Chapter 15

Unit Quaternions and Rotations in SO(3)

This chapter is devoted to the representation of rotations in $\mathbf{SO}(3)$ in terms of unit quaternions. Since we already defined the unitary groups $\mathbf{SU}(n)$, the quickest way to introduce the *unit quaternions* is to define them as the elements of the group $\mathbf{SU}(2)$.

The skew field $\mathbb{H}$ of quaternions and the group $\mathbf{SU}(2)$ of unit quaternions are discussed in Section 15.1. In Section 15.2, we define a homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ and prove that its kernel is $\{-I, I\}$. We compute the rotation matrix R_q associated with the rotation r_q induced by a unit quaternion q in Section 15.3. In Section 15.4, we prove that the homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ is surjective by providing an algorithm to construct a quaternion from a rotation matrix. In Section 15.5 we define the exponential map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$ where $\mathfrak{su}(2)$ is the real vector space of skew-Hermitian 2×2 matrices with zero trace. We prove that exponential map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$ is surjective and give an algorithm for finding a logarithm. We discuss quaternion interpolation and prove the famous *slerp interpolation formula* due to Ken Shoemake in Section 15.6. This formula is used in robotics and computer graphics to deal with interpolation problems. In Section 15.7, we prove that there is no "nice" section $s\colon \mathbf{SO}(3) \to \mathbf{SU}(2)$ of the homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$, in the sense that any section of r is neither a homomorphism nor continuous.

15.1 The Group SU(2) of Unit Quaternions and the Skew Field $\mathbb{H}$ of Quaternions

Definition 15.1. The *unit quaternions* are the elements of the group $\mathbf{SU}(2)$, namely the group of 2×2 complex matrices of the form

$$\begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \quad \alpha, \beta \in \mathbb{C},\ \alpha\overline{\alpha} + \beta\overline{\beta} = 1.$$

The *quaternions* are the elements of the real vector space $\mathbb{H} = \mathbb{R}\,\mathbf{SU}(2)$.

Let $\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k}$ be the matrices

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{i} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \qquad \mathbf{j} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad \mathbf{k} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix},$$

then $\mathbb{H}$ is the set of all matrices of the form

$$X = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}, \quad a, b, c, d \in \mathbb{R}.$$

Indeed, every matrix in $\mathbb{H}$ is of the form

$$X = \begin{pmatrix} a + ib & c + id \\ -(c - id) & a - ib \end{pmatrix}, \quad a, b, c, d \in \mathbb{R}.$$

It is easy (but a bit tedious) to verify that the quaternions $\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k}$ satisfy the famous identities discovered by Hamilton:

$$\begin{aligned} &\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -\mathbf{1}, \\ &\mathbf{ij} = -\mathbf{ji} = \mathbf{k}, \\ &\mathbf{jk} = -\mathbf{kj} = \mathbf{i}, \\ &\mathbf{ki} = -\mathbf{ik} = \mathbf{j}. \end{aligned}$$

Thus, the quaternions are a generalization of the complex numbers, but there are three square roots of $-\mathbf{1}$ and multiplication is not commutative.

Given any two quaternions $X = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ and $Y = a'\mathbf{1} + b'\mathbf{i} + c'\mathbf{j} + d'\mathbf{k}$, Hamilton's famous formula

$$\begin{aligned} XY = &(aa' - bb' - cc' - dd')\mathbf{1} + (ab' + ba' + cd' - dc')\mathbf{i} \\ &+ (ac' + ca' + db' - bd')\mathbf{j} + (ad' + da' + bc' - cb')\mathbf{k} \end{aligned}$$

looks mysterious, but it is simply the result of multiplying the two matrices

$$X = \begin{pmatrix} a + ib & c + id \\ -(c - id) & a - ib \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} a' + ib' & c' + id' \\ -(c' - id') & a' - ib' \end{pmatrix}.$$

It is worth noting that this formula was discovered independently by Olinde Rodrigues in 1840, a few years before Hamilton (Veblen and Young [Veblen and Young (1946)]). However, Rodrigues was working with a different formalism, homogeneous transformations, and he did not discover the quaternions.

If

$$X = \begin{pmatrix} a+ib & c+id \\ -(c-id) & a-ib \end{pmatrix}, \quad a, b, c, d \in \mathbb{R},$$

it is immediately verified that

$$XX^* = X^*X = (a^2+b^2+c^2+d^2)\mathbf{1}.$$

Also observe that

$$X^* = \begin{pmatrix} a-ib & -(c+id) \\ c-id & a+ib \end{pmatrix} = a\mathbf{1} - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}.$$

This implies that if $X \neq 0$, then X is invertible and its inverse is given by

$$X^{-1} = (a^2+b^2+c^2+d^2)^{-1}X^*.$$

As a consequence, it can be verified that $\mathbb{H}$ is a skew field (a noncommutative field). It is also a real vector space of dimension 4 with basis $(\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k})$; thus as a vector space, $\mathbb{H}$ is isomorphic to $\mathbb{R}^4$.

Definition 15.2. A concise notation for the quaternion X defined by $\alpha = a + ib$ and $\beta = c + id$ is

$$X = [a, (b, c, d)].$$

We call a the *scalar part* of X and (b, c, d) the *vector part* of X. With this notation, $X^* = [a, -(b, c, d)]$, which is often denoted by $\overline{X}$. The quaternion $\overline{X}$ is called the *conjugate* of X. If q is a unit quaternion, then $\overline{q}$ is the multiplicative inverse of q.

15.2 Representation of Rotations in SO(3) by Quaternions in SU(2)

The key to representation of rotations in $\mathbf{SO}(3)$ by unit quaternions is a certain group homomorphism called the *adjoint representation of* $\mathbf{SU}(2)$. To define this mapping, first we define the real vector space $\mathfrak{su}(2)$ of skew Hermitian matrices.

Definition 15.3. The (real) vector space $\mathfrak{su}(2)$ of 2×2 *skew Hermitian matrices with zero trace* is given by

$$\mathfrak{su}(2) = \left\{ \begin{pmatrix} ix & y+iz \\ -y+iz & -ix \end{pmatrix} \;\middle|\; (x, y, z) \in \mathbb{R}^3 \right\}.$$

Observe that for every matrix $A \in \mathfrak{su}(2)$, we have $A^* = -A$, that is, A is skew Hermitian, and that $\mathrm{tr}(A) = 0$.

Definition 15.4. The *adjoint representation* of the group $\mathbf{SU}(2)$ is the group homomorphism
$\mathrm{Ad}\colon \mathbf{SU}(2) \to \mathbf{GL}(\mathfrak{su}(2))$ defined such that for every $q \in \mathbf{SU}(2)$, with

$$q = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \in \mathbf{SU}(2),$$

we have

$$\mathrm{Ad}_q(A) = qAq^*, \quad A \in \mathfrak{su}(2),$$

where q^* is the inverse of q (since $\mathbf{SU}(2)$ is a unitary group) and is given by

$$q^* = \begin{pmatrix} \overline{\alpha} & -\beta \\ \overline{\beta} & \alpha \end{pmatrix}.$$

One needs to verify that the map Ad_q is an invertible linear map from $\mathfrak{su}(2)$ to itself, and that Ad is a group homomorphism, which is easy to do.

In order to associate a rotation ρ_q (in $\mathbf{SO}(3)$) to q, we need to embed $\mathbb{R}^3$ into $\mathbb{H}$ as the pure quaternions, by

$$\psi(x, y, z) = \begin{pmatrix} ix & y + iz \\ -y + iz & -ix \end{pmatrix}, \quad (x, y, z) \in \mathbb{R}^3.$$

Then q defines the map ρ_q (on $\mathbb{R}^3$) given by

$$\rho_q(x, y, z) = \psi^{-1}(q\psi(x, y, z)q^*).$$

Therefore, modulo the isomorphism ψ, the linear map ρ_q is the linear isomorphism Ad_q. In fact, it turns out that ρ_q is a rotation (and so is Ad_q), which we will prove shortly. So, the representation of rotations in $\mathbf{SO}(3)$ by unit quaternions is just the adjoint representation of $\mathbf{SU}(2)$; its image is a subgroup of $\mathbf{GL}(\mathfrak{su}(2))$ isomorphic to $\mathbf{SO}(3)$.

Technically, it is a bit simpler to embed $\mathbb{R}^3$ in the (real) vector spaces of Hermitian matrices with zero trace,

$$\left\{ \begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix} \;\middle|\; x, y, z \in \mathbb{R} \right\}.$$

Since the matrix $\psi(x, y, z)$ is skew-Hermitian, the matrix $-i\psi(x, y, z)$ is Hermitian, and we have

$$-i\psi(x, y, z) = \begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix} = x\sigma_3 + y\sigma_2 + z\sigma_1,$$

where $\sigma_1, \sigma_2, \sigma_3$ are the *Pauli spin matrices*

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Matrices of the form $x\sigma_3 + y\sigma_2 + z\sigma_1$ are Hermitian matrices with zero trace.

It is easy to see that every 2×2 Hermitian matrix with zero trace must be of this form. (observe that $(i\sigma_1, i\sigma_2, i\sigma_3)$ forms a basis of $\mathfrak{su}(2)$. Also, $\mathbf{i} = i\sigma_3$, $\mathbf{j} = i\sigma_2$, $\mathbf{k} = i\sigma_1$.)

Now, if $A = x\sigma_3 + y\sigma_2 + z\sigma_1$ is a Hermitian 2×2 matrix with zero trace, we have

$$(qAq^*)^* = qA^*q^* = qAq^*,$$

so qAq^* is also Hermitian, and

$$\mathrm{tr}(qAq^*) = \mathrm{tr}(Aq^*q) = \mathrm{tr}(A),$$

and qAq^* also has zero trace. Therefore, the map $A \mapsto qAq^*$ preserves the Hermitian matrices with zero trace. We also have

$$\det(x\sigma_3 + y\sigma_2 + z\sigma_1) = \det \begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix} = -(x^2 + y^2 + z^2),$$

and

$$\det(qAq^*) = \det(q)\det(A)\det(q^*) = \det(A) = -(x^2 + y^2 + z^2).$$

We can embed $\mathbb{R}^3$ into the space of Hermitian matrices with zero trace by

$$\varphi(x, y, z) = x\sigma_3 + y\sigma_2 + z\sigma_1.$$

Note that

$$\varphi = -i\psi \quad \text{and} \quad \varphi^{-1} = i\psi^{-1}.$$

Definition 15.5. The unit quaternion $q \in \mathbf{SU}(2)$ induces a map r_q on $\mathbb{R}^3$ by

$$r_q(x, y, z) = \varphi^{-1}(q\varphi(x, y, z)q^*) = \varphi^{-1}(q(x\sigma_3 + y\sigma_2 + z\sigma_1)q^*).$$

The map r_q is clearly linear since φ is linear.

Proposition 15.1. *For every unit quaternion* $q \in \mathbf{SU}(2)$*, the linear map* r_q *is orthogonal, that is,* $r_q \in \mathbf{O}(3)$*.*

Proof. Since

$$-\|(x,y,z)\|^2 = -(x^2+y^2+z^2) = \det(x\sigma^3 + y\sigma^2 + z\sigma_1) = \det(\varphi(x,y,z)),$$

we have

$$\begin{aligned} -\|r_q(x,y,z)\|^2 &= \det(\varphi(r_q(x,y,z))) = \det(q(x\sigma_3 + y\sigma_2 + z\sigma_1)q^*) \\ &= \det(x\sigma_3 + y\sigma_2 + z\sigma_1) = -\left\|(x,y,z)^2\right\|, \end{aligned}$$

and we deduce that r_q is an isometry. Thus, $r_q \in \mathbf{O}(3)$. □

In fact, r_q is a rotation, and we can show this by finding the fixed points of r_q. Let q be a unit quaternion of the form

$$q = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix}$$

with $\alpha = a + ib$, $\beta = c + id$, and $a^2 + b^2 + c^2 + d^2 = 1$ $(a, b, c, d \in \mathbb{R})$.

If $b = c = d = 0$, then $q = I$ and r_q is the identity so we may assume that $(b, c, d) \neq (0, 0, 0)$.

Proposition 15.2. *If* $(b, c, d) \neq (0, 0, 0)$*, then the fixed points of* r_q *are solutions* (x, y, z) *of the linear system*

$$\begin{aligned} -dy + cz &= 0 \\ cx - by &= 0 \\ dx - bz &= 0. \end{aligned}$$

This linear system has the nontrivial solution (b, c, d) *and has rank* 2*. Therefore,* r_q *has the eigenvalue* 1 *with multiplicity* 1*, and* r_q *is a rotation whose axis is determined by* (b, c, d)*.*

Proof. We have $r_q(x, y, z) = (x, y, z)$ iff

$$\varphi^{-1}(q(x\sigma_3 + y\sigma_2 + z\sigma_1)q^*) = (x, y, z)$$

iff

$$q(x\sigma_3 + y\sigma_2 + z\sigma_1)q^* = \varphi(x, y, z),$$

and since

$$\varphi(x, y, z) = x\sigma_3 + y\sigma_2 + z\sigma_1 = A$$

with

$$A = \begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix},$$

we see that $r_q(x, y, z) = (x, y, z)$ iff

$$qAq^* = A \quad \text{iff} \quad qA = Aq.$$

We have

$$qA = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix} = \begin{pmatrix} \alpha x + \beta z + i\beta y & \alpha z - i\alpha y - \beta x \\ -\overline{\beta} x + \overline{\alpha} z + i\overline{\alpha} y & -\overline{\beta} z + i\overline{\beta} y - \overline{\alpha} x \end{pmatrix}$$

and

$$Aq = \begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} = \begin{pmatrix} \alpha x - \overline{\beta} z + i\overline{\beta} y & \beta x + \overline{\alpha} z - i\overline{\alpha} y \\ \alpha z + i\alpha y + \overline{\beta} x & \beta z + i\beta y - \overline{\alpha} x \end{pmatrix}.$$

By equating qA and Aq, we get

$$\begin{aligned} i(\beta - \overline{\beta})y + (\beta + \overline{\beta})z &= 0 \\ 2\beta x + i(\alpha - \overline{\alpha})y + (\overline{\alpha} - \alpha)z &= 0 \\ 2\overline{\beta} x + i(\alpha - \overline{\alpha})y + (\alpha - \overline{\alpha})z &= 0 \\ i(\beta - \overline{\beta})y + (\beta + \overline{\beta})z &= 0. \end{aligned}$$

The first and the fourth equation are identical and the third equation is obtained by conjugating the second, so the above system reduces to

$$\begin{aligned} i(\beta - \overline{\beta})y + (\beta + \overline{\beta})z &= 0 \\ 2\beta x + i(\alpha - \overline{\alpha})y + (\overline{\alpha} - \alpha)z &= 0. \end{aligned}$$

Replacing α by $a + ib$ and β by $c + id$, we get

$$\begin{aligned} -dy + cz &= 0 \\ cx - by + i(dx - bz) &= 0, \end{aligned}$$

which yields the equations

$$\begin{aligned} -dy + cz &= 0 \\ cx - by &= 0 \\ dx - bz &= 0. \end{aligned}$$

This linear system has the nontrivial solution (b, c, d) and the matrix of this system is

$$\begin{pmatrix} 0 & -d & c \\ c & -b & 0 \\ d & 0 & -b \end{pmatrix}.$$

Since $(b, c, d) \neq (0, 0, 0)$, this matrix always has a 2×2 submatrix which is nonsingular, so it has rank 2, and consequently its kernel is the one-dimensional space spanned by (b, c, d). Therefore, r_q has the eigenvalue 1 with multiplicity 1. If we had $\det(r_q) = -1$, then the eigenvalues of r_q would be either $(-1, 1, 1)$ or $(-1, e^{i\theta}, e^{-i\theta})$ with $\theta \neq k2\pi$ (with $k \in \mathbb{Z}$), contradicting the fact that 1 is an eigenvalue with multiplicity 1. Therefore, r_q is a rotation; in fact, its axis is determined by (b, c, d). □

In summary, $q \mapsto r_q$ is a map r from $\mathbf{SU}(2)$ to $\mathbf{SO}(3)$.

Theorem 15.1. *The map* $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ *is a homomorphism whose kernel is* $\{I, -I\}$.

Proof. This map is a homomorphism, because if $q_1, q_2 \in \mathbf{SU}(2)$, then

$$\begin{aligned} r_{q_2}(r_{q_1}(x,y,z)) &= \varphi^{-1}(q_2\varphi(r_{q_1}(x,y,z))q_2^*) \\ &= \varphi^{-1}(q_2\varphi(\varphi^{-1}(q_1\varphi(x,y,z)q_1^*))q_2^*) \\ &= \varphi^{-1}((q_2q_1)\varphi(x,y,z)(q_2q_1)^*) \\ &= r_{q_2q_1}(x,y,z). \end{aligned}$$

The computation that showed that if $(b,c,d) \neq (0,0,0)$, then r_q has the eigenvalue 1 with multiplicity 1 implies the following: if $r_q = I_3$, namely r_q has the eigenvalue 1 with multiplicity 3, then $(b,c,d) = (0,0,0)$. But then $a = \pm 1$, and so $q = \pm I_2$. Therefore, the kernel of the homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ is $\{I, -I\}$. □

Remark: Perhaps the quickest way to show that r maps $\mathbf{SU}(2)$ into $\mathbf{SO}(3)$ is to observe that the map r is continuous. Then, since it is known that $\mathbf{SU}(2)$ is connected, its image by r lies in the connected component of I, namely $\mathbf{SO}(3)$.

The map r is surjective, but this is not obvious. We will return to this point after finding the matrix representing r_q explicitly.

15.3 Matrix Representation of the Rotation r_q

Given a unit quaternion q of the form

$$q = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix}$$

with $\alpha = a + ib$, $\beta = c + id$, and $a^2 + b^2 + c^2 + d^2 = 1$ $(a, b, c, d \in \mathbb{R})$, to find the matrix representing the rotation r_q we need to compute

$$q(x\sigma_3 + y\sigma_2 + z\sigma_1)q^* = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix} \begin{pmatrix} \overline{\alpha} & -\beta \\ \overline{\beta} & \alpha \end{pmatrix}.$$

First we have

$$\begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix} \begin{pmatrix} \overline{\alpha} & -\beta \\ \overline{\beta} & \alpha \end{pmatrix} = \begin{pmatrix} x\overline{\alpha} + z\overline{\beta} - iy\overline{\beta} & -x\beta + z\alpha - iy\alpha \\ z\overline{\alpha} + iy\overline{\alpha} - x\overline{\beta} & -z\beta - iy\beta - x\alpha \end{pmatrix}.$$

Next, we have

$$\begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \begin{pmatrix} x\overline{\alpha} + z\overline{\beta} - iy\overline{\beta} & -x\beta + z\alpha - iy\alpha \\ z\overline{\alpha} + iy\overline{\alpha} - x\overline{\beta} & -z\beta - iy\beta - x\alpha \end{pmatrix} = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix},$$

with

$$\begin{aligned} A_1 &= (\alpha\overline{\alpha} - \beta\overline{\beta})x + i(\overline{\alpha}\beta - \alpha\overline{\beta})y + (\alpha\overline{\beta} + \overline{\alpha}\beta)z \\ A_2 &= -2\alpha\beta x - i(\alpha^2 + \beta^2)y + (\alpha^2 - \beta^2)z \\ A_3 &= -2\overline{\alpha}\overline{\beta}x + i(\overline{\alpha}^2 + \overline{\beta}^2)y + (\overline{\alpha}^2 - \overline{\beta}^2)z \\ A_4 &= -(\alpha\overline{\alpha} - \beta\overline{\beta})x - i(\overline{\alpha}\beta - \alpha\overline{\beta})y - (\alpha\overline{\beta} + \overline{\alpha}\beta)z. \end{aligned}$$

Since $\alpha = a + ib$ and $\beta = c + id$, with $a, b, c, d \in \mathbb{R}$, we have

$$\begin{aligned} \alpha\overline{\alpha} - \beta\overline{\beta} &= a^2 + b^2 - c^2 - d^2 \\ i(\overline{\alpha}\beta - \alpha\overline{\beta}) &= 2(bc - ad) \\ \alpha\overline{\beta} + \overline{\alpha}\beta &= 2(ac + bd) \\ -\alpha\beta &= -ac + bd - i(ad + bc) \\ -i(\alpha^2 + \beta^2) &= 2(ab + cd) - i(a^2 - b^2 + c^2 - d^2) \\ \alpha^2 - \beta^2 &= a^2 - b^2 - c^2 + d^2 + i2(ab - cd). \end{aligned}$$

Using the above, we get

$$\begin{aligned} &(\alpha\overline{\alpha} - \beta\overline{\beta})x + i(\overline{\alpha}\beta - \alpha\overline{\beta})y + (\alpha\overline{\beta} + \overline{\alpha}\beta)z \\ &\qquad = (a^2 + b^2 - c^2 - d^2)x + 2(bc - ad)y + 2(ac + bd)z, \end{aligned}$$

and

$$\begin{aligned} &-2\alpha\beta x - i(\alpha^2 + \beta^2)y + (\alpha^2 - \beta^2)z \\ &\qquad = 2(-ac + bd)x + 2(ab + cd)y + (a^2 - b^2 - c^2 + d^2)z \\ &\qquad\quad - i[2(ad + bc)x + (a^2 - b^2 + c^2 - d^2)y + 2(-ab + cd)z]. \end{aligned}$$

If we write

$$q(x\sigma_3 + y\sigma_2 + z\sigma_1)q^* = \begin{pmatrix} x' & z' - iy' \\ z' + iy' & -x' \end{pmatrix},$$

we obtain

$$\begin{aligned} x' &= (a^2 + b^2 - c^2 - d^2)x + 2(bc - ad)y + 2(ac + bd)z \\ y' &= 2(ad + bc)x + (a^2 - b^2 + c^2 - d^2)y + 2(-ab + cd)z \\ z' &= 2(-ac + bd)x + 2(ab + cd)y + (a^2 - b^2 - c^2 + d^2)z. \end{aligned}$$

In summary, we proved the following result.

Proposition 15.3. *The matrix representing* r_q *is*

$$R_q = \begin{pmatrix} a^2+b^2-c^2-d^2 & 2bc-2ad & 2ac+2bd \\ 2bc+2ad & a^2-b^2+c^2-d^2 & -2ab+2cd \\ -2ac+2bd & 2ab+2cd & a^2-b^2-c^2+d^2 \end{pmatrix}.$$

Since $a^2+b^2+c^2+d^2=1$*, this matrix can also be written as*

$$R_q = \begin{pmatrix} 2a^2+2b^2-1 & 2bc-2ad & 2ac+2bd \\ 2bc+2ad & 2a^2+2c^2-1 & -2ab+2cd \\ -2ac+2bd & 2ab+2cd & 2a^2+2d^2-1 \end{pmatrix}.$$

The above is the rotation matrix in Euler form induced by the quaternion q, which is the matrix corresponding to ρ_q. This is because

$$\varphi = -i\psi, \quad \varphi^{-1} = i\psi^{-1},$$

so

$$\begin{aligned} r_q(x,y,z) = \varphi^{-1}(q\varphi(x,y,z)q^*) &= i\psi^{-1}(q(-i\psi(x,y,z))q^*) \\ &= \psi^{-1}(q\psi(x,y,z)q^*) = \rho_q(x,y,z), \end{aligned}$$

and so $r_q = \rho_q$.

We showed that every unit quaternion $q \in \mathbf{SU}(2)$ induces a rotation $r_q \in \mathbf{SO}(3)$, but it is not obvious that every rotation can be represented by a quaternion. This can shown in various ways.

One way to is use the fact that every rotation in $\mathbf{SO}(3)$ is the composition of two reflections, and that every reflection σ of $\mathbb{R}^3$ can be represented by a quaternion q, in the sense that

$$\sigma(x,y,z) = -\varphi^{-1}(q\varphi(x,y,z)q^*).$$

Note the presence of the negative sign. This is the method used in Gallier [Gallier (2011b)] (Chapter 9).

15.4 An Algorithm to Find a Quaternion Representing a Rotation

Theorem 15.2. *The homomorphism* $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ *is surjective.*

Here is an algorithmic method to find a unit quaternion q representing a rotation matrix R, which provides a proof of Theorem 15.2.

Let

$$q = \begin{pmatrix} a+ib & c+id \\ -(c-id) & a-ib \end{pmatrix}, \quad a^2+b^2+c^2+d^2 = 1,\ a,b,c,d \in \mathbb{R}.$$

First observe that the trace of R_q is given by

$$\mathrm{tr}(R_q) = 3a^2 - b^2 - c^2 - d^2,$$

but since $a^2+b^2+c^2+d^2 = 1$, we get $\mathrm{tr}(R_q) = 4a^2 - 1$, so

$$a^2 = \frac{\mathrm{tr}(R_q)+1}{4}.$$

If $R \in \mathbf{SO}(3)$ is any rotation matrix and if we write

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33}, \end{pmatrix}$$

we are looking for a unit quaternion $q \in \mathbf{SU}(2)$ such that $R_q = R$. Therefore, we must have

$$a^2 = \frac{\mathrm{tr}(R)+1}{4}.$$

We also know that

$$\mathrm{tr}(R) = 1 + 2\cos\theta,$$

where $\theta \in [0,\pi]$ is the angle of the rotation R, so we get

$$a^2 = \frac{\cos\theta + 1}{2} = \cos^2\left(\frac{\theta}{2}\right),$$

which implies that

$$|a| = \cos\left(\frac{\theta}{2}\right) \quad (0 \leq \theta \leq \pi).$$

Note that we may assume that $\theta \in [0,\pi]$, because if $\pi \leq \theta \leq 2\pi$, then $\theta - 2\pi \in [-\pi, 0]$, and then the rotation of angle $\theta - 2\pi$ and axis determined by the vector (b,c,d) is the same as the rotation of angle $2\pi - \theta \in [0,\pi]$ and axis determined by the vector $-(b,c,d)$. There are two cases.

Case 1. $\mathrm{tr}(R) \neq -1$, or equivalently $\theta \neq \pi$. In this case $a \neq 0$. Pick

$$a = \frac{\sqrt{\mathrm{tr}(R)+1}}{2}.$$

Then by equating $R - R^\top$ and $R_q - R_q^\top$, we get

$$\begin{aligned} 4ab &= r_{32} - r_{23} \\ 4ac &= r_{13} - r_{31} \\ 4ad &= r_{21} - r_{12}, \end{aligned}$$

which yields

$$b = \frac{r_{32} - r_{23}}{4a}, \quad c = \frac{r_{13} - r_{31}}{4a}, \quad d = \frac{r_{21} - r_{12}}{4a}.$$

Case 2. $\mathrm{tr}(R) = -1$, or equivalently $\theta = \pi$. In this case $a = 0$. By equating $R + R^\top$ and $R_q + R_q^\top$, we get

$$\begin{aligned} 4bc &= r_{21} + r_{12} \\ 4bd &= r_{13} + r_{31} \\ 4cd &= r_{32} + r_{23}. \end{aligned}$$

By equating the diagonal terms of R and R_q, we also get

$$\begin{aligned} b^2 &= \frac{1 + r_{11}}{2} \\ c^2 &= \frac{1 + r_{22}}{2} \\ d^2 &= \frac{1 + r_{33}}{2}. \end{aligned}$$

Since $q \neq 0$ and $a = 0$, at least one of b, c, d is nonzero.

If $b \neq 0$, let

$$b = \frac{\sqrt{1 + r_{11}}}{\sqrt{2}},$$

and determine c, d using

$$\begin{aligned} 4bc &= r_{21} + r_{12} \\ 4bd &= r_{13} + r_{31}. \end{aligned}$$

If $c \neq 0$, let

$$c = \frac{\sqrt{1 + r_{22}}}{\sqrt{2}},$$

and determine b, d using

$$\begin{aligned} 4bc &= r_{21} + r_{12} \\ 4cd &= r_{32} + r_{23}. \end{aligned}$$

If $d \neq 0$, let

$$d = \frac{\sqrt{1 + r_{33}}}{\sqrt{2}},$$

and determine b, c using

$$4bd = r_{13} + r_{31}$$
$$4cd = r_{32} + r_{23}.$$

It is easy to check that whenever we computed a square root, if we had chosen a negative sign instead of a positive sign, we would obtain the quaternion $-q$. However, both q and $-q$ determine the same rotation r_q.

The above discussion involving the cases $\mathrm{tr}(R) \neq -1$ and $\mathrm{tr}(R) = -1$ is reminiscent of the procedure for finding a logarithm of a rotation matrix using the Rodrigues formula (see Section 11.7). This is not surprising, because if

$$B = \begin{pmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix}$$

and if we write $\theta = \sqrt{u_1^2 + u_2^2 + u_3^2}$ (with $0 \leq \theta \leq \pi$), then the Rodrigues formula says that

$$e^B = I + \frac{\sin\theta}{\theta} B + \frac{(1 - \cos\theta)}{\theta^2} B^2, \quad \theta \neq 0,$$

with $e^0 = I$. It is easy to check that $\mathrm{tr}(e^B) = 1 + 2\cos\theta$. Then it is an easy exercise to check that the quaternion q corresponding to the rotation $R = e^B$ (with $B \neq 0$) is given by

$$q = \left[\cos\left(\frac{\theta}{2}\right), \sin\left(\frac{\theta}{2}\right)\left(\frac{u_1}{\theta}, \frac{u_2}{\theta}, \frac{u_3}{\theta}\right)\right].$$

So the method for finding the logarithm of a rotation R is essentially the same as the method for finding a quaternion defining R.

Remark: Geometrically, the group $\mathbf{SU}(2)$ is homeomorphic to the 3-sphere S^3 in $\mathbb{R}^4$,

$$S^3 = \{(x, y, z, t) \in \mathbb{R}^4 \mid x^2 + y^2 + z^2 + t^2 = 1\}.$$

However, since the kernel of the surjective homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ is $\{I, -I\}$, as a topological space, $\mathbf{SO}(3)$ is homeomorphic to the quotient of S^3 obtained by identifying antipodal points (x, y, z, t) and $-(x, y, z, t)$. This quotient space is the (real) projective space $\mathbb{RP}^3$, and it is more complicated than S^3. The space S^3 is simply-connected, but $\mathbb{RP}^3$ is not.

