2.4. Orthogonal Complements and Projection onto a Subspace

Definition 2.14. If $\mathcal{U}$ is a subspace of a unitary space $\mathcal{V}$, the *orthogonal complement* of $\mathcal{U}$ is $\mathcal{U}^\perp = \{\mathbf{x} \in \mathcal{V} \mid \langle \mathbf{x}, \mathbf{u} \rangle = 0 \text{ for all } \mathbf{u} \in \mathcal{U}\}$.

Here are some basic properties of the orthogonal complement.

Theorem 2.15. *Let $\mathcal{V}$ be a unitary space, and let $\mathcal{U}$ be a subspace of $\mathcal{V}$. Then*

(1) *The orthogonal complement, $\mathcal{U}^\perp$, is a subspace of $\mathcal{V}$.*
(2) $\mathcal{U} \cap \mathcal{U}^\perp = \{\mathbf{0}\}$.
(3) *If $\mathcal{U} \subset \mathcal{W}$, then $\mathcal{W}^\perp \subset \mathcal{U}^\perp$.*
(4) *If $\mathcal{V}$ is finite dimensional, then* $\dim(\mathcal{U}) + \dim(\mathcal{U}^\perp) = \dim(\mathcal{V})$, *and* $\mathcal{V} = \mathcal{U} \oplus \mathcal{U}^\perp$.
(5) *If $\mathcal{V}$ is finite dimensional, then $\mathcal{U}^{\perp\perp} = \mathcal{U}$.*

Proof. We leave (1) and (3) as exercises for the reader. For (2), note that if $\mathbf{x} \in \mathcal{U} \cap \mathcal{U}^\perp$, then $\langle \mathbf{x}, \mathbf{x} \rangle = 0$, and hence $\mathbf{x} = 0$.

Now suppose $\mathcal{V}$ is n-dimensional and $\mathcal{U}$ is k-dimensional. Let $\mathbf{u}_1, \ldots, \mathbf{u}_k$ be a basis for $\mathcal{U}$. Then $\mathbf{x} \in \mathcal{U}^\perp$ if an only if $\langle \mathbf{x}, \mathbf{u}_i \rangle = 0$ for $i = 1, \ldots, k$. Let A be the $n \times k$ matrix with columns $\mathbf{u}_1, \ldots, \mathbf{u}_k$. We know there is positive definite Hermitian matrix P such that $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^* P \mathbf{x}$, so $\langle \mathbf{x}, \mathbf{u}_i \rangle = \mathbf{u}_i{}^* P \mathbf{x}$. Hence, $\mathbf{x} \in \mathcal{U}^\perp$ if and only if $A^* P \mathbf{x} = \mathbf{0}$, so $\mathcal{U}^\perp$ is the null space of $A^* P$. Since A has rank k, and P is nonsingular, the $k \times n$ matrix $A^* P$ also has rank k, and thus the null space of $A^* P$ has dimension $n - k$. This, combined with (2), establishes (4). For (5), first note that $\mathcal{U} \subseteq \mathcal{U}^{\perp\perp}$. Then from (4), we have $\dim(\mathcal{U}^{\perp\perp}) = n - \dim(\mathcal{U}^\perp) = n - (n - k) = k = \dim(\mathcal{U})$, so $\mathcal{U}^{\perp\perp} = \mathcal{U}$. □

For a one-dimensional subspace $\mathcal{U}$ of $\mathbb{C}^n$, that is, a line through the origin, the orthogonal complement $\mathcal{U}^\perp$ is an $(n-1)$-dimensional subspace. An $(n-1)$-dimensional subspace of an n-dimensional space is called a *hyperplane*. If we fix a nonzero vector $\mathbf{a}$ in an n-dimensional space, then the equation $\langle \mathbf{a}, \mathbf{x} \rangle = 0$ defines a hyperplane through the origin. Over the real numbers and using the dot product, this equation takes the form $a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = 0$.

Definition 2.8 gives the orthogonal projection of $\mathbf{y}$ onto the one-dimensional subspace spanned by $\mathbf{x}$. We now look at orthogonal projection onto a general subspace. Let $\mathcal{U}$ be a subspace of an inner product space $\mathcal{V}$, and let $\mathbf{y} \in \mathcal{V}$. There are two approaches to the definition of orthogonal projection: we can seek $\mathbf{p} \in \mathcal{U}$ such that $\mathbf{y} - \mathbf{p} \in \mathcal{U}^\perp$, or we can seek $\mathbf{p} \in \mathcal{U}$ such that $\|\mathbf{y} - \mathbf{p}\|$ is minimized. Later in this section, we will see these two characterizations are equivalent; for now we take the first as the definition. There are several issues to consider: uniqueness, existence, and computing $\mathbf{p}$. We start with the uniqueness question.

Theorem 2.16. *Let $\mathcal{U}$ be a subspace of an inner product space $\mathcal{V}$, and let $\mathbf{y} \in \mathcal{V}$. Then there is at most one vector $\mathbf{p}$ in $\mathcal{U}$ such that $\mathbf{y} - \mathbf{p} \in \mathcal{U}^\perp$.*

Proof. Suppose we have $\mathbf{p}, \mathbf{q} \in \mathcal{U}$ with $\mathbf{y} - \mathbf{p} \in \mathcal{U}^\perp$ and $\mathbf{y} - \mathbf{q} \in \mathcal{U}^\perp$. Since $\mathcal{U}$ and $\mathcal{U}^\perp$ are both subspaces, $\mathbf{q} - \mathbf{p} \in \mathcal{U}$ and $(\mathbf{y} - \mathbf{p}) - (\mathbf{y} - \mathbf{q}) = \mathbf{q} - \mathbf{p} \in \mathcal{U}^\perp$. But $\mathcal{U} \cap \mathcal{U}^\perp = \mathbf{0}$, so $\mathbf{q} - \mathbf{p} = \mathbf{0}$ and hence $\mathbf{q} = \mathbf{p}$. □

Definition 2.17. Let $\mathcal{U}$ be a subspace of an inner product space $\mathcal{V}$, and let $\mathbf{y} \in \mathcal{V}$. If there exists a vector $\mathbf{p} \in \mathcal{U}$ such that $\mathbf{y} - \mathbf{p} \in \mathcal{U}^\perp$, then $\mathbf{p}$ is called the orthogonal projection of $\mathbf{y}$ onto $\mathcal{U}$ and denoted $\mathbf{p} = \text{proj}_{\mathcal{U}}\mathbf{y}$.

Note that $\mathbf{y} = \mathbf{p} + (\mathbf{y} - \mathbf{p})$ decomposes $\mathbf{y}$ into the sum of its projection onto $\mathcal{V}$ and a vector orthogonal to $\mathcal{V}$. We have $\|\mathbf{y}\|^2 = \|\mathbf{p}\|^2 + \|\mathbf{y}-\mathbf{p}\|^2$ and thus $\|\mathbf{p}\| \leq \|\mathbf{y}\|$, with equality if and only if $\mathbf{p} = \mathbf{y}$.

The more subtle question is whether $\text{proj}_{\mathcal{U}}\mathbf{y}$ always exists. For a one-dimensional subspace $\mathcal{U}$, we have seen the answer is yes; if $\mathbf{x}$ spans $\mathcal{U}$, then formula (2.3) tells us how to compute $\text{proj}_{\mathcal{U}}\mathbf{y}$. The next theorem generalizes this result to any finite-dimensional subspace $\mathcal{U}$, giving a formula for $\text{proj}_{\mathcal{U}}\mathbf{y}$ in terms of an orthonormal basis for $\mathcal{U}$.

Theorem 2.18. *Let $\mathcal{U}$ be a nonzero, finite-dimensional subspace of an inner product space $\mathcal{V}$. Let $\mathbf{y} \in \mathcal{V}$, and let $\mathbf{u}_1, \ldots, \mathbf{u}_k$ be an orthogonal basis for $\mathcal{U}$. Then*

$$\text{proj}_{\mathcal{U}}\mathbf{y} = \sum_{i=1}^{k} \text{proj}_{\mathbf{u}_i}\mathbf{y} = \sum_{i=1}^{k} \frac{\langle \mathbf{y}, \mathbf{u}_i \rangle}{\langle \mathbf{u}_i, \mathbf{u}_i \rangle} \mathbf{u}_i. \tag{2.6}$$

Setting $\mathbf{p} = \text{proj}_{\mathcal{U}}\mathbf{y}$, the following holds: for any $\mathbf{x}$ in $\mathcal{U}$ with $\mathbf{x} \neq \mathbf{p}$, we have $\|\mathbf{y} - \mathbf{p}\| < \|\mathbf{y} - \mathbf{x}\|$.

Proof. Set $\mathbf{p} = \sum_{i=1}^{k} \text{proj}_{\mathbf{u}_i}\mathbf{y}$. We need to show that $\mathbf{p} = \text{proj}_{\mathcal{U}}\mathbf{y}$. Since $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is an orthogonal basis, $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ when $i \neq j$. Then for any basis vector $\mathbf{u}_j$,

$$\langle \mathbf{p}, \mathbf{u}_j \rangle = \sum_{i=1}^{k} \frac{\langle \mathbf{y}, \mathbf{u}_i \rangle}{\langle \mathbf{u}_i, \mathbf{u}_i \rangle} \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \langle \mathbf{y}, \mathbf{u}_j \rangle. \tag{2.7}$$

Let $\mathbf{u} \in \mathcal{U}$. Then $\mathbf{u} = \sum_{j=1}^{k} a_j \mathbf{u}_j$ and

$$\langle \mathbf{y}, \mathbf{u} \rangle = \sum_{j=1}^{k} \overline{a}_j \langle \mathbf{y}, \mathbf{u}_j \rangle = \sum_{j=1}^{k} \overline{a}_j \langle \mathbf{p}, \mathbf{u}_j \rangle = \Big\langle \mathbf{p}, \sum_{j=1}^{k} a_j \mathbf{u}_j \Big\rangle = \langle \mathbf{p}, \mathbf{u} \rangle.$$

So $\langle \mathbf{y} - \mathbf{p}, \mathbf{u} \rangle = 0$ for all $\mathbf{u} \in \mathcal{U}$, which means $\mathbf{y} - \mathbf{p} \in \mathcal{U}^\perp$. Hence, $\mathbf{p} = \text{proj}_{\mathcal{U}}\mathbf{y}$.

Now suppose $\mathbf{x} \in \mathcal{U}$. Then $\mathbf{p} - \mathbf{x} \in \mathcal{U}$ and $\mathbf{y} - \mathbf{p} \in \mathcal{U}^\perp$, and so

$$\|(\mathbf{y} - \mathbf{p})\|^2 + \|(\mathbf{p} - \mathbf{x})\|^2 = \|\mathbf{y} - \mathbf{x}\|^2.$$

So $\|(\mathbf{y} - \mathbf{p})\| \leq \|\mathbf{y} - \mathbf{x}\|$ and equality holds only when $\mathbf{x} = \mathbf{p}$. □

When the basis $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is orthonormal, formula (2.6) simplifies to

$$\mathbf{p} = \text{proj}_{\mathcal{U}}\mathbf{y} = \sum_{i=1}^{k} \langle \mathbf{y}, \mathbf{u}_i \rangle \mathbf{u}_i.$$

We now express this in matrix form for the standard inner product on $\mathbb{C}^n$. Let B be the $n \times k$ matrix with the vector $\mathbf{u}_j$ in column j. The reader may check that

$$\mathbf{p} = \text{proj}_{\mathcal{U}}\mathbf{y} = \sum_{i=1}^{k} \langle \mathbf{y}, \mathbf{u}_i \rangle_I \mathbf{u}_i = BB^*\mathbf{y}. \tag{2.8}$$

Earlier we said that an equivalent way to define the orthogonal projection of $\mathbf{y}$ on a subspace $\mathcal{U}$ is to find the point in $\mathcal{U}$ which is closest to $\mathbf{y}$ (provided such a point exists). Theorem 2.18 establishes this equivalence for a finite-dimensional $\mathcal{U}$; we now use this to show the equivalence holds in general.

Theorem 2.19. *Let $\mathcal{U}$ be a subspace of an inner product space $\mathcal{V}$. Let $\mathbf{y} \in \mathcal{V}$. Then for $\mathbf{x} \in \mathcal{U}$, the following are equivalent:*

(1) $\mathbf{y} - \mathbf{x} \in \mathcal{U}^\perp$.

(2) *For any $\mathbf{u} \in \mathcal{U}$ with $\mathbf{u} \neq \mathbf{x}$, we have $\|\mathbf{y} - \mathbf{x}\| < \|\mathbf{y} - \mathbf{u}\|$.*

Proof. Suppose there exists $\mathbf{x} \in \mathcal{U}$ such $\mathbf{y} - \mathbf{x} \in \mathcal{U}^\perp$. Let $\mathbf{u} \in \mathcal{U}$. Then $\mathbf{y} - \mathbf{u} = (\mathbf{y} - \mathbf{x}) + (\mathbf{x} - \mathbf{u})$. Since $\mathbf{y} - \mathbf{x} \in \mathcal{U}^\perp$ and $(\mathbf{x} - \mathbf{u}) \in \mathcal{U}$, the vectors $(\mathbf{y} - \mathbf{x})$ and $(\mathbf{x} - \mathbf{u})$ are orthogonal. Hence,

$$\|\mathbf{y} - \mathbf{u}\|^2 = \|(\mathbf{y} - \mathbf{x})\|^2 + \|(\mathbf{x} - \mathbf{u})\|^2 \geq \|(\mathbf{y} - \mathbf{x})\|^2. \tag{2.9}$$

Equality holds in equation (2.9) if and only if $\mathbf{x} = \mathbf{u}$, so $\|\mathbf{y} - \mathbf{x}\| < \|\mathbf{y} - \mathbf{u}\|$ for any $\mathbf{u} \in \mathcal{U}$ with $\mathbf{u} \neq \mathbf{x}$. (Yes, we did just repeat the argument at the end of the previous proof.)

Conversely, suppose $\mathbf{x} \in \mathcal{U}$ satisfies the condition in statement (2). Let $\mathbf{u} \in \mathcal{U}$, and let $\mathcal{P}$ be the subspace spanned by $\mathbf{x}$ and $\mathbf{u}$. Since $\mathcal{P} \subseteq \mathcal{U}$, we know that $\|\mathbf{y} - \mathbf{x}\| < \|\mathbf{y} - \mathbf{v}\|$ for any $\mathbf{v} \in \mathcal{P}$ with $\mathbf{v} \neq \mathbf{x}$. Since $\mathcal{P}$ is finite dimensional, Theorem 2.18 tells us $\text{proj}_{\mathcal{P}}\mathbf{y} = \mathbf{p}$ must exist, and $\|\mathbf{y} - \mathbf{p}\| < \|\mathbf{y} - \mathbf{v}\|$ for any $\mathbf{v} \in \mathcal{P}$ with $\mathbf{v} \neq \mathbf{p}$. But then we must have $\mathbf{p} = \mathbf{x}$. So $\mathbf{y} - \mathbf{x} \in \mathcal{P}^\perp$. Hence, $(\mathbf{y} - \mathbf{x})$ is orthogonal to $\mathbf{u}$. Since this holds for any $\mathbf{u} \in \mathcal{U}$, we have $\mathbf{y} - \mathbf{x} \in \mathcal{U}^\perp$. □

If $\mathcal{U}$ is infinite dimensional, we are not guaranteed the existence of $\text{proj}_{\mathcal{U}}\mathbf{y}$.

Example 2.20. Let l^2 be the set of all square summable complex sequences, that is, the space of all sequences $\{a_n\}_{n=1}^\infty$ of complex numbers such that $\sum_{i=1}^{\infty} |a_i|^2$ converges. Using the Cauchy–Schwarz inequality, one can show that for sequences $\{a_n\}_{n=1}^\infty$, $\{b_n\}_{n=1}^\infty$ in l^2, the series $\sum_{i=1}^{\infty} a_i\overline{b_i}$ converges. The set l^2 is an inner product space with vector sum and scalar multiplication defined component-wise and the inner product

$$\langle \{a_n\}_{n=1}^\infty, \{b_n\}_{n=1}^\infty \rangle = \sum_{i=1}^{\infty} a_i\overline{b_i}.$$

Let $\mathcal{U}$ be the subspace of all sequences which have only a finite number of nonzero entries. The subspace $\mathcal{U}$ contains all of the unit coordinate sequences $\mathbf{e}_j$, where $\mathbf{e}_j$ denotes the sequence with a one in position j and zeroes elsewhere. Then $\mathcal{U}^\perp = \{\mathbf{0}\}$. Let $\mathbf{y} = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots)$ be the sequence with $\frac{1}{n}$ in position n. Then $\mathbf{y} \in l^2$, but for any $\mathbf{x} \in \mathcal{U}$ the vector $\mathbf{y} - \mathbf{x}$ has nonzero entries, so it cannot be in $\mathcal{U}^\perp$. Hence, $\mathcal{U}$ contains no vector $\mathbf{x}$ such that $\mathbf{y} - \mathbf{x}$ is in $\mathcal{U}^\perp$, and so $\text{proj}_{\mathcal{U}}\mathbf{y}$ does not exist.

A *Hilbert space* is an inner product space which is a complete metric space with respect to the metric induced by the inner product (that is, the distance from $\mathbf{x}$ to $\mathbf{y}$ is $\|\mathbf{x} - \mathbf{y}\| = \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle^{\frac{1}{2}}$). For any n, the space $\mathbb{C}^n$ is a Hilbert space. The space l^2 is an infinite-dimensional Hilbert space. In a Hilbert space one has

the following "closest point property" for closed, convex subsets. See [**Youn88**, pp. 26–28] for a proof.

Theorem 2.21. *Let S be a nonempty, closed, convex set in a Hilbert space $\mathcal{H}$. For any $\mathbf{x} \in \mathcal{H}$, there is a unique point $\mathbf{y} \in S$ such that $\|\mathbf{x}-\mathbf{y}\| < \|\mathbf{x}-\mathbf{a}\|$ for all $\mathbf{a} \neq \mathbf{y}$ in $\mathcal{S}$.*

The problem with the subspace $\mathcal{U}$ in Example 2.20 is that it is not closed. The vector $\mathbf{y} = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots)$ is not in $\mathcal{U}$ but is the limit of a sequence of points in $\mathcal{U}$. Although $\mathbf{y}$ is not in $\mathcal{U}$, it is in the closure of $\mathcal{U}$. However, any subspace is a convex set, so when $\mathcal{W}$ is a *closed* subspace of a Hilbert space $\mathcal{H}$, the closest point property holds for $\mathcal{W}$, and thus for any $\mathbf{x} \in \mathcal{H}$, the orthogonal projection $\text{proj}_{\mathcal{W}}\mathbf{x}$ will exist.

Returning to the finite-dimensional world, formula (2.6) gives $\text{proj}_{\mathcal{U}}\mathbf{y}$ if we have an orthogonal basis for $\mathcal{U}$. For the standard inner product on $\mathbb{C}^n$, we now obtain a formula for computing $\text{proj}_{\mathcal{U}}\mathbf{y}$ from an arbitrary basis for $\mathcal{U}$. First we need the following fact about matrices.

Theorem 2.22. *Let A be an $n \times k$ matrix. Then the four matrices A, A^*, AA^*, and A^*A all have the same rank.*

Proof. We first show that A and A^*A have the same null space. We clearly have $\ker(A) \subseteq \ker(A^*A)$. Now suppose $A^*A\mathbf{x} = \mathbf{0}$. Then $\mathbf{x}^*A^*A\mathbf{x} = 0$. But $\mathbf{x}^*A^*A\mathbf{x} = (A\mathbf{x})^*(A\mathbf{x}) = \|A\mathbf{x}\|^2$. So $\|A\mathbf{x}\| = 0$ and hence $A\mathbf{x} = \mathbf{0}$. So we have $\ker(A^*A) \subseteq \ker(A)$. Hence, $\ker(A) = \ker(A^*A)$. Since A and A^*A also have the same number of columns, k, the rank plus nullity theorem then tells us they must have the same rank. Now, since the row and column space of a matrix have the same dimension, we know A^* has the same rank as A. Using A^* in the first argument then tells us that A^* and $A^{**}A^* = AA^*$ have the same rank. □

In particular, when $k \leq n$ and the $n \times k$ matrix A has linearly independent columns, the $k \times k$ matrix A^*A has rank k and thus is invertible.

