629.

(2) What is the asymptotic ratio of males to females? (It can be determined from an idempotent.)

14.2 Burgers

We continue with the model of the previous section, but we assume that each year a certain number h_n of the adult females are harvested. The equations describing the female population become

$$\begin{aligned} c_n &= 0.95c_{n-1} + 0.12c_{n-2} - h_n \\ b_n &= 0.95b_{n-1} + 0.14c_{n-2} \end{aligned}$$

which we write in matrix form as

$$C_{n+1} = AC_n - h_n \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Let us assume that $h_n = hc_n$, for some constant h. Then we can write the resulting system as

$$C_{n+1} = A(h)C_n,$$

where

$$A(h) = \begin{pmatrix} 0.95 - h & 0.12 \\ 1 & 0 \end{pmatrix}.$$

The minimal polynomial of A_h is

$$t^2 - (0.95 - h)t - 0.12. \tag{14.2.1}$$

Given our model, we must have $0 \le h \le 0.95$. Let θ_h and τ_h denote the eigenvalues of $A(h)$. Then $\theta_h \tau_h = -0.12$, since this is the constant term of the minimal polynomial. It follows that θ_h and τ_h are distinct and therefore $A(h)$ is diagonalizable, for all h. For small values of h, we may assume $\theta_h \approx 1$ and τ_h is small and negative. The population will grow as a power in θ_h, and will be asymptotically constant if and only if $\theta_h = 1$. If this happens, then

$$1 - (0.95 - h) - 0.12 = 0,$$

implying that $h = 0.07$. In this case the eigenvalues are 1 and -0.12, and idempotent corresponding to 1 is

$$\begin{pmatrix} 0.8929 & 0.1071 \\ 0.8929 & 0.1071 \end{pmatrix}.$$

(1) Explain why the female population can stay constant when we harvest 7% of the animals annually, even though the uncontrolled growth rate is only 6.3%.

14.3 Controllability

Consider the linear system given by

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

where A is $n \times n$ and B is $n \times k$. If the initial state of the system is $x(0)$, then we have the equations

$$\begin{aligned} x(1) &= Ax(0) + Bu(0) \\ x(2) &= A^2 x(0) + ABu(0) + Bu(1) \end{aligned}$$

which leads us to the general formula

$$x(n) = A^n x(0) + \sum_{i=1}^{n} A^{n-i} Bu(i-1).$$

Thus the state at time n is the sum of two terms, namely the state of the uncontrolled system at time n and the state of the controlled system with zero initial state. (This decomposition is an important property of linear systems.)

Define the *controllability matrix* to be

$$\mathscr{R} = \begin{pmatrix} B & AB & \cdots & A^{n-1}B \end{pmatrix}.$$

Since A is $n \times n$, its minimal polynomial has degree at most n, and so if $i \geq 0$, then A^{n+i} is a linear combination of

$$I, A, \ldots, A^{n-1}.$$

Therefore the column space of $\mathscr{R}$ is the sum of the subspaces $A^r \operatorname{col}(B)$, where $0 \leq r < n$. It follows that if our initial state is zero, then the state of the system is always an element of $\operatorname{col}(CM)$.

We say the pair (A, B) is *controllable* if, given any vector v in $\mathbb{F}^n$ and starting with $x(0) = 0$, we can choose inputs $u(0), u(1), \ldots, u(n-1)$ so that $x(n) = v$. We will call the system itself controllable if (A, B) is.

14.3.1 Theorem. *For a linear system, the following are equivalent:*

(a) The pair (A, B) is controllable.

(b) The rows of the controllability matrix are linearly independent.

(c) The only A-invariant subspace that contains $\operatorname{col}(B)$ is $\mathbb{R}^n$.

(d) No non-zero subspace of $\ker(B^T)$ is A^T-invariant.

Proof. By the previous lemma, (a) and (b) are equivalent. The column space of the controllability matrix is the smallest A-invariant subspace that contains the columns of B, hence (b) holds if and only if (c) holds. We show that (c) and (d) are equivalent too.

Suppose $\mathrm{rk}(\mathscr{R}) < n$. Then there is a non-zero vector f such that $f^T\mathscr{R} = 0$, and so

$$f^T A^r B = 0, \quad r = 0, 1, \ldots, n-1.$$

Consequently $f^T A^r B = 0$ for all non-negative r, and therefore the A^T-invariant subspace generated by f lies in $\ker(B^T)$.

Conversely, if the A^T-invariant subspace generated by the non-zero vector f in $\ker(B^T)$ is contained in $\ker(B^T)$, that $f^T A^r B = 0$ for all r, and $\mathrm{rk}(\mathscr{R}) < n$. □

14.3.2 Corollary. *If B is $n \times 1$ and (A, B) is controllable, then the minimal polynomial of A has degree n.* □

There is another concept related to controllability, sometimes called *controllability to the origin*. Suppose our system starts in some state $x(0)$ and we wish to know if there is a sequence of inputs which will drive it to the zero state.

Now the state at time r will be

$$A^r x(0) + \sum_{i=1}^{n} A^{r-i} Bu(i).$$

Since

$$\sum_{i=1}^{n} A^{r-i} Bu(i) \in \mathrm{col}(\mathscr{R}),$$

we see that if there is a sequence of inputs that takes the state to zero in r steps, then $A^r x(0)$ must lie in $\mathrm{col}(\mathscr{R})$. If $r \geq n$ and $A^r x(0) \in \mathrm{col})\mathscr{R})$, then there is a sequence of inputs of length r that sends the system to zero.

Thus we see, for example, that if $\mathrm{rk}\mathscr{R} = n$, then we can bring the system to rest in n steps. To be more precise, we investigate the range of A. We note that

$$\mathrm{col}(A^r)$$

is a nested sequence of A-invariant subspaces which is first strictly deceasing, then constant. Since $\dim(\mathrm{col}(A)) \leq n$, it follows that when $r \geq n$,

$$\mathrm{col}(A^r) = \mathrm{col}(A^n).$$

We conclude that our system can be brought to rest in n steps if and only if $A^n x(0) \in \mathrm{col}(\mathscr{R})$. Further it can be brought to rest in n steps no matter what the initial state is, if and only if

$$\mathrm{col}(A^n) \subseteq \mathrm{col}\mathscr{R}.$$

We conclude that any controllable system is controllable to the origin, but the latter condition is weaker.

(1) Show that (A, B) is controllable if and only if $\mathscr{R}$ has a right inverse.

(2) Show that the column space of $\mathscr{R}$ is the smallest A-invariant subspace of $\mathbb{F}^n$ that contains the columns of B.

(3) If A is invertible, show that (A, B) is controllable to the origin if and only if it is controllable.

14.4 Observability

Consider the linear system given by the matrix

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

We consider the problem of determining the initial state from the observed values of y. We have

$$x(r+1) = Ax(r) + Bu(r), \qquad y(r) = Cx(r) + Du(r).$$

Since we know the values of the input vectors $u(r)$, our problem reduces to that of reconstructing $x(0)$ from the vectors $Cx(r)$. Now

$$x(r+1) = A^r x(0) + \sum_{i=1}^{r} A^{r-i} Bu(i);$$

since the vectors $A^{r-i}Bu(i)$ are known, the final form of our problem is to reconstruct $x(0)$ from the sequence $CA^r x(0)$ for $r = 0, 1, \ldots$. Since A is $n \times n$, it follows that the first n values of this sequence determine the rest.

We say that the pair (C, A) is *observable* if the sequence

$$Cx, CAx, \ldots, CA^{n-1}x$$

determines x (in all cases). The system itself is observable if (C, A) is. Define the *observability matrix* $\mathscr{O}$ by

$$\mathscr{O} := \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix}.$$

14.4.1 Theorem. *The pair* (C, A) *is observable if and only if the columns of the observability matrix are linearly independent.*

Proof. If the columns of $\mathscr{O}$ are linearly independent, then it has a left inverse N. So $N\mathscr{O}x = x$, and thus we recover x. □

14.4.2 Corollary. *The pair* (C, A) *is observable if and only if* (A^T, C^T) *is controllable.* □

This implies for example, that (C, A) is observable if and only if no subspace of $\ker(C)$ is A-invariant.

14.5 Feedback and Controllability

Consider the system

$$x(n+1) = Ax(n) + Bu(n).$$

If we take u to be given by

$$u(n) = Kx(n) + v(n),$$

then our system becomes

$$x(n+1) = (A+BK)x(n) + Bv(n).$$

The $Kx(n)$ term is called *feedback*; the behaviour of the original system is governed by A, the behaviour of the system with feedback is governed by the matrix $A+BK$. We show that feedback does not effect controllability.

14.5.1 Lemma. *Suppose A is $n \times n$ and B is $n \times k$. Then for any $k \times n$ matrix K, the pair (A,B) is controllable if and only if $(A+BK,B)$ is.*

Proof. We show that $\operatorname{col}(\mathscr{R}(A+BK,B) \subseteq \operatorname{col}(\mathscr{R}(A,B)$. Since

$$A = (A+BK) + B(-K),$$

it follows that these two column spaces are equal.

If $v \in \operatorname{col}(\mathscr{R})$, then $Av \in \operatorname{col}(\mathscr{R})$ and

$$BKv \in \operatorname{col}(B) \subseteq \operatorname{col}(\mathscr{R}),$$

whence $(A+BK)v \in \operatorname{col}(\mathscr{R})$. It follows that $\operatorname{col}(\mathscr{R})$ is an $(A+BK)$-invariant subspace that contains $\operatorname{col}(B)$, and therefore it contains the column space of $\mathscr{R}(A+BK,B)$. □

14.5.2 Lemma. *If (A,B) is controllable and b is a non-zero column of B, then there is a matrix K such that $(A+BK,b)$ is controllable.*

Proof. Assume A is $n \times n$ and that (A,B) is controllable. We aim first to find columns $b_1,\dots,b_k$ of B and integers $r_1,\dots,r_k$, such that the union of the sets

$$S(b_i,r_i) := \{b_i, Ab_i, \dots, A^{r_i-1}b_i\}$$

is a basis for $\mathbb{F}^n$. This is straightforward. Choose b_1 equal to b and choose r_1 to be the greatest integer such that the vectors

$$b_1, Ab_1, \dots, A^{r_1-1}b_1$$

are linearly independent. Next, assume inductively that we have found $b_1,\dots,b_{j-1}$ and $r_1,\dots,r_{j-1}$ such that

$$\bigcup_{i<j} S(b_i,r_i)$$

is linearly independent. The span of this set of vectors in A-invariant and so, if this set contains fewer than n vectors, there must be a column of B which it does not contain. Take b_j to be such a vector, and let r_j be the greatest integer such that the span of $S(b_j, r_j)$ contains no non-zero vectors from the span of the above union.

There is a unique linear mapping $\mathscr{L}$ such that

$$\mathscr{L}(A^j b_i) = \begin{cases} b_{i+1}, & \text{if } j = r_i - 1; \\ 0, & \text{otherwise.} \end{cases}$$

Let L be the matrix representing $\mathscr{L}$. We claim that the vectors

$$b, (A+L)b, \ldots, (A+L)^{n-1} b$$

are linearly independent.

Since $LA^i b_1 = 0$ if $i < r_1 - 1$ and $LA^{r_1-1} b_1 = b_2$, we see that if $i > 1$, then

$$(A+L)^{r_1-i} b = A^{r_1-i} b = A^{r_1-i} b_1$$

and

$$(A+L)^{r_1} b = A^{r_1} b_1 + b_2.$$

Starting from this, a reasonably easy induction argument, which we omit, shows that the span of the m vectors

$$(A+L)^i b, \qquad i = 0, 1, \ldots, m-1$$

is equal to the span of the first m vectors from

$$S(b_1, r_1) \cup \cdots \cup S(b_k, r_k).$$

This proves our claim.

To complete the proof, we note that the image of $\mathscr{L}$ is spanned by columns of B, and therefore there is a matrix K such that $L = BK$. □

14.5.3 Corollary. *Let b be a non-zero column of B. The pair (A, B) is controllable if and only if there is a matrix K such that $(A + BK, b)$ is controllable.*

Proof. The previous lemma shows that if (A, B) is controllable and b is a non-zero column of B, then there is a matrix K such that $(A + BK, b)$ is controllable. For the converse we note that if $(A + BK, b)$ is controllable, then certainly $(A + BK, B)$ is controllable. By Lemma 14.5.1, this implies that (A, B) is controllable. □

(1) Let b be a non-zero element of $\operatorname{col} B$. Show that (A, B) is controllable if and only if there is a matrix K such that $(A + BK, b)$ is controllable.

14.6 Canonical Forms

We consider first the general system

$$\begin{aligned} x(n+1) &= Ax(n) + Bu(n), \\ y(n) &= Cx(n) + Du(n). \end{aligned}$$

Suppose M is invertible and $x(n) = Mz(n)$ for all non-negative n. Then we rewrite our system as

$$\begin{aligned} z(n+1) &= M^{-1}AMz(n) + M^{-1}Bu(n), \\ y(n) &= MCz(n) + Du(n). \end{aligned}$$

These two systems correspond respectively to the block matrices

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}, \qquad \begin{pmatrix} M^{-1}AM & M^{-1}B \\ MC & D \end{pmatrix}.$$

We say two systems related in this way are *equivalent*. We will also say that the pairs (A, B) and $(M^{-1}AM, M^{-1}B)$ are equivalent.

We now confine ourselves to the single-input case, where B is $n \times 1$. Suppose A is $n \times n$ and set $M = \mathscr{R}(A, b)$. Then

$$AM = \begin{pmatrix} Ab & A^2b & \cdots A^nb \end{pmatrix} = MP,$$

where F is the companion matrix of ψ_b, the minimal polynomial of A relative to b. If (A, b) is controllable, then $\operatorname{rk}(M) = n$, and so M is invertible. It follows that $M^{-1}AM = F$. Since $\mathscr{R}(A, b)e_1 = b$, we also find that $M^{-1}B = e_1$. We conclude that if a pair (A, b) is controllable, then our original system is equivalent to the system

$$\begin{pmatrix} F & e_1 \\ CM & D \end{pmatrix}.$$

where F is the companion matrix of the minimal polynomial of A. If we also have a single output, that is, if C is $1 \times n$, then $C = c^T$ and

$$c^TM = \begin{pmatrix} c^Tb & c^TAb & \cdots & c^TA^{n-1}b \end{pmatrix}.$$

It follows that our system is determined by the minimal polynomial of A and the entries of this vector.

From **??**, we know that if F is a companion matrix of order $n \times n$, there is an symmetric invertible matrix Q such that $Q^{-1}FQ = F^T$. We see that $Qe_n = e_1$, and therefore the pair (F, e_1) is equivalent to the pair (F^T, e_n). The pairs (C, e_1) and (C^T, e_n) are called the *controllability canonical forms* of the pair (A, b).

There are analogous canonical forms for observable pairs (c^T, A), but these can be deduced from our work above, applied to the controllable pair (A^T, c).

14.7 Eigenvalues and Controllability

In this section our matrices are real matrices, but our subspaces may be complex. (For example, eigenspaces.)

14.7.1 Lemma. *The pair (A,B) is controllable if and only if the rows of $\begin{pmatrix} A-\lambda I & B\end{pmatrix}$ are linearly independent for all complex numbers λ.*

Proof. First suppose $z \neq 0$ and

$$z^* \begin{pmatrix} A-\lambda I & B\end{pmatrix} = 0. \tag{14.7.1}$$

Then $z^* A^r = \lambda^r z^*$ and $z^* B = 0$, so $z^* \mathscr{R} = 0$. Hence the rows of the controllability matrix are linearly dependent, and therefore (A,B) is not controllable.

On the other hand if (A,B) is not controllable, then by Theorem 14.3.1 there is an A^T-invariant subspace of $\ker(B^T)$, and this subspace must contain an eigenvector z of A^T. If the eigenvalue belonging to z is λ, then (14.7.1) is satisfied. □

The *spectrum* of a matrix is the multiset formed by its eigenvalues, and their algebraic multiplicities. The spectrum of a real-matrix is conjugate closed—if θ is an eigenvalue, then its complex conjugate $\bar{\theta}$ is an eigenvalue with the same algebraic multiplicity.

14.7.2 Theorem. *Let A be an $n \times n$ real matrix. The pair (A,B) is controllable if and only each conjugate-closed multiset of complex numbers with size n occurs as the spectrum of some matrix $A+BK$.*

Proof. Assume first that we can choose K so that $A+BK$ has any given conjugate-closed set of complex numbers as its eigenvalues.

Suppose there is a vector z such that

$$z^T \begin{pmatrix} B & AB & \cdots & A^{n-1}B\end{pmatrix} = 0.$$

Then, for all r and any K,

$$z^T(A+BK)^r = z^T A^r$$

and therefore

$$z^T[(A+BK_0)^r - (A+BK_1)^r] = 0.$$

Choose K_0 so that all eigenvalues of $A+BK_0$ lie inside the unit circle, and choose K_1 so that the eigenvalues of $A+BK_1$ are the distinct n-th roots of unity. Then

$$(A+BK_1)^{ns} = I$$

for all non-negative integers s, while

$$(A+BK_0)^{ns} \to 0$$

as $s \to \infty$. It follows that $z = 0$, whence the rows of $\mathscr{R}(A,B)$ are linearly independent, and (A,B) is controllable.

We turn to the converse. We first prove the result holds in the single-input case. Suppose (A,b) is a controllable pair. We work with the equivalent canonical form (F^T, e_n), where F^T is the transpose of the companion matrix of the minimal polynomial of A. If K is a $1 \times n$ matrix, then $e_n K$ is an $n \times n$ matrix with its first $n-1$ rows zero, and with last row equal to K. Therefore $F^T + e_n K$ is also the transpose of a companion matrix. By varying our choice of K, we can arrange to the last row of $F^T + e_n K$ to be any desired vector, and so force $F^T + e_n K$ to have any desired conjugate-closed set of complex numbers as its eigenvalues.

We consider the general case. Suppose (A,B) is controllable and b is a non-zero column of B. Then by Lemma 14.5.2 there is a matrix K such that $(A+BK, b)$ is controllable. By what we have just proved, for each conjugate-closed set of complex numbers, there is a $1 \times n$ matrix K_1 such that

$$A + BK + bK_1$$

has this set as its eigenvalues. But $b = Be_r$ for some r, and so

$$BK + bK_1 = BK + Be_r K_1 = B(K + e_r K_1),$$

and so our result is proved. □

14.8 Observers

Consider the discrete dynamical system given by the equations

$$\begin{aligned} x(n+1) &= Ax(n) + Bu(n) \\ y(n) &= Cx(n) + Du(n). \end{aligned}$$

We want to construct a second system which will accept both the input and the output of the first system as its inputs, and as produce as its own output at least an approximation to the state of our first system. To construct such a system, we consider a second system based on the one above:

$$\begin{aligned} \hat{x}(n+1) &= A\hat{x}(n) + Bu(n) + L(y(n) - \hat{y}(n)) \\ \hat{y}(n) &= C\hat{x}(n) + Du(n). \end{aligned}$$

If this system has the property that $x(n) - \hat{x}(n) \to 0$ as $n \to \infty$, we call it an *asymptotic observer*. If A is $m \times m$ and $x(n) = \hat{x}(n)$ when $n > m$, we call it an *exact* observer. The choice of L is up to us. We calculate

$$\begin{aligned} x(n+1) - \hat{x}(n+1) &= A(x(n) - \hat{x}(n)) - L(y(n) - \hat{y}(n)) \\ &= A(x(n) - \hat{x}(n)) - L(Cx(n) - C\hat{x}(n)) \\ &= (A - LC)(x(n) - \hat{x}(n)). \end{aligned}$$

Now the initial difference $x(0)-\hat{x}(0)$ can be any vector, so we conclude that we have an asymptotic observer if and only if $A-LC$ is a contraction, and an exact observer if and only if $A-LC$ is nilpotent.

If A is $n\times n$ and B is $n\times m$, we say that the pair (A,B) is *stabilizable* if there is an $m\times n$ matrix K such that all eigenvalues of $A+BK$ lie inside the unit circle. (In other terms, $A+BK$ is a contraction.) Every controllable pair is stabilizable. If C is $\ell\times n$, we say that (C,A) is *detectable* if (A^T,C^T) is stabilizable or, equivalently, if there is a matrix L such that $A+LC$ is a contraction.

14.8.1 Theorem. *An asymptotic observer exists if and only if* (C,A) *is detectable. An observer exists if and only if* (C,A) *is observable.* □

14.9 Transfer Matrices

We introduce a very important tool in the study of discrete dynamical systems: transfer matrices.

We first present this in a special case, coming from coding theory. We suppose that a sequence $(u_i)_{i\geq 0}$ of binary vectors is encoded by a device as a second sequence $(y_i)_{i\geq 0}$ of binary vectors. In the simplest case, we have a matrix D and u_i is mapped to Du_i. But we are going to assume that our device has a state x_i (another binary vector) and that y_i is computed according to the system

$$x_{i+1} = Ax_i + Bu_i \tag{14.9.1}$$

$$y_i = Cx_i + Du_i. \tag{14.9.2}$$

Here A, B, C and D are binary matrices and A is square. (For a coding theorist it might be natural to assume D is $n\times k$; the matrix A is square.) The first problem that arises is to reconstruct the inputs u_i given the outputs y_i (and the four matrices A, B, C, D. In the real applications, the vectors y_i are corrupted by noise, and we also have the harder task of first determining the uncorrupted values of the outputs.

We say that the system described by the four matrices is a *convolutional encoder*. The space of possible output sequences is a *convolutional code*. Convolutional codes are important in practice.

To make further progress, we introduce generating functions. A convolutional encoder takes an input sequence

$$u_0,u_1,u_2,\ldots$$

and converts it to an output sequence

$$y_0,y_1,y_2,\ldots.$$

In any practical situation, the vectors u_i will be zero for all sufficiently large i, but we defer imposing this as a requirement. One standard way to deal with infinite sequences is to encode them as formal power series, and so we define

$$U(z) := \sum_{i\ge0} z^{-i} u_i, \qquad Y(t) := \sum_{i\ge0} z^{-i} y_i.$$

These can be viewed as formal power series in the variable z^{-1} with vectors as coefficients, or as vectors whose entries are formal power series of $\mathbb{F}$. (We tend to prefer the latter view.) We say that $U(z)$ is a generating function for the sequence $(u_i)_{i\ge0}$.

Next we assume that $x_0 = 0$ and introduce the generating function $X(t)$. The defining equations for our encoder give us

$$zX(z) = AX(z) + BU(z), \quad Y(t) = CX(z) + DU(z),$$

and consequently

$$Y(z) = (D + C(zI - A)^{-1}B)U(z).$$

It follows that our encoder is completely specified by the proper rational matrix

$$G(z) := D + C(zI - A)^{-1}B.$$

If we have a discrete dynamical system over a field $\mathbb{F}$, given by the matrix

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \tag{14.9.3}$$

we define the *transfer matrix* of the system to be

$$D + C(zI - A)^{-1}B.$$

The transfer matrix completely determines the response of our system to a given input sequence (given that $x_0 = 0$). If e_i denotes the i-th standard basis vector, then the generating function of the output sequence corresponding to the input sequence

$$e_i, 0, 0, \ldots$$

is the i-th column of $G(z)$. This provides a very natural interpretation of the columns of $G(z)$, and shows that we can find the transfer matrix of a system by determining its response to each of the above input sequences. In particular it is not unusual to be given the transfer matrix of a system, rather than the state-space description.

It may seem more natural to use formal power series in z rather than z^{-1}, but the above choice is standard in control theory.

Part III

Convexity

15
Norms

15.1 Convexity

We work over $\mathbb{R}$. We say that a vector v is an *affine combination* of vectors $x_1, \dots, x_n$ if

$$v = \sum_i a_i x_i$$

and $\sum a_i = 1$. An affine combination is *proper* if it has at least two non-zero coefficients. The set of all affine combinations of a set of vectors is the *affine hull* of the set. The affine hull of x is x itself. The affine hull of $\{x, y\}$ (where $x \neq y$) is

$$\{tx + (1-t)y : t \in \mathbb{R}\}.$$

Geometrically this set is the unique line passing through the points represented by x and y. Note that this line contains 0 if and only if x and y are linearly dependent.

If U is a subspace of V and then a *coset* of U is a set of the form

$$\{a + u : u \in U\},$$

for some a in V.

15.1.1 Lemma. *The affine hull of a set of vectors* $\{x_1, \dots, x_m\}$ *is a coset of the subspace spanned by* $x_2 - x_1, \dots, x_m - x_1$. □

An *affine subspace* is a set S that is closed under affine combinations.

We say that vectors $x_1, \dots, x_m$ are *affinely dependent* if there are scalars a_i, not all zero, such that

$$\sum_i a_i = 0, \qquad \sum_i a_i x_i = 0.$$

If a set is not affinely dependent, it is *affinely independent*. Note that any single vector, including the zero vector, is affinely independent.

A vector v is a *convex combination* of vectors $x_1, \dots, x_m$ if there are scalars $a_1, \dots, a_m$ such that

$$\sum_i a_i = 1, \qquad a_i \ge 0 \ (i = 1, \dots, m)$$

and

$$v = \sum a_i x_i.$$

Thus a convex combination is a non-negative affine combination. A convex combination is *proper* if its has at least two non-zero coefficients. The *convex hull* of a subset S is the set of all convex combinations of elements of S. A set S is *convex* if any convex combinations of its elements is contained in S, that is, if S is equal to its convex hull.

The convex hull of two distinct vectors consists of the line segment that joins them. Hence a set S is convex if, whenever x and y belong to S, so do all points on the line segment joining them. We also see that the intersection of two convex sets is convex.

A real-valued function f on $\mathbb{R}^n$ is *convex* if

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y), \qquad 0 \le t \le 1.$$

(1) If $a \in \mathbb{R}^n$, show that $f(x) := \exp(a^T x)$ is a convex function.

(2) Show that set of positive semidefinite matrices is the convex hull of the matrices with rank 1.

(3) Suppose $a_i \ge 0$ and $\sum_i a_i = 1$. If f is convex, prove that

$$f\left(\sum_i a_i x_i\right) \le \sum_i a_i f(x_i).$$

(4) Use the result of the previous exercise with $f(x) = x^p$ $(p > 1)$ to show that

$$\sum_i |x_i y_i| \le \left(|x_i|^p\right)^{1/p} \left(|y_i|^q\right)^{1/q},$$

where $1/p + 1/q = 1$. (This is Hölder's inequality.)

15.2 Extreme Points

Let C be a convex set. A point x is C is extreme if it cannot be expressed as the convex combination of points in $C \setminus x$. The extreme points of a line segment are its endpoints. Suppose C is convex and $x \in C$. Let ℓ be a line through x. Then $\ell \cap C$ is a line segment. The interior points of this line segment are not extreme. A closed convex set is the convex hull of its extreme points. We will not prove this, but we consider two cases that will be useful.

15.2.1 Lemma. *Let S be the set of vectors x in $\mathbb{R}^n$ such that $|x_i| \le 1$ for all i. Then S is the convex hull of the vectors with all entries ± 1.*

Proof. It is easy to verify that S is convex, we leave this as an exercise. We show that it is the convex hull of the ±1-vectors.

We prove this by induction on n, asserting that it is trivial when $n = 1$. Assume $v \in S$ and that $v_1 = 1$. Let v' be the vector we get by deleting the first entry of v. Then v' lies in the set of vectors x in $\mathbb{R}^{n-1}$ such that $|x_i| \le 1$, and so by induction it is a convex combination of the ±1-vectors in $\mathbb{R}^{n-1}$. It follows that v is a convex combination of those ±1-vectors in $\mathbb{R}^n$ with first entry equal to 1. If $v_1 = -1$, then $-v'$ is a convex combination of ±1-vectors $x_1, \dots, x_m$, and so v' is a convex combination of the vectors $-x_1, \dots, -x_m$, but these are ±1-vectors too. It follows that if $|v_i| = 1$, then v is a convex combination of ±1-vectors.

Now suppose that $|v_i| < 1$ for all i. Let v^+ be the vector such that $(v^+)_i = 1$ if $v_i \ge 0$ and $(v^+)_i = -1$ if $v_i < 0$. Then

$$((1-t)v^+ + tv)_i = \begin{cases} 1 - t + tv_i, & \text{if } v_i \ge 0; \\ t - 1 + tv_i, & \text{otherwise.} \end{cases}$$

from which we eventually deduce that $w = (1-t)v^+ + tv \in S$ provided

$$0 \le t \le \frac{2}{1 - |v_i|}.$$

Choose t so that $t = 2/(1 - |v_j|)$ for some j. Then $|w_j| = 1$, and therefore v is a convex combination of v^+ and w. Since $|w_j| = 1$, it is the convex combination of ±1-vectors, and therefore v is too. □

15.2.2 Lemma. *Let S be the set of vectors x such that*

$$\sum_i |x_i| \le 1.$$

Then S is the convex hull of the vectors $\pm e_i$ for $i = 1, \dots, n$. □

(1) Show that if x is a proper convex combination of points from C, it is the proper convex combination of two points.

(2) Let C be a convex set and let f be a convex function. If the point x_0 in C maximizes the value of f, show that it is an extreme point.

(3) Prove (**??**).

15.3 Norms

Let V be a vector space over $\mathbb{F}$, where $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$. A *norm* on V is a function from V to $\mathbb{R}$, whose value on x is written $\|x\|$, such that

(1) $\|x\| \ge 0$ and $\|x\| = 0$ if and only if $x = 0$.