15.5 The Exponential Map exp: $\mathfrak{su}(2) \to \mathbf{SU}(2)$

Given any matrix $A \in \mathfrak{su}(2)$, with

$$A = \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix},$$

it is easy to check that

$$A^2 = -\theta^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

with $\theta = \sqrt{u_1^2 + u_2^2 + u_3^2}$. Then we have the following formula whose proof is very similar to the proof of the formula given in Proposition 8.17.

Proposition 15.4. *For every matrix* $A \in \mathfrak{su}(2)$*, with*

$$A = \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix},$$

if we write $\theta = \sqrt{u_1^2 + u_2^2 + u_3^2}$*, then*

$$e^A = \cos\theta I + \frac{\sin\theta}{\theta} A, \quad \theta \neq 0,$$

and $e^0 = I$*.*

Therefore, by the discussion at the end of the previous section, e^A is a unit quaternion representing the rotation of angle 2θ and axis (u_1, u_2, u_3) (or I when $\theta = k\pi$, $k \in \mathbb{Z}$). The above formula shows that we may assume that $0 \leq \theta \leq \pi$. Proposition 15.4 shows that the exponential yields a map exp: $\mathfrak{su}(2) \to \mathbf{SU}(2)$. It is an analog of the exponential map exp: $\mathfrak{so}(3) \to \mathbf{SO}(3)$.

Remark: Because $\mathfrak{so}(3)$ and $\mathfrak{su}(2)$ are real vector spaces of dimension 3, they are isomorphic, and it is easy to construct an isomorphism. In fact, $\mathfrak{so}(3)$ and $\mathfrak{su}(2)$ are isomorphic as Lie algebras, which means that there is a linear isomorphism preserving the the Lie bracket $[A, B] = AB - BA$. However, as observed earlier, the groups $\mathbf{SU}(2)$ and $\mathbf{SO}(3)$ are *not isomorphic*.

An equivalent, but often more convenient, formula is obtained by assuming that $u = (u_1, u_2, u_3)$ is a unit vector, equivalently $\det(A) = 1$, in which case $A^2 = -I$, so we have

$$e^{\theta A} = \cos\theta I + \sin\theta A.$$

Using the quaternion notation, this is read as

$$e^{\theta A} = [\cos\theta, \sin\theta\, u].$$

Proposition 15.5. *The exponential map* $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$ *is surjective*

Proof. We give an algorithm to find the logarithm $A \in \mathfrak{su}(2)$ of a unit quaternion

$$q = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix}$$

with $\alpha = a + bi$ and $\beta = c + id$.

If $q = I$ (*i.e.* $a = 1$), then $A = 0$. If $q = -I$ (*i.e.* $a = -1$), then

$$A = \pm\pi \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}.$$

Otherwise, $a \neq \pm 1$ and $(b, c, d) \neq (0, 0, 0)$, and we are seeking some $A = \theta B \in \mathfrak{su}(2)$ with $\det(B) = 1$ and $0 < \theta < \pi$, such that, by Proposition 15.4,

$$q = e^{\theta B} = \cos\theta I + \sin\theta B.$$

Let

$$B = \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix},$$

with $u = (u_1, u_2, u_3)$ a unit vector. We must have

$$a = \cos\theta, \quad e^{\theta B} - (e^{\theta B})^* = q - q^*.$$

Since $0 < \theta < \pi$, we have $\sin\theta \neq 0$, and

$$2\sin\theta \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix} = \begin{pmatrix} \alpha - \overline{\alpha} & 2\beta \\ -2\overline{\beta} & \overline{\alpha} - \alpha \end{pmatrix}.$$

Thus, we get

$$u_1 = \frac{1}{\sin\theta}\, b, \quad u_2 + iu_3 = \frac{1}{\sin\theta}(c + id);$$

that is,

$$\cos\theta = a \qquad (0 < \theta < \pi)$$
$$(u_1, u_2, u_3) = \frac{1}{\sin\theta}(b, c, d).$$

Since $a^2 + b^2 + c^2 + d^2 = 1$ and $a = \cos\theta$, the vector $(b, c, d)/\sin\theta$ is a unit vector. Furthermore if the quaternion q is of the form $q = [\cos\theta, \sin\theta u]$ where $u = (u_1, u_2, u_3)$ is a unit vector (with $0 < \theta < \pi$), then

$$A = \theta \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix} \qquad (*_{\log})$$

is a logarithm of q. □

Observe that not only is the exponential map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$ surjective, but the above proof shows that it is injective on the open ball

$$\{\theta B \in \mathfrak{su}(2) \mid \det(B) = 1, 0 \leq \theta < \pi\}.$$

Also, unlike the situation where in computing the logarithm of a rotation matrix $R \in \mathbf{SO}(3)$ we needed to treat the case where $\mathrm{tr}(R) = -1$ (the angle of the rotation is π) in a special way, computing the logarithm of a quaternion (other than $\pm I$) does not require any case analysis; no special case is needed when the angle of rotation is π.

15.6 Quaternion Interpolation ⊛

We are now going to derive a formula for interpolating between two quaternions. This formula is due to Ken Shoemake, once a Penn student and my TA! Since rotations in $\mathbf{SO}(3)$ can be defined by quaternions, this has applications to computer graphics, robotics, and computer vision.

First we observe that multiplication of quaternions can be expressed in terms of the inner product and the cross-product in $\mathbb{R}^3$. Indeed, if $q_1 = [a, u_1]$ and $q_2 = [a_2, u_2]$, it can be verified that

$$q_1 q_2 = [a_1, u_1][a_2, u_2] = [a_1 a_2 - u_1 \cdot u_2,\ a_1 u_2 + a_2 u_1 + u_1 \times u_2]. \quad (*_{\mathrm{mult}})$$

We will also need the identity

$$u \times (u \times v) = (u \cdot v)u - (u \cdot u)v.$$

Given a quaternion q expressed as $q = [\cos\theta, \sin\theta\, u]$, where u is a unit vector, we can interpolate between I and q by finding the logs of I and q, interpolating in $\mathfrak{su}(2)$, and then exponentiating. We have

$$A = \log(I) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \log(q) = \theta \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix},$$

and so $q = e^B$. Since $\mathbf{SU}(2)$ is a compact Lie group and since the inner product on $\mathfrak{su}(2)$ given by

$$\langle X, Y \rangle = \mathrm{tr}(X^\top Y)$$

is $\mathrm{Ad}(\mathbf{SU}(2))$-invariant, it induces a biinvariant Riemannian metric on $\mathbf{SU}(2)$, and the curve

$$\lambda \mapsto e^{\lambda B}, \quad \lambda \in [0, 1]$$

is a geodesic from I to q in $\mathbf{SU}(2)$. We write $q^\lambda = e^{\lambda B}$. Given two quaternions q_1 and q_2, because the metric is left invariant, the curve

$$\lambda \mapsto Z(\lambda) = q_1 (q_1^{-1} q_2)^\lambda, \quad \lambda \in [0, 1]$$

is a geodesic from q_1 to q_2. Remarkably, there is a closed-form formula for the interpolant $Z(\lambda)$.

Say $q_1 = [\cos\theta, \sin\theta\, u]$ and $q_2 = [\cos\varphi, \sin\varphi\, v]$, and assume that $q_1 \neq q_2$ and $q_1 \neq -q_2$. First, we compute $q^{-1}q_2$. Since $q^{-1} = [\cos\theta, -\sin\theta\, u]$, we have

$$\begin{aligned} q^{-1}q_2 = [&\cos\theta\cos\varphi + \sin\theta\sin\varphi(u\cdot v),\\ &-\sin\theta\cos\varphi\, u + \cos\theta\sin\varphi\, v - \sin\theta\sin\varphi(u\times v)]. \end{aligned}$$

Define Ω by

$$\cos\Omega = \cos\theta\cos\varphi + \sin\theta\sin\varphi(u\cdot v). \tag{$*_\Omega$}$$

Since $q_1 \neq q_2$ and $q_1 \neq -q_2$, we have $0 < \Omega < \pi$, so we get

$$q_1^{-1}q_2 = \left[\cos\Omega,\ \sin\Omega\,\frac{(-\sin\theta\cos\varphi\, u + \cos\theta\sin\varphi\, v - \sin\theta\sin\varphi(u\times v)}{\sin\Omega}\right],$$

where the term multiplying $\sin\Omega$ is a unit vector because q_1 and q_2 are unit quaternions, so $q_1^{-1}q_2$ is also a unit quaternion. By $(*_{\log})$, we have

$$\begin{aligned} &(q_1^{-1}q_2)^\lambda\\ &\quad= \left[\cos\lambda\Omega,\ \sin\lambda\Omega\,\frac{(-\sin\theta\cos\varphi\, u + \cos\theta\sin\varphi\, v - \sin\theta\sin\varphi(u\times v)}{\sin\Omega}\right]. \end{aligned}$$

Next we need to compute $q_1(q_1^{-1}q_2)^\lambda$. The scalar part of this product is

$$\begin{aligned} s = \cos\theta\cos\lambda\Omega + \frac{\sin\lambda\Omega}{\sin\Omega}\sin^2\theta\cos\varphi(u\cdot u) - \frac{\sin\lambda\Omega}{\sin\Omega}\sin\theta\sin\varphi\cos\theta(u\cdot v)\\ + \frac{\sin\lambda\Omega}{\sin\Omega}\sin^2\theta\sin\varphi(u\cdot(u\times v)). \end{aligned}$$

Since $u\cdot(u\times v) = 0$, the last term is zero, and since $u\cdot u = 1$ and

$$\sin\theta\sin\varphi(u\cdot v) = \cos\Omega - \cos\theta\cos\varphi,$$

we get

$$\begin{aligned} s &= \cos\theta\cos\lambda\Omega + \frac{\sin\lambda\Omega}{\sin\Omega}\sin^2\theta\cos\varphi - \frac{\sin\lambda\Omega}{\sin\Omega}\cos\theta(\cos\Omega - \cos\theta\cos\varphi)\\ &= \cos\theta\cos\lambda\Omega + \frac{\sin\lambda\Omega}{\sin\Omega}(\sin^2\theta + \cos^2\theta)\cos\varphi - \frac{\sin\lambda\Omega}{\sin\Omega}\cos\theta\cos\Omega\\ &= \frac{(\cos\lambda\Omega\sin\Omega - \sin\lambda\Omega\cos\Omega)\cos\theta}{\sin\Omega} + \frac{\sin\lambda\Omega}{\sin\Omega}\cos\varphi\\ &= \frac{\sin(1-\lambda)\Omega}{\sin\Omega}\cos\theta + \frac{\sin\lambda\Omega}{\sin\Omega}\cos\varphi. \end{aligned}$$

The vector part of the product $q_1(q_1^{-1}q_2)^\lambda$ is given by

$$\begin{aligned}\nu = &-\frac{\sin\lambda\Omega}{\sin\Omega}\cos\theta\sin\theta\cos\varphi\, u + \frac{\sin\lambda\Omega}{\sin\Omega}\cos^2\theta\sin\varphi\, v\\ &- \frac{\sin\lambda\Omega}{\sin\Omega}\cos\theta\sin\theta\sin\varphi(u\times v) + \cos\lambda\Omega\sin\theta\, u\\ &- \frac{\sin\lambda\Omega}{\sin\Omega}\sin^2\theta\cos\varphi(u\times u) + \frac{\sin\lambda\Omega}{\sin\Omega}\cos\theta\sin\theta\sin\varphi(u\times v)\\ &- \frac{\sin\lambda\Omega}{\sin\Omega}\sin^2\theta\sin\varphi(u\times(u\times v)).\end{aligned}$$

We have $u\times u = 0$, the two terms involving $u\times v$ cancel out,

$$u\times(u\times v) = (u\cdot v)u - (u\cdot u)v,$$

and $u\cdot u = 1$, so we get

$$\begin{aligned}\nu = &-\frac{\sin\lambda\Omega}{\sin\Omega}\cos\theta\sin\theta\cos\varphi\, u + \cos\lambda\Omega\sin\theta\, u + \frac{\sin\lambda\Omega}{\sin\Omega}\cos^2\theta\sin\varphi\, v\\ &+ \frac{\sin\lambda\Omega}{\sin\Omega}\sin^2\theta\sin\varphi\, v - \frac{\sin\lambda\Omega}{\sin\Omega}\sin^2\theta\sin\varphi(u\cdot v)u.\end{aligned}$$

Using

$$\sin\theta\sin\varphi(u\cdot v) = \cos\Omega - \cos\theta\cos\varphi,$$

we get

$$\begin{aligned}\nu &= -\frac{\sin\lambda\Omega}{\sin\Omega}\cos\theta\sin\theta\cos\varphi\, u + \cos\lambda\Omega\sin\theta\, u + \frac{\sin\lambda\Omega}{\sin\Omega}\sin\varphi\, v\\ &\quad - \frac{\sin\lambda\Omega}{\sin\Omega}\sin\theta(\cos\Omega - \cos\theta\cos\varphi)u\\ &= \cos\lambda\Omega\sin\theta\, u + \frac{\sin\lambda\Omega}{\sin\Omega}\sin\varphi\, v - \frac{\sin\lambda\Omega}{\sin\Omega}\sin\theta\cos\Omega\, u\\ &= \frac{(\cos\lambda\Omega\sin\Omega - \sin\lambda\Omega\cos\Omega)}{\sin\Omega}\sin\theta\, u + \frac{\sin\lambda\Omega}{\sin\Omega}\sin\varphi\, v\\ &= \frac{\sin(1-\lambda)\Omega}{\sin\Omega}\sin\theta\, u + \frac{\sin\lambda\Omega}{\sin\Omega}\sin\varphi\, v.\end{aligned}$$

Putting the scalar part and the vector part together, we obtain

$$\begin{aligned}q_1(q_1^{-1}q_2)^\lambda = &\left[\frac{\sin(1-\lambda)\Omega}{\sin\Omega}\cos\theta + \frac{\sin\lambda\Omega}{\sin\Omega}\cos\varphi,\right.\\ &\left.\frac{\sin(1-\lambda)\Omega}{\sin\Omega}\sin\theta\, u + \frac{\sin\lambda\Omega}{\sin\Omega}\sin\varphi\, v\right],\\ &= \frac{\sin(1-\lambda)\Omega}{\sin\Omega}[\cos\theta,\, \sin\theta\, u] + \frac{\sin\lambda\Omega}{\sin\Omega}[\cos\varphi,\, \sin\varphi\, v].\end{aligned}$$

This yields the celebrated *slerp interpolation formula*

$$Z(\lambda) = q_1(q_1^{-1}q_2)^\lambda = \frac{\sin(1-\lambda)\Omega}{\sin\Omega}q_1 + \frac{\sin\lambda\Omega}{\sin\Omega}q_2,$$

with

$$\cos\Omega = \cos\theta\cos\varphi + \sin\theta\sin\varphi(u\cdot v).$$

15.7 Nonexistence of a "Nice" Section from SO(3) to SU(2)

We conclude by discussing the problem of a consistent choice of sign for the quaternion q representing a rotation $R = \rho_q \in \mathbf{SO}(3)$. We are looking for a "nice" section $s\colon \mathbf{SO}(3) \to \mathbf{SU}(2)$, that is, a function s satisfying the condition

$$\rho \circ s = \mathrm{id},$$

where ρ is the surjective homomorphism $\rho\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$.

Proposition 15.6. *Any section* $s\colon \mathbf{SO}(3) \to \mathbf{SU}(2)$ *of* ρ *is neither a homomorphism nor continuous.*

Intuitively, this means that there is no "nice and simple " way to pick the sign of the quaternion representing a rotation.

The following proof is due to Marcel Berger.

Proof. Let Γ be the subgroup of $\mathbf{SU}(2)$ consisting of all quaternions of the form $q = [a, (b, 0, 0)]$. Then, using the formula for the rotation matrix R_q corresponding to q (and the fact that $a^2 + b^2 = 1$), we get

$$R_q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2a^2 - 1 & -2ab \\ 0 & 2ab & 2a^2 - 1 \end{pmatrix}.$$

Since $a^2 + b^2 = 1$, we may write $a = \cos\theta, b = \sin\theta$, and we see that

$$R_q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos 2\theta & -\sin 2\theta \\ 0 & \sin 2\theta & \cos 2\theta \end{pmatrix},$$

a rotation of angle 2θ around the x-axis. Thus, both Γ and its image are isomorphic to $\mathbf{SO}(2)$, which is also isomorphic to $\mathbf{U}(1) = \{w \in \mathbb{C} \mid |w| = 1\}$. By identifying $\mathbf{i}$ and i, and identifying Γ and its image to $\mathbf{U}(1)$, if we write $w = \cos\theta + i\sin\theta \in \Gamma$, the restriction of the map ρ to Γ is given by $\rho(w) = w^2$.

We claim that any section s of ρ is not a homomorphism. Consider the restriction of s to $\mathbf{U}(1)$. Then since $\rho \circ s = \mathrm{id}$ and $\rho(w) = w^2$, for $-1 \in \rho(\Gamma) \approx \mathbf{U}(1)$, we have

$$-1 = \rho(s(-1)) = (s(-1))^2.$$

On the other hand, if s is a homomorphism, then

$$(s(-1))^2 = s((-1)^2) = s(1) = 1,$$

contradicting $(s(-1))^2 = -1$.

We also claim that s is not continuous. Assume that $s(1) = 1$, the case where $s(1) = -1$ being analogous. Then s is a bijection inverting ρ on Γ whose restriction to $\mathbf{U}(1)$ must be given by

$$s(\cos\theta + i\sin\theta) = \cos(\theta/2) + \mathbf{i}\sin(\theta/2), \quad -\pi \leq \theta < \pi.$$

If θ tends to π, that is $z = \cos\theta + i\sin\theta$ tends to -1 in the upper-half plane, then $s(z)$ tends to $\mathbf{i}$, but if θ tends to $-\pi$, that is z tends to -1 in the lower-half plane, then $s(z)$ tends to $-\mathbf{i}$, which shows that s is not continuous. □

Another way (due to Jean Dieudonné) to prove that a section s of ρ is not a homomorphism is to prove that any unit quaternion is the product of two unit pure quaternions. Indeed, if $q = [a, u]$ is a unit quaternion, if we let $q_1 = [0, u_1]$, where u_1 is any unit vector orthogonal to u, then

$$q_1 q = [-u_1 \cdot u, au_1 + u_1 \times u] = [0, au_1 + u_1 \times u] = q_2$$

is a nonzero unit pure quaternion. This is because if $a \neq 0$ then $au_1 + u_1 \times u \neq 0$ (since $u_1 \times u$ is orthogonal to $au_1 \neq 0$), and if $a = 0$ then $u \neq 0$, so $u_1 \times u \neq 0$ (since u_1 is orthogonal to u). But then, $q_1^{-1} = [0, -u_1]$ is a unit pure quaternion and we have

$$q = q_1^{-1} q_2,$$

a product of two pure unit quaternions.

We also observe that for any two pure quaternions q_1, q_2, there is some unit quaternion q such that

$$q_2 = q q_1 q^{-1}.$$

This is just a restatement of the fact that the group $\mathbf{SO}(3)$ is transitive. Since the kernel of $\rho\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ is $\{I, -I\}$, the subgroup $s(\mathbf{SO}(3))$ would be a normal subgroup of index 2 in $\mathbf{SU}(2)$. Then we would have a surjective homomorphism η from $\mathbf{SU}(2)$ onto the quotient group $\mathbf{SU}(2)/s(\mathbf{SO}(3))$, which is isomorphic to $\{1, -1\}$. Now, since any two pure quaternions are conjugate of each other, η would have a constant value on the unit pure quaternions. Since $\mathbf{k} = \mathbf{ij}$, we would have

$$\eta(\mathbf{k}) = \eta(\mathbf{ij}) = (\eta(\mathbf{i}))^2 = 1.$$

Consequently, η would map all pure unit quaternions to 1. But since every unit quaternion is the product of two pure quaternions, η would map every unit quaternion to 1, contradicting the fact that it is surjective onto $\{-1, 1\}$.

15.8 Summary

The main concepts and results of this chapter are listed below:

- The group $\mathbf{SU}(2)$ of unit quaternions.
- The skew field $\mathbb{H}$ of quaternions.
- Hamilton's identities.
- The (real) vector space $\mathfrak{su}(2)$ of 2×2 skew Hermitian matrices with zero trace.
- The adjoint representation of $\mathbf{SU}(2)$.
- The (real) vector space $\mathfrak{su}(2)$ of 2×2 Hermitian matrices with zero trace.
- The group homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$; $\mathrm{Ker}\,(r) = \{+I, -I\}$.
- The matrix representation R_q of the rotation r_q induced by a unit quaternion q.
- Surjectivity of the homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$.
- The exponential map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$.
- Surjectivity of the exponential map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$.
- Finding a logarithm of a quaternion.
- Quaternion interpolation.
- Shoemake's slerp interpolation formula.
- Sections $s\colon \mathbf{SO}(3) \to \mathbf{SU}(2)$ of $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$.

15.9 Problems

Problem 15.1. Verify the quaternion identities

$$\begin{aligned}
&\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -\mathbf{1},\\
&\mathbf{ij} = -\mathbf{ji} = \mathbf{k},\\
&\mathbf{jk} = -\mathbf{kj} = \mathbf{i},\\
&\mathbf{ki} = -\mathbf{ik} = \mathbf{j}.
\end{aligned}$$

Problem 15.2. Check that for every quaternion $X = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$, we have

$$XX^* = X^*X = (a^2 + b^2 + c^2 + d^2)\mathbf{1}.$$

Conclude that if $X \neq 0$, then X is invertible and its inverse is given by

$$X^{-1} = (a^2 + b^2 + c^2 + d^2)^{-1}X^*.$$

Problem 15.3. Given any two quaternions $X = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ and $Y = a'\mathbf{1} + b'\mathbf{i} + c'\mathbf{j} + d'\mathbf{k}$, prove that

$$\begin{aligned} XY = &(aa' - bb' - cc' - dd')\mathbf{1} + (ab' + ba' + cd' - dc')\mathbf{i} \\ &+ (ac' + ca' + db' - bd')\mathbf{j} + (ad' + da' + bc' - cb')\mathbf{k}. \end{aligned}$$

Also prove that if $X = [a, U]$ and $Y = [a', U']$, the quaternion product XY can be expressed as

$$XY = [aa' - U \cdot U',\ aU' + a'U + U \times U'].$$

Problem 15.4. Let $\mathrm{Ad}\colon \mathbf{SU}(2) \to \mathbf{GL}(\mathfrak{su}(2))$ be the map defined such that for every $q \in \mathbf{SU}(2)$,

$$\mathrm{Ad}_q(A) = qAq^*, \quad A \in \mathfrak{su}(2),$$

where q^* is the inverse of q (since $\mathbf{SU}(2)$ is a unitary group) Prove that the map Ad_q is an invertible linear map from $\mathfrak{su}(2)$ to itself and that Ad is a group homomorphism.

Problem 15.5. Prove that every Hermitian matrix with zero trace is of the form $x\sigma_3 + y\sigma_2 + z\sigma_1$, with

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Check that $\mathbf{i} = i\sigma_3$, $\mathbf{j} = i\sigma_2$, and that $\mathbf{k} = i\sigma_1$.

Problem 15.6. If

$$B = \begin{pmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix},$$

and if we write $\theta = \sqrt{u_1^2 + u_2^2 + u_3^2}$ (with $0 \leq \theta \leq \pi$), then the Rodrigues formula says that

$$e^B = I + \frac{\sin\theta}{\theta} B + \frac{(1 - \cos\theta)}{\theta^2} B^2, \quad \theta \neq 0,$$

with $e^0 = I$. Check that $\mathrm{tr}(e^B) = 1 + 2\cos\theta$. Prove that the quaternion q corresponding to the rotation $R = e^B$ (with $B \neq 0$) is given by

$$q = \left[\cos\left(\frac{\theta}{2}\right), \sin\left(\frac{\theta}{2}\right)\left(\frac{u_1}{\theta}, \frac{u_2}{\theta}, \frac{u_3}{\theta}\right)\right].$$

Problem 15.7. For every matrix $A \in \mathfrak{su}(2)$, with

$$A = \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix},$$

prove that if we write $\theta = \sqrt{u_1^2 + u_2^2 + u_3^2}$, then

$$e^A = \cos\theta I + \frac{\sin\theta}{\theta} A, \quad \theta \neq 0,$$

and $e^0 = I$. Conclude that e^A is a unit quaternion representing the rotation of angle 2θ and axis (u_1, u_2, u_3) (or I when $\theta = k\pi$, $k \in \mathbb{Z}$).

Problem 15.8. Write a `Matlab` program implementing the method of Section 15.4 for finding a unit quaternion corresponding to a rotation matrix.

Problem 15.9. Show that there is a very simple method for producing an orthonormal frame in $\mathbb{R}^4$ whose first vector is any given nonnull vector (a, b, c, d).

Problem 15.10. Let i, j, and k, be the unit vectors of coordinates $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ in $\mathbb{R}^3$.

(1) Describe geometrically the rotations defined by the following quaternions:

$$p = (0, i), \quad q = (0, j).$$

Prove that the interpolant $Z(\lambda) = p(p^{-1}q)^\lambda$ is given by

$$Z(\lambda) = (0,\, \cos(\lambda\pi/2)i + \sin(\lambda\pi/2)j)\,.$$

Describe geometrically what this rotation is.

(2) Repeat Question (1) with the rotations defined by the quaternions

$$p = \left(\frac{1}{2}, \frac{\sqrt{3}}{2} i\right), \quad q = (0, j).$$

Prove that the interpolant $Z(\lambda)$ is given by

$$Z(\lambda) = \left(\frac{1}{2}\cos(\lambda\pi/2),\, \frac{\sqrt{3}}{2}\cos(\lambda\pi/2)i + \sin(\lambda\pi/2)j\right).$$

Describe geometrically what this rotation is.

(3) Repeat Question (1) with the rotations defined by the quaternions

$$p = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} i\right), \quad q = \left(0, \frac{1}{\sqrt{2}}(i + j)\right).$$

Prove that the interpolant $Z(\lambda)$ is given by

$$Z(\lambda) = \left(\frac{1}{\sqrt{2}}\cos(\lambda\pi/3) - \frac{1}{\sqrt{6}}\sin(\lambda\pi/3),\right.$$
$$\left.(1/\sqrt{2}\cos(\lambda\pi/3) + 1/\sqrt{6}\sin(\lambda\pi/3))i + \frac{2}{\sqrt{6}}\sin(\lambda\pi/3)j\right).$$

Problem 15.11. Prove that

$$w \times (u \times v) = (w \cdot v)u - (u \cdot w)v.$$

Conclude that

$$u \times (u \times v) = (u \cdot v)u - (u \cdot u)v.$$

Chapter 16

Spectral Theorems in Euclidean and Hermitian Spaces

16.1 Introduction

The goal of this chapter is to show that there are nice normal forms for symmetric matrices, skew-symmetric matrices, orthogonal matrices, and normal matrices. The spectral theorem for symmetric matrices states that symmetric matrices have real eigenvalues and that they can be diagonalized over an orthonormal basis. The spectral theorem for Hermitian matrices states that Hermitian matrices also have real eigenvalues and that they can be diagonalized over a complex orthonormal basis. Normal real matrices can be block diagonalized over an orthonormal basis with blocks having size at most two and there are refinements of this normal form for skew-symmetric and orthogonal matrices.

The spectral result for real symmetric matrices can be used to prove two characterizations of the eigenvalues of a symmetric matrix in terms of the *Rayleigh ratio*. The first characterization is the *Rayleigh–Ritz theorem* and the second one is the *Courant–Fischer theorem*. Both results are used in optimization theory and to obtain results about perturbing the eigenvalues of a symmetric matrix.

In this chapter all vector spaces are finite-dimensional real or complex vector spaces.

16.2 Normal Linear Maps: Eigenvalues and Eigenvectors

We begin by studying normal maps, to understand the structure of their eigenvalues and eigenvectors. This section and the next three were inspired by Lang [Lang (1993)], Artin [Artin (1991)], Mac Lane and Birkhoff [Mac Lane and Birkhoff (1967)], Berger [Berger (1990a)], and Bertin [Bertin

(1981)].

Definition 16.1. Given a Euclidean or Hermitian space E, a linear map $f\colon E \to E$ is *normal* if

$$f \circ f^* = f^* \circ f.$$

A linear map $f\colon E \to E$ is *self-adjoint* if $f = f^*$, *skew-self-adjoint* if $f = -f^*$, and *orthogonal* if $f \circ f^* = f^* \circ f = \mathrm{id}$.

Obviously, a self-adjoint, skew-self-adjoint, or orthogonal linear map is a normal linear map. Our first goal is to show that for every normal linear map $f\colon E \to E$, there is an orthonormal basis (w.r.t. $\langle -, - \rangle$) such that the matrix of f over this basis has an especially nice form: it is a block diagonal matrix in which the blocks are either one-dimensional matrices (i.e., single entries) or two-dimensional matrices of the form

$$\begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix}.$$

This normal form can be further refined if f is self-adjoint, skew-self-adjoint, or orthogonal. As a first step we show that f and f^* have the same kernel when f is normal.

Proposition 16.1. *Given a Euclidean space E, if $f\colon E \to E$ is a normal linear map, then* $\operatorname{Ker} f = \operatorname{Ker} f^*$.

Proof. First let us prove that

$$\langle f(u), f(v) \rangle = \langle f^*(u), f^*(v) \rangle$$

for all $u, v \in E$. Since f^* is the adjoint of f and $f \circ f^* = f^* \circ f$, we have

$$\begin{aligned} \langle f(u), f(u) \rangle &= \langle u, (f^* \circ f)(u) \rangle, \\ &= \langle u, (f \circ f^*)(u) \rangle, \\ &= \langle f^*(u), f^*(u) \rangle. \end{aligned}$$

Since $\langle -, - \rangle$ is positive definite,

$$\begin{aligned} \langle f(u), f(u) \rangle = 0 \quad &\text{iff} \quad f(u) = 0, \\ \langle f^*(u), f^*(u) \rangle = 0 \quad &\text{iff} \quad f^*(u) = 0, \end{aligned}$$

and since

$$\langle f(u), f(u) \rangle = \langle f^*(u), f^*(u) \rangle,$$

we have

$$f(u) = 0 \quad \text{iff} \quad f^*(u) = 0.$$

Consequently, $\operatorname{Ker} f = \operatorname{Ker} f^*$. □

Assuming again that E is a Hermitian space, observe that Proposition 16.1 also holds. We deduce the following corollary.