Let $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ be a basis for a k-dimensional subspace $\mathcal{U}$ of an n-dimensional inner product space $\mathcal{V}$. Let A be the $n \times k$ matrix with $\mathbf{a}_j$ in column j. The columns of A are linearly independent, and the $k \times k$ matrix A^*A is invertible. For $\mathbf{y} \in \mathcal{V}$, set $\mathbf{p} = \text{proj}_{\mathcal{U}}\mathbf{y}$. Since $\mathbf{p}$ is a linear combination of the columns of A, we have $\mathbf{p} = A\mathbf{x}$ for some $\mathbf{x}$ in $\mathbb{C}^k$. In the proof of Theorem 2.18 we saw that $\langle \mathbf{p}, \mathbf{u} \rangle = \langle \mathbf{y}, \mathbf{u} \rangle$ for any $\mathbf{u} \in \mathcal{U}$. Since the columns of A are vectors from $\mathcal{U}$, and entry j of $A^*\mathbf{y}$ is $\langle \mathbf{y}, \mathbf{a}_j \rangle_I$, we have $A^*\mathbf{y} = A^*\mathbf{p} = A^*A\mathbf{x}$. Hence $\mathbf{x} = (A^*A)^{-1}A^*\mathbf{y}$ and

$$\mathbf{p} = \text{proj}_{\mathcal{U}}\mathbf{y} = A\mathbf{x} = A(A^*A)^{-1}A^*\mathbf{y}. \tag{2.10}$$

For $k = 1$, equation (2.10) reduces to (2.3). If the columns of A are orthogonal, then A^*A is a diagonal matrix, and equation (2.10) reduces to (2.6). If the columns of A are orthonormal, then $A^*A = I_k$, and (2.10) reduces to (2.8).

Formula (2.10) typically appears as the solution to finding the best least squares fit. Consider a system of linear equations, $A\mathbf{x} = \mathbf{b}$, where A is $n \times k$. This system has a solution if and only if $\mathbf{b}$ is in the column space of A. When the system has no solution, we want to find the vector $\mathbf{w}$ in the column space of A which is as close as possible to $\mathbf{b}$. Thus, letting $\mathcal{U}$ be the subspace spanned by the columns of A, we are looking for $\mathbf{w} = \text{proj}_{\mathcal{U}}\mathbf{b}$. Equivalently, we want the vector $\mathbf{w} = A\mathbf{x}$

which minimizes $\|\mathbf{Ax}-\mathbf{b}\|$, or we want to minimize $\|\mathbf{Ax}-\mathbf{b}\|^2$, which is the sum of the squares of the differences between the coordinates of $A\mathbf{x}$ and $\mathbf{b}$; hence the term "best least squares fit". If the columns of A are linearly independent, then $\mathbf{x}=(A^*A)^{-1}A^*\mathbf{b}$ and $\mathbf{w}=A\mathbf{x}=A(A^*A)^{-1}A^*\mathbf{b}$.

We conclude this section with a fact needed in later chapters, both in the proof of the Jordan canonical form and in the proof of the spectral theorem for normal matrices.

Theorem 2.23. *Let A be an $n\times n$ complex matrix. Then a subspace $\mathcal{U}$ of $\mathbb{C}^n$ is invariant under A if and only if $\mathcal{U}^\perp$ is invariant under A^*.*

Proof. First we note that we are working here with the standard inner product in $\mathbb{C}^n$ (although one can formulate this theorem using the more general definition of adjoint). Suppose $\mathcal{U}$ is an A-invariant subspace. Let $\mathbf{y}\in\mathcal{U}^\perp$, and let $\mathbf{u}$ be any vector in $\mathcal{U}$. Then $A\mathbf{u}\in\mathcal{U}$ and so $\langle\mathbf{u},A^*\mathbf{y}\rangle=\langle A\mathbf{u},\mathbf{y}\rangle=0$, which shows that $A^*\mathbf{y}\in\mathcal{U}^\perp$; hence $\mathcal{U}^\perp$ is invariant under A^*. Conversely, if $\mathcal{U}^\perp$ is invariant under A^*, then $\mathcal{U}^{\perp\perp}$ is invariant under A^{**}. But $\mathcal{U}^{\perp\perp}=\mathcal{U}$ and $A^{**}=A$, so $\mathcal{U}$ is A-invariant. □

2.5. Hilbert Spaces and Fourier Series

The Hilbert space l^2 appeared in Example 2.20. Recall that a Hilbert space is a unitary space which is a complete metric space under the distance induced by the inner product. (A metric space is *complete* if every Cauchy sequence of points in the space converges to a point in the space.) We say an infinite sequence of vectors $\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\dots$ converges to $\mathbf{x}$ if $\|\mathbf{x}-\mathbf{x}_n\|\to 0$ as $n\to\infty$.

Suppose $\mathcal{H}$ is an infinite-dimensional Hilbert space and $\{\mathbf{e}_n\}=\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3,\dots$ is an infinite sequence of orthonormal vectors in $\mathcal{H}$. For each positive integer k, let $\mathcal{U}_k$ denote the k-dimensional subspace spanned by $\mathbf{e}_1,\dots,\mathbf{e}_k$. For each k and any $\mathbf{x}$ in $\mathcal{H}$, we put

$$\mathbf{p}_k=\mathrm{proj}_{\mathcal{U}_k}\mathbf{x}=\sum_{i=1}^{k}\langle\mathbf{x},\mathbf{e}_i\rangle\mathbf{e}_i.$$

Since $\|\mathbf{p}_k\|^2\le\|\mathbf{x}\|^2$, we have

$$\sum_{i=1}^{k}\|\langle\mathbf{x},\mathbf{e}_i\rangle\|^2\le\|\mathbf{x}\|^2 \tag{2.11}$$

for any positive integer k. The inequality (2.11) says that the partial sums of the infinite series $\sum_{i=1}^{\infty}\|\langle\mathbf{x},\mathbf{e}_i\rangle\|^2$ are bounded above by $\|\mathbf{x}\|^2$. Hence, the series converges and

$$\sum_{i=1}^{\infty}\|\langle\mathbf{x},\mathbf{e}_i\rangle\|^2\le\|\mathbf{x}\|^2; \tag{2.12}$$

The Jordan and Weyr Canonical Forms

Similarity is an equivalence relation on the set of $n \times n$ matrices over the field $\mathbb{F}$, and thus partitions $M_n(\mathbb{F})$ into equivalence classes called *similarity classes*. By a *canonical form* for matrices under similarity, we mean a rule for selecting exactly one representative from each equivalence class. This is the canonical form for the matrices in that class, and two matrices are similar if and only if they have the same canonical form. We might want various things from this canonical form. We might want it to be as "simple" as possible, or to display important information about the matrix (such as rank, eigenvalues, etc.), or to be useful for some specific computations. For example, consider computing the exponential of a matrix A, defined via the Taylor series for the exponential function

$$\exp(A) = \sum_{n=0}^{\infty} \frac{A^k}{k!}.$$

Direct calculation of A^k is impractical for large k, and then there is the issue of summing the terms. However, we have

$$\exp(S^{-1}AS) = \sum_{n=0}^{\infty} \frac{(S^{-1}AS)^k}{k!} = \sum_{n=0}^{\infty} \frac{S^{-1}A^kS}{k!} = S^{-1}(\exp A)S,$$

so if we can find S such that $(S^{-1}AS)^k$ is easy to compute, we can use this to compute $\exp A$. When $S^{-1}AS = D$ is diagonal, it is easy to compute D^k and show

$$\exp \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix} = \begin{pmatrix} e^{d_1} & 0 & \cdots & 0 \\ 0 & e^{d_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{d_n} \end{pmatrix}. \tag{4.1}$$

We then find $\exp A$ from $\exp A = \exp(SDS^{-1}) = S(\exp D)S^{-1}$. If A is not diagonable, we would like a canonical form which is close to being diagonal and for which calculations like the one above are easily done.

The Jordan canonical form is probably the best known of various canonical forms for matrices under similarity and is valid whenever the field $\mathbb{F}$ contains the eigenvalues of the matrix. In this form, the eigenvalues of the matrix appear on the main diagonal, the entries on the super-diagonal (i.e., the diagonal line directly above the main diagonal) are either zeroes or ones, and all other entries of the matrix are zero. Thus, it is close to being diagonal. Furthermore, it has a block diagonal structure which dictates the positions of the super-diagonal ones and reflects important information about the matrix. Here is an example of a matrix in Jordan form.

$$\begin{pmatrix} 7 & 1 & 0 & & & & & & & \\ 0 & 7 & 1 & & & & & & & \\ 0 & 0 & 7 & & & & & & & \\ & & & 7 & 1 & & & & & \\ & & & 0 & 7 & & & & & \\ & & & & & 8 & 1 & 0 & 0 & \\ & & & & & 0 & 8 & 1 & 0 & \\ & & & & & 0 & 0 & 8 & 1 & \\ & & & & & 0 & 0 & 0 & 8 & \\ & & & & & & & & & 10 \\ & & & & & & & & & & -4 \end{pmatrix} \tag{4.2}$$

This example has two diagonal blocks (of sizes 3×3 and 2×2) belonging to eigenvalue 7, one 4×4 block belonging to eigenvalue 8, and 1×1 blocks for the eigenvalues 10 and -4.

There are various approaches to the Jordan canonical form. In abstract algebra, it comes as an application of the classification theorem for finitely generated modules over principal ideal domains; see [**Rot88, Art11**]. Another classical approach is via the equivalence theory of matrices over polynomial rings, using elementary divisors and invariant factors, as done, for example, in [**Gan59, Mal63, HK71**]. There are several ways to obtain the Jordan form using only techniques from linear algebra, as is done in [**Hal87, HJ85**] and other texts. For a more combinatorial approach, see Brualdi's article [**Bru87**]. Our approach here is based on three key steps: a theorem of Sylvester, the Schur triangularization theorem, and a final step presented by Halmos in his 1993 lecture in Pensacola, Florida. (In that lecture, Halmos said he had never been quite satisfied with the proof in his book [**Hal87**].) Except for the last part, much of the development here is similar to that in Horn and Johnson [**HJ85**].

We begin with a theorem of Sylvester, of significant interest in its own right. This theorem enables us to use matrix methods to reach a block diagonal form, where each diagonal block corresponds to an eigenvalue of the transformation. The more standard way to achieve this is by decomposing the vector space into a direct sum of generalized eigenspaces. In a sense, we are using the Sylvester theorem to bypass that algebraic argument. Of course, one could argue that the decomposition into generalized eigenspaces is essential to understanding the structure of the linear

transformation, and thus should be approached directly; the reader may find that method elsewhere, for example in [**HK71**].

4.1. A Theorem of Sylvester and Reduction to Block Diagonal Form

Let A be an $m \times m$ matrix, and let B be an $n \times n$ matrix. Let C be $m \times n$. Consider the matrix equation

$$AX - XB = C, \tag{4.3}$$

where the unknown matrix X is $m \times n$. Considered entry by entry, this is a system of mn linear equations in the mn unknowns x_{ij}. Sylvester's theorem is about whether, for a given pair of matrices A, B, equation (4.3) has a solution for any right-hand side C. We state two equivalent versions; the equivalence of these two forms of the theorem follows from basic results about systems of linear equations (specifically, the fact that for a square matrix M, the system $M\mathbf{x} = \mathbf{b}$ is consistent for every vector $\mathbf{b}$ if and only if the homogeneous system $M\mathbf{x} = \mathbf{0}$ has no nontrivial solutions.)

Theorem 4.1 (Sylvester [**Syl84**])**.** *Let A be $m \times m$, and let B be $n \times n$. The matrix equation $AX - XB = 0$ has nontrivial solutions if and only if* $\operatorname{spec}(A) \cap \operatorname{spec}(B)$ *is nonempty.*

Theorem 4.2. *Let A be $m \times m$, and let B be $n \times n$. Then the matrix equation $AX - XB = C$ is consistent for every choice of $m \times n$ matrix C if and only if* $\operatorname{spec}(A) \cap \operatorname{spec}(B)$ *is empty.*

We will obtain Sylvester's theorem as a consequence of Theorem 4.3 below. First, note that if $\mathcal{V} = \mathcal{M}(m, n)$ is the mn-dimensional space of $m \times n$ matrices, we can define a linear transformation on $\mathcal{V}$ by

$$T_{A,B}(X) = AX - XB \tag{4.4}$$

for any matrix $X \in \mathcal{M}(m, n)$. Suppose $\alpha \in \operatorname{spec}(A)$ and $\beta \in \operatorname{spec}(B)$. Let $\mathbf{y}$ be an eigenvector of A associated with α; thus, $A\mathbf{y} = \alpha\mathbf{y}$. Note that $\mathbf{y}$ is a nonzero column vector with m coordinates. Let $\mathbf{z}^T$ be a *left* eigenvector of B associated with β; i.e., $\mathbf{z}^T$ is a nonzero row vector with n coordinates such that $\mathbf{z}^T B = \beta\mathbf{z}^T$. Put $X = \mathbf{y}\mathbf{z}^T$; note that X is $m \times n$. Then

$$T_{A,B}(X) = A\mathbf{y}\mathbf{z}^T - \mathbf{y}\mathbf{z}^T B = \alpha\mathbf{y}\mathbf{z}^T - \beta\mathbf{y}\mathbf{z}^T = (\alpha - \beta)X. \tag{4.5}$$

So the $m \times n$ nonzero matrix $X = \mathbf{y}\mathbf{z}^T$ is an eigenvector of $T_{A,B}$ with associated eigenvalue $\alpha - \beta$. We claim that all of the eigenvalues of $T_{A,B}$ are obtained in this way.

Theorem 4.3. *Let A be $m \times m$, and let B be $n \times n$ with* $\operatorname{eigen}(A) = \alpha_1, \ldots, \alpha_m$ *and* $\operatorname{eigen}(B) = \beta_1, \ldots, \beta_n$*. Then the eigenvalues of the map $T_{A,B}$, defined by $T_{A,B}(X) = AX - XB$ for $X \in \mathcal{M}(m, n)$, are the mn numbers $\alpha_i - \beta_j$, where $i = 1, \ldots, m$ and $j = 1, \ldots, n$ and repeated values are to be listed according to multiplicities.*

Remark 4.4. As an example of how to treat repeated eigenvalues, suppose A is 2×2 with $\operatorname{eigen}(A) = 5, 6$ and B is 3×3 with $\operatorname{eigen}(B) = 2, 2, 3$. Then we would have $\operatorname{eigen}(T_{A,B}) = 5-2, 5-2, 5-3, 6-2, 6-2, 6-3$, or $3, 3, 2, 4, 4, 3$.

Proof. First, assume the m eigenvalues of A are distinct and the n eigenvalues of B are distinct. Let $\mathbf{y}_1, \ldots, \mathbf{y}_m$ be a basis of eigenvectors for A and $\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T$ be a basis of left eigenvectors for B. So $A\mathbf{y}_i = \alpha_i \mathbf{y}_i$ for $i = 1, \ldots, m$ and $\mathbf{z}_j^T B = \beta_j \mathbf{z}_j^T$ for $j = 1, \ldots, n$. From equation (4.5), we have $T_{A,B}(\mathbf{y}_i \mathbf{z}_j^T) = (\alpha_i - \beta_j)\mathbf{y}_i \mathbf{z}_j^T$. Now, the mn matrices $\mathbf{y}_i \mathbf{z}_j^T$ correspond to the mn vectors $\mathbf{y}_i \otimes \mathbf{z}_j$. Since $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ and $\{\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T\}$ are each linearly independent sets, the set

$$\{\mathbf{y}_i \otimes \mathbf{z}_j \mid i = 1, \ldots, m;\ j = 1, \ldots, n\}$$

is linearly independent, and hence forms a basis of eigenvectors for the map $T_{A,B}$.

We can now use a continuity argument to deal with repeated eigenvalues. Let $\{A_p\}_{p=1}^{\infty}$ be a sequence of matrices with distinct eigenvalues which converges to A, and let $\{B_p\}_{p=1}^{\infty}$ be a sequence of matrices with distinct eigenvalues which converges to B. Then $\{T_{A_p,B_p}\}_{p=1}^{\infty}$ converges to $T_{A,B}$, and the result follows from the fact that the eigenvalues of a matrix are continuous functions of the matrix entries. □

As a consequence of Theorem 4.3, the map $T_{A,B}$ is nonsingular if and only if spec(A) ∩ spec(B) is empty, giving Theorem 4.2.

Those who prefer to deal with the case of repeated eigenvalues with an algebraic argument may prefer the following approach. We can rearrange the entries of X into a column vector of length mn by stacking the columns of X, starting with the first column on top, then placing column two below it, and so on. We denote this column vector formed from X as X^{stack}. For example if

$$X = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix},$$

then

$$X^{stack} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{pmatrix}.$$

For a $p \times q$ matrix R and an $m \times n$ matrix S, we have the tensor product

$$R \otimes S = \begin{pmatrix} Rs_{11} & Rs_{12} & \cdots & Rs_{1n} \\ Rs_{21} & Rs_{22} & \cdots & Rs_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Rs_{m1} & Rs_{m2} & \cdots & Rs_{mn} \end{pmatrix}.$$

We then have $(AX)^{stack} = (A \otimes I_n)X^{stack}$ and $(XB)^{stack} = (I_m \otimes B^T)X^{stack}$. We leave the verification of these formulas to to the reader—reluctance to write out the details is one reason we did the proof with the continuity argument. The first equation, $(AX)^{stack} = (A \otimes I_n)X^{stack}$ is easy to check if you simply work with the columns of X. The second one is a bit messier to deal with, but persevere and keep your rows and columns straight, and it will work out.

Combining these equations gives

$$(AX - XB)^{stack} = (A \otimes I_n - I_m \otimes B^T)X^{stack}. \tag{4.6}$$

This shows that the matrix for the linear transformation $T_{A,B}$, considered as acting on the column vectors, X^{stack}, is $A \otimes I_n - I_m \otimes B^T$. By Schur's triangularization theorem, there are nonsingular matrices, P and Q, of sizes $m \times m$ and $n \times n$, respectively, such that $P^{-1}AP = \text{triang}(\alpha_1, \ldots, \alpha_m)$ and $Q^{-1}B^TQ = \text{triang}(\beta_1, \ldots, \beta_n)$ are upper triangular. Then,

$$(P \otimes Q)^{-1}(A \otimes I_n - I_m \otimes B^T)(P \otimes Q) = P^{-1}AP \otimes I_n - I_m \otimes Q^{-1}B^TQ.$$

Now $P^{-1}AP \otimes I_n$ is block diagonal; each of the n diagonal blocks is a copy of the triangular matrix $P^{-1}AP$. From the tensor product formula, and the fact that $Q^{-1}B^TQ$ is triangular, we see $I_m \otimes Q^{-1}B^TQ$ is triangular, and the main diagonal consists of the scalar matrices $\beta_1 I_m, \beta_2 I_m, \ldots, \beta_n I_m$ in that order. So $P^{-1}AP \otimes I_n - I_m \otimes Q^{-1}B^TQ$ is upper triangular and the diagonal entries are exactly the mn numbers $\alpha_i - \beta_j$, where $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

We need the following corollary of Theorem 4.2.