(2) If $c \in \mathbb{R}$, then $\|cx\| = |c|\,\|x\|$.

(3) If $\|x+y\| \le \|x\| + \|y\|$.

The third axiom is called the *triangle inequality*. It implies that any norm is a convex function on V. The set

$$\{x : \|x\| \le 1\}$$

is called the *unit ball* of the norm, but it need not be very round.

We consider some examples over $\mathbb{R}$. If we have an inner product on V, then we can define a norm by

$$\|x\| := \sqrt{\langle x, x\rangle}$$

The only difficulty here is to verify the triangle inequality. We note that

$$\|x + ty\|^2 = \langle x + ty, x + ty\rangle = \langle x,,\rangle x + 2\langle x, y\rangle t + \langle y, y\rangle t^2.$$

This is a quadratic in t which is non-negative for all t, and consequently

$$\langle x, y\rangle^2 - \langle x, x\rangle\langle y, y\rangle \le 0,$$

which is usually called the Cauchy-Schwarz inequality. It follows that

$$\begin{aligned}\langle x, x\rangle + 2\langle x, y\rangle t + \langle y, y\rangle t^2 &\le \langle x, x\rangle + 2\langle x, y\rangle t + \langle y, y\rangle t^2 \\ &\le \langle x, x\rangle + 2\|x\|\,\|y\| t + \langle y, y\rangle t^2 \\ &= (\|x\| + t\|y\|)^2.\end{aligned}$$

We conclude that $\|x\| + ty \le \|x\| + \|ty\|$, which yields the triangle inequality.

If our inner product is the dot product our norm is the usual Euclidean norm or ℓ_2-norm and is denoted by $\|\cdot\|_2$ or, sometimes, by $\|\cdot\|$. The unit ball for the Euclidean norm is the unit ball.

If $\langle\cdot,\cdot\rangle$ is a complex inner product, the function

$$\sqrt{\langle x, x\rangle}$$

is a norm. Note that $\langle x, x\rangle$ is guaranteed to be real and non-negative.

Once we have a norm, we can declare that a sequence $x_0, x_1, \ldots$ of vectors *converges* to x if the sequence of real numbers

$$\|x - x_0\|, \|x - x_1\|, \ldots$$

converges to 0. It is a somewhat surprising fact that if a sequence of vectors in a finite-dimensional vector space converges with respect to one norm, then it converges with respect to all. (This is false if the dimension is infinite, as the exercises show.)

(1) Prove that a norm is a convex function.

(2) Let $V = C[0,1]$, the space of continuous functions on the interval $[0,1]$. If $f \in V$, let $\|f\|$ be the norm asssociated with the inner product

$$\langle f,g\rangle := \int_0^1 f(x)g(x)\,dx$$

and let $\|f\|_\infty$ be the norm defined by

$$\|f\|_\infty = \max\{f(x) : x \in [0,1]\}.$$

(You may prove that this is a norm.) Define

$$g_r(x) := (4x(1-x))^r.$$

Prove that $\|g_r\| \to 0$ as $r \to \infty$, but $\|g_r\|_\infty = 1$ for all r.

15.4 Dual Norms

We introduce two further norms. We define $\|x\|_1$ by

$$\|x\|_1 := \max_i \sum_i |x_i|$$

and $\|x\|_\infty$ by

$$\|x\|_\infty := \max_i |x_i|.$$

These are known respectively as the ℓ_1 and ℓ_∞-norms on $\mathbb{R}^n$. As we saw in the previous section, the unit ball for the ℓ_1-norm is the convex hull of the vectors $\pm e_i$ and the unit ball for the ℓ_∞-norm is the convex hull of the ± 1 vectors. (These definitions work over both $\mathbb{R}$ and $\mathbb{C}$, we will only use them over $\mathbb{R}$ though.)

If $\|\cdot\|$ is a norm, we define the *dual norm* $\|\cdot\|^*$ by

$$\|a\|^* := \max_{\|x\|=1} x^T a.$$

We leave the proof that this is a norm as an exercise. As another exercise, we leave you to prove that $\|x\|^{**} = \|x\|$, for any x.

By way of example, we determine the dual of the ℓ_∞-norm. Our problem is compute the maximum value of the function $x^T a$ over the vectors x in the unit ball of the ℓ_∞-norm. This is linear in x, and hence convex; therefore its maximum value occurs at an extreme point of this ball. By Lemma 15.2.1, the extreme points are the ± 1-vectors and hence $\|a\|_\infty^*$ is equal to the maximum value of $x^T a$, as x ranges over the set of ± 1-vectors. Clearly this maximum is realized when $x_i a_i > 0$ for each i, and therefore

$$\|a\|_\infty^* = \sum_i |a_i| = \|a\|_1.$$

(1) Let V be the Euclidean space $\mathbb{R}^n$. Determine the largest C and the smallest D such that

$$C\|x\|_\infty \le \|x\| \le D\|x\|_\infty.$$

(2) If the function $\|\cdot\|^*$ is defined on $\mathbb{R}^n$ by

$$\|y\|^* = \max_{\|x\|=1} x^T y,$$

show that it is a norm.

(3) Prove that $\|x\|^{**} = \|x\|$, for any x.

(4) Prove that $y^T x \leq \|x\| \, \|y\|^*$, and show that this bound is tight.

(5) Show that the ℓ_1-norm is dual to the ℓ_∞-norm, and vice versa.

15.5 Matrix Norms

Let $\mathscr{B}$ be an algebra over the reals. A norm on $\mathscr{B}$ is a function $\|\cdot\|$ from $\mathscr{B}$ to $\mathbb{R}$ that is norm, when we view $\mathscr{B}$ as a vector space, and in addition satisfies:

$$\|AB\| \leq \|A\| \, \|B\|.$$

If $\|\cdot\|$ is a norm on a vector space V then the unit ball

$$\{x \in V : \|x\| \leq 1\}$$

is a closed convex set. If $\|\cdot\|$ is a norm on an algebra then the unit ball must be closed under multiplication, hence forms a semigroup.

Now suppose $\|\cdot\|$ is a norm on $L(V)$, viewed as a vector space. The unit ball is compact and so, if $A \in L(V)$ then there is a constant γ_A such that, if $\|X\| \leq 1$,

$$\|AX\| \leq \gamma_A.$$

If we define γ to be the maximum value of γ_A, where $\|A\| \leq 1$, then

$$\|AB\| = \|A\| \, \|B\| \gamma.$$

From this it follows that $\gamma^{-1}\|\cdot\|$ is a norm on $L(V)$, viewed as an algebra. We will refer to a norm on an algebra as an *operator norm* or *matrix norm*, according as the elements of our algebra are linear mappings or matrices.

Let V be a normed vector space, with norm $\| \cdot \|$. If T is an endomorphism of V, we define the *induced norm* of T by

$$\|T\| = \max\{\|Tx\| : \|x\| = 1\}.$$

Equivalently, it is the maximum value of $\|Tx\|/\|x\|$, for all non-zero vectors x in V. It is straightforward to verify that this is a norm on $L(V)$, with the useful properties:

$$\|Tx\| \leq \|T\| \, \|x\|$$

and

$$\|ST\| \leq \|S\| \, \|T\|.$$

Unless explicitly stated otherwise, we use the same symbol to denote a norm on $\mathbb{R}^n$ and the norm it induces on $n \times n$ matrices. If $\|\cdot\|$ is an induced norm, then $\|I\| = 1$.

If $\|\cdot\|$ is an induced norm then for any matrix A and vector x, we have the very useful inequality:

$$\|Ax\| \le \|A\|\,\|x\|.$$

If $\|\cdot\|_a$ and $\|\cdot\|_b$ are any two norms on a vector space, we say that $\|\cdot\|_b$ *dominates* $\|\cdot\|_a$ if, for all v in V,

$$\|v\|_a \le \|v\|_b.$$

A norm is *minimal* if it does not dominate any other norm. Generally minimal norms are more useful than general norms.

15.5.1 Lemma. *Every matrix norm dominates an induced norm.*

Proof. Suppose $\|\cdot\|$ is a matrix norm. We use this to construct a norm on $\mathbb{R}^n$ whose induced norm is dominated by $\|\cdot\|$.

Let a be a fixed non-zero vector in $\mathbb{R}^n$. We define $\|\cdot\|_a$ by

$$\|b\|_a := \|ba^T\|.$$

Then

$$\|Ax\|_a = \|Axa^T\| \le \|A\|\,\|x\|_a$$

and the matrix norm induced by $\|\cdot\|_a$ is dominated by $\|\cdot\|$. □

15.5.2 Theorem. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ with dual norm $\|\cdot\|^*$. If A is a square matrix then $\|A\|^* = \|A^T\|$.*

Proof. We have

$$\|Ax\|^* = \max_{\|y\|=1} y^T Ax = \max_{\|y\|=1} x^T A^T y$$

and so

$$\|A\|^* = \max_{\|x\|^*=1} \max_{\|y\|=1} x^T A^T y.$$

Now

$$\max_{\|x\|^*=1} x^T b = \|b\|^{**} = \|b\|$$

and consequently

$$\|A\|^* = \max_{\|y\|=1} \|A^T y\| = \|A^T\|.$$

In the sequel any norm we use on matrices will be a matrix norm. If $(A_n)_{n\ge0}$ is a sequence of matrices and we write that $A_n \to 0$, we mean that $\|A_n\| \to 0$, for some norm $\|\cdot\|$.

(1) Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, and let $\|\cdot\|$ also denote the induced matrix norm. Prove that $\|ab^T\| = \|a\|\,\|b\|^*$ and hence that $b^T a \le \|ab^T\|$.

(2) Prove that if $n \ge 1$, then $\|A^n\|^{1/n} \le \|A\|$.

15.6 Examples

The *Euclidean* or *trace* norm of a matrix is the norm associated with the inner product

$$\langle A, B\rangle := \operatorname{tr} A^T B.$$

We denote this norm by $\|\cdot\|_2$ or, sometimes, by $\|\cdot\|$. Note that $\|A\|_2^2$ is the sum of the squares of the entries of A. We have

$$\begin{aligned}\|AB\|_2^2 = \sum_{i,j}\left|\sum_r A_{i,r}B_{r,j}\right|^2 &\le \sum_{i,j}\left(\sum_r |A_{i,r}|^2\right)\left(\sum_r |B_{r,j}|^2\right)\\ &= \left(\sum_{i,r}|A_{i,r}|^2\right)\left(\sum_{r,j}|B_{r,j}|^2\right)\\ &= \|A\|_2^2\,\|B\|_2^2.\end{aligned}$$

We have $\|I_n\| = n$ and so the trace norm is not an induced norm.

We turn next to induced matrix norms. First we note that

$$\|Ax\|_2^2 = (Ax)^T Ax = x^T A^T Ax$$

and therefore

$$\max_{\|x\|_2=1}\|Ax\|_2$$

is equal to $\sqrt{\rho}$, where ρ is the largest eigenvalue of $A^T A$. (But since we have not discussed eigenvalues at any length yet, we defer any further discussion.)

Both of above norms have the useful property that, if Q is orthogonal, then $\|QA\| = \|A\|$.

15.6.1 Lemma. *Let A be a square matrix. Then*

$$\|A\|_\infty = \max_i \|e_i^T A\|_1.$$

Proof. The function $x \mapsto \|Ax\|_\infty$ is convex and hence realizes its maximum at an extreme point of the unit ball relative to the ℓ_∞ norm. These extreme points are the ± 1-vectors. If x is a ± 1-vector then

$$|(Ax)_i| = |\sum_j A_{i,j}x_j| \le \sum_j |A_{i,j}x_j| \le \sum_j |A_{i,j}| = \|e_i^T A\|_1.$$

Further, equality holds throughout if we choose x so that $A_{i,j}x_j \ge 0$. This proves the lemma. □

15.6.2 Lemma. *Let A be a square matrix. Then*

$$\|A\|_1 = \max_i \|Ae_i\|_1.$$

Proof. Since ℓ_1 and ℓ_∞-norms are dual, we can apply Theorem 15.5.2 to the previous lemma, concluding that

$$\|A\|_1 = \|A^T\|_\infty = \max_i \|e_i^T A^T\|_1 = \max_i \|Ae_i\|_1. \qquad \square$$

(1) If $\|\cdot\|$ is the trace norm or the induced ℓ_2-norm, and Q is an orthogonal matrix, show that $\|QA\| = \|A\|$

15.7 Matrix Functions

We say a matrix is a function of a variable t if each element of the matrix is. This makes sense over any field, but here we work over $\mathbb{R}$ or $\mathbb{C}$. If the matrix $A(t)$ is a function of t then

$$\frac{d}{dt}A(t)$$

is the matrix we get by differentiating each entry of $A(t)$ with respect to t.

As an example, we consider the differential equation

$$f'' + af' + b = 0. \tag{15.7.1}$$

This is equivalent to the following pair of equations:

$$\begin{aligned}\frac{d}{dt}f' &= -af' - b,\\ \frac{d}{dt}f &= f',\end{aligned}$$

which we can rewrite as

$$\frac{d}{dt}\begin{pmatrix} f' \\ f \end{pmatrix} = \begin{pmatrix} -a & -b \\ 1 & 0 \end{pmatrix}\begin{pmatrix} f' \\ f \end{pmatrix}.$$

We can solve this using the matrix exponential.

For any square matrix A we define

$$\exp(tA) := \sum_{n=0}^{\infty} \frac{t^n}{n!}A^n.$$

But we need to see that this makes sense. We have

$$\|A^n\|_\infty \le \|A\|_\infty^n$$

and so, if $a := \|A\|_\infty$, each entry of A^n is bounded in absolute value by a^n. Therefore each entry of

$$\sum_{n=0}^{m} \frac{t^n}{n!}A^n$$

converges as $m \to \infty$, for any value of t. Moreover we are entitled to differentiate the series term-by-term, with the result that

$$\frac{d}{dt}\exp(tA) = \sum_{n=1}^{\infty} \frac{t^{n-1}}{(n-1)!}A^n = A\exp(tA).$$

Now define the vector $F(t)$ by

$$F(t) = \begin{pmatrix} f' \\ f \end{pmatrix}$$

and suppose

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

Then (15.7.1) becomes

$$\frac{d}{dt}F(t) = AF(t)$$

and it is easy to see that this has the solution

$$F(t) = \exp(tA)\,F(0).$$

Although this method of solving differential equations is very important, it is of limited use as a tool for solving particular equations. It is computationally difficult to compute $\exp(A)$ because, even though

$$\frac{1}{n!}A^n \to 0$$

as $n \to \infty$, for moderate values of n this ratio can be very large. The difficulty is essentially the same as attempting to compute $\exp(100)$ using the power series for the exponential.

(1) Show that

$$\exp t(A+B) = \exp(tA)\exp(tB)$$

if and only if $AB - BA = 0$.

(2) If S is skew symmetric, show that $\exp(S)$ is orthogonal.

(3) If

$$H := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

show that $\exp(\pi H) = -I$.

15.8 Powers

We have seen the exponential series in a matrix A is well-defined and useful. We will find useful to consider other power series with matrix arguments. Our next result provides a basic tool.

15.8.1 Lemma. *If A is a non-zero matrix and $\|\cdot\|$ is a matrix norm, then the sequence $\|A^n\|^{1/n}$ converges to a limit ρ. Further $\rho \le \|A^n\|^{1/n}$ for all n.*

Proof. By way of abbreviation, let $f(n) = \|A^n\|^{1/n}$. Note first that

$$\|A^{km}\| \le \|A^m\|^k,$$

and therefore $f(km) \le f(m)$. Assume $n = km + \ell$, where $0 \le \ell < m$. Then

$$f(km+\ell) \le f(km)^{\frac{km}{km+\ell}} f(\ell)^{\frac{\ell}{km+\ell}} \le f(m)^{\frac{km}{km+\ell}} f(\ell)^{\frac{\ell}{km+\ell}}$$

Given $\epsilon > 0$ and fixed m, it follows that for all but finitely many n, we have

$$f(n) \le (1+\epsilon)f(m).$$

We say that $f(m)$ is a *record* for f if, when $k < m$,

$$f(m) < f(k).$$

Consider the sequence of records for f. If it is finite, let ρ denote its last member. If it is not finite, then it is a strictly decreasing sequence, bounded below by 0 and therefore it has a limit, which we denote by ρ. From the previous paragraph it follows that if $\epsilon > 0$, then $f(n) \le (1+\epsilon)\rho$ for all but finitely many values of n. Consequently the sequence $\|A^n\|^{1/n}$ converges to ρ and $\rho \le \|A^n\|^{1/n}$ for all n. □

This lemma does not guarantee that $\rho^{-n}A^n$ converges. For example, if

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

then

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

and, using the trace norm

$$\|A^n\| = \|A\| = 2.$$

Therefore $\|A^n\|^{1/n} = 1$ but, nonetheless, the sequence $(A^n)_{n\ge0}$ does not converge except in special cases.

The quantity

$$lim_{n\to\infty}\|A^n\|^{1/n}$$

is known as the *spectral radius* of A.

We want to work with the geometric series

$$\sum_{r\ge0} t^r A^r.$$

15.8.2 Lemma. *The series $\sum_{r\ge0} t^r A^r$ converges if and only if $t^n A^n \to 0$ as $n\to\infty$. If it does converge, its limit is $(I-tA)^{-1}$.*

Proof. We have

$$(I-tA)(I+tA+\cdots+t^{n-1}A^{n-1}) = I - t^n A^n.$$

Suppose $I-tA$ is not invertible. Then there is a non-zero vector u such that $(I-tA)u = 0$. Therefore $tAu = u$ and $t^r A^r u = u$ for all r. So $t^n A^n$ does not converge to 0 and, since

$$(I+tA+\cdots+t^{n-1}A^{n-1})u = nu,$$

the series $\sum_{r\ge0} t^r A^r$ does not converge.

Hence we may suppose that $I-tA$ is invertible and consequently

$$I+tA+\cdots+t^{n-1}A^{n-1} = (I-tA)^{-1}(I-t^n A^n).$$

The lemma follows immediately. □

15.8.3 Corollary. *Let ρ be the spectral radius of A. The series $\sum_r t^r A^r$ converges (to $(I-tA)^{-1}$) if $|t| < \rho^{-1}$ and diverges if $|t| > \rho^{-1}$.*

Proof. We observe that $t^n A^n$ converges to 0 if and only if $\|t^n A^n\|$ does. By Lemma 15.8.1 we see that $t^n A^n \to 0$ if $|t| < \rho^{-1}$ and that it does not converge if $|t| > \rho$. □

This result shows that ρ^{-1} is the radius of convergence of the series $\sum_n t^n A^n$.

15.9 Contractions

We call a linear map T a *contraction* relative to the norm $\|\cdot\|$ if $\|T^n\| \to 0$ as n increases. Our first result shows that being a contraction is independent of the norm we use.

15.9.1 Lemma. *The linear map T is a contraction if and only if its spectral radius is less than 1.*

Proof. Let ρ be the spectral radius of T. Let $\|\cdot\|$ be an operator norm, and suppose $\epsilon > 0$. By Lemma 15.8.1, for all sufficiently large values of n,

$$\rho^n \le \|T^n\| \le (\rho+\epsilon)^n.$$

The result follows at once. □

While this result has its uses, it does not provide an effective means of deciding if a particular map is a contraction. But contractions are important, and so we need effective ways of recognizing them. If there is an operator norm such that $\|T\| < 1$, then since

$$\|T^n\| \le \|T\|^n,$$

it follows that T is a contraction. Our work in this section shows that, if T is a contraction, there is a norm $\|\cdot\|$ such that $\|T\| < 1$.

If B is a positive definite matrix then the bilinear form

$$\langle u,,\rangle v = u^T B v$$

is an inner product, and $\sqrt{u^T B u}$ is a norm. (See Lemma 16.2.2.)

15.9.2 Lemma. *A matrix A is a contraction if and only if there is a positive definite matrix B such that $B - A^T BA$ is positive definite.*

Proof. Suppose first that B is positive definite and $B - A^T BA$ is positive definite. Then for any non-zero vector v,

$$0 < v^T(B - A^T BA)v = v^T B v^T - v^T A^T BAv.$$

If $\|\cdot\|_B$ denotes the norm determined by B, this shows that, for any non-zero vector v,

$$\|Av\|_B < \|v\|_B$$

and therefore $\|A\|_B < 1$.

To complete the proof, we show that if C is positive definite and the equation

$$X - A^T XA = C \tag{15.9.1}$$

has a positive definite solution X, then A is a contraction. If X satisfies (15.9.1), then

$$\begin{aligned} X &= C + A^T XA \\ A^T XA &= A^T CA + (A^T)^2 XA^2 \\ (A^T)^2 XA^2 &= (A^T)^2 CA^2 + (A^T)^3 XA^3, \end{aligned}$$

which leads us to conjecture (and you to prove, by summing enough of these equations) that

$$X - (C + A^T CA + \cdots + (A^T)^{n-1} CA^{n-1}) = (A^T)^n CA^n.$$

Since the right side of this identity goes to 0 as n increases, we conclude that

$$X = \sum_{r \geq 0} (A^T)^r CA^r$$

is a solution to (15.9.1). Because C is positive definite, $v^T Cv > 0$ for all non-zero vectors v, and therefore

$$v^T (A^T)^r CA^r v > 0$$

for all non-zero vectors v. Consequently X is positive definite. □

Equation (15.9.1) is known as *Stein's equation*. It is a system of linear equations in the entries of X, and so can readily be solved. Since all we need of C is that it be positive definite, we may choose $C = I$. The proof of the lemma shows that if A is a contraction, then Stein's equation has a unique solution. Therefore we could determine if A is a contraction by solving $X - A^T XA = I$, and then testing whether the solution X is positive definite. (This can be decided by Cholesky factorization.)

(1) If C is symmetric and $X - A^T XA = C$ has a solution, show that it has symmetric solution.

(2) Read up on Kronecker products (in Corollary **??**), and then show that, if A does not have distinct eigenvalues whose product is equal to 1, then $X - A^T XA = I$ has a solution.

15.10 Projections

We study subspaces and projections in $\mathbb{R}^n$; our results extend to any inner product space. Suppose U is a k-dimensional subspace of $\mathbb{R}^n$, and let Y be an $n \times k$ matrix whose columns form a basis for U. The Gram-Schmidt algorithm implies that there is a $k \times k$ upper-triangular matrix P such that the columns of YP are orthogonal. As Y and YP have the same column space, it follows that the columns of YP form an orthonormal basis for U.

For our purposes we may as well assume that we chose Y so that $Y^TY = I_k$, without further ado. If we define

$$P = Y^TY$$

then we see that P is symmetric and

$$P^2 = Y^TYY^TY = Y^TY = P.$$

Hence P represents orthogonal projection onto its column space. As $\operatorname{rk} P = \operatorname{rk} Y = k$ and as the column space of P is contained in the column space of Y, it follows that the column space of P equals U. So P represents orthogonal projection onto U. One consequence of this is that the properties of the collection of k-dimensional subspaces of $\mathbb{R}^n$ are mirrored by the properties of the $n \times n$ orthogonal projections with rank k.

Our projections are symmetric and there is a natural inner product on the space of symmetric matrices:

$$\langle A, B\rangle = \operatorname{tr}(AB).$$

If $P_i = Y_i^TY_i$ where Y_i is $n \times k$ and $Y_i^TY_i = I_k$ then

$$\begin{aligned}\langle P_1, P_2\rangle = \operatorname{tr}(Y_1Y_1^TY_2Y_2^T) &= \operatorname{tr}(Y_2^TY_1Y_1^TY_2)\\ &= \operatorname{tr}((Y_1^TY_2)^T(Y_1^TY_2))\\ &\geq 0.\end{aligned}$$

Further

$$\begin{aligned}\langle P_1 - P_2, P_1 - P_2\rangle &= \operatorname{tr}(P_1^2 - P_1P_2 - P_2P_1 + P_2^2)\\ &= \operatorname{tr}(P_1 + P_2 - 2P_1P_2)\\ &= 2k - 2\langle P_1, P_2\rangle.\end{aligned}$$

Thus the value of $k - \operatorname{tr}(P_1P_2)$ can be viewed as a measure of how close the subspaces represented by P_1 and P_2 are.

If P and Q are projections defining two subspaces U and V of $\mathbb{R}^n$ and x is a unit vector in $\mathbb{R}^n$ then $\|Px - Qx\|$ is a measure of distance of U from V. Now

$$\|Px - Qx\|^2 = x^T(P - Q)^2x,$$

whence all information of this sort is contained in the matrix $(P - Q)^2$. The maximum value over all unit vectors x of

$$\|Px - Qx\|^2 = x^T(P - Q)^2x$$

is the largest eigenvalue of the (real symmetric) matrix $(P - Q)^2$. Our next result bounds this.

15.10.1 Lemma. *Let P and Q be projections. Then $\|Px - Qx\| \le \|x\|$ and, if equality holds, $x = Px + Qx$ and $\langle Px, Qx\rangle = 0$.*

Proof. The vectors Px and $(I - P)x$ are orthogonal, so the points represented by the vectors 0, Px and x are the vertices of a right-angled triangle with hypotenuse joining 0 to x. Thus (why??) they lie on the circle with this hypotenuse as a diameter. Similarly the vectors 0, Qx and x form a second right-angled triangle, and also lie on a circle. Now, if two triangles in $\mathbb{R}^n$ share a side then the distance between their third vertices is maximal when they lie in the same plane (and on opposite sides of their shared side). Hence $\|Px - Qx\| \le \|x\|$; if equality holds then the two triangles are coplanar, the two circles coincide and Px and Qx must be diametrically opposed on the circle. Since the origin is on a circle with the line segment from Px to Qx as a diameter, Px, 0 and Qx form a right triangle and Px must be orthogonal to Qx. Further, 0, Px, x and Qx form the vertices of a rectangle; by the parallelogram rule for addition of vectors in the plane, $x = Px + Qx$. □

(1) Show that if P and Q are projections and $\operatorname{rk} P = \operatorname{rk} Q$, then $\operatorname{tr}(P - Q)^3 = 0$.

(2) Show that $(P - Q)^2$ commutes with P and Q.

15.11 Contractions

In this section, we derive the characterization of contractions in terms of eigenvalues. If M is a square matrix, we use $\|M\|_1$ to denote the induced ℓ_1 norm of M—this equals the maximum value of the ℓ_1-norms of the columns of M, as we saw in **??**.

15.11.1 Theorem. *Let A be a square matrix. If $|\theta| < 1$ for all eigenvalues θ of A, then A is a contraction.*

Proof. As a first step, we prove the theorem when A is lower triangular. Suppose A is $n \times n$ and let D_t be the $n \times n$ diagonal matrix with $(D_t)_{i,i} = t^{i-1}$. Let Δ denote the diagonal matrix with $\Delta_{i,i} = A_{i,i}$. The ij-entry of $D_t^{-1} A D_t$ is $t^{j-i} A_{i,j}$ and so

$$\lim_{t \to \infty} D_t^{-1} A D_t = \Delta.$$

In particular, given $\epsilon > 0$, we can choose t large enough that $\|D_t^{-1} A D_t\|_1$ lies within ϵ of $\|\Delta\|_1$. Consequently, if $|\theta| < 1$ for each eigenvalue θ, then we can choose t so that $\|D_t^{-1} A D_t\|_1 < 1$.

This implies that

$$\|D_t^{-1} A^n D_t\|_1 \to 0$$

as $n \to \infty$. Since

$$\|A^n\|_1 = \|D_t^{-1} D_t^{-1} A^n D_t D_t\| \le \|D_t^{-1}\|_1 \|D_t^{-1} A^n D_t\|_1 \|D_t\|_1,$$

it follows that $\|A^n\|_1 \to 0$ as $n \to \infty$.

If A is not triangular, then $A = LTL^{-1}$, where T is triangular. Since

$$\|A^n\| \le \|L\|_1 \|T^n\|_1 \|L^{-1}\|_1$$

and

$$\|T^n\|_1 = \|L^{-1}A^nL\|_1 \le \|L^{-1}\|_1 \|A^n\|_1 \|L\|_1,$$

we see that A is a contraction if and only if T is. To complete the proof, we recall that A and T have the same eigenvalues. □

There is another proof of this result using root vectors.

15.11.2 Lemma. *Let A be an $n \times n$ matrix over $\mathbb{C}$, let θ be an eigenvalue of A and let v be a root vector for θ. If $|\theta| < 1$, then $A^m v \to 0$ as $m \to \infty$.*

Proof. Since v is a root vector for θ, we have $(A - \theta I)^n v = 0$. Then

$$A^m = (A - \theta I + \theta I)^m$$

and so using the binomial theorem, we find that

$$A^m v = \theta^{m-n+1} \Big[\binom{m}{n-1} (A - \theta I)^{n-1} \\ + \binom{m}{n-2} (A - \theta I)^{n-2} \theta + \cdots + \theta^{n-1} I \Big] v.$$

Hence we have

$$A^m v = \theta^{m-n+1} P(m) v,$$

where $P(m)$ is a matrix whose entries are polynomials in m with degree at most $n - 1$. Since $|\theta| < 1$, it follows that

$$\theta^{m-n+1} P(m) \to 0$$

as $m \to \infty$. □

Now suppose A is an $n \times n$ matrix with all eigenvalues inside the unit circle. Since each vector in $\mathbb{C}^n$ is a linear combination of root vectors, it follows that for any vector v,

$$A^m v \to 0$$

as $m \to \infty$.