Proposition 16.2. *Given a Hermitian space E, for any normal linear map $f: E \to E$, we have* $\mathrm{Ker}\,(f) \cap \mathrm{Im}(f) = (0)$.

Proof. Assume $v \in \mathrm{Ker}\,(f) \cap \mathrm{Im}(f)$, which means that $v = f(u)$ for some $u \in E$, and $f(v) = 0$. By Proposition 16.1, $\mathrm{Ker}\,(f) = \mathrm{Ker}\,(f^*)$, so $f(v) = 0$ implies that $f^*(v) = 0$. Consequently,

$$\begin{aligned} 0 &= \langle f^*(v), u\rangle \\ &= \langle v, f(u)\rangle \\ &= \langle v, v\rangle, \end{aligned}$$

and thus, $v = 0$. □

We also have the following crucial proposition relating the eigenvalues of f and f^*.

Proposition 16.3. *Given a Hermitian space E, for any normal linear map $f: E \to E$, a vector u is an eigenvector of f for the eigenvalue λ (in $\mathbb{C}$) iff u is an eigenvector of f^* for the eigenvalue $\overline{\lambda}$.*

Proof. First it is immediately verified that the adjoint of $f - \lambda\,\mathrm{id}$ is $f^* - \overline{\lambda}\,\mathrm{id}$. Furthermore, $f - \lambda\,\mathrm{id}$ is normal. Indeed,

$$\begin{aligned} (f - \lambda\,\mathrm{id}) \circ (f - \lambda\,\mathrm{id})^* &= (f - \lambda\,\mathrm{id}) \circ (f^* - \overline{\lambda}\,\mathrm{id}), \\ &= f \circ f^* - \overline{\lambda} f - \lambda f^* + \lambda\overline{\lambda}\,\mathrm{id}, \\ &= f^* \circ f - \lambda f^* - \overline{\lambda} f + \overline{\lambda}\lambda\,\mathrm{id}, \\ &= (f^* - \overline{\lambda}\,\mathrm{id}) \circ (f - \lambda\,\mathrm{id}), \\ &= (f - \lambda\,\mathrm{id})^* \circ (f - \lambda\,\mathrm{id}). \end{aligned}$$

Applying Proposition 16.1 to $f - \lambda\,\mathrm{id}$, for every nonnull vector u, we see that

$$(f - \lambda\,\mathrm{id})(u) = 0 \quad \text{iff} \quad (f^* - \overline{\lambda}\,\mathrm{id})(u) = 0,$$

which is exactly the statement of the proposition. □

The next proposition shows a very important property of normal linear maps: eigenvectors corresponding to distinct eigenvalues are orthogonal.

Proposition 16.4. *Given a Hermitian space E, for any normal linear map $f: E \to E$, if u and v are eigenvectors of f associated with the eigenvalues λ and μ (in $\mathbb{C}$) where $\lambda \neq \mu$, then $\langle u, v\rangle = 0$.*

Proof. Let us compute $\langle f(u), v\rangle$ in two different ways. Since v is an eigenvector of f for μ, by Proposition 16.3, v is also an eigenvector of f^* for $\overline{\mu}$, and we have

$$\langle f(u), v\rangle = \langle \lambda u, v\rangle = \lambda\langle u, v\rangle,$$

and

$$\langle f(u), v\rangle = \langle u, f^*(v)\rangle = \langle u, \overline{\mu} v\rangle = \mu\langle u, v\rangle,$$

where the last identity holds because of the semilinearity in the second argument. Thus

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

that is,

$$(\lambda - \mu)\langle u, v\rangle = 0,$$

which implies that $\langle u, v\rangle = 0$, since $\lambda \neq \mu$. □

We can show easily that the eigenvalues of a self-adjoint linear map are real.

Proposition 16.5. *Given a Hermitian space E, all the eigenvalues of any self-adjoint linear map $f\colon E \to E$ are real.*

Proof. Let z (in $\mathbb{C}$) be an eigenvalue of f and let u be an eigenvector for z. We compute $\langle f(u), u\rangle$ in two different ways. We have

$$\langle f(u), u\rangle = \langle zu, u\rangle = z\langle u, u\rangle,$$

and since $f = f^*$, we also have

$$\langle f(u), u\rangle = \langle u, f^*(u)\rangle = \langle u, f(u)\rangle = \langle u, zu\rangle = \overline{z}\langle u, u\rangle.$$

Thus,

$$z\langle u, u\rangle = \overline{z}\langle u, u\rangle,$$

which implies that $z = \overline{z}$, since $u \neq 0$, and z is indeed real. □

There is also a version of Proposition 16.5 for a (real) Euclidean space E and a self-adjoint map $f\colon E \to E$ since every real vector space E can be embedded into a complex vector space $E_{\mathbb{C}}$, and every linear map $f\colon E \to E$ can be extended to a linear map $f_{\mathbb{C}}\colon E_{\mathbb{C}} \to E_{\mathbb{C}}$.

Definition 16.2. Given a real vector space E, let $E_{\mathbb{C}}$ be the structure $E \times E$ under the addition operation

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

and let multiplication by a complex scalar $z = x + iy$ be defined such that

$$(x + iy) \cdot (u, v) = (xu - yv, yu + xv).$$

The space $E_{\mathbb{C}}$ is called the *complexification* of E.

It is easily shown that the structure $E_{\mathbb{C}}$ is a complex vector space. It is also immediate that

$$(0,\, v) = i(v,\, 0),$$

and thus, identifying E with the subspace of $E_{\mathbb{C}}$ consisting of all vectors of the form $(u,\, 0)$, we can write

$$(u,\, v) = u + iv.$$

Observe that if $(e_1, \ldots, e_n)$ is a basis of E (a real vector space), then $(e_1, \ldots, e_n)$ is also a basis of $E_{\mathbb{C}}$ (recall that e_i is an abbreviation for $(e_i, 0)$).

A linear map $f\colon E \to E$ is extended to the linear map $f_{\mathbb{C}}\colon E_{\mathbb{C}} \to E_{\mathbb{C}}$ defined such that

$$f_{\mathbb{C}}(u + iv) = f(u) + if(v).$$

For any basis $(e_1, \ldots, e_n)$ of E, the matrix $M(f)$ representing f over $(e_1, \ldots, e_n)$ is *identical* to the matrix $M(f_{\mathbb{C}})$ representing $f_{\mathbb{C}}$ over $(e_1, \ldots, e_n)$, where we view $(e_1, \ldots, e_n)$ as a basis of $E_{\mathbb{C}}$. As a consequence, $\det(zI - M(f)) = \det(zI - M(f_{\mathbb{C}}))$, which means that f and $f_{\mathbb{C}}$ have the *same* characteristic polynomial (which has real coefficients). We know that every polynomial of degree n with real (or complex) coefficients always has n complex roots (counted with their multiplicity), and the roots of $\det(zI - M(f_{\mathbb{C}}))$ that are real (if any) are the eigenvalues of f.

Next we need to extend the inner product on E to an inner product on $E_{\mathbb{C}}$.

The inner product $\langle -, - \rangle$ on a Euclidean space E is extended to the Hermitian positive definite form $\langle -, - \rangle_{\mathbb{C}}$ on $E_{\mathbb{C}}$ as follows:

$$\langle u_1 + iv_1, u_2 + iv_2 \rangle_{\mathbb{C}} = \langle u_1, u_2 \rangle + \langle v_1, v_2 \rangle + i(\langle v_1, u_2 \rangle - \langle u_1, v_2 \rangle).$$

It is easily verified that $\langle -, - \rangle_{\mathbb{C}}$ is indeed a Hermitian form that is positive definite, and it is clear that $\langle -, - \rangle_{\mathbb{C}}$ agrees with $\langle -, - \rangle$ on real vectors. Then given any linear map $f\colon E \to E$, it is easily verified that the map $f^*_{\mathbb{C}}$ defined such that

$$f^*_{\mathbb{C}}(u + iv) = f^*(u) + if^*(v)$$

for all $u, v \in E$ is the adjoint of $f_{\mathbb{C}}$ w.r.t. $\langle -, - \rangle_{\mathbb{C}}$.

Proposition 16.6. *Given a Euclidean space E, if $f\colon E \to E$ is any self-adjoint linear map, then every eigenvalue λ of $f_{\mathbb{C}}$ is real and is actually an eigenvalue of f (which means that there is some real eigenvector $u \in E$ such that $f(u) = \lambda u$). Therefore, all the eigenvalues of f are real.*

Proof. Let $E_{\mathbb{C}}$ be the complexification of E, $\langle -, - \rangle_{\mathbb{C}}$ the complexification of the inner product $\langle -, - \rangle$ on E, and $f_{\mathbb{C}} \colon E_{\mathbb{C}} \to E_{\mathbb{C}}$ the complexification of $f \colon E \to E$. By definition of $f_{\mathbb{C}}$ and $\langle -, - \rangle_{\mathbb{C}}$, if f is self-adjoint, we have

$$\begin{aligned}
\langle f_{\mathbb{C}}(u_1 + iv_1), u_2 + iv_2 \rangle_{\mathbb{C}} &= \langle f(u_1) + if(v_1), u_2 + iv_2 \rangle_{\mathbb{C}} \\
&= \langle f(u_1), u_2 \rangle + \langle f(v_1), v_2 \rangle \\
&\quad + i(\langle u_2, f(v_1) \rangle - \langle f(u_1), v_2 \rangle) \\
&= \langle u_1, f(u_2) \rangle + \langle v_1, f(v_2) \rangle \\
&\quad + i(\langle f(u_2), v_1 \rangle - \langle u_1, f(v_2) \rangle) \\
&= \langle u_1 + iv_1, f(u_2) + if(v_2) \rangle_{\mathbb{C}} \\
&= \langle u_1 + iv_1, f_{\mathbb{C}}(u_2 + iv_2) \rangle_{\mathbb{C}},
\end{aligned}$$

which shows that $f_{\mathbb{C}}$ is also self-adjoint with respect to $\langle -, - \rangle_{\mathbb{C}}$.

As we pointed out earlier, f and $f_{\mathbb{C}}$ have the same characteristic polynomial $\det(zI - f_{\mathbb{C}}) = \det(zI - f)$, which is a polynomial with real coefficients. Proposition 16.5 shows that the zeros of $\det(zI - f_{\mathbb{C}}) = \det(zI - f)$ are all real, and for each real zero λ of $\det(zI - f)$, the linear map $\lambda \mathrm{id} - f$ is singular, which means that there is some nonzero $u \in E$ such that $f(u) = \lambda u$. Therefore, all the eigenvalues of f are real. □

Proposition 16.7. *Given a Hermitian space E, for any linear map $f \colon E \to E$, if f is skew-self-adjoint, then f has eigenvalues that are pure imaginary or zero, and if f is unitary, then f has eigenvalues of absolute value* 1.

Proof. If f is skew-self-adjoint, $f^* = -f$, and then by the definition of the adjoint map, for any eigenvalue λ and any eigenvector u associated with λ, we have

$$\begin{aligned}
\lambda \langle u, u \rangle = \langle \lambda u, u \rangle = \langle f(u), u \rangle = \langle u, f^*(u) \rangle = \langle u, -f(u) \rangle \\
= -\langle u, \lambda u \rangle = -\overline{\lambda} \langle u, u \rangle,
\end{aligned}$$

and since $u \neq 0$ and $\langle -, - \rangle$ is positive definite, $\langle u, u \rangle \neq 0$, so

$$\lambda = -\overline{\lambda},$$

which shows that $\lambda = ir$ for some $r \in \mathbb{R}$.

If f is unitary, then f is an isometry, so for any eigenvalue λ and any eigenvector u associated with λ, we have

$$|\lambda|^2 \langle u, u \rangle = \lambda \overline{\lambda} \langle u, u \rangle = \langle \lambda u, \lambda u \rangle = \langle f(u), f(u) \rangle = \langle u, u \rangle,$$

and since $u \neq 0$, we obtain $|\lambda|^2 = 1$, which implies

$$|\lambda| = 1.$$

□

16.3 Spectral Theorem for Normal Linear Maps

Given a Euclidean space E, our next step is to show that for every linear map $f: E \to E$ there is some subspace W of dimension 1 or 2 such that $f(W) \subseteq W$. When $\dim(W) = 1$, the subspace W is actually an eigenspace for some real eigenvalue of f. Furthermore, when f is normal, there is a subspace W of dimension 1 or 2 such that $f(W) \subseteq W$ **and** $f^*(W) \subseteq W$. The difficulty is that the eigenvalues of f are not necessarily real. One way to get around this problem is to complexify both the vector space E and the inner product $\langle -, - \rangle$ as we did in Section 16.2.

Given any subspace W of a Euclidean space E, recall that the *orthogonal complement* $W^\perp$ of W is the subspace defined such that

$$W^\perp = \{u \in E \mid \langle u, w \rangle = 0, \text{ for all } w \in W\}.$$

Recall from Proposition 11.9 that $E = W \oplus W^\perp$ (this can be easily shown, for example, by constructing an orthonormal basis of E using the Gram–Schmidt orthonormalization procedure). The same result also holds for Hermitian spaces; see Proposition 13.12.

As a warm up for the proof of Theorem 16.2, let us prove that every self-adjoint map on a Euclidean space can be diagonalized with respect to an orthonormal basis of eigenvectors.

Theorem 16.1. *(Spectral theorem for self-adjoint linear maps on a Euclidean space) Given a Euclidean space E of dimension n, for every self-adjoint linear map $f: E \to E$, there is an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors of f such that the matrix of f w.r.t. this basis is a diagonal matrix*

$$\begin{pmatrix} \lambda_1 & & \cdots & \\ & \lambda_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \lambda_n \end{pmatrix},$$

with $\lambda_i \in \mathbb{R}$.

Proof. We proceed by induction on the dimension n of E as follows. If $n = 1$, the result is trivial. Assume now that $n \geq 2$. From Proposition 16.6, all the eigenvalues of f are real, so pick some eigenvalue $\lambda \in \mathbb{R}$, and let w be some eigenvector for λ. By dividing w by its norm, we may assume that w is a unit vector. Let W be the subspace of dimension 1 spanned by w. Clearly, $f(W) \subseteq W$. We claim that $f(W^\perp) \subseteq W^\perp$, where $W^\perp$ is the orthogonal complement of W.

Indeed, for any $v \in W^\perp$, that is, if $\langle v, w\rangle = 0$, because f is self-adjoint and $f(w) = \lambda w$, we have

$$\begin{aligned}\langle f(v), w\rangle &= \langle v, f(w)\rangle \\ &= \langle v, \lambda w\rangle \\ &= \lambda\langle v, w\rangle = 0\end{aligned}$$

since $\langle v, w\rangle = 0$. Therefore,

$$f(W^\perp) \subseteq W^\perp.$$

Clearly, the restriction of f to $W^\perp$ is self-adjoint, and we conclude by applying the induction hypothesis to $W^\perp$ (whose dimension is $n-1$). □

We now come back to normal linear maps. One of the key points in the proof of Theorem 16.1 is that we found a subspace W with the property that $f(W) \subseteq W$ implies that $f(W^\perp) \subseteq W^\perp$. In general, this does not happen, *but normal maps satisfy a stronger property which ensures that such a subspace exists*.

The following proposition provides a condition that will allow us to show that a normal linear map can be diagonalized. It actually holds for any linear map. We found the inspiration for this proposition in Berger [Berger (1990a)].

Proposition 16.8. *Given a Hermitian space E, for any linear map $f\colon E \to E$ and any subspace W of E, if $f(W) \subseteq W$, then $f^*\big(W^\perp\big) \subseteq W^\perp$. Consequently, if $f(W) \subseteq W$ and $f^*(W) \subseteq W$, then $f\big(W^\perp\big) \subseteq W^\perp$ and $f^*\big(W^\perp\big) \subseteq W^\perp$.*

Proof. If $u \in W^\perp$, then

$$\langle w, u\rangle = 0 \quad \text{for all } w \in W.$$

However,

$$\langle f(w), u\rangle = \langle w, f^*(u)\rangle,$$

and $f(W) \subseteq W$ implies that $f(w) \in W$. Since $u \in W^\perp$, we get

$$0 = \langle f(w), u\rangle = \langle w, f^*(u)\rangle,$$

which shows that $\langle w, f^*(u)\rangle = 0$ for all $w \in W$, that is, $f^*(u) \in W^\perp$. Therefore, we have $f^*(W^\perp) \subseteq W^\perp$.

We just proved that if $f(W) \subseteq W$, then $f^*\big(W^\perp\big) \subseteq W^\perp$. If we also have $f^*(W) \subseteq W$, then by applying the above fact to f^*, we get $f^{**}(W^\perp) \subseteq W^\perp$, and since $f^{**} = f$, this is just $f(W^\perp) \subseteq W^\perp$, which proves the second statement of the proposition. □

It is clear that the above proposition also holds for Euclidean spaces.

Although we are ready to prove that for every normal linear map f (over a Hermitian space) there is an orthonormal basis of eigenvectors (see Theorem 16.3 below), we now return to real Euclidean spaces.

Proposition 16.9. *If $f\colon E \to E$ is a linear map and $w = u + iv$ is an eigenvector of $f_{\mathbb{C}}\colon E_{\mathbb{C}} \to E_{\mathbb{C}}$ for the eigenvalue $z = \lambda + i\mu$, where $u, v \in E$ and $\lambda, \mu \in \mathbb{R}$, then*

$$f(u) = \lambda u - \mu v \quad \textit{and} \quad f(v) = \mu u + \lambda v. \tag{$*$}$$

As a consequence,

$$f_{\mathbb{C}}(u - iv) = f(u) - if(v) = (\lambda - i\mu)(u - iv),$$

which shows that $\overline{w} = u - iv$ is an eigenvector of $f_{\mathbb{C}}$ for $\overline{z} = \lambda - i\mu$.

Proof. Since

$$f_{\mathbb{C}}(u + iv) = f(u) + if(v)$$

and

$$f_{\mathbb{C}}(u + iv) = (\lambda + i\mu)(u + iv) = \lambda u - \mu v + i(\mu u + \lambda v),$$

we have

$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v. \qquad \square$$

Using this fact, we can prove the following proposition.

Proposition 16.10. *Given a Euclidean space E, for any normal linear map $f\colon E \to E$, if $w = u + iv$ is an eigenvector of $f_{\mathbb{C}}$ associated with the eigenvalue $z = \lambda + i\mu$ (where $u, v \in E$ and $\lambda, \mu \in \mathbb{R}$), if $\mu \neq 0$ (i.e., z is not real) then $\langle u, v\rangle = 0$ and $\langle u, u\rangle = \langle v, v\rangle$, which implies that u and v are linearly independent, and if W is the subspace spanned by u and v, then $f(W) = W$ and $f^*(W) = W$. Furthermore, with respect to the (orthogonal) basis (u, v), the restriction of f to W has the matrix*

$$\begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix}.$$

If $\mu = 0$, then λ is a real eigenvalue of f, and either u or v is an eigenvector of f for λ. If W is the subspace spanned by u if $u \neq 0$, or spanned by $v \neq 0$ if $u = 0$, then $f(W) \subseteq W$ and $f^*(W) \subseteq W$.

Proof. Since $w = u + iv$ is an eigenvector of $f_{\mathbb{C}}$, by definition it is nonnull, and either $u \neq 0$ or $v \neq 0$. Proposition 16.9 implies that $u - iv$ is an eigenvector of $f_{\mathbb{C}}$ for $\lambda - i\mu$. It is easy to check that $f_{\mathbb{C}}$ is normal. However, if $\mu \neq 0$, then $\lambda + i\mu \neq \lambda - i\mu$, and from Proposition 16.4, the vectors $u + iv$ and $u - iv$ are orthogonal w.r.t. $\langle -, - \rangle_{\mathbb{C}}$, that is,

$$\langle u + iv, u - iv \rangle_{\mathbb{C}} = \langle u, u \rangle - \langle v, v \rangle + 2i\langle u, v \rangle = 0.$$

Thus we get $\langle u, v \rangle = 0$ and $\langle u, u \rangle = \langle v, v \rangle$, and since $u \neq 0$ or $v \neq 0$, u and v are linearly independent. Since

$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v$$

and since by Proposition 16.3 $u + iv$ is an eigenvector of $f^*_{\mathbb{C}}$ for $\lambda - i\mu$, we have

$$f^*(u) = \lambda u + \mu v \quad \text{and} \quad f^*(v) = -\mu u + \lambda v,$$

and thus $f(W) = W$ and $f^*(W) = W$, where W is the subspace spanned by u and v.

When $\mu = 0$, we have

$$f(u) = \lambda u \quad \text{and} \quad f(v) = \lambda v,$$

and since $u \neq 0$ or $v \neq 0$, either u or v is an eigenvector of f for λ. If W is the subspace spanned by u if $u \neq 0$, or spanned by v if $u = 0$, it is obvious that $f(W) \subseteq W$ and $f^*(W) \subseteq W$. Note that $\lambda = 0$ is possible, and this is why $\subseteq$ cannot be replaced by $=$. □

The beginning of the proof of Proposition 16.10 actually shows that for every linear map $f : E \to E$ there is some subspace W such that $f(W) \subseteq W$, where W has dimension 1 or 2. In general, it doesn't seem possible to prove that $W^{\perp}$ is invariant under f. *However, this happens when f is normal.*

We can finally prove our first main theorem.

Theorem 16.2. *(Main spectral theorem) Given a Euclidean space E of dimension n, for every normal linear map $f : E \to E$, there is an orthonormal basis $(e_1, \ldots, e_n)$ such that the matrix of f w.r.t. this basis is a block diagonal matrix of the form*

$$\begin{pmatrix} A_1 & & \cdots & \\ & A_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & A_p \end{pmatrix}$$

such that each block A_j is either a one-dimensional matrix (i.e., a real scalar) or a two-dimensional matrix of the form

$$A_j = \begin{pmatrix} \lambda_j & -\mu_j \\ \mu_j & \lambda_j \end{pmatrix},$$

where $\lambda_j, \mu_j \in \mathbb{R}$, with $\mu_j > 0$.

Proof. We proceed by induction on the dimension n of E as follows. If $n = 1$, the result is trivial. Assume now that $n \geq 2$. First, since $\mathbb{C}$ is algebraically closed (i.e., every polynomial has a root in $\mathbb{C}$), the linear map $f_{\mathbb{C}} \colon E_{\mathbb{C}} \to E_{\mathbb{C}}$ has some eigenvalue $z = \lambda + i\mu$ (where $\lambda, \mu \in \mathbb{R}$). Let $w = u + iv$ be some eigenvector of $f_{\mathbb{C}}$ for $\lambda + i\mu$ (where $u, v \in E$). We can now apply Proposition 16.10.

If $\mu = 0$, then either u or v is an eigenvector of f for $\lambda \in \mathbb{R}$. Let W be the subspace of dimension 1 spanned by $e_1 = u/\|u\|$ if $u \neq 0$, or by $e_1 = v/\|v\|$ otherwise. It is obvious that $f(W) \subseteq W$ and $f^*(W) \subseteq W$. The orthogonal $W^\perp$ of W has dimension $n - 1$, and by Proposition 16.8, we have $f\big(W^\perp\big) \subseteq W^\perp$. But the restriction of f to $W^\perp$ is also normal, and we conclude by applying the induction hypothesis to $W^\perp$.

If $\mu \neq 0$, then $\langle u, v\rangle = 0$ and $\langle u, u\rangle = \langle v, v\rangle$, and if W is the subspace spanned by $u/\|u\|$ and $v/\|v\|$, then $f(W) = W$ and $f^*(W) = W$. We also know that the restriction of f to W has the matrix

$$\begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix}$$

with respect to the basis $(u/\|u\|, v/\|v\|)$. If $\mu < 0$, we let $\lambda_1 = \lambda$, $\mu_1 = -\mu$, $e_1 = u/\|u\|$, and $e_2 = v/\|v\|$. If $\mu > 0$, we let $\lambda_1 = \lambda$, $\mu_1 = \mu$, $e_1 = v/\|v\|$, and $e_2 = u/\|u\|$. In all cases, it is easily verified that the matrix of the restriction of f to W w.r.t. the orthonormal basis (e_1, e_2) is

$$A_1 = \begin{pmatrix} \lambda_1 & -\mu_1 \\ \mu_1 & \lambda_1 \end{pmatrix},$$

where $\lambda_1, \mu_1 \in \mathbb{R}$, with $\mu_1 > 0$. However, $W^\perp$ has dimension $n - 2$, and by Proposition 16.8, $f\big(W^\perp\big) \subseteq W^\perp$. Since the restriction of f to $W^\perp$ is also normal, we conclude by applying the induction hypothesis to $W^\perp$. □

After this relatively hard work, we can easily obtain some nice normal forms for the matrices of self-adjoint, skew-self-adjoint, and orthogonal linear maps. However, for the sake of completeness (and since we have all the tools to so do), we go back to the case of a Hermitian space and show that

normal linear maps can be diagonalized with respect to an orthonormal basis. The proof is a slight generalization of the proof of Theorem 16.6.

Theorem 16.3. *(Spectral theorem for normal linear maps on a Hermitian space) Given a Hermitian space E of dimension n, for every normal linear map $f: E \to E$ there is an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors of f such that the matrix of f w.r.t. this basis is a diagonal matrix*

$$\begin{pmatrix} \lambda_1 & & \cdots & \\ & \lambda_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \lambda_n \end{pmatrix},$$

where $\lambda_j \in \mathbb{C}$.

Proof. We proceed by induction on the dimension n of E as follows. If $n = 1$, the result is trivial. Assume now that $n \geq 2$. Since $\mathbb{C}$ is algebraically closed (i.e., every polynomial has a root in $\mathbb{C}$), the linear map $f: E \to E$ has some eigenvalue $\lambda \in \mathbb{C}$, and let w be some unit eigenvector for λ. Let W be the subspace of dimension 1 spanned by w. Clearly, $f(W) \subseteq W$. By Proposition 16.3, w is an eigenvector of f^* for $\overline{\lambda}$, and thus $f^*(W) \subseteq W$. By Proposition 16.8, we also have $f(W^\perp) \subseteq W^\perp$. The restriction of f to $W^\perp$ is still normal, and we conclude by applying the induction hypothesis to $W^\perp$ (whose dimension is $n - 1$). □

Theorem 16.3 implies that (complex) self-adjoint, skew-self-adjoint, and orthogonal linear maps can be diagonalized with respect to an orthonormal basis of eigenvectors. In this latter case, though, an orthogonal map is called a *unitary* map. Proposition 16.5 also shows that the eigenvalues of a self-adjoint linear map are real, and Proposition 16.7 shows that the eigenvalues of a skew self-adjoint map are pure imaginary or zero, and that the eigenvalues of a unitary map have absolute value 1.

Remark: There is a converse to Theorem 16.3, namely, if there is an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors of f, then f is normal. We leave the easy proof as an exercise.

In the next section we specialize Theorem 16.2 to self-adjoint, skew-self-adjoint, and orthogonal linear maps. Due to the additional structure, we obtain more precise normal forms.

16.4 Self-Adjoint, Skew-Self-Adjoint, and Orthogonal Linear Maps

We begin with self-adjoint maps.

Theorem 16.4. *Given a Euclidean space E of dimension n, for every self-adjoint linear map $f: E \to E$, there is an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors of f such that the matrix of f w.r.t. this basis is a diagonal matrix*

$$\begin{pmatrix} \lambda_1 & & \cdots & \\ & \lambda_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \lambda_n \end{pmatrix},$$

where $\lambda_i \in \mathbb{R}$.

Proof. We already proved this; see Theorem 16.1. However, it is instructive to give a more direct method not involving the complexification of $\langle -, - \rangle$ and Proposition 16.5.

Since $\mathbb{C}$ is algebraically closed, $f_{\mathbb{C}}$ has some eigenvalue $\lambda + i\mu$, and let $u + iv$ be some eigenvector of $f_{\mathbb{C}}$ for $\lambda + i\mu$, where $\lambda, \mu \in \mathbb{R}$ and $u, v \in E$. We saw in the proof of Proposition 16.9 that

$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v.$$

Since $f = f^*$,

$$\langle f(u), v \rangle = \langle u, f(v) \rangle$$

for all $u, v \in E$. Applying this to

$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v,$$

we get

$$\langle f(u), v \rangle = \langle \lambda u - \mu v, v \rangle = \lambda \langle u, v \rangle - \mu \langle v, v \rangle$$

and

$$\langle u, f(v) \rangle = \langle u, \mu u + \lambda v \rangle = \mu \langle u, u \rangle + \lambda \langle u, v \rangle,$$

and thus we get

$$\lambda \langle u, v \rangle - \mu \langle v, v \rangle = \mu \langle u, u \rangle + \lambda \langle u, v \rangle,$$

that is,

$$\mu(\langle u, u \rangle + \langle v, v \rangle) = 0,$$

which implies $\mu = 0$, since either $u \neq 0$ or $v \neq 0$. Therefore, λ is a real eigenvalue of f.

Now going back to the proof of Theorem 16.2, only the case where $\mu = 0$ applies, and the induction shows that all the blocks are one-dimensional. □

Theorem 16.4 implies that if $\lambda_1, \ldots, \lambda_p$ are the distinct real eigenvalues of f, and E_i is the eigenspace associated with λ_i, then

$$E = E_1 \oplus \cdots \oplus E_p,$$

where E_i and E_j are orthogonal for all $i \neq j$.