Corollary 4.5. *Let A be $k \times k$, let B be $r \times r$, and assume* $\text{spec}(A) \cap \text{spec}(B) = \emptyset$. *Then for any $k \times r$ matrix C, the block triangular matrix $M = \begin{pmatrix} A & C \\ 0 & B \end{pmatrix}$ is similar to the block diagonal matrix $A \oplus B = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$.*

Proof. Since $\text{spec}(A) \cap \text{spec}(B) = \emptyset$, Sylvester's theorem guarantees the existence of a $k \times r$ matrix X such that $AX - XB = -C$. Put $S = \begin{pmatrix} I_k & X \\ 0 & I_r \end{pmatrix}$. Then $S^{-1} = \begin{pmatrix} I_k & -X \\ 0 & I_r \end{pmatrix}$ and

$$\begin{aligned} S^{-1}MS &= \begin{pmatrix} I_k & -X \\ 0 & I_r \end{pmatrix} \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} \begin{pmatrix} I_k & X \\ 0 & I_r \end{pmatrix} \\ &= \begin{pmatrix} A & C - XB \\ 0 & B \end{pmatrix} \begin{pmatrix} I_k & X \\ 0 & I_r \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}. \end{aligned}$$

□

One can extend Corollary 4.5 to apply to any number of blocks by using an induction argument. Thus, if

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1t} \\ 0 & A_{22} & \cdots & A_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{tt} \end{pmatrix},$$

and $\text{spec}(A_{ii}) \cap \text{spec}(A_{jj}) = \emptyset$ whenever $i \neq j$, then A is similar to the block diagonal matrix $A_{11} \oplus A_{22} \oplus \cdots \oplus A_{tt}$.

Now consider an $n \times n$ matrix A over an algebraically closed field $\mathbb{F}$, with $\text{spec}(A) = \{\lambda_1, \ldots, \lambda_t\}$, where eigenvalue λ_i has multiplicity m_i. From Theorem 3.13, we know A is similar to a triangular matrix in which the eigenvalues

appear in the order $\lambda_1, \ldots, \lambda_1, \lambda_2, \ldots, \lambda_2, \ldots, \lambda_t, \ldots, \lambda_t$. We then have

$$S^{-1}AS = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1t} \\ 0 & A_{22} & \cdots & A_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{tt} \end{pmatrix}, \tag{4.7}$$

where each block A_{ij} is $m_i \times m_j$, and the diagonal block A_{ii} is triangular of size m_i with λ_i along the main diagonal. Using Corollary 4.5 and abbreviating A_{ii} as A_i, we see that A is similar to the block diagonal matrix $A_1 \oplus A_2 \oplus \cdots \oplus A_t$. The next step is to find a canonical form for a typical block A_i, i.e., for a matrix which has a single eigenvalue. This takes some work. Note that if there are no repeated eigenvalues, then $t = n$ and the block diagonal matrix $A_1 \oplus A_2 \oplus \cdots \oplus A_t$ is just an ordinary diagonal matrix. All of the fuss in the next two sections is to deal with repeated eigenvalues.

Sylvester's theorem is also useful for simultaneously reducing a commuting pair of matrices to triangular or block form.

Theorem 4.6. *Suppose A, B are a pair of $n \times n$ matrices which commute and*

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1t} \\ 0 & A_{22} & \cdots & A_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{tt} \end{pmatrix}, \tag{4.8}$$

where for $i \neq j$, we have $\operatorname{spec}(A_i) \cap \operatorname{spec}(A_j) = \emptyset$. Then B must have the form

$$B = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1t} \\ 0 & B_{22} & \cdots & B_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_{tt} \end{pmatrix},$$

where A_{ii} and B_{ii} have the same size for each $i = 1, \ldots, t$.

Proof. We do the proof for the case $t = 2$, and leave it as an exercise to the reader to do the induction argument to extend to the general case. So, suppose we have $A = \begin{pmatrix} A_1 & A_{12} \\ 0 & A_2 \end{pmatrix}$. Partition B conformally with A, and write $B = \begin{pmatrix} X & Y \\ Z & W \end{pmatrix}$. Then the $(2,1)$ block of AB is A_2Z and the $(2,1)$ block of BA is ZA_1. Since $AB = BA$, we must have $A_2Z = ZA_1$. Since $\operatorname{spec}(A_1) \cap \operatorname{spec}(A_2) = \emptyset$, Sylvester's theorem tells us $Z = 0$. □

Consequently, if $AB = BA$, and S is a similarity such that $S^{-1}AS$ has the form (4.8), with $\operatorname{spec}(A_i) \cap \operatorname{spec}(A_j) = \emptyset$ when $i \neq j$, then $S^{-1}BS$ must also be in a conformal block triangular form. When the diagonal blocks, A_{ii}, are all scalar blocks, we can apply a further block diagonal similarity $R = R_1 \oplus R_2 \oplus \cdots \oplus R_t$ to both A and B so that each diagonal block of B is put in triangular form; observe that R will preserve the scalar diagonal blocks of A.

4.5. Weyr Normal Form

The Segre characteristic lists the sizes of the diagonal blocks in the Jordan normal form of a matrix. We showed these block sizes are uniquely determined by the matrix by using the Weyr characteristic, which comes from the dimensions of null spaces of powers of a nilpotent matrix. For a nilpotent matrix, the Segre characteristic is the conjugate partition of the Weyr characteristic. An alternative normal form, due to Weyr, has block sizes given directly by the Weyr characteristic. One could well argue that the Weyr form is more natural, as it comes directly from the dimensions of the null spaces of the powers.

Let A be an $n \times n$ nilpotent matrix of index p, and let $\omega_1, \ldots, \omega_p$ be the Weyr characteristic of A. So $\omega_1(A) = \nu(A)$ and $\omega_k(A) = \nu(A^k) - \nu(A^{k-1})$ for $k \geq 2$. Note that

$$\sum_{i=1}^{p} \omega_i = \sum_{i=1}^{p} (\nu(A^i) - \nu(A^{i-1})) = \nu(A^p) = n.$$

We will show below that $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_p$. For now, we assume this is true, so as to describe the Weyr normal form of A. For $r \geq s$, let $I_{r,s}$ denote the $r \times s$ matrix which has I_s in the first s rows, and zeroes in the remaining $r - s$ rows. Let W be the $n \times n$ matrix, partitioned into blocks of sizes $\omega_i \times \omega_j$, in which all blocks are zero except the blocks directly above the diagonal blocks, and the $\omega_k \times \omega_{k+1}$ block directly above the $(k+1)$-st diagonal block is the matrix $I_{\omega_k, \omega_{k+1}}$. Note the diagonal blocks are zero blocks of sizes $\omega_i \times \omega_i$. Direct calculation shows $\nu(W) = \omega_1$, $\nu(W^2) = \omega_1 + \omega_2$, and in general, $\nu(W^k) = \omega_1 + \omega_2 + \omega_3 + \cdots + \omega_k$, for $k = 1, \ldots, p$. So the matrix W has the same Weyr characteristic as A, and hence the nilpotent matrices W and A must have the same Segre characteristic and thus the same Jordan form. Therefore, A and W are similar. We define W to be the Weyr normal form of the nilpotent matrix A. Note that the numbers of the Weyr characteristic give the sizes of the diagonal blocks of W (which are blocks of zeros).

Example 4.29. Suppose $N = J_n(0)$. Then for $k = 1, \ldots, n$, we have $\nu(N^k) = k$ and so $\omega_k = 1$ for each k. So $J_n(0)$ is already in Weyr canonical form.

Example 4.30. Suppose $N = J_3(0) \oplus J_2(0)$. We have $\nu(N) = 2$, $\nu(N^2) = 4$, and $\nu(N^3) = 5$. The Weyr characteristic is $\omega_1 = 2$, $\omega_1 = 2$, and $\omega_3 = 1$. The Weyr canonical form of N is

$$\left(\begin{array}{cc|cc|c} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

Example 4.31. Suppose $N = J_2(0) \oplus J_2(0) \oplus J_1(0)$. Then $N^2 = 0$; we have $\nu(N) = 3$ and $\nu(N^2) = 5$. The Weyr characteristic is $\omega_1 = 3$ and $\omega_2 = 2$. The

Weyr canonical form is

$$\left(\begin{array}{ccc|cc} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

One can then obtain a Weyr normal form for a general matrix A, with eigenvalues $\lambda_1, \ldots, \lambda_t$ of multiplicities $m_1, \ldots, m_t$. As we did for the Jordan form, A is similar to $A_1 \oplus A_2 \oplus \cdots \oplus A_t$, where A_i is $m_i \times m_i$, and $\text{spec}(A_i) = \{\lambda_i\}$. So $N_i = A_i - \lambda_i I$ is nilpotent and has a Weyr normal form, W_i, as described above. The Weyr normal form of A_i is then $\lambda_i I + W_i$, and the Weyr form of A is the direct sum of the blocks $\lambda_i I + W_i$. In the Weyr form, the diagonal blocks are scalar matrices, and the ones appear in the super-diagonal blocks.

It remains to show that $\omega_k \geq \omega_{k+1}$ for $1 \leq k \leq p-1$. Let $T : \mathcal{V} \to \mathcal{V}$ be a nilpotent linear transformation of index p on an n-dimensional vector space $\mathcal{V}$. For any positive integer k, the null space of T^k is contained in the null space of T^{k+1}. Moreover, for $k = 1, \ldots, p$, these null spaces form a strictly increasing sequence. That is, using "$\subset$" to mean "strict subset of", we have

$$\ker(T) \subset \ker(T^2) \subset \ker(T^3) \subset \cdots \subset \ker(T^{p-1}) \subset \ker(T^p) = \mathcal{V}. \tag{4.16}$$

This follows from the following fact.

Lemma 4.32. *Let T be a nilpotent linear transformation on a vector space $\mathcal{V}$. If for some positive integer k, we have $\ker(T^k) = \ker(T^{k+1})$, then $\ker(T^k) = \ker(T^{k+r})$ for every positive integer r.*

Proof. Suppose $\ker(T^k) = \ker(T^{k+1})$. Let $\mathbf{v} \in \ker(T^{k+2})$. Then $T^{k+2}\mathbf{v} = 0$. But $T^{k+2}\mathbf{v} = T^{k+1}(T\mathbf{v})$, so $T\mathbf{v} \in \ker(T^{k+1})$. Hence $T\mathbf{v} \in \ker(T^k)$ and so $T^{k+1}\mathbf{v} = 0$. So $\ker(T^{k+2}) \subseteq \ker(T^{k+1})$. Since $\ker(T^{k+1}) \subseteq \ker(T^{k+2})$, we have $\ker(T^{k+2}) = \ker(T^{k+1})$. Repeating the argument r times (or, more formally, doing an induction argument) gives the result. □

Getting back to the Weyr characteristic, $\omega_1, \ldots, \omega_p$, we have

$$\dim(\ker(T^k)) = \dim(\ker(T^{k-1})) + \omega_k,$$

for $k = 1, \ldots, p$ We want to show $\omega_k \geq \omega_{k+1}$ for each $k \leq p-1$. Before doing the proof for general k, we illustrate the key idea of the argument with the case $k = 1$. Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_{\omega_1}\}$ be a basis for $\ker(T)$. Since the dimension of $\ker(T^2)$ is $\omega_1 + \omega_2$, and $\ker(T) \subset \ker(T^2)$, we can choose ω_2 linearly independent vectors $\mathbf{y}_1, \ldots, \mathbf{y}_{\omega_2}$ in $\ker(T^2)$ such that

$$\mathcal{B} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{\omega_1}\} \cup \{\mathbf{y}_1, \ldots, \mathbf{y}_{\omega_2}\}$$

is a basis for $\ker(T^2)$. Note the vectors $\mathbf{y}_1, \ldots, \mathbf{y}_{\omega_2}$ are not in $\ker(T)$. Hence, the ω_2 vectors $T(\mathbf{y}_1), \ldots, T(\mathbf{y}_{\omega_2})$ are ω_2 nonzero vectors in $\ker(T)$; we now show they are linearly independent. Since ω_1 is the dimension of $\ker(T)$, this will prove $\omega_1 \geq \omega_2$. Suppose then that $\sum_{i=1}^{\omega_2} c_i T(\mathbf{y}_i) = 0$. Then $\sum_{i=1}^{\omega_2} c_i \mathbf{y}_i$ is in $\ker(T)$, and so

$\sum_{i=1}^{\omega_2} c_i \mathbf{y}_i = \sum_{j=1}^{\omega_1} b_j \mathbf{x}_j$. But then, since $\mathcal{B}$ is linearly independent, all of the coefficients must be zero. Hence, the vectors $T(\mathbf{y}_1), \ldots, T(\mathbf{y}_{\omega_2})$ are linearly independent.

For $k > 1$, we use a similar argument, but work in the quotient spaces $\ker(T^k)/\ker(T^{k-1})$ and $\ker(T^{k+1})/\ker(T^k)$. For a proof which does not use quotient spaces, see [**Sha99**].

For the remainder of this section, we use $\mathcal{U}_k$ to denote $\ker(T^k)$. Rewriting (4.16) in this notation,

$$\mathcal{U}_1 \subset \mathcal{U}_2 \subset \mathcal{U}_3 \subset \cdots \subset \mathcal{U}_{p-1} \subset \mathcal{U}_p = \mathcal{V}. \tag{4.17}$$

The next result and its proof set the foundation for the rest of this section.

Theorem 4.33. *Let $T : \mathcal{V} \to \mathcal{V}$ be a nilpotent linear transformation of index p on an n-dimensional vector space $\mathcal{V}$; let $\omega_1, \ldots, \omega_p$ be the Weyr characteristic of T. Then $\omega_k \geq \omega_{k+1}$ for $k = 1, \ldots, p-1$.*

Proof. Let $\mathcal{U}_k$ denote the null space of T^k. The space $\mathcal{U}_0$ is the zero space and $\mathcal{U}_p = \mathcal{V}$. For any $1 \leq k \leq p-1$, we have $\mathcal{U}_{k-1} \subset \mathcal{U}_k \subset \mathcal{U}_{k+1}$. The dimension of the quotient space $\mathcal{U}_k/\mathcal{U}_{k-1}$ is ω_k, and the dimension of the quotient space $\mathcal{U}_{k+1}/\mathcal{U}_k$ is ω_{k+1}.

Choose ω_{k+1} vectors $\mathbf{y}_1, \ldots, \mathbf{y}_{\omega_{k+1}}$ in $\mathcal{U}_{k+1}$ such that the cosets

$$\left\{ \mathbf{y}_i + \mathcal{U}_k \;\middle|\; 1 \leq i \leq \omega_{k+1} \right\}$$

are a basis for the quotient space $\mathcal{U}_{k+1}/\mathcal{U}_k$. Note that for each i, the vector $\mathbf{y}_i$ is in $\mathcal{U}_{k+1}$ but not in $\mathcal{U}_k$. Consequently, $T(\mathbf{y}_i)$ is in $\mathcal{U}_k$, but not in $\mathcal{U}_{k-1}$. We claim that the ω_{k+1} cosets

$$T(\mathbf{y}_i) + \mathcal{U}_{k-1}, \quad 1 \leq i \leq \omega_{k+1},$$

are linearly independent in the quotient space $\mathcal{U}_k/\mathcal{U}_{k-1}$. For, suppose the sum

$$\sum_{i=1}^{\omega_{k+1}} a_i \big(T(\mathbf{y}_i) + \mathcal{U}_{k-1} \big)$$

is the zero vector in the quotient space $\mathcal{U}_k/\mathcal{U}_{k-1}$. Then $T\Big(\sum_{i=1}^{\omega_{k+1}} a_i \mathbf{y}_i \Big)$ is in $\mathcal{U}_{k-1}$ and so $T^k\Big(\sum_{i=1}^{\omega_{k+1}} a_i \mathbf{y}_i \Big) = 0$. But then $\sum_{i=1}^{\omega_{k+1}} a_i \mathbf{y}_i \in \mathcal{U}_k$, which means that the linear combination $\sum_{i=1}^{\omega_{k+1}} a_i \big(\mathbf{y}_i + \mathcal{U}_k \big)$ is the zero vector in the quotient space $\mathcal{U}_{k+1}/\mathcal{U}_k$. Since the cosets $\mathbf{y}_i + \mathcal{U}_k$, for $1 \leq i \leq \omega_{k+1}$, are linearly independent, we have $a_i = 0$ for all i, and thus $\left\{ T(\mathbf{y}_i) + \mathcal{U}_{k-1} \;\middle|\; 1 \leq i \leq \omega_{k+1} \right\}$ is linearly independent in the quotient space $\mathcal{U}_k/\mathcal{U}_{k-1}$. Since ω_k is the dimension of $\mathcal{U}_k/\mathcal{U}_{k-1}$, we must have $\omega_k \geq \omega_{k+1}$. □

Our assertion that two $n \times n$ nilpotent matrices are similar if they have the same Weyr characteristic was based on the fact that the Weyr characteristic determines the Jordan canonical form. We now see how to show this directly with the Weyr theory, without invoking Jordan canonical form. First, some preliminaries.

Definition 4.34. We say a matrix has *full column rank* if it has linearly independent columns.

Note that a matrix B has full column rank if and only if the only solution to $B\mathbf{x} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$.

Lemma 4.35. *If B and C are matrices of full column rank of sizes $r \times s$ and $s \times t$, respectively, then BC has full column rank.*

Proof. Suppose $(BC)\mathbf{x} = \mathbf{0}$. Then $B(C\mathbf{x}) = \mathbf{0}$. Since B has full column rank, this gives $C\mathbf{x} = \mathbf{0}$. But C also has full column rank, so $\mathbf{x} = 0$. Hence, the columns of BC are linearly independent. □

Lemma 4.36. *Suppose $m_1 \geq m_2 \geq \cdots \geq m_p$. Suppose A is a block triangular matrix, with block (i, j) of size $m_i \times m_j$, where the diagonal blocks of A are all blocks of zeroes, and each super-diagonal block $A_{k,(k+1)}$ has full column rank. Thus, A has the form*

$$A = \begin{pmatrix} 0_{m_1} & A_{12} & A_{13} & A_{14} & \cdots & A_{1p} \\ 0 & 0_{m_2} & A_{23} & A_{24} & \cdots & A_{2p} \\ 0 & 0 & 0_{m_3} & A_{34} & \cdots & A_{3p} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0_{m_{p-1}} & A_{(p-1),p} \\ 0 & 0 & 0 & 0 & \cdots & 0_{m_p} \end{pmatrix}, \tag{4.18}$$

where the $m_k \times m_{k+1}$ block $A_{k,(k+1)}$ has full column rank. Then $m_1, \ldots, m_p$ is the Weyr characteristic of A.

Proof. Since each block $A_{k,(k+1)}$ has rank m_{k+1}, the last $n - m_1$ columns of A are linearly independent, and $\dim(\ker(A)) = m_1$. Squaring A, we see that A^2 has blocks of zeroes in both the diagonal and super-diagonal blocks, while the next diagonal line (i.e., the diagonal line above the super-diagonal) contains the products

$$A_{12}A_{23}, \quad A_{23}A_{34}, \quad A_{34}A_{45}, \quad \ldots, \quad A_{(p-2),(p-1)}A_{(p-1),p}.$$

From Lemma 4.35, each of these products has full column rank, and hence the dimension of $\ker(A^2)$ is $m_1 + m_2$. With each successive power of A, we get an additional diagonal line of zero blocks. In A^k, each block in the first diagonal line of nonzero blocks is a product of k consecutive $A_{i,(i+1)}$ blocks, hence has linearly independent columns. So the dimension of the null space of A^k is $m_1 + m_2 + \cdots + m_k$ and $m_1, \ldots, m_p$ is the Weyr characteristic of A. □

Now we show that any nilpotent linear transformation on a finite-dimensional vector space has a matrix representation of the form (4.18). We again use quotient spaces for this argument, but the result can be obtained in other ways [**Sha99**].