We have two methods now for determining if a square matrix A is a contraction. We can solve Stein's equation, as discussed in Section 15.9, or we can compute the spectral radius from the eigenvalues of A. This second alternative is useful if A is symmetric, or if A is real and its entries are positive.

15.12 Perron

We say a real matrix M is *non-negative* if all its entries are non-negative. We write $M \geq N$ is $M - N$ is non-negative. We say M is *positive* if all its entries are positive. If M is a real matrix of any order, then we define $|M|$ to be the matrix we get by replacing each entry by its absolute value.

15.12.1 Lemma. *Let A be an $n \times n$ matrix with spectral radius ρ, and suppose A is real and all its entries are positive. Suppose that θ is an eigenvalue such that $|\theta| = \rho$ and let x be an eigenvector wih eigenvalue θ. Then $|x|$ is an eigenvector for A with eigenvalue ρ.*

Proof. We have

$$\rho|x| = |\theta x| = |Ax| \leq |A|\,|x|$$

and therefore

$$A|x| \geq \rho|x|.$$

First, suppose there is a non-negative non-zero vector z such that $Az \geq \sigma z$ and $\sigma > \rho$. Then

$$A^n z \geq \sigma^n z$$

and therefore

$$\|A^n\| \geq \sigma^n$$

for all n. This implies that the spectral radius of A is at least σ, which contradicts the fact that the spectral radius equals ρ.

Now suppose that z is a non-negative non-zero vector such that $Az \geq \rho z$ and, for some index k, we have

$$e_k^T Az > \rho e_k^T z.$$

Consider the vector $z + te_k$, where t is small. Then

$$A(z + te_k) \geq \rho z + tAe_k.$$

Since all entries of A are positive, it follows that, if $i \neq k$, then

$$e_i^T A(z + te_k) > \rho e_i^T z = \rho e_i^T (z + te_k).$$

On the other hand

$$\begin{aligned} e_k^T A(z + te_k) = e_k^T Az + te_k^T Ae_k > \rho e_k^T z + te_k^T Ae_k \\ = \rho e_k^T (z + te_k) + t(A_{k,k} - 1). \end{aligned}$$

It follows that the are positive values of t such that

$$e_k^T A(z + te_k) > \rho e_k^T (z + te_k)$$

and, for these values of t, we have

$$A(z + te_k) > \rho z + te_k.$$

Since this is impossible, we are forced to conclude that $Az = \rho z$. □

15.12.2 Theorem. *Let A be a real square matrix with positive entries. Then the spectral radius of A is an eigenvalue of A with algebraic multiplicity 1, and the corresponding eigenspace is spanned by an eigenvector with all entries positive. If θ is an eigenvalue of A not equal to ρ, then $|\theta| < \rho$.*

Proof. We have seen that there is an eigenvector x with eigenvalue ρ and all its entries non-negative. We show that the entries of any non-negative eigenvector with eigenvalue ρ must all be positive. Suppose $\rho y = Ay$ and $y \geq 0$. Then

$$\rho e_i^T y = e_i^T Ay = \sum_j A_{i,j} y_j.$$

However all entries of A are positive and y is non-negative and not zero, so the above sum is positive. As $\rho > 0$, it follows that $e_i^T y > 0$.

Next we show that ρ has geometric multiplicity 1. Assume $Ay = \rho y$, where y is not a scalar multiple of x. Then there is a real number t such that $x + ty \geq 0$ and some entry of $x + ty$ equals 0. But $x + ty$ is an eigenvector for A with eigenvalue ρ, and so we have a contradiction. We conclude that ρ has geometric multiplicity 1.

Finally we show that ρ has algebraic multiplicity 1. Suppose that $(A - \rho I)^2 w = 0$ and w is not in $\ker(A - \rho I)$. Then, replacing w by $-w$ if needed, we may assume that $x = (A - \rho I)w$ is a positive eigenvector for A with eigenvalue ρ. Note now that A^T is a positive matrix with spectral radius ρ. (It has the same minimal polynomial as A, hence has the same eigenvalues.) Let y be a positive eigenvector for A^T with eigenvalue ρ. Then $y^T(A - \rho I) = 0$, and consequently

$$y^T x = y^T (A - \rho I) w = 0.$$

But y and x are positive, and therefore $y^T x > 0$. Thus we conclude that, if $(A - \rho I)^2 w = 0$ then $w = 0$. Therefore the algebraic multiplicity of ρ is 1.

Now suppose that θ is an eigenvalue of A distinct from ρ, and let x be an eigenvector for θ. Then, using the triangle inequality,

$$|\theta|\,|x_i| = |(Ax)_i| = \Big|\sum_j A_{i,j} x_j\Big| \leq \sum_j |A_{i,j} x_j| = (A|x|)_i.$$

This implies that $|\theta| \leq \rho$. If equality holds, then

$$\Big|\sum_j A_{i,j} x_j\Big| \leq \sum_j |A_{i,j} x_j|.$$

Thus we have n possibly complex numbers $z_j := A_{i,j}$ such that

$$\Big|\sum_j z_j\Big| \leq \sum_j |z_j|,$$

which implies that there is a root of unity ξ such that ξz_j is real and positive for all j. Therefore ξx is a positive eigenvector and $\theta = \rho$. □

If $y^T A = \rho y^T$ and $Ax = \theta x$, where $\theta \neq \rho$, then $y^T x = 0$. This implies that any non-negative eigenvector for A must be an eigenvector for ρ.

15.12.3 Lemma. *Let A be a real square matrix with all entries positive, and let x be a positive eigenvector for A with eigenvalue ρ, such that $\mathbf{1}^T x = 1$. If u is a non-zero non-negative vector, then*

$$\lim_{n\to\infty} \frac{A^n u}{\mathbf{1}^T A^n u} = x.$$

Proof. Let x be a positive eigenvector for A with eigenvalue ρ, and let y be a positive eigenvector for A^T with eigenvalue ρ. Let B be defined by

$$B := A - \frac{\rho}{y^T x} x y^T.$$

If $Az = \theta z$ and $\theta \neq \rho$, then $y^T z = 0$ and $Bz = \theta z$. Also $Bx = 0$ and therefore if θ is an eigenvalue of B, then $|\theta| < \rho$. Consequently $\rho^{-1}B$ is a contraction. Let E be given by

$$E := \frac{1}{y^T x} x y^T.$$

Then $E^2 = E$ and $AE = EA$ and $BE = EB = 0$. Accordingly

$$(B + \rho E)^n = B^n + \rho^n E$$

and, for any vector u,

$$A^n u - \rho^n E u = B^n u.$$

Therefore, since $\rho^{-1}B$ is a contraction, $\rho^{-n}B^n u \to 0$ as $n \to \infty$ and, provided $y^T u \neq 0$.

$$\lim_{n\to\infty} \frac{A^n u}{\mathbf{1}^T A^n u} = \lim_{n\to\infty} \frac{\rho^{-n} A^n u}{\rho^{-n} \mathbf{1}^T A^n u} = \frac{1}{\mathbf{1}^T E u} E u = \frac{1}{\mathbf{1}^T x} x.$$

(1) Let A be a positive square matrix. Show that there is a non-negative vector x such $(I - A)x$ is non-negative and not zero if and only if A is a contraction.

16
Positive Semidefinite Matrices

16.1 Gram Matrices

The *Gram matrix* G of a subset $x_1, \dots, x_n$ of U is the matrix with entries given by

$$G_{i,j} = \langle x_i, x_j \rangle.$$

If $a^T = (a_1, \dots, a_n)^T$, then

$$a^T Ga = \left\langle \sum_i a_i x_i, \sum_i a_i x_i \right\rangle$$

and therefore $a^T Ga > 0$ for any non-zero vector a. We say a matrix G is *positive definite* if it is self-adjoint and $a^T Ga > 0$ for any non-zero vector a; if it is self-adjoint and $a^T AGa \geq 0$ for all a, then G is *positive semidefinite*. We have just seen that Gram matrices are positive semidefinite.

16.1.1 Lemma. *A set of vectors in an inner product space is linearly independent if and only if their Gram matrix is invertible.*

Proof. Suppose G is the Gram matrix for $x_1, \dots, x_n$. Then the entries of Ga are the inner products

$$\langle x_r, \sum_r a_r x_r \rangle$$

Hence if U is the span of the vectors $x_1, \dots, x_n$, then $Ga = 0$ if and only if $\sum_r a_r x_r = 0$. Thus $\ker(G)$ is zero if and only if $x_1, \dots, x_n$ are linearly independent. □

16.2 Factorizing Positive Semidefinite Matrices

If H is a matrix with linearly independent columns, then the product $H^T H$ is the Gram matrix for a basis of $\operatorname{col}(H)$ and therefore it is positive definite. Our next result provides a converse to this.

16.2.1 Theorem. *If G is a positive definite matrix, there is a lower triangular matrix L with diagonal entries equal to* 1 *and a diagonal matrix D with positive diagonal entries, such that $LGL^T = D$.*

Proof. If G is positive definite, then $e_i^T G e_i > 0$ for all i; hence the diagonal entries of G are positive.

Since L and G are invertible, $D = LGL^T$ is necessarily invertible. We must show that L exists. We write G in partitioned form:

$$G = \begin{pmatrix} a & b^T \\ b & G_1 \end{pmatrix}.$$

If we also define

$$L_1 = \begin{pmatrix} 1 & 0 \\ -a^{-1}b & I \end{pmatrix}$$

then

$$L_1 G L_1^T = \begin{pmatrix} a & 0 \\ 0 & G_1 - a^{-1}bb^T \end{pmatrix}.$$

Note that $a \neq 0$, because G is positive definite. It follows from the exercises below that $G_1 - a^{-1}bb^T$ is positive definite. By induction, we have that there is a lower triangular matrix L_2 with diagonal entries equal to 1 such that

$$L_2(G - a^{-1}bb^T)L_2^T$$

is diagonal. Taking L to be given by

$$L := \begin{pmatrix} 1 & 0 \\ 0 & L_2 \end{pmatrix} L_1,$$

our result follows. □

This result implies that $G = L^{-1}DL^{-T}$. Since the diagonal entries of D are positive, there is a unique non-negative diagonal matrix $D^{1/2}$ such that $(D^{1/2})^2 = D$ and therefore

$$G = (L^{-1}D^{1/2})(L^{-1}D^{1/2})^T.$$

A factorization of a positive-definite matrix G in the form MM^T, where M is lower triangular with positive diagonal entries, is known as a *Cholesky factorization*. Any reasonable software package for linear algebra will have a command to compute the matrix M from G.

If G is presented as a matrix X^TX and $LGL^T = D$, then

$$(XL^T)^T(XL^T) = D,$$

whence we see that the columns of XL^T are orthogonal (with respect to the dot product). Thus they form an orthogonal basis for $\operatorname{col}(X)$, and so we may use the Cholesky decomposition to find orthogonal bases. We illustrate this in the next section.

We record an important property of positive definite matrices—it is basically a reformulation of the definition.

16.2.2 Lemma. *If A is a positive definite matrix, the bilinear form*

$$\langle x, y\rangle = x^T Ay$$

is an inner product.

Proof. Exercise. □

(1) If G is positive definite and the columns of L are linearly independent, show that LGL^T is positive definite.

(2) Show that a principal submatrix of a positive definite matrix is positive definite.

(3) Prove that if G has Cholesky factorizations MM^T and NN^T, then $M = N$.

16.3 Computing Cholesky

The Cholesky decomposition of a positive definite matrix can be useful, in particular it may be used to find orthogonal bases. In this section we describe an algorithm for computing the Cholesky factorization using elementary row operations. (But outside linear algebra courses, we recommend using methods based on the QR-factorization, which we address later. Our point is that we can carry out Gram-Schmidt by using Gaussian elimination.)

As a first step, we need to to note one consequence of Theorem 16.2.1. This result shows that if G is positive definite, then by successively subracting multiples of higher rows from lower rows, we can convert G to an invertible upper triangular matrix. The product of the elementary matrices corresponding to these operations is the lower triangular matrix L. Our next result asserts that if we use elementary operations as described to bring G to row echelon form, we obtain the Cholesky factorization of G.

16.3.1 Lemma. *Let G be a positive definite matrix. If K is lower triangular with diagonal entries equal to 1 and KG is upper triangular, then $KG = DK^{-T}$, where D is a diagonal matrix with positive diagonal entries.*

Proof. Suppose that K is lower triangular with diagonal entries equal to 1, and that $KG = DM$, where D is diagonal and M is upper triangular, with diagonal entries 0 or 1. Then

$$KGK^T = DMK^T.$$

Here the left side is a symmetric matrix, while the right side is the product of three upper triangular matrices, and is therefore upper triangular. It follows that MK^T is diagonal. Since KGK^T is invertible, both D and MK^T are invertible. Therefore $MK^T = I$. Finally KGK^T is positive definite and equal to D. So D is positive definite, and therefore its diagonal entries are positive. □

Suppose we are given a Gram matrix G. If we bring the partitioned matrix

$$\begin{pmatrix} G & I \end{pmatrix}$$

to row-echelon form, then the resulting matrix equals

$$\begin{pmatrix} LG & L \end{pmatrix}.$$

As noted at the end of the previous section, if $G = X^T X$, then the columns of XL^T are orthogonal (with respect to the dot product). The i-th column of XL^T is a linear combination of the first i columns of X and consequently the columns of XL^T are the orthogonal set we would compute using the usual approach to Gram-Schmidt. (Using exact arithmetic—in fact we have developed the so-called modified Gram-Schmidt method.)

We first illustrate this in $\mathbb{R}^n$, with the dot product. The row echelon form of the partitioned matrix

$$M = \begin{pmatrix} X^T X & X^T \end{pmatrix}$$

is

$$\begin{pmatrix} LX^T X & LX^T \end{pmatrix}$$

and so the transposes of the rows of LX^T are an orthogonal basis for the column space of X. Suppose for example that

$$x_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad x_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \quad x_3 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}.$$

Let X be the matrix with x_1, x_2 and x_3 as its columns. Then

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

has row echelon form

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

Hence

$$XL^T = \begin{pmatrix} 1 & -\frac{1}{2} & \frac{2}{3} \\ 1 & \frac{1}{2} & -\frac{2}{3} \\ 0 & 1 & \frac{2}{3} \end{pmatrix}$$

and its columns are an orthogonal basis for $\operatorname{col}(X)$.

16.4 Polynomial Examples

We consider the situation where we want to find an orthogonal basis for an inner product space of polynomials. By way of example, we take V to be the space of all polynomials, with inner product:

$$\langle p, q\rangle := \int_0^\infty p(x)q(x)e^{-x}\,dx.$$

Let U be the subspace consisting of the polynomials with degree at most n, let $p_0, \dots, p_n$ be basis for U and let G be the Gram matrix of this basis. (Thus the rows and columns of G are indexed by $0, 1, \dots, n$, rather than $1, \dots, n$—good news for C programmers anyway.)

If $[q]$ denotes the coordinate vector of q in U relative to the given basis, then

$$[p]^T G[q] = \langle p, q\rangle.$$

Suppose $LGL^T = D$. Then

$$e_i^T LGL^T D = e_i^T De_j$$

whence the columns of L^T are the coordinate vectors of an orthogonal basis for U.

Turning to a concrete case, suppose U is the space of polynomials with degree at most three. We start with the basis 1, x, x^2, x^3. It can be shown (by integration by parts) that

$$\int_0^\infty x^n e^{-x}\,dx = n!,$$

and therefore the Gram matrix of this set of polynomials is

$$G = \begin{pmatrix} 1 & 1 & 2 & 6 \\ 1 & 2 & 6 & 24 \\ 2 & 6 & 24 & 120 \\ 6 & 24 & 120 & 720 \end{pmatrix}.$$

Let M be given by

$$M = \begin{pmatrix} 1 & 1 & 2 & 6 & 1 & 0 & 0 & 0 \\ 1 & 2 & 6 & 24 & 0 & 1 & 0 & 0 \\ 2 & 6 & 24 & 120 & 0 & 0 & 1 & 0 \\ 6 & 24 & 120 & 720 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We convert the first four columns to an upper triangular matrix:

$$\begin{pmatrix} 1 & 1 & 2 & 6 & 1 & 0 & 0 & 0 \\ 0 & 1 & 4 & 18 & -1 & 1 & 0 & 0 \\ 0 & 0 & 4 & 36 & 2 & -4 & 1 & 0 \\ 0 & 0 & 0 & 36 & -6 & 18 & -9 & 1 \end{pmatrix},$$

and thus obtain the following set of four orthogonal polynomials:

$$1, \quad x-1, \quad x^2-4x+2, \quad x^3-9x^2+18x-6.$$

16.5 Positive Semidefinite Matrices

We develop some further properties of positive semidefinite matrices.

16.5.1 Lemma. *If A and B are positive semidefinite, so is $A+B$. If A is positive and B is positive definite, then $A+B$ is positive definite.* □

We leave the proof as an exercise. Note that it implies that if A is positive semidefinite, then $A+I$ is positive definite.

16.5.2 Lemma. *A self-adjoint matrix is positive semidefinite if and only if its eigenvalues are non-negative. It is positive definite if and only if its eigenvalues are positive.*

Proof. If x is an eigenvector of A with eigenvalue θ, then $x^TAx = \theta x^Tx$, and therefore if A is positive semidefinite, its eigenvalues are non-negative. If A is positive definite then 0 is not an eigenvalue.

Suppose we have the spectral decomposition

$$A = \sum_\theta \theta E_\theta.$$

Each projection E_θ is positive semidefinite, because

$$x^TE_\theta x = x^TE_\theta^2x = x^TE_\theta^TE_\theta x = \|E_\theta x\|^2.$$

If each eigenvalue of A is non-negative, it follows that x^TAx is a sum of non-negative terms $\theta x^TE_\theta x$, and therefore $x^TAX \ge 0$.

If the eigenvalues of A are positive, we see that $x^TAx = 0$ if and only if $x^TE_\theta x = 0$ for each eigenvalue θ. Hence

$$0 = \sum_\theta x^TE_\theta x = x^T\left(\sum_\theta E_\theta\right)x = x^TIx,$$

and therefore $x = 0$. Consequently A is positive definite. □

Note that I_n has 2^n distinct square roots, that is, there are 2^n matrices S such that $S^2 = I$. However it has only one positive semidefinite square root. This is typical:

16.5.3 Corollary. *If A is positive semidefinite, there is a unique positive semidefinite matrix S such that $S^2 = A$.*

Proof. Using the spectral decomposition we have

$$A = \sum_\theta \theta E_\theta,$$

where the sum is over all eigenvalues of A. If A is positive semidefinite, its eigenvalues are non-negative and we may define S by

$$S = \sum_\theta \sqrt{\theta}E_\theta.$$

Since the eigenvalues of S are non-negative, it is positive semidefinite.

We turn to uniqueness. Let T be a positive semidefinite square root of A and suppose x is an eigenvector for A. If $Ax = 0$ then $T^2x = 0$, so $x^TTTx = 0$ and therefore $Tx = 0$. Assume now that $Ax = \sigma^2x$, where $\sigma > 0$, then

$$0 = (T^2 - \sigma^2I)x = (T - \sigma I)(T + \sigma I)x.$$

If the subspace spanned by x is T-invariant, it follows that $Tx = \pm\sigma x$ and x is an eigenvector for T. Otherwise x and Tx span a T-invariant subspace on which T acts wih minimal polynomial $t^2 - \sigma^2$. If $(T - \sigma I)x \neq 0$ then $y = (T - \sigma I)x$ is an eigenvector for $T + \sigma I$ with eigenvalue $-\sigma$. Therefore if T is positive semidefinite and $Ax = \sigma^2x$, then $Tx = \sigma x$.

Thus we have shown that, if $Ax = \sigma^2x$ then $Tx = \sigma x$. Since the eigenvectors of A span, this shows that T is determined by A. □

The next result is known as the *polar decomposition* of a matrix. It is analogous to the fact that each complex number is the product of a positive real number and a complex number with norm 1.

16.5.4 Theorem. *If A is a square matrix, there is a positive semidefinite matrix M and an orthogonal matrix Q such that $A = MQ$.*

Proof. We use the singular value decomposition, which yields that

$$A = Y\Sigma X^T,$$

where X and Y are orthogonal and Σ is positive semidefinite. Hence

$$A = Y\Sigma Y^TYX^T,$$

where $Y\Sigma Y^T$ is positive semidefinite and YX^T is orthogonal. □

Note that $AA^T = (MQ)(MQ)^T = M^2$; hence the positive definite factor in the above theorem is unique, and the orthogonal factor is unique if A is invertible.

17
Channels

We concern ourselves with linear maps from the space $\mathscr{M}_d$ of $d \times d$ complex matrices to the the space $\mathscr{M}_e$ of $e \times e$ matrices. Special classes of such maps are known to physicists as *channels* and we will address these too.

17.1 Matrix Maps

Suppose A is a $k \times \ell$ matrix and B is an $m \times n$ matrix over a field $\mathbb{F}$. Then the map

$$\Phi : M \mapsto AMB$$

on $\ell \times m$ matrices is linear, with domain $\operatorname{Mat}_{\ell \times m}(\mathbb{F})$ and codomain $\operatorname{Mat}_{k \times n}(\mathbb{F})$. More generally, if

$$\Phi(M) := \sum_r A_r M B_r$$

then Φ is a linear map from $\operatorname{Mat}_{\ell \times m}(\mathbb{F})$ to $\operatorname{Mat}_{k \times n}(\mathbb{F})$, and it is not hard to show that all such linear maps can be represented in this form.

We assume henceforth that we are working over $\mathbb{C}$. We denote $\operatorname{Mat}_{d \times d}(\mathbb{C})$ and $\operatorname{Mat}_{e \times e}(\mathbb{C})$ by $\mathscr{M}_d$ and $\mathscr{M}_e$ respectively, and we note that any linear map Φ from $\mathscr{M}_d$ to $\mathscr{M}_e$ can be expressed in the form

$$\Phi(M) = \sum_r A_r M B_r^*.$$

A linear map $\Psi : \mathscr{M}_e \to \mathscr{M}_d$ is *adjoint* to Φ if, for all matrices M and N we have

$$\langle \Psi(N), M \rangle = \langle N, \Phi(M) \rangle.$$

Since

$$\langle N, \Phi(M) \rangle = \sum_r \operatorname{tr}(N^* A_r M B_r^*) = \sum_r \operatorname{tr}(B_r^* N^* A_r M) = \sum_r \langle A_r^* N B_r, M \rangle$$

and it is easy to verify that

$$\Psi(N) = \sum_r A_r^* N B_r.$$

As is traditional, we use Φ^* to denote the adjoint of Φ. (Which is a truly unfortunate choice! $\Phi^*(X) \neq \Phi(X)^*$.)

A linear map Φ on a matrix algebra is *unital* if it maps I to I. It is *trace preserving* if $\operatorname{tr}(\Phi(M)) = \operatorname{tr}(M)$ for all M. Now

$$\Phi(I) = \sum_r A_r B_r^*,$$

whence Φ is unital if and only $\sum_r A_r B_r^* = I$. We also see that

$$\operatorname{tr}(\Phi(M)) = \sum_r \operatorname{tr}(A_r M B_r^*) = \operatorname{tr}\left(M\left(\sum_r B_r^* A_r\right)\right)$$

and therefore Φ is trace preserving if and only if $\sum_r B_r^* A_r = I$.

If $\sum_r A_r B_r^* = I$, then

$$I = I^* = \sum_r B_r A_r^*$$

and thus Φ^* is trace preserving if and only if Φ is unital. (Similarly Φ^* is unital if and only if Φ is trace preserving.)

17.2 Norms

A *Banach space* is a complete normed vector space. If V and W are Banach spaces and $L : V \to W$ is linear, the *operator norm* of L is

$$\sup_{\|v\|=1} \|Av\|.$$

and we denote it by $\|L\|$. If the norm of L is finite we say that L is *bounded* and we use $B(V,W)$ to denote the set of bounded operators from V to W. When $V = W$ (as will usually be the case), we write simply $B(V)$. Note that an operator is bounded if and only if is continuous. You may show that

$$\|LM\| \le \|L\|\|M\|,$$

whence it follows that $B(V)$ is an algebra. Note that $\|I\| = 1$.

A norm on an algebra is *sub-multiplicative* if

$$\|AB\| \le \|A\|\|B\|.$$

Operator norms are necessarily sub-multiplicative, as you may verify.

If our Banach space V is $\mathbb{C}^d$, we can express the norm of an operator in $B(V)$, (i.e., of a $d \times d$ matrix) using singular values.

17.2.1 Theorem. *Let A be a $d \times e$ matrix over $\mathbb{C}$ with $d \ge e$. Then there is a $d \times d$ matrix U, a diagonal $n \times n$ matrix Σ and an $e \times e$ matrix V such that:*

(a) $UU^* = I_d$, $VV^* = I_e$.

(b) *Σ is real and non-negative.*

(c) $A = U^* \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V$. □

The diagonal entries of Σ are the *singular values* of A, we set $\sigma_i = \Sigma_{i,i}$ with the assumption that

$$\sigma_1 \geq \cdots \sigma_e.$$

17.2.2 Lemma. *If $A \in \mathscr{M}_d$, then $\sigma_1(A) = \|A\|$.*

Proof. Assume A has singular value decomposition $A = U^* \Sigma V$. We have

$$\|Ax\|^2 = x^* A^* Ax = x^* V^* \Sigma U U^* \Sigma V x = x^* V^* \Sigma^2 V x$$

and so $\|Ax\|^2 = \sigma_1^2$. □

The *Hilbert-Schmidt norm* on an operator A on a Hilbert space is defined to be $\mathrm{tr}(A^* A)^{1/2}$. (If the underlying Hilbert space is infinite-dimensional, it may be be infinite for some operators; the operators for which it is finite are said to be *trace-class*.) If A has singular value decomposition $U^* \Sigma V$, then

$$\mathrm{tr}(A^* A) = \mathrm{tr}(V^* \Sigma U U^* \Sigma V) = \mathrm{tr}(V^* \Sigma^2 V) = \mathrm{tr}(\Sigma^2).$$

Thus the sum of the squares of the singular values of A is equal to $\langle A, A \rangle$.

17.3 Positive Maps

A linear map $\Phi : \mathscr{M}_d \to \mathscr{M}_e$ is *positive* if $\Phi(M) \succcurlyeq 0$ whenever $M \succcurlyeq 0$. We note some examples:

(a) $\Phi(M) = I \circ M$,

(b) $\Phi(M) = \sum_r V_r M V_r^*$,

(c) $\Phi(M) = M^T$.

There is no known characterization of positive maps; this is less of an issue that it might be, because only a special class of positive maps is of interest in quantum computing: the so-called *completely positive maps*. (We will discuss these in Section 17.5.)

A linear map $\Phi : \mathscr{M}_d \to \mathscr{M}_e$ is *Hermitian preserving* if $\Phi(M^*) = \Phi(M)^*$.

17.3.1 Lemma. *A positive linear map is Hermitian preserving.*

Proof. First we show that positive maps take Hermitian matrices to Hermitian matrices. By the spectral decomposition a Hermitian matrix is a linear combination of positive semidefinite matrices and so a positive linear map sends a Hermitian matrix to a linear combination of positive semidefinite matrices. Since positive semidefinite matrices are Hermitian, our claim follows.

Next, if M is a square complex matrix, we have

$$M = \frac{1}{2}[(M + M^*) - i(iM - iM^*)]$$

and this shows that M is a linear combination of Hermitian matrices. If A and B are Hermitian and $M = A + iB$

$$\Phi(M^*) = \Phi(A - iB) = \Phi(A) - i\Phi(B) = \Phi(M)^*. \qquad \square$$

Our next goal is to show that if Φ is positive, then $\|\Phi\| = \|\Phi(I)\|$.

17.3.2 Lemma. *Assume Φ is positive. If A is normal, then*

$$\Phi(A)\Phi(A)^* \preccurlyeq \Phi(AA^*).$$

Proof. Let $A = \sum_r \lambda_r F_r$ be the spectral decomposition of A. Then

$$\begin{pmatrix} \Phi(A^*A) & \Phi(A) \\ \Phi(A^*) & I \end{pmatrix} = \sum_r \begin{pmatrix} \lambda_r\overline{\lambda}_r & \lambda_r \\ \overline{\lambda}_r & 1 \end{pmatrix} \otimes \Phi(F_r)$$

where the right side is positive semidefinite. Now noting the identity

$$\begin{pmatrix} I & -C \\ 0 & I \end{pmatrix}\begin{pmatrix} B & C \\ C^* & I \end{pmatrix}\begin{pmatrix} I & 0 \\ -C^* & I \end{pmatrix} = \begin{pmatrix} B - CC^* & 0 \\ 0 & I \end{pmatrix},$$

we conclude that

$$0 \preccurlyeq \Phi(A^*A) - \Phi(A)\Phi(A^*) = \Phi(A^*A) - \Phi(A)\Phi(A)^*. \qquad \square$$

The next result is known as the Russo-Dye theorem.

17.3.3 Theorem. *If Φ is positive, then $\|\Phi\| \le \|\Phi(I)\|$.*

Proof. We first prove that if Φ is unital, then $\|\Phi\| = 1$.

Assume $\|M\| \le 1$. Then

$$\|MM^*\| \le \|M\|\|M^*\| = \|M\|^2 \le 1$$

and therefore the matrices $I - MM^*$ and $I - M^*M$ are positive semidefinite. The matrix

$$\widehat{M} = \begin{pmatrix} M & -(I - MM^*)^{1/2} \\ (I - M^*M)^{1/2} & M^* \end{pmatrix}$$

is unitary (check this!).