Remark: Another way to prove that a self-adjoint map has a real eigenvalue is to use a little bit of calculus. We learned such a proof from Herman Gluck. The idea is to consider the real-valued function $\Phi\colon E \to \mathbb{R}$ defined such that

$$\Phi(u) = \langle f(u), u\rangle$$

for every $u \in E$. This function is C^∞, and if we represent f by a matrix A over some orthonormal basis, it is easy to compute the gradient vector

$$\nabla\Phi(X) = \left(\frac{\partial\Phi}{\partial x_1}(X), \ldots, \frac{\partial\Phi}{\partial x_n}(X)\right)$$

of Φ at X. Indeed, we find that

$$\nabla\Phi(X) = (A + A^\top)X,$$

where X is a column vector of size n. But since f is self-adjoint, $A = A^\top$, and thus

$$\nabla\Phi(X) = 2AX.$$

The next step is to find the maximum of the function Φ on the sphere

$$S^{n-1} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \cdots + x_n^2 = 1\}.$$

Since S^{n-1} is compact and Φ is continuous, and in fact C^∞, Φ takes a maximum at some X on S^{n-1}. But then it is well known that at an extremum X of Φ we must have

$$d\Phi_X(Y) = \langle\nabla\Phi(X), Y\rangle = 0$$

for all tangent vectors Y to S^{n-1} at X, and so $\nabla\Phi(X)$ is orthogonal to the tangent plane at X, which means that

$$\nabla\Phi(X) = \lambda X$$

for some $\lambda \in \mathbb{R}$. Since $\nabla\Phi(X) = 2AX$, we get

$$2AX = \lambda X,$$

and thus $\lambda/2$ is a real eigenvalue of A (i.e., of f).

Next we consider skew-self-adjoint maps.

Theorem 16.5. *Given a Euclidean space E of dimension n, for every skew-self-adjoint linear map $f\colon E \to E$ there is an orthonormal basis $(e_1, \ldots, e_n)$ such that the matrix of f w.r.t. this basis is a block diagonal matrix of the form*

$$\begin{pmatrix} A_1 & & \cdots & \\ & A_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & A_p \end{pmatrix}$$

such that each block A_j is either 0 or a two-dimensional matrix of the form

$$A_j = \begin{pmatrix} 0 & -\mu_j \\ \mu_j & 0 \end{pmatrix},$$

where $\mu_j \in \mathbb{R}$, with $\mu_j > 0$. In particular, the eigenvalues of $f_{\mathbb{C}}$ are pure imaginary of the form $\pm i\mu_j$ or 0.

Proof. The case where $n = 1$ is trivial. As in the proof of Theorem 16.2, $f_{\mathbb{C}}$ has some eigenvalue $z = \lambda + i\mu$, where $\lambda, \mu \in \mathbb{R}$. We claim that $\lambda = 0$. First we show that

$$\langle f(w), w\rangle = 0$$

for all $w \in E$. Indeed, since $f = -f^*$, we get

$$\langle f(w), w\rangle = \langle w, f^*(w)\rangle = \langle w, -f(w)\rangle = -\langle w, f(w)\rangle = -\langle f(w), w\rangle,$$

since $\langle -, -\rangle$ is symmetric. This implies that

$$\langle f(w), w\rangle = 0.$$

Applying this to u and v and using the fact that

$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v,$$

we get

$$0 = \langle f(u), u\rangle = \langle \lambda u - \mu v, u\rangle = \lambda\langle u, u\rangle - \mu\langle u, v\rangle$$

and

$$0 = \langle f(v), v\rangle = \langle \mu u + \lambda v, v\rangle = \mu\langle u, v\rangle + \lambda\langle v, v\rangle,$$

from which, by addition, we get

$$\lambda(\langle v, v\rangle + \langle v, v\rangle) = 0.$$

Since $u \neq 0$ or $v \neq 0$, we have $\lambda = 0$.

Then going back to the proof of Theorem 16.2, unless $\mu = 0$, the case where u and v are orthogonal and span a subspace of dimension 2 applies, and the induction shows that all the blocks are two-dimensional or reduced to 0. □

Remark: One will note that if f is skew-self-adjoint, then $if_{\mathbb{C}}$ is self-adjoint w.r.t. $\langle -, -\rangle_{\mathbb{C}}$. By Proposition 16.5, the map $if_{\mathbb{C}}$ has real eigenvalues, which implies that the eigenvalues of $f_{\mathbb{C}}$ are pure imaginary or 0.

Finally we consider orthogonal linear maps.

Theorem 16.6. *Given a Euclidean space E of dimension n, for every orthogonal linear map $f\colon E \to E$ there is an orthonormal basis $(e_1, \ldots, e_n)$ such that the matrix of f w.r.t. this basis is a block diagonal matrix of the form*

$$\begin{pmatrix} A_1 & & \cdots & \\ & A_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & A_p \end{pmatrix}$$

such that each block A_j is either 1, -1, or a two-dimensional matrix of the form

$$A_j = \begin{pmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{pmatrix}$$

where $0 < \theta_j < \pi$. In particular, the eigenvalues of $f_{\mathbb{C}}$ are of the form $\cos\theta_j \pm i\sin\theta_j$, 1, or -1.

Proof. The case where $n = 1$ is trivial. It is immediately verified that $f \circ f^* = f^* \circ f = \mathrm{id}$ implies that $f_{\mathbb{C}} \circ f_{\mathbb{C}}^* = f_{\mathbb{C}}^* \circ f_{\mathbb{C}} = \mathrm{id}$, so the map $f_{\mathbb{C}}$ is unitary. By Proposition 16.7, the eigenvalues of $f_{\mathbb{C}}$ have absolute value 1. As a consequence, the eigenvalues of $f_{\mathbb{C}}$ are of the form $\cos\theta \pm i\sin\theta$, 1, or -1. The theorem then follows immediately from Theorem 16.2, where the condition $\mu > 0$ implies that $\sin\theta_j > 0$, and thus, $0 < \theta_j < \pi$. □

It is obvious that we can reorder the orthonormal basis of eigenvectors given by Theorem 16.6, so that the matrix of f w.r.t. this basis is a block diagonal matrix of the form

$$\begin{pmatrix} A_1 & \cdots & & & \\ \vdots & \ddots & \vdots & & \vdots \\ & \cdots & A_r & & \\ & & & -I_q & \\ \cdots & & & & I_p \end{pmatrix}$$

where each block A_j is a two-dimensional rotation matrix $A_j \neq \pm I_2$ of the form

$$A_j = \begin{pmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{pmatrix}$$

with $0 < \theta_j < \pi$.

The linear map f has an eigenspace $E(1, f) = \mathrm{Ker}\,(f - \mathrm{id})$ of dimension p for the eigenvalue 1, and an eigenspace $E(-1, f) = \mathrm{Ker}\,(f + \mathrm{id})$ of dimension q for the eigenvalue -1. If $\det(f) = +1$ (f is a rotation), the dimension q of $E(-1, f)$ must be even, and the entries in $-I_q$ can be paired to form two-dimensional blocks, if we wish. In this case, every rotation in $\mathbf{SO}(n)$ has a matrix of the form

$$\begin{pmatrix} A_1 & \cdots & & \\ \vdots & \ddots & \vdots & \\ & \cdots & A_m & \\ \cdots & & & I_{n-2m} \end{pmatrix}$$

where the first m blocks A_j are of the form

$$A_j = \begin{pmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{pmatrix}$$

with $0 < \theta_j \leq \pi$.

Theorem 16.6 can be used to prove a version of the Cartan–Dieudonné theorem.

Theorem 16.7. *Let E be a Euclidean space of dimension $n \geq 2$. For every isometry $f \in \mathbf{O}(E)$, if $p = \dim(E(1, f)) = \dim(\mathrm{Ker}\,(f - \mathrm{id}))$, then f is the composition of $n - p$ reflections, and $n - p$ is minimal.*

Proof. From Theorem 16.6 there are r subspaces $F_1, \ldots, F_r$, each of dimension 2, such that

$$E = E(1, f) \oplus E(-1, f) \oplus F_1 \oplus \cdots \oplus F_r,$$

and all the summands are pairwise orthogonal. Furthermore, the restriction r_i of f to each F_i is a rotation $r_i \neq \pm\mathrm{id}$. Each 2D rotation r_i can be written as the composition $r_i = s'_i \circ s_i$ of two reflections s_i and s'_i about lines in F_i (forming an angle $\theta_i/2$). We can extend s_i and s'_i to hyperplane reflections in E by making them the identity on $F_i^{\perp}$. Then

$$s'_r \circ s_r \circ \cdots \circ s'_1 \circ s_1$$

agrees with f on $F_1 \oplus \cdots \oplus F_r$ and is the identity on $E(1, f) \oplus E(-1, f)$. If $E(-1, f)$ has an orthonormal basis of eigenvectors $(v_1, \ldots, v_q)$, letting s''_j be the reflection about the hyperplane $(v_j)^{\perp}$, it is clear that

$$s''_q \circ \cdots \circ s''_1$$

agrees with f on $E(-1, f)$ and is the identity on $E(1, f) \oplus F_1 \oplus \cdots \oplus F_r$. But then

$$f = s''_q \circ \cdots \circ s''_1 \circ s'_r \circ s_r \circ \cdots \circ s'_1 \circ s_1,$$

the composition of $2r + q = n - p$ reflections.

If

$$f = s_t \circ \cdots \circ s_1,$$

for t reflections s_i, it is clear that

$$F = \bigcap_{i=1}^{t} E(1, s_i) \subseteq E(1, f),$$

where $E(1, s_i)$ is the hyperplane defining the reflection s_i. By the Grassmann relation, if we intersect $t \leq n$ hyperplanes, the dimension of their intersection is at least $n - t$. Thus, $n - t \leq p$, that is, $t \geq n - p$, and $n - p$ is the smallest number of reflections composing f. □

As a corollary of Theorem 16.7, we obtain the following fact: If the dimension n of the Euclidean space E is odd, then every rotation $f \in \mathbf{SO}(E)$ admits 1 as an eigenvalue.

Proof. The characteristic polynomial $\det(XI - f)$ of f has odd degree n and has real coefficients, so it must have some real root λ. Since f is an isometry, its n eigenvalues are of the form, $+1, -1$, and $e^{\pm i\theta}$, with $0 < \theta < \pi$, so $\lambda = \pm 1$. Now the eigenvalues $e^{\pm i\theta}$ appear in conjugate pairs, and since n is odd, the number of real eigenvalues of f is odd. This implies that $+1$ is an eigenvalue of f, since otherwise -1 would be the only real eigenvalue of f, and since its multiplicity is odd, we would have $\det(f) = -1$, contradicting the fact that f is a rotation. □

When $n = 3$, we obtain the result due to Euler which says that every 3D rotation R has an invariant axis D, and that restricted to the plane orthogonal to D, it is a 2D rotation. Furthermore, if (a, b, c) is a unit vector defining the axis D of the rotation R and if the angle of the rotation is θ, if B is the skew-symmetric matrix

$$B = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix},$$

then the Rodigues formula (Proposition 11.13) states that

$$R = I + \sin\theta B + (1 - \cos\theta)B^2.$$

The theorems of this section and of the previous section can be immediately translated in terms of matrices. The matrix versions of these theorems is often used in applications so we briefly present them in the section.

16.5 Normal and Other Special Matrices

First we consider real matrices. Recall the following definitions.

Definition 16.3. Given a real $m \times n$ matrix A, the *transpose* $A^\top$ *of* A is the $n \times m$ matrix $A^\top = (a^\top_{i\,j})$ defined such that

$$a^\top_{i\,j} = a_{j\,i}$$

for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$. A real $n \times n$ matrix A is

- *normal* if
$$A\,A^\top = A^\top A,$$
- *symmetric* if
$$A^\top = A,$$
- *skew-symmetric* if
$$A^\top = -A,$$
- *orthogonal* if
$$A\,A^\top = A^\top A = I_n.$$

Recall from Proposition 11.12 that when E is a Euclidean space and $(e_1, \ldots, e_n)$ is an orthonormal basis for E, if A is the matrix of a linear map $f\colon E \to E$ w.r.t. the basis $(e_1, \ldots, e_n)$, then $A^\top$ is the matrix of the adjoint f^* of f. Consequently, a normal linear map has a normal matrix, a self-adjoint linear map has a symmetric matrix, a skew-self-adjoint linear map has a skew-symmetric matrix, and an orthogonal linear map has an orthogonal matrix.

Furthermore, if $(u_1, \ldots, u_n)$ is another orthonormal basis for E and P is the change of basis matrix whose columns are the components of the u_i w.r.t. the basis $(e_1, \ldots, e_n)$, then P is orthogonal, and for any linear map $f\colon E \to E$, if A is the matrix of f w.r.t $(e_1, \ldots, e_n)$ and B is the matrix of f w.r.t. $(u_1, \ldots, u_n)$, then

$$B = P^\top A P.$$

As a consequence, Theorems 16.2 and 16.4–16.6 can be restated as follows.

Theorem 16.8. *For every normal matrix A there is an orthogonal matrix P and a block diagonal matrix D such that $A = PD\,P^\top$, where D is of the form*

$$D = \begin{pmatrix} D_1 & & \cdots & \\ & D_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & D_p \end{pmatrix}$$

such that each block D_j is either a one-dimensional matrix (i.e., a real scalar) or a two-dimensional matrix of the form

$$D_j = \begin{pmatrix} \lambda_j & -\mu_j \\ \mu_j & \lambda_j \end{pmatrix},$$

where $\lambda_j, \mu_j \in \mathbb{R}$, with $\mu_j > 0$.

Theorem 16.9. *For every symmetric matrix A there is an orthogonal matrix P and a diagonal matrix D such that $A = PD\,P^\top$, where D is of the form*

$$D = \begin{pmatrix} \lambda_1 & & \cdots & \\ & \lambda_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \lambda_n \end{pmatrix},$$

where $\lambda_i \in \mathbb{R}$.

Theorem 16.10. *For every skew-symmetric matrix A there is an orthogonal matrix P and a block diagonal matrix D such that $A = PDP^\top$, where D is of the form*

$$D = \begin{pmatrix} D_1 & & \cdots & \\ & D_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & D_p \end{pmatrix}$$

such that each block D_j is either 0 or a two-dimensional matrix of the form

$$D_j = \begin{pmatrix} 0 & -\mu_j \\ \mu_j & 0 \end{pmatrix},$$

where $\mu_j \in \mathbb{R}$, with $\mu_j > 0$. In particular, the eigenvalues of A are pure imaginary of the form $\pm i\mu_j$, or 0.

Theorem 16.11. *For every orthogonal matrix A there is an orthogonal matrix P and a block diagonal matrix D such that $A = PDP^\top$, where D is of the form*

$$D = \begin{pmatrix} D_1 & & \cdots & \\ & D_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & D_p \end{pmatrix}$$

such that each block D_j is either 1, -1, or a two-dimensional matrix of the form

$$D_j = \begin{pmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{pmatrix}$$

where $0 < \theta_j < \pi$. In particular, the eigenvalues of A are of the form $\cos\theta_j \pm i \sin\theta_j$, 1, or -1.

Theorem 16.11 can be used to show that the exponential map $\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$ is surjective; see Gallier [Gallier (2011b)].

We now consider complex matrices.

Definition 16.4. Given a complex $m \times n$ matrix A, the *transpose* $A^\top$ *of* A is the $n \times m$ matrix $A^\top = \left(a_{ij}^\top\right)$ defined such that

$$a_{ij}^\top = a_{ji}$$

for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$. The *conjugate* $\overline{A}$ *of* A is the $m \times n$ matrix $\overline{A} = (b_{ij})$ defined such that

$$b_{ij} = \overline{a}_{ij}$$

for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$. Given an $m \times n$ complex matrix A, the *adjoint* A^* *of* A is the matrix defined such that

$$A^* = \overline{(A^\top)} = (\overline{A})^\top.$$

A complex $n \times n$ matrix A is

- *normal* if
$$AA^* = A^*A,$$
- *Hermitian* if
$$A^* = A,$$
- *skew-Hermitian* if
$$A^* = -A,$$
- *unitary* if
$$AA^* = A^*A = I_n.$$

Recall from Proposition 13.14 that when E is a Hermitian space and $(e_1, \ldots, e_n)$ is an orthonormal basis for E, if A is the matrix of a linear map $f\colon E \to E$ w.r.t. the basis $(e_1, \ldots, e_n)$, then A^* is the matrix of the adjoint f^* of f. Consequently, a normal linear map has a normal matrix, a self-adjoint linear map has a Hermitian matrix, a skew-self-adjoint linear map has a skew-Hermitian matrix, and a unitary linear map has a unitary matrix.

Furthermore, if $(u_1, \ldots, u_n)$ is another orthonormal basis for E and P is the change of basis matrix whose columns are the components of the u_i w.r.t. the basis $(e_1, \ldots, e_n)$, then P is unitary, and for any linear map $f\colon E \to E$, if A is the matrix of f w.r.t $(e_1, \ldots, e_n)$ and B is the matrix of f w.r.t. $(u_1, \ldots, u_n)$, then

$$B = P^*AP.$$

Theorem 16.3 and Proposition 16.7 can be restated in terms of matrices as follows.

Theorem 16.12. *For every complex normal matrix A there is a unitary matrix U and a diagonal matrix D such that $A = UDU^*$. Furthermore, if A is Hermitian, then D is a real matrix; if A is skew-Hermitian, then the entries in D are pure imaginary or zero; and if A is unitary, then the entries in D have absolute value* 1.

16.6 Rayleigh–Ritz Theorems and Eigenvalue Interlacing

A fact that is used frequently in optimization problems is that the eigenvalues of a symmetric matrix are characterized in terms of what is known as the *Rayleigh ratio*, defined by

$$R(A)(x) = \frac{x^\top Ax}{x^\top x}, \quad x \in \mathbb{R}^n, x \neq 0.$$

The following proposition is often used to prove the correctness of various optimization or approximation problems (for example PCA; see Section 21.4). It is also used to prove Proposition 16.13, which is used to justify the correctness of a method for graph-drawing (see Chapter 19).

Proposition 16.11. *(Rayleigh–Ritz) If A is a symmetric $n \times n$ matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and if $(u_1, \ldots, u_n)$ is any orthonormal basis of eigenvectors of A, where u_i is a unit eigenvector associated with λ_i, then*

$$\max_{x \neq 0} \frac{x^\top Ax}{x^\top x} = \lambda_n$$

(with the maximum attained for $x = u_n$), and

$$\max_{x \neq 0, x \in \{u_{n-k+1}, \ldots, u_n\}^\perp} \frac{x^\top Ax}{x^\top x} = \lambda_{n-k}$$

(with the maximum attained for $x = u_{n-k}$), where $1 \leq k \leq n-1$. Equivalently, if V_k is the subspace spanned by $(u_1, \ldots, u_k)$, then

$$\lambda_k = \max_{x \neq 0, x \in V_k} \frac{x^\top Ax}{x^\top x}, \quad k = 1, \ldots, n.$$

Proof. First observe that

$$\max_{x \neq 0} \frac{x^\top Ax}{x^\top x} = \max_x \{x^\top Ax \mid x^\top x = 1\},$$

and similarly,

$$\max_{x \neq 0, x \in \{u_{n-k+1}, \ldots, u_n\}^\perp} \frac{x^\top Ax}{x^\top x}$$
$$= \max_x \left\{x^\top Ax \mid (x \in \{u_{n-k+1}, \ldots, u_n\}^\perp) \wedge (x^\top x = 1)\right\}.$$

Since A is a symmetric matrix, its eigenvalues are real and it can be diagonalized with respect to an orthonormal basis of eigenvectors, so let $(u_1, \ldots, u_n)$ be such a basis. If we write

$$x = \sum_{i=1}^n x_i u_i,$$

a simple computation shows that

$$x^\top Ax = \sum_{i=1}^{n} \lambda_i x_i^2.$$

If $x^\top x = 1$, then $\sum_{i=1}^{n} x_i^2 = 1$, and since we assumed that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, we get

$$x^\top Ax = \sum_{i=1}^{n} \lambda_i x_i^2 \leq \lambda_n \left(\sum_{i=1}^{n} x_i^2 \right) = \lambda_n.$$

Thus,

$$\max_x \left\{ x^\top Ax \mid x^\top x = 1 \right\} \leq \lambda_n,$$

and since this maximum is achieved for $e_n = (0, 0, \ldots, 1)$, we conclude that

$$\max_x \left\{ x^\top Ax \mid x^\top x = 1 \right\} = \lambda_n.$$

Next observe that $x \in \{u_{n-k+1}, \ldots, u_n\}^\perp$ and $x^\top x = 1$ iff $x_{n-k+1} = \cdots = x_n = 0$ and $\sum_{i=1}^{n-k} x_i^2 = 1$. Consequently, for such an x, we have

$$x^\top Ax = \sum_{i=1}^{n-k} \lambda_i x_i^2 \leq \lambda_{n-k} \left(\sum_{i=1}^{n-k} x_i^2 \right) = \lambda_{n-k}.$$

Thus,

$$\max_x \left\{ x^\top Ax \mid (x \in \{u_{n-k+1}, \ldots, u_n\}^\perp) \wedge (x^\top x = 1) \right\} \leq \lambda_{n-k},$$

and since this maximum is achieved for $e_{n-k} = (0, \ldots, 0, 1, 0, \ldots, 0)$ with a 1 in position $n - k$, we conclude that

$$\max_x \left\{ x^\top Ax \mid (x \in \{u_{n-k+1}, \ldots, u_n\}^\perp) \wedge (x^\top x = 1) \right\} = \lambda_{n-k},$$

as claimed. □

For our purposes we need the version of Proposition 16.11 applying to min instead of max, whose proof is obtained by a trivial modification of the proof of Proposition 16.11.

Proposition 16.12. *(Rayleigh–Ritz) If A is a symmetric $n \times n$ matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and if $(u_1, \ldots, u_n)$ is any orthonormal basis of eigenvectors of A, where u_i is a unit eigenvector associated with λ_i, then*

$$\min_{x \neq 0} \frac{x^\top Ax}{x^\top x} = \lambda_1$$

(with the minimum attained for $x = u_1$), and

$$\min_{x \neq 0, x \in \{u_1, \ldots, u_{i-1}\}^\perp} \frac{x^\top A x}{x^\top x} = \lambda_i$$

(with the minimum attained for $x = u_i$), where $2 \leq i \leq n$. Equivalently, if $W_k = V_{k-1}^\perp$ denotes the subspace spanned by $(u_k, \ldots, u_n)$ (with $V_0 = (0)$), then

$$\lambda_k = \min_{x \neq 0, x \in W_k} \frac{x^\top A x}{x^\top x} = \min_{x \neq 0, x \in V_{k-1}^\perp} \frac{x^\top A x}{x^\top x}, \quad k = 1, \ldots, n.$$

Propositions 16.11 and 16.12 together are known the *Rayleigh–Ritz theorem*.

As an application of Propositions 16.11 and 16.12, we prove a proposition which allows us to compare the eigenvalues of two symmetric matrices A and $B = R^\top A R$, where R is a rectangular matrix satisfying the equation $R^\top R = I$.

First we need a definition.

Definition 16.5. Given an $n \times n$ symmetric matrix A and an $m \times m$ symmetric B, with $m \leq n$, if $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of A and $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_m$ are the eigenvalues of B, then we say that the eigenvalues of B *interlace* the eigenvalues of A if

$$\lambda_i \leq \mu_i \leq \lambda_{n-m+i}, \quad i = 1, \ldots, m.$$

For example, if $n = 5$ and $m = 3$, we have

$$\begin{aligned} \lambda_1 &\leq \mu_1 \leq \lambda_3 \\ \lambda_2 &\leq \mu_2 \leq \lambda_4 \\ \lambda_3 &\leq \mu_3 \leq \lambda_5. \end{aligned}$$

Proposition 16.13. *Let A be an $n \times n$ symmetric matrix, R be an $n \times m$ matrix such that $R^\top R = I$ (with $m \leq n$), and let $B = R^\top A R$ (an $m \times m$ matrix). The following properties hold:*

(a) The eigenvalues of B interlace the eigenvalues of A.
(b) If $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of A and $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_m$ are the eigenvalues of B, and if $\lambda_i = \mu_i$, then there is an eigenvector v of B with eigenvalue μ_i such that Rv is an eigenvector of A with eigenvalue λ_i.

Proof. (a) Let $(u_1, \ldots, u_n)$ be an orthonormal basis of eigenvectors for A, and let $(v_1, \ldots, v_m)$ be an orthonormal basis of eigenvectors for B. Let U_j be the subspace spanned by $(u_1, \ldots, u_j)$ and let V_j be the subspace spanned by $(v_1, \ldots, v_j)$. For any i, the subspace V_i has dimension i and the subspace $R^\top U_{i-1}$ has dimension at most $i-1$. Therefore, there is some nonzero vector $v \in V_i \cap (R^\top U_{i-1})^\perp$, and since

$$v^\top R^\top u_j = (Rv)^\top u_j = 0, \quad j = 1, \ldots, i-1,$$

we have $Rv \in (U_{i-1})^\perp$. By Proposition 16.12 and using the fact that $R^\top R = I$, we have

$$\lambda_i \leq \frac{(Rv)^\top ARv}{(Rv)^\top Rv} = \frac{v^\top Bv}{v^\top v}.$$

On the other hand, by Proposition 16.11,

$$\mu_i = \max_{x \neq 0, x \in \{v_{i+1}, \ldots, v_n\}^\perp} \frac{x^\top Bx}{x^\top x} = \max_{x \neq 0, x \in \{v_1, \ldots, v_i\}} \frac{x^\top Bx}{x^\top x},$$

so

$$\frac{w^\top Bw}{w^\top w} \leq \mu_i \quad \text{for all } w \in V_i,$$

and since $v \in V_i$, we have

$$\lambda_i \leq \frac{v^\top Bv}{v^\top v} \leq \mu_i, \quad i = 1, \ldots, m.$$

We can apply the same argument to the symmetric matrices $-A$ and $-B$, to conclude that

$$-\lambda_{n-m+i} \leq -\mu_i,$$

that is,

$$\mu_i \leq \lambda_{n-m+i}, \quad i = 1, \ldots, m.$$

Therefore,

$$\lambda_i \leq \mu_i \leq \lambda_{n-m+i}, \quad i = 1, \ldots, m,$$

as desired.

(b) If $\lambda_i = \mu_i$, then

$$\lambda_i = \frac{(Rv)^\top ARv}{(Rv)^\top Rv} = \frac{v^\top Bv}{v^\top v} = \mu_i,$$

so v must be an eigenvector for B and Rv must be an eigenvector for A, both for the eigenvalue $\lambda_i = \mu_i$. □

Proposition 16.13 immediately implies the *Poincaré separation theorem*. It can be used in situations, such as in quantum mechanics, where one has information about the inner products $u_i^\top A u_j$.

Proposition 16.14. *(Poincaré separation theorem) Let A be a $n \times n$ symmetric (or Hermitian) matrix, let r be some integer with $1 \leq r \leq n$, and let $(u_1, \ldots, u_r)$ be r orthonormal vectors. Let $B = (u_i^\top A u_j)$ (an $r \times r$ matrix), let $\lambda_1(A) \leq \ldots \leq \lambda_n(A)$ be the eigenvalues of A and $\lambda_1(B) \leq \ldots \leq \lambda_r(B)$ be the eigenvalues of B; then we have*

$$\lambda_k(A) \leq \lambda_k(B) \leq \lambda_{k+n-r}(A), \quad k = 1, \ldots, r.$$

Observe that Proposition 16.13 implies that

$$\lambda_1 + \cdots + \lambda_m \leq \mathrm{tr}(R^\top A R) \leq \lambda_{n-m+1} + \cdots + \lambda_n.$$

If P_1 is the the $n \times (n-1)$ matrix obtained from the identity matrix by dropping its last column, we have $P_1^\top P_1 = I$, and the matrix $B = P_1^\top A P_1$ is the matrix obtained from A by deleting its last row and its last column. In this case the interlacing result is

$$\lambda_1 \leq \mu_1 \leq \lambda_2 \leq \mu_2 \leq \cdots \leq \mu_{n-2} \leq \lambda_{n-1} \leq \mu_{n-1} \leq \lambda_n,$$

a genuine interlacing. We obtain similar results with the matrix P_{n-r} obtained by dropping the last $n-r$ columns of the identity matrix and setting $B = P_{n-r}^\top A P_{n-r}$ (B is the $r \times r$ matrix obtained from A by deleting its last $n-r$ rows and columns). In this case we have the following interlacing inequalities known as *Cauchy interlacing theorem*:

$$\lambda_k \leq \mu_k \leq \lambda_{k+n-r}, \quad k = 1, \ldots, r. \tag{$*$}$$

16.7 The Courant–Fischer Theorem; Perturbation Results

Another useful tool to prove eigenvalue equalities is the Courant–Fischer characterization of the eigenvalues of a symmetric matrix, also known as the Min-max (and Max-min) theorem.

Theorem 16.13. *(Courant–Fischer) Let A be a symmetric $n \times n$ matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. If $\mathcal{V}_k$ denotes the set of subspaces of $\mathbb{R}^n$ of dimension k, then*

$$\lambda_k = \max_{W \in \mathcal{V}_{n-k+1}} \min_{x \in W, x \neq 0} \frac{x^\top A x}{x^\top x}$$

$$\lambda_k = \min_{W \in \mathcal{V}_k} \max_{x \in W, x \neq 0} \frac{x^\top A x}{x^\top x}.$$

Proof. Let us consider the second equality, the proof of the first equality being similar. Let $(u_1, \ldots, u_n)$ be any orthonormal basis of eigenvectors of A, where u_i is a unit eigenvector associated with λ_i. Observe that the space V_k spanned by $(u_1, \ldots, u_k)$ has dimension k, and by Proposition 16.11, we have

$$\lambda_k = \max_{x \neq 0, x \in V_k} \frac{x^\top A x}{x^\top x} \geq \min_{W \in \mathcal{V}_k} \max_{x \in W, x \neq 0} \frac{x^\top A x}{x^\top x}.$$

Therefore, we need to prove the reverse inequality; that is, we have to show that

$$\lambda_k \leq \max_{x \neq 0, x \in W} \frac{x^\top A x}{x^\top x}, \quad \text{for all} \quad W \in \mathcal{V}_k.$$

Now for any $W \in \mathcal{V}_k$, if we can prove that $W \cap V_{k-1}^\perp \neq (0)$, then for any nonzero $v \in W \cap V_{k-1}^\perp$, by Proposition 16.12 , we have

$$\lambda_k = \min_{x \neq 0, x \in V_{k-1}^\perp} \frac{x^\top A x}{x^\top x} \leq \frac{v^\top A v}{v^\top v} \leq \max_{x \in W, x \neq 0} \frac{x^\top A x}{x^\top x}.$$

It remains to prove that $\dim(W \cap V_{k-1}^\perp) \geq 1$. However, $\dim(V_{k-1}) = k-1$, so $\dim(V_{k-1}^\perp) = n-k+1$, and by hypothesis $\dim(W) = k$. By the Grassmann relation,

$$\dim(W) + \dim(V_{k-1}^\perp) = \dim(W \cap V_{k-1}^\perp) + \dim(W + V_{k-1}^\perp),$$

and since $\dim(W + V_{k-1}^\perp) \leq \dim(\mathbb{R}^n) = n$, we get

$$k + n - k + 1 \leq \dim(W \cap V_{k-1}^\perp) + n;$$

that is, $1 \leq \dim(W \cap V_{k-1}^\perp)$, as claimed. □

The Courant–Fischer theorem yields the following useful result about perturbing the eigenvalues of a symmetric matrix due to Hermann Weyl.