Theorem 4.37. *Suppose $T : \mathcal{V} \to \mathcal{V}$ is a nilpotent linear transformation on an n-dimensional vector space $\mathcal{V}$. Let $\omega_1, \ldots, \omega_p$ be the Weyr characteristic of T. Then there is a basis $\mathcal{B}$ for $\mathcal{V}$ such that $[T]_{\mathcal{B}}$ has the block triangular form of* (4.18) *with*

$m_k = \omega_k$. *That is,*

$$A = [T]_{\mathcal{B}} = \begin{pmatrix} 0_{\omega_1} & A_{12} & A_{13} & A_{14} & \cdots & A_{1p} \\ 0 & 0_{\omega_2} & A_{23} & A_{24} & \cdots & A_{2p} \\ 0 & 0 & 0_{\omega_3} & A_{34} & \cdots & A_{3p} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0_{\omega_{p-1}} & A_{(p-1),p} \\ 0 & 0 & 0 & 0 & \cdots & 0_{\omega_p} \end{pmatrix}, \tag{4.19}$$

where the $\omega_k \times \omega_{k+1}$ *block* $A_{k,(k+1)}$ *has full column rank.*

Proof. We use the method in the proof of Theorem 4.33 to construct $\mathcal{B}$. For each $k = 0, \ldots, (p-1)$, choose ω_{k+1} vectors $\mathbf{b}_{(k+1),1}, \ldots, \mathbf{b}_{(k+1),\omega_{k+1}}$ in $\mathcal{U}_{k+1} \setminus \mathcal{U}_k$ such that $\{\mathbf{b}_{(k+1),i} + \mathcal{U}_k \mid 1 \le i \le \omega_{k+1}\}$ is a basis for the quotient space $\mathcal{U}_{k+1}/\mathcal{U}_k$. Set $\mathcal{B}_{k+1} = \{\mathbf{b}_{(k+1),1}, \ldots \mathbf{b}_{(k+1),\omega_{k+1}}\}$. Then $\mathcal{B} = \bigcup_{k=1}^{p} \mathcal{B}_k$ is a basis for $\mathcal{V}$.

Let $A = [T]_{\mathcal{B}}$. For each k, the transformation T maps the vectors of $\mathcal{B}_k$ into the space $\mathcal{U}_{k-1}$, so, for each i, the vector $T(\mathbf{b}_{k,i})$ is a linear combination of vectors in $\bigcup_{j=1}^{k-1} \mathcal{B}_j$. Consequently, A is block triangular with blocks of sizes $\omega_i \times \omega_j$, and the diagonal blocks are zero. From the proof of Theorem 4.33, we know that

$$\left\{T(\mathbf{b}_{(k+1),i}) + \mathcal{U}_{k-1} \;\middle|\; i = 1, \ldots, \omega_{k+1}\right\}$$

is linearly independent in the quotient space $\mathcal{U}_k/\mathcal{U}_{k-1}$, so each super-diagonal block $A_{k,(k+1)}$ has full column rank. □

The quotient space and coset notation in this argument may be obscuring the essential idea, so let us describe it in a simpler way. The set $\mathcal{B}_1$ is a basis for $\ker(T)$ and has ω_1 vectors in it. The set $\mathcal{B}_2$ has ω_2 vectors and is chosen so that $\mathcal{B}_1 \cup \mathcal{B}_2$ is a basis for $\ker(T^2)$. Thus, we start with a basis for the null space of T, and then extend it to get a basis for the null space of T^2. Then we adjoin the ω_3 vectors of $\mathcal{B}_3$ to get a basis for the null space of T^3, and so on. In general, at stage k, we have a basis $\bigcup_{i=1}^{k} \mathcal{B}_i$ for the null space of T^k, and adjoin ω_{k+1} additional vectors to get a basis for the null space of T^{k+1}; the set $\mathcal{B}_{k+1}$ is the set of these $k+1$ additional vectors. We used the quotient space language as a convenient tool to prove that the blocks $A_{k,(k+1)}$ have full column rank.

We now examine two particular ways of choosing the sets $\mathcal{B}_k$. One yields an orthonormal basis, $\mathcal{B}$. The other will yield a basis such that $[T]_{\mathcal{B}}$ is in Weyr canonical form.

Theorem 4.38. *Let* $T : \mathcal{V} \to \mathcal{V}$ *be a nilpotent linear transformation on an* n*-dimensional inner product space* $\mathcal{V}$*. Let* $\omega_1, \ldots, \omega_p$ *be the Weyr characteristic of* T*. Then there is an orthonormal basis* $\mathcal{B}$ *for* $\mathcal{V}$ *such that* $[T]_{\mathcal{B}}$ *has the block triangular form of* (4.19)*.*

Proof. Select the set $\mathcal{B}_1$ so that it is an orthonormal basis for $\mathcal{U}_1 = \ker(T)$. Since $\mathcal{U}_1$ is an ω_1-dimensional subspace of the $(\omega_1 + \omega_2)$-dimensional subspace $\mathcal{U}_2$, the

subspace $\mathcal{U}_1^\perp \cap \mathcal{U}_2$ is an ω_2-dimensional subspace of $\mathcal{U}_2$. Choose $\mathcal{B}_2$ to be an orthonormal basis for $\mathcal{U}_1^\perp \cap \mathcal{U}_2$; then $\mathcal{B}_1 \cup \mathcal{B}_2$ is an orthonormal basis for $\mathcal{U}_2$.

In general, $\mathcal{U}_{k-1}$ is an $(\omega_1 + \omega_2 + \cdots + \omega_{k-1})$-dimensional subspace of the subspace $\mathcal{U}_k$, so $\mathcal{U}_{k-1}^\perp \cap \mathcal{U}_k$ is an ω_k-dimensional subspace of $\mathcal{U}_k$; choose $\mathcal{B}_k$ to be an orthonormal basis for $\mathcal{U}_{k-1}^\perp \cap \mathcal{U}_k$. Now, for $i < k$, we have $\mathcal{U}_i \subset \mathcal{U}_k$, so $\mathcal{U}_k^\perp \subset \mathcal{U}_i^\perp$. Hence, the fact that $\mathcal{B}_k \subseteq \mathcal{U}_{k-1}^\perp$ tells us that vectors of $\mathcal{B}_k$ are orthogonal to the vectors in $\mathcal{B}_i$ whenever $i < k$. Hence, $\mathcal{B} = \bigcup_{i=1}^{p} \mathcal{B}_k$ is an orthonormal basis for $\mathcal{V}$. □

Theorem 4.38 is important for numerical computation where stability issues are important. It is desirable to work with unitary similarity and stick with orthonormal change of basis. See [**Sha99**] for references.

Finally, to show any nilpotent matrix is similar to one in Weyr normal form, we need to show that there is a basis $\mathcal{B}$ such that the block triangular matrix of (4.19) has $A_{k,(k+1)} = I_{\omega_k,\omega_{k+1}}$, with all the other blocks being zero. This will involve starting with the set $\mathcal{B}_p$ and working backwards. Before plunging into the morass of notation for the general case, let us see the argument for the case $p = 2$. Let $\mathcal{B}_2 = \{\mathbf{y}_1, \ldots, \mathbf{y}_{\omega_2}\}$ be a linearly independent set of ω_2 vectors from $\mathcal{U}_2 \setminus \mathcal{U}_1$. Then, as seen in the proof of Theorem 4.33, the vectors $T(\mathbf{y}_1), T(\mathbf{y}_2), \ldots, T(\mathbf{y}_{\omega_2})$ are linearly independent and are in $\mathcal{U}_1$. We extend the set $\{T(\mathbf{y}_1), T(\mathbf{y}_2), \ldots, T(\mathbf{y}_{\omega_2})\}$ to a basis for $\mathcal{U}_1$ by adjoining $m = \omega_1 - \omega_2$ additional vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ from $\mathcal{U}_1$. (If $m = 0$, we do not need to adjoin any vectors.) We now have the following basis $\mathcal{B}$ for $\mathcal{V}$:

$$\mathcal{B} = \{T(\mathbf{y}_1), T(\mathbf{y}_2), \ldots, T(\mathbf{y}_{\omega_2}), \quad \mathbf{v}_1, \ldots, \mathbf{v}_m, \quad \mathbf{y}_1, \ldots, \mathbf{y}_{\omega_2}\}.$$

Remember that $\omega_2 + m = \omega_1$. Since $T(T(\mathbf{y}_i)) = T^2(\mathbf{y}_i) = \mathbf{0}$ and $T(\mathbf{v}_i) = \mathbf{0}$, we see T maps the first ω_1 vectors in this set to zero. (These first ω_1 vectors are a basis for $\mathcal{U}_1$.) And clearly, T maps the last ω_2 vectors of this basis onto the first ω_2 vectors, retaining the order. Hence,

$$[T]_{\mathcal{B}} = \begin{pmatrix} 0_{\omega_1} & I_{\omega_1,\omega_2} \\ 0 & 0_{\omega_2} \end{pmatrix}.$$

This is the basic idea of the proof for the general case, but we need to repeat the procedure $p - 1$ times, working backwards from $\mathcal{U}_p$ to $\mathcal{U}_1$. Try to keep this main idea in mind as the notation escalates.

Theorem 4.39. *Let $T : \mathcal{V} \to \mathcal{V}$ be a nilpotent linear transformation on an n-dimensional vector space $\mathcal{V}$. Let $\omega_1, \ldots, \omega_p$ be the Weyr characteristic of T. Then there is a basis, $\mathcal{B}$, for $\mathcal{V}$ such that $[T]_{\mathcal{B}}$ is block triangular with blocks of sizes $\omega_i \times \omega_j$, with $A_{k,(k+1)} = I_{\omega_k,\omega_{k+1}}$ and all other blocks are zero. Thus,*

$$A = [T]_{\mathcal{B}} = \begin{pmatrix} 0_{\omega_1} & I_{\omega_1,\omega_2} & 0 & 0 & \cdots & 0 \\ 0 & 0_{\omega_2} & I_{\omega_2,\omega_3} & 0 & \cdots & 0 \\ 0 & 0 & 0_{\omega_3} & I_{\omega_3,\omega_4} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0_{\omega_{p-1}} & I_{\omega_{p-1},\omega_p} \\ 0 & 0 & 0 & 0 & \cdots & 0_{\omega_p} \end{pmatrix}. \tag{4.20}$$

Proof. First we set up some notation. If $\mathcal{S}$ is a set of vectors, write $T(\mathcal{S})$ for $\{T(\mathbf{x}) \mid \mathbf{x} \in \mathcal{S}\}$. For a set of vectors, $\mathcal{S} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ in $\mathcal{U}_{k+1}$, we write

$$\overline{\mathcal{S}} = \{\mathbf{y}_i + \mathcal{U}_k \mid 1 \leq i \leq m\}$$

for the corresponding cosets of $\mathcal{U}_{k+1}/\mathcal{U}_k$.

Let $\mathcal{B}_p$ be a set of ω_p vectors in $\mathcal{U}_p \setminus \mathcal{U}_{p-1}$ such that $\overline{\mathcal{B}_p}$ is a basis for the quotient space $\mathcal{U}_p/\mathcal{U}_{p-1}$. Then $T(\mathcal{B}_p) \subseteq \mathcal{U}_{p-1}$, and from the proof of Theorem 4.33, we know $\overline{T(\mathcal{B}_p)}$ is linearly independent in the quotient space $\mathcal{U}_{p-1}/\mathcal{U}_{p-2}$. Hence, we can extend $\overline{T(\mathcal{B}_p)}$ to a basis for $\mathcal{U}_{p-1}/\mathcal{U}_{p-2}$ by adjoining $m_p = \omega_{(p-1)} - \omega_p$ additional cosets from vectors in $\mathcal{U}_{p-1} \setminus \mathcal{U}_{p-2}$. Label these additional vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{m_p}$ and set

$$\mathcal{B}_{p-1} = T(\mathcal{B}_p) \cup \{\mathbf{v}_1, \ldots, \mathbf{v}_{m_p}\}.$$

Then $\overline{\mathcal{B}_{p-1}}$ is a basis for $\mathcal{U}_{p-1}/\mathcal{U}_{p-2}$.

Repeating the process, we have $T(\mathcal{B}_{p-1}) \subseteq \mathcal{U}_{p-2}$ and $\overline{T(\mathcal{B}_{p-1})}$ is linearly independent in the quotient space $\mathcal{U}_{p-2}/\mathcal{U}_{p-3}$. We adjoin $m_{p-1} = \omega_{p-2} - \omega_{p-1}$ cosets from vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{m_{p-1}}$ in $\mathcal{U}_{p-2} \setminus \mathcal{U}_{p-3}$ and get

$$\mathcal{B}_{p-2} = T(\mathcal{B}_{p-1}) \cup \{\mathbf{w}_1, \ldots, \mathbf{w}_{m_{p-1}}\},$$

where $\overline{\mathcal{B}_{p-2}}$ is a basis for $\mathcal{U}_{p-2}/\mathcal{U}_{p-3}$.

So far, we have

$$\mathcal{B}_{p-2} \cup \mathcal{B}_{p-1} \cup \mathcal{B}_p = \{T(\mathcal{B}_{p-1}),\ \mathbf{w}_1, \ldots, \mathbf{w}_{m_{p-1}},\ T(\mathcal{B}_p),\ \mathbf{v}_1, \ldots, \mathbf{v}_{m_p}, \mathcal{B}_p\}.$$

Continue in this fashion. For each k, with $\overline{\mathcal{B}_k}$ a basis for $\mathcal{U}_k/\mathcal{U}_{k-1}$, the set $\overline{T(\mathcal{B}_k)}$ is linearly independent in the quotient space $\mathcal{U}_{k-1}/\mathcal{U}_{k-2}$ and thus can be extended to a basis for $\mathcal{U}_{k-1}/\mathcal{U}_{k-2}$ by adjoining cosets from $m_k = \omega_{(k-1)} - \omega_k$ vectors from $\mathcal{U}_{k-1} \setminus \mathcal{U}_{k-2}$.

Here is the key point: for $k \geq 2$, the transformation T sends the ω_k vectors of $\mathcal{B}_k$ to the first ω_k vectors of $\mathcal{B}_{k-1}$. The ω_1 vectors of $\mathcal{B}_1$ are sent to zero. Hence, if we use the basis $\mathcal{B} = \bigcup_{i=1}^{p} \mathcal{B}_i$, the matrix $[T]_{\mathcal{B}}$ has the desired form. □

As a consequence of Theorem 4.39, we have now shown, directly from the Weyr characteristic, that two nilpotent matrices are similar if and only if they have the same Weyr characteristic. Note also that if we re-order the vectors of the basis $\mathcal{B}$, constructed in the proof of Theorem 4.39, we can recover the Jordan canonical form. Specifically, for each of the ω_p vectors $\mathbf{x}$ in $\mathcal{B}_p$, the p vectors, $T^{p-1}\mathbf{x}, T^{p-2}\mathbf{x}, \ldots, T\mathbf{x}, \mathbf{x}$, are in the basis $\mathcal{B}$. The matrix representing the action of T on these p vectors is the Jordan block $J_p(0)$. If $m_p > 0$, then for each of the $m_p = \omega_{p-1} - \omega_p$ vectors $\mathbf{v}_i$, we have a chain of $p-1$ vectors, $T^{p-2}\mathbf{v},\ T^{p-3}\mathbf{v},\ \ldots,\ T\mathbf{v},\ \mathbf{v}$ such that the action of T on these $p-1$ vectors is represented by the Jordan block $J_{p-1}(0)$. Hence by re-ordering the vectors of $\mathcal{B}$, starting with the chains generated by vectors in $\mathcal{B}_p$, then the chains generated by the vectors in $\mathcal{B}_{p-1} \setminus T(\mathcal{B}_p)$, next the chains generated by the vectors in $\mathcal{B}_{p-2} \setminus T(\mathcal{B}_{p-1})$, and so on, we get a matrix representation for T which is in Jordan canonical form. Note that ω_p gives the number of Jordan blocks of size p, and then $m_p = \omega_{p-1} - \omega_p$ gives the number of

Chapter 8

Some Matrix Factorizations

This chapter concerns some matrix factorizations. The singular value decomposition (SVD) was introduced in Section 6.8; we now examine the SVD in more detail. Next, we look at Householder transformations (introduced in Chapter 2, Exercise 18 and Chapter 3, Exercise 10) and use them to achieve the QR factorization, and to transform matrices to special forms, such as Hessenberg and tridiagonal forms, with unitary transformations. To give some idea of why these techniques are useful in computational matrix theory, we briefly describe the basics for a few classic methods for the numerical computation of eigenvalues. For serious discussion of these methods with analysis of the error and the efficiency for each algorithm, see one of the many texts on computational and numerical linear algebra, some classic, and some more recent, for example [**Hou64, Wilk65, LawHan74, GvL89, Stew73, Parl80, StewSun90, Tref97, Demm97**].

The chapter concludes with a brief review of the well-known LDU factorization, which comes from the Gaussian elimination process.

Except for the section on the LDU factorization, matrices in this section are over $\mathbb{C}$, unless we specifically say that the matrix is real.

8.1. Singular Value Decomposition

Section 6.8 obtained the singular value decomposition for nonsingular square matrices from the polar factorization. Here we approach the SVD directly for $m \times n$ matrices.

Recall the following fact (Theorem 2.22) from Chapter 2: If A is a complex matrix, then the four matrices A, A^*, AA^*, and A^*A all have the same rank. Let A be an $m \times n$ matrix of rank r. Then A^*A and AA^* are both positive semi-definite matrices of rank r; each has r positive eigenvalues and then $n - r$ and $m - r$ zero eigenvalues, respectively. Theorem 3.24 tells us that A^*A and AA^* have the same nonzero eigenvalues.

Definition 8.1. Let A be an $m \times n$ matrix of rank r. Let $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_r^2$ be the positive eigenvalues of A^*A (equivalently, of AA^*). The numbers $\sigma_1, \ldots, \sigma_r$ are called the *singular values* of A.

Remark 8.2. Since the $n \times n$ matrix A^*A has n eigenvalues, one may also put $\sigma_j = 0$ for $r+1 \leq j \leq n$ and define the singular values of A to be $\sigma_1, \ldots, \sigma_n$. Another option is use the $m \times m$ matrix AA^*, with the list $\sigma_1, \ldots, \sigma_m$, where $\sigma_j = 0$ for $r+1 \leq j \leq m$. We use all of these, depending on what is convenient.

Theorem 8.3. *Let A be an $m \times n$ matrix of rank r; let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ be the singular values of A. Let Σ_A be the $m \times n$ diagonal matrix with $\sigma_1, \ldots, \sigma_r$ in the first r diagonal entries and zeroes elsewhere. Then there exist unitary matrices U and V, of sizes $m \times m$ and $n \times n$, respectively, such that $A = U\Sigma_A V$.*

Proof. Note that $D = \Sigma_A^* \Sigma_A = \Sigma_A^T \Sigma_A$ is a real $n \times n$ diagonal matrix with $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_r^2$ in the first r diagonal entries and zeroes elsewhere. The $n \times n$ matrix A^*A is Hermitian, with r positive eigenvalues $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_r^2$ and $(n-r)$ eigenvalues equal to zero. Hence, from the spectral theorem, there is an $n \times n$ unitary matrix V such that

$$VA^*AV^* = D = \Sigma_A^* \Sigma_A. \tag{8.1}$$

The ij entry of VA^*AV^* is the inner product of columns j and i of AV^*, so equation (8.1) tells us the columns of AV^* are pairwise orthogonal. Furthermore, when $1 \leq j \leq r$ the length of column j is σ_j. For $j > r$, equation (8.1) tells us the jth column of AV^* is a column of zeroes. For $1 \leq j \leq r$, divide column j of AV^* by its length, σ_j, and let U_r denote the $m \times r$ matrix with $\frac{1}{\sigma_j}$ (column j of AV^*) as its jth column. The r columns of U_r are then an orthonormal set. Now complete U_r to an $m \times m$ unitary matrix by using an orthonormal basis for the orthogonal complement of the column space of U_r for the remaining $m - r$ columns. We then have $AV^* = U\Sigma_A$. Multiply both sides on the right by V to obtain $A = U\Sigma_A V$. □

Definition 8.4. Let A be $m \times n$ complex matrix of rank r, with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. A factorization of the form $A = U\Sigma_A V$, where Σ_A is the $m \times n$ diagonal matrix with $\sigma_1, \ldots, \sigma_r$ in the first r diagonal entries and zeroes elsewhere, and U and V are unitary matrices of sizes $m \times m$ and $n \times n$, respectively, is called a *singular value decomposition* for A. We write *SVD* for a singular value decomposition.