Assume M is $d \times d$. The map that sends a $2d \times 2d$ matrix to its leading $d \times d$ block is linear, unital and positive. (It is a *compression*.) If Ψ denotes the image of this compression under Φ, then Ψ is positive and unital. Hence we may apply Lemma 17.3.2 to conclude that

$$\Psi(\widehat{M})\Psi(\widehat{M}^*) \preccurlyeq \Psi(\widehat{M}\widehat{M}^*) = \Psi(I) = I;$$

this immediately implies that $\Phi(M)\Phi(M)^* \preccurlyeq I$.

Now suppose Φ is not unitary, but $D = \Phi(I)$ is invertible. Then

$$\Psi := D^{-1/2}\Phi D^{-1/2}$$

is positive and unitary and

$$\|\Phi(M)\| = \|D^{1/2}\Psi(M)D^{1/2}\| \le \|D\|\|\Psi(M)\| \le \|D\|\|M\|.$$

It follows that $\|\Phi\| \le \|\Phi(I)\|$.

Finally, if D is not invertible, we consider the maps

$$\Phi_\epsilon = \Phi + \epsilon I;$$

since $\|\Phi_\epsilon\| \le \|\Phi_\epsilon(I)\|$ the full result follows by continuity. □

There is a simpler proof that, if Φ is a positive linear functional, then $\|\Phi\| = \|\Phi(I)\|$. For this we need the dual norm $\|\cdot\|^*$ to the operator norm $\|\cdot\|$. The operator norm is the largest singular value, the dual norm is the sum of all singular values. Thus the dual norm of a positive semidefinite matrix is its trace. If Φ is a postive linear functional, there is a positive semidefinite matrix Q such that $\Phi(A) = \langle Q, A\rangle$. Now

$$\|\Phi(A)\| = |\langle A, Q\rangle| \le \|A\|\|Q\|^* = \|A\| \operatorname{tr}(Q)$$

and as

$$\operatorname{tr}(Q) = \langle Q, I\rangle = \Phi(I),$$

we conclude that $\|\Phi\| = \|\Phi(I)\|$.

A linear functional φ on a C^* algebra is a *state* if it is positive and has norm 1.

A square matrix M is a *contraction* if $\|M\| \le 1$ The matrix $\widehat{M}$ defined above is referred to as a *unitary dilation* of M. Our argument shows that any contraction of order $n \times n$ is a principal submatrix of a unitary matrix of order $2n \times 2n$. Conversely, any principal submatrix of a unitary matrix is a contraction (and this is straightforward to prove).

17.4 Contractions and Positive Maps

We are going to prove the unital maps with norm one are positive. We will work at a more general level. An *operator system* is a *-closed subspace of a matrix algebra that contains I.

17.4.1 Theorem. *Let $\mathscr{S}$ be an operator system and let $\Phi : \mathscr{S} \to \mathscr{M}_e$ be a unital linear map with norm one. Then Φ is positive.*

Proof. We first prove the result under the assumption that Φ is a linear functional (i.e., $e = 1$). If Φ is a linear functional on S, by the Hahn-Banach theorem it can be extended to a linear functional on $\mathscr{M}_d$ with the same norm. Therefore there is a matrix M in $\mathscr{M}_d$ such that

$$\Phi(X) = \langle M, X\rangle$$

for all X in S We have

$$1 = \Phi(I) = \langle M, I\rangle = \operatorname{tr}(M).$$

Now

$$\|\Phi\| = \sup_{\|X\|=1} \langle M, X\rangle$$

and hence if $\|\Phi\| = 1$, then whenever $\|X\| = 1$,

$$\langle M, I\rangle \geq \langle M, X\rangle$$

and so

$$\langle M, I - X\rangle \geq 0.$$

The singular values of a positive semidefinite matrix are its eigenvalues (and its norm is its largest eigenvalue). So if $M \succcurlyeq 0$, then $I - \|M\|^{-1}M \succcurlyeq 0$. Suppse $Y \in \mathscr{S}$ and $Y \succcurlyeq 0$. If $\nu = \|Y\|$, then $\|I - \nu^{-1}Y\| \leq 1$, whence

$$0 \leq \langle M, I - (I - \nu^{-1}Y)\rangle = \nu^{-1}\Phi(Y)$$

and we conclude that $\Phi(X)$ if X is positive.

We now show that the theorem holds for arbitrary linear maps. If x is a unit vector in $\mathscr{M}_e$, define a linear functional ϕ_x on S by

$$\phi_x(A) = \langle x, \Phi(A)x\rangle.$$

This is a unital linear functional and since

$$|\phi_x(A)| \leq \|\Phi(A)\| \leq \|A\|,$$

we see that $\|\phi_x\| \leq 1$. Consequently ϕ_x is positive, and thus we have shown that if $AS \succcurlyeq 0$, then for each unit vector x,

$$0 \leq \phi_x(A) = \langle x, \Phi(A)x\rangle.$$

Therefore Φ is positive. □

17.5 Completely Positive Maps

A linear map $\Phi : \mathscr{M}_d \to \mathscr{M}_e$ is *completely positive* if $I_m \otimes \Phi$ is positive for all (positive integers) m. With the exception of transpose, all the examples of positive maps we have met are completely positive. We will give one concrete class of examples, after the following remarks.

Working from the assumption that $I_m \otimes \Phi$ is positive can take some getting used to, so we offer an alternative viewpoint. If F is an $md \times md$ matrix with blocks of order $d \times d$, we can identify $(I \otimes \Phi)F$ with matrix we get by applying Φ to each of the blocks of F. (The key is that $I \otimes \Phi$ is an operator on $\mathrm{Mat}_{md\times md}(\mathbb{C})$, it is not a matrix.)

17.5.1 Lemma. *If $\Phi(M) = \sum_r A_r M A_r^*$ for M in $\mathscr{M}_d$, then Φ is completely positive.*

Proof. Assume F is $md \times md$. If we define Ψ_r by $\Psi_r(M) = V_r M V_r^*$, then

$$(I_m \otimes \Psi_r)F = (I \otimes V_r)F(I \otimes V_r)^*.$$

(Note that here $I_m \otimes \Phi$ is an operator, the other terms are all matrices.) If $F \succcurlyeq 0$, then $(I \otimes V_r)F(I \otimes V_r)^* \succcurlyeq 0$. As $\sum_r \Psi_r = \Phi$, the lemma follows. □

Aside from providing a wide range of examples of completely positive maps, the significance of this lemma is that all completely positive maps arise from this construction. This result is due to Choi, and we turn to its proof.

We use $E_{i,j}$ to denote the elementary matrix $e_i e_j^*$ (of order $d \times d$). If

$$\Phi : \mathcal{M}_d \to \mathcal{M}_e$$

is linear, we define its *Choi matrix* $\Gamma(\Phi)$ to be the $de \times de$ matrix with d^2 blocks of order $e \times e$, where the ij-block is $\Phi(E_{i,j})$. If M is $d \times d$, then

$$M = \sum_{i,j} M_{i,j} E_{i,j}$$

from which we see that Φ is determined by its Choi matrix.

17.5.2 Lemma. *If $\Phi : \mathcal{M}_d \to \mathcal{M}_e$ is completely positive, its Choi matrix is positive semidefinite.*

Proof. If z is the vector in $\mathbb{C}^{d^2}$ given by

$$z = \begin{pmatrix} e_1 \\ \vdots \\ e_d \end{pmatrix}$$

then $(I \otimes \Phi)zz^*$ is the Choi matrix of Φ. As $zz^* \succcurlyeq 0$, it follows that if Φ is completely positive, its Choi matrix is positive semidefinite. □

To get some practice, show that the Choi matrix of transpose operating on $\mathrm{Mat}_{2\times 2}(\mathbb{C})$ is not positive semidefinite.

17.5.3 Theorem. *If $\Phi : \mathcal{M}_d \to \mathcal{M}_e$ is completely positive, there are $e \times d$ matrices $A_1, \ldots, A_{de}$ such that*

$$\Phi(M) = \sum_r A_r M A_r^*.$$

Proof. Let Γ denote $\Gamma(\Phi)$; since this is positive semidefinite we may write as a sum

$$\Gamma = \sum_r w_r w_r^*.$$

The vector w_r is formed from e blocks, each of length d. Let W_r be the $e \times d$ matrix with the i-th block of w_r as its i-th column. Then the ij-block of the matrix $w_r w_r^*$ is equal to $W_r E_{i,j} W_r^*$, and therefore the ij-block of Γ is equal to

$$\sum_r W_r E_{i,j} W_r^*,$$

and this is the value of $\Phi(E_{i,j})$. □

17.5.4 Corollary. *The map Φ is completely positive if and only if its Choi matrix is positive semidefinite.*

We note that a map $\Phi : \mathcal{M}_d \to \mathcal{M}_e$ is Hermitian preserving if and only if its Choi matrix is Hermitian. Hence it is diagonalizable, and it follows that we can write it in the form

$$\Gamma = \sum_r \lambda_r w_r w_r^*$$

where each λ_r's is an eigenvalue of Γ and w_r is the corresponding eigenvector. Following the proof of Choi's theorem we deduce the following (due to Hill).

17.5.5 Theorem. *A map $\Phi : \mathcal{M}_d \to \mathcal{M}_e$ is Hermitian preserving if and only there are matrices W_r and real numbers λ_r such that*

$$\Phi(M) = \sum_r \lambda_r A_r M A_r^*. \qquad \square$$

17.6 States

Recall that a state is positive linear functional with norm one.

If τ is a linear functional on an algebra A, then the map

$$(x, y) \mapsto \tau(x^* y)$$

is semilinear. Our next result is known as *Schwarz's inequality*.

17.6.1 Lemma. *If τ is a state on the algebra A and $x, y \in A$, then*

$$|\tau(x^* y)|^2 \le \tau(x^* x)\tau(y^* y).$$

Proof. For any complex numbers a and b, we have

$$\begin{pmatrix} \overline{a} & \overline{b} \end{pmatrix} \begin{pmatrix} \tau(x^* x) & \tau(x^* y) \\ \tau(y^* x) & \tau(y^* y) \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \tau((ax + by)^* (ax + by)) \ge 0,$$

whence we see that the matrix

$$\begin{pmatrix} \tau(x^* x) & \tau(x^* y) \\ \tau(y^* x) & \tau(y^* y) \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

is positive semidefinite and therefore its determinant is non-negative. □

We point out that if τ is a linear functional, the map

$$(x, y) \mapsto \tau(x^* y)$$

is sesquilinear. If τ is positive, it is Hermitian preserving and accordingly

$$\tau(y^* x) = \tau((x^* y)^*) = \tau(x^* y)^*$$

and so we have an inner product if the form is non-degenerate.

17.6.2 Theorem. *If τ is a linear functional of norm one on an algebra A, then τ is a state if and only if $\tau(1) = 1$.*

Proof. Assume τ is a state on A. Since τ is positive and $\|1\| = 1$, we have

$$0 \le \tau(1^*1) = \tau(1) \le \|\tau\|\|1\| = 1.$$

Thus $\tau(1) \le 1$, to show that $\tau(1) \ge 1$ we choose x in A such that $\|x\| \le 1$. Then $1 - xx^*$ is positive semidefinite and so

$$0 \le \tau(1 - xx^*) = \tau(1) - \tau(xx^*)$$

and hence $\tau(xx^*) \le \tau(1)$. Using the previous lemma, we find that

$$|\tau(x)|^2 = |\tau(1^*x)|^2 \le \tau(x^*x)\tau(1^*1) \le \tau(1)^2 \le 1.$$

Therefore $|\tau(x)| \le 1$ if $\|x\| \le 1$ and so $\|\tau\| \le 1$.

Now assume $\tau(1) = 1$. We claim that if $h \in A$ is Hermitian, that $\tau(h) \in \mathbb{R}$. For otherwise there is a Hermitian element h of A such that

$$\tau(h) = a + ib$$

where $a, b \in \mathbb{R}$ and $b \neq 0$. Then

$$\tau(b^{-1}(h - a)) = i$$

and if we set $z = b^{-1}(h - a)$ and $c \in \mathbb{R}$,

$$(c+1)^2 = |i + ci|^2 = \|\tau(z + ci1)\|^2 \le \|\tau\|^2\|c + zi1\|^2 = \|z + ci1\|^2.$$

As z is Hermitian,

$$\|z + ci1\|^2 = \|(z + ci1)^*(z + ci1)\| = \|z^2\| + c^2$$

from which it follows that $2c + 1 \le \|z^2\|$. This forces us to the conclusion that $\tau(h)$ is real if h is Hermitian.

Finally, if h is positive semidefinite and $\|h\| \le 1$, then $\|1 - h\| \le 1$ and so $\tau(1 - h) \le 1$. Since

$$\tau(h) = \tau(1 - (1 - h)) = \tau(1) - \tau(1 - h) = 1 - \tau(1 - h),$$

it follows that $\tau(h) \ge 0$. We conclude that τ is positive. □

17.7 Positive Definite Block Matrices

We consider $md \times nd$ matrices, viewed as $m \times n$ with $d \times d$ blocks as entries; equivalently as $m \times n$ matrices over the algebra of $d \times d$ matrices. The following theorem is not deep, but it will simnplify our calculations.

17.7.1 Theorem. *Let A be a matrix of order $nd \times nd$. The following statements are equivalent:*

(a) $A \succcurlyeq 0$.

(b) *There are* $nd \times d$ *matrices* $X_1, \dots, X_m$ *such that* $A = \sum_i X_i X_i^*$.

(c) *For any* $nd \times d$ *matrix* Y*, we have* $Y^* AY \succcurlyeq 0$.

Proof. If $A \succcurlyeq 0$, then $A = MM^*$ for some $nd \times nd$ matrix M. By viewing M as a block matrix, we find that (a) implies (b). Similarly it is easy to see that (b) implies (c). If $z \in \mathbb{C}^d$, then

$$z^* Y^* AYz = (Yz)^* A(Yz) \succcurlyeq 0,$$

whence $Y^* AY \succcurlyeq 0$. □

17.8 Conditional Expectations

We will consider linear maps from an algebra to a subalgebra. The grown-up version of these results are stated in terms of C^*-algebras, but we restrict ourselves to *-closed subalgebras of full matrix algebras, i.e., to *-closed subalgebras of $\mathrm{Mat}_{n\times n}(\mathbb{C})$. We point out that our algebras (and subalgebras) always have an identity element.

First some remarks about block matrices. Suppose A is $md \times md$, viewed as an $m \times m$ matrix with blocks of size $d \times d$. Let $[A]_{i,j}$ denote the ij-block of A. If $A \succcurlyeq 0$, then $A = C^* C$, for some $md \times md$ matrix C and we can view C in turn as a block matrix (with blocks of size $d \times d$). Then

$$[A]_{i,j} = \sum_{r=1}^{m} [C]_{r,i}^* [C]_{r,j},$$

from which it follows that there are $d \times md$ matrices $D_1, \dots, D_m$ such that

$$A = \sum_r D_r^* D_r.$$

If B is $md \times d$ with $d \times d$ blocks, then

$$B^* AB = \sum_{i,j} \sum_r [B]_i^* C_{r,i}^* C_{r,j} [B]_j = \sum_r \Big(\sum_i C_{r,i} B_i\Big)^* \Big(\sum_j C_{r,jB_j}\Big) \succcurlyeq 0.$$

Suppose $\mathscr{A}$ and $\mathscr{B}$ are algebras as above and $\mathscr{A} \le \mathscr{B}$. Then $\mathscr{B}$ is a bimodule over $\mathscr{A}$ (an algebra is just an up-market ring). A linear map $E : \mathscr{B} \to \mathscr{A}$ is a bimodule map if, given a_1 and a_2 in $\mathscr{A}$ and b in $\mathscr{B}$, we have

$$E(a_1 b a_2) = a_1 E(b) a_2.$$

We say E is a *conditional expectation* from B onto A if $E(a) = a$ for all a in A and it is:

(a) completely positive and contractive,

(b) a bimodule map.

The following result is a form of Tomiyama's theorem (which holds for matrices over C^*-algebras).

17.8.1 Theorem. *Suppose $\mathscr{A}$ and $\mathscr{B}$ are algebras and $\mathscr{A} \le \mathscr{B}$. Let $E : \mathscr{B} \to \mathscr{A}$ be a linear map such that $E(a) = a$ for all a in A. Then if E is a contraction, it is a conditional expectation.*

Proof. Assume E is a contraction. We first prove that this implies E is a bimodule map. Since any element of a $*$-closed matrix algebra is a linear combination of projections, it well suffice to prove that if p is a projection in A, then $E(pb) = pE(b)$ and $E(bp) = E(b)p$ for all b in B.

So assume p in A is a projection and that $p^\perp := 1_B - p$. Since E acts on A as the identity, for all b in B.

$$pE(p^\perp b) = E(pE(p^\perp b)).$$

Hence, for any t in $\mathbb{R}$,

$$\begin{aligned}(1+t)^2\|pE(p^\perp b)\|^2 &= \|pE(p^\perp b + tpE(p^\perp b))\|^2\\ &\le \|p^\perp b + tpzE(p^\perp b)\|^2\\ &\le \|p^\perp b\|^2 + t^2\|pE(p^\perp b)\|\end{aligned}$$

and therefore

$$\|pE(p^\perp b)\|^2 + 2t\|pE(p^\perp b)\|^2 \le \|p^\perp b\|^2$$

for all real t. This implies that $pE(p^\perp b) = 0$, and the same reasoning shows that $(1_A - p)E(p^\perp b) = 0$.

As 1_A is a projection,

$$0 = 1_A E(1_A^\perp b) = E(1_A^\perp b).$$

Accordingly

$$E(px) = pE(pb) = pE(b - p^\perp b) = p(Eb)$$

for each projection p in A and each b in B. Swapping sides, we also find that $E(bp) = E(b)p$ and hence E is a bimodule map.

Since E is a unital linear map with norm 1, it is positive by Theorem 17.4.1. It remains for us to prove that it is completely positive, which means we must show that if $M \in \mathrm{Mat}_{m\times m}(\mathscr{B})$ and $M \succcurlyeq 0$, then $(I \otimes E)(M) \succcurlyeq 0$.

If $M \succcurlyeq 0$, there are matrices $X_1, \dots, X_m$ with entries from $\mathscr{B}$ such that

$$M = \sum_i X_i X_i^*.$$

Then

$$((I \otimes E)M)_{i,j} = E(X_i X_j^*).$$

Now if $Y_1, \dots, Y_m$ are matrices in $\mathscr{B}$, we have

$$\sum_{i,j} Y_i^* E(X_i X_j^*) Y_j = \sum_{i,j} E(Y_i^* X_i X_j^* Y_j) = E\Big(\Big(\sum_i X_i^* Y_i\Big)^* \sum_j X_j^* Y_j\Big)$$

and, since E is positive, the last term is positive. (Note that here we have made use of Theorem 17.7.1) and of the fact that E is a bimodule map.) □

Part IV

Geometry

18
Lines and Frames

We study some geometric questions related to the geometry of lines.

18.1 Equiangular Lines

We work in the vector space V, which is $\mathbb{R}^d$ or $\mathbb{C}^d$ with the usual Euclidean inner product. If x and y are nonzero vectors, the *cosine* of the angle between the lines spanned by x and y is

$$\frac{|\langle x, y\rangle|}{\|x\|\|y\|}.$$

We will often work with the squared cosine

$$\frac{\langle x, y\rangle\langle y, x\rangle}{\langle x, x\rangle\langle y, y\rangle}.$$

A set of lines in V is *equiangular* is the cosine of the angle between any two distinct lines is the same.

18.1.1 Theorem. *The maximum size of a set of equiangular lines in $\mathbb{C}^d$ is d^2; in $\mathbb{R}^d$ it is $\binom{d+1}{2}$.*

Proof. Suppose we have lines spanned by unit vectors $x_1, \dots, x_m$. Define matrices $P_1, \dots, P_m$ by

$$P_r = x_r x_r^*.$$

Then P_r represents orthogonal projection onto the line spanned by x_r, and if $r \neq s$,

$$\langle P_r, P_s\rangle = \operatorname{tr}(P_r P_s) = \langle x_r, x_s, \langle\rangle, x_s\rangle x_r = |\langle x_r, x_s\rangle|^2.$$

We assume that $\alpha = |\langle P_r, P_s\rangle|$. We see also that $\langle P_r, P_r\rangle = 1$ for all r.

The projections P_r lie in the space of Harmitian matrices. If G is their Gram matrix, then

$$G = (1-\alpha^2)I + \alpha^2 J.$$

We can prove, in a number of ways, that G is invertible, which implies that the matrices $P_1, \dots, P_m$ form a linearly independent set in the space of Hermitian matrices. We complete the proof by noting that this space has dimension d^2 (over $\mathbb{C}$) and the dimension in the real case is $\binom{d+1}{2}$. □

In $\mathbb{R}^2$ it is easy to find three lines with pairwise cosine $1/2$, and the diagonals of the icosahedron give six lines with pairwise cosine $1/\sqrt{5}$. Examples of sets of size $\binom{d+1}{2}$ are known in $\mathbb{R}^d$ when $d = 7$ and $d = 23$. In the complex case, examples of tight sets are known for d in $\{1,\dots,15,19,24,35,48\}$.

18.2 Tight Frames

Suppose we have a set of equiangular lines of maximum size. Then the associated projections $P_1,\dots,P_m$ form a basis for the space of Hermitian matrices. Hence there are scalars c_r such that

$$I = \sum_r c_r P_r.$$

If we multiply both sides by P_k and take traces, we get

$$1 = (1-\alpha^2)c_k + \alpha^2 \sum_r c_r.$$

It follows that $c_1 = \cdots = c_m$ and hence that

$$I = \frac{d}{m}\sum_r P_r.$$

In a slightly different format, we have established that if $x_1,\dots,x_m$ are unit vectors spanning a set of equiangular lines of maximum size, then

$$\sum_r x_r x_r^* = \frac{m}{d} I.$$

Such a set of vectors is an example of a *tight frame*.

We will see that tight frames are more common than set of lines meeting the absolute bound. Consider a set of projections $P_1,\dots,P_m$ corresponding to a set of equiangular lines with squared cosine α^2, and define

$$M = \sum_r P_r - \frac{m}{d} I.$$

Then

$$\begin{aligned} 0 \le \langle M, M\rangle &= \langle \sum_r P_r, \sum_r P_r\rangle - \frac{2m}{d}\langle \sum_r P_r, I\rangle + \frac{d^2}{m^2}\operatorname{tr}(I) \\ &= m + m(m-1)\alpha^2 - \frac{m^2}{d}. \end{aligned}$$

If equality holds we have

$$\alpha^2 = \frac{m-d}{md-d}.$$

This yields the following, sometimes known as the *relative bound*.

18.2.1 Theorem. *If there is a set of m lines in $\mathbb{F}^d$ with squared cosine α^2, where $d\alpha^2 < 1$, then*

$$m \le \frac{d - d\alpha^2}{1 - d\alpha^2}.$$

If equality holds, then a set of unit vectors spanning the lines forms a tight frame. □

Note that if we have d^2 lines in $\mathbb{C}^d$, then $\alpha^2 = (d+1)^{-1}$, and for $\binom{d+1}{2}$ lines in $\mathbb{R}^d$, then $\alpha^2 = (d+2)^{-1}$.

18.3 Another Gram Matrix

Suppose $x_1, \dots, x_m$ form a tight frame in dimension d. Then

$$\sum_r x_r x_r^* = \frac{m}{d} I.$$

If U is the $d \times m$ matrix with the vectors $x_1, \dots, x_m$ as its columns then we have

$$UU^* = \sum_r x_r x_r^* = \frac{m}{d} I,$$

which implies that the rows of U are orthogonal (and of the same length).

Set $H = U^* U$. Then

$$H^2 = U^* U U^* U = \frac{m}{d} U^* U = \frac{m}{d} H$$

and therefore the minimal polynomial of H divides

$$t\left(t - \frac{m}{d}\right).$$

(If the minimal polynomial is a proper divisor of this polynomial that $H = 0$ or $H = I$.) We can write H as $I + \alpha S$, where S is Hermitian with diagonal entries zero and all off-diagonal entries have absolute value 1. (In the real case, this means the off-diagonal entries are ±1.) The eigenvalues of S are

$$\frac{1}{\alpha}\left(\frac{m}{d} - 1\right), \quad -\frac{1}{\alpha}$$

with respective multiplicities d and $m - d$.

18.4 The Orthogonal Group

Let V be a vector space with a bilinear form. We say that an endomorphism A of V *preserves* the form if $\langle Ax, Ay \rangle = \langle x, y \rangle$, for all x and y. If the form is symmetric and the characteristic of our field is odd, then

$$\langle x, y \rangle = \frac{1}{2}(\langle x + y, x + y \rangle - \langle x, x \rangle - \langle y, y \rangle.$$

Hence A preserves the form if and only if $\langle Ax, Ax \rangle = \langle x, x \rangle$ for all x.

Now assume V is $\mathbb{R}^n$ and that our form is the dot product. A matrix which preserves dot product is called *orthogonal*. If v and w are orthogonal vectors in V and A is orthogonal, then Av and Aw are orthogonal.

18.4.1 Lemma. *A matrix A is orthogonal if and only if $A^T A = I$.*

Proof. If $v_1, \dots, v_n$ is an orthogonal basis for V, then so is $Av_1, \dots, Av_n$. Since the standard basis $e_1, \dots, e_n$ for V is orthogonal, it follows that

$Ae_1, \dots, Ae_n$ is an orthogonal set of vectors. Therefore the columns of an orthogonal matrix A form an orthogonal basis. This also implies that

$$A^T A = I.$$

Since A is square, we see that $A^T = A^{-1}$ and $AA^T = I$. Conversely, if $A^T = A^{-1}$, then

$$\langle Av, Aw\rangle = (Av)^T Aw = v^T A^T Aw = v^t w = \langle v, w\rangle. \qquad \square$$

We see from this result that, if A is orthogonal, then it columns form an orthonormal set. Also, if A is orthogonal, then $A^T = A^{-1}$ and therefore $AA^T = I$. Hence the rows of A also form an orthonormal set.

We consider the complex version of orthogonal matrices. A complex matrix is *unitary* if it preserves the complex dot-product. This means that

$$y^* x = (Ay)^*(Ax) = y^* A^* Ax$$

for all x and y, and hence that

$$A^* A = I.$$

A real matrix is unitary if and only if it is orthogonal.

We turn to examples of orthogonal matrices. Any permutation matrix is orthogonal, and a diagonal matrix A is orthogonal if and only if $A_{i,i} = \pm 1$ for all i. The matrices

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

are orthogonal, for any value of θ. It is easy to verify that the product of two orthogonal matrices is orthogonal, and that the inverse of an orthogonal matrix is orthogonal. Therefore the set of all orthogonal matrices is an example of a *group*, known as the *orthogonal group*.

18.5 Skew-Symmetric to Orthogonal

We define a matrix A to be *skew symmetric* if $A^T = -A$ and $A_{i,i} = 0$ for all i. (The last condition is only needed if our field has characteristic two.) The set of $n \times n$ skew-symmetric matrices is a subspace of the space of square matrices.

18.5.1 Lemma. *If S is a real skew-symmetric matrix, then $(I-S)^{-1}(I+S)$ is orthogonal.*

Proof. We first show that $I - S$ is invertible for all real t. Suppose $x \neq 0$ and $Ax = \theta x$. Then

$$\theta x^T x = x^T Ax = (A^T x)x = (-Ax)^T x = (-\theta x)^T x = -\theta x^T x.$$

It follows that 0 is the only possible real eigenvalue for S. Therefore $I - S$ is invertible for all real t and we can define

$$M := (I - S)^{-1}(I + S).$$

The matrices $I + S$ and $I - S$ commute, and from this it follows that $I + S$ and $(I - S)^{-1}$ commute. Hence we find that

$$\begin{aligned} M^T = (I + S^T)(I - S^T)^{-1} &= (I - S)(I + S)^{-1} \\ &= (I + S)^{-1}(I - S) \\ &= M^{-1}. \end{aligned}$$

Therefore M is orthogonal. □

The matrix M above is sometimes known as the *Cayley transform* of S. Note that, since tS is skew-symmetric if S is, the matrix

$$(I - tS)^{-1}(I + tS)$$

is orthogonal for real t.

(1) If H is hermitian and $S = iH$, show that $(I - S)^{-1}(I + S)$ is unitary.

18.6 Reflections

Suppose a is a fixed non-zero vector in V. Define the map ρ_a by

$$\rho_a(v) = v - 2\frac{\langle a, v\rangle}{\langle a, a\rangle} a.$$

Note that ρ_a is the sum of two linear mappings (the identity and a scalar multiple of the orthogonal projection onto the line spanned by a) and therefore it is linear. We check that

$$\rho_a(a) = -a$$

and, using this, that

$$\rho_a^2 = I.$$

If $v \in a^\perp$, then $\rho_a(v) = v$. It follows that ρ_a corresponds to the geometric operation of reflection in the hyperplane perpendicular to a.