Proposition 16.15. *Given two $n \times n$ symmetric matrices A and $B = A + \Delta A$, if $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$ are the eigenvalues of A and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$ are the eigenvalues of B, then*

$$|\alpha_k - \beta_k| \leq \rho(\Delta A) \leq \|\Delta A\|_2, \quad k = 1, \ldots, n.$$

Proof. Let $\mathcal{V}_k$ be defined as in the Courant–Fischer theorem and let V_k be the subspace spanned by the k eigenvectors associated with $\lambda_1, \ldots, \lambda_k$. By

the Courant–Fischer theorem applied to B, we have

$$\begin{aligned}\beta_k &= \min_{W\in\mathcal{V}_k}\max_{x\in W, x\neq 0}\frac{x^\top Bx}{x^\top x}\\ &\leq \max_{x\in V_k}\frac{x^\top Bx}{x^\top x}\\ &= \max_{x\in V_k}\left(\frac{x^\top Ax}{x^\top x}+\frac{x^\top \Delta\, Ax}{x^\top x}\right)\\ &\leq \max_{x\in V_k}\frac{x^\top Ax}{x^\top x}+\max_{x\in V_k}\frac{x^\top \Delta A\, x}{x^\top x}.\end{aligned}$$

By Proposition 16.11, we have

$$\alpha_k = \max_{x\in V_k}\frac{x^\top Ax}{x^\top x},$$

so we obtain

$$\begin{aligned}\beta_k &\leq \max_{x\in V_k}\frac{x^\top Ax}{x^\top x}+\max_{x\in V_k}\frac{x^\top \Delta A\, x}{x^\top x}\\ &= \alpha_k+\max_{x\in V_k}\frac{x^\top \Delta A\, x}{x^\top x}\\ &\leq \alpha_k+\max_{x\in \mathbb{R}^n}\frac{x^\top \Delta A\, x}{x^\top x}.\end{aligned}$$

Now by Proposition 16.11 and Proposition 8.6, we have

$$\max_{x\in\mathbb{R}^n}\frac{x^\top \Delta A\, x}{x^\top x} = \max_i \lambda_i(\Delta A)\leq \rho(\Delta A)\leq \|\Delta A\|_2,$$

where $\lambda_i(\Delta A)$ denotes the ith eigenvalue of ΔA, which implies that

$$\beta_k\leq \alpha_k+\rho(\Delta A)\leq \alpha_k+\|\Delta A\|_2.$$

By exchanging the roles of A and B, we also have

$$\alpha_k\leq \beta_k+\rho(\Delta A)\leq \beta_k+\|\Delta A\|_2,$$

and thus,

$$|\alpha_k-\beta_k|\leq \rho(\Delta A)\leq \|\Delta A\|_2,\quad k=1,\ldots,n,$$

as claimed. □

Proposition 16.15 also holds for Hermitian matrices.

A pretty result of Wielandt and Hoffman asserts that

$$\sum_{k=1}^n(\alpha_k-\beta_k)^2\leq \|\Delta A\|_F^2,$$

where $\| \ \|_F$ is the Frobenius norm. However, the proof is significantly harder than the above proof; see Lax [Lax (2007)].

The Courant–Fischer theorem can also be used to prove some famous inequalities due to Hermann Weyl. These can also be viewed as perturbation results. Given two symmetric (or Hermitian) matrices A and B, let $\lambda_i(A), \lambda_i(B)$, and $\lambda_i(A+B)$ denote the ith eigenvalue of A, B, and $A+B$, respectively, arranged in nondecreasing order.

Proposition 16.16. *(Weyl) Given two symmetric (or Hermitian) $n \times n$ matrices A and B, the following inequalities hold: For all i, j, k with $1 \leq i, j, k \leq n$:*

(1) If $i + j = k + 1$, then

$$\lambda_i(A) + \lambda_j(B) \leq \lambda_k(A + B).$$

(2) If $i + j = k + n$, then

$$\lambda_k(A + B) \leq \lambda_i(A) + \lambda_j(B).$$

Proof. Observe that the first set of inequalities is obtained from the second set by replacing A by $-A$ and B by $-B$, so it is enough to prove the second set of inequalities. By the Courant–Fischer theorem, there is a subspace H of dimension $n - k + 1$ such that

$$\lambda_k(A + B) = \min_{x \in H, x \neq 0} \frac{x^\top (A + B)x}{x^\top x}.$$

Similarly, there exists a subspace F of dimension i and a subspace G of dimension j such that

$$\lambda_i(A) = \max_{x \in F, x \neq 0} \frac{x^\top Ax}{x^\top x}, \quad \lambda_j(B) = \max_{x \in G, x \neq 0} \frac{x^\top Bx}{x^\top x}.$$

We claim that $F \cap G \cap H \neq (0)$. To prove this, we use the Grassmann relation twice. First,

$$\begin{aligned} \dim(F \cap G \cap H) &= \dim(F) + \dim(G \cap H) - \dim(F + (G \cap H)) \\ &\geq \dim(F) + \dim(G \cap H) - n, \end{aligned}$$

and second,

$$\dim(G \cap H) = \dim(G) + \dim(H) - \dim(G + H) \geq \dim(G) + \dim(H) - n,$$

so

$$\dim(F \cap G \cap H) \geq \dim(F) + \dim(G) + \dim(H) - 2n.$$

However,

$$\dim(F) + \dim(G) + \dim(H) = i + j + n - k + 1$$

and $i + j = k + n$, so we have

$$\dim(F \cap G \cap H) \geq i + j + n - k + 1 - 2n = k + n + n - k + 1 - 2n = 1,$$

which shows that $F \cap G \cap H \neq (0)$. Then for any unit vector $z \in F \cap G \cap H \neq (0)$, we have

$$\lambda_k(A + B) \leq z^\top (A + B)z, \quad \lambda_i(A) \geq z^\top Az, \quad \lambda_j(B) \geq z^\top Bz,$$

establishing the desired inequality $\lambda_k(A + B) \leq \lambda_i(A) + \lambda_j(B)$. □

In the special case $i = j = k$, we obtain

$$\lambda_1(A) + \lambda_1(B) \leq \lambda_1(A + B), \quad \lambda_n(A + B) \leq \lambda_n(A) + \lambda_n(B).$$

It follows that λ_1 (as a function) is concave, while λ_n (as a function) is convex.

If $i = 1$ and $j = k$, we obtain

$$\lambda_1(A) + \lambda_k(B) \leq \lambda_k(A + B),$$

and if $i = k$ and $j = n$, we obtain

$$\lambda_k(A + B) \leq \lambda_k(A) + \lambda_n(B),$$

and combining them, we get

$$\lambda_1(A) + \lambda_k(B) \leq \lambda_k(A + B) \leq \lambda_k(A) + \lambda_n(B).$$

In particular, if B is positive semidefinite, since its eigenvalues are nonnegative, we obtain the following inequality known as the *monotonicity theorem* for symmetric (or Hermitian) matrices: if A and B are symmetric (or Hermitian) and B is positive semidefinite, then

$$\lambda_k(A) \leq \lambda_k(A + B) \quad k = 1, \ldots, n.$$

The reader is referred to Horn and Johnson [Horn and Johnson (1990)] (Chapters 4 and 7) for a very complete treatment of matrix inequalities and interlacing results, and also to Lax [Lax (2007)] and Serre [Serre (2010)].

16.8 Summary

The main concepts and results of this chapter are listed below:

- *Normal* linear maps, *self-adjoint* linear maps, *skew-self-adjoint* linear maps, and *orthogonal* linear maps.
- Properties of the eigenvalues and eigenvectors of a normal linear map.
- The *complexification* of a real vector space, of a linear map, and of a Euclidean inner product.
- The eigenvalues of a self-adjoint map in a Hermitian space are *real*.
- The eigenvalues of a self-adjoint map in a Euclidean space are *real*.
- Every self-adjoint linear map on a Euclidean space has an orthonormal basis of eigenvectors.
- Every normal linear map on a Euclidean space can be block diagonalized (blocks of size at most 2×2) with respect to an orthonormal basis of eigenvectors.
- Every normal linear map on a Hermitian space can be diagonalized with respect to an orthonormal basis of eigenvectors.
- The spectral theorems for self-adjoint, skew-self-adjoint, and orthogonal linear maps (on a Euclidean space).
- The spectral theorems for normal, symmetric, skew-symmetric, and orthogonal (real) matrices.
- The spectral theorems for normal, Hermitian, skew-Hermitian, and unitary (complex) matrices.
- The *Rayleigh ratio* and the *Rayleigh–Ritz theorem*.
- *Interlacing inequalities* and the *Cauchy interlacing theorem*.
- The *Poincaré separation theorem*.
- The *Courant–Fischer theorem*.
- Inequalities involving perturbations of the eigenvalues of a symmetric matrix.
- The *Weyl inequalities*.

16.9 Problems

Problem 16.1. Prove that the structure $E_{\mathbb{C}}$ introduced in Definition 16.2 is indeed a complex vector space.

Problem 16.2. Prove that the formula

$$\langle u_1 + iv_1, u_2 + iv_2 \rangle_{\mathbb{C}} = \langle u_1, u_2 \rangle + \langle v_1, v_2 \rangle + i(\langle v_1, u_2 \rangle - \langle u_1, v_2 \rangle)$$

defines a Hermitian form on $E_{\mathbb{C}}$ that is positive definite and that $\langle -, - \rangle_{\mathbb{C}}$ agrees with $\langle -, - \rangle$ on real vectors.

Problem 16.3. Given any linear map $f\colon E \to E$, prove the map $f^*_{\mathbb{C}}$ defined such that

$$f^*_{\mathbb{C}}(u + iv) = f^*(u) + if^*(v)$$

for all $u, v \in E$ is the adjoint of $f_{\mathbb{C}}$ w.r.t. $\langle -, - \rangle_{\mathbb{C}}$.

Problem 16.4. Let A be a real symmetric $n \times n$ matrix whose eigenvalues are nonnegative. Prove that for every $p > 0$, there is a real symmetric matrix S whose eigenvalues are nonnegative such that $S^p = A$.

Problem 16.5. Let A be a real symmetric $n \times n$ matrix whose eigenvalues are positive.

(1) Prove that there is a real symmetric matrix S such that $A = e^S$.

(2) Let S be a real symmetric $n \times n$ matrix. Prove that $A = e^S$ is a real symmetric $n \times n$ matrix whose eigenvalues are positive.

Problem 16.6. Let A be a complex matrix. Prove that if A can be diagonalized with respect to an orthonormal basis, then A is normal.

Problem 16.7. Let $f\colon \mathbb{C}^n \to \mathbb{C}^n$ be a linear map.

(1) Prove that if f is diagonalizable and if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of f, then $\lambda_1^2, \ldots, \lambda_n^2$ are the eigenvalues of f^2, and if $\lambda_i^2 = \lambda_j^2$ implies that $\lambda_i = \lambda_j$, then f and f^2 have the same eigenspaces.

(2) Let f and g be two real self-adjoint linear maps $f, g\colon \mathbb{R}^n \to \mathbb{R}^n$. Prove that if f and g have nonnegative eigenvalues (f and g are positve semidefinite) and if $f^2 = g^2$, then $f = g$.

Problem 16.8. (1) Let $\mathfrak{so}(3)$ be the space of 3×3 skew symmetric matrices

$$\mathfrak{so}(3) = \left\{ \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} \;\middle|\; a, b, c \in \mathbb{R} \right\}.$$

For any matrix

$$A = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} \in \mathfrak{so}(3),$$

if we let $\theta = \sqrt{a^2 + b^2 + c^2}$, recall from Section 11.7 (the Rodrigues formula) that the exponential map $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ is given by

$$e^A = I_3 + \frac{\sin\theta}{\theta} A + \frac{(1 - \cos\theta)}{\theta^2} A^2, \quad \text{if } \theta \neq 0,$$

with $\exp(0_3) = I_3$.

(2) Prove that e^A is an orthogonal matrix of determinant $+1$, i.e., a rotation matrix.

(3) Prove that the exponential map $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ is surjective. For this proceed as follows: Pick any rotation matrix $R \in \mathbf{SO}(3)$;

(1) The case $R = I$ is trivial.
(2) If $R \neq I$ and $\mathrm{tr}(R) \neq -1$, then
$$\exp^{-1}(R) = \left\{ \frac{\theta}{2\sin\theta}(R - R^T) \;\middle|\; 1 + 2\cos\theta = \mathrm{tr}(R) \right\}.$$
(Recall that $\mathrm{tr}(R) = r_{1\,1} + r_{2\,2} + r_{3\,3}$, the *trace* of the matrix R).
Show that there is a unique skew-symmetric B with corresponding θ satisfying $0 < \theta < \pi$ such that $e^B = R$.
(3) If $R \neq I$ and $\mathrm{tr}(R) = -1$, then prove that the eigenvalues of R are $1, -1, -1$, that $R = R^\top$, and that $R^2 = I$. Prove that the matrix
$$S = \frac{1}{2}(R - I)$$
is a symmetric matrix whose eigenvalues are $-1, -1, 0$. Thus S can be diagonalized with respect to an orthogonal matrix Q as
$$S = Q \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} Q^\top.$$
Prove that there exists a skew symmetric matrix
$$U = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix}$$
so that
$$U^2 = S = \frac{1}{2}(R - I).$$
Observe that
$$U^2 = \begin{pmatrix} -(c^2 + d^2) & bc & bd \\ bc & -(b^2 + d^2) & cd \\ bd & cd & -(b^2 + c^2) \end{pmatrix},$$
and use this to conclude that if $U^2 = S$, then $b^2 + c^2 + d^2 = 1$. Then show that
$$\exp^{-1}(R) = \left\{ (2k+1)\pi \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix},\ k \in \mathbb{Z} \right\},$$
where (b, c, d) is any unit vector such that for the corresponding skew symmetric matrix U, we have $U^2 = S$.

(4) To find a skew symmetric matrix U so that $U^2 = S = \frac{1}{2}(R - I)$ as in (3), we can solve the system

$$\begin{pmatrix} b^2 - 1 & bc & bd \\ bc & c^2 - 1 & cd \\ bd & cd & d^2 - 1 \end{pmatrix} = S.$$

We immediately get b^2, c^2, d^2, and then, since one of b, c, d is nonzero, say b, if we choose the positive square root of b^2, we can determine c and d from bc and bd.

Implement a computer program in `Matlab` to solve the above system.

Problem 16.9. It was shown in Proposition 14.10 that the exponential map is a map $\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$, where $\mathfrak{so}(n)$ is the vector space of real $n \times n$ skew-symmetric matrices. Use the spectral theorem to prove that the map $\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$ is surjective.

Problem 16.10. Let $\mathfrak{u}(n)$ be the space of (complex) $n \times n$ skew-Hermitian matrices ($B^* = -B$) and let $\mathfrak{su}(n)$ be its subspace consisting of skew-Hermitian matrice with zero trace ($\mathrm{tr}(B) = 0$).

(1) Prove that if $B \in \mathfrak{u}(n)$, then $e^B \in \mathbf{U}(n)$, and if if $B \in \mathfrak{su}(n)$, then $e^B \in \mathbf{SU}(n)$. Thus we have well-defined maps $\exp\colon \mathfrak{u}(n) \to \mathbf{U}(n)$ and $\exp\colon \mathfrak{su}(n) \to \mathbf{SU}(n)$.

(2) Prove that the map $\exp\colon \mathfrak{u}(n) \to \mathbf{U}(n)$ is surjective.

(3) Prove that the map $\exp\colon \mathfrak{su}(n) \to \mathbf{SU}(n)$ is surjective.

Problem 16.11. Recall that a matrix $B \in \mathrm{M}_n(\mathbb{R})$ is skew-symmetric if $B^\top = -B$. Check that the set $\mathfrak{so}(n)$ of skew-symmetric matrices is a vector space of dimension $n(n-1)/2$, and thus is isomorphic to $\mathbb{R}^{n(n-1)/2}$.

(1) Given a rotation matrix

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

where $0 < \theta < \pi$, prove that there is a skew symmetric matrix B such that

$$R = (I - B)(I + B)^{-1}.$$

(2) Prove that the eigenvalues of a skew-symmetric matrix are either 0 or pure imaginary (that is, of the form $i\mu$ for $\mu \in \mathbb{R}$.).

Let $C\colon \mathfrak{so}(n) \to \mathrm{M}_n(\mathbb{R})$ be the function (called the *Cayley transform* of B) given by

$$C(B) = (I - B)(I + B)^{-1}.$$

Prove that if B is skew-symmetric, then $I - B$ and $I + B$ are invertible, and so C is well-defined. Prove that

$$(I + B)(I - B) = (I - B)(I + B),$$

and that

$$(I + B)(I - B)^{-1} = (I - B)^{-1}(I + B).$$

Prove that

$$(C(B))^\top C(B) = I$$

and that

$$\det C(B) = +1,$$

so that $C(B)$ is a rotation matrix. Furthermore, show that $C(B)$ does not admit -1 as an eigenvalue.

(3) Let $\mathbf{SO}(n)$ be the group of $n \times n$ rotation matrices. Prove that the map

$$C \colon \mathfrak{so}(n) \to \mathbf{SO}(n)$$

is bijective onto the subset of rotation matrices that do not admit -1 as an eigenvalue. Show that the inverse of this map is given by

$$B = (I + R)^{-1}(I - R) = (I - R)(I + R)^{-1},$$

where $R \in \mathbf{SO}(n)$ does not admit -1 as an eigenvalue.

Problem 16.12. Please refer back to Problem 3.6. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A (not necessarily distinct). Using Schur's theorem, A is similar to an upper triangular matrix B, that is, $A = PBP^{-1}$ with B upper triangular, and we may assume that the diagonal entries of B in descending order are $\lambda_1, \ldots, \lambda_n$.

(1) If the E_{ij} are listed according to total order given by

$$(i, j) < (h, k) \quad \text{iff} \quad \begin{cases} i = h \text{ and } j > k \\ \text{or } i < h. \end{cases}$$

prove that R_B is an upper triangular matrix whose diagonal entries are

$$(\underbrace{\lambda_n, \ldots, \lambda_1, \ldots, \lambda_n, \ldots, \lambda_1}_{n^2}),$$

and that L_B is an upper triangular matrix whose diagonal entries are

$$(\underbrace{\lambda_1, \ldots, \lambda_1}_{n} \ldots, \underbrace{\lambda_n, \ldots, \lambda_n}_{n}).$$

Hint. Figure out what are $R_B(E_{ij}) = E_{ij}B$ and $L_B(E_{ij}) = BE_{ij}$.

(2) Use the fact that

$$L_A = L_P \circ L_B \circ L_P^{-1}, \quad R_A = R_P^{-1} \circ R_B \circ R_P,$$

to express $\mathrm{ad}_A = L_A - R_A$ in terms of $L_B - R_B$, and conclude that the eigenvalues of ad_A are $\lambda_i - \lambda_j$, for $i = 1, \ldots, n$, and for $j = n, \ldots, 1$.

Chapter 17

Computing Eigenvalues and Eigenvectors

After the problem of solving a linear system, the problem of computing the eigenvalues and the eigenvectors of a real or complex matrix is one of most important problems of numerical linear algebra. Several methods exist, among which we mention Jacobi, Givens–Householder, divide-and-conquer, QR iteration, and Rayleigh–Ritz; see Demmel [Demmel (1997)], Trefethen and Bau [Trefethen and Bau III (1997)], Meyer [Meyer (2000)], Serre [Serre (2010)], Golub and Van Loan [Golub and Van Loan (1996)], and Ciarlet [Ciarlet (1989)]. Typically, better performing methods exist for special kinds of matrices, such as symmetric matrices.

In theory, given an $n \times n$ complex matrix A, if we could compute a Schur form $A = UTU^*$, where T is upper triangular and U is unitary, we would obtain the eigenvalues of A, since they are the diagonal entries in T. However, this would require finding the roots of a polynomial, but methods for doing this are known to be numerically very unstable, so this is not a practical method.

A common paradigm is to construct a sequence (P_k) of matrices such that $A_k = P_k^{-1} A P_k$ converges, in some sense, to a matrix whose eigenvalues are easily determined. For example, $A_k = P_k^{-1} A P_k$ could become upper triangular in the limit. Furthermore, P_k is typically a product of "nice" matrices, for example, orthogonal matrices.

For general matrices, that is, matrices that are not symmetric, the QR iteration algorithm, due to Rutishauser, Francis, and Kublanovskaya in the early 1960s, is one of the most efficient algorithms for computing eigenvalues. A fascinating account of the history of the QR algorithm is given in Golub and Uhlig [Golub and Uhlig (2009)]. The QR algorithm constructs a sequence of matrices (A_k), where A_{k+1} is obtained from A_k by performing a QR-decomposition $A_k = Q_k R_k$ of A_k and then setting

$A_{k+1} = R_k Q_k$, the result of swapping Q_k and R_k. It is immediately verified that $A_{k+1} = Q_k^* A_k Q_k$, so A_k and A_{k+1} have *the same eigenvalues*, which are the eigenvalues of A.

The basic version of this algorithm runs into difficulties with matrices that have several eigenvalues with the same modulus (it may loop or not "converge" to an upper triangular matrix). There are ways of dealing with some of these problems, but for ease of exposition, we first present a simplified version of the QR algorithm which we call basic QR algorithm. We prove a convergence theorem for the basic QR algorithm, under the rather restrictive hypothesis that the input matrix A is diagonalizable and that its eigenvalues are nonzero and have distinct moduli. The proof shows that the part of A_k strictly below the diagonal converges to zero and that the diagonal entries of A_k converge to the eigenvalues of A.

Since the convergence of the QR method depends crucially only on the fact that the part of A_k below the diagonal goes to zero, it would be highly desirable if we could replace A by a similar matrix U^*AU easily computable from A and having lots of zero strictly below the diagonal. It turns out that there is a way to construct a matrix $H = U^*AU$ which is almost triangular, except that it may have an extra nonzero diagonal below the main diagonal. Such matrices called, *Hessenberg matrices*, are discussed in Section 17.2. An $n \times n$ diagonalizable Hessenberg matrix H having the property that $h_{i+1i} \neq 0$ for $i = 1, \ldots, n-1$ (such a matrix is called *unreduced*) has the nice property that its eigenvalues are all distinct. Since every Hessenberg matrix is a block diagonal matrix of unreduced Hessenberg blocks, *it suffices to compute the eigenvalues of unreduced Hessenberg matrices.* There is a special case of particular interest: symmetric (or Hermitian) positive definite tridiagonal matrices. Such matrices must have real positive distinct eigenvalues, so the QR algorithm converges to a diagonal matrix.

In Section 17.3, we consider techniques for making the basic QR method practical and more efficient. The first step is to convert the original input matrix A to a similar matrix H in Hessenberg form, and to apply the QR algorithm to H (actually, to the unreduced blocks of H). The second and crucial ingredient to speed up convergence is to add shifts.

A shift is the following step: pick some σ_k, hopefully close to some eigenvalue of A (in general, λ_n), QR-factor $A_k - \sigma_k I$ as

$$A_k - \sigma_k I = Q_k R_k,$$

and then form

$$A_{k+1} = R_k Q_k + \sigma_k I.$$

It is easy to see that we still have $A_{k+1} = Q_k^* A_k Q_k$. A judicious choice of σ_k can speed up convergence considerably. If H is real and has pairs of complex conjugate eigenvalues, we can perform a double shift, and it can be arranged that we work in real arithmetic.

The last step for improving efficiency is to compute $A_{k+1} = Q_k^* A_k Q_k$ without even performing a QR-factorization of $A_k - \sigma_k I$. This can be done when A_k is unreduced Hessenberg. Such a method is called QR iteration with implicit shifts. There is also a version of QR iteration with implicit double shifts.

If the dimension of the matrix A is very large, we can find approximations of some of the eigenvalues of A by using a truncated version of the reduction to Hessenberg form due to Arnoldi in general and to Lanczos in the symmetric (or Hermitian) tridiagonal case. *Arnoldi iteration* is discussed in Section 17.4. If A is an $m \times m$ matrix, for $n \ll m$ (n much smaller than m) the idea is to generate the $n \times n$ Hessenberg submatrix H_n of the full Hessenberg matrix H (such that $A = UHU^*$) consisting of its first n rows and n columns; the matrix U_n consisting of the first n columns of U is also produced. The Rayleigh–Ritz method consists in computing the eigenvalues of H_n using the QR- method with shifts. These eigenvalues, called *Ritz values*, are approximations of the eigenvalues of A. Typically, extreme eigenvalues are found first.

Arnoldi iteration can also be viewed as a way of computing an orthonormal basis of a *Krylov subspace*, namely the subspace $\mathcal{K}_n(A, b)$ spanned by $(b, Ab, \ldots, A^n b)$. We can also use Arnoldi iteration to find an approximate solution of a linear equation $Ax = b$ by minimizing $\|b - Ax_n\|_2$ for all x_n is the Krylov space $\mathcal{K}_n(A, b)$. This method named GMRES is discussed in Section 17.5.

The special case where H is a symmetric (or Hermitian) tridiagonal matrix is discussed in Section 17.6. In this case, Arnoldi's algorithm becomes *Lanczos' algorithm*. It is much more efficient than Arnoldi iteration.

We close this chapter by discussing two classical methods for computing a single eigenvector and a single eigenvalue: power iteration and inverse (power) iteration; see Section 17.7.

17.1 The Basic QR Algorithm

Let A be an $n \times n$ matrix which is assumed to be diagonalizable and invertible. The basic QR algorithm makes use of two very simple steps. Starting with $A_1 = A$, we construct sequences of matrices (A_k), (Q_k) (R_k) and (P_k)

as follows:

$$\begin{array}{ll} \text{Factor} & A_1 = Q_1 R_1 \\ \text{Set} & A_2 = R_1 Q_1 \\ \text{Factor} & A_2 = Q_2 R_2 \\ \text{Set} & A_3 = R_2 Q_2 \\ & \vdots \\ \text{Factor} & A_k = Q_k R_k \\ \text{Set} & A_{k+1} = R_k Q_k \\ & \vdots \end{array}$$

Thus, A_{k+1} is obtained from a QR-factorization $A_k = Q_k R_k$ of A_k by swapping Q_k and R_k. Define P_k by

$$P_k = Q_1 Q_2 \cdots Q_k.$$

Since $A_k = Q_k R_k$, we have $R_k = Q_k^* A_k$, and since $A_{k+1} = R_k Q_k$, we obtain

$$A_{k+1} = Q_k^* A_k Q_k. \tag{$*_1$}$$

An obvious induction shows that

$$A_{k+1} = Q_k^* \cdots Q_1^* A_1 Q_1 \cdots Q_k = P_k^* A P_k,$$

that is

$$A_{k+1} = P_k^* A P_k. \tag{$*_2$}$$

Therefore, A_{k+1} and A are similar, so they have the same eigenvalues.

The basic QR iteration method consists in computing the sequence of matrices A_k, and in the ideal situation, to expect that A_k "converges" to an upper triangular matrix, more precisely that the part of A_k below the main diagonal goes to zero, and the diagonal entries converge to the eigenvalues of A.

This ideal situation is only achieved in rather special cases. For one thing, if A is unitary (or orthogonal in the real case), since in the QR decomposition we have $R = I$, we get $A_2 = IQ = Q = A_1$, so the method does *not* make any progress. Also, if A is a real matrix, since the A_k are also real, if A has complex eigenvalues, then the part of A_k below the main diagonal can't go to zero. Generally, the method runs into troubles whenever A has distinct eigenvalues with the same modulus.

The convergence of the sequence (A_k) is only known under some fairly restrictive hypotheses. Even under such hypotheses, this is not really genuine convergence. Indeed, it can be shown that the part of A_k below the main diagonal goes to zero, and the diagonal entries converge to the eigenvalues of A, but the part of A_k above the diagonal *may not converge*. However, for the purpose of finding the eigenvalues of A, this does not matter.

The following convergence result is proven in Ciarlet [Ciarlet (1989)] (Chapter 6, Theorem 6.3.10 and Serre [Serre (2010)] (Chapter 13, Theorem 13.2). It is rarely applicable in practice, except for symmetric (or Hermitian) positive definite matrices, as we will see shortly.

Theorem 17.1. *Suppose the (complex) $n \times n$ matrix A is invertible, diagonalizable, and that its eigenvalues $\lambda_1, \ldots, \lambda_n$ have different moduli, so that*

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0.$$

If $A = P\Lambda P^{-1}$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, and if P^{-1} has an LU-factorization, then the strictly lower-triangular part of A_k converges to zero, and the diagonal of A_k converges to Λ.

Proof. We reproduce the proof in Ciarlet [Ciarlet (1989)] (Chapter 6, Theorem 6.3.10). The strategy is to study the asymptotic behavior of the matrices $P_k = Q_1 Q_2 \cdots Q_k$. For this, it turns out that we need to consider the powers A^k.

Step 1. Let $\mathcal{R}_k = R_k \cdots R_2 R_1$. We claim that

$$A^k = (Q_1 Q_2 \cdots Q_k)(R_k \cdots R_2 R_1) = P_k \mathcal{R}_k. \tag{$*_3$}$$

We proceed by induction. The base case $k = 1$ is trivial. For the induction step, from $(*_2)$, we have

$$P_k A_{k+1} = A P_k.$$

Since $A_{k+1} = R_k Q_k = Q_{k+1} R_{k+1}$, we have

$$P_{k+1}\mathcal{R}_{k+1} = P_k Q_{k+1} R_{k+1} \mathcal{R}_k = P_k A_{k+1} \mathcal{R}_k = A P_k \mathcal{R}_k = A A^k = A^{k+1}$$

establishing the induction step.