Theorem 8.3 states that any complex matrix has an SVD. Note that while Σ_A is uniquely determined by A, the unitary matrices U, V are not. The equation $AV^* = U\Sigma_A$ tells us how the transformation A acts:

$$A \text{ (column } j \text{ of } V^*) = \sigma_j (\text{column } j \text{ of } U).$$

Thus, using the columns of V^* as an orthonormal basis for $\mathbb{C}^n$ and the columns of U as an orthonormal basis for $\mathbb{C}^m$, the effect of the transformation A is to map the jth basis vector of $\mathbb{C}^n$ to a multiple of the jth basis vector of $\mathbb{C}^m$; the multiplier is the singular value σ_j. The basis vectors have length one, so the jth basis vector in $\mathbb{C}^n$ is mapped to a vector of length σ_j in $\mathbb{C}^m$. The largest singular value σ_1 is the largest factor by which the length of a basis vector is multiplied. We now show that σ_1 is the largest factor by which the length of any vector is multiplied.

Theorem 8.5. *Let A be an $m \times n$ matrix with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. Then* $\max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| = \sigma_1$.

Proof. We have already seen, from the equation $AV^* = U\Sigma_A$, that if $\mathbf{x}$ is the first column of V^*, then $\|\mathbf{x}\| = 1$ and $\|A\mathbf{x}\| = \sigma_1$.

Now suppose $\mathbf{x}$ is any vector in $\mathbb{C}^n$ of length one. From the singular value decomposition $A = U\Sigma_A V$, we have $A\mathbf{x} = U\Sigma_A V\mathbf{x}$. Set $\mathbf{y} = V\mathbf{x}$; since V is unitary, $\|\mathbf{y}\| = 1$. In the product $\Sigma_A\mathbf{y}$, coordinate i of $\mathbf{y}$ gets multiplied by σ_i; hence $\|\Sigma_A\mathbf{y}\| \leq \sigma_1\|\mathbf{y}\| = \sigma_1$. Since U is unitary,

$$\|A\mathbf{x}\| = \|U\Sigma_A V\mathbf{x}\| = \|U\Sigma_A\mathbf{y}\| = \|\Sigma_A\mathbf{y}\| \leq \sigma_1. \qquad \square$$

Definition 8.6. The *spectral norm* of an $m \times n$ matrix A is $\max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| = \sigma_1$. We denote the spectral norm of A as $\|A\|_2$.

The spectral norm is also called the *operator norm*; in Section 7.2 we saw it is the norm induced by the 2-norm on $\mathbb{C}^n$. We have $\|A\mathbf{x}\| \leq \|A\|_2\|\mathbf{x}\|$ for all $\mathbf{x}$. When A is square and $\mathbf{v}$ is an eigenvector of A, we have $\|A\mathbf{v}\| = \|\lambda\mathbf{v}\| = |\lambda|\|\mathbf{v}\|$, so $|\lambda| \leq \|A\|_2$ for any eigenvalue λ of A. Hence, $\rho(A) \leq \|A\|_2 = \sigma_1$.

Recall the Frobenius norm $\|A\|_F = \big(\mathrm{trace}(A^*A)\big)^{\frac{1}{2}} = \Big(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\Big)^{\frac{1}{2}}$. It is often more convenient to work with $\|A\|_F^2 = \mathrm{trace}(A^*A)$. Since $\|A\|_F = \|UAV\|_F$ for any unitary matrices U, V, we have

$$\|A\|_F^2 = \|\Sigma_A\|_F^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2.$$

There are other ways to write the factorization $A = U\Sigma_A V$. Since only the first r diagonal entries of Σ_A are nonzero, the last $(m-r)$ columns of U and the last $(n-r)$ rows of V are superfluous. Let $\hat{\Sigma}_A$ be the $r \times r$ diagonal matrix, $\mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. Replace the $n \times n$ unitary matrix V with the $r \times n$ matrix V_r consisting of the first r rows of V; these rows form an orthonormal set. And instead of completing the $m \times r$ matrix U_r in the proof of Theorem 8.3 to get a square unitary matrix U, simply use U_r, which has r orthonormal columns. We then have

$$A = U_r\hat{\Sigma}_r V_r, \tag{8.2}$$

where U_r is $m \times r$ with orthonormal columns, $\hat{\Sigma}_r$ is an $r \times r$ diagonal matrix with positive diagonal entries, and V_r is $r \times n$ with orthonormal rows. We may also decompose (8.2) into a sum of r rank one matrices. Let $\mathbf{u}_i$ denote column i of U_r, and let $\mathbf{v}_i$ be column i of V_r^*. Then row i of V_r is $\mathbf{v}_i^*$, and

$$A = \sum_{i=1}^{r} \sigma_i\mathbf{u}_i\mathbf{v}_i^* = \sigma_1\mathbf{u}_1\mathbf{v}_1^* + \sigma_2\mathbf{u}_2\mathbf{v}_2^* + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r^*. \tag{8.3}$$

Since $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, while all of the vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ have the same length (namely, one), the sum (8.3) shows that the earlier terms in the sum (that is, the terms corresponding to larger σ_i's) make larger contributions to A. This suggests that to approximate A with a matrix of lower rank (i.e., one with fewer terms in the sum), one should use the terms corresponding to larger singular values

and drop terms with smaller singular values. Let $1 \le k \le r$. From $A = U\Sigma_A V$, let U_k be the $m \times k$ matrix consisting of the first k columns of U, and let V_k be the $k \times n$ matrix consisting of the first k rows of V. Set

$$A_k = U_k \mathrm{diag}(\sigma_1, \ldots, \sigma_k) V_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^*.$$

We shall see that of all matrices of rank k, the matrix A_k is the one which is "closest" to A, where distance is measured by the Frobenius norm. Let Σ_k denote the $m \times n$ matrix with $\sigma_1, \ldots, \sigma_k$ in the first k diagonal positions and zeroes elsewhere. Then $A_k = U\Sigma_k V$ and

$$\|A - A_k\|_F^2 = \|U(\Sigma_A - \Sigma_k)V\|_F^2 = \|\Sigma_A - \Sigma_k\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2. \tag{8.4}$$

The following theorem is usually attributed to Eckart and Young [**EcYo36**]. Stewart [**Stew93**] points out that the result is older and is contained in [**Schm07**].

Theorem 8.7 (Schmidt [**Schm07**])**.** *Let A be an $m \times n$ matrix of rank r with singular value decomposition $A = U\Sigma_A V = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^*$, where $\mathbf{u}_i$ is column i of U and $\mathbf{v}_i^*$ is row i of V. For $1 \le k \le r$, set $A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^*$. Then for any $m \times n$ matrix B of rank at most k, we have $\|A - A_k\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2 \le \|A - B\|_F^2$.*

We give two proofs. The first, from [**Stew73**, pp. 322–323], has the advantage of being direct and natural. The second proof, based on Schmidt's argument as presented in [**Stew93**], seems sneakier and more complicated, but we feel it has its own charm.

First proof. Let B be an $m \times n$ matrix of rank at most k, which gives the minimum value of $\|A - X\|_F$ for all matrices X of rank at most k. At this point, one must ask how we are guaranteed the existence of such a B, and need to do a bit of analysis. Define the real-valued function, f, on the space of $m \times n$ complex matrices by $f(X) = \|A - X\|_F$. The function f is continuous. The set of matrices of rank at most k is a closed set. Since we seek to minimize $\|A - X\|_F$, we may restrict our attention to matrices X satisfying $\|A - X\|_F \le \|A\|_F$. The set

$$\{X \mid \mathrm{rank}(X) \le k \text{ and } \|A - X\|_F \le \|A\|_F\}$$

is then a closed bounded set in a finite-dimensional normed space; hence, it is compact and the continuous function f attains its minimum value for some B in this set. Now let $\beta_1 \ge \beta_2 \ge \cdots \ge \beta_k \ge 0$ be the singular values of B, and let $U\Sigma_B V$ be the singular value decomposition of B. The Frobenius norm is invariant under unitary transformations, so we may, without loss of generality, assume

$$B = \Sigma_B = \begin{pmatrix} D_k & 0 \\ 0 & 0 \end{pmatrix},$$

where $D_k = \operatorname{diag}(\beta_1, \ldots, \beta_k)$. Partition A conformally with B, thus

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

where A_{11} is $k \times k$. Then

$$\|A - B\|_F^2 = \|A_{11} - D_k\|_F^2 + \|A_{12}\|_F^2 + \|A_{21}\|_F^2 + \|A_{22}\|_F^2.$$

If $A_{12} \neq 0$, then $B_1 = \begin{pmatrix} D_k & A_{12} \\ 0 & 0 \end{pmatrix}$ is a matrix of rank at most k with

$$\|A - B_1\|_F^2 = \|A_{11} - D_k\|_F^2 + \|A_{21}\|_F^2 + \|A_{22}\|_F^2 < \|A - B\|_F^2.$$

So, by the choice of B, we must have $A_{12} = 0$. The same argument shows $A_{21} = 0$ and $D_k = A_{11}$. Hence, we have $A = \begin{pmatrix} D_k & 0 \\ 0 & A_{22} \end{pmatrix}$ and the numbers $\beta_1, \ldots, \beta_k$ are k of the singular values of A. Also, $\|A - B\|_F = \|A_{22}\|_F$. Since $\|A_{22}\|_F^2$ is the sum of the squares of the remaining singular values of A, and since $k \leq r$, it is clear that the minimum value possible for $\|A_{22}\|_F^2$ is $\sum_{i=k+1}^{r} \sigma_i^2$. To achieve this value we must have $\beta_i = \sigma_i$ for $i = 1, \ldots, k$. □

Now for the second proof, based on the presentation of Schmidt's argument in [**Stew93**]. First, a lemma is needed in the proof, but it is of interest in its own right.

Lemma 8.8. *Let A be an $m \times n$ complex matrix of rank r with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. For $k > r$, put $\sigma_k = 0$. Suppose $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is a set of k orthonormal vectors in $\mathbb{C}^n$. Then $\sum_{i=1}^{k} \|A\mathbf{x}_i\|^2 \leq \sum_{i=1}^{k} \sigma_i^2$.*

Remark 8.9. For $k = 1$ this gives Theorem 8.5.

Proof. Let X be the $n \times k$ matrix with column vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$. Since the columns are orthonormal, $X^*X = I_k$. We have $AX = (\, A\mathbf{x}_1 \quad A\mathbf{x}_2 \quad \cdots \quad A\mathbf{x}_k \,)$ and

$$\sum_{i=1}^{k} \|A\mathbf{x}_i\|^2 = \|AX\|_F^2 = \operatorname{trace}(X^*A^*AX).$$

Let $A = U\Sigma V$ be the singular value decomposition of A, where U, V are unitary matrices of sizes $m \times m$ and $n \times n$, respectively, and Σ is the $m \times n$ diagonal matrix with the singular values on the diagonal. Put $V_k = VX$; the $n \times k$ matrix V_k has orthonormal columns, and $AX = U\Sigma V_k$. Using $U^*U = I$, we have

$$X^*A^*AX = V_k^*\Sigma^T U^* U \Sigma V_k = V_k^* \Sigma^T \Sigma V_k.$$

Recall that $\operatorname{trace}(BC) = \operatorname{trace}(CB)$; apply this with $B = V_k^*\Sigma^T\Sigma$ and $C = V_k$ to get $\operatorname{trace}(X^*A^*AX) = \operatorname{trace}(V_k V_k^* \Sigma^T \Sigma)$. Let $\mathbf{v}_i$ denote the ith row of V_k. The (i,i) entry of $V_k V_k^*$ is then $\|\mathbf{v}_i\|^2$ and so

$$\sum_{i=1}^{k} \|A\mathbf{x}_i\|^2 = \operatorname{trace}(V_k V_k^* \Sigma^T \Sigma) = \sum_{i=1}^{r} \sigma_i^2 \|\mathbf{v}_i\|^2. \tag{8.5}$$

Now $\sum_{i=1}^{k} \|\mathbf{v}_i\|^2 = \text{trace}(V_k V_k^*) = \text{trace}(V_k^* V_k) = k$, because $V_k^* V_k = I_k$. So the last sum in (8.5) has the form $\sum_{i=1}^{r} c_i \sigma_i^2$, where the coefficents $c_i = \|\mathbf{v}_i\|^2$ satisfy $0 \le c_i \le 1$ and sum to k. Since $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r$, a sum of this form is maximized by setting $c_i = 1$ for $i = 1, \ldots, k$ and thus $\sum_{i=1}^{r} \sigma_i^2 \|\mathbf{v}_i\|^2 \le \sum_{i=1}^{k} \sigma_i^2$. □

Now for the second proof of Theorem 8.7.

Second proof. Since $\|A - A_k\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2$, we need to show that for any $m \times n$ matrix B with $\text{rank}(B) \le k$, we have $\sum_{i=k+1}^{r} \sigma_i^2 \le \|A - B\|_F^2$. Since $\text{rank}(B) \le k$, we can factor B as $B = XY^*$, where X is $m \times k$ and Y is $k \times n$. Furthermore, this can be done with an X which has orthonormal columns. (One way to get this is from the SVD of B. We have $B = U_k \Sigma V_k$ where U_k is $m \times k$ with orthonormal columns and V_k is $k \times n$ with orthonormal rows, and Σ is diagonal. We can then put $X = U_k$ and $Y^* = \Sigma V_k$.)

Since X has orthonormal columns, $X^*X = I_k$ and

$$\begin{aligned} (A-B)^*(A-B) &= A^*A - A^*B - B^*A + B^*B \\ &= A^*A - A^*XY^* - YX^*A + YX^*XY^* \\ &= A^*A - A^*XY^* - YX^*A + YY^*. \end{aligned} \tag{8.6}$$

Now the key trick:

$$(Y - A^*X)(Y - A^*X)^* = YY^* - A^*XY^* - YX^*A + A^*XX^*A.$$

Hence,

$$-A^*XY^* - YX^*A = (Y - A^*X)(Y - A^*X)^* - YY^* - A^*XX^*A. \tag{8.7}$$

Substitute (8.7) in (8.6) to get

$$(A-B)^*(A-B) = A^*A + (Y - A^*X)(Y - A^*X)^* - A^*XX^*A. \tag{8.8}$$

Since $\text{trace}(Y - A^*X)(Y - A^*X)^* \ge 0$ and $\text{trace}(A^*XX^*A) = \text{trace}(X^*AA^*X)$, we have

$$\|A - B\|_F^2 \ge \text{trace}(A^*A) - \text{trace}(X^*AA^*X).$$

Let $\mathbf{x}_i$ denote column i of X; then the (i,i) entry of X^*AA^*X is $\|A^*\mathbf{x}_i\|^2$. Since $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are orthonormal, Lemma 8.8 tells us $\text{trace}(X^*AA^*X) \le \sum_{i=1}^{k} \sigma_i^2$, and so

$$\|A - B\|_F^2 \ge \text{trace}(A^*A) - \sum_{i=1}^{k} \sigma_i^2 = \sum_{i=k+1}^{r} \sigma_i^2. \qquad \square$$

Theorem 8.7 is for the Frobenius norm. Mirsky [**Mirs60**] showed that the result holds for any unitarily invariant norm. See also [**StewSun90**, Chapter IV].

In Section 6.8, we got the singular value decomposition for nonsingular square matrices by starting with the square root of a positive definite Hermitian matrix and then using the polar decomposition of a square matrix. We now reverse the process.

Suppose A is a square $n \times n$ matrix with singular value decomposition $A = U\Sigma_A V$. Insert $I = VV^*$ after the factor U to get $A = U\Sigma_A V = (UV)(V^*\Sigma_A V)$. Then $U_1 = UV$ is unitary and $P = V^*\Sigma_A V$ is Hermitian, positive semi-definite, so we obtain a polar factorization $A = U_1 P$. Recall that for a nonsingular A, the factors of the polar decomposition are uniquely determined by A. If A is singular, then $A^*A = P^2$ is positive semi-definite, and since A^*A has a unique positive semi-definite square root, we see the P is uniquely determined by A. The U, however, is not.

Finally, we mention an analogue of the Courant–Fischer minimax/maximin characterization (Theorem 6.18) of the eigenvalues of a Hermitian matrix for singular values.

Theorem 8.10. *Let A be an $m \times n$ matrix with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. Then*

$$\sigma_k = \max_{\mathcal{V}_k} \min_{\substack{\mathbf{x} \in \mathcal{V}_k \\ \mathbf{x} \neq 0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|},$$

where the maximum is over all k-dimensional subspaces $\mathcal{V}_k$ of $\mathbb{C}^n$. Also,

$$\sigma_k = \min_{\mathcal{V}_{n-k+1}} \max_{\substack{\mathbf{x} \in \mathcal{V}_{n-k+1} \\ \mathbf{x} \neq 0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|},$$

where the minimum is over all $(n-k+1)$-dimensional subspaces $\mathcal{V}_{n-k+1}$.

These may be used to obtain inequalities for singular values, such as the following analogue of Corollary 6.24.

Theorem 8.11. *Let A, B be $m \times n$ matrices. Let the singular values of A and B be $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ and $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_n$, respectively. Then $|\sigma_k - \tau_k| \leq \|A - B\|_2$.*

We have focused on proving the existence of the SVD. For numerical computation, one must consider the efficiency and stability of the algorithms used to compute the factorization $A = U\Sigma_A V$. It is generally not a good strategy to compute A^*A and then try to find the eigenvalues. We refer the reader elsewhere [**LawHan74, GvL89, Tref97, Demm97, Hou64**] for discussion of the issues involved and descriptions of algorithms. However, we will discuss one basic tool: using Householder transformations to obtain a QR factorization and to put a matrix in Hessenberg or tridiagonal form.

8.2. Householder Transformations

Householder transformations [**Hou64**] are orthogonal reflections across hyperplanes. Given two linearly independent vectors in $\mathbb{R}^n$ of the same length, there is a simple formula for constructing a Householder transformation which sends one to the other. These transformations provide another way to compute the QR factorization, presented in Chapter 2 as a byproduct of the Gram–Schmidt orthogonalization process. They are also useful for transforming matrices to special forms via unitary change of basis.

Those familiar with Lie algebras will recall that $\mathcal{M}_n$ with the product $[A, B]$ is a Lie algebra. This product is not associative, but does satisfy the *Jacobi identity*

$$[[A, B], C] + [[C, A], B] + [[B, C], A] = 0.$$

If a set of matrices $\mathcal{S}$ has Property P, then all of the eigenvalues of $[A, B]$ are zero for any $A, B \in \mathcal{S}$. So $[A, B]$ is nilpotent. Furthermore, for any matrices $C_1, \ldots, C_t$ in $\mathcal{S}$ and any polynomial $p(X_1, \ldots, X_t)$ in the noncommuting variables $X_1, \ldots, X_t$, the matrix $p(C_1, \ldots, C_t)[A, B]$ is nilpotent.

10.4. McCoy's Theorem

We now come to the main result of this chapter.

Theorem 10.20 (McCoy [**McCoy36**])**.** *Let $\mathcal{S}$ be a nonempty set of $n \times n$ matrices over an algebraically closed field $\mathcal{F}$. Then the following are equivalent.*

(1) *The set $\mathcal{S}$ is simultaneously triangularizable.*

(2) *The set $\mathcal{S}$ has Property P.*

(3) *For any matrices $A, B, C_1, \ldots, C_t$ in $\mathcal{S}$ and any polynomial $p(X_1, \ldots, X_t)$ in the noncommuting variables $X_1, \ldots, X_t$, the matrix $p(C_1, \ldots, C_t)[A, B]$ is nilpotent.*

We have already observed that (1) implies (2) and that (2) implies (3); the hard part of the proof is showing that (3) implies (1). This can be done in several ways. Again, we follow the approach in [**Flan56**]. We first need the following.