We have

$$\langle \rho_a(v), \rho_a(v)\rangle = \langle v, v\rangle - 4\frac{\langle a, v^2\rangle}{ipaa} + 4\frac{\langle a, v\rangle}{\langle a, a\rangle}^2 \langle a, a\rangle = \langle v, v\rangle.$$

Therefore ρ_a is orthogonal. The matrix R_a representing it is given by

$$R_a = I - \frac{2}{\langle a, a\rangle} aa^T.$$

If v and w have the same length then

$$\langle v-w, v+w\rangle = 0.$$

Therefore R_{v-w} fixes $v+w$ and maps $v-w$ to $w-v$. Consequently

$$R_{v-w}(2v) = R_{v-w}((v+w)+(v-w)) = 2w,$$

and so, after a very modest amount of extra work, we find that R_{v-w} swaps v and w.

18.6.1 Theorem. *Every non-identity orthogonal matrix is a product of at most n matrices R_a.*

Proof. If A is a matrix, let $F(A)$ be the subspace

$$\{v \in V : Av = v\}.$$

We prove by induction that A is the product of at most $\dim(V) - \dim(F(A))$ matrices R_a.

Suppose A is orthogonal and $\dim(F(A)) = k$. If $k = \dim V$, then $A = I$. Suppose $k < \dim V$, and let v be a vector in V such that $Av \neq v$. If $w := Av$ and $Ax = x$, then $\langle v, x\rangle = \langle w, x\rangle$ and so $\langle v-w, x\rangle = 0$. Therefore $F(A) \subseteq (v-w)^{\perp}$, and R_{v-w} fixes each vector in $F(A)$. Now R_{v-w} swaps v and W, whence the product $R_{v-w}A$ fixes each vector in $F(A)$, and fixes v. As $v \notin F(A)$, we see that

$$\dim(F(R_{v-w}A)) > \dim(F(A)).$$

The lemma follows. □

A matrix A is an involution if $A^2 = I$. Diagonal matrices with diagonal entries equal to ± 1 provide a fairly trivial class of examples. If P is an idempotent then

$$(I-2P)^2 = I - 4P + 4P = I,$$

and thus $I - 2P$ is an involution.

18.6.2 Theorem. *Every orthogonal matrix is the product of two involutions.*

Proof. We actually prove a stronger result: A and A^{-1} are similar if and only if A is the product of two involutions. Since any square matrix is similar to its transpose, orthogonal matrices satisfy this condition.

Suppose $S^2 = T^2 = I$ and $A = ST$. Then $(ST)(TS) = I$ and

$$S^{-1}AS = SAS = S(ST)S = TS.$$

Therefore a product of two involutions is similar to its inverse.

So assume now that A and A^{-1} are similar and let F be the Frobenius normal form of A. By **??**, there is a permutation matrix T such that $T^2 = I$ and

$$F^{-1} = TFT.$$

Then $I = FTFT$, whence FT and T are involutions whose product is F. As any matrix that is similar to an involution is an involution, the general result follows. □

19
Isoclinic Subspaces, Covers and Codes

19.1 Isoclinic Subspaces

Let U and V be two k-dimensional subspace of an inner product space W, and let P and Q be the corresponding orthogonal projections. Then P maps the unit sphere in V to an ellipsoid in U. The shape of this ellipsoid is determined by the extreme points of the function

$$\|Pv\|^2 = v^* P^* Pv = v^* Pv,$$

where v runs over the unit vectors in V. We say that V is *isoclinic* to U is there is a constant λ such that

$$v^* Pv = \lambda v^* v.$$

If V is isoclinic to U with parameter λ, then

$$x^* Q^* PQx = \lambda x^* Q^* QX = \lambda x^* Qx$$

for all x in w. Hence we see see that U and V are isoclinic with parameter λ if and only if

$$QPQ = \lambda Q.$$

Thus we have translated a geometric condition into a linear algebraic one. Our next result shows that is a symmetric relation.

19.1.1 Lemma. *The subspace U is isoclinic to V if and only if V is isoclinic to U.*

Proof. Let R be a matrix whose columns form an orthonormal basis for U, and let S be a matrix whose columns form an orthonormal basis for V. Then

$$RR^* = P, \quad SS^* = Q$$

and

$$QPQ = SS^* RR^* SS^* = S(S^* RR^* S)S^*.$$

If $QPQ = \lambda Q$, then it follows that

$$\lambda SS^* = S(S^* RR^* S)S^*$$

and therefore

$$\lambda I = S^* S(S^* RR^* S)S^* S = S^* RR^* S.$$

Hence $R^* SS^* R = \lambda I$ and so

$$\lambda P = \lambda RR^* = R(R^* SS^* R)R^* = PQP. \qquad \square$$

Note that $\mathrm{tr}(PQP) = \mathrm{tr}(QPQ)$, and so if $\mathrm{rk}(P) = \mathrm{rk}(Q)$ and $QPQ = \lambda P$, then $PQP = \lambda P$. A consequence of the proof is that U and V are isoclinic if and only the matrix $\lambda^{-1/2} R^* S$ is orthogonal.

As exercises, prove that if P and Q are projections then $(P-Q)^2$ commutes with P and Q. Also if U and V are isoclinic with parameter λ, then

$$(P-Q)^3 = (1-\lambda)(P-Q).$$

This implies that the eigenvalues of $P-Q$ are

$$0,\ \pm\sqrt{1-\lambda};$$

since $\mathrm{tr}(P-Q) = 0$, the non-zero eigenvalues have equal multiplicity.

19.2 Matrices

We investigate sets of pairwise isoclinic k-subspaces in $\mathbb{R}^n$. Let U be the column space of the matrix

$$R = \begin{pmatrix} I_k \\ 0 \end{pmatrix}.$$

Suppose S is the $n \times k$ matrix

$$S = \begin{pmatrix} Y \\ Z \end{pmatrix}$$

where $S^* S = I_k$. Then the column spaces of R and S are λ-isoclinic if and only if

$$\lambda I = S^* RR^* S = Y^* Y.$$

Since

$$I = S^* S = Y^* Y + Z^* Z$$

we then have $Z^* Z = (1-\lambda)I$. If

$$T = \begin{pmatrix} \lambda^{1/2} I \\ \lambda^{-1/2} ZY^* \end{pmatrix}$$

then $T = \lambda^{-1/2} SY^*$, so $\mathrm{col}(T) = \mathrm{col}(S)$ and $T^* T = I$.

19.2.1 Lemma. *If V is λ-isoclinic to the column space of*

$$\begin{pmatrix} I_k \\ 0 \end{pmatrix}$$

then V is the column space of a matrix

$$\begin{pmatrix} \lambda^{1/2} I_k \\ \lambda^{-1/2} Z \end{pmatrix}$$

where $Z^* Z = (1-\lambda)I$. □

Assume $Y^*Y = Z^*Z = (1-\lambda)I$. Then the column spaces of the matrices

$$\begin{pmatrix} \lambda^{1/2} I_k \\ \lambda^{-1/2} Y \end{pmatrix}, \quad \begin{pmatrix} \lambda^{1/2} I_k \\ \lambda^{-1/2} Z \end{pmatrix}$$

are ν-isoclinic if and only if the matrix

$$\nu^{-1/2}(\lambda I + \lambda^{-1} Y^* Z)$$

is orthogonal.

19.3 Equiangular Subspaces

Suppose that $P_1, \ldots, P_m$ are projections onto e-dimensional subspaces of d-dimensional vector space. We say that they are *equiangular* if there is a scalar α^2 such that

$$\operatorname{tr}(P_i P_j) = \alpha^2$$

whenever $i \neq j$. We note that

$$\operatorname{tr}(P-Q)^2 = 2e - 2\operatorname{tr}(PQ)$$

where $\operatorname{tr}(P-Q)^2$ is the Euclidean distance between the matrices P and Q. So we could have used "equidistant" in place of "equiangular".

19.3.1 Lemma. *An equiangular set of projections is linearly independent.*

Proof. Suppose we have scalars $c_1, \ldots, c_m$ such that

$$0 = \sum_i c_i P_i.$$

Then

$$0 = \sum_i \operatorname{tr}(P_r P_i) = c_r e + \alpha^2 \sum_{i \neq r} c_i = e(c_r - \alpha^2) + \alpha_2 \sum_i c_i.$$

From this we deduce that c_r is independent of r and hence that $c_r = 0$ for all r. □

The projections P_i are Hermitian and so, if we work over $\mathbb{C}$, they lie in a real vector space of dimension d^2. Over $\mathbb{R}$ they lie in a space of dimension $d(d+1)/2$. These upper bounds are known as the *absolute bounds*. The bound supplied by the following theorem is the *relative bound*.

19.3.2 Theorem. *If the projections $P_1, \dots, P_m$ are equiangular with angle α^2 and $d\alpha^2 \le e$, then*

$$m \le \frac{d(e-\alpha^2)}{e^2 - d\alpha^2},$$

equality holds if and only if

$$\sum_i P_i = \frac{me}{d} I.$$

Proof. We set

$$S := \sum_i \Big(P_i - \frac{e}{d} I\Big)$$

Then $S = S^*$ and therefore $\operatorname{tr}(S^2) \ge 0$, which yields

$$\begin{aligned} 0 &\le \sum_i \operatorname{tr}\Big(P_i - \frac{e}{d}I\Big)^2 + \sum_{i \ne j} \operatorname{tr}\Big[\Big(P_i - \frac{e}{d}I\Big)\Big(P_j - \frac{e}{d}I\Big)\Big] \\ &= m\Big(e - \frac{e^2}{d}\Big) + m(m-1)\Big(\alpha^2 - \frac{e^2}{d}\Big). \end{aligned}$$

Our bound follows from this. If equality holds that $\operatorname{tr}(S^2) = 0$ and therefore $S = 0$. □

If P and Q are projections onto isoclinic spaces with parameter λ, then

$$\lambda e = \operatorname{tr}(\lambda P) = \operatorname{tr}(PQP) = \operatorname{tr}(PQ) = \alpha^2.$$

Thus $\lambda = \alpha^2/e$ and our expression for m becomes

$$m = \frac{d(1-\lambda)}{e - d\lambda}.$$

This bound (for equi-isoclinic subspaces) is due to Lemmens and Seidel. They also note that the absolute bound cannot be tight if $e > 1$, because the projections P_i lie in the subspace of mappings Q such that $P_1 Q P_1$ is a scalar multiple of P_1 and this has codimension $e(e+1)/2$.

A set $P_1, \dots, P_m$ of projections with rank e such that

$$\sum P_i = \frac{me}{d}$$

is known as a *tight fusion frame*. If $e = 1$, it is a *tight frame*.

If R_i is a matrix whose columns form an orthonormal basis for $\operatorname{im}(P_i)$, then

$$P_i = R_i R_i^*.$$

So if $\sum_i P_i = (me/d)I$, then

$$\frac{me}{d} I = \sum_i R_i R_i^*.$$

If $\mathscr{R}$ denotes the $d \times me$ matrix

$$\begin{pmatrix} R_1 & \dots & R_m \end{pmatrix}$$

then

$$\mathscr{R}\mathscr{R}^* = \sum_i R_i R_i^* = \frac{me}{d} I$$

and accordingly $\mathscr{R}^*\mathscr{R}$ is a scalar multiple of a projection of order $me \times me$. (It has a block decomposition where the ij-block is $R_i^* R_j$; this block is a scalar multiple of an orthogonal matrix.

19.4 Error Correction

Let $\mathscr{C}$ be an e-dimensional subspace of $\mathbb{C}^d$. A matrix A in $U(d)$ is *detectable* if for any two vectors x and y in $\mathscr{C}$, we have $\langle x, y\rangle = 0$ if and only if $\langle x, Ay\rangle = 0$. We note that $A^{-1} = A^*$ is detectable if and only if A is.

If $x, y \in \mathscr{C}$, then

$$\langle x, Ay\rangle = \langle Px, APy\rangle = \langle x, PAPy\rangle$$

and hence PAP maps $\mathscr{C}$ to itself. Therefore A is detectable if and only if PAP maps $x^\perp \cap \mathscr{C}$ into itself, for each x in $\mathscr{C}$.

19.4.1 Theorem. *Let $\mathscr{C}$ be an e-dimensional subspace of $\mathbb{C}^d$, where $e \ge 3$. A matrix A is detectable if and only if U and AU are isoclinic.*

Proof. Let P represent orthogonal projection onto U and assume A is detectable. Then PAP fixes $\mathscr{C}$ and fixes the subspace $x^\perp \cap \mathscr{C}$ for each x in $\mathscr{C}$. Thus it fixes each hyperplane in $\mathscr{C}$, and therefore it must be a scalar matrix. If the columns of R are an orthonormal basis for $\mathscr{C}$, we have

$$\alpha I = PAP = RR^* ARR^*$$

and hence $R^* AR = \lambda I$. Now

$$PA^* PAP = RR^* A^* RR^* ARR^* = \alpha\bar{\alpha} P$$

and we conclude that $\mathscr{C}$ and $A\mathscr{C}$ are isoclinic.

We turn to the converse. Assume $A \in U(d)$ and $\mathscr{C}$ and $A\mathscr{C}$ are isoclinic. Assume further that R is a $d \times e$ matrix whose columns form an orthonormal basis for $\mathscr{C}$. Then RR^* represents projection onto $\mathscr{C}$ and $A^* RR^* A$ represents projection onto $A\mathscr{C}$. Therefore

$$\lambda RR^* = RR^* A^* RR^* ARR^*$$

and accordingly

$$\lambda I = R^* A^* RR^* AR.$$

This implies that $\lambda^{-1/2} R^* AR$ is unitary. Assume $x = Rw$ and $y = Rz$. Then $x, y \in \mathscr{C}$ and $x \perp y$, then $x^* Ay = 0$ (because A is unitary). We conclude that A is detectable. □

19.5 Isoclinic Subspaces from Covers

Two subsets of $\mathbb{C}^d$ are *congruent* if there is a unitary mapping that takes the first subset to the second. If the subsets are finite, and are given as the columns of matrices M and N, then they are congruent if and only if there is a unitary matrix A and a permutation matrix P such that $AMP = N$.

19.5.1 Lemma. *Two spanning sets of vectors $x_1, \dots, x_m$ and $y_1, \dots, y_m$ are congruent if and only if their Gram matrices are permutation equivalent.* □

Proof. Let U_1 and U_2 be the matrices with the vectors $x_1, \dots, x_m$ and $y_1, \dots, y_m$ repectively as columns. Reordering the columns of U_1 as needed, the two sets of vectors are congruent if and only if there is an orthogonal matrix Q such that $QU_1 = U_2$. If such Q exists,

$$U_2^T U_2 = U_1^T Q^T Q U_1 = U_1^T U_1$$

and the Gram matrices are equal.

So now we assume that $U_1^T U_1 = U_2^T U_2$. Since our vectors span, the rows of U_1 are linearly independent and hence U_1 has a right inverse R. Then

$$I = R^T U_1^T U_1 R = R^T U_2^T U_2 R$$

and therefore $Q = U_2 R$ is orthogonal.

Next, since $U_1 R = I$, the matrix RU_1 is idempotent and, as $U_1 R U_1 = U_1$, it acts as the identity on the row space of U_1.

We now note that, since $U_1^T U_1 = U_2^T U_2$, the row spaces of U_1 and U_2 are equal. Therefore

$$QU_1 = U_2 R U_1 = U_2.$$ □

A set of vectors $x_0, \dots, x_r$ in $\mathbb{R}^d$ forms a *regular r-simplex* if its Gram matrix is a non-zero scalar multiple of $rI_r - J_r$. The vectors $x_0, \dots, x_r$ are the *vertices* of the simplex. The span of a regular r-simplex has dimension $r-1$. Any two regular r-simplices are congruent, in fact any bijection from the vertices of one simplex to the vertices of the other extends to an orthogonal mapping (by the previous lemma).

Suppose $\mathscr{C}$ and $\mathscr{D}$ are subspaces with dimension e, with associated projections P and Q respectively. Then $\mathscr{C}$ and $\mathscr{D}$ are isoclinic if the restriction of P to $\mathscr{D}$ is a scalar multiple of an orthogonal operator.

We can construct isoclinic subspaces from antipodal distance regular graphs. Suppo0se X is distance-regular on n vertices. Assume θ is an eigenvalue of X with multiplicity d and corresponding spectral idempotent E. If $u \in V(X)$, then the map $u \mapsto Ee_u$ assigns a vector in $\mathbb{R}^m$ to each vertex of X—we call it a representation of X on the θ-eigenspace of X. Since X is distance regular, $E_{u,v}$ is determined by the distance between x and y in X, in particular the vectors Ee_u all have same length (namely $\sqrt{d/n}$).

19.5.2 Theorem. *Let X an antipodal distance-regular graph with fibres of size r and let θ be an eigenvalue of X that is not an eigenvalue of the quotient. Then the images of the fibres under the representation on the θ-eigenspace are pairwise isoclinic subspaces of dimension $r-1$. The parameter of isoclinism is determined by the distance between the fibres in X.*

Proof. Let F be a fibre with vertices $1,\dots,r$, set $E = E_\theta$ and $y_i = Ee_i$ for $i \in F$. Let $\mathscr{F}$ denote the span of the vectors Ea_i.

As

$$\sum_i y_i = (A_d + I)y_1$$

we see that

$$\sum_i y_i = E(A_d + I)y_1.$$

Since $r^{-1}(A_d + I)$ is an idempotent, representing projection onto the space of vectors constant on the fibres of X, it follows that $E(I + Y_d) = 0$. We conclude that vectors y_i sum to zero and, since the vertices in F are pairwise equidistant, their image is a regular simplex.

Suppose b is a vertex in X at distance i from F and that $2i < D$. Set $x = Ee_b$. Assume b is at distance i from a_1; then it is at distance $D-i$ from each of $a_2,\dots,a_r$. Accordingly

$$0 = \langle x, \sum_{i=1}^r y_i\rangle = \langle x, y_1\rangle + (r-1)\langle x, y_2\rangle$$

and similarly

$$0 = \langle y_1, \sum_{i=1}^r y_i\rangle = \langle y_1, y_1\rangle + (r-1)\langle y_1, y_2\rangle$$

Now we calculate that

$$x - \frac{\langle y_1, x\rangle}{\langle y_1, y_1\rangle} y_1$$

is orthogonal to the vectors $y_1 - y_i$ for $i = 2,\dots,r$, and therefore the vector

$$\frac{\langle y_1, x\rangle}{\langle y_1, y_1\rangle} y_1$$

is the projection of x onto $\mathscr{F}$.

Since each vertex in the fibre of x is at distance i from a vertex in F, we deduce that orthogonal projection P onto $\mathscr{F}$ maps the regular simplex spanned by the fibre of X onto $\alpha = \langle y_1, x\rangle / \langle y_1, y_1\rangle$ times the image of F. Therefore the restriction of $\alpha^{-1}P$ to the span of the fibre of x is an orthogonal mapping, and so the spans of two fibres at distance i are isoclinic.

If $2i = d$, then x is at the same distance from each vertex in $\mathscr{F}$, whence $\langle x, y_i\rangle = 0$ and therefore the images of distinct fibres are orthogonal subspaces—still isoclinic. □

If X is a distance-regular antipodal r-fold cover of Y, then fibres in the preimage of a clique in Y give rise to a set of equi-isoclinic subspaces of dimension $r-1$.

A distance-regular antipodal r-fold cover of $K_{n,n}$ has diameter four. It follows that that the images of the fibres corresponding to vertices in one of the colour classes are pairwise orthogonal. The eigenvalues of this cover are the eigenvalues of $K_{n,n}$ and $\pm\sqrt{n}$, each with multiplicity $(r-1)n$. Hence the images of the fibres in a given colour class form an orthogonal decomposition of $\mathbb{R}^{(r-1)n}$ into n subspaces of dimension $(r-1)$.

19.6 Equi-isoclinic Subspaces and Unitary Covers

Let $\mathscr{C}_1,\dots,\mathscr{C}_m$ be a set of pairwise λ-isoclinic e-dimensional subspaces of $\mathbb{C}^d$, and let $R_1,\dots,R_m$ be $d\times e$ matrices such that $R_i^*R_i = I_e$ and $P_i = R_iR_i^*$ is the projection onto $\mathscr{C}_i$. Let G be the $me\times me$ block matrix with ij-block equal to $R_i^*R_j$; we might privately think of G as a kind of Gram matrix.

We see that $G^* = G$. The projections $P_1,\dots,P_m$ form a tight fusion frame if and only if G is a scalar times an idempotent. The subspaces $\mathscr{C}_1,\dots,\mathscr{C}_m$ are pairwise isoclinic if and only if each block is a scalar times a unitary matrix. If the subspaces are pairwise λ-isoclinic, then each off-diagonal block of $\lambda^{-1/2}(G-I)$ is unitary. Thus a set of m pairwise equi-isoclinic d-dimensional subspaces determines a map d from the arcs of the complete graph K_m into the unitary group into the unitary group, such that $f(i,j)f(j,i)=1$ for each arc ij. We call it a *unitary arc function* on K_m. We extend f to a function on the walks in K_m: if $w = v_0\cdots v_n$ is a walk, then

$$f(w) = f(v_0v_1)\cdots f(v_{n-1}v_n).$$

If $A_1,\dots,A_m$ are matrices from $U(d)$, then the function

$$(i,j)\mapsto A_i^* f(i,j)A_i$$

is a function on the arcs that takes the same value on closed walks that f does. The corresponding block matrix G is similar to G. It follows that we may assume that f takes the value I on the arcs from a spanning tree, in which case we say the function is *normalized*. In particular we may assume that

$$f(1,i) = I = f(i,1)$$

for all $i\neq 1$.

The reduced closed walks at a given graph form the fundamental group of the graph; a normalized unitary arc function determines a homomorphism from the fundamental group into the unitary group. Hence it gives a unitary representation of the fundamental group.

19.7 Lines from Subspaces

If x and y are unit vectors and $\langle x,y\rangle\langle y,x\rangle = \lambda$

$$xx^* yy^* xx^* = x\langle x,y\rangle\langle y,x\rangle x^* = \lambda xx^*$$

and so the spans of x and y are 1-dimensional λ-isoclinic subspaces.

19.7.1 Lemma. *Let $\mathscr{C}$ and $\mathscr{D}$ be a pair of λ-isoclinic subspaces and let x and y be unit vectors such that $x \in \mathscr{C}$ and $y \in \mathscr{D}$ and $|\langle x,y\rangle|^2 = \lambda$. If P represents orthogonal projection onto $\mathscr{C}$, then $Py = \langle x,y\rangle x$.*

Proof. Set $\gamma = \langle x,y\rangle$. We have

$$\langle Py-\gamma x, Py-\gamma x\rangle = \langle Py,Py\rangle - \gamma\langle Py,x\rangle - \bar\gamma\langle x,Py\rangle + \gamma\bar\gamma\langle x,x\rangle.$$

Here, because $\mathscr{C}$ and $\mathscr{D}$ are λ-isoclinic,

$$\langle Py,Py\rangle = \lambda\langle y,y\rangle = \lambda,$$

and

$$\gamma\langle Py,x\rangle = \gamma\langle y,Px\rangle = \gamma\langle y,x\rangle = \lambda,$$

similarly $\bar\gamma\langle x,Py\rangle = \bar\gamma\langle x,y\rangle = \lambda$. Hence $\langle Py-\gamma x, Py-\gamma x\rangle = 0$. □

The following result is an extension of a result from Lemmens and Seidel. It gives a necessary condition for a set of equi-isoclinic subsapces to contain a set of equiangular lines.

19.7.2 Theorem. *Let $\mathscr{C}_1,\dots,\mathscr{C}_m$ be a set of pairwise λ-isoclinic subspaces in $\mathbb{C}^d$, with associated projections $P_1,\dots,P_m$. Let $R_1,\dots,R_m$ be matrices with orthonormal columns such that $P_i = R_iR_i^*$. Let f denote the corresponding unitary arc function. If $z_1,\dots,z_m$ are unit vectors such that $z_i \in \mathscr{C}_i$ and*

$$\langle z_i,z_j\rangle\langle z_j,z_i\rangle = \lambda,\quad i \neq j,$$

then for any closed walk w on K_m starting at vertex 1, the vector $R_1^* z_1$ is an eigenvector for $f(w)$.

Proof. We have $f(i,j) = R_i^* R_j$ for each arc (i,j). Now

$$P_1P_{i_1}\cdots P_{i_k}P_1z_1 = z_1z_1^*z_{i_1}z_{i_1}^*\cdots z_{i_k}z_{i_k}^*z_1 = \gamma z_1$$

and

$$P_1P_{i_1}\cdots P_{i_k}P_1z_1 = R_1R_1^*(P_{i_1}\cdots P_{i_k})R_1R_1^*z_1,$$

it follows that $R_1^*z_1$ is an eigenvector for the product

$$f(1,i_1)\cdots f(i_k,1).$$ □

This result tells us that if a set of equi-isoclinic subspaces contains a set of equiangular lines, then the group generated by the arc function on closed walks has a 1-dimensional invariant subspace. Equivalently it has a non-trivial linear representation.

20
Forms

20.1 Semilinear Forms

A *semilinear form* on a vector space V is a map from $V \times V$ to the underlying field. It maps the pair (x, y) to $\langle x, y\rangle$, and saisfies the following:

(a) For each vector a, the map $x \mapsto \langle a, x\rangle$ is linear.

(b) For each vector b, the map $x \mapsto \langle x, b\rangle$ is semilinear.

It follows that for all vectors x and y and all scalars a,

$$\langle ax, y\rangle = a^\sigma \langle x, y\rangle.$$

The standard inner product on $\mathbb{C}^d$ is semilinear; in this case σ is complex conjugation. For a wider class of examples, take a square matrix A and define

$$\langle x, y\rangle = (x^\sigma)^T Ay.$$

(For a matrix or vector M, we use A^σ to denote the result of applying σ to each entry of M.)

Since the map $\psi_a : x \mapsto \langle a, x\rangle$ is a linear map from V to the 1-dimensional space $\mathbb{F}$ we see that either ψ_a is onto and its kernel has codimension 1 in V, or ψ_a is the zero map and its kernel is V. We denote the kernel of ψ_a by $x^\perp$. The radical of V (relative to our form) is the set of vectors a such that ψ_a is the zero map. It is a subspace of V. We say that the form is *non-degenerate* if its radical is zero. The radical of an inner product is zero.

If $U \le V$, we define

$$U^\perp = \cap_{u\in U} u^\perp.$$

This is again a subspace of V.

20.1.1 Lemma. *If* $U \le V$ *and our form is non-degenerate, then* $\dim(U) + \dim(U^\perp) = \dim(V)$.

Let $u_1, \dots, u_k$ be a basis for U and define a map $\rho : U \to \mathbb{F}^k$ by

$$\rho(x) = \begin{pmatrix} \langle u_1, x\rangle & \dots & \langle u_k, x\rangle \end{pmatrix}.$$

We see that ρ is linear and that $\ker(\rho) = U^\perp$. If ρ is not surjective, there are scalars $a_1, \ldots, a_k$ such that

$$0 = \sum_{r=1}^{k} a_r \langle u_r, x\rangle = \langle \sum_{r=1}^{k} a_r u_r, x\rangle.$$

Since our form is non-degenerate, it follows that $\sum_{r=1}^{k} a_r u_r = 0$ and, since $u_1, \ldots, u_k$ is a basis, $a_r = 0$ for all r. We conclude that ρ is surjective, and the lemma follows from the rank-nullity theorem. □

We say that a subspace U of V is *isotropic* if $U \le U^\perp$. The the zero subspace is the only isotropic subspace of an inner product space.

20.2 The Classification of Forms

There are three classes of semilinear forms.

For the first, the associated automorphism is not trivial, and

$$\langle y, x\rangle = (\langle x, y\rangle)^\sigma.$$

In this case we have a *Hermitian form*. For a Hermitian form there is a matrix H such that $(H^\sigma)^T = H$ and

$$\langle x, y\rangle = (x^\sigma)^T Hy.$$

Otherwise σ is trivial. The next possibility is that

$$\langle y, x\rangle = \langle x, y\rangle.$$

In this case we have a *symmetric form*, for which there is always a symmetric matrix A such that $\langle x, y\rangle = x^T Ay$. Finally we may have an *alternating form*, where

$$\langle x, x\rangle = 0$$

for all x. Here

$$0 = \langle x + y, x + y\rangle = \langle x, x\rangle + \langle x, y\rangle + \langle y, x\rangle + \langle y, y\rangle$$

and since $\langle x, x\rangle = \langle y, y\rangle$, it follows that

$$\langle y, x, =\rangle - \langle x, y\rangle$$

For an alternating form there is a matrix S such that $S^T = -S$ and all diagonal entries are zero; then $\langle x, y\rangle = x^T Sy$.

Alternating forms are also known as *symplectic forms*. In odd characteristic it is reasonable to describe the matrix S as skew symmetric. In even characteristic, S is symmetric with zero diagonal.

Under natural geometric assumptions it can be shown (with some effort) that the above three families of semilinear forms are the only interesting possibiities.

We say two forms $\langle,\rangle_1$ and $\langle,\rangle_2$ are *equivalent* if if there is an invertible matrix M such that

$$\langle x, y\rangle_2 = \langle Mx, My\rangle_1.$$

This raises the problem of determining the equivalence classes of forms of a given type on vector space.

Over finite fields it can be shown that there is only one class of non-degenerate Hermitian forms, and only one class of non-degenerate alternating forms. It cannot be shown that there is only one class of non-degenerate symmetric forms—because this is false.

21 Groups

For each bilinear form $\langle , \rangle$, we have the group of matrices M that preserve the form, that is, the matrices M such that

$$\langle x, y \rangle = \langle Mx, My \rangle.$$

We spend some time with these groups.