Step 2. We will express the matrix P_k as $P_k = Q\widetilde{Q}_k D_k$, in terms of a diagonal matrix D_k with unit entries, with Q and $\widetilde{Q}_k$ unitary.

Let $P = QR$, a QR-factorization of P (with R an upper triangular matrix with positive diagonal entries), and $P^{-1} = LU$, an LU-factorization of P^{-1}. Since $A = P\Lambda P^{-1}$, we have

$$A^k = P\Lambda^k P^{-1} = QR\Lambda^k LU = QR(\Lambda^k L \Lambda^{-k})\Lambda^k U. \tag{$*_4$}$$

Here, Λ^{-k} is the diagonal matrix with entries λ_i^{-k}. The reason for introducing the matrix $\Lambda^k L \Lambda^{-k}$ is that its asymptotic behavior is easy to determine. Indeed, we have

$$(\Lambda^k L \Lambda^{-k})_{ij} = \begin{cases} 0 & \text{if } i < j \\ 1 & \text{if } i = j \\ \left(\dfrac{\lambda_i}{\lambda_j}\right)^k L_{ij} & \text{if } i > j. \end{cases}$$

The hypothesis that $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$ implies that

$$\lim_{k \mapsto \infty} \Lambda^k L \Lambda^{-k} = I. \tag{$\dagger$}$$

Note that it is to obtain this limit that we made the hypothesis on the moduli of the eigenvalues. We can write

$$\Lambda^k L \Lambda^{-k} = I + F_k, \quad \text{with} \quad \lim_{k \mapsto \infty} F_k = 0,$$

and consequently, since $R(\Lambda^k L \Lambda^{-k}) = R(I + F_k) = R + RF_k R^{-1} R = (I + RF_k R^{-1})R$, we have

$$R(\Lambda^k L \Lambda^{-k}) = (I + RF_k R^{-1})R. \tag{$*_5$}$$

By Proposition 8.8(1), since $\lim_{k \mapsto \infty} F_k = 0$, and thus $\lim_{k \mapsto \infty} RF_k R^{-1} = 0$, the matrices $I + RF_k R^{-1}$ are invertible for k large enough. Consequently for k large enough, we have a QR-factorization

$$I + RF_k R^{-1} = \widetilde{Q}_k \widetilde{R}_k, \tag{$*_6$}$$

with $(\widetilde{R}_k)_{ii} > 0$ for $i = 1, \ldots, n$. Since the matrices $\widetilde{Q}_k$ are unitary, we have $\left\|\widetilde{Q}_k\right\|_2 = 1$, so the sequence $(\widetilde{Q}_k)$ is bounded. It follows that it has a convergent subsequence $(\widetilde{Q}_\ell)$ that converges to some matrix $\widetilde{Q}$, which is also unitary. Since

$$\widetilde{R}_\ell = (\widetilde{Q}_\ell)^*(I + RF_\ell R^{-1}),$$

we deduce that the subsequence $(\widetilde{R}_\ell)$ also converges to some matrix $\widetilde{R}$, which is also upper triangular with positive diagonal entries. By passing to the limit (using the subsequences), we get $\widetilde{R} = (\widetilde{Q})^*$, that is,

$$I = \widetilde{Q}\widetilde{R}.$$

By the uniqueness of a QR-decomposition (when the diagonal entries of R are positive), we get

$$\widetilde{Q} = \widetilde{R} = I.$$

Since the above reasoning applies to any subsequences of $(\widetilde{Q}_k)$ and $(\widetilde{R}_k)$, by the uniqueness of limits, we conclude that the "full" sequences $(\widetilde{Q}_k)$ and $(\widetilde{R}_k)$ converge:

$$\lim_{k\mapsto\infty} \widetilde{Q}_k = I, \quad \lim_{k\mapsto\infty} \widetilde{R}_k = I.$$

Since by $(*_4)$,

$$A^k = QR(\Lambda^k L \Lambda^{-k})\Lambda^k U,$$

by $(*_5)$,

$$R(\Lambda^k L \Lambda^{-k}) = (I + RF_k R^{-1})R,$$

and by $(*_6)$

$$I + RF_k R^{-1} = \widetilde{Q}_k \widetilde{R}_k,$$

we proved that

$$A^k = (Q\widetilde{Q}_k)(\widetilde{R}_k R\Lambda^k U). \tag{$*_7$}$$

Observe that $Q\widetilde{Q}_k$ is a unitary matrix, and $\widetilde{R}_k R\Lambda^k U$ is an upper triangular matrix, as a product of upper triangular matrices. However, some entries in Λ may be negative, so we can't claim that $\widetilde{R}_k R\Lambda^k U$ has positive diagonal entries. Nevertheless, we have another QR-decomposition of A^k,

$$A^k = (Q\widetilde{Q}_k)(\widetilde{R}_k R\Lambda^k U) = P_k \mathcal{R}_k.$$

It is easy to prove that there is diagonal matrix D_k with $|(D_k)_{ii}| = 1$ for $i = 1, \ldots, n$, such that

$$P_k = Q\widetilde{Q}_k D_k. \tag{$*_8$}$$

The existence of D_k is consequence of the following fact: If an invertible matrix B has two QR factorizations $B = Q_1 R_1 = Q_2 R_2$, then there is a diagonal matrix D with unit entries such that $Q_2 = DQ_1$.

The expression for P_k in $(*_8)$ is that which we were seeking.

Step 3. Asymptotic behavior of the matrices $A_{k+1} = P_k^* A P_k$.

Since $A = P\Lambda P^{-1} = QR\Lambda R^{-1}Q^{-1}$ and by $(*_8)$, $P_k = Q\widetilde{Q}_k D_k$, we get

$$A_{k+1} = D_k^*(\widetilde{Q}_k)^* Q^* QR\Lambda R^{-1}Q^{-1}Q\widetilde{Q}_k D_k = D_k^*(\widetilde{Q}_k)^* R\Lambda R^{-1}\widetilde{Q}_k D_k. \tag{$*_9$}$$

Since $\lim_{k\mapsto\infty} \widetilde{Q}_k = I$, we deduce that

$$\lim_{k\mapsto\infty} (\widetilde{Q}_k)^* R\Lambda R^{-1}\widetilde{Q}_k = R\Lambda R^{-1} = \begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & \lambda_2 & \cdots & * \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix},$$

an upper triangular matrix with the eigenvalues of A on the diagonal. Since R is upper triangular, the order of the eigenvalues is preserved. If we let

$$\mathcal{D}_k = (\widetilde{Q}_k)^* R\Lambda R^{-1}\widetilde{Q}_k, \tag{$*_{10}$}$$

then by $(*_9)$ we have $A_{k+1} = D_k^*\mathcal{D}_k D_k$, and since the matrices D_k are diagonal matrices, we have

$$(A_{k+1})_{jj} = (D_k^*\mathcal{D}_k D_k)_{ij} = \overline{(D_k)_{ii}}(D_k)_{jj}(\mathcal{D}_k)_{ij},$$

which implies that

$$(A_{k+1})_{ii} = (\mathcal{D}_k)_{ii}, \quad i = 1, \ldots, n, \tag{$*_{11}$}$$

since $|(D_k)_{ii}| = 1$ for $i = 1, \ldots, n$. Since $\lim_{k\mapsto\infty}\mathcal{D}_k = R\Lambda R^{-1}$, we conclude that the strictly lower-triangular part of A_{k+1} converges to zero, and the diagonal of A_{k+1} converges to Λ. □

Observe that if the matrix A is real, then the hypothesis that the eigenvalues have distinct moduli implies that the eigenvalues are all real and simple.

The following `Matlab` program implements the basic QR-method using the function `qrv4` from Section 11.8.

```
function T = qreigen(A,m)
T = A;
for k = 1:m
  [Q R] = qrv4(T);
  T = R*Q;
end
end
```

Example 17.1. If we run the function `qreigen` with 100 iterations on the 8×8 symmetric matrix

$$A = \begin{pmatrix} 4&1&0&0&0&0&0&0\\ 1&4&1&0&0&0&0&0\\ 0&1&4&1&0&0&0&0\\ 0&0&1&4&1&0&0&0\\ 0&0&0&1&4&1&0&0\\ 0&0&0&0&1&4&1&0\\ 0&0&0&0&0&1&4&1\\ 0&0&0&0&0&0&1&4 \end{pmatrix},$$

we find the matrix

$$T = \begin{pmatrix} 5.8794 & 0.0015 & 0.0000 & -0.0000 & 0.0000 & -0.0000 & 0.0000 & -0.0000 \\ 0.0015 & 5.5321 & 0.0001 & 0.0000 & -0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0 & 0.0001 & 5.0000 & 0.0000 & -0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0 & 0 & 0.0000 & 4.3473 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0 & 0 & 0 & 0.0000 & 3.6527 & 0.0000 & 0.0000 & -0.0000 \\ 0 & 0 & 0 & 0 & 0.0000 & 3.0000 & 0.0000 & -0.0000 \\ 0 & 0 & 0 & 0 & 0 & 0.0000 & 2.4679 & 0.0000 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.0000 & 2.1.206 \end{pmatrix}.$$

The diagonal entries match the eigenvalues found by running the `Matlab` function `eig(A)`.

If several eigenvalues have the same modulus, then the proof breaks down, we can no longer claim (†), namely that

$$\lim_{k \mapsto \infty} \Lambda^k L \Lambda^{-k} = I.$$

If we assume that P^{-1} has a suitable "block LU-factorization," it can be shown that the matrices A_{k+1} converge to a block upper-triangular matrix, where each block corresponds to eigenvalues having the same modulus. For example, if A is a 9×9 matrix with eigenvalues λ_i such that $|\lambda_1| = |\lambda_2| = |\lambda_3| > |\lambda_4| > |\lambda_5| = |\lambda_6| = |\lambda_7| = |\lambda_8| = |\lambda_9|$, then A_k converges to a block diagonal matrix (with three blocks, a 3×3 block, a 1×1 block, and a 5×5 block) of the form

$$\begin{pmatrix} \star & \star & \star & * & * & * & * & * & * \\ \star & \star & \star & * & * & * & * & * & * \\ \star & \star & \star & * & * & * & * & * & * \\ 0 & 0 & 0 & \star & * & * & * & * & * \\ 0 & 0 & 0 & 0 & \star & \star & \star & \star & \star \\ 0 & 0 & 0 & 0 & \star & \star & \star & \star & \star \\ 0 & 0 & 0 & 0 & \star & \star & \star & \star & \star \\ 0 & 0 & 0 & 0 & \star & \star & \star & \star & \star \\ 0 & 0 & 0 & 0 & \star & \star & \star & \star & \star \end{pmatrix}.$$

See Ciarlet [Ciarlet (1989)] (Chapter 6 Section 6.3) for more details.

Under the conditions of Theorem 17.1, in particular, if A is a symmetric (or Hermitian) positive definite matrix, the eigenvectors of A can be approximated. However, when A is not a symmetric matrix, since the upper triangular part of A_k does not necessarily converge, one has to be cautious that a rigorous justification is lacking.

Suppose we apply the QR algorithm to a matrix A satisfying the hypotheses of Theorem Theorem 17.1. For k large enough, $A_{k+1} = P_k^* A P_k$ is nearly upper triangular and the diagonal entries of A_{k+1} are all distinct, so we can consider that they are the eigenvalues of A_{k+1}, and thus of A. To avoid too many subscripts, write T for the upper triangular matrix obtained by settting the entries of the part of A_{k+1} below the diagonal to 0. Then we can find the corresponding eigenvectors by solving the linear system

$$Tv = t_{ii}v,$$

and since T is upper triangular, this can be done by bottom-up elimination. We leave it as an exercise to show that the following vectors $v^i = (v_1^i, \ldots, v_n^i)$ are eigenvectors:

$$v^1 = e_1,$$

and if $i = 2, \ldots, n$, then

$$v_j^i = \begin{cases} 0 & \text{if } i+1 \leq j \leq n \\ 1 & \text{if } j = i \\ -\dfrac{t_{jj+1}v_{j+1}^i + \cdots + t_{ji}v_i^i}{t_{jj} - t_{ii}} & \text{if } i-1 \geq j \geq 1. \end{cases}$$

Then the vectors $(P_k v^1, \ldots, P_k v^n)$ are a basis of (approximate) eigenvectors for A. In the special case where T is a diagonal matrix, then $v^i = e_i$ for $i = 1, \ldots, n$ and the columns of P_k are an orthonormal basis of (approximate) eigenvectors for A.

If A is a real matrix whose eigenvalues are not all real, then there is some complex pair of eigenvalues $\lambda + i\mu$ (with $\mu \neq 0$), and the QR-algorithm cannot converge to a matrix whose strictly lower-triangular part is zero. There is a way to deal with this situation using upper Hessenberg matrices which will be discussed in the next section.

Since the convergence of the QR method depends crucially only on the fact that the part of A_k below the diagonal goes to zero, it would be highly desirable if we could replace A by a similar matrix U^*AU easily computable from A having lots of zero strictly below the diagonal. We can't expect U^*AU to be a diagonal matrix (since this would mean that A was easily diagonalized), but it turns out that there is a way to construct a matrix $H = U^*AU$ which is almost triangular, except that it may have an extra nonzero diagonal below the main diagonal. Such matrices called Hessenberg matrices are discussed in the next section.

17.2 Hessenberg Matrices

Definition 17.1. An $n \times n$ matrix (real or complex) H is an *(upper) Hessenberg matrix* if it is almost triangular, except that it may have an extra nonzero diagonal below the main diagonal. Technically, $h_{jk} = 0$ for all (j, k) such that $j - k \geq 2$.

The 5×5 matrix below is an example of a Hessenberg matrix.

$$H = \begin{pmatrix} * & * & * & * & * \\ h_{21} & * & * & * & * \\ 0 & h_{32} & * & * & * \\ 0 & 0 & h_{43} & * & * \\ 0 & 0 & 0 & h_{54} & * \end{pmatrix}.$$

The following result can be shown.

Theorem 17.2. *Every $n \times n$ complex or real matrix A is similar to an upper Hessenberg matrix H, that is, $A = UHU^*$ for some unitary matrix U. Furthermore, H can be constructed as a product of Householder matrices (the definition is the same as in Section 12.1, except that W is a complex vector, and that the inner product is the Hermitian inner product on $\mathbb{C}^n$). If A is a real matrix, then H is an orthogonal matrix (and H is a real matrix).*

Theorem 17.2 and algorithms for converting a matrix to Hessenberg form are discussed in Trefethen and Bau [Trefethen and Bau III (1997)] (Lecture 26), Demmel [Demmel (1997)] (Section 4.4.6, in the real case), Serre [Serre (2010)] (Theorem 13.1), and Meyer [Meyer (2000)] (Example 5.7.4, in the real case). The proof of correctness is not difficult and will be the object of a homework problem.

The following functions written in `Matlab` implement a function to compute a Hessenberg form of a matrix.

The function `house` constructs the normalized vector u defining the Householder reflection that zeros all but the first entries in a vector x.

```
function [uu, u] = house(x)
tol = 2*10^(-15);   % tolerance
uu = x;
p = size(x,1);
% computes l^1-norm of x(2:p,1)
```

```
n1 = sum(abs(x(2:p,1)));
if n1 <= tol
   u = zeros(p,1);  uu = u;
else
   l = sqrt(x'*x);  % l^2 norm of x
   uu(1) = x(1) + signe(x(1))*l;
   u = uu/sqrt(uu'*uu);
end
end
```

The function `signe(z)` returms -1 if $z < 0$, else $+1$.

The function `buildhouse` builds a Householder reflection from a vector uu.

```
function P = buildhouse(v,i)
% This function builds a Householder reflection
%   [I 0 ]
%   [0 PP]
%   from a Householder reflection
%   PP = I - 2uu*uu'
%   where uu = v(i:n)
%   If uu = 0 then P - I
%

n = size(v,1);
if v(i:n) == zeros(n - i + 1,1)
   P = eye(n);
else
   PP = eye(n - i + 1) - 2*v(i:n)*v(i:n)';
   P = [eye(i-1) zeros(i-1, n - i + 1);
        zeros(n - i + 1, i - 1) PP];
end
end
```

The function `Hessenberg1` computes an upper Hessenberg matrix H and an orthogonal matrix Q such that $A = Q^\top HQ$.

```
function [H, Q] = Hessenberg1(A)
%
%   This function constructs an upper Hessenberg
```

```
%   matrix H and an orthogonal matrix Q such that
%   A = Q' H Q

n = size(A,1);
H = A;
Q = eye(n);
for i = 1:n-2
  %  H(i+1:n,i)
    [~,u] = house(H(i+1:n,i));
  %  u
    P =  buildhouse(u,1);
    Q(i+1:n,i:n) = P*Q(i+1:n,i:n);
    H(i+1:n,i:n) = H(i+1:n,i:n) - 2*u*(u')*H(i+1:n,i:n);
    H(1:n,i+1:n) = H(1:n,i+1:n) - 2*H(1:n,i+1:n)*u*(u');
end
end
```

Example 17.2. If

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{pmatrix},$$

running `Hessenberg1` we find

$$H = \begin{pmatrix} 1.0000 & -5.3852 & 0 & 0 \\ -5.3852 & 15.2069 & -1.6893 & -0.0000 \\ -0.0000 & -1.6893 & -0.2069 & -0.0000 \\ 0 & -0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

$$Q = \begin{pmatrix} 1.0000 & 0 & 0 & 0 \\ 0 & -0.3714 & -0.5571 & -0.7428 \\ 0 & 0.8339 & 0.1516 & -0.5307 \\ 0 & 0.4082 & -0.8165 & 0.4082 \end{pmatrix}.$$

An important property of (upper) Hessenberg matrices is that if some subdiagonal entry $H_{p+1p} = 0$, then H is of the form

$$H = \begin{pmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{pmatrix},$$

where both H_{11} and H_{22} are upper Hessenberg matrices (with H_{11} a $p \times p$ matrix and H_{22} a $(n-p) \times (n-p)$ matrix), and the eigenvalues of H are

the eigenvalues of H_{11} and H_{22}. For example, in the matrix

$$H = \begin{pmatrix} * & * & * & * & * \\ h_{21} & * & * & * & * \\ 0 & h_{32} & * & * & * \\ 0 & 0 & h_{43} & * & * \\ 0 & 0 & 0 & h_{54} & * \end{pmatrix},$$

if $h_{43} = 0$, then we have the block matrix

$$H = \begin{pmatrix} * & * & * & * & * \\ h_{21} & * & * & * & * \\ 0 & h_{32} & * & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & h_{54} & * \end{pmatrix}.$$

Then the list of eigenvalues of H is the concatenation of the list of eigenvalues of H_{11} and the list of the eigenvalues of H_{22}. This is easily seen by induction on the dimension of the block H_{11}.

More generally, every upper Hessenberg matrix can be written in such a way that it has diagonal blocks that are Hessenberg blocks whose subdiagonal is not zero.

Definition 17.2. An upper Hessenberg $n \times n$ matrix H is *unreduced* if $h_{i+1i} \neq 0$ for $i = 1, \ldots, n-1$. A Hessenberg matrix which is not unreduced is said to be *reduced*.

The following is an example of an 8×8 matrix consisting of three diagonal unreduced Hessenberg blocks:

$$H = \begin{pmatrix} \star & \star & \star & * & * & * & * & * \\ \mathbf{h_{21}} & \star & \star & * & * & * & * & * \\ \mathbf{0} & \mathbf{h_{32}} & \star & * & * & * & * & * \\ 0 & 0 & 0 & \star & \star & \star & * & * \\ 0 & 0 & 0 & \mathbf{h_{54}} & \star & \star & * & * \\ 0 & 0 & 0 & \mathbf{0} & \mathbf{h_{65}} & \star & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & \star & \star \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{h_{87}} & \star \end{pmatrix}.$$

An interesting and important property of unreduced Hessenberg matrices is the following.

Proposition 17.1. *Let H be an $n \times n$ complex or real unreduced Hessenberg matrix. Then every eigenvalue of H is geometrically simple, that is,*

$\dim(E_\lambda) = 1$ *for every eigenvalue* λ*, where* E_λ *is the eigenspace associated with* λ*. Furthermore, if* H *is diagonalizable, then every eigenvalue is simple, that is,* H *has* n *distinct eigenvalues.*

Proof. We follow Serre's proof [Serre (2010)] (Proposition 3.26). Let λ be any eigenvalue of H, let $M = \lambda I_n - H$, and let N be the $(n-1) \times (n-1)$ matrix obtained from M by deleting its first row and its last column. Since H is upper Hessenberg, N is a diagonal matrix with entries $-h_{i+1i} \neq 0$, $i = 1, \ldots, n-1$. Thus N is invertible and has rank $n-1$. But a matrix has rank greater than or equal to the rank of any of its submatrices, so $\text{rank}(M) = n-1$, since M is singular. By the rank-nullity theorem, $\text{rank}(\text{Ker}\, N) = 1$, that is, $\dim(E_\lambda) = 1$, as claimed.

If H is diagonalizable, then the sum of the dimensions of the eigenspaces is equal to n, which implies that the eigenvalues of H are distinct. □

As we said earlier, a case where Theorem 17.1 applies is the case where A is a symmetric (or Hermitian) positive definite matrix. This follows from two facts.

The first fact is that if A is Hermitian (or symmetric in the real case), then it is easy to show that the Hessenberg matrix similar to A is a Hermitian (or symmetric in real case) *tridiagonal matrix*. The conversion method is also more efficient. Here is an example of a symmetric tridiagonal matrix consisting of three unreduced blocks:

$$H = \begin{pmatrix} \alpha_1 & \beta_1 & \mathbf{0} & 0 & 0 & 0 & 0 & 0 \\ \beta_1 & \alpha_2 & \beta_2 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{0} & \beta_2 & \alpha_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha_4 & \beta_4 & \mathbf{0} & 0 & 0 \\ 0 & 0 & 0 & \beta_4 & \alpha_5 & \beta_5 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{0} & \beta_5 & \alpha_6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha_7 & \beta_7 \\ 0 & 0 & 0 & 0 & 0 & 0 & \beta_7 & \alpha_8 \end{pmatrix}.$$

Thus the problem of finding the eigenvalues of a symmetric (or Hermitian) matrix reduces to the problem of finding the eigenvalues of a symmetric (resp. Hermitian) tridiagonal matrix, and this can be done much more efficiently.

The second fact is that if H is an upper Hessenberg matrix and if it is diagonalizable, then there is an invertible matrix P such that $H = P\Lambda P^{-1}$ with Λ a diagonal matrix consisting of the eigenvalues of H, such that P^{-1} has an LU-decomposition; see Serre [Serre (2010)] (Theorem 13.3).

As a consequence, since any symmetric (or Hermitian) tridiagonal matrix is a block diagonal matrix of unreduced symmetric (resp. Hermitian) tridiagonal matrices, by Proposition 17.1, we see that the QR algorithm applied to a tridiagonal matrix which is symmetric (or Hermitian) positive definite converges to a diagonal matrix consisting of its eigenvalues. Let us record this important fact.

Theorem 17.3. *Let H be a symmetric (or Hermitian) positive definite tridiagonal matrix. If H is unreduced, then the QR algorithm converges to a diagonal matrix consisting of the eigenvalues of H.*

Since every symmetric (or Hermitian) positive definite matrix is similar to tridiagonal symmetric (resp. Hermitian) positive definite matrix, we deduce that we have a method for finding the eigenvalues of a symmetric (resp. Hermitian) positive definite matrix (more accurately, to find approximations as good as we want for these eigenvalues).

If A is a symmetric (or Hermitian) matrix, since its eigenvalues are real, for some $\mu > 0$ large enough (pick $\mu > \rho(A)$), $A + \mu I$ is symmetric (resp. Hermitan) positive definite, so we can apply the QR algorithm to an upper Hessenberg matrix similar to $A + \mu I$ to find its eigenvalues, and then the eigenvalues of A are obtained by subtracting μ.

The problem of finding the eigenvalues of a symmetric matrix is discussed extensively in Parlett [Parlett (1997)], one of the best references on this topic.

The upper Hessenberg form also yields a way to handle singular matrices. First, checking the proof of Proposition 13.20 that an $n \times n$ complex matrix A (possibly singular) can be factored as $A = QR$ where Q is a unitary matrix which is a product of Householder reflections and R is upper triangular, it is easy to see that if A is upper Hessenberg, then Q is also upper Hessenberg. If H is an unreduced upper Hessenberg matrix, since Q is upper Hessenberg and R is upper triangular, we have $h_{i+1i} = q_{i+1i} r_{ii}$ for $i = 1 \ldots, n-1$, and since H is unreduced, $r_{ii} \neq 0$ for $i = 1, \ldots, n-1$. Consequently H is singular iff $r_{nn} = 0$. Then the matrix RQ is a matrix whose last row consists of zero's thus we can deflate the problem by considering the $(n-1) \times (n-1)$ unreduced Hessenberg matrix obtained by deleting the last row and the last column. After finitely many steps (not larger that the multiplicity of the eigenvalue 0), there remains an invertible unreduced Hessenberg matrix. As an alternative, see Serre [Serre (2010)] (Chapter 13, Section 13.3.2).

As is, the QR algorithm, although very simple, is quite inefficient for

several reasons. In the next section, we indicate how to make the method more efficient. This involves a lot of work and we only discuss the main ideas at a high level.

17.3 Making the QR Method More Efficient Using Shifts

To improve efficiency and cope with pairs of complex conjugate eigenvalues in the case of real matrices, the following steps are taken:

(1) Initially reduce the matrix A to upper Hessenberg form, as $A = UHU^*$. Then apply the QR-algorithm to H (actually, to its unreduced Hessenberg blocks). It is easy to see that the matrices H_k produced by the QR algorithm remain upper Hessenberg.
(2) To accelerate convergence, use *shifts*, and to deal with pairs of complex conjugate eigenvalues, use *double shifts*.
(3) Instead of computing a QR-factorization explicitly while doing a shift, perform an *implicit shift* which computes $A_{k+1} = Q_k^* A_k Q_k$ without having to compute a QR-factorization (of $A_k - \sigma_k I$), and similarly in the case of a double shift. This is the most intricate modification of the basic QR algorithm and we will not discuss it here. This method is usually referred as *bulge chasing*. Details about this technique for real matrices can be found in Demmel [Demmel (1997)] (Section 4.4.8) and Golub and Van Loan [Golub and Van Loan (1996)] (Section 7.5). Watkins discusses the QR algorithm with shifts as a bulge chasing method in the more general case of complex matrices [Watkins (1982, 2008)].

Let us repeat an important remark made in the previous section. If we start with a matrix H in upper Hessenberg form, if at any stage of the QR algorithm we find that some subdiagonal entry $(H_k)_{p+1p} = 0$ *or is very small*, then H_k is of the form

$$H_k = \begin{pmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{pmatrix},$$

where both H_{11} and H_{22} are upper Hessenberg matrices (with H_{11} a $p \times p$ matrix and H_{22} a $(n-p) \times (n-p)$ matrix), and the eigenvalues of H_k are

the eigenvalues of H_{11} and H_{22}. For example, in the matrix

$$H = \begin{pmatrix} * & * & * & * & * \\ h_{21} & * & * & * & * \\ 0 & h_{32} & * & * & * \\ 0 & 0 & h_{43} & * & * \\ 0 & 0 & 0 & h_{54} & * \end{pmatrix},$$

if $h_{43} = 0$, then we have the block matrix

$$H = \begin{pmatrix} * & * & * & * & * \\ h_{21} & * & * & * & * \\ 0 & h_{32} & * & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & h_{54} & * \end{pmatrix}.$$

Then we can recursively apply the QR algorithm to H_{11} and H_{22}.

In particular, if $(H_k)_{nn-1} = 0$ or is very small, then $(H_k)_{nn}$ is a good approximation of an eigenvalue, so we can delete the last row and the last column of H_k and apply the QR algorithm to this submatrix. This process is called *deflation*. If $(H_k)_{n-1n-2} = 0$ or is very small, then the 2×2 "corner block"

$$\begin{pmatrix} (H_k)_{n-1n-1} & (H_k)_{n-1n} \\ (H_k)_{nn-1} & (H_k)_{nn} \end{pmatrix}$$

appears, and its eigenvalues can be computed immediately by solving a quadratic equation. Then we deflate H_k by deleting its last two rows and its last two columns and apply the QR algorithm to this submatrix.

Thus it would seem desirable to modify the basic QR algorithm so that the above situations arises, and this is what shifts are designed for. More precisely, under the hypotheses of Theorem 17.1, it can be shown (see Ciarlet [Ciarlet (1989)], Section 6.3) that the entry $(A_k)_{ij}$ with $i > j$ converges to 0 as $|\lambda_i/\lambda_j|^k$ converges to 0. Also, if we let r_i be defined by

$$r_1 = \left|\frac{\lambda_2}{\lambda_1}\right|, \quad r_i = \max\left\{\left|\frac{\lambda_i}{\lambda_{i-1}}\right|, \left|\frac{\lambda_{i+1}}{\lambda_i}\right|\right\}, \quad 2 \leq i \leq n-1, \quad r_n = \left|\frac{\lambda_n}{\lambda_{n-1}}\right|,$$

then there is a constant C (independent of k) such that

$$|(A_k)_{ii} - \lambda_i| \leq C r_i^k, \quad 1 \leq i \leq n.$$

In particular, if H is upper Hessenberg, then the entry $(H_k)_{i+1i}$ converges to 0 as $|\lambda_{i+1}/\lambda_i|^k$ converges to 0. Thus if we pick σ_k close to λ_i, we expect that $(H_k - \sigma_k I)_{i+1i}$ converges to 0 as $|(\lambda_{i+1} - \sigma_k)/(\lambda_i - \sigma_k)|^k$ converges to 0, and this ratio is much smaller than 1 as σ_k is closer to λ_i.

Typically, we apply a shift to accelerate convergence to λ_n (so $i = n-1$). In this case, both $(H_k - \sigma_k I)_{nn-1}$ and $|(H_k - \sigma_k I)_{nn} - \lambda_n|$ converge to 0 as $|(\lambda_n - \sigma_k)/(\lambda_{n-1} - \sigma_k)|^k$ converges to 0.