Theorem 10.21. *Let $\mathbb{F}$ be an algebraically closed field, let $\mathcal{V}$ be a vector space over $\mathbb{F}$ with $\dim(\mathcal{V}) > 1$, and let $\mathcal{A}$ be an algebra of linear transformations of $\mathcal{V}$. Suppose there is a nil ideal $\mathcal{B}$ of $\mathcal{A}$ such that $\mathcal{A}/\mathcal{B}$ is commutative. Then $\mathcal{V}$ has a proper $\mathcal{A}$-invariant subspace.*

Proof. If $\mathcal{B} = 0$, then $\mathcal{A}$ is commutative, and the result follows from Theorem 10.1. If $\mathcal{B} \neq 0$, then, as shown in the proof of Theorem 10.15, there exists a vector $\mathbf{v}$ such that $\mathcal{U} = \mathcal{B}\mathbf{v}$ is a proper $\mathcal{B}$-invariant subspace of $\mathcal{V}$. Since $\mathcal{B}$ is an ideal of $\mathcal{A}$, we have $\mathcal{A}\mathcal{B} \subseteq \mathcal{B}$, and so $\mathcal{A}\mathcal{U} = \mathcal{A}\mathcal{B}\mathbf{v} \subseteq \mathcal{B}\mathbf{v} = \mathcal{U}$. Hence the proper subspace $\mathcal{U}$ is $\mathcal{A}$-invariant. □

Proof of Theorem 10.20. We now complete the proof of McCoy's theorem by showing that (3) implies (1). We do this with Theorem 10.21, using condition (3) to obtain the nil ideal $\mathcal{B}$.

First, note that if $\mathcal{S}' = \mathcal{S} \cup \{I\}$, then each of the properties (1), (2), (3) holds for $\mathcal{S}$ if and only if it holds for $\mathcal{S}'$. So, without loss of generality, we may assume $\mathcal{S}$ includes the identity I. Let $\mathcal{A} = \mathcal{A}(\mathcal{S})$ be the algebra generated by $\mathcal{S}$. Assume $\mathcal{S}$ satisfies property (3). Note that for any pair of $n \times n$ matrices R, S, the matrix RS is nilpotent if and only if SR is nilpotent. Hence, $[A, B]p(C_1, \ldots, C_t)$ is nilpotent for any matrices $A, B, C_1, \ldots, C_t$ in $\mathcal{S}$. If we replace the elements $C_1, \ldots, C_t$ of $\mathcal{S}$ by elements $D_1, \ldots, D_t$ of $\mathcal{A}$, the resulting expression $p(D_1, \ldots, D_t)$ is still some polynomial in elements of $\mathcal{S}$. Therefore, every element of the ideal generated by $[A, B]$ in $\mathcal{A}$ is nilpotent. So $\mathcal{I}([A, B])$ is a nil ideal, and hence, by Corollary 10.17, is

nilpotent. Now let $\mathcal{R}$ be a maximal nilpotent ideal of $\mathcal{A}$. Then the sum $\mathcal{R}+\mathcal{I}([A,B])$ is also a nilpotent ideal; since $\mathcal{R}$ is maximal, we have $\mathcal{R}+\mathcal{I}([A,B]) \subseteq \mathcal{R}$, and so $\mathcal{I}([A,B]) \subseteq \mathcal{R}$. Hence, $[A,B] \in \mathcal{R}$. Consider the quotient $\mathcal{A}/\mathcal{R}$. Since $[A,B] \in \mathcal{R}$ for all $A, B \in \mathcal{S}$, the elements $A+\mathcal{R}$ and $B+\mathcal{R}$ commute in $\mathcal{A}/\mathcal{R}$. Since $\mathcal{S}$ generates $\mathcal{A}$, this means that $\mathcal{A}/\mathcal{R}$ is commutative. Hence, by Theorem 10.21, there is a proper $\mathcal{A}$-invariant subspace. We now proceed as in the proof of Theorem 10.7. Use the $\mathcal{A}$-invariant subspace to put the algebra $\mathcal{A}$ into block triangular form. The algebras formed by the diagonal blocks will satisfy condition (3), so one can use an induction argument and assume the blocks can be triangularized. Thus $\mathcal{A}$ has property (1). □

Those familiar with Lie algebras will recognize that this result is closely related to the theorems of Lie and Engel.

10.5. Property L

We now consider a weaker property, called Property L.

Definition 10.22. The $n \times n$ matrices A and B are said to have *Property L* if there is an ordering of the eigenvalues $\alpha_1, \ldots, \alpha_n$ of A and $\beta_1, \ldots, \beta_n$ of B such that for any scalars x and y, the eigenvalues of $xA + yB$ are $x\alpha_i + y\beta_i$ for $i = 1, \ldots, n$.

The set of all matrices of the form $xA + yB$ is called the *pencil* of A and B. Matrices which have Property P certainly have Property L, but in general, the converse is not true.

Example 10.23. This example comes from [**MT52**], the first of two papers on Property L by Motzkin and Taussky. Let $A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}$ and $B = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$. Then $xA+yB = \begin{pmatrix} 0 & x & 0 \\ y & 0 & -x \\ 0 & y & 0 \end{pmatrix}$ and $(xA+yB)^3 = 0$. Hence, for every x and y, all of the eigenvalues of $xA + yB$ are zero, so the pair A, B has Property L. However, the product $AB = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ has eigenvalues $1, -1$, and 0, so A and B do not have Property P.

In [**MT52, MT55**], Motzkin and Taussky establish significant results about the pair A, B, the pencil $xA+yB$, and Property L, and they use algebraic geometry to study the characteristic curve associated with the polynomial $\det(zI - xA - yB)$. Here are a few of the main results.

Theorem 10.24. *Let A and B be $n \times n$ matrices over an algebraically closed field $\mathbb{F}$ of characteristic p. Assume all the matrices in the pencil $xA+yB$ are diagonalizable, and, for $p \neq 0$, assume that $n \leq p$ or that A and B have Property L. Then A and B commute.*

The proof of Theorem 10.24 uses methods from algebraic geometry and is beyond our scope here. However, the next two results are proven with matrix theory tools.

Theorem 10.25 (Motzkin and Taussky [**MT52**])**.** *Let A and B be $n \times n$ matrices with Property L, and assume A is diagonalizable. Let $\alpha_1, \ldots, \alpha_n$ be the eigenvalues of A, listed so that repeated eigenvalues appear together; assume there are t distinct eigenvalues of multiplicities $m_1, \ldots, m_t$. Let $\beta_1, \beta_2, \ldots, \beta_n$ be the corresponding eigenvalues of B. Let $A' = P^{-1}AP$ be in Jordan form, and let $B' = P^{-1}BP$. Write*

$$B' = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1t} \\ B_{21} & B_{22} & \cdots & B_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ B_{t1} & B_{t2} & \cdots & B_{tt} \end{pmatrix},$$

where B_{ij} is size $m_i \times m_j$. Then $\det(zI - B') = \prod_{i=1}^{t} \det(zI - B_{ii})$ and $\sum b'_{ik}b'_{ki} = 0$, where the sum is over all $i < k$ for which the entry b'_{ik} lies outside every diagonal block B_{jj}.

Proof. Property L is not affected by a translation, so we may assume $\alpha_1 = 0$. The matrix A' is diagonal, and we write $A' = \begin{pmatrix} 0 & 0 \\ 0 & A_{22} \end{pmatrix}$ and $B' = \begin{pmatrix} B_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$, where A_{22} has size $(n - m_1) \times (n - m_1)$, the zeros indicate blocks of zeros, and B' is partitioned conformally with A. Note that A_{22} will be nonsingular. Consider the polynomial

$$\det(zI - xA' - B') = \det \begin{pmatrix} zI - B_{11} & -C_{12} \\ -C_{21} & zI - xA_{22} - C_{22} \end{pmatrix}$$

in x and z, and note that the coefficient of x^{n-m_1} is $\det(zI - B_{11}) \det(-A_{22})$. However, since A and B have Property L, we also have

$$\det(zI - xA' - B') = \prod_{i=1}^{m_1}(z - \beta_i) \prod_{i=m_1+1}^{n} (z - x\alpha_i - \beta_i).$$

From this we see the coefficient of x^{n-m_1} is $\prod_{i=1}^{m_1}(z-\beta_i) \prod_{i=m_1+1}^{n} (-\alpha_i)$. Equating these two expressions for the coefficient of x^{n-m_1} gives

$$\det(zI - B_{11}) \det(-A_{22}) = \prod_{i=1}^{m_1}(z - \beta_i) \prod_{i=m_1+1}^{n} (-\alpha_i).$$

But $\det(-A_{22}) = \prod_{i=m_1+1}^{n} (-\alpha_i)$, and this is nonzero, so $\det(zI - B_{11}) = \prod_{i=1}^{m_1}(z - \beta_i)$. Applying this argument to each of the eigenvalues of A yields

$$\det(zI - B') = \prod_{i=1}^{t} \det(zI - B_{ii}). \tag{10.3}$$

The second part of the theorem comes from examining the coefficient of z^{n-2} on both sides of (10.3). In each case, the coefficient of z^{n-2} is the sum of the

determinants of all 2×2 principal submatrices. The diagonal entries of these submatrices are the same for both sides, but for the off-diagonal entries, the left-hand side gives the sum of all products $b'_{ik}b'_{ki}$, where $i \neq k$, while on the right-hand side we have the sum of all such products where b'_{ik} is inside one of the diagonal blocks. So the sum of all such products where b'_{ik} comes from outside the diagonal blocks must be zero. □

One can think of the theorem as telling us that when A and B have Property L they retain some of the behavior of a pair of matrices which have Property P. We now use Theorem 10.25 to show that Hermitian matrices with Property L must commute.

Theorem 10.26 (Motzkin and Taussky)**.** *If A and B are Hermitian matrices with Property L, then $AB = BA$.*

Proof. Since A is Hermitian, we can reduce it to Jordan form with a unitary similarity P, so $A' = P^{-1}AP$ is diagonal and $B' = P^{-1}AP$ is still Hermitian. Hence, $b'_{ik}b'_{ki} = |b'_{ik}|^2$, and the second part of Theorem 10.25 tells us all of the off-diagonal blocks of B' are zero. Since the diagonal blocks of B' pair up with scalar blocks in A', we see that A' and B' commute, and hence so do A and B. □

Wielandt [**Wiel53**] generalized Theorem 10.26 to pairs of normal matrices. We prove this generalization using Theorem 5.12, which we restate and prove below.

Theorem 10.27. *Let A be a normal matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. Partition A into t^2 blocks, where the diagonal blocks $A_{11}, \ldots, A_{tt}$ are square. Suppose the direct sum of the diagonal blocks $A_{11} \oplus A_{22} \oplus \cdots \oplus A_{tt}$ has eigenvalues $\lambda_1, \ldots, \lambda_n$. Then $A_{ij} = 0$ when $i \neq j$, and so $A = A_{11} \oplus A_{22} \oplus \cdots \oplus A_{tt}$.*

Proof. Since A is normal, we have

$$\|A\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2 = \sum_{i=1}^{n} |\lambda_i|^2. \tag{10.4}$$

Let S denote the sum of the squares, $|a_{ij}|^2$, of the entries a_{ij} which are in the diagonal blocks, $A_{11}, \ldots, A_{tt}$. Since $A_{11} \oplus A_{22} \oplus \cdots \oplus A_{tt}$ has eigenvalues $\lambda_1, \ldots, \lambda_n$, we have $S \geq \sum_{i=1}^{n} |\lambda_i|^2$. But clearly, $\|A\|_F^2 \geq S$. Combine this with (10.4) to get

$$\sum_{i=1}^{n} |\lambda_i|^2 = \|A\|_F^2 \geq S \geq \sum_{i=1}^{n} |\lambda_i|^2.$$

Hence, we must have $\|A\|_F^2 = S$. This means all of the entries of A which are outside the diagonal blocks must be zeros, and hence $A_{ij} = 0$ when $i \neq j$. □

Theorem 10.28 (Wielandt [**Wiel53**])**.** *If A and B are $n \times n$ normal matrices with Property L, then $AB = BA$.*

Proof. Since A is normal, we can reduce it to Jordan form with a unitary similarity U, so $A' = U^*AU$ is diagonal and $B' = U^*BU$ is still normal. From Theorem 10.25, we have $\det(zI - B') = \prod_{i=1}^{t} \det(zI - B_{ii})$, and so B' has the same eigenvalues as

$B_{11} \oplus B_{22} \oplus \cdots \oplus B_{tt}$. Theorem 10.27 then tells us the off-diagonal blocks of B' are zero. The diagonal blocks of B' pair up with scalar blocks in A'. So A' and B' commute, and hence so do A and B. □

Exercises

1. Let $\mathcal{A}$ be a nonempty set of linear transformations of an n-dimensional vector space $\mathcal{V}$ over a field $\mathbb{F}$, and let $\mathcal{U}$ be an $\mathcal{A}$-invariant subspace of $\mathcal{V}$. For $T \in \mathcal{A}$, show that we can define an action of T on the quotient space $\mathcal{V}/\mathcal{U}$ by defining $T(\mathbf{v}+\mathcal{U}) = T\mathbf{v}+\mathcal{U}$. (The main thing you need to show here is that this is well defined.)
2. Find a pair of upper-triangular matrices which do not commute.
3. For an $n \times n$ matrix A, let $\mathcal{C}(A)$ denote the set of matrices which commute with A. Thus, $\mathcal{C}(A) = \{B \in \mathcal{M}_n \big| AB = BA\}$.
   (a) What is $\mathcal{C}(I)$?
   (b) Show that $\mathcal{C}(A)$ is a subspace of $\mathcal{M}_n$.
   (c) Show that if $p(x)$ is any polynomial, then $p(A) \in \mathcal{C}(A)$.
   (d) Revisit part (a) to see that there can be matrices in $\mathcal{C}(A)$ which are not polynomials in A.
   (e) Show that if the matrix A is a single Jordan block $A = J_p(\lambda)$, then every matrix in $\mathcal{C}(A)$ is a polynomial in A. In this case, what is the dimension of $\mathcal{C}(A)$?
   Hint: Note that B commutes with $J_p(\lambda)$ if and only if B commutes with the nilpotent matrix $N = J_p(0)$, so it suffices to find $\mathcal{C}(N)$.
   (f) Find a matrix which commutes with $A = J_2(0) \oplus J_1(0)$ which is not a polynomial in A.
4. Give the details of the proof of Theorem 10.16.
5. Show that if $\mathcal{B}$ and $\mathcal{C}$ are two nilpotent ideals of $\mathcal{A}$, then the sum $\mathcal{B}+\mathcal{C}$ is also a nilpotent ideal.
6. Let $\mathcal{A}$ be an associative algebra. For $x, y \in \mathcal{A}$, define $[x, y] = xy - yx$. Show this Lie bracket product satisfies the Jacobi identity
$$[[x, y], z] + [[z, x], y] + [[y, z], x] = 0.$$

have r choices for $\mathcal{B}$. Now, $\mathcal{B}$ has k points, one of which is p, leaving $(k-1)$ choices for the point y. So, the number of 2-flags of the form $(p, y, \mathcal{B})$ must be $r(k-1)$. Hence, $\lambda(v-1) = r(k-1)$. Note that this argument shows that $r = \frac{\lambda(v-1)}{(k-1)}$, and hence, the number r does not depend on the choice of the point p. □

We now generalize the process to prove Theorem 13.5.

Proof. Let $0 \le s \le t$, and let $\mathcal{R}_s$ be a fixed set of s points. Let m denote the number of blocks $\mathcal{B}$ which contain $\mathcal{R}_s$. We will show that m depends only on s and not on the choice of the set $\mathcal{R}_s$ by showing $m\binom{k-s}{t-s} = \lambda\binom{v-s}{t-s}$. This will prove Theorem 13.5 with $m = \lambda_s$.

Define an *admissible pair* $(\mathcal{T},\ \mathcal{B})$ to be a t-set $\mathcal{T}$ and a block $\mathcal{B}$ such that $\mathcal{R}_s \subseteq \mathcal{T} \subseteq \mathcal{B}$. We count the number of admissible pairs in two different ways.

First, the number of ways to choose $(t-s)$ points out of the $(v-s)$ points which are not in $\mathcal{R}_s$ is $\binom{v-s}{t-s}$, so this gives the number of t-sets $\mathcal{T}$ such that $\mathcal{R}_s \subseteq \mathcal{T}$. There are then exactly λ blocks $\mathcal{B}$ which contain $\mathcal{T}$. So, the number of admissible pairs is $\lambda\binom{v-s}{t-s}$.

On the other hand, there are m blocks $\mathcal{B}$ such that $\mathcal{R}_s \subseteq \mathcal{B}$. For each such block $\mathcal{B}$ there are $\binom{k-s}{t-s}$ ways to choose $(t-s)$ points of $\mathcal{B}$ which are not in $\mathcal{R}_s$, which gives $\binom{k-s}{t-s}$ subsets $\mathcal{T}$ of size t such that $\mathcal{R}_s \subseteq \mathcal{T} \subseteq \mathcal{B}$. So, the number of admissible pairs is $m\binom{k-s}{t-s}$. Hence, $m\binom{k-s}{t-s} = \lambda\binom{v-s}{t-s}$, and the number $m = \lambda_s$ depends only on s and not on the particular set $\mathcal{R}_s$. □

Corollary 13.9. *Let $\mathcal{D}$ be a t-(v, k, λ) design. Let $b = \lambda_0$ be the number of blocks, and let $\lambda_1 = r$ be the number of blocks which contain a point x. Then*

(1) $b = \lambda_0 = \lambda\ \dfrac{v(v-1)(v-2)\cdots(v-t+1)}{k(k-1)(k-2)\cdots(k-t+1)}$.

(2) $bk = rv$.

(3) *If $t > 1$, then $r(k-1) = \lambda_2(v-1)$.*

Example 13.10. Suppose $\mathcal{D}$ is a k-$(v, k, 1)$ design. Then

$$b = \frac{v(v-1)(v-2)\cdots(v-k+1)}{k(k-1)(k-2)\cdots(k-k+1)} = \frac{v(v-1)(v-2)\cdots(v-k+1)}{k!} = \binom{v}{k}.$$

So $\mathcal{D}$ is the "trivial" design of all k-subsets of the v-set.

13.2. Incidence Matrices for 2-Designs

The term *block design* is often used for 2-designs. The design consisting of all k-subsets of a v-set is the trivial 2-design; we are interested in the case where not every k-subset is a block. These are called *balanced incomplete block designs* or BIBDs.

Let A be the incidence matrix of a 2-(v, k, λ) design having b blocks and $\lambda_1 = r$. The matrix A is size $b \times v$. Each row of A corresponds to a block of the design

and contains exactly k ones. Each column corresponds to a point; since each point belongs to exactly r blocks, each column of A has r ones. Furthermore, the inner product of columns i and j of A counts the number of rows which have ones in both positions i and j, and hence gives the number of blocks which contain both of the points x_i and x_j. So the inner product of columns i and j of A is λ when $i \neq j$, and it is r when $i = j$. These facts can be stated as the following three matrix equations:

$$\begin{aligned} AJ_v &= kJ_{b\times v}, \\ J_bA &= rJ_{b\times v}, \\ A^TA &= (r-\lambda)I + \lambda J_v. \end{aligned}$$

These equations will be the key in proving several important results about 2-designs. We will need the following facts, which appeared in earlier chapters.

- The determinant of the matrix $aI + bJ_n$ is $(a+nb)a^{n-1}$ (see Section 11.1).
- For any matrix A, the matrices A, A^T, AA^T, and A^TA all have the same rank.

Consider a 2-(v,k,λ) design with b blocks and $\lambda_1 = r$. The case $k = v$ simply gives b repeated copies of the full v-set $\{x_1,\ldots,x_v\}$, which is of little interest. Hence, we shall assume $k \leq v-1$. Recall the equations

$$\begin{aligned} bk &= vr, \\ r(k-1) &= \lambda(v-1). \end{aligned}$$

Since $k-1 < v-1$, the second equation gives $r > \lambda$. We now prove Fischer's inequality, $b \geq v$.

Theorem 13.11 (Fischer's inequality)**.** *Let b be the number of blocks in a 2-(v,k,λ) design $\mathcal{D}$. If $k \leq v-1$, then $b \geq v$.*

Proof. Let A be the incidence matrix for the design. Then A is $b \times v$ and satisfies $A^TA = (r-\lambda)I + \lambda J_v$. Now,

$$\det\big((r-\lambda)I + \lambda J_v\big) = (r-\lambda)^{v-1}(r-\lambda+v\lambda).$$

But $r-\lambda+v\lambda = r+\lambda(v-1) = r+r(k-1) = rk$. So, $\det(A^TA) = (r-\lambda)^{v-1}rk$. Since $k \leq v-1$, we know $r > \lambda$, and hence $\det(A^TA) \neq 0$. Therefore, the rank of A^TA is v, and so A also has rank v. Since A has b rows, we must have $b \geq v$. □

Definition 13.12. A 2-(v,k,λ) design with $k \leq v-1$ is called *symmetric* if $b = v$.