Part V

Algebras

22
Lie Algebras

We study Lie algebras because they force themselves on us when we study the Terwilliger algebra of the binary Hamming scheme. As we will see, there are other combinatorial applications. Additionally we will work with the universal enveloping algebra of a Lie algebra, which provides a useful example of an infinite dimensional algebra.

22.1 Basics

A *Lie algebra* over a field $\mathbb{F}$ is a vector space with a multiplication $[a,b]$ such that

(a) $[b,a] = -[a,b]$.

(b) For all a, b and c, we have the *Jacobi identity*:

$$[a,[b,c]] + [b,[c,a]] + [c,[a,b]] = 0.$$

The only fields we will use in this context are $\mathbb{R}$ and $\mathbb{C}$, whence we see that $[a,a] = 0$ for all a. We call $[a,b]$ the *Lie bracket* or *commutator* of a and b, and we abbreviate $[a,[b,c]]$ to $[a,b,c]$. A Lie algebra is *abelian* if $[a,b] = 0$ for all a and b.

Note that a Lie algebra is **not** an algebra in the sense we have used elsewhere—the multiplication is not even associative in general.

We offer examples:

(a) $gl(n,\mathbb{F})$, the Lie algebra of all $n \times n$ matrices over $\mathbb{F}$, where

$$[A,B] := AB - BA.$$

(b) The real skew symmetric matrices of order $n \times n$ form a Lie algebra over $\mathbb{R}$.

(c) $\mathbb{R}^3$ with the cross product. We will use $a \wedge b$ to denote the cross product.

(d) A *derivation* of a commutative algebra $\mathscr{A}$ over $\mathbb{F}$ is a map $\delta : \mathscr{A} \to \mathbb{F}$ such that

$$\delta(fg) = \delta(f)g + f\delta(g).$$

You may check that the product of two derivations is not in general a derivation, but their Lie bracket is, and further the set of derivations of $\mathscr{A}$ is a Lie algebra. By way of a more specific example take $\mathscr{A}$ to be the polynomial ring $\mathbb{F}[x_1,\dots,x_d]$ and note that, for each i, partial differentiation with respect to x_i is a derivation.

The construction in (a) can be usefully generalized: if $\mathscr{A}$ is an algebra over $\mathbb{F}$, then the multiplication

$$[a,b] := ab - ba$$

gives us a Lie algebra. Thus if V is a vector space, then $\mathrm{End}(V)$ is a Lie algebra under this operation. For fixed a in $\mathscr{A}$, the map from $\mathscr{A}$ to itself given by

$$x :\mapsto [a,x]$$

is a derivation (as you should check).

A subspace of a Lie algebra $\mathscr{L}$ is subalgebra if it is closed under the Lie bracket. You could check that the subspace of skew symmetric matrices is a subalgbra of $gl(n,\mathbb{F})$. A subspace U of $\mathscr{L}$ is an *ideal* if $[a,u] \in U$, for all u in U. The subspace of strictly upper triangular matrices is an ideal in the Lie algebra formed by the set of all upper triangular matrices.

If $\mathscr{L}$ is a Lie algebra and S,T are subsets of $\mathscr{L}$, then we define $[S,T]$ to be the subspace of $\mathscr{L}$ spanned by the set

$$\{[x,y] : x \in S, y \in T\}.$$

In particular the subspace $[\mathscr{L},\mathscr{L}]$ is a subalgebra of $\mathscr{L}$, called the *commutator subalgebra*.

For example, suppose $\mathscr{L} = gl(V)$. Then for any A and B in $\mathscr{L}$, we have

$$\mathrm{tr}[A,B] = \mathrm{tr}(AB) - \mathrm{tr}(BA) = 0.$$

So the commutator of $gl(V)$ consists of matrices with zero trace. It can be shown that it contains all matrices with zero trace. It is known as the special linear Lie algebra and is denoted by $sl(V)$. You may show that $sl(V)$ is equal to its commutator subalgebra.

22.2 Enveloping Algebras

The construction of the Lie algebra $gl(V)$ from the algebra $\mathrm{End}(V)$ can be generalized: if $\mathscr{A}$ is an algebra and $a,b \in \mathscr{L}$, we can define their Lie bracket by

$$[a,b] := ab - ba.$$

This leads us to ask which Lie algebras arise in this way, and the answer is that they all do. Let us denote the Lie algebra we get from $\mathscr{A}$ by Lie$\mathscr{A}$. The *universal enveloping algebra* of $\mathscr{L}$ is essentially the smallest algebra $\mathscr{U}$ such that $\mathscr{L} = \text{Lie}\mathscr{U}$. Of course the adjective 'universal' indicates that a category theorist has escaped. What we should say is that $\mathscr{U}$ is defined by the condition that if $\psi : \mathscr{L} \to \text{Lie}\mathscr{A}$ for some algebra $\mathscr{A}$, then ψ can be factored into a Lie homomorphism from $\mathscr{L}$ to Lie$\mathscr{U}$ and a Lie homomorphism from Lie$\mathscr{U}$ to Lie$\mathscr{A}$ induced by an algebra homomorphism from $\mathscr{U}$ to $\mathscr{A}$.

We consider a particular example, using the the Lie algebra $sl(2,\mathbb{R})$. The elements of this are the 2 × 2 matrices of trace zero, which form a vector space of dimension three, with basis

$$X = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad H = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

We note that

$$[X,Y] = H$$

and that

$$[H,X] = 2X, \quad [H,Y] = -2Y.$$

The universal enveloping algebra of $sl(2,\mathbb{F})$ is the quotient of the free polynomial algebra in variables X, Y modulo the relations

$$XY - YX - H = 0, \quad HX - XH - 2H = 0, \quad HY - YH + 2Y = 0.$$

Note that this is an infinite-dimensional algebra—it can be shown that the elements $X^k Y^\ell H^m$ form a basis.

22.3 Posets

A poset is *ranked* if all elements covered by an element have the same height. If P is ranked then the i-th *level number* is the number of elements with height i. Thus the poset formed by the subsets of $\{1,\dots,n\}$, ordered by inclusion, is ranked and the i-th level number of $\binom{n}{i}$. If P is ranked with height d and the i-th level number is w_i, we say that P is *rank symmetric* if $w_i = w_{d-i}$ for all i, and we say P is *unimodal* if the sequence of level numbers is unimodal. The lattice of subsets of $\{1,\dots,n\}$ is rank symmetric and unimodal.

An *antichain* in a poset P is set of elements such that no two are comparable. (Equivalently it is a coclique in the comparability graph of P.) The elements of given height in a poset form an antichain, and we say P is *Sperner* if the maximum size of an antichain is equal to the maximum level number. More generally we call P *strongly Sperner* if the maximum size of a subset that does not contain a chain of length $k+1$ is equal to the sum of the k largest level numbers. A *Peck poset* is a ranked poset that is rank symmetric, unimodal and strongly Sperner. The lattice of subsets of a finite set is Peck.

We use $\mathbb{P}$ to denote the the vector space $\mathbb{R}^P$. We can represent subsets of P by their characteristic vectors, which belong to $\mathbb{P}$. If $a \in P$ we will often denote the characteristic vector of a by a. The subspace of $\mathbb{P}$ spanned by the (characteristic vectors of) the elements of height i will be denoted by $\mathbb{P}(i)$.

Suppose P is a finite ranked poset. An element of $\operatorname{End}(\mathbb{P})$ is a *raising operator* if for each element a of P, the support of Ra is a subset of the elements of P that cover a. Similarly we define *lowering operators*. If R is a raising operator then R^T is lowering. Both raising and lowering operators are nilpotent: if P has height d and R is a raising operator, then $R^{d+1} = 0$.

The following result is due to Stanley and Griggs.

22.3.1 Theorem. *Let P be a rank-symmetric poset with height h. Then P is Peck if and only if there is an order-raising operator R such that the mappings*

$$R^{h-i}\restriction\mathbb{P}(i) : \mathbb{P}(i) \to \mathbb{P}(h-i), \qquad i = 0, \ldots, \left\lfloor \frac{h}{2} \right\rfloor \qquad \square$$

are invertible.

Using the above result, Proctor showed the following.

22.3.2 Theorem. *A ranked poset is Peck if and only if it has raising and lowering operators R and L such that the Lie algebra generated by R and L is isomorphic to $sl(2,\mathbb{C})$.* □

We derive an important consequence of these results.

22.3.3 Corollary. *If P_1 and P_2 are Peck posets, then so is $P_1 \times P_2$.*

Proof. If $_1$ and P_2 are Peck then the vector spaces $\mathbb{P}_1$ and $\mathbb{P}_2$ are modules for $sl(2,\mathbb{C})$. Now

$$\mathbb{C}^{P_1 \times P_2} = \mathbb{P}_1 \otimes \mathbb{P}_2$$

and therefore $\mathbb{C}^{P_1 \times P_2}$ is a module for $sl(2,\mathbb{C})$. We conclude that $P_1 \times P_2$ is Peck. □

If U and V are modules for an algebra $\mathscr{A}$ then $U \otimes V$ is a module for $\mathscr{A} \times \mathscr{A}$, but it is **not** in general a module for $\mathscr{A}$. However it is module for $\mathscr{A}$ when $\mathscr{A}$ is an enveloping algebra of a Lie algebra (and when $\mathscr{A}$ is a group algebra).

22.4 Representations of Lie Algebras

A linear map ψ from a Lie algebra $\mathscr{L}_1$ to a Lie algebra $\mathscr{L}_2$ is a *homomorphism* if

$$\psi([a,b]) = [\psi(a), \psi(b)].$$

A *representation* of a Lie algebra $\mathscr{L}$ is a homomorphism into $gl(n,\mathbb{F})$. More generally ψ could be a homomorphism into $\mathrm{End}(V)$ for some vector space V; in this case we may say that V is a *module* over $\mathscr{L}$. A subspace of V that is invariant under the operators in $\psi(\mathscr{L})$ is a *submodule*. (Calling V a module for $\mathscr{L}$ is a courtesy, since modules are defined over rings—if we wish to be precise, it is a module for the enveloping algebra.)

If $\mathscr{L}$ is a Lie algebra and $A \in \mathscr{L}$, we define the *adjoint map* ad_A by

$$\mathrm{ad}_A(X) := [A, X].$$

This is a linear map, and is a derivation of the enveloping algebra. By Jacobi's identity

$$\begin{aligned}\mathrm{ad}_A([X,Y]) = [A,[X,Y]] &= -[X,[Y,A]] - [Y,[A,X]]\\ &= [X,[A,Y]] + [[A,X],Y].\end{aligned}$$

We also have, by appeal to Jacobi

$$\begin{aligned}(\mathrm{ad}_X\,\mathrm{ad}_Y - \mathrm{ad}_Y\,\mathrm{ad}_X)(Z) &= [X,[Y,Z]] - [Y,[X,Z]]\\ &= [X,[Y,Z]] + [Y,[Z,X]]\\ &= [[X,Y],Z]\\ &= \mathrm{ad}_{[X,Y]}(Z),\end{aligned}$$

which shows that ad_A is a homomorphism from $\mathscr{L}$ into the Lie algebra $\mathrm{End}(L)$.

An element A of $\mathscr{L}$ is *ad-nilpotent* if ad_A is nilpotent. We observe that

$$\begin{aligned}\mathrm{ad}_A(X) &= [A,X],\\ (\mathrm{ad}_A)^2(X) &= [A,[A,X]],\\ (\mathrm{ad}_A)^3(X) &= [A,[A,[A,X]]]\end{aligned}$$

and in general, $(\mathrm{ad}_A)^{k+1}(X) = [A, (\mathrm{ad}_A)^k(X)]$. If $A \in gl(n,\mathbb{F})$, then we may represent the linear map ad_A by

$$A \otimes I - I \otimes A.$$

It follows that if $A^k = 0$, then $(\mathrm{ad}_A)^{2k} = 0$. In particular if A in $gl(V)$ is nilpotent then ad_A is nilpotent. Thus we have the fortunate conclusion that nilpotent elements of $gl(V)$ are ad-nilpotent.

22.5 Bilinear Forms

Suppose ψ is a representation of the Lie algebra $\mathscr{L}$ in $\mathrm{End}(V)$. A bilinear form β on V is *invariant* if

$$\beta(\psi(X)u, v) + \beta(u, \psi(X)v) = 0$$

for all u and v from V. By way of example, if V is $\mathscr{L}$ itself then

$$\beta(X,Y) := \operatorname{tr}(\mathrm{ad}_X\, \mathrm{ad}_Y)$$

is a symmetric bilinear form, known as the *Killing form*. We check that it is invariant.

$$\begin{aligned}\beta([A,X],Y) &= \operatorname{tr}(\mathrm{ad}_{[A,X]}\, \mathrm{ad}_Y)\\ &= \operatorname{tr}([\mathrm{ad}_X, \mathrm{ad}_Y]\, \mathrm{ad}_Y)\\ &= \operatorname{tr}(\mathrm{ad}_A\, \mathrm{ad}_X\, \mathrm{ad}_Y - \mathrm{ad}_X\, \mathrm{ad}_A\, \mathrm{ad}_Y)\end{aligned}$$

Similarly

$$\beta(X,[A,Y]) = \operatorname{tr}(\mathrm{ad}_X\, \mathrm{ad}_A\, \mathrm{ad}_Y - \mathrm{ad}_X\, \mathrm{ad}_Y\, \mathrm{ad}_A)$$

from which we see that β is invariant. (Thus the adjoint of ad_X relative to the Killing form is $-\mathrm{ad}_X$.)

Suppose $\mathscr{L}$ is a Lie algebra with a non-degenerate invariant bilinear form. If $X_1,\ldots,X_d$ is a basis for $\mathscr{L}$, there is a dual basis $Y_1,\ldots,Y_d$ such that

$$\beta(X_i,Y_j) = \delta_{i,j}.$$

The *Casimir element* of the universal enveloping algebra is defined to be

$$\sum_{i=1}^{d} X_i Y_i.$$

22.5.1 Theorem. *Let $\mathscr{L}$ be a Lie algebra with a non-degenerate invariant bilinear form β. Then the Casimir element is independent of the choice of basis for $\mathscr{L}$, and lies in the center of the universal enveloping algebra.*

Proof. Let $X_1,\ldots,X_d$ be a basis for $\mathscr{L}$ with dual basis $Y_1,\ldots,Y_d$ and let Δ be the Casimir element defined using this pair of bases. Let $U_1,\ldots,U_d$ and $V_1,\ldots,V_d$ be a second pair of dual bases. Then there are scalars $\rho_{i,j}$ and $\sigma_{i,j}$ such that

$$\begin{aligned}U_i &= \sum_k \rho_{i,k} X_k,\\ V_j &= \sum_\ell \sigma_{j,\ell} Y_\ell.\end{aligned}$$

We have

$$\sum_i U_i V_i = \sum_{i,k,\ell} \rho_{i,k}\sigma_{i,\ell} X_i Y_i \tag{22.5.1}$$

Since $\beta(X_i,Y_j) = \delta_{i,j}$, we have

$$\delta_{i,j} = \beta(U_i,V_j) = \sum_k \rho_{i,k}\sigma_{j,k}$$

So if we define matrices R and S by $R := (\rho_{i,j})$ and $S := (\sigma_{i,j})$ then $RS^T = 0$. Consequently $SR^T = 0$ and therefore

$$\delta_{k,\ell} = \sum_i \rho_{i,k}\sigma_{i,\ell}.$$

Hence (22.5.1) implies that $\sum_i U_i V_i = \Delta$.

We now prove Δ lies is central. Suppose $A \in \mathscr{L}$. There are scalars $\alpha_{i,j}$ and $\beta_{i,j}$ such that

$$[A, X_i] = \sum_j \alpha_{i,j} X_j$$

and

$$[A, Y_i] = \sum_j \beta_{i,j} Y_j$$

Since β is invariant,

$$0 = \beta([A, X_i], Y_j) + \beta(X_i, [A, Y_j]) = \alpha_{i,j} + \beta_{j,i}.$$

This implies that

$$\sum_i [A, X_i] Y_i = \sum_{i,j} \alpha_{i,j} X_j Y_i = -\sum_{i,j} \beta_{j,i} X_j Y_i = -\sum_i X_i [A, Y_i].$$

Now we compute that

$$A\Delta = \sum_i A X_i Y_i = \sum_i [A, X_i] Y_i + \sum_i X_i A Y_i$$

and

$$\Delta A = \sum_i X_i Y_i A = -\sum_i X_i [A, Y_i] + \sum_i X_i A Y_i,$$

whence we conclude that $A\Delta = \Delta A$. □

22.5.2 Lemma. *If Δ is the Casimir element of the Lie algebra $\mathscr{L}$ and φ is a representation of $\mathscr{L}$, then* $\operatorname{tr}(\varphi(\Delta)) = \dim(\varphi(\mathscr{L}))$. □

22.6 An Example

We compute the Casimir element for $sl(2,\mathbb{C})$, relative to the form

$$\beta(X, Y) := \operatorname{tr}(\operatorname{ad}_X \operatorname{ad}_Y).$$

Recall that X, H and Y form a basis, where

$$X = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad H = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

and

$$[X, Y] = H, \quad [H, X] = 2X, \quad [H, Y] = -2Y.$$

It follows that

$$\operatorname{ad}_X = \begin{pmatrix} 0 & -2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \operatorname{ad}_Y = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix}, \quad \operatorname{ad}_H = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

If

$$A = \begin{pmatrix} a & b \\ c & -a \end{pmatrix},$$

then

$$\mathrm{ad}_A = \begin{pmatrix} 2a & -2b & 0 \\ -c & 0 & b \\ 0 & 2c & -2a \end{pmatrix}$$

and now it is easy verify that if

$$\beta(A, X) = \beta(A, H) = \beta(A, Y) = 0,$$

then $A = 0$. Therefore β is nondegenerate.

Next we calculate that

$$\beta(X, Y) = \beta(Y, X) = 4, \quad \beta(H, H) = 8$$

and all other inner products are zero. So the dual basis to (X, H, Y) is

$$\Big(\frac{1}{4}Y, \frac{1}{4}X, \frac{1}{8}H\Big)$$

and the Casimir element is

$$\Delta := \frac{1}{4}(XY + YX + \frac{1}{2}H^2).$$

Using the fact that

$$[A, BC] = [A, B]C + B[A, C],$$

it is not hard to verify directly that Δ is central.

22.7 Irreducible Modules

We construct a family of irreducible modules for $sl(2,\mathbb{C})$, by constructing irreducible modules for its enveloping algebra.

22.7.1 Lemma. *Let $\mathcal{U}$ denote the enveloping algebra of $sl(2,\mathbb{C})$, with generators X, Y and H, and suppose V is a module for $\mathcal{U}$ with finite dimension. If v is an eigenvector for H in its action on V, then there are integers k and ℓ such that $X^k v = 0$ and $Y^\ell v = 0$.*

Proof. Suppose $Hv = \lambda v$. Recalling that $[H, X] = 2X$, we have

$$HXv = (XH + 2X)v = \lambda Xv + 2Xv = (\lambda + 2)Xv.$$

Hence if $Xv \neq 0$ and λ is an eigenvalue of H, then $\lambda + 2$ is also an eigenvalue of H. A similar calculation shows that if $Yv \neq 0$, then Yv is an eigenvector for H with eigenvalue $\lambda - 2$. □

Note that XYv is an eigenvector for H with eigenvalue λ, consistent with the fact that H and XY commute.

If V is a module for $\mathcal{U}$, an element v of V has *weight* λ if $Hv = \lambda v$. If $Hv = \lambda v$ and also $Xv = 0$, we say that v is a *highest weight vector* of weight λ. The eigenspaces of H are often called *weight spaces*. We have seen that every finite-dimensional module for $\mathcal{U}$ must contain a highest weight vector; the following theorem completely specifies the structure of the cyclic $\mathcal{U}$-module generated by a highest weight vector.

22.7.2 Theorem. *Suppose V is a module for $\mathcal{U}$ and v is a highest weight vector in V with eigenvalue λ. Let d be the least non-negative integer such that $Y^d v = 0$. Then $\lambda = d-1$, the cyclic $\mathcal{U}$-module generated by v is simple and the vectors*

$$v, Yv, \ldots, Y^{d-1}v$$

form a basis for it. Further, for $k = 0, 1, \ldots, d-1$,

$$HY^k v = (d-1-2k)Y^k v, \qquad XY^k v = k(d-k)Y^{k-1}.$$

Proof. The adjoint map ad_H is a derivation of $\mathcal{U}$ whence

$$[H, Y^k] = [H, Y]Y^{k-1} + Y[H, Y^{k-1}]$$

and a trivial induction yields that

$$[H, Y^k] = -2kY^k.$$

If $Hv = \lambda v$, we have

$$HY^k v = [H, Y^k]v + Y^k Hv = -2kY^k v + \lambda Y^k v = (\lambda - 2k)Y^k v.$$

Let d be the least integer such that $Y^d v = 0$. Then the vector space V_1 spanned by the vectors

$$v, Yv, \ldots, Y^{d-1}v$$

has dimension d, and these vectors for a basis for it. Since these vectors are all eigenvectors for H, we see that V_1 is invariant under both Y and H. We prove that it is X-invariant.

We have

$$[X, Y^k] = [X, Y]Y^{k-1} + Y[X, Y^{k-1}]$$

Since $Xv = 0$, it follows that

$$XY^k v = [X, Y^k]v = HY^{k-1}v + Y[X, Y^{k-1}]v$$

and so by induction we have

$$XY^k v = HY^{k-1}v + YHY^{k-2}v + \cdots + Y^{k-1}Hv.$$

Since the vectors $Y^k v$ are eigenvectors for H, this implies that $XY^k v = c_k Y^{k-1}v$, for some constant c_k and therefore V_1 is a module for $\mathcal{U}$. We have

$$c_k = (\lambda + 2 - 2k) + (\lambda + 4 - 2k) + \cdots + \lambda = k\lambda - (k^2 - k) = k(\lambda - k + 1).$$

We see that c_d is the sum of the eigenvalues of H on V_1 and so $c_d = \operatorname{tr}(H)$. As $H = XY - YX$ we have $\operatorname{tr}(H) = 0$, and therefore $\lambda = d - 1$.

It remains to prove that V is simple. Suppose V_1 is a non-zero submodule of V. Then V_1 contains a highest weight vector u, and since u is an eigenvector for H it must be a non-zero scalar multiple of one of the vectors $Y^i v$. Since $Xu = 0$, we see see that u is a non-zero scalar multiple of v. Hence the cyclic module generate by u is equal to V and therefore $V_1 = V$. □

This result implies that the $\mathscr{U}$ module generated by a highest weight vector v is determined by its dimension (or by the eigenvalue of v). Also note that any simple module is isomorphic to one of the modules described in this theorem, since any module contains a highest weight vector.

22.7.3 Corollary. *If C is the Casimir element of $sl(2,\mathbb{C})$, then $CY^k v = (d^2 - 1)Y^k v$.*

Proof. From above we have

$$\begin{aligned} XYY^k v &= (k+1)(d-k-1)Y^k v \\ YXY^k v &= k(d-k)Y^k v \\ HY^k v &= (d-1-2k)Y^k v \end{aligned}$$

and the claim follows easily from these. □

22.8 Semisimple Elements

We derive two useful identities that hold in the enveloping algebra of $sl(2,\mathbb{C})$. We define

$$H_k := H + kI$$

and we define $H_{k;r}$ recursively by $H_{k;0} = I$ and

$$H_{k;i+1} := H_{k;i} H_{k-i+1}.$$

22.8.1 Lemma. *We have*

$$X^m Y^n = \sum_{r=0}^{m\wedge n} r! \binom{m}{r} \binom{n}{r} Y^{n-r} X^{m-r} H_{n-m;r}$$

Proof. First prove by induction that if $n \geq 1$, then

$$X^n Y = YX^n + nX^{n-1} H_{n-1} \tag{22.8.1}$$

and then, by a second induction, derive the lemma. □

22.8.2 Lemma. *In a finite dimensional representation of $\mathscr{U}(sl(2,\mathbb{C}))$, if $X^k = 0$ then*

$$\prod_{r=-k+1}^{k-1} (H - rI) = 0.$$

Proof. We do not give a complete proof, but offer a generous hint and leave the details as an exercise.

Suppose V is a finite-dimensional representation for $\mathscr{U}$. The idea is to prove that, if $X^k = 0$, then for $i = 1, \dots, k$ we have

$$X^{k-i} H_{k-1;2i-1} = 0.$$

Setting $i = k$ in this yields the result.

For convenience we prove the above claim in the case $k = 4$. We have the following equations:

$$\begin{aligned} X^4 Y &= Y X^4 + 4X^3 H_{3;1} && (22.8.2)\\ X^4 Y^2 &= Y^2 X^4 + 8Y X^3 H_{2;1} + 12X^2 H_{2;2} && (22.8.3)\\ X^4 Y^3 &= Y^3 X^4 + 12Y^2 X^3 H_{1;1} + 36Y X^2 H_{1;2} + 24X H_{1;3} && (22.8.4)\\ X^4 Y^4 &= Y^4 X^4 + 16Y^3 X^3 H_{0;1} + 72Y^2 X^2 H_{0;2} + 216Y X H_{0;3} + 24H_{0;4} && (22.8.5) \end{aligned}$$

Since $X^4 = 0$ we see that (22.8.2) implies

$$X^3 H_{3;1} = 0.$$

Now multiply (22.8.3) on the right by H_3; since $XH_i = H_{i-2}X$, we get

$$0 = 8Y X^3 H_3 H_2 + 12X^2 H_1 H_2 H_3$$

and since $Y X^3 H_3 = 0$, we deduce that

$$X^2 H_{3;3} = 0.$$

Next multiply (22.8.4) on the right by $H_2 H_3$ and deduce that since

$$Y X^2 H_{1;2} H_2 H_3 = Y X^2 H_0 H_1 H_2 H_3 = Y X^2 H_{3;3} H_0 = 0,$$

that

$$X H_{3;5} = 0.$$

Finally multiply (22.8.5) on the right by $H_{1;3}$ to deduce that

$$H_{3;7} = 0.$$ □

Recall that H, XY and YX all commute.

22.8.3 Lemma. *If $1 \le k \le n$, then*

$$X^n Y^k = \Big(\prod_{i=0}^{k-1} (Y X + (n-i) H_{-n+i+1}) \Big) X^{n-k}$$

Proof. From (22.8.1) we have

$$X^n Y = Y X^n + n X^{n-1} H_{n-1} = Y X^n + n H_{-n+1} X^{n-1} = (Y X + n H_{-n+1}) X^{n-1}$$

and use induction on k. □

22.8.4 Theorem. *In a finite-dimensional representation of $\mathscr{U}(sl(2,\mathbb{C}))$, the images of H, XY and YX are semisimple.*

Proof. Since H and XY commute and $YX = XY - H$, it is enough to show that H and YX are semisimple. By 22.7.1, there is an integer k such that $X^k = 0$. From 22.8.2 it follows that H is semisimple, and so the underlying vector space V is a direct sum of eigenspaces of H. Suppose V_λ is one of these eigenspaces, where λ is the eigenvalue of H.

By 22.8.3 we have

$$0 = X^k Y^k = (YX + k(H-(k-1)I)\cdots(YX+H)$$

and if $z \in V_\lambda$, then

$$0 = (YX + k(\lambda-(k-1)I)\cdots(YX+\lambda)z.$$

Hence the minimal polynomial of YX on V_λ has only simple zeros, and therefore YX is semisimple on V_λ. We conclude that YX must be semisimple. □

22.9 Semisimple Modules

22.9.1 Theorem. *Any finite dimensional module for $\mathscr{U}(sl(2,\mathbb{C}))$ is semisimple.*

Proof. Let $\mathscr{U}$ denote $\mathscr{U}(sl(2,\mathbb{C}))$, let M be a finite-dimensional $\mathscr{U}$-module, and let C be the Casimir element of $\mathscr{U}$. Since C is central and semisimple, M is the direct sum of eigenspaces of C, and so to prove the theorem it will suffice if we show that any eigenspace for C semisimple.

Hence we assume that M itself is an eigenspace for C. Since H also is semisimple, M is the direct sum of weight spaces M_σ and, if $N \le M$, then N is the direct sum of its weight space N_σ, where

$$N_\sigma = N \cap M_\sigma.$$

We have

$$\dim(M_\sigma) = \dim(N_\sigma) + \dim(M_\sigma/N_\sigma).$$

Note that M/N is a $\mathscr{U}$-module and

$$(M/N)_\sigma = M_\sigma/N_\sigma.$$

Next assume we have the composition series for M:

$$0 = M_0 < M_1 < \cdots < M_r = M.$$

Then

$$\dim(M_\sigma) = \sum_{i=1}^{r} \dim(M_i/M_{i-1})_\sigma$$

but M_i/M_{i-1} is a simple $\mathscr{U}$-module and consequently $\dim(M_i/M_{i-1})_\sigma = 1$. We conclude that $\dim(M_\sigma) = r$ and that $\dim(M)$ is r times the number of eigenvalues of H. The cyclic $\mathscr{U}$-submodule of M generated by a non-zero element is simple and, since all non-zero elements of M are eigenvectors for C with the same eigenvalue, all these simple modules have the same dimension.