A *shift* is the following modified QR-steps (switching back to an arbitrary matrix A, since the shift technique applies in general). Pick some σ_k, hopefully close to some eigenvalue of A (in general, λ_n), and QR-factor $A_k - \sigma_k I$ as

$$A_k - \sigma_k I = Q_k R_k,$$

and then form

$$A_{k+1} = R_k Q_k + \sigma_k I.$$

Since

$$\begin{aligned} A_{k+1} &= R_k Q_k + \sigma_k I \\ &= Q_k^* Q_k R_k Q_k + Q_k^* Q_k \sigma_k \\ &= Q_k^* (Q_k R_k + \sigma_k I) Q_k \\ &= Q_k^* A_k Q_k, \end{aligned}$$

A_{k+1} is similar to A_k, as before. If A_k is upper Hessenberg, then it is easy to see that A_{k+1} is also upper Hessenberg.

If A is upper Hessenberg and if σ_i is exactly equal to an eigenvalue, then $A_k - \sigma_k I$ is singular, and forming the QR-factorization will detect that R_k has some diagonal entry equal to 0. Assuming that the QR-algorithm returns $(R_k)_{nn} = 0$ (if not, the argument is easily adapted), then the last row of $R_k Q_k$ is 0, so the last row of $A_{k+1} = R_k Q_k + \sigma_k I$ ends with σ_k (all other entries being zero), so we are in the case where we can deflate A_k (and σ_k is indeed an eigenvalue).

The question remains, what is a good choice for the shift σ_k?

Assuming again that H is in upper Hessenberg form, it turns out that when $(H_k)_{nn-1}$ is small enough, then a good choice for σ_k is $(H_k)_{nn}$. In fact, the rate of convergence is quadratic, which means roughly that the number of correct digits doubles at every iteration. The reason is that shifts are related to another method known as inverse iteration, and such a method converges very fast. For further explanations about this connection, see Demmel [Demmel (1997)] (Section 4.4.4) and Trefethen and Bau [Trefethen and Bau III (1997)] (Lecture 29).

One should still be cautious that the QR method with shifts does not necessarily converge, and that our convergence proof no longer applies, because instead of having the identity $A^k = P_k \mathcal{R}_k$, we have

$$(A - \sigma_k I) \cdots (A - \sigma_2 I)(A - \sigma_1 I) = P_k \mathcal{R}_k.$$

Of course, the QR algorithm loops immediately when applied to an orthogonal matrix A. This is also the case when A is symmetric but not positive definite. For example, both the QR algorithm and the QR algorithm with shifts loop on the matrix

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

In the case of symmetric matrices, Wilkinson invented a shift which helps the QR algorithm with shifts to make progress. Again, looking at the lower corner of A_k, say

$$B = \begin{pmatrix} a_{n-1} & b_{n-1} \\ b_{n-1} & a_n \end{pmatrix},$$

the *Wilkinson shift* picks the eigenvalue of B closer to a_n. If we let

$$\delta = \frac{a_{n-1} - a_n}{2},$$

it is easy to see that the eigenvalues of B are given by

$$\lambda = \frac{a_n + a_{n-1}}{2} \pm \sqrt{\delta^2 + b_{n-1}^2}.$$

It follows that

$$\lambda - a_n = \delta \pm \sqrt{\delta^2 + b_{n-1}^2},$$

and from this it is easy to see that the eigenvalue closer to a_n is given by

$$\mu = a_n - \frac{\operatorname{sign}(\delta) b_{n-1}^2}{(|\delta| + \sqrt{\delta^2 + b_{n-1}^2})}.$$

If $\delta = 0$, then we pick arbitrarily one of the two eigenvalues. Observe that the Wilkinson shift applied to the matrix

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

is either $+1$ or -1, and in one step, deflation occurs and the algorithm terminates successfully.

We now discuss double shifts, which are intended to deal with pairs of complex conjugate eigenvalues.

Let us assume that A is a real matrix. For any complex number σ_k with nonzero imaginary part, a *double shift* consists of the following steps:

$$\begin{aligned} A_k - \sigma_k I &= Q_k R_k \\ A_{k+1} &= R_k Q_k + \sigma_k I \\ A_{k+1} - \overline{\sigma}_k I &= Q_{k+1} R_{k+1} \\ A_{k+2} &= R_{k+1} Q_{k+1} + \overline{\sigma}_k I. \end{aligned}$$

From the computation made for a single shift, we have $A_{k+1} = Q_k^* A_k Q_k$ and $A_{k+2} = Q_{k+1}^* A_{k+1} Q_{k+1}$, so we obtain

$$A_{k+2} = Q_{k+1}^* Q_k^* A_k Q_k Q_{k+1}.$$

The matrices Q_k are complex, so we would expect that the A_k are also complex, but remarkably we can keep the products $Q_k Q_{k+1}$ real, and so the A_k also real. This is highly desirable to avoid complex arithmetic, which is more expensive.

Observe that since

$$Q_{k+1} R_{k+1} = A_{k+1} - \overline{\sigma}_k I = R_k Q_k + (\sigma_k - \overline{\sigma}_k) I,$$

we have

$$\begin{aligned} Q_k Q_{k+1} R_{k+1} R_k &= Q_k (R_k Q_k + (\sigma_k - \overline{\sigma}_k) I) R_k \\ &= Q_k R_k Q_k R_k + (\sigma_k - \overline{\sigma}_k) Q_k R_k \\ &= (A_k - \sigma_k I)^2 + (\sigma_k - \overline{\sigma}_k)(A_k - \sigma_k I) \\ &= A_k^2 - 2(\Re \sigma_k) A_k + |\sigma_k|^2 I. \end{aligned}$$

If we assume by induction that matrix A_k is real (with $k = 2\ell+1, \ell \geq 0$), then the matrix $S = A_k^2 - 2(\Re \sigma_k) A_k + |\sigma_k|^2 I$ is also real, and since $Q_k Q_{k+1}$ is unitary and $R_{k+1} R_k$ is upper triangular, we see that

$$S = Q_k Q_{k+1} R_{k+1} R_k$$

is a QR-factorization of the real matrix S, thus $Q_k Q_{k+1}$ and $R_{k+1} R_k$ can be chosen to be real matrices, in which case $(Q_k Q_{k+1})^*$ is also real, and thus

$$A_{k+2} = Q_{k+1}^* Q_k^* A_k Q_k Q_{k+1} = (Q_k Q_{k+1})^* A_k Q_k Q_{k+1}$$

is real. Consequently, if $A_1 = A$ is real, then $A_{2\ell+1}$ is real for all $\ell \geq 0$.

The strategy that consists in picking σ_k and $\overline{\sigma}_k$ as the complex conjugate eigenvalues of the corner block

$$\begin{pmatrix} (H_k)_{n-1\,n-1} & (H_k)_{n-1\,n} \\ (H_k)_{n\,n-1} & (H_k)_{nn} \end{pmatrix}$$

is called the *Francis shift* (here we are assuming that A has be reduced to upper Hessenberg form).

It should be noted that there are matrices for which neither a shift by $(H_k)_{nn}$ nor the Francis shift works. For instance, the permutation matrix

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

has eigenvalues $e^{i2\pi/3}, e^{i4\pi/3}, +1$, and neither of the above shifts apply to the matrix

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

However, a shift by 1 does work. There are other kinds of matrices for which the QR algorithm does not converge. Demmel gives the example of matrices of the form

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & h & 0 \\ 0 & -h & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

where h is small.

Algorithms implementing the QR algorithm with shifts and double shifts perform "exceptional" shifts every 10 shifts. Despite the fact that the QR algorithm has been perfected since the 1960's, it is still an open problem to find a shift strategy that ensures convergence of all matrices.

Implicit shifting is based on a result known as the *implicit Q theorem*. This theorem says that if A is reduced to upper Hessenberg form as $A = UHU^*$ and if H is unreduced ($h_{i+1i} \neq 0$ for $i = 1, \ldots, n-1$), then the columns of index $2, \ldots, n$ of U are determined by the first column of U up to sign; see Demmel [Demmel (1997)] (Theorem 4.9) and Golub and Van Loan [Golub and Van Loan (1996)] (Theorem 7.4.2) for the proof in the case of real matrices. Actually, the proof is not difficult and will be the object of a homework exercise. In the case of a single shift, an implicit shift generates $A_{k+1} = Q_k^* A_k Q_k$ without having to compute a QR-factorization of $A_k - \sigma_k I$. For real matrices, this is done by applying a sequence of Givens rotations which perform a bulge chasing process (a Givens rotation is an orthogonal block diagonal matrix consisting of a single block which is a 2D rotation, the other diagonal entries being equal to 1). Similarly, in the case of a double shift, $A_{k+2} = (Q_k Q_{k+1})^* A_k Q_k Q_{k+1}$ is generated without having to compute the QR-factorizations of $A_k - \sigma_k I$ and $A_{k+1} - \overline{\sigma}_k I$. Again, $(Q_k Q_{k+1})^* A_k Q_k Q_{k+1}$ is generated by applying some simple orthogonal matrices which perform a bulge chasing process. See Demmel [Demmel (1997)] (Section 4.4.8) and Golub and Van Loan [Golub and Van Loan (1996)] (Section 7.5) for further explanations regarding implicit shifting involving bulge chasing in the case of real matrices. Watkins [Watkins (1982, 2008)] discusses bulge chasing in the more general case of complex matrices.

The `Matlab` function for finding the eigenvalues and the eigenvectors of a matrix A is `eig` and is called as `[U, D] = eig(A)`. It is implemented using an optimized version of the QR-algorithm with implicit shifts.

If the dimension of the matrix A is very large, we can find approximations of some of the eigenvalues of A by using a truncated version of the reduction to Hessenberg form due to Arnoldi in general and to Lanczos in the symmetric (or Hermitian) tridiagonal case.

17.4 Krylov Subspaces; Arnoldi Iteration

In this section, we denote the dimension of the square real or complex matrix A by m rather than n, to make it easier for the reader to follow Trefethen and Bau exposition [Trefethen and Bau III (1997)], which is particularly lucid.

Suppose that the $m \times m$ matrix A has been reduced to the upper Hessenberg form H, as $A = UHU^*$. For any $n \leq m$ (typically much smaller than m), consider the $(n+1) \times n$ upper left block

$$\widetilde{H}_n = \begin{pmatrix} h_{11} & h_{12} & h_{13} & \cdots & h_{1n} \\ h_{21} & h_{22} & h_{23} & \cdots & h_{2n} \\ 0 & h_{32} & h_{33} & \cdots & h_{3n} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_{nn-1} & h_{nn} \\ 0 & \cdots & 0 & 0 & h_{n+1n} \end{pmatrix}$$

of H, and the $n \times n$ upper Hessenberg matrix H_n obtained by deleting the last row of $\widetilde{H}_n$,

$$H_n = \begin{pmatrix} h_{11} & h_{12} & h_{13} & \cdots & h_{1n} \\ h_{21} & h_{22} & h_{23} & \cdots & h_{2n} \\ 0 & h_{32} & h_{33} & \cdots & h_{3n} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_{nn-1} & h_{nn} \end{pmatrix}.$$

If we denote by U_n the $m \times n$ matrix consisting of the first n columns of U, denoted $u_1, \ldots, u_n$, then matrix consisting of the first n columns of the matrix $UH = AU$ can be expressed as

$$AU_n = U_{n+1}\widetilde{H}_n. \tag{$*_1$}$$

It follows that the nth column of this matrix can be expressed as

$$Au_n = h_{1n}u_1 + \cdots + h_{nn}u_n + h_{n+1n}u_{n+1}. \tag{$*_2$}$$

Since $(u_1, \ldots, u_n)$ form an orthonormal basis, we deduce from $(*_2)$ that

$$\langle u_j, Au_n \rangle = u_j^* A u_n = h_{jn}, \quad j = 1, \ldots, n. \tag{$*_3$}$$

Equations $(*_2)$ and $(*_3)$ show that U_{n+1} and $\widetilde{H}_n$ can be computed iteratively using the following algorithm due to Arnoldi, known as *Arnoldi iteration*:

Given an arbitrary nonzero vector $b \in \mathbb{C}^m$, let $u_1 = b/\|b\|$;
for $n = 1, 2, 3, \ldots$ **do**
  $z := Au_n$;
  for $j = 1$ **to** n **do**
    $h_{jn} := u_j^* z$;
    $z := z - h_{jn} u_j$
  endfor
  $h_{n+1n} := \|z\|$;
  if $h_{n+1n} = 0$ **quit**
  $u_{n+1} = z/h_{n+1n}$

When $h_{n+1n} = 0$, we say that we have a *breakdown* of the Arnoldi iteration.

Arnoldi iteration is an algorithm for producing the $n \times n$ Hessenberg submatrix H_n of the full Hessenberg matrix H consisting of its first n rows and n columns (the first n columns of U are also produced), not using Householder matrices.

As long as $h_{j+1j} \neq 0$ for $j = 1, \ldots, n$, Equation $(*_2)$ shows by an easy induction that u_{n+1} belong to the span of $(b, Ab, \ldots, A^n b)$, and obviously Au_n belongs to the span of $(u_1, \ldots, u_{n+1})$, and thus the following spaces are identical:

$$\mathrm{Span}(b, Ab, \ldots, A^n b) = \mathrm{Span}(u_1, \ldots, u_{n+1}).$$

The space $\mathcal{K}_n(A, b) = \mathrm{Span}(b, Ab, \ldots, A^{n-1}b)$ is called a *Krylov subspace*. We can view Arnoldi's algorithm as the construction of an orthonormal basis for $\mathcal{K}_n(A, b)$. It is a sort of Gram–Schmidt procedure.

Equation $(*_2)$ shows that if K_n is the $m \times n$ matrix whose columns are the vectors $(b, Ab, \ldots, A^{n-1}b)$, then there is a $n \times n$ upper triangular matrix R_n such that

$$K_n = U_n R_n. \tag{$*_4$}$$

The above is called a *reduced QR factorization* of K_n.

Since $(u_1, \ldots, u_n)$ is an orthonormal system, the matrix $U_n^* U_{n+1}$ is the $n \times (n+1)$ matrix consisting of the identity matrix I_n plus an extra column of 0's, so $U_n^* U_{n+1} \widetilde{H}_n = U_n^* A U_n$ is obtained by deleting the last row of $\widetilde{H}_n$, namely H_n, and so

$$U_n^* A U_n = H_n. \tag{$*_5$}$$

We summarize the above facts in the following proposition.

Proposition 17.2. *If Arnoldi iteration run on an $m \times m$ matrix A starting with a nonzero vector $b \in \mathbb{C}^m$ does not have a breakdown at stage $n \leq m$, then the following properties hold:*

(1) If K_n is the $m \times n$ Krylov matrix associated with the vectors $(b, Ab, \ldots, A^{n-1}b)$ and if U_n is the $m \times n$ matrix of orthogonal vectors produced by Arnoldi iteration, then there is a QR-factorization

$$K_n = U_n R_n,$$

for some $n \times n$ upper triangular matrix R_n.

(2) The $m \times n$ upper Hessenberg matrices H_n produced by Arnoldi iteration are the projection of A onto the Krylov space $\mathcal{K}_n(A, b)$, that is,

$$H_n = U_n^* A U_n.$$

(3) The successive iterates are related by the formula

$$A U_n = U_{n+1} \widetilde{H}_n.$$

Remark: If Arnoldi iteration has a breakdown at stage n, that is, $h_{n+1} = 0$, then we found the first unreduced block of the Hessenberg matrix H. It can be shown that the eigenvalues of H_n are eigenvalues of A. So a breakdown is actually a good thing. In this case, we can pick some new nonzero vector u_{n+1} orthogonal to the vectors $(u_1, \ldots, u_n)$ as a new starting vector and run Arnoldi iteration again. Such a vector exists since the $(n+1)$th column of U works. So repeated application of Arnoldi yields a full Hessenberg reduction of A. However, this is not what we are after, since m is very large an we are only interested in a "small" number of eigenvalues of A.

There is another aspect of Arnoldi iteration, which is that it solves an optimization problem involving polynomials of degree n. Let $\mathcal{P}^n$ denote the set of (complex) monic polynomials of degree n, that is, polynomials of the form

$$p(z) = z^n + c_{n-1} z^{n-1} + \cdots + c_1 z + c_0 \quad (c_i \in \mathbb{C}).$$

For any $m \times m$ matrix A, we write

$$p(A) = A^n + c_{n-1}A^{n-1} + \cdots + c_1 A + c_0 I.$$

The following result is proven in Trefethen and Bau [Trefethen and Bau III (1997)] (Lecture 34, Theorem 34.1).

Theorem 17.4. *If Arnoldi iteration run on an $m \times m$ matrix A starting with a nonzero vector b does not have a breakdown at stage $n \leq m$, then there is a unique polynomial $p \in \mathcal{P}^n$ such that $\|p(A)b\|_2$ is minimum, namely the characteristic polynomial* $\det(zI - H_n)$ *of H_n.*

Theorem 17.4 can be viewed as the "justification" for a method to find some of the eigenvalues of A (say $n \ll m$ of them). Intuitively, the closer the roots of the characteristic polynomials of H_n are to the eigenvalues of A, the smaller $\|p(A)b\|_2$ should be, and conversely. In the extreme case where $m = n$, by the Cayley–Hamilton theorem, $p(A) = 0$ (where p is the characteristic polynomial of A), so this idea is plausible, but this is far from constituting a proof (also, b should have nonzero coordinates in all directions associated with the eigenvalues).

The method known as the *Rayleigh–Ritz method* is to run Arnoldi iteration on A and some $b \neq 0$ chosen at random for $n \ll m$ steps before or until a breakdown occurs. Then run the QR algorithm with shifts on H_n. The eigenvalues of the Hessenberg matrix H_n may then be considered as approximations of the eigenvalues of A. The eigenvalues of H_n are called *Arnoldi estimates* or *Ritz values*. One has to be cautious because H_n is a truncated version of the full Hessenberg matrix H, so not all of the Ritz values are necessary close to eigenvalues of A. It has been observed that the eigenvalues that are found first are the *extreme* eigenvalues of A, namely those close to the boundary of the spectrum of A plotted in $\mathbb{C}$. So if A has real eigenvalues, the largest and the smallest eigenvalues appear first as Ritz values. In many problems where eigenvalues occur, the extreme eigenvalues are the one that need to be computed. Similarly, the eigenvectors of H_n may be considered as approximations of eigenvectors of A.

The `Matlab` function `eigs` is based on the computation of Ritz values. It computes the six eigenvalues of largest magnitude of a matrix A, and the call is `[V, D] = eigs(A)`. More generally, to get the top k eigenvalues, use `[V, D] = eigs(A, k)`.

In the absence of rigorous theorems about error estimates, it is hard to make the above statements more precise; see Trefethen and Bau [Trefethen and Bau III (1997)] (Lecture 34) for more on this subject.

However, if A is a symmetric (or Hermitian) matrix, then H_n is a symmetric (resp. Hermitian) tridiagonal matrix and more precise results can be shown; see Demmel [Demmel (1997)] (Chapter 7, especially Section 7.2). We will consider the symmetric (and Hermitan) case in the next section, but first we show how Arnoldi iteration can be used to find approximations for the solution of a linear system $Ax = b$ where A is invertible but of very large dimension m.

17.5 GMRES

Suppose A is an invertible $m \times m$ matrix and let b be a nonzero vector in $\mathbb{C}^m$. Let $x_0 = A^{-1}b$, the unique solution of $Ax = b$. It is not hard to show that $x_0 \in \mathcal{K}_n(A, b)$ for some $n \leq m$. In fact, there is a unique monic polynomial $p(z)$ of minimal degree $s \leq m$ such that $p(A)b = 0$, so $x_0 \in \mathcal{K}_s(A, b)$. Thus it makes sense to search for a solution of $Ax = b$ in Krylov spaces of dimension $m \leq s$. The idea is to find an approximation $x_n \in \mathcal{K}_n(A, b)$ of x_0 such that $r_n = b - Ax_n$ is minimized, that is, $\|r_n\|_2 = \|b - Ax_n\|_2$ is minimized over $x_n \in \mathcal{K}_n(A, b)$. This minimization problem can be stated as

$$\text{minimize} \quad \|r_n\|_2 = \|Ax_n - b\|_2, \quad x_n \in \mathcal{K}_n(A, b).$$

This is a least-squares problem, and we know how to solve it (see Section 21.1). The quantity r_n is known as the *residual* and the method which consists in minimizing $\|r_n\|_2$ is known as GMRES, for *generalized minimal residuals*.

Now since $(u_1, \ldots, u_n)$ is a basis of $\mathcal{K}_n(A, b)$ (since $n \leq s$, no breakdown occurs, except for $n = s$), we may write $x_n = U_n y$, so our minimization problem is

$$\text{minimize} \quad \|AU_n y - b\|_2, \quad y \in \mathbb{C}^n.$$

Since by $(*_1)$ of Section 17.4, we have $AU_n = U_{n+1}\widetilde{H}_n$, minimizing $\|AU_n y - b\|_2$ is equivalent to minimizing $\|U_{n+1}\widetilde{H}_n y - b\|_2$ over $\mathbb{C}^m$. Since $U_{n+1}\widetilde{H}_n y$ and b belong to the column space of U_{n+1}, minimizing $\|U_{n+1}\widetilde{H}_n y - b\|_2$ is equivalent to minimizing $\|\widetilde{H}_n y - U_{n+1}^* b\|_2$. However, by construction,

$$U_{n+1}^* b = \|b\|_2 e_1 \in \mathbb{C}^{n+1},$$

so our minimization problem can be stated as

$$\text{minimize} \quad \|\widetilde{H}_n y - \|b\|_2 e_1\|_2, \quad y \in \mathbb{C}^n.$$

The approximate solution of $Ax = b$ is then

$$x_n = U_n y.$$

Starting with $u_1 = b/\left\|b\right\|_2$ and with $n = 1$, the GMRES method runs $n \leq s$ Arnoldi iterations to find U_n and $\widetilde{H}_n$, and then runs a method to solve the least squares problem

$$\text{minimize} \quad \|\widetilde{H}_n y - \|b\|_2 e_1\|_2, \quad y \in \mathbb{C}^n.$$

When $\left\|r_n\right\|_2 = \|\widetilde{H}_n y - \|b\|_2 e_1\|_2$ is considered small enough, we stop and the approximate solution of $Ax = b$ is then

$$x_n = U_n y.$$

There are ways of improving efficiency of the "naive" version of GMRES that we just presented; see Trefethen and Bau [Trefethen and Bau III (1997)] (Lecture 35). We now consider the case where A is a Hermitian (or symmetric) matrix.

17.6 The Hermitian Case; Lanczos Iteration

If A is an $m \times m$ symmetric or Hermitian matrix, then Arnoldi's method is simpler and much more efficient. Indeed, in this case, it is easy to see that the upper Hessenberg matrices H_n are also symmetric (Hermitian respectively), and thus tridiagonal. Also, the eigenvalues of A and H_n are real. It is convenient to write

$$H_n = \begin{pmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \beta_2 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \beta_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{pmatrix}.$$

The recurrence $(*_2)$ of Section 17.4 becomes the three-term recurrence

$$Au_n = \beta_{n-1}u_{n-1} + \alpha_n u_n + \beta_n u_{n+1}. \qquad (*_6)$$

We also have $\alpha_n = u_n^* A U_n$, so Arnoldi's algorithm become the following algorithm known as *Lanczos' algorithm* (or *Lanczos iteration*). The inner loop on j from 1 to n has been eliminated and replaced by a single assignment.

Given an arbitrary nonzero vector $b \in \mathbb{C}^m$, let $u_1 = b/\left\|b\right\|$;

for $n = 1, 2, 3, \ldots$ **do**

$z := Au_n$;
  $\alpha_n := u_n^* z$;
  $z := z - \beta_{n-1} u_{n-1} - \alpha_n u_n$
$\beta_n := \|z\|$;
if $\beta_n = 0$ **quit**
$u_{n+1} = z/\beta_n$

When $\beta_n = 0$, we say that we have a *breakdown* of the Lanczos iteration.

Versions of Proposition 17.2 and Theorem 17.4 apply to Lanczos iteration.

Besides being much more efficient than Arnoldi iteration, Lanczos iteration has the advantage that the *Rayleigh–Ritz method* for finding some of the eigenvalues of A as the eigenvalues of the symmetric (respectively Hermitian) tridiagonal matrix H_n applies, but there are more methods for finding the eigenvalues of symmetric (respectively Hermitian) tridiagonal matrices. Also theorems about error estimates exist. The version of Lanczos iteration given above may run into problems in floating point arithmetic. What happens is that the vectors u_j may lose the property of being orthogonal, so it may be necessary to reorthogonalize them. For more on all this, see Demmel [Demmel (1997)] (Chapter 7, in particular Section 7.2-7.4). The version of GMRES using Lanczos iteration is called MINRES.

We close our brief survey of methods for computing the eigenvalues and the eigenvectors of a matrix with a quick discussion of two methods known as power methods.

17.7 Power Methods

Let A be an $m \times m$ complex or real matrix. There are two power methods, both of which yield one eigenvalue and one eigenvector associated with this vector:

(1) *Power iteration.*
(2) *Inverse (power) iteration.*

Power iteration only works if the matrix A has an eigenvalue λ of largest modulus, which means that if $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of A, then

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_m| \geq 0.$$

In particular, if A is a real matrix, then λ_1 must be real (since otherwise there are two complex conjugate eigenvalues of the same largest modulus).

If the above condition is satisfied, then power iteration yields λ_1 and some eigenvector associated with it. The method is simple enough:

Pick some initial unit vector x^0 and compute the following sequence (x^k), where

$$x^{k+1} = \frac{Ax^k}{\|Ax^k\|}, \quad k \geq 0.$$

We would expect that (x^k) converges to an eigenvector associated with λ_1, but this is not quite correct. The following results are proven in Serre [Serre (2010)] (Section 13.5.1). First assume that $\lambda_1 \neq 0$.

We have

$$\lim_{k \mapsto \infty} \left\|Ax^k\right\| = |\lambda_1|.$$

If A is a complex matrix which has a unique complex eigenvalue λ_1 of largest modulus, then

$$v = \lim_{k \mapsto \infty} \left(\frac{\overline{\lambda_1}}{|\lambda_1|} \right)^k x^k$$

is a unit eigenvector of A associated with λ_1. If λ_1 is real, then

$$v = \lim_{k \mapsto \infty} x^k$$

is a unit eigenvector of A associated with λ_1. Actually some condition on x^0 is needed: x^0 must have a nonzero component in the eigenspace E associated with λ_1 (in any direct sum of $\mathbb{C}^m$ in which E is a summand).

The eigenvalue λ_1 is found as follows. If λ_1 is complex, and if $v_j \neq 0$ is any nonzero coordinate of v, then

$$\lambda_1 = \lim_{k \mapsto \infty} \frac{(Ax^k)_j}{x^k_j}.$$

If λ_1 is real, then we can define the sequence $(\lambda^{(k)})$ by

$$\lambda^{(k+1)} = (x^{k+1})^* A x^{k+1}, \quad k \geq 0,$$

and we have

$$\lambda_1 = \lim_{k \mapsto \infty} \lambda^{(k)}.$$

Indeed, in this case, since $v = \lim_{k \mapsto \infty} x^k$ and v is a unit eigenvector for λ_1, we have

$$\lim_{k \mapsto \infty} \lambda^{(k)} = \lim_{k \mapsto \infty} (x^{k+1})^* A x^{k+1} = v^* A v = \lambda_1 v^* v = \lambda_1.$$

Note that since x^{k+1} is a unit vector, $(x^{k+1})^* A x^{k+1}$ is a Rayleigh ratio.

If A is a Hermitian matrix, then the eigenvalues are real and we can say more about the rate of convergence, which is not great (only linear). For details, see Trefethen and Bau [Trefethen and Bau III (1997)] (Lecture 27).

If $\lambda_1 = 0$, then there is some power $\ell < m$ such that $Ax^\ell = 0$.

The *inverse iteration method* is designed to find an eigenvector associated with an eigenvalue λ of A for which we know a good approximation μ.

Pick some initial unit vector x^0 and compute the following sequences (w^k) and (x^k), where w^{k+1} is the solution of the system

$$(A - \mu I)w^{k+1} = x^k \quad \text{equivalently} \quad w^{k+1} = (A - \mu I)^{-1}x^k, \quad k \geq 0,$$

and

$$x^{k+1} = \frac{w^{k+1}}{\|w^{k+1}\|}, \quad k \geq 0.$$

The following result is proven in Ciarlet [Ciarlet (1989)] (Theorem 6.4.1).

Proposition 17.3. *Let A be an $m \times m$ diagonalizable (complex or real) matrix with eigenvalues $\lambda_1, \ldots, \lambda_m$, and let $\lambda = \lambda_\ell$ be an arbitrary eigenvalue of A (not necessary simple). For any μ such that*

$$\mu \neq \lambda \quad \textit{and} \quad |\mu - \lambda| < |\mu - \lambda_j| \quad \textit{for all } j \neq \ell,$$

if x^0 does not belong to the subspace spanned by the eigenvectors associated with the eigenvalues λ_j with $j \neq \ell$, then

$$\lim_{k \mapsto \infty} \left(\frac{(\lambda - \mu)^k}{|\lambda - \mu|^k} \right) x^k = v,$$

where v is an eigenvector associated with λ. Furthermore, if both λ and μ are real, we have

$$\lim_{k \mapsto \infty} x^k = v \qquad \textit{if } \mu < \lambda,$$
$$\lim_{k \mapsto \infty} (-1)^k x^k = v \qquad \textit{if } \mu > \lambda.$$

Also, if we define the sequence $(\lambda^{(k)})$ by

$$\lambda^{(k+1)} = (x^{k+1})^* A x^{k+1},$$

then

$$\lim_{k \mapsto \infty} \lambda^{(k+1)} = \lambda.$$

The condition of x^0 may seem quite stringent, but in practice, a vector x^0 chosen at random usually satisfies it.

If A is a Hermitian matrix, then we can say more. In particular, the inverse iteration algorithm can be modified to make use of the newly computed $\lambda^{(k+1)}$ instead of μ, and an even faster convergence is achieved. Such a method is called the *Rayleigh quotient iteration*. When it converges (which is for almost all x^0), this method eventually achieves cubic convergence, which is remarkable. Essentially, this means that the number of correct digits is tripled at every iteration. For more details, see Trefethen and Bau [Trefethen and Bau III (1997)] (Lecture 27) and Demmel [Demmel (1997)] (Section 5.3.2).