In a symmetric BIBD, the number of blocks equals the number of points. The projective plane of order two, Example 13.3, is a symmetric block design.

Theorem 13.13. *In a 2-(v,k,λ) design with $k \leq v-1$, the following are equivalent:*

(1) $b = v$.

(2) $r = k$.

(3) *Any pair of blocks have exactly λ points in common.*

Proof. The equation $bk = vr$ shows the equivalence of (1) and (2).

Now suppose $b = v$. Then the incidence matrix A of the design is square. Since we also have $r = k$, we get $AJ = kJ = rJ = JA$. (Note both A and J are $v \times v$.) So, A commutes with J. But then A commutes with $(r - \lambda)I + \lambda J = A^T A$, and so $A(A^T A) = (A^T A)A$. From the proof of Fisher's inequality we know A has rank v and hence is invertible. Multiply both sides of $A(A^T A) = (A^T A)A$ on the right by A^{-1} to get $AA^T = A^T A = (r - \lambda)I + \lambda J$. Since entry i, j of AA^T is the dot product of rows i and j of A, the number of points in the intersection of blocks i and j is λ when $i \neq j$.

Conversely, suppose any pair of blocks intersect in exactly λ points. Then, since each block has k points, $AA^T = (k - \lambda)I + \lambda J$. If $(k - \lambda) = 0$, then $AA^T = \lambda J$, and so A has rank one. But we know A has rank v and $v > 1$. So $(k - \lambda) \neq 0$ and

$$\det(AA^T) = (k - \lambda)^{b-1}(k - \lambda + \lambda b) \neq 0.$$

So AA^T has rank b, and hence $b = v$. □

In summary, in a symmetric block design, the number of blocks equals the number of points. Each pair of points belongs to exactly λ blocks, and each pair of blocks intersects in exactly λ points. Each block has k points, and each point is in exactly k blocks. We then have

$$A^T A = AA^T = (k - \lambda)I + \lambda J. \tag{13.2}$$

Theorem 13.13 enables us to set up a dual design by reversing the roles of points and blocks. The points of the dual design correspond to the blocks of the original design. If x is a point in the original design, the set of blocks which contain x becomes a block in the dual design. If A is the incidence matrix of the original design, then A^T is the incidence matrix of the dual design.

13.3. Finite Projective Planes

The last section concluded with a comment on duality. This concept is more typically encountered in projective geometry.

Definition 13.14. A *finite projective plane* is a symmetric 2-$(v, k, 1)$ design. The *points* of the plane are the elements of the v-set, and the *lines* of the plane are the blocks of the design.

In a finite projective plane, we have $\lambda = 1, b = v$, and $r = k$. Each pair of points is contained in exactly one line, and each pair of lines intersects in exactly one point. The equation $r(k - 1) = \lambda(v - 1)$ then becomes $k(k - 1) = (v - 1)$, and so $v = k^2 - k + 1$. We also have $v = (k - 1)^2 + (k - 1) + 1$; the number $(k - 1)$ is called the *order* of the projective plane. Example 13.3 is a projective plane of order two; it is, in fact, the only projective plane of order two, so we may call it "the" projective plane of order two.

A major question in this area is, For which numbers $n = k - 1$ do there exist projective planes of order n? There is a method for constructing a projective plane

Hence, the sequence $\{\mathbf{p}_k\}_{k=1}^{\infty}$ has a convergence subsequence, $\mathbf{p}_{\nu_1}, \mathbf{p}_{\nu_2}, \mathbf{p}_{\nu_3}, \dots$. Let $\mathbf{p} = \lim_{i\to\infty} \mathbf{p}_{\nu_i}$. We have $\mathbf{p} \geq 0$ and $\mathbf{e}^T\mathbf{p} = 1$, so $\mathbf{p} \neq 0$. Then,

$$A\mathbf{p} = \lim_{i\to\infty} A_{\nu_i}\mathbf{p}_{\nu_i} = \lim_{i\to\infty} \rho_{\nu_i}\mathbf{p}_{\nu_i} = \mu\mathbf{p}.$$

So μ is an eigenvalue of A. Hence, $\mu \leq \rho$. But we already had $\mu \geq \rho$, so $\mu = \rho$. Therefore, ρ is an eigenvalue of A and the nonnegative vector $\mathbf{p}$ is an associated eigenvalue.

Part (3) follows by noting that $A^T \geq 0$ and $\rho(A) = \rho(A^T)$. □

17.5. Irreducible Matrices

In the last section, we saw that only some parts of Perron's theorem for positive matrices hold for general nonnegative matrices. We now see that parts (2) and (3) of Theorem 17.16 hold for irreducible matrices.

Recall the following Definition 16.3 from Chapter 16.

Definition 17.20. An $n \times n$ matrix A is said to be *reducible* if there exists a permutation matrix P such that $P^TAP = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$, where A_{11} is $k \times k$, A_{22} is $(n-k)\times(n-k)$, and $1 \leq k \leq n-1$. If A is not reducible, we say A is *irreducible*.

Note that reducibility is determined by the location of the zero and nonzero entries of the matrix. In Chapter 16, we saw that A is irreducible if and only if the directed graph $\mathcal{D}(A)$ is strongly connected (Theorem 16.4). We then used graphs to prove the following result (Theorem 16.5).

Theorem 17.21. *Let A be an $n \times n$ nonnegative matrix. Then A is irreducible if and only if $(I + A)^{n-1} > 0$.*

Remark 17.22. For a matrix with negative entries, note that A is irreducible if and only if $|A|$ is irreducible. So this theorem is equivalent to the statement that A is irreducible if and only if $(I+|A|)^{n-1} > 0$. Theorem 17.21 can also be proven with matrix methods, rather than by using the directed graph $\mathcal{D}(A)$. See Exercise 9 for this matrix approach.

When a nonnegative matrix is irreducible, the Perron root is a simple root and the Perron eigenvector is positive.

Theorem 17.23 (Frobenius [**Frob12**])**.** *If A is an $n \times n$, nonnegative, irreducible matrix, then the following hold.*

(1) *The spectral radius $\rho = \rho(A)$ of A is positive and is an eigenvalue of A.*

(2) *The eigenvalue ρ is a simple root of the characteristic polynomial $p_A(x)$ of A.*

(3) *There is a positive vector $\mathbf{p}$ such that $A\mathbf{p} = \rho\mathbf{p}$. Furthermore, $A\mathbf{x} = \rho\mathbf{x}$ if and only if $\mathbf{x}$ is a scalar multiple of $\mathbf{p}$.*

(4) *There is a positive vector $\mathbf{q}$ such that $\mathbf{q}^TA = \rho\mathbf{q}^T$, and $\mathbf{y}^TA = \rho\mathbf{y}^T$ if and only if $\mathbf{y}$ is a scalar multiple of $\mathbf{q}$.*

Proof. Since $A \geq 0$, we know there exists a nonzero, nonnegative vector $\mathbf{p}$ such that $A\mathbf{p} = \rho\mathbf{p}$. Let $B = (I + A)^{n-1}$. Since A is irreducible, $B > 0$. Also, $B\mathbf{p} = (I + A)^{n-1}\mathbf{p} = (1 + \rho)^{n-1}\mathbf{p}$. Part (5) of Theorem 17.16 then tells us that $\mathbf{p} > 0$. But then $A\mathbf{p} \neq 0$, and so ρ cannot be zero. So $\rho > 0$.

Now, let $\lambda_1 = \rho, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of A. The eigenvalues of B are $(1 + \rho)^{n-1}, (1 + \lambda_2)^{n-1}, \ldots, (1 + \lambda_n)^{n-1}$, with $(1 + \rho)^{n-1} = \rho(B)$. If ρ were a multiple root of $p_A(x)$, then $(1+\rho)^{n-1}$ would be a multiple root of $p_B(x)$, which is not possible by Perron's Theorem 17.16 for positive matrices. Hence, ρ is a simple root of $p_A(x)$. So, we have established the first three parts. Part (4) follows from the fact that A^T is irreducible if and only if A is irreducible. (This may not be obvious directly from the definition, but it follows easily from Theorem 17.21.) □

Theorem 17.23 says nothing about $\lim_{k\to\infty} A^k$. Observe that the n-cycle matrix in Example 17.18 is an irreducible, nonnegative matrix with spectral radius 1 for which this limit does not exist.

17.6. Primitive and Imprimitive Matrices

We now obtain more information about the eigenvalues of irreducible, nonnegative matrices. The structure of the set of eigenvalues depends on the zero-nonzero structure of the matrix, i.e., on the directed graph associated with the matrix. If A is a nonnegative, irreducible matrix, then the associated directed graph $\mathcal{D}(A)$ is strongly connnected. In Chapter 16, we defined the terms *primitive* and *index of imprimitivity* for strongly connected directed graphs (Definition 16.10). We extend these definitions to nonnegative, irreducible matrices by using the associated digraph. There is also another way to define the index of imprimitivity of a matrix; we discuss this alternative definition later and will see it is equivalent to the graph definition.

Definition 17.24. Let A be an $n \times n$, irreducible, nonnegative matrix. We say k is the *index of imprimitivity* of A if k is the index of imprimitivity of the associated directed graph $\mathcal{D}(A)$. If $k = 1$, we say A is *primitive*.

Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of an $n \times n$ matrix A where repeated eigenvalues are listed according to their multiplicities. For any positive integer k, the eigenvalues of A^k are ${\lambda_1}^k, {\lambda_2}^k, \ldots, {\lambda_n}^k$ and $\rho(A^k) = [\rho(A)]^k$.

Theorem 17.25. *Let A be a nonnegative, irreducible, primitive matrix. Then the eigenvalue $\rho = \rho(A)$ is a simple eigenvalue, and is the only eigenvalue of modulus ρ.*

Proof. From Theorem 17.23, we know ρ is a simple eigenvalue. Since A is primitive, Theorem 16.16 tells us there is a positive integer M such that $A^M > 0$. By Perron's theorem for positive matrices, ρ^M is a simple eigenvalue of A^M, and every other eigenvalue of A^M has modulus less than ρ^M. Hence, if λ is an eigenvalue of A and $\lambda \neq \rho$, we must have $|\lambda|^M < \rho^M$ and so $|\lambda| < \rho$. □

We now reap the benefits of the work done in Chapter 16 on the structure of imprimitive graphs, and in Section 11.3 on the eigenvalues of block cycle matrices.

Chapter 18

Error-Correcting Codes

When information is transmitted over a communications channel, errors may be introduced. The purpose of error-correcting codes is to enable the recipient to detect and correct errors. The theory of error-correcting codes involves linear algebra, finite field theory, block designs, and other areas of combinatorics and graph theory. This chapter is a brief introduction to give some idea of how linear algebra is used to construct binary linear codes. We consider a message to be a string of zeros and ones, and work in a vector space over the binary field $\mathbb{Z}_2$. More generally, one can work with an alphabet of q distinct symbols. If $q = p^r$ is a power of a prime number p, then the set of alphabet symbols can be the elements of the finite field $GF(q)$ We refer the reader to books on error correcting codes, such as [**Berle68, MacSlo77, Hamming80, CVL80, VanLi82**], for comprehensive accounts.

18.1. Introduction

Consider a message given in the form of a string of zeros and ones. When the message is sent via a communications channel, errors may occur, either from random noise or other sorts of error. It is possible that you send a zero, but your recipient receives a one, or vice versa. A simple way to protect against error is by using repetition. You could send a block of two zeroes (0 0) for each zero, and a block of two ones (1 1) for each one. Should noise distort one of the bits, so that the receiver gets (1 0) or (0 1), she knows an error has occurred and can ask you to resend the message. Of course, if both bits get changed by errors, the receiver cannot detect this. Thus, this scheme allows the receiver to detect one error but not two. Increasing the number of repetitions will increase the number of detectable errors and also enable some error correction. For example, sending (000) for each zero and (111) for each one enables the receiver to detect and correct a single error—for example if the string (010) arrives, the receiver would assume the middle digit was wrong and decode the message as zero. However, if two digits are changed by error, then the message will be incorrectly decoded. More generally, if you send a block of

k zeroes, $(000\cdots0)$, for each zero in the message, and a block of k ones, $(111\cdots1)$, for each one in the message, and the decoding algorithm is "majority rule", then up to $\frac{(k-1)}{2}$ errors can be corrected when k is odd. If k is even, then up to $\frac{(k-2)}{2}$ errors can be corrected, while $\frac{k}{2}$ errors can be detected but not corrected. Increasing the block length enables correction of more errors. The cost is transmission rate; you are using k bits for each bit of the message.

Now consider the string of zeros and ones as separated into words with k bits in each; thus, each message word is a block of k digits; each digit is either zero or one. We can view each block as a vector with k coordinates over the binary field $\mathbb{Z}_2$. We write the k-digit block as a vector, $\mathbf{x} = (x_1, \ldots, x_k)$. The number of possible words of this type is 2^k. We now adjoin an extra bit x_{k+1} defined by

$$x_{k+1} = \sum_{i=1}^{k} x_i. \tag{18.1}$$

Since we are working in the binary field $\mathbb{Z}_2$, the sum is 0 when $(x_1, \ldots, x_k)$ has an even number of ones and is 1 when there are an odd number of ones; hence this additional bit x_{k+1} is called a "parity check". We adjoin the digit x_{k+1} to the original block $(x_1, \ldots, x_k)$ and transmit the codeword $(x_1, \ldots, x_k, x_{k+1})$. Since equation (18.1) is equivalent to the equation

$$x_1 + \cdots + x_k + x_{k+1} = 0,$$

the code words are those vectors in $\mathbb{Z}_2^{k+1}$ which satisfy the linear equation

$$\sum_{i=1}^{k+1} x_i = 0.$$

The set of code words is then a k-dimensional subspace of $\mathbb{Z}_2^{k+1}$; it is, in fact, the null space of the $1 \times (k+1)$ matrix $A = (\,1 \quad 1 \quad \cdots \quad 1\,)$. This is an example of a linear code. The receiver checks each block of $(k+1)$ digits by summing the digits. If they sum to zero, it is assumed no errors were made; if they sum to one it is assumed one error was made. This code can detect a single error, but we will not know which coordinate of the block has the error. However, only one extra bit is needed to get this error detection capability. We are going to generalize this procedure to construct codes which can detect more errors, and also correct errors. The idea is to use more linear equations as checks.

18.2. The Hamming Code

Our story begins in the late 1940s when Richard Hamming created a family of single error-correcting codes, (published in [**Hamming50**]). (These codes also appeared in papers by Shannon (1948) and Golay (1949). See [**Berle68**, page 8]). Frustrated by the shut-down of long computer programs running on weekends (when no operators were present to restart a program shut-down due to a detected error), Hamming wanted a system which could not only detect an error, but correct it. The $[7, 4]$ Hamming code has four message bits and three parity check bits. Denoting the message word as (x_1, x_2, x_3, x_4), we adjoin three additional check bits x_5, x_6, x_7

defined by the equations

$$\begin{array}{rcccccccc} x_5 & = & x_1 & + & x_2 & & & + & x_4 \\ x_6 & = & x_1 & + & & & x_3 & + & x_4 \\ x_7 & = & & & x_2 & + & x_3 & + & x_4. \end{array} \tag{18.2}$$

Since $+1$ and -1 are the same in $\mathbb{Z}_2$, we may rewrite the system (18.2) as

$$\begin{array}{ccccccccccccc} x_1 & + & x_2 & + & & & x_4 & + & x_5 & & & = & 0 \\ x_1 & + & & & x_3 & + & x_4 & + & & x_6 & & = & 0 \\ & & x_2 & + & x_3 & + & x_4 & + & & & x_7 & = & 0. \end{array}$$

Rewriting this system of three linear equations in seven unknowns in matrix form, we see that the codewords are the solutions to

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix} \mathbf{x} = \mathbf{0}. \tag{18.3}$$

Now put

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

and

$$A = (\, C \;\; | \;\; I_3 \,).$$

Equation (18.3) then becomes $A\mathbf{x} = 0$. The 3×7 matrix A has rank 3, so it has a four-dimensional nullspace. The 16 codewords, formed from the 16 possible messages using equations (18.2), are exactly the vectors in this nullspace. The set of codewords forms a four-dimensional subspace of $\mathbb{Z}_2^7$.

How does the decoding work? Suppose we send the codeword $\mathbf{x}$, but, due to error, the receiver gets $\mathbf{y}$. Put $\mathbf{e} = \mathbf{y} - \mathbf{x}$; equivalently, $\mathbf{y} = \mathbf{x} + \mathbf{e}$. The vector $\mathbf{e}$ is the *error vector*. A 1 in entry j of $\mathbf{e}$ means that the jth-received entry is wrong. Since $\mathbf{x}$ is a codeword, $A\mathbf{x} = 0$, and $A\mathbf{y} = A\mathbf{x} + A\mathbf{e} = A\mathbf{e}$. When no errors occur, $\mathbf{y} = \mathbf{x}$ and $A\mathbf{y} = 0$. Now, suppose exactly one error occurs, and that it is in entry j. Then $A\mathbf{e}$ will be the jth column of the matrix A. Here is the key point: the seven columns of A are all different; in fact they are the seven nonzero binary vectors with three coordinates. If only one error occurs, computing $A\mathbf{y}$ will tell us exactly where the error was—find the column of A which matches $A\mathbf{y} = A\mathbf{e}$, and the position of that column is the position of the error. Hence, this code can detect *and correct* a single error in any position. It uses a 7-bit code word for a 4-bit message. To achieve this error correction by simply repeating each message bit three times would require 12 bits for the 4-bit message.

18.3. Linear Codes: Parity Check and Generator Matrices

With this famous example under our belt, we move on to some general definitions. We start with linear equations and matrices and will see that this leads to a more abstract definition of a linear code as a subspace of $\mathbb{Z}_2^n$. For the remainder of this chapter we use the term *length* of a vector $\mathbf{x}$ to mean the number of coordinates of $\mathbf{x}$. So, vectors in $\mathbb{Z}_2^n$ have length n.

Clearly, a nonnegative matrix A and a nonnegative vector $\mathbf{b}$ will yield a nonnegative system. Happily, it turns out that this is the only way we can have a nonnegative system.

Theorem 19.14. *The system* $\mathbf{x}(k+1) = A\mathbf{x}(k) + \mathbf{b}$ *is nonnegative if and only if* $A \geq 0$ *and* $\mathbf{b} \geq 0$.

Proof. As already noted, if A and $\mathbf{b}$ are both nonnegative, then the system is clearly nonnegative. We need to prove the converse. So, suppose $\mathbf{x}(k+1) = A\mathbf{x}(k) + \mathbf{b}$ is nonnegative. If we choose $\mathbf{x}(0) = 0$, then $\mathbf{x}(1) = \mathbf{b}$, so we must have $\mathbf{b} \geq 0$. Now, suppose the matrix A has a negative entry in position i, j. Let c be a positive constant large enough that $c|a_{ij}| > b_i$. Since $a_{ij} < 0$, we then have $c(a_{ij}) + b_i < 0$. Choose $\mathbf{x}(0) = c(\mathbf{e}_j)$ where $\mathbf{e}_j$ is the jth unit coordinate vector. Then $\mathbf{x}(1) = A\mathbf{x}(0) + \mathbf{b} = c(\text{column } j \text{ of } A) + \mathbf{b}$, so $\mathbf{x}(1)$ has a negative entry in the ith coordinate. This contradicts the fact that the system is nonnegative. Hence, every entry of A must be nonnegative. □

For nonnegative systems, we are interested in knowing if there is a nonnegative equilibrium point. It turns out that the criterion for a nonnegative equilibrium point is the same as that for stability.

Theorem 19.15. *Suppose* $\mathbf{x}(k+1) = A\mathbf{x}(k) + \mathbf{b}$ *is a nonnegative system and* $\mathbf{b} > 0$. *Then there is a nonnegative equilibrium point if and only if* $\rho(A) < 1$.