Choose a basis $x_1, \dots, x_d$ for M. Then

$$M = x_1\mathscr{U} + \cdots + x_d\mathscr{U}.$$

where each submodule $x_i\mathscr{U}$ contains a simple submodule S_i (say). (We do not assume that this is a direct sum.) Since $\dim(M_\sigma) = r$, we have $d = r$. Since $x_1, \dots, x_r$ is a basis, the sum

$$S_1 + \cdots + S_r$$

is direct and therefore $\dim(M)$ is bounded below by r times the number of eigenvalues of H. But we saw that equality holds, and therefore M is a direct sum of simple modules as required. □

This proof follows Jantzen [1].

[1]

23
Terwilliger Algebras

Let $\mathscr{A}$ be an association scheme with d classes and let π be an equitable partition of its vertex set with e classes. Define the diagonal 01-matrix F_i by setting $(F_i)_{u,u} = 1$ if u lies in the i-th class of π. Then the matrices F_i are symmetric idempotents and

$$\sum_i F_i = I.$$

We will study the algebra generated by $\mathscr{A}$ together with the matrices F_i.

If u is a vertex in the scheme and the i-th cell of π consists of the vertices x such that (u, x) lies in the i-th relation, the algebra we get is the *Terwilliger algebra* of the scheme relative to the vertex u.

23.1 Modules

Our basic task is to determine the irreducible modules of the Terwilliger algebra. Suppose $\mathscr{A}$ is an association scheme with d classes $A_0, \ldots, A_d$ and vertex set V, and assume $|V| = v$. Let $\mathbb{T}$ denote the Terwilliger algebra of this scheme and suppose W is an irreducible $\mathbb{T}$-module. Since W is invariant under $\mathscr{A}$, it must have basis that consists of eigenvectors for $\mathscr{A}$. Similarly it must have basis that consists of eigenvectors for the matrices F_i, that is, vectors whose supports are subsets of the cells of the partition π.

The subspace spanned by the characteristic vectors of the cells of π is $\mathbb{T}$-invariant and has dimension equal to $|\pi|$, the number of cells of π. We call it the *standard module* It is a cyclic $\mathbb{T}$-module, generated by $\mathbf{1}$. You may prove that it is irreducible.

This may seem an encouraging start to determining the irreducible modules for the Terwilliger algebra, but unfortunately further progress will require much more effort. Since $\mathbb{T}$ is transpose-closed, $\mathbb{R}^v$ decomposes into an orthogonal sum of irreducible $\mathbb{T}$-modules. Hence if W is irreducible and is not the standard module, we may assume that it is orthogonal to it. Thus each element of W will be orthogonal to the vectors $F_i\mathbf{1}$—it sums to zero on the cells of π.

23.1.1 Lemma. *If W is an irreducible module for an algebra $\mathcal{B}$ and f is an idempotent in $\mathcal{B}$, then Wf is an irreducible module for $f\mathcal{B}f$.*

Proof. We may assume $\dim(W) \geq 2$, or there is nothing to prove. Since

$$Wf\,f\mathcal{B}f = Wf\mathcal{B}f \leq Wf,$$

we see that Wf is a module for $f\mathcal{B}f$.

Suppose U is an $f\mathcal{B}f$-submodule of Wf. Each element of U can be written as wf where $w \in W$ and as $f^2 = f$, it follows that $Uf = U$. Since $Uf\mathcal{B}$ is a $\mathcal{B}$-submodule of W, it is either zero or equal to W. If it is equal to W, then

$$U = Uf\mathcal{B}f = Wf$$

and therefore Wf is irreducible for $f\mathcal{B}f$.

To complete the proof, we show that $Uf\mathcal{B}$ cannot be zero. The key is to note that the set

$$\{u \in W : u\mathcal{B} = 0\}$$

is a $\mathcal{B}$-submodule of W. Since W is simple and not zero, it follows that this set must be the zero module. Consequently $Uf\mathcal{B}$ cannot be zero. □

Note that $f\mathcal{B}f$ is a subspace of $\mathcal{B}$ and is closed under multiplication, but fails to be a subalgebra because it does not contain I (in general). However

$$f\mathcal{B}f + (I-f)\mathcal{B}(I-f)$$

is a subalgebra of $\mathcal{B}$.

When we want to use 23.1.1, we will have two possible sources of idempotents: the matrices F_i and the principal matrix idempotents E_j.

23.2 Thinness

Let $\mathbb{T}$ be the Terwilliger algebra for an association scheme $\mathcal{A}$ and let W be a $\mathbb{T}$-submodule of $\mathbb{R}^v$. We say that W is *thin* if for each i we have

$$\dim(F_iW) \leq 1.$$

We also say that W is *dual thin* if for each j,

$$\dim(E_jW) \leq 1.$$

We generalise the concept of thinness. Suppose $\mathcal{B}$ is an algebra. We say that a set of idempotents $F_1, \ldots, F_r$ is a *resolution of the identity* if they are pairwise orthogonal ($F_iF_j = 0$ when $i \neq j$) and

$$\sum_i F_i = I.$$

A module W for $\mathcal{B}$ is *thin relative to the resolution* $F_1, \ldots, F_r$ if $\dim(F_iW) \leq 1$ for all i.

Being thin is not easy, but it is a desirable property that holds in many interesting cases.

23.2.1 Lemma. *If $\mathscr{A}$ is an association scheme then the standard modules are thin and dual thin,*

Proof. Exercise. □

23.2.2 Theorem. *If the algebra $\mathscr{B}$ is self-adjoint, then it is thin relative to the resolution $F_1, \dots, F_r$ if and only if the subalgebra*

$$F_1\mathscr{B}F_1 + \cdots + F_r\mathscr{B}F_r$$

is commutative.

23.2.3 Lemma. *Suppose $\mathbb{T}$ is the Terwilliger algebra of an association scheme relative to some vertex. If each matrix in*

$$F_0\mathbb{T}F_0 + \cdots + F_e\mathbb{T}F_e$$

is symmetric, or if $\operatorname{Aut}(X)_1$ is generously transitive on each cell of π, then $\mathbb{T}$ is thin.

Proof. For the first, two symmetric matrices commute if and only if their product is symmetric. The second condition implies that each $F_i\mathbb{T}F_i$ is the Bose-Mesner algebra of a symmetric association scheme. □

23.3 Jaeger Algebras

We define some endomorphisms of $\operatorname{Mat}_{v\times v}(\mathbb{C})$. If A is a $v \times v$ matrix define the operators X_A and Y_A on $\operatorname{Mat}_{v\times v}(\mathbb{C})$ by

$$X_A(M) := AM, \qquad Y_A(M) = MA^*$$

and if B is a $v \times v$ matrix, then we define Δ_B by

$$\Delta_B(M) := B \circ M.$$

Note that

$$Y_A(Y_B(M)) = MB^*A^* = M(AB)^* = Y_{AB}(M),$$

which explains the A^* in the definition of Y_A. Also X_A and Y_B commute, for any A and B.

If $\mathscr{A}$ is an association scheme, we define $\mathscr{J}_2$ to be the algebra generated by the matrices X_A for A in $\mathbb{C}[\mathscr{A}]$. We define $\mathscr{J}_3(\mathscr{A})$ to be the algebra generated by the operators

$$X_A,\ \Delta_B, \qquad A, B \in \mathbb{C}[\mathscr{A}].$$

We obtain $\mathscr{J}_4(\mathscr{A})$ by adjoining the right multiplication operators Y_A as well

The vector space $\operatorname{Mat}_{v\times v}(\mathbb{C})$ is a module M for $\mathscr{J}_3$, and the subspace of matrices with all but the i-th column zero is a submodule, which we denote by $M(i)$. We see that M is the direct sum of the modules $M(i)$.

Our first result shows that $\mathscr{J}_3(\mathscr{A})$ is a kind of global Terwilliger algebra.

23.3.1 Lemma. *The algebra generated by the restriction to $M(i)$ of the operators in $\mathscr{J}_3$ is isomorphic to the Terwilliger algebra of $\mathscr{A}$ relative to the i-th vertex.*

Proof. We have

$$X_A(e_i e_j^T) = (Ae_i)e_j^T$$

and

$$\Delta_B(e_i e_j^T) = (B_{i,j} e_i)e_j^T.$$

So X_A is represented on $M(j)$ by the matrix A, and Δ_B by the diagonal matrix formed from the vector Be_j. □

We say that a $\mathscr{J}_3$-submodule U of $\mathrm{Mat}_{v\times v}(\mathbb{C})$ is *thin* if the subspaces Δ_{A_i} are 1-dimensional, and say that it is *dual thin* if the subspaces $X_{E_j}U$ are 1-dimensional.

23.3.2 Lemma. *If $\mathscr{A}$ is metric, then a thin submodule of* $\mathrm{Mat}_{v\times v}(\mathbb{C})$ *is dual thin; if $\mathscr{A}$ is cometric then a dual thin submodule of* $\mathrm{Mat}_{v\times v}(\mathbb{C})$ *is thin.*

Proof. Suppose $\mathscr{A}$ is metric relative to the Schur idempotent A_1. If C is a $v \times v$ matrix, then

$$(A_1(A_i \circ C)) \circ A_j = 0$$

if $|i - j| > 1$. Hence if M is submodule of $\mathrm{Mat}_{v\times v}(\mathbb{C})$, then

$$A_1(A_i \circ M) \le A_{i-1} \circ M + A_i \circ M + A_{i+1} \circ M. \qquad (23.3.1)$$

Now let r denote the least positive integer such that $A_r \circ M \ne 0$, and let d be the greatest positive integer such that $A_{r+d-1} \circ M \ne 0$. From (23.3.1) it follows that if $r \le i \le r + d - 1$ then $A_i \circ M \ne 0$. We also see that M is generated by the subspace $A_d \circ M$ as an X_{A_1}-module. In other terms,

$$M = \langle A_1\rangle(A_d \circ M).$$

If E_j is a matrix idempotent, then

$$E_j M = E_j \langle A_1\rangle(A_r \circ M) = E_j(A_r \circ M)$$

If M is thin, then $\dim(A_r \circ M) = 1$ and therefore $\dim(E_j M) \le 1$ for all j. Therefore M is dual thin.

Suppose $\mathscr{A}$ is cometric relative to E_1 and let s be the least integer such that $E_s M \ne 0$. Then each column of a matrix in $E_j M$ lies in $\mathrm{col}(E_j)$, and so if $C \in M$, then each column of $E_1 \circ (E_i M)$ is the Schur product of a column of E_1 with a vector in $\mathrm{col}(E_i)$. Hence by ??? we have

$$E_1 \circ (E_i M) \le E_{i-1} M + E_i M + E_{i+1} M.$$

Given this, it is easy to prove the second part of the theorem. □

24
Hamming Schemes

24.1 The Binary Hamming Scheme

The Hamming scheme $H(d,2)$ is a metric and cometric association scheme. The matrix $A = A_1$ is the adjacency matrix of the d-cube, and its eigenvalues are the integers

$$d-2i, \quad i=0,\dots,d$$

with respective multiplicities

$$\binom{d}{i}.$$

The automorphism group of the Hamming scheme is vertex-transitive, and so the Terwilliger algebra is the same for each vertex.

We can write

$$A = R+L$$

where $L = R^T$ and R is the natural raising operator on the lattice of subsets of $\{1,\dots,d\}$. (So L is the natural lowering operator.)

24.1.1 Theorem. *The Terwilliger algebra of the binary Hamming scheme is a quotient of the enveloping algebra $U(sl(2,\mathbb{C}))$.*

Proof. View the vertices of the Hamming scheme as subsets of $\{1,\dots,d\}$. Define

$$H = RL-LR.$$

We note that

$$R_{\alpha,\beta} = 1$$

if and only if $\alpha \subseteq \beta$ and $|\beta| = |\alpha|+1$. Further $H_{\alpha,\beta} = 0$ if $|\alpha| \neq |\beta|$ and, if $|\alpha| = |\beta| = i$, then

$$H_{\alpha,\beta} = d-2i.$$

It follows that

$$H = \sum_{i=0}^{d} (d-2i)F_i$$

and hence the the algebra of all polynomials in H is the equal to the algebra generated by the diagonal matrices F_i.

Since

$$[R,L]=H, \quad [H,R]=2R, \quad [H,L]=-2L$$

the algebra generated by R, L and H is a homomorphic image of $U(sl(2,\mathbb{C}))$.

To complete the proof we must show that R and L generate the Terwilliger algebra of $H(n,d)$. But since the scheme is metric, each element of the Bose-Mesner algebra is a polynomial in A and since the algebra generated by H contains each F_i, we conclude that R and L generate the Terwilliger algebra. □

24.2 *Modules*

With what we know about the representation theory of $sl(2,\mathbb{C})$, it is easy to determine the irreducible $\mathbb{T}$-modules for the binary Hamming scheme $H(d,2)$. If u is a vertex of $H(d,2)$ with Hamming weight i, then the vectors

$$v, Rv, \ldots, R^{d-2i}v$$

are a basis for an irreducible module of dimension $d-2i+1$. If u and v are binary vectors then the irreducible modules they generate are isomorphic if and only if u and v have the same Hamming weight.

24.2.1 Lemma. *We have*

$$\dim(\mathbb{T}(H(d,2))) = \frac{1}{6}(d+1)(d+2)(d+3).$$

Proof. If $0 \le 2i \le d$, then our Terwilliger algebra has one isomorphism class of irreducible module with dimension $d-2i+1$, whence

$$\dim(\mathbb{T}(H(d,2))) = \sum_{i\le d/2}(d-2i+1)^2 = \frac{1}{6}(d+1)(d+2)(d+3). \qquad \square$$

24.2.2 Lemma. *The Terwilliger algebra of the Hamming scheme is thin and dual thin.*

Proof. If v has Hamming weight i, then the Hamming weight of each vector in $\operatorname{supp}(R^j v)$ is $i+j$. Hence $F_{i+j}R^j v = R_j v$, and therefore the R-module generated by v is thin. Since the Hamming schemes are metric, it follows from 23.3.2 that this module is also dual thin. □

Part VI

Background

25 Determinants

The determinant is a function on square matrices which plays many roles. If A is a square matrix over $\mathbb{R}$, its determinant is a measure of 'what A does to volume'. More precisely, if S is a region in $\mathbb{R}^n$ with unit volume, then the volume of set of points

$$\{Ax : x \in S\}$$

is $|\det(A)|$. Because of this, the determinant plays an important role in integration of functions of several variables.

25.1 Permutations

Let Ω be a set. A *permutation* of Ω is a bijection from Ω to itself. The set of all permutations of Ω is called the *symmetric group* on Ω. If $|\Omega| = n$, then $|\mathrm{Sym}(\Omega)| = n!$. We use $\mathrm{Sym}(n)$ to denote the set of all permutations on some set of size n, usually $\{1,\dots,n\}$. If $i \in \Omega$ and $\sigma \in \mathrm{Sym}(\Omega)$, then we denote the image of i under σ by i^σ.

Permutations of Ω are functions from Ω to Ω, so if ρ and σ are permutations, their *product* $\sigma\rho$ is defined by

$$i^{\sigma\rho} = (i^\sigma)^\rho.$$

This is again a permutation of Ω. As we will see, the order matters: usually $\sigma\rho \neq \rho\sigma$. Since a permutation is a bijection, it has an inverse. If $\sigma \in \mathrm{Sym}(\Omega)$, we denote the inverse of σ by σ^{-1}. We have

$$\sigma\sigma^{-1} = \sigma^{-1}\sigma.$$

The identity mapping on Ω is a bijection; we call it the *identity permutation* and denote it by 1. Finally, if ρ, σ and τ are permutations of Ω, then

$$(\rho\sigma)\tau = \rho(\sigma\tau).$$

In other words, multiplication of permutations is associative.

If $\Omega = \{1,\dots,n\}$ and $\sigma \in \Omega$, we can specify σ by writing down the sequence

$$1^\sigma, 2^\sigma, \dots, n^\sigma.$$

This is sometimes called the *Cartesian form* of the permutation. There is a second useful way to present permutations, which we develop now. Suppose $i \in \Omega$ and consider the infinite sequence of elements

$$i, i^{\sigma}, i^{\sigma^2}, \dots$$

by successively applying σ. Since Ω is finite there are integers r and s such that $r < s$ and

$$i^{\sigma^r} = i^{\sigma^s}.$$

Then

$$i = i^{\sigma^s \sigma^{-r}} = i^{\sigma^{s-r}}.$$

This shows that $r = 0$ and that s is the least integer such that $i^{\sigma^s} = i$. Hence the elements

$$i, i^{\sigma}, \dots, i^{\sigma^{s-1}}$$

are distinct. We call the cyclic sequence

$$(i, i^{\sigma}, \dots, i^{\sigma^{s-1}})$$

the *cycle* of σ that contains i. We can view σ as rotating the elements of this cycle.

We consider an example. Suppose $n = 7$ and the Cartesian form of σ is

$$2\,3\,1\,5\,6\,7\,4.$$

Then the cycle of σ that contains 1 is

$$(123)$$

and the cycle of σ that contains 5 is

$$(5674).$$

We regard this as equal to each of the cycles

$$(4567), (6745), (7456).$$

The distinct cycles of Ω form a partition of Ω. Together they determine σ—we can specify σ by simply listing its cycles. In the example at hand we may write

$$\sigma = (123)(4567).$$

The order in which we list the cycles is irrelevant. This is the *cyclic form* of σ. A permutation may have cycles of length one; it is conventional to omit this from the cyclic form if the underlying set is clear. (The cyclic form of the identity permutation is often denoted by (1).) Note that i lies in a cycle of length one if and only if it is fixed by σ, that is, $i^{\sigma} = i$.

Each cycle of a permutation is a permutation in its own right, and a permutation is the product of the permutations corresponding to its cycles.

A permutation is a *transposition* if it has one cycle of length two, and all other cycles have length one.

25.1.1 Theorem. *If $\sigma \in \mathrm{Sym}(n)$ and σ has exactly k cycles, then it is the product of $n-k$ transpositions.* □

We leave the proof as an exercise. By way of a hint we note that

$$(1234) = (12)(13)(14),$$

from which we see that a cycle of length m is the product of $m-1$ transpositions. We must count cycles of length one.

25.2 The Sign of a Permutation

A function of $x_1, \dots, x_n$ is *alternating* if, when τ is a transposition in $\mathrm{Sym}(n)$,

$$f^\tau = -f.$$

Thus $x_1 - x_2$ is an alternating function of two variables. If f is symmetric and g alternating in $x_1, \dots, x_n$, then fg is alternating. Define the function $V(x_1, \dots, x_n)$ by

$$V(x_1, \dots, x_n) = \prod_{i<j} (x_i - x_j).$$

Clearly V is alternating. Further, if $\sigma \in \mathrm{Sym}(n)$, then

$$V^\sigma = \mathrm{sign}(\sigma) V,$$

where $\mathrm{sign}(\sigma) = \pm 1$. The value of $\mathrm{sign}(\sigma)$ is called the *sign* of σ. If σ is a transposition, $\mathrm{sign}(\sigma) = -1$.

25.2.1 Theorem. *If $\sigma, \tau \in \mathrm{Sym}(n)$, then $\mathrm{sign}(\sigma\tau) = \mathrm{sign}(\sigma)\,\mathrm{sign}(\tau)$.*

Proof. We have

$$V^{\sigma\tau} = (\mathrm{sign}(\sigma) V)^\tau = \mathrm{sign}(\sigma)\,\mathrm{sign}(\tau) V$$

and therefore $\mathrm{sign}(\sigma\tau) = \mathrm{sign}(\sigma)\,\mathrm{sign}(\tau)$. □

By Theorem 25.1.1, each permutation is a product of transpositions, and therefore we have the following:

25.2.2 Corollary. *If f is an alternating function of n variables and $\sigma \in \mathrm{Sym}(n)$, then $f^\sigma = \mathrm{sign}(\sigma) f$.* □

The set of even permutations is known as the *alternating group*.

Since each permutation is a product of cycles, if we know the sign of these cycles, we can use the previous lemma to get the sign of the permutation itself.

25.2.3 Lemma. *The sign of a cycle is odd if and only if its length is even.*

Proof. It follows from Theorem 25.1.1 that a cycle of length k can be written as the product of $k-1$ transpositions. Since the sign of a transposition is odd, the sign of a cycle of length k is $(-1)^{k-1}$. □

25.2.4 Corollary. *If a permutation has exactly e even cycles, its sign is $(-1)^e$.* □

25.3 Permutation Matrices

Let $\mathbb{F}$ be a field. If $\sigma \in \mathrm{Sym}(n)$, let $P(\sigma)$ be the linear transformation that maps

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} x_{1^\sigma} \\ x_{2^\sigma} \\ \vdots \\ x_{n^\sigma} \end{pmatrix}.$$

Thus if $e_1, \dots, e_n$ is the standard basis for $\mathbb{F}^{n\times 1}$, then $P(\sigma)$ maps e_j to $e_{j^{\sigma^{-1}}}$. Hence the coordinate matrix for $P(\sigma)$ is

$$\begin{pmatrix} e_{1^{\sigma^{-1}}} & e_{2^{\sigma^{-1}}} & \dots & e_{n^{\sigma^{-1}}} \end{pmatrix}.$$

The inverses are annoying, it may help to note that the i-th row of this matrix is $e_{i^\sigma}^T$. We call $P(\sigma)$ are permutation operator and the matrix which represents it is a *permutation matrix*.

The product of two permutation operators is a permutation operator, and consequently the product of two permutation matrices is a permutation matrix.

If P is a permutation matrix then $PP^T = I$, and therefore $P^{-1} = P^T$.

A matrix is a permutation matrix if it is a 01-matrix, and exactly one entry in each row and column is equal to 1. We define a matrix to be a *monomial matrix* there is at most one non-zero entry in each row and each column. It is not hard to verify that a matrix M is monomial if $M = PD$, where P is a permutation matrix and D is diagonal. Similarly DP is monomial. If P is a permutation matrix and D is diagonal, then

$$P^{-1}DP$$

is diagonal.

25.3.1 Lemma. *The product of two monomial matrices of the same order is a monomial matrix.*

Proof. Suppose P_1 and P_2 are permutation matrices and D_1 and D_2 are diagonal. Then P_1D_1 and P_2D_2 are monomial and

$$(P_1D_1)(P_2D_2) = P_1P_2(P_2^{-1}D_1P_2)D_2.$$

Here P_1P_2 is a permutation matrix and $(P_2^{-1}D_1P_2)D_2$ is a product of diagonal matrices, and so is diagonal. Hence $(P_1D_1)(P_2D_2)$ is a monomial matrix. □

25.4 Definition of the Determinant

In this section we define the determinant of a square matrix, and develop some of its properties.

For this we will use a somewhat unusual matrix product: it is commutative and associative and distributes over addition. If A and B are $m \times n$ matrices, we define their *Schur product* $A \circ B$ by

$$(A \circ B)_{i,j} = A_{i,j} B_{i,j}.$$

There are no difficulties in working with this product. If A and P are $n \times n$ matrices and P is a permutation matrix, then $A \circ P$ is a monomial matrix.

The *determinant* is a function from the set of $n \times n$ matrices over a field (e.g., $\mathbb{R}$ or $\mathbb{C}$) to the field itself. We define it in stages. If D is diagonal, then

$$\det(D) := \prod_{i=1}^{n} D_{i,i}.$$

If M is monomial, then $M = DP$ where D is diagonal and P is a permutation matrix. If $P = P(\sigma)$ for some permutation σ, we define $\operatorname{sign}(P)$ to be $\operatorname{sign}(\sigma)$ and then

$$\det(M) := \det(D) \operatorname{sign}(P).$$

Note that

$$PD = (PDP^{-1})P$$

where PDP^{-1} is diagonal. Since PDP^{-1} is diagonal and $\det(PDP^{-1}) = \det(D)$,

$$\det(PD) = \det(PDP^{-1}) \operatorname{sign}(P) = \det(D) \operatorname{sign}(P).$$

It is implicit in this that, if P is a permutation matrix, then $\det(P) = \operatorname{sign}(P)$.

To complete the definition of the determinant, let $\operatorname{Perm}(n)$ denote the set of all $n \times n$ permutation matrices. If $A \in \operatorname{Mat}_{n\times n}(\mathbb{F})$, we define

$$\det(A) := \sum_{\pi \in \operatorname{Sym}(n)} \det(A \circ P(\pi)).$$

By way of example, if $n = 2$ then $\operatorname{Perm}(2)$ consists of the two matrices

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

and so if

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

then

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

25.4.1 Lemma. *Let A be an $n \times n$ matrix. If A is lower triangular, then*

$$\det(A) = \prod_{i=1}^{n} A_{i,i}.$$

Proof. Suppose $P \in \operatorname{Perm}(n)$. If $\det(A \circ P) \neq 0$, then P must be lower triangular, but the identity matrix is the only lower triangular permutation matrix. Therefore $\det(A) = \det(A \circ I)$, and the lemma follows. □

25.4.2 Lemma. *If A is a square matrix,* $\det(A^T) = \det(A)$.

Proof. We note first that if M is monomial, so is M^T. Further, if $M = DP$ where D is diagonal and P is a permutation matrix, then

$$\det(M^T) = \det(P^T D) = \det((P^T DP)P^{-1}),$$

Since $P^{-1} = P^T$, we see that $P^T DP$ is diagonal, and therefore

$$\det(M^T) = \det(P^T DP)\operatorname{sign}(P^{-1}) = \det(D)\operatorname{sign}(P) = \det(M).$$

Now

$$\begin{aligned}\det(A^T) &= \sum_{P\in\mathrm{Perm}(n)} \det(A^T \circ P)\\ &= \sum_{P\in\mathrm{Perm}(n)} \det(A\circ P^T)^T\\ &= \sum_{P\in\mathrm{Perm}(n)} \det(A\circ P^T)\\ &= \sum_{P\in\mathrm{Perm}(n)} \det(A\circ P)\\ &= \det(A).\end{aligned}$$

25.5 The Determinant is Multiplicative

The determinant is useful in particular because, if A and B are square matrices of the same order, then $\det(AB) = \det(A)\det(B)$. We work towards a proof of this.

We work with functions on $n \times n$ matrices. We may think of such a function δ as a function of n variables, the columns of the matrix. To indicate this, if A is $n \times n$ and $e_1, \ldots, e_n$ is the standard basis of $\mathbb{F}^{n\times 1}$, we may use $\delta(Ae_1, \ldots, Ae_n)$ in place of of $\delta(A)$. A function $\delta : \mathrm{Mat}_{n\times n}(\mathbb{F}) \to \mathbb{F}$ is *multilinear* if $\delta(A)$ is a linear function of each column of A. If δ is multilinear and $Ae_1 = x + y$, then

$$\delta(A) = \delta(Ae_1, \ldots, Ae_n) = \delta(x, Ae_2, \ldots, Ae_n) + \delta(x, Ae_2, \ldots, Ae_n).$$

Note that trace, although it is a linear function of A, is not multilinear. However, if P is a permutation matrix then the function δ_P given by

$$\delta_P(A) = \det(A\circ P)$$

is multilinear. (Prove it.) If δ_1 and δ_2 are multilinear, then their sum, given by

$$(\delta_1 + \delta_2)(A) = \delta_1(A) + \delta_2(A),$$

is multilinear.

A function $\delta : \mathrm{Mat}_{n\times n}(\mathbb{F})$ to $\mathbb{F}$ is *alternating* if $\delta(A) = 0$ whenever two columns of A are equal. This usage is different from the one used in Section 25.2, but we will see that it is consistent with it.

We need two preliminary results.

25.5.1 Lemma. *If M_1 and M_2 are $n \times n$ monomial matrices, then* $\det(M_1 M_2) = \det(M_1)\det(M_2)$.

Proof. We may suppose that for $i = 1, 2$,

$$M_i = D_i P_i$$

where D_i is diagonal and P_i is a permutation matrix. Then

$$M_1 M_2 = D_1 P_1 D_2 P_2 = D_1 (P_1 D P_1^{-1}) P_1 P_2.$$

Here $P_1 D P_1^{-1}$ is diagonal, so $D_1 (P_1 D P_1^{-1})$ is diagonal and also $P_1 P_2$ is a permutation matrix. Therefore $M_1 M_2$ is monomial and

$$\begin{aligned}\det(M_1 M_2) &= \det(D_1 (P_1 D P_1^{-1}))\,\mathrm{sign}(P_1 P_2)\\ &= \det(D_1)\det(D_2)\,\mathrm{sign}(P_1)\,\mathrm{sign}(P_2)\\ &= \det(D_1 P_1)\det(D_2 P_2)\\ &= \det(M_1)\det(M_2).\end{aligned}$$

This completes the proof. □

25.5.2 Lemma. *If A, B and P are $n \times n$ matrices and P is a permutation matrix, then* $(A \circ B)P = (AP) \circ (BP)$.

Proof. Suppose $e_1, \dots, e_n$ is the standard basis and $Pe_i = e_j$. Then

$$\begin{aligned}((AP) \circ (BP))e_i &= (AP)e_i \circ (BP)e_i\\ &= Ae_j \circ Be_j\\ &= (A \circ B)e_j\\ &= (A \circ B)Pe_i.\end{aligned}$$

Since this works for all i, we have proved the lemma. □

25.5.3 Theorem. *The determinant is an alternating multilinear function of the columns of a matrix.*

Proof. Since the functions δ_P are multilinear and since det is the sum of the functions δ_P, it follows that det is multilinear.

To show that det is alternating, we first prove that if Q is a permutation matrix, then $\det(AQ) = \det(A)\,\mathrm{sign}(Q)$. Using the previous two lemmas, we have

$$\begin{aligned}\det(AQ) &= \sum_{P \in \mathrm{Perm}(n)} \det((AQ) \circ P)\\ &= \sum_{P \in \mathrm{Perm}(n)} \det[(A \circ PQ^{-1}))Q]\\ &= \sum_{P \in \mathrm{Perm}(n)} \det(A \circ (PQ^{-1}))\det(Q)\\ &= \det(Q) \sum_{P \in \mathrm{Perm}(n)} \det(A \circ (PQ^{-1})).\end{aligned}$$

Since

$$\{P : P \in \text{Perm}(n)\} = \{PQ^{-1} : P \in \text{Perm}(n)\},$$

the last sum above equals $\det(A)$, we have proved that $\det(AQ) = \det(A)\,\text{sign}(Q)$, as claimed.

Now suppose columns i and j of A are equal, let τ be the transposition $(i\,j)$ and let $T = P(\tau)$. Then $\text{sign}(T) = -1$, $T^2 = I$ and $AT = A$; hence

$$(A \circ P)T = (AT) \circ PT = A \circ PT$$

and consequently

$$\begin{aligned}\det(A \circ P) + \det(A \circ PT) &= \det(A \circ P) + \det((A \circ P)T) \\ &= \det(A \circ P) + \det(A \circ P)\det(T) \\ &= \det(A \circ P) - \det(A \circ P) \\ &= 0.\end{aligned}$$

The set $\{P, PT\}$ is the left coset of the subgroup $\{I, T\}$ of $\text{Perm}(n)$. For fixed T, the set $\text{Perm}(n)$ can be partitioned into pairs of the form $\{P, PT\}$ (prove this), and therefore it follows that $\det(A) = 0$. □

One corollary of this proof is that if $P \in \text{Perm}(n)$, then $\det(AP) = \det(A)\,\text{sign}(P)$. Hence the determinant is an alternating function in the sense we used in Section 25.2. More generally, the same argument shows that if δ is an alternating function on $n \times n$ matrices and P is a permutation matrix, then

$$\delta(AP) = \delta(A)\,\text{sign}(P).$$

Therefore a function that is alternating in the sense of this section is alternating in the sense we used in Section 25.2, but the current definition is more useful if we work over fields such as $\mathbb{Z}_2$.

Our next result is a converse to the previous theorem.

25.5.4 Theorem. *If δ is an alternating multilinear function on $n \times n$ matrices and $\delta(I) = 1$, then $\delta(A) = \det(A)$ for all $n \times n$ matrices.*

Proof. We have

$$Ae_j = \sum_{i=1}^{n} A_{i,j} e_i.$$

Since δ is multilinear,

$$\delta(A) = \delta(Ae_1, \ldots, Ae_n) = \sum_{i=1}^{n} \delta(A_{i,1} e_i, Ae_2, \ldots, Ae_n)$$

and, using even more subscripts,

$$\delta(A) = \sum_{1 \le i_1, \ldots, i_n \le n} \delta(A_{i_1,1} e_{i_1}, \ldots, A_{i_n,n} e_{i_n}). \tag{25.5.1}$$

Since δ is multilinear,

$$\delta(A_{i_1,1}e_{i_1},\dots,A_{i_n,n}e_{i_n}) = \delta(e_{i_1},\dots,e_{i_n})\prod_{k=1}^{n} A_{i_k,k};$$

and since δ is alternating if $r < s$ and $i_r = i_s$, then

$$\delta(e_{i_1},\dots,e_{i_n}) = 0$$

Hence in (25.5.1), the summands indexed by the sequences $i_1,\dots,i_n$ that are not permutations are zero, and therefore

$$\delta(A) = \sum_{P\in\operatorname{Perm}(n)} \delta(A\circ P).$$

This shows that δ is determined by the values it takes on monomial matrices.

If D is diagonal and P is a permutation matrix, then since δ is alternating,

$$\delta(DP) = \delta(D)\operatorname{sign}(P).$$

Further, since δ is multilinear,

$$\delta(D) = \prod_{i=1}^{n} D_{i,i}\delta(I)$$

and therefore

$$\delta(DP) = \det(DP)\delta(I).$$

This completes the argument. □

25.5.5 Corollary. *If A and B are $n\times n$ matrices, then* $\det(AB) = \det(A)\det(B)$.

Proof. Consider the function δ from $\operatorname{Mat}_{n\times n}(\mathbb{F})$ to $\mathbb{F}$, given by

$$\delta(B) := \det(AB).$$

It is easy to verify that this is alternating and multilinear, and therefore

$$\delta(B) = c_A \det(B)$$

for some scalar c_A. Taking $B = I$ in the definition of δ, we see that $c_A = \det(A)$ and therefore $\det(AB) = \det(A)\det(B)$. □

25.6 The Laplace Expansion

The determinant is remarkable for the number of different ways in which we can compute it. Here we describe an approach due to Laplace. You may be familiar with the case when $k = 1$, because this is the well-known expansion by cofactors.

If $T = \{t_1, \dots, t_k\}$, define $\|T\|$ by

$$\|T\| = \sum_{i=1}^{k} (t_i - i).$$

Let $A_{S,T}$ denote the submatrix of A with rows indexed by S and columns by T. If $|S| = |T| = 1$, then $A_{S,T}$ is just an entry of A. We use $\overline{S}$ to denote the complement of S in $\{1, \dots, n\}$. Now we can state and prove a result known as Laplace's expansion of the determinant.

25.6.1 Theorem. *Let A be an $n \times n$ matrix and let S and S' be two subsets of $\{1, \dots, n\}$, with sizes k and $n-k$ respectively. Then*

$$\sum_{T : |T| = k} (-1)^{\|T\|} \det(A_{S,T}) \det(A_{S',\overline{T}}) = \begin{cases} (-1)^{\|S\|} \det(A), & \text{if } S' = \overline{S}; \\ 0, & \text{otherwise.} \end{cases}$$

Proof. We first consider the case where $S' = \overline{S}$. Let S and T be subsets of $\{1, \dots, n\}$ with size k. Then

$$\det A = \sum_{T} \sum_{\sigma : S^\sigma = T} \det(A \circ P(\sigma)).$$

Note that if σ maps S to T then it must map $\overline{S}$ to $\overline{T}$. Hence

$$\sum_{\sigma : S^\sigma = T} \det(A \circ P)(\sigma) = (-1)^{\|T\|} \det(A_{S,T}) \det(A_{\overline{S},\overline{T}}).$$

Now suppose that $S' \neq \emptyset$. Let A' be the matrix whose first k rows are the rows of A indexed by S_1, and whose last $n-k$ rows are the rows of A indexed by S_2. Since we know that Laplace's expansion holds when $S \cap S' = \emptyset$, we see that $\det(A')$ is equal to the sum on the left on the statement of the theorem. On the other hand, A' has a repeated row, and therefore $\det(A') = 0$. □

Let $A(i|j)$ denote the matrix we get from the square matrix A by deleting row i and column j. Then $(-1)^{i+j} \det(A(i|j)$ is called the *ij-cofactor* of A. The following special case of the Laplace expansion is known the *expansion by cofactors* of $\det(A)$. This is somtimes used as a definition of the determinant.

25.6.2 Corollary. *Let A be an $n \times n$ matrix. Then*

$$\det(A) = (-1)^{i-1} \sum_{j=1}^{n} (-1)^{j-1} A_{i,j} \det(A(i|j)). \qquad \square$$

Let A be an $n \times n$ matrix. We define the *adjugate* $\operatorname{adj}(A)$ of A as follows:

$$\operatorname{adj}(A)_{i,j} = (-1)^{i+j} \det A(i|j).$$

Thus if

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

then

$$\operatorname{adj}(A) = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

If

$$J = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

then $\operatorname{adj}(J) = 0$.

Applying the previous theorem with $k = 1$, we obtain:

25.6.3 Corollary. *If A is a square matrix, then $A \operatorname{adj}(A) = \det(A) I$.* □

It is also true that $\operatorname{adj}(A)A = \det(A)$; this can be proved using the transpose. We leave the proof as an exercise.

25.6.4 Corollary. *If A is a square matrix, then it is invertible if and only if $\det(A)$ is.*

Proof. If $\det(A)$ is invertible, the previous corollary implies that

$$A^{-1} = \det(A)^{-1} \operatorname{adj}(A).$$

If A ia invertible then

$$1 = \det(I) = \det(AA^{-1}) = \det(A)\det(A^{-1})$$

and therefore $\det(A)$ is invertible. □

The following identity is due to Jacobi.

25.6.5 Theorem. *Let A be an $n \times n$ matrix and suppose $S \subseteq \{1, \dots, n\}$. If $s = |S|$, then*

$$\det(\operatorname{adj}(A)_{\bar{S},\bar{S}}) = \det(A)^{n-1-s} \det(A_{S,S}).$$

Proof. If M is $n \times n$, we have $\operatorname{adj}(M)M = \det(M)I$ and, taking determinants of both sides yields

$$\det(\operatorname{adj}(M)) \det(M) = \det(M)^n.$$

Therefore $\det(\operatorname{adj}(M)) = \det(M)^{n-1}$. Assume S consists of the first s elements of $\{1, \dots, n\}$. We have $\operatorname{adj}(A)A = \det(A)I$ whence $\operatorname{adj}(A)Ae_i = \det(A)e_i$ and

$$\operatorname{adj}(A) \begin{pmatrix} Ae_1 & \dots & Ae_s & e_{s+1} & \dots & e_n \end{pmatrix} = \begin{pmatrix} \det(A)I_s & ? \\ 0 & \operatorname{adj}(A)_{\bar{S},\bar{S}} \end{pmatrix}$$

Taking the determinant of each side, we get

$$\det(A)^{n-1} \det(A_{S,S}) = \det(A)^s \det(\operatorname{adj}(A_{\bar{S},\bar{S}})).$$

This yields the theorem. □

25.7 The Characteristic Polynomial of a Matrix

If A is a square matrix then $\det(tI - A)$ is a polynomial in t. It is called the *characteristic polynomial* of A. It is not too difficult to verify that if A is $n \times n$, then its characteristic polynomial is a monic polynomial of degree n. If

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

then

$$\det(tI - A) = t^2 - (a + c)t + (ac - bd).$$

The constant term of the characteristic polynomial of A is

$$\det(-A) = (-1)^n \det(A).$$

Suppose $A = LBL^{-1}$. Then

$$\begin{aligned} \det(tI - A) = \det(tI - LBL^{-1}) &= \det[L(tI - B)L^{-1}] \\ &= \det(B)\det(tI - B)\det(L^{-1}) \\ &= \det(tI - B). \end{aligned}$$

Thus we see that similar matrices have the same characteristic polynomial.

We leave the proof of the following as an exercise.

25.7.1 Lemma. *If $\phi(t)$ is the characteristic polynomial of the square matrix A, then the coefficient of t^{n-1} is $-\operatorname{tr}(A)$.* □

Our next result is called the Cayley-Hamilton theorem. Cayley proved it for 2 × 2 and 3 × 3 matrices.

25.7.2 Theorem. *If $\phi(t)$ is the characteristic polynomial of the square matrix A, then $\phi(A) = 0$.*

Proof. Each entry of $\operatorname{adj}(tI - A)$ is a polynomial in t with degree at most $n - 1$. Hence there are matrices $B_1, \dots, B_n$ such that

$$\operatorname{adj}(tI - A) = B_n + tB_{n-1} + \cdots + t^{n-1}B_1$$

We want to show that each of the matrices $B_1, \dots, B_n$ is a polynomial in A.

We have

$$\begin{aligned} &(tI - A)\operatorname{adj}(tI - A) \\ &\qquad = t^n B_1 + t^{n-1}(B_2 - AB_1) + \cdots + t(B_n - AB_{n-1}) + (-A)B_n. \qquad (25.7.1) \end{aligned}$$

Assume that

$$\phi(t) = t^n + a_1 t^{n-1} + \cdots + a_n.$$

From Corollary 25.6.3 we have

$$(tI - A)\operatorname{adj}(tI - A) = (t^n + a_1 t^{n-1} + \cdots + a_n)I. \qquad (25.7.2)$$

If we equate the coefficients of the powers of t, we obtain:

$$B_1 = I, \quad B_{i+1} = AB_i + a_i I \quad (i = 1, \ldots, n-1)$$

whence

$$\begin{aligned} B_1 &= I \\ B_2 &= A + a_1 I \\ B_3 &= AB_2 + a_2 I = A^2 + a_1 A + a_2 I \end{aligned}$$

and, in general,

$$B_{k+1} = A^k + a_1 A^{k-1} + \cdots + a_k I.$$

Thus B_k is a polynomial of degree $k-1$ in A.

From (25.7.1) and (25.7.2), we see that $a_n I = -AB_n$. So

$$\begin{aligned} 0 = AB_n + a_n I &= A(A^{n-1} + a_1 A^{n-2} + \cdots + a_{n-1} I) + a_n I \\ &= \phi(A). \end{aligned}$$

This completes the proof. □

It is tempting to argue that if we substitute A for t in the equation

$$(tI - A)\,\mathrm{adj}(tI - A) = \phi(t)I,$$

then $tI - A$ becomes zero, and therefore $\phi(A) = 0$. It is true that if $f(t)$ is a polynomial in t with coefficients in a field and $t - a$ divides $f(t)$, then $f(a) = 0$. It need not be true that if $f(t)$ and $f_1(t)$ are polynomials in t with matrices as coefficients and

$$(tI - A) f_1(t) = f(t)$$

then $f(A) = 0$. The basic problem is, for example, that if b is a scalar then

$$t^2 b = tbt = bt^2,$$

but if A and B are square matrices, then the products A^2B, ABA and BA^2 can all be different.

25.8 An Algorithm

If we attempt to compute the determinant of a matrix in $\mathrm{Mat}_{n\times n}(\mathbb{Z})$ using our definition, we may be obliged to sum $n!$ products. This is already unpleasant when $n = 4$. There is a second algorithm using elementary row operations; the only disadvantage of this is that its intermediate stages often require the use of rational numbers, even though the final answer is an integer. (This is the algorithm usually taught.) We are going to describe a a third algorithm that does not suffer from this disadvantage, and still runs in polynomial time.

Let A be an $m \times n$ matrix and suppose $k \leq m, n$. We construct an $(m+1-k) \times (n+1-k)$ matrix $D_k(A)$ from A as follows. If $k \leq r \leq m$ and $k \leq s \leq n$, there is a unique $k \times k$ submatrix of A that contains the rs-entry of A along with all entries in the first $k-1$ rows and columns. Define $D_k(A)_{r-k,s-k}$ to be the determinant of this submatrix. So $D_1(A) = A$ and if

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

then

$$D_2(A) = \begin{pmatrix} a_{1,1}a_{2,2} - a_{1,2}a_{2,1} & a_{1,1}a_{2,3} - a_{2,1}a_{1,3} \\ a_{1,1}a_{3,2} - a_{1,2}a_{3,1} & a_{1,1}a_{3,3} - a_{3,1}a_{1,3} \end{pmatrix}.$$

If A is $n \times n$, then $D_n(A) = \det(A)$. For any matrix A, let $d_k(A)$ denote the determinant of the submatrix formed by the first k rows and columns; we assume $d_0(A) = 1$.

25.8.1 Lemma. *If A is an $m \times n$ matrix, then $D_2(D_k(A)) = d_{k-1}(A)D_{k+1}(A)$.*

Proof. We prove the result by induction on the size of A. Since $D_1(A) = A$, the lemma holds when $k = 1$ and we assume $k \geq 2$.

First we consider a special case. Suppose A is $(k+1) \times (k+1)$. Then $D_{k+1}(A) = \det(A)$ and

$$D_k(A) = \begin{pmatrix} \det(A(k+1|k+1) & \det(A(k+1|k)) \\ \det(A(k|k+1)) & \det(A(k|k)) \end{pmatrix}$$

Therefore

$$\begin{aligned} \det(D_k(A)) = \det(A(k|k)) & \det(A(k+1|k+1)) \\ & - \det(A(k+1|k)) \det(A(k|k+1)) \end{aligned}$$

and so if $S := 1, \dots, k-1$, then

$$D_2(D_k(A)) = \det(\operatorname{adj}(A)_{\overline{S},\overline{S}}).$$

By Jacobi's identity (Theorem 25.6.5),

$$\det(\operatorname{adj}(A)_{\overline{S},\overline{S}} = \det(A)\det(A_{S,S}) = d_{k-1}D_{k+1}(A).$$

Now we verify that the result follows from this special case. If $i \geq k$ we and B is the matrix we get by deleting the i-th row of A, then $D_k(B)$ is obtained from $D_k(A)$ by deleting its $(i+1-k)$-th row. Since that $D_k(A^T) = D_k(A)$, a similar claim holds when we delete columns.

If $i, j \geq k+1$, then $(D_{k+1}(M))_{i-k,j-k}$ is the determinant of the submatrix M of A formed by the intersection of rows 1 through k and i with columns 1 through k and j. Since $d_{k-1}(A) = d_{k-1}(M)$, we have

$$\begin{aligned} d_{k-1}(A)D_{k+1}(A)_{i-k,j-k} &= d_{k-1}(M)D_{k+1}(M) \\ &= D_2(D_k(M)) \\ &= D_2(D_k(A))_{i-k,j-k} \end{aligned}$$

and so the result follows. □

The algorithm to compute $\det(A)$ runs as follows. The input is an $n \times n$ matrix A. We also use a scalar δ, which is initially set to 1.

1. If $n = 1$, then $\det(A) = A$; halt.

2. If the first row or column of A is zero, then $\det(A) = 0$; halt.

3. If necessary, swap two columns of A so that $A_{1,1} \neq 0$ and replace δ by $-\delta$.

4. Compute $\delta^{-1}D_2(A)$ and let $\delta = (A)_{1,1}$. Return to the first step with $\delta^{-1}D_2(A)$ in place of A.

After $n-1$ steps of this kind, we obtain $D_n(A) = \det(A)$.

We give one example. If

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

then

$$D_2(A) = \begin{pmatrix} x^2-1 & -x \\ -x & x^2 \end{pmatrix}$$

Since $d_1(A) = x$,

$$\det(A) = D_3(A) = x^{-1}(x^4 - 2x^2) = x^3 - 2x.$$

This algorithm is sometimes attributed to C. Dodgson, better known as Lewis Carroll.

25.9 Summary

The most useful facts are (c), (f) and (g). You are not required to know anything about the proofs of (f), (g), (h), (i) and (j). You might need to use them. Note that (d) and (e) together yield an algorithm for computing the determinant, since we can bring a matrix to triangular form by elementary row operations.

(a) Permutations, sign of a permutation, permutation and monomial matrices.

(b) Definition of determinant.

(c) $\det(A^T) = \det(A)$

(d) If A is triangular, $\det(A) = \prod_i A_{i,i}$.

(e) Adding a scalar multiple of one row of A to another does not change $\det(A)$. Swapping rows changes the sign. Ditto for columns. If we get B from A by multiplying a column by c, then $\det(B) = c\det(A)$.

(f) Multilinear and alternating functions on matrices, $\det(AB) = \det(A)\det(B)$.

(g) The adjugate of a matrix, $A\,\mathrm{adj}(A) = \det(A)I$.

(h) Cofactor expansion of $\det(A)$.

(i) The Cayley-Hamilton theorem.

(j) Bareiss algorithm.

(k) When the products AB and BA are both defined, $\det(I - AB) = \det(I - BA)$.

(l) Binet-Cauchy.

(m) $\det(\exp(M)) = \exp(\mathrm{tr}(M))$.

(We did not treat the last three items.)

25.10 Groups

In this chapter we met the 'symmetric group' and the 'alternating group'. As we continue with the course, we will meet other 'groups'. For the sake of background information, we explain the terminology.

A *group* is a set G with a multiplication $\circ$ defined on it. If $a, b \in G$, then $a \circ b$ denotes the product of a and b. (In many cases the elements of G are operations on some structure, and $a \circ b$ denotes "do a, then b".) The multiplication must satisfy the following axioms.

1. If $a, b \in G$, then $a \circ b \in G$.

2. If $a, b, c \in G$, then $(a \circ b) \circ c = a \circ (b \circ c)$.

3. There is an element θ in G such that $\theta \circ a = a$ for all a in G.

4. For each element $a \in G$, there is an element a^{-1} in G such that $a^{-1} \circ a = \theta$.

The first axiom states that G is closed under multiplication. The element θ is the *identity element* of the group. The element a^{-1} is the *inverse* of a. We do **not** assume that $a \circ b = b \circ a$; if this does hold for all a and b the group is *commutative* (or *abelian*).

One example of a group is the integers, with + as the 'multiplication'. A second example is the set of invertible $n \times n$ matrices over a field with the usual matrix multiplication.

We usually write ab in place of $a \circ b$ unless G is commutative, in which case we write $a + b$. We usually use 1 to denote the identity unless G is commutative, when we use 0.

Suppose $a, x, y \in G$ and $ax = ay$. Then

$$x = 1x = (a^{-1}a)x = a^{-1}(ax) = a^{-1}(ay) = (a^{-1}a)y = 1y = y.$$

Thus in a group we may 'cancel on the left'. Since

$$a^{-1}(a1) = (a^{-1}a)1 = 1^2 = 1 = a^{-1}a,$$

it follows (by left cancellation) that $a1 = a$ for all a. Since

$$(aa^{-1})a = a(a^{-1}a) = a1 = a = 1a$$

we also see that $aa^{-1} = 1$ for any a. Now if $xa = ya$, then

$$x = x1 = x(aa^{-1}) = (xa)a^{-1} = (ya)a^{-1} = y(aa^{-1}) = y1 = y;$$

therefore we may also cancel on the right.

A subset of G is a *subgroup* if it contains the inverse of each of its elements and is closed under multiplication. The alternating group is a subgroup of the symmetric group.

Finally we point out that a group is a set with three operations. A binary operation which, given (a, b) as input, returns $a \circ b$. A unary operation which, given a as input, returns a^{-1}. And a nullary operation which, given no input, returns the identity θ. (It may help to understand the last statement if you think of a button on a calculator labelled π—this takes no input and returns π.)

26
Rings, Fields, Algebras

Thus chapter is meant to to provide some background, to help you deal with linear algebra over fields other than $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$.

26.1 Rings

A ring R consists of a set R on which an addition operation $+$ is defined, such that $(R, +)$ is a commutative group; in addition there is an associative multiplication in R that satisfies the usual distributive laws relative to addition. The multiplication is usually denoted by juxtaposition, i.e., the product of a and b is denoted ab (and ab need not equal ba.). We always assume that there is multiplicative identity, denoted by 1 (so $1x = x1 = x$ for all x in R).

The canonical examples are $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$. Polynomials over $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$ form a ring, and so do power series. Further, matrices with entries from a ring R form a ring which is not normally commutative. Continuous real functions on $\mathbb{R}$ form a ring.

Rings were first introduced in number theory, but now it is somewhat unusual for a mathematician not to be working in the context of some ring.

As a general principal, any operation we can carry out on abelian groups can be carried out on rings. So we have subrings, products and homomorphisms/quotients. Somewhat surprisingly, subrings do not play a big role, except for ideals (which you can look up). Also finite rings seem to be less useful than finite groups.

26.2 Fields

A field is a ring in which every non-zero element has a multiplicative inverse. The canonical examples are $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$. We see that $\mathbb{Z}$ is not a field, and the rings of polynomials we referred to above are not fields (although they can be used to construct fields). If $\mathbb{F}$ is a field then $\mathbb{F}(t)$, the ring of rational functions with coefficients from $\mathbb{F}$ is a field.

As just defined, the multiplication in a field need not be commutative.

However all fields we need are commutative and so henceforth field means commutative field.

The integers modulo a prime p form a field $\mathbb{Z}_p$. We consider this in some detail. Stricly speaking, the elements of $\mathbb{Z}_p$ are equivalence classes of integers, where integers m and n are equivalent, i.e., $m \equiv n$, if p divides $m - n$. Each equivalence class contains exactly one element from the set of integers

$$]\{0, 1, \dots, p-1\}$$

and so we can identify the equivalence classes with the members of this set. It is not too difficult to show that the equivalence classes form a ring, with addition mod p and multiplication mod p as its operations. In fact we can show that, for any positive integer n, the set $\mathbb{Z}_n$ forms a ring. But if n is not a prime we can write $n = ab$ where a and b both greater than 1, and therefore $ab = 0$ in $\mathbb{Z}_n$. It follows that the equivalence class of a does not have a multiplicative inverse—if $xa = 1$ and $ab = 0$ then

$$0 = x(ab) = (xa)b = 1b = b.$$

Therefore if n is not prime, then $\mathbb{Z}_n$ is not a field.

If p is a prime then each non-zero element of $\mathbb{Z}_p$ does have a multiplicative inverse. For if $a \in \mathbb{Z}_p$ and $a \neq 0$, then the gcd of a and p is 1, and hence there are integers x and y such that

$$xa + yp = 1,$$

and therefore $xa = 1$. Thus we can find the multiplicative inverse of a using the Euclidean algorithm. We have been a little sloppy here: when we apply the Eulidean algorithm we are viewing a and p as integers, but we originally chose a to be a non-zero element of the ring $\mathbb{Z}_p$. To avoid this we should use some notation like $[a]$ to denote the equivalence class of a, but the sloppiness is easier, and traditional.

We can also construct fields from rings of polynomials. Let $\mathbb{F}$ be a field and let $\mathbb{F}[t]$ denote the ring of polynomials with coefficients from $\mathbb{F}$. If $p(t)$ is a monic polynomial in $\mathbb{F}[t]$, define a relation $\equiv$ on $\mathbb{F}[t]$ by declaring polynomials g and h to be equivalent if their difference is divisible by p. Then this is an equivalence relation and the equivalence classes form a ring. You may show that this ring is a field if and only if p is irreducible over $\mathbb{F}$ (has no non-trivial factors).

If we take $\mathbb{F} = \mathbb{R}$ and $p(t) = t^2 + 1$, this construction produces a field isomorphic to the complex numbers. If $\mathbb{F} = \mathbb{Z}_2$ and $p(t) = t^2 + t + 1$, we obtain a field with four elements.

Exercise: Let $\mathbb{E}$ be a field and let F be the subset of $\mathbb{E}$ consisting of all the elements of $\mathbb{E}$ we can get by adding 1 to itself any number of times. (By assumption, $0 \in F$; thus F is the additive subgroup of $\mathbb{E}$ generated by 1.) Show that F is a ring. If $|F|$ is finite, prove that it is a prime, and deduce that it is a field.

26.3 Algebras

A ring R is an algebra over a field $\mathbb{F}$ if R is a vector space over $\mathbb{F}$ such that if $x, y \in R$ and $a \in \mathbb{F}$, then

$$(ax)y = x(ay) = a(xy).$$

If 1 is the multiplicative identity in R, then the set $\{a1 : a \in \mathbb{F}\}$ forms a subring of R that is isomorphic to $\mathbb{F}$. Each element of this subring commutes with each element of R (it lies in the *center* of R).

The term 'algebra' has changed its meaning over the years, and it still has more than one interpretation. As we have just defined it, every algebra contains a multiplicative unit, but, in analysis for example, this requirement can be dropped.

The set of $d \times d$ matrices over of field $\mathbb{F}$ forms an algebra. More generally, the set of linear mappings of a vector space to itself is an algebra. The complex numbers are an algebra over the reals.

The *dimension* of an algebra is its dimension as a vector space over the underlying field.

Let M denote the subset of the algebra of 2×2 matrices over $\mathbb{Q}$ consisting of the matrices of the form

$$\begin{pmatrix} a & 2b \\ b & a \end{pmatrix} = a\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + b\begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}.$$

It is not hard to show that this set is a subspace of $\mathrm{Mat}_{2\times 2}(\mathbb{Q})$ and this it is closed under multiplication. Hence it is a *subalgebra* of $\mathrm{Mat}_{2\times 2}(\mathbb{Q})$, but you can also show that it is commutative and that every non-zero element is invertible. Therefore it is a field, isomorphic to the field usually denoted by $\mathbb{Q}(\sqrt{2})$.

If $\mathbb{A}$ is an algebra of dimension d over $\mathbb{F}$ and $M \in \mathbb{A}$, then the $d+1$ powers $I, M, \dots, M^d$ are linearly dependent, whence there is a polynomial f such that $f(M) = 0$. Consequently there is a monic polynomial ψ of least degree such that $\psi(M) = 0$. It is called the minimal polynomial of M and degree at most d.

Exercise: If $\mathbb{A}$ is a finite-dimensional algebra over $\mathbb{F}$ and $x \in \mathbb{A}$, show that multiplication by x is a linear mapping (over $\mathbb{F}$).

Exercise: If $\mathbb{A}$ is a finite-dimensional algebra over $\mathbb{F}$, prove that $\mathbb{A}$ is isomorphic to an algebra of matrices over $\mathbb{F}$.

Exercise: Suppose K, L, M are fields with $K \le L \le M$. Then L and M are algebras over K; let ℓ and m respectively denote the dimensions of L and M over K. Prove that ℓ divides m.

Exercise: Let $\mathbb{F}$ be a field. If S is a subspace of $\mathrm{Mat}_{d\times d}(\mathbb{F})$ such that each non-zero element is invertible, prove that $\dim(S) \le d$.

Exercise: If $\mathbb{A}$ is a finite-dimensional algebra over a field $\mathbb{F}$ and each non-zero element of $\mathbb{A}$ is invertible, prove that the minimal polynomial of each non-zero element is irreducible over $\mathbb{F}$.

27
Index