Proof. Suppose $\rho(A) < 1$. Theorem 19.8 then tells us there is a stable equilibrium point; furthermore, the equilibrium point $\bar{\mathbf{x}}$ is unique and given by the formula $\bar{\mathbf{x}} = (I - A)^{-1}\mathbf{b}$. Since $\rho(A) < 1$, the geometric series $\sum_{j=0}^{\infty} A^j$ converges to $(I - A)^{-1}$ and $\bar{\mathbf{x}} = \sum_{j=0}^{\infty} A^j \mathbf{b}$. Since $A \geq 0$ and $\mathbf{b} \geq 0$, it is clear that $\bar{\mathbf{x}} \geq 0$.

Conversely, suppose there is a nonnegative equilibrium point, that is, a nonnegative vector $\bar{\mathbf{x}}$ such that $\bar{\mathbf{x}} = A\bar{\mathbf{x}} + \mathbf{b}$. Since $A \geq 0$, the Perron–Frobenius theorem tells us that $\rho = \rho(A)$ is an eigenvalue of A; furthermore, there is a corresponding nonnegative left eigenvector $\mathbf{v}^T$. We then have

$$\mathbf{v}^T\bar{\mathbf{x}} = \mathbf{v}^T A\bar{\mathbf{x}} + \mathbf{v}^T\mathbf{b} = \rho(\mathbf{v}^T\bar{\mathbf{x}}) + \mathbf{v}^T\mathbf{b}.$$

Rearrange this to get

$$(1 - \rho)\mathbf{v}^T\bar{\mathbf{x}} = \mathbf{v}^T\mathbf{b}. \tag{19.34}$$

Since $\mathbf{v}$ is nonnegative and is not the zero vector and $\mathbf{b} > 0$, we must have $\mathbf{v}^T\mathbf{b} > 0$. Since we also have $\bar{\mathbf{x}} \geq 0$, equation (19.34) tells us $1 - \rho > 0$, hence $\rho = \rho(A) < 1$. □

Now consider a continuous time system, $\mathbf{x}'(t) = A\mathbf{x}(t) + \mathbf{b}$. We first want to find the conditions on A and $\mathbf{b}$ which guarantee that the system preserves nonnegativity of the state vector. We clearly must have $\mathbf{b} \geq 0$, because if the initial state $\mathbf{x}(0)$ is the zero vector, then $\mathbf{x}'(0) = \mathbf{b}$ must be nonnegative to preserve nonnegativity. However, it turns out that it is not necessary that all of the entries of A be nonnegative, but only the off-diagonal entries.

Definition 19.16. A square matrix A is a *Metzler matrix* if $a_{ij} \geq 0$ for all $i \neq j$.

Theorem 19.17. *The continuous time system* $\mathbf{x}'(t) = A\mathbf{x}(t) + \mathbf{b}$ *preserves nonnegativity if and only if* A *is a Metzler matrix and* $\mathbf{b} \geq 0$.

Proof. Suppose the system preserves nonnegativity. Then when a nonnegative state vector $\mathbf{x}(t)$ has a zero in the ith coordinate, we must have $x_i'(t) \geq 0$ in order for that coordinate to remain nonnegative. If we start at $\mathbf{x}(0) = 0$, then $\mathbf{x}'(0) = \mathbf{b}$, so we must have $\mathbf{b} \geq 0$. Next, consider the i, j entry of A, where $i \neq j$. Suppose a_{ij} is negative. Choose a positive constant c such that $ca_{ij} + b_i < 0$. Let $\mathbf{x}(0) = c\mathbf{e}_j$. Entry i of $\mathbf{x}(0)$ is then 0, while entry i of $\mathbf{x}'(0)$ is $ca_{ij} + b_i < 0$, which contradicts the assumption that the system preserves nonnegativity. Therefore, for $i \neq j$, we have $a_{ij} \geq 0$, and A is a Metzler matrix.

Conversely, suppose A is a Metzler matrix and $\mathbf{b} \geq 0$. If we start at a nonnegative vector $\mathbf{x}(0)$, then the trajectory, $\mathbf{x}(t)$ will stay in the nonnegative region of $\mathbb{R}^n$ unless there is some value of t and some coordinate i such that $x_i(t) = 0$ but $x_i'(t) < 0$; that is, unless the trajectory reaches a boundary point of the region and the derivative of the boundary coordinate is negative. So, suppose $\mathbf{x}(t) \geq 0$ with $x_i(t) = 0$ for some t and some coordinate i. Since A is a Metzler matrix, the ith coordinate of $A\mathbf{x}(t)$ will then be nonnegative, because the zero in the ith coordinate of $\mathbf{x}(t)$ will pair up with the ith diagonal entry of A and thus $x_i'(t) \geq 0$. Hence, $A\mathbf{x}(t) + \mathbf{b} \geq 0$ and the system preserves nonnegativity. □

There is a close relationship between Metzler matrices and nonnegative matrices. If A is a Metzler matrix, then for a sufficiently large positive constant c the matrix $P = cI + A$ is nonnegative. Therefore, P has a Perron–Frobenius eigenvalue $\lambda_0 = \rho(P)$ with a corresponding nonnegative eigenvector $\mathbf{v}_0$. We then have $A\mathbf{v}_0 = (\lambda_0 - c)\mathbf{v}_0$, so $\mathbf{v}_0$ is an eigenvector of A corresponding to the eigenvalue $\lambda_0 - c$. Set $\mu_0 = \lambda_0 - c = \rho(P) - c$. Then μ_0 is real. Since the eigenvalues of A are obtained from those of P by shifting them distance c to the left, we can take the circle of radius $\rho(P)$ centered at the origin, and shift it c units to the left to obtain a circle centered at $-c$ with radius $\rho(P)$ which contains the eigenvalues of A. This shifted circle is then tangent to the vertical line $x = \mu_0$, which tells us that μ_0 is the eigenvalue of A with largest real part.

Theorem 19.18. *Let* A *be a Metzler matrix. Then* A *has a real eigenvalue* μ_0 *which satisfies the following.*

(1) *There is a nonnegative eigenvector* $\mathbf{v}_0$ *corresponding to* μ_0.

(2) *If* $\mu \neq \mu_0$ *is any other eigenvalue of* A, *then* $\Re(\mu) < \mu_0$.

The next result is the analogue of Theorem 19.15 for continuous systems.

Theorem 19.19. *Let* A *be a Metzler matrix, and let* $\mathbf{b} > \mathbf{0}$. *Then the system* $\mathbf{x}'(t) = A\mathbf{x}(t) + \mathbf{b}$ *has a nonnegative equilibrium point* $\bar{\mathbf{x}}$ *if and only if all of the eigenvalues of* A *are strictly in the left half of the complex plane* (*that is,* $\Re(\lambda) < 0$ *for every eigenvalue* λ *of* A).

Proof. If $\Re(\lambda) < 0$ for every eigenvalue λ of A, Theorem 19.11 tells us the system has an asymptotically stable equilibrium point $\bar{\mathbf{x}}$. For any trajectory $\mathbf{x}(t)$ we then have $\lim_{t\to\infty} \mathbf{x}(t) \to \bar{\mathbf{x}}$. However, since the system is nonnegative, we know that for any

nonnegative starting point $\mathbf{x}(0)$, every point of the trajectory $\mathbf{x}(t)$ is nonnegative. Therefore, $\bar{\mathbf{x}} \geq \mathbf{0}$.

Conversely, suppose the system has a nonnegative equilibrium point $\bar{\mathbf{x}}$. Then $\bar{\mathbf{x}} \geq \mathbf{0}$ and $A\bar{\mathbf{x}} + \mathbf{b} = \mathbf{0}$. Since $\mathbf{b} > \mathbf{0}$, we must have $A\bar{\mathbf{x}} < \mathbf{0}$. Let μ_0 be the eigenvalue of A with largest real part. Since A is a Metzler matrix, we know that μ_0 is a real number and that there is a corresponding left nonnegative eigenvector $\mathbf{w}_0$. We then have $\mathbf{w}_0^T A\bar{\mathbf{x}} < 0$ and $\mathbf{w}_0^T A\bar{\mathbf{x}} = \mu_0 \mathbf{w}_0^T \bar{\mathbf{x}}$. Since $\mathbf{w}_0^T \bar{\mathbf{x}} > 0$, this tells us $\mu_0 < 0$. Hence, every eigenvalue of A has a negative real part. □

19.7. Markov Chains

Suppose we have a set of n *states* and a process (also called a chain) which moves successively from state to state in discrete time periods. Each move is called a *step*. If the chain is currently in state i, we let p_{ij} denote the probability that it will move to state j at the next step. Note that $0 \leq p_{ij} \leq 1$ and $\sum_{j=1}^{n} p_{ij} = 1$. Let P be the $n \times n$ matrix with p_{ij} in entry i, j. The probabilities p_{ij} are called *transition probabilities* and the matrix P is called the *transition matrix*. The condition $\sum_{j=1}^{n} p_{ij} = 1$ says that in each row of P, the entries sum to one.

Definition 19.20. We say an $n \times n$ matrix P is *row stochastic* if the entries of P are all nonnegative real numbers and all of the row sums of P are one.

Observe that P is row stochastic if and only if $P \geq 0$ and $P\mathbf{e} = \mathbf{e}$, where $\mathbf{e}$ denotes the all-one vector.

Definition 19.21. We say a vector $\mathbf{x}$ is a *probability vector* if all of the entries of $\mathbf{x}$ are nonnegative and the sum of the coordinates of $\mathbf{x}$ is 1. Equivalently, $\mathbf{x} \geq 0$, and if $\mathbf{e}$ denotes the all-one vector, then $\mathbf{x}^T\mathbf{e} = 1$.

We interpret the coordinates of a probability vector $\mathbf{x}$ as representing the probabilities of being in each of the n states; thus x_j represents the probability the chain is in state j. A matrix P is row stochastic if and only if each row of P is a probability vector.

Proposition 19.22. *If $\mathbf{x}$ is a probability vector and P is row stochastic, then $\mathbf{x}^T P$ is a probability vector. If P and Q are $n \times n$ row stochastic matrices, then PQ is row stochastic.*

Proof. Suppose $\mathbf{x}$ is a probability vector and P is row stochastic. Then $\mathbf{x}^T P$ is clearly nonnegative, and $(\mathbf{x}^T P)\mathbf{e} = \mathbf{x}^T(P\mathbf{e}) = \mathbf{x}^T\mathbf{e} = 1$, so $\mathbf{x}^T P$ is a probability vector.

Now suppose P and Q are $n \times n$ row stochastic matrices. Then PQ is clearly nonnegative, and $(PQ)\mathbf{e} = P(Q\mathbf{e}) = P\mathbf{e} = \mathbf{e}$, so PQ is row stochastic. □

Now suppose $\mathbf{x}(k)$ is the probability vector in which $x_j(k)$ gives the probability the chain is in state j at step (or time) k. Then $p_{ij}x_i(k)$ gives the probability that

the chain is in state i at time k and in state j at time $k+1$. If we sum the quantities $p_{ij}x_i(k)$ over i, we get the probability the chain is in state j at time $k+1$. Thus,

$$x_j(k+1) = \sum_{i=1}^{n} p_{ij}x_i(k), \quad j = 1, \ldots, n. \tag{19.35}$$

Writing the probability vectors as row vectors, the matrix form of (19.35) is

$$\mathbf{x}^T(k+1) = \mathbf{x}^T(k)P. \tag{19.36}$$

Our apologies for this notational change to row vectors and placing the matrix transformation on the right; however, this seems to be the usual notation for Markov chains.

As we have seen before, equation (19.36) leads to the formula

$$\mathbf{x}^T(k) = \mathbf{x}^T(0)P^k. \tag{19.37}$$

We now use the fact that P is row stochastic to study the behavior of $\mathbf{x}(k)$ as $k \to \infty$. Since $P \geq 0$, the Perron–Frobenius theorem applies. From $P\mathbf{e} = \mathbf{e}$, we see $\lambda = 1$ is an eigenvalue of P and $\mathbf{e}$ is an associated eigenvector. The Geršgorin Circle Theorem (more specifically, Theorem 3.16) tells us $\rho(P) \leq 1$. Since 1 is an eigenvalue, $\rho(P) = 1$, and we have the positive eigenvector $\mathbf{e}$. Of critical importance to analyzing the behavior of $\mathbf{x}(k)$ as $k \to \infty$ is whether there are other eigenvalues of modulus one—i.e., whether the matrix P is irreducible or not, and, in the case of an irreducible P, whether P is primitive or not.

Theorem 19.23. *Suppose A is an $n \times n$ row stochastic matrix which is irreducible and primitive. Then there exists a unique, positive probability vector $\mathbf{q}$ such that*

(1) $\mathbf{q}^T A = \mathbf{q}^T$.
(2) $\lim_{k\to\infty} A^k = \mathbf{e}\ \mathbf{q}^T$.

Proof. Since A is row stochastic, we know $\rho(A) = 1$; since A is irreducible, we know $\lambda = 1$ is a simple eigenvalue of A. Also, we know that A has a positive left eigenvector $\mathbf{x}$ corresponding to the eigenvalue 1. Since 1 is a simple eigenvalue, the corresponding eigenspace is one dimensional. Divide $\mathbf{x}$ by $\sum_{i=1}^{n} x_i$ to get the unique probability vector, $\mathbf{q} = \frac{1}{\sum_{i=1}^{n} x_i}\mathbf{x}$, which satisfies $\mathbf{q}^T A = \mathbf{q}^T$.

For the second part, consider the Jordan canonical form of A. Since A is primitive, we know that for any eigenvalue $\lambda \neq 1$, we have $|\lambda| < 1$. Hence, the Jordan canonical form of A may be written as $J = 1 \oplus \bar{J}$, where $\bar{J}$ is a sum of Jordan blocks corresponding to eigenvalues of modulus less than one. We then have $\lim_{k\to\infty} \bar{J}^k = 0$ and $\lim_{k\to\infty} J^k = 1 \oplus 0_{n-1}$. This tells us $\lim_{k\to\infty} A^k$ exists and is a matrix of rank 1. Let $B = \lim_{k\to\infty} A^k$; then B is row stochastic and has rank 1. Also, $BA = B$, so every row of B must be a left eigenvector of A corresponding to the eigenvalue 1. Since every row of B is also a probability vector, we see every row of B is the vector $\mathbf{q}^T$. Hence, $B = \lim_{k\to\infty} A^k = \mathbf{e}\mathbf{q}^T$. □

Theorem 19.23 tells us the following. If the transition matrix P of a Markov chain is irreducible and primitive, then in the long run, the probability distribution

for the chain is given by the left eigenvector corresponding to the eigenvalue 1. In particular, note that if $P > 0$, then P is irreducible and primitive.

Example 19.24. Let us return to Hesh's bakery and consider a group of regular customers, each of whom buys exactly one of the following items each week: the French babka, the chocolate chip pound cake (referred to as the CCC), or the sticky buns. Suppose that if a customer buys a babka one week, there is a probability of .30 that she buys the babka the following week, a probability of .50 she buys the CCC the following week, and then a .20 probability she buys the sticky buns. Of those who buy the sticky buns one week, 60% will buy the babka the following week, and 30% will buy the CCC, while 10% stick with the buns. Finally, of the CCC buyers, half will buy the CCC the following week, while 30% switch to sticky buns, and the remaining 20% buy babka the following week. If our three states are

(1) State 1: French babka,

(2) State 2: Sticky buns,

(3) State 3: CCC,

then the transition matrix for this Markov chain is

$$P = \begin{pmatrix} .3 & .2 & .5 \\ .6 & .1 & .3 \\ .2 & .3 & .5 \end{pmatrix}.$$

In the long run, what proportion of sales will be babka, sticky buns, and CCC? This corresponds to the left eigenvector for P; thus, we want to find the probability vector $\mathbf{q}$ satisfying $\mathbf{q}^T P = \mathbf{q}^T$. A straightforward calculation gives

$$\mathbf{q} = \frac{1}{112}\begin{pmatrix} 36 \\ 25 \\ 51 \end{pmatrix} = \begin{pmatrix} .32 \\ .22 \\ .46 \end{pmatrix}$$

rounded off to two decimal places. Thus, over a long period of time, we may expect 32% of customers to buy babka, 22% to buy sticky buns, and 46% to buy chocolate chip cake.

Things are more complicated if P is reducible or imprimitive. We refer the reader elsewhere for more complete treatments of the general case. Here, we consider one more special situation: the absorbing Markov chain.

Definition 19.25. We say a state of a Markov chain is an *absorbing state* if once the process is in that state, it remains in that state.

When state i is an absorbing state, the ith row of the transition matrix P has a 1 in the ith entry and zeroes elsewhere.

Definition 19.26. We say a Markov chain is an *absorbing* chain if it has at least one absorbing state, and, from any nonabsorbing state, there is a positive probability of reaching an absorbing state in a finite number of steps. In this case we call the nonabsorbing states *transient* states.

The definition of absorbing chain can also be stated in terms of the directed graph of the matrix P: there are no edges coming out of any absorbing state, while from any nonabsorbing state, there is a path to some absorbing state.

Suppose that we have an absorbing n-state Markov chain with t transient states and r absorbing states; so $t+r=n$. We order the states so that states $1, 2, \dots t$ are the transient states and states $t+1, t+1, \dots, t+r=n$ are the absorbing states. The transition matrix P then takes the form

$$P = \begin{pmatrix} Q & R \\ 0 & I_r \end{pmatrix}, \tag{19.38}$$

where Q is $t \times t$, while R is $t \times r$ and the block of zeros in the lower left-hand corner is $r \times t$. Then P^k has the form

$$P^k = \begin{pmatrix} Q^k & R_k \\ 0 & I_r \end{pmatrix},$$

where R_k is $t \times r$; we now find a formula for R_k in terms of Q and R. From

$$P^2 = \begin{pmatrix} Q^2 & QR+R \\ 0 & I_r \end{pmatrix},$$

we see

$$R_2 = QR + R = (Q+I)R.$$

Computing P^3, we have

$$P^3 = \begin{pmatrix} Q^3 & Q^2R+QR+R \\ 0 & I_r \end{pmatrix},$$

so $R_3 = Q^2R+QR+R = (Q^2+Q+I)R$. One may now guess the following formula, which may be proven by induction:

$$R_k = (I + Q + Q^2 + \cdots + Q^{k-1})R. \tag{19.39}$$

From (19.39), we see that $R_m \geq R_k$ when $m \geq k$. In particular, note that if R_k has a nonzero entry in position i, j, then for any $m > k$, the matrix R_m will also have a nonzero entry in position i, j.

Theorem 19.27. *Suppose* $P = \begin{pmatrix} Q & R \\ 0 & I_r \end{pmatrix}$ *is the transition matrix for an absorbing Markov chain with* r *absorbing states and* t *transient states, with* Q *being* $t \times t$. *Then* $\rho(Q) < 1$ *and* $\lim_{k\to\infty} P^k = \begin{pmatrix} 0_t & (I-Q)^{-1}R \\ 0 & I_r \end{pmatrix}$.

Proof. Let $1 \leq i \leq t$; then state i is nonabsorbing. Hence, starting from state i, there is a positive integer k_i such that the process has positive probability of being in one of the absorbing states after k_i steps. This means that the ith row of P^{k_i} will have a positive entry in at least one of the last r columns. So for any $m \geq k_i$, the matrix R_m has a positive entry in row i. Choose $m = \max\{k_1, k_2, \dots, k_t\}$. Then every row of R_m must have a positive entry. Now P is row stochastic; hence any power of P is also row stochastic. So $P^m = \begin{pmatrix} Q^m & R_m \\ 0 & I_r \end{pmatrix}$ is row stochastic and every row of R_m has a positive entry. This tells us that each row sum of the nonnegative matrix Q^m is less than one, and hence $\rho(Q^m) < 1$. Since $\rho(Q^m) = (\rho(Q))^m$, we have $\rho(Q) < 1$.

Now $\rho(Q) < 1$ tells us $\lim\limits_{k\to\infty} Q^k = 0$ and the infinite series

$$I + Q + Q^2 + Q^3 + \cdots$$