17.8 Summary

The main concepts and results of this chapter are listed below:

- QR iteration, QR algorithm.
- Upper Hessenberg matrices.
- Householder matrix.
- Unreduced and reduced Hessenberg matrices.
- Deflation.
- Shift.
- Wilkinson shift.
- Double shift.
- Francis shift.
- Implicit shifting.
- Implicit Q-theorem.
- Arnoldi iteration.
- Breakdown of Arnoldi iteration.
- Krylov subspace.
- Rayleigh–Ritz method.
- Ritz values, Arnoldi estimates.
- Residual.
- GMRES
- Lanczos iteration.
- Power iteration.
- Inverse power iteration.
- Rayleigh ratio.

17.9 Problems

Problem 17.1. Prove Theorem 17.2; see Problem 12.7.

Problem 17.2. Prove that if a matrix A is Hermitian (or real symmetric), then any Hessenberg matrix H similar to A is Hermitian tridiagonal (real symmetric tridiagonal).

Problem 17.3. For any matrix (real or complex) A, if $A = QR$ is a QR-decomposition of A using Householder reflections, prove that if A is upper Hessenberg then so is Q.

Problem 17.4. Prove that if A is upper Hessenberg, then the matrices A_k obtained by applying the QR-algorithm are also upper Hessenberg.

Problem 17.5. Prove the *implicit Q theorem*. This theorem says that if A is reduced to upper Hessenberg form as $A = UHU^*$ and if H is unreduced ($h_{i+1i} \neq 0$ for $i = 1, \ldots, n-1$), then the columns of index $2, \ldots, n$ of U are determined by the first column of U up to sign;

Problem 17.6. Read Section 7.5 of Golub and Van Loan [Golub and Van Loan (1996)] and implement their version of the QR-algorithm with shifts.

Problem 17.7. If an Arnoldi iteration has a breakdown at stage n, that is, $h_{n+1} = 0$, then we found the first unreduced block of the Hessenberg matrix H. Prove that the eigenvalues of H_n are eigenvalues of A.

Problem 17.8. Prove Theorem 17.4.

Problem 17.9. Implement GRMES and test it on some linear systems.

Problem 17.10. State and prove versions of Proposition 17.2 and Theorem 17.4 for the Lanczos iteration.

Problem 17.11. Prove the results about the power iteration method stated in Section 17.7.

Problem 17.12. Prove the results about the inverse power iteration method stated in Section 17.7.

Problem 17.13. Implement and test the power iteration method and the inverse power iteration method.

Problem 17.14. Read Lecture 27 in Trefethen and Bau [Trefethen and Bau III (1997)] and implement and test the Rayleigh quotient iteration method.

Chapter 18

Graphs and Graph Laplacians; Basic Facts

In this chapter and the next we present some applications of linear algebra to graph theory. Graphs (undirected and directed) can be defined in terms of various matrices (incidence and adjacency matrices), and various connectivity properties of graphs are captured by properties of these matrices. Another very important matrix is associated with a (undirected) graph: the *graph Laplacian*. The graph Laplacian is symmetric positive definite, and its eigenvalues capture some of the properties of the underlying graph. This is a key fact that is exploited in graph clustering methods, the most powerful being the method of normalized cuts due to Shi and Malik [Shi and Malik (2000)]. This chapter and the next constitute an introduction to algebraic and spectral graph theory. We do not discuss normalized cuts, but we discuss graph drawings. Thorough presentations of algebraic graph theory can be found in Godsil and Royle [Godsil and Royle (2001)] and Chung [Chung (1997)].

We begin with a review of basic notions of graph theory. Even though the graph Laplacian is fundamentally associated with an undirected graph, we review the definition of both directed and undirected graphs. For both directed and undirected graphs, we define the degree matrix D, the incidence matrix B, and the adjacency matrix A. Then we define a *weighted graph*. This is a pair (V, W), where V is a finite set of nodes and W is a $m \times m$ symmetric matrix with nonnegative entries and zero diagonal entries (where $m = |V|$).

For every node $v_i \in V$, the *degree* $d(v_i)$ (or d_i) of v_i is the sum of the weights of the edges adjacent to v_i:

$$d_i = d(v_i) = \sum_{j=1}^{m} w_{i\,j}.$$

The *degree matrix* is the diagonal matrix

$$D = \mathrm{diag}(d_1, \ldots, d_m).$$

The notion of degree is illustrated in Figure 18.1. Then we introduce the

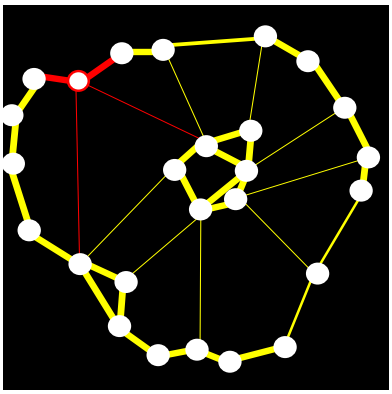

Fig. 18.1 Degree of a node.

(unnormalized) *graph Laplacian* L of a directed graph G in an "old-fashion" way, by showing that for any orientation of a graph G,

$$BB^\top = D - A = L$$

is an invariant. We also define the (unnormalized) *graph Laplacian* L of a weighted graph $G = (V, W)$ as $L = D - W$. We show that the notion of incidence matrix can be generalized to weighted graphs in a simple way. For any graph G^σ obtained by orienting the underlying graph of a weighted graph $G = (V, W)$, there is an incidence matrix B^σ such that

$$B^\sigma (B^\sigma)^\top = D - W = L.$$

We also prove that

$$x^\top L x = \frac{1}{2} \sum_{i,j=1}^{m} w_{ij}(x_i - x_j)^2 \quad \text{for all } x \in \mathbb{R}^m.$$

Consequently, $x^\top L x$ does not depend on the diagonal entries in W, and if $w_{ij} \geq 0$ for all $i, j \in \{1, \ldots, m\}$, then L is positive semidefinite. Then if W consists of nonnegative entries, the eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m$ of L are real and nonnegative, and there is an orthonormal basis of eigenvectors of L. We show that the number of connected components of the graph

$G = (V, W)$ is equal to the dimension of the kernel of L, which is also equal to the dimension of the kernel of the transpose $(B^\sigma)^\top$ of any incidence matrix B^σ obtained by orienting the underlying graph of G.

We also define the normalized graph Laplacians L_{sym} and L_{rw}, given by

$$L_{\mathrm{sym}} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$$
$$L_{\mathrm{rw}} = D^{-1} L = I - D^{-1} W,$$

and prove some simple properties relating the eigenvalues and the eigenvectors of L, L_{sym} and L_{rw}. These normalized graph Laplacians show up when dealing with normalized cuts.

Next, we turn to *graph drawings* (Chapter 19). Graph drawing is a very attractive application of so-called spectral techniques, which is a fancy way of saying that that eigenvalues and eigenvectors of the graph Laplacian are used. Furthermore, it turns out that graph clustering using normalized cuts can be cast as a certain type of graph drawing.

Given an undirected graph $G = (V, E)$, with $|V| = m$, we would like to draw G in $\mathbb{R}^n$ for n (much) smaller than m. The idea is to assign a point $\rho(v_i)$ in $\mathbb{R}^n$ to the vertex $v_i \in V$, for every $v_i \in V$, and to draw a line segment between the points $\rho(v_i)$ and $\rho(v_j)$. Thus, a *graph drawing* is a function $\rho\colon V \to \mathbb{R}^n$.

We define the *matrix of a graph drawing ρ (in $\mathbb{R}^n$)* as a $m \times n$ matrix R whose ith row consists of the row vector $\rho(v_i)$ corresponding to the point representing v_i in $\mathbb{R}^n$. Typically, we want $n < m$; in fact n should be much smaller than m.

Since there are infinitely many graph drawings, it is desirable to have some criterion to decide which graph is better than another. Inspired by a physical model in which the edges are springs, it is natural to consider a representation to be better if it requires the springs to be less extended. We can formalize this by defining the *energy* of a drawing R by

$$\mathcal{E}(R) = \sum_{\{v_i, v_j\} \in E} \|\rho(v_i) - \rho(v_j)\|^2,$$

where $\rho(v_i)$ is the ith row of R and $\|\rho(v_i) - \rho(v_j)\|^2$ is the square of the Euclidean length of the line segment joining $\rho(v_i)$ and $\rho(v_j)$.

Then "good drawings" are drawings that minimize the energy function $\mathcal{E}$. Of course, the trivial representation corresponding to the zero matrix is optimum, so we need to impose extra constraints to rule out the trivial solution.

We can consider the more general situation where the springs are not necessarily identical. This can be modeled by a symmetric weight (or stiffness) matrix $W = (w_{ij})$, with $w_{ij} \geq 0$. In this case, our energy function becomes

$$\mathcal{E}(R) = \sum_{\{v_i, v_j\} \in E} w_{ij} \left\| \rho(v_i) - \rho(v_j) \right\|^2.$$

Following Godsil and Royle [Godsil and Royle (2001)], we prove that

$$\mathcal{E}(R) = \mathrm{tr}(R^\top L R),$$

where

$$L = D - W,$$

is the familiar unnormalized Laplacian matrix associated with W, and where D is the degree matrix associated with W.

It can be shown that there is no loss in generality in assuming that the columns of R are pairwise orthogonal and that they have unit length. Such a matrix satisfies the equation $R^\top R = I$ and the corresponding drawing is called an *orthogonal drawing*. This condition also rules out trivial drawings.

Then we prove the main theorem about graph drawings (Theorem 19.1), which essentially says that the matrix R of the desired graph drawing is constituted by the n eigenvectors of L associated with the smallest nonzero n eigenvalues of L. We give a number examples of graph drawings, many of which are borrowed or adapted from Spielman [Spielman (2012)].

18.1 Directed Graphs, Undirected Graphs, Incidence Matrices, Adjacency Matrices, Weighted Graphs

Definition 18.1. A *directed graph* is a pair $G = (V, E)$, where $V = \{v_1, \ldots, v_m\}$ is a set of *nodes* or *vertices*, and $E \subseteq V \times V$ is a set of ordered pairs of distinct nodes (that is, pairs $(u, v) \in V \times V$ with $u \neq v$), called *edges*. Given any edge $e = (u, v)$, we let $s(e) = u$ be the *source* of e and $t(e) = v$ be the *target* of e.

Remark: Since an edge is a pair (u, v) with $u \neq v$, self-loops are not allowed. Also, there is at most one edge from a node u to a node v. Such graphs are sometimes called *simple graphs*.

An example of a directed graph is shown in Figure 18.2.

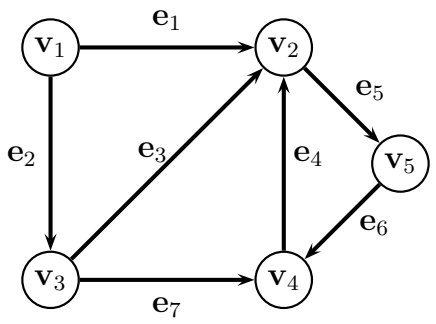

Fig. 18.2 Graph G_1.

Definition 18.2. For every node $v \in V$, the *degree* $d(v)$ of v is the number of edges leaving or entering v:

$$d(v) = |\{u \in V \mid (v,u) \in E \text{ or } (u,v) \in E\}|.$$

We abbreviate $d(v_i)$ as d_i. The *degree matrix*, $D(G)$, is the diagonal matrix

$$D(G) = \mathrm{diag}(d_1, \ldots, d_m).$$

For example, for graph G_1, we have

$$D(G_1) = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}.$$

Unless confusion arises, we write D instead of $D(G)$.

Definition 18.3. Given a directed graph $G = (V, E)$, for any two nodes $u, v \in V$, a *path from u to v* is a sequence of nodes $(v_0, v_1, \ldots, v_k)$ such that $v_0 = u$, $v_k = v$, and (v_i, v_{i+1}) is an edge in E for all i with $0 \leq i \leq k-1$. The integer k is the *length* of the path. A path is *closed* if $u = v$. The graph G is *strongly connected* if for any two distinct nodes $u, v \in V$, there is a path from u to v and there is a path from v to u.

Remark: The terminology *walk* is often used instead of *path*, the word path being reserved to the case where the nodes v_i are all distinct, except that $v_0 = v_k$ when the path is closed.

The binary relation on $V \times V$ defined so that u and v are related iff there is a path from u to v and there is a path from v to u is an equivalence relation whose equivalence classes are called the *strongly connected components* of G.

Definition 18.4. Given a directed graph $G = (V, E)$, with $V = \{v_1, \ldots, v_m\}$, if $E = \{e_1, \ldots, e_n\}$, then the *incidence matrix* $B(G)$ of G is the $m \times n$ matrix whose entries $b_{i\,j}$ are given by

$$b_{i\,j} = \begin{cases} +1 & \text{if } s(e_j) = v_i \\ -1 & \text{if } t(e_j) = v_i \\ 0 & \text{otherwise.} \end{cases}$$

Here is the incidence matrix of the graph G_1:

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

Observe that every column of an incidence matrix contains exactly two nonzero entries, $+1$ and -1. Again, unless confusion arises, we write B instead of $B(G)$.

When a directed graph has m nodes $v_1, \ldots, v_m$ and n edges $e_1, \ldots, e_n$, a vector $x \in \mathbb{R}^m$ can be viewed as a function $x\colon V \to \mathbb{R}$ assigning the value x_i to the node v_i. Under this interpretation, $\mathbb{R}^m$ is viewed as $\mathbb{R}^V$. Similarly, a vector $y \in \mathbb{R}^n$ can be viewed as a function in $\mathbb{R}^E$. This point of view is often useful. For example, the incidence matrix B can be interpreted as a linear map from $\mathbb{R}^E$ to $\mathbb{R}^V$, the *boundary map*, and $B^\top$ can be interpreted as a linear map from $\mathbb{R}^V$ to $\mathbb{R}^E$, the *coboundary map*.

Remark: Some authors adopt the opposite convention of sign in defining the incidence matrix, which means that their incidence matrix is $-B$.

Undirected graphs are obtained from directed graphs by forgetting the orientation of the edges.

Definition 18.5. A *graph* (or *undirected graph*) is a pair $G = (V, E)$, where $V = \{v_1, \ldots, v_m\}$ is a set of *nodes* or *vertices*, and E is a set of two-element subsets of V (that is, subsets $\{u, v\}$, with $u, v \in V$ and $u \neq v$), called *edges*.

Remark: Since an edge is a set $\{u, v\}$, we have $u \neq v$, so self-loops are not allowed. Also, for every set of nodes $\{u, v\}$, there is at most one edge between u and v. As in the case of directed graphs, such graphs are sometimes called *simple graphs*.

An example of a graph is shown in Figure 18.3.

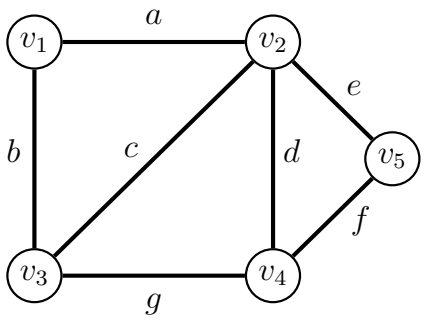

Fig. 18.3 The undirected graph G_2.

Definition 18.6. For every node $v \in V$, the *degree* $d(v)$ of v is the number of edges incident to v:

$$d(v) = |\{u \in V \mid \{u, v\} \in E\}|.$$

The degree matrix $D(G)$ (or simply, D) is defined as in Definition 18.2.

Definition 18.7. Given a (undirected) graph $G = (V, E)$, for any two nodes $u, v \in V$, a *path from u to v* is a sequence of nodes $(v_0, v_1, \ldots, v_k)$ such that $v_0 = u$, $v_k = v$, and $\{v_i, v_{i+1}\}$ is an edge in E for all i with $0 \leq i \leq k-1$. The integer k is the *length* of the path. A path is *closed* if $u = v$. The graph G is *connected* if for any two distinct nodes $u, v \in V$, there is a path from u to v.

Remark: The terminology *walk* or *chain* is often used instead of *path*, the word path being reserved to the case where the nodes v_i are all distinct, except that $v_0 = v_k$ when the path is closed.

The binary relation on $V \times V$ defined so that u and v are related iff there is a path from u to v is an equivalence relation whose equivalence classes are called the *connected components* of G.

The notion of incidence matrix for an undirected graph is not as useful as in the case of directed graphs

Definition 18.8. Given a graph $G = (V, E)$, with $V = \{v_1, \ldots, v_m\}$, if $E = \{e_1, \ldots, e_n\}$, then the *incidence matrix* $B(G)$ of G is the $m \times n$ matrix whose entries $b_{i\,j}$ are given by

$$b_{i\,j} = \begin{cases} +1 & \text{if } e_j = \{v_i, v_k\} \text{ for some } k \\ 0 & \text{otherwise.} \end{cases}$$

Unlike the case of directed graphs, the entries in the incidence matrix of a graph (undirected) are nonnegative. We usually write B instead of $B(G)$.

Definition 18.9. If $G = (V, E)$ is a directed or an undirected graph, given a node $u \in V$, any node $v \in V$ such that there is an edge (u, v) in the directed case or $\{u, v\}$ in the undirected case is called *adjacent to* u, and we often use the notation

$$u \sim v.$$

Observe that the binary relation $\sim$ is symmetric when G is an undirected graph, but in general it is not symmetric when G is a directed graph.

The notion of adjacency matrix is basically the same for directed or undirected graphs.

Definition 18.10. Given a directed or undirected graph $G = (V, E)$, with $V = \{v_1, \ldots, v_m\}$, the *adjacency matrix* $A(G)$ of G is the symmetric $m \times m$ matrix $(a_{i\,j})$ such that

(1) If G is directed, then

$$a_{i\,j} = \begin{cases} 1 & \text{if there is some edge } (v_i, v_j) \in E \text{ or some edge } (v_j, v_i) \in E \\ 0 & \text{otherwise.} \end{cases}$$

(2) Else if G is undirected, then

$$a_{i\,j} = \begin{cases} 1 & \text{if there is some edge } \{v_i, v_j\} \in E \\ 0 & \text{otherwise.} \end{cases}$$

As usual, unless confusion arises, we write A instead of $A(G)$. Here is the adjacency matrix of both graphs G_1 and G_2:

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

If $G = (V, E)$ is an undirected graph, the adjacency matrix A of G can be viewed as a linear map from $\mathbb{R}^V$ to $\mathbb{R}^V$, such that for all $x \in \mathbb{R}^m$, we have

$$(Ax)_i = \sum_{j \sim i} x_j;$$

that is, the value of Ax at v_i is the sum of the values of x at the nodes v_j adjacent to v_i. The adjacency matrix can be viewed as a *diffusion operator*. This observation yields a geometric interpretation of what it means for a vector $x \in \mathbb{R}^m$ to be an eigenvector of A associated with some eigenvalue λ; we must have

$$\lambda x_i = \sum_{j \sim i} x_j, \quad i = 1, \ldots, m,$$

which means that the the sum of the values of x assigned to the nodes v_j adjacent to v_i is equal to λ times the value of x at v_i.

Definition 18.11. Given any undirected graph $G = (V, E)$, an *orientation* of G is a function $\sigma \colon E \to V \times V$ assigning a source and a target to every edge in E, which means that for every edge $\{u, v\} \in E$, either $\sigma(\{u, v\}) = (u, v)$ or $\sigma(\{u, v\}) = (v, u)$. The *oriented graph* G^σ obtained from G by applying the orientation σ is the directed graph $G^\sigma = (V, E^\sigma)$, with $E^\sigma = \sigma(E)$.

The following result shows how the number of connected components of an undirected graph is related to the rank of the incidence matrix of any oriented graph obtained from G.

Proposition 18.1. *Let $G = (V, E)$ be any undirected graph with m vertices, n edges, and c connected components. For any orientation σ of G, if B is the incidence matrix of the oriented graph G^σ, then $c = \dim(\mathrm{Ker}\,(B^\top))$, and B has rank $m - c$. Furthermore, the nullspace of $B^\top$ has a basis consisting of indicator vectors of the connected components of G; that is, vectors $(z_1, \ldots, z_m)$ such that $z_j = 1$ iff v_j is in the ith component K_i of G, and $z_j = 0$ otherwise.*

Proof. (After Godsil and Royle [Godsil and Royle (2001)], Section 8.3). The fact that $\text{rank}(B) = m - c$ will be proved last.

Let us prove that the kernel of $B^\top$ has dimension c. A vector $z \in \mathbb{R}^m$ belongs to the kernel of $B^\top$ iff $B^\top z = 0$ iff $z^\top B = 0$. In view of the definition of B, for every edge $\{v_i, v_j\}$ of G, the column of B corresponding to the oriented edge $\sigma(\{v_i, v_j\})$ has zero entries except for a $+1$ and a -1 in position i and position j or vice-versa, so we have

$$z_i = z_j.$$

An easy induction on the length of the path shows that if there is a path from v_i to v_j in G (unoriented), then $z_i = z_j$. Therefore, z has a constant value on any connected component of G. It follows that every vector $z \in \text{Ker}\,(B^\top)$ can be written uniquely as a linear combination

$$z = \lambda_1 z^1 + \cdots + \lambda_c z^c,$$

where the vector z^i corresponds to the ith connected component K_i of G and is defined such that

$$z^i_j = \begin{cases} 1 & \text{iff } v_j \in K_i \\ 0 & \text{otherwise.} \end{cases}$$

This shows that $\dim(\text{Ker}\,(B^\top)) = c$, and that $\text{Ker}\,(B^\top)$ has a basis consisting of indicator vectors.

Since $B^\top$ is a $n \times m$ matrix, we have

$$m = \dim(\text{Ker}\,(B^\top)) + \text{rank}(B^\top),$$

and since we just proved that $\dim(\text{Ker}\,(B^\top)) = c$, we obtain $\text{rank}(B^\top) = m - c$. Since B and $B^\top$ have the same rank, $\text{rank}(B) = m - c$, as claimed. □

Definition 18.12. Following common practice, we denote by $\mathbf{1}$ the (column) vector (of dimension m) whose components are all equal to 1.

Since every column of B contains a single $+1$ and a single -1, the rows of $B^\top$ sum to zero, which can be expressed as

$$B^\top \mathbf{1} = 0.$$

According to Proposition 18.1, the graph G is connected iff B has rank $m - 1$ iff the nullspace of $B^\top$ is the one-dimensional space spanned by $\mathbf{1}$.

In many applications, the notion of graph needs to be generalized to capture the intuitive idea that two nodes u and v are linked with a degree

of certainty (or strength). Thus, we assign a nonnegative weight $w_{i\,j}$ to an edge $\{v_i, v_j\}$; the smaller $w_{i\,j}$ is, the weaker is the link (or similarity) between v_i and v_j, and the greater $w_{i\,j}$ is, the stronger is the link (or similarity) between v_i and v_j.

Definition 18.13. A *weighted graph* is a pair $G = (V, W)$, where $V = \{v_1, \ldots, v_m\}$ is a set of *nodes* or *vertices*, and W is a symmetric matrix called the *weight matrix*, such that $w_{i\,j} \geq 0$ for all $i, j \in \{1, \ldots, m\}$, and $w_{i\,i} = 0$ for $i = 1, \ldots, m$. We say that a set $\{v_i, v_j\}$ is an edge iff $w_{i\,j} > 0$. The corresponding (undirected) graph (V, E) with $E = \{\{v_i, v_j\} \mid w_{i\,j} > 0\}$, is called the *underlying graph* of G.

Remark: Since $w_{i\,i} = 0$, these graphs have no self-loops. We can think of the matrix W as a generalized adjacency matrix. The case where $w_{i\,j} \in \{0, 1\}$ is equivalent to the notion of a graph as in Definition 18.5.

We can think of the weight $w_{i\,j}$ of an edge $\{v_i, v_j\}$ as a degree of similarity (or affinity) in an image, or a cost in a network. An example of a weighted graph is shown in Figure 18.4. The thickness of an edge corresponds to the magnitude of its weight.

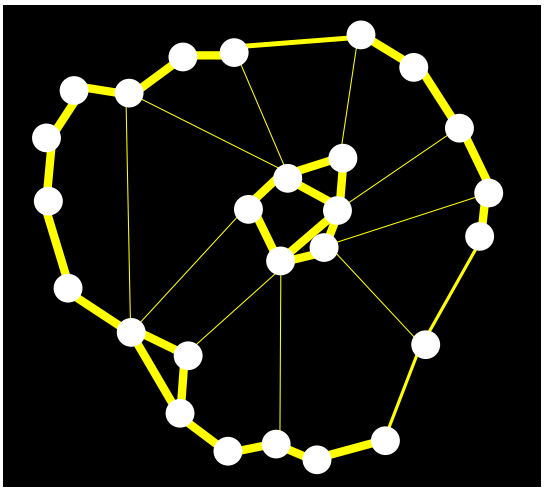

Fig. 18.4 A weighted graph.

Definition 18.14. Given a weighted graph $G = (V, W)$, for every node

$v_i \in V$, the *degree* $d(v_i)$ of v_i is the sum of the weights of the edges adjacent to v_i:

$$d(v_i) = \sum_{j=1}^{m} w_{i\,j}.$$

Note that in the above sum, only nodes v_j such that there is an edge $\{v_i, v_j\}$ have a nonzero contribution. Such nodes are said to be *adjacent* to v_i, and we write $v_i \sim v_j$. The degree matrix $D(G)$ (or simply, D) is defined as before, namely by $D(G) = \mathrm{diag}(d(v_1), \ldots, d(v_m))$.

The weight matrix W can be viewed as a linear map from $\mathbb{R}^V$ to itself. For all $x \in \mathbb{R}^m$, we have

$$(Wx)_i = \sum_{j \sim i} w_{ij} x_j;$$

that is, the value of Wx at v_i is the weighted sum of the values of x at the nodes v_j adjacent to v_i.

Observe that $W\mathbf{1}$ is the (column) vector $(d(v_1), \ldots, d(v_m))$ consisting of the degrees of the nodes of the graph.

We now define the most important concept of this chapter: the Laplacian matrix of a graph. Actually, as we will see, it comes in several flavors.

18.2 Laplacian Matrices of Graphs

Let us begin with directed graphs, although as we will see, graph Laplacians are fundamentally associated with undirected graph. The key proposition below shows how given an undirected graph G, for any orientation σ of G, $B^\sigma (B^\sigma)^\top$ relates to the adjacency matrix A (where B^σ is the incidence matrix of the directed graph G^σ). We reproduce the proof in Gallier [Gallier (2011a)] (see also Godsil and Royle [Godsil and Royle (2001)]).

Proposition 18.2. *Given any undirected graph G, for any orientation σ of G, if B^σ is the incidence matrix of the directed graph G^σ, A is the adjacency matrix of G^σ, and D is the degree matrix such that $D_{i\,i} = d(v_i)$, then*

$$B^\sigma (B^\sigma)^\top = D - A.$$

Consequently, $L = B^\sigma (B^\sigma)^\top$ is independent of the orientation σ of G, and $D - A$ is symmetric and positive semidefinite; that is, the eigenvalues of $D - A$ are real and nonnegative.

Proof. The entry $B^\sigma(B^\sigma)^\top_{ij}$ is the inner product of the ith row b_i^σ, and the jth row b_j^σ of B^σ. If $i = j$, then as

$$b^\sigma_{i\,k} = \begin{cases} +1 & \text{if } s(e_k) = v_i \\ -1 & \text{if } t(e_k) = v_i \\ 0 & \text{otherwise} \end{cases}$$

we see that $b_i^\sigma \cdot b_i^\sigma = d(v_i)$. If $i \neq j$, then $b_i^\sigma \cdot b_j^\sigma \neq 0$ iff there is some edge e_k with $s(e_k) = v_i$ and $t(e_k) = v_j$ or vice-versa (which are mutually exclusive cases, since G^σ arises by orienting an undirected graph), in which case, $b_i^\sigma \cdot b_j^\sigma = -1$. Therefore,

$$B^\sigma(B^\sigma)^\top = D - A,$$

as claimed.

For every $x \in \mathbb{R}^m$, we have

$$x^\top Lx = x^\top B^\sigma(B^\sigma)^\top x = ((B^\sigma)^\top x)^\top (B^\sigma)^\top x = \left\|(B^\sigma)^\top x\right\|_2^2 \geq 0,$$

since the Euclidean norm $\| \ \|_2$ is positive (definite). Therefore, $L = B^\sigma(B^\sigma)^\top$ is positive semidefinite. It is well-known that a real symmetric matrix is positive semidefinite iff its eigenvalues are nonnegative. □

Definition 18.15. The matrix $L = B^\sigma(B^\sigma)^\top = D - A$ is called the *(unnormalized) graph Laplacian* of the graph G^σ. The *(unnormalized) graph Laplacian* of an undirected graph $G = (V, E)$ is defined by

$$L = D - A.$$

For example, the graph Laplacian of graph G_1 is

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

Observe that each row of L sums to zero (because $(B^\sigma)^\top \mathbf{1} = 0$). Consequently, the vector $\mathbf{1}$ is in the nullspace of L.

Remarks:

(1) With the unoriented version of the incidence matrix (see Definition 18.8), it can be shown that

$$BB^\top = D + A.$$

(2) As pointed out by Evangelos Chatzipantazis, Proposition 18.2 in which the incidence matrix B^σ is replaced by the incidence matrix B of any *arbitrary* directed graph G does not hold. The problem is that such graphs may have both edges (v_i, v_j) and (v_j, v_i) between two distinct nodes v_i and v_j, and as a consequence, the inner product $b_i \cdot b_j = -2$ instead of -1. A simple counterexample is given by the directed graph with three vertices and four edges whose incidence matrix is given by

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

We have

$$BB^\top = \begin{pmatrix} 3 & -2 & -1 \\ -2 & 3 & -1 \\ -1 & -1 & 2 \end{pmatrix} \neq \begin{pmatrix} 3\,0\,0 \\ 0\,3\,0 \\ 0\,0\,2 \end{pmatrix} - \begin{pmatrix} 0\,1\,1 \\ 1\,0\,1 \\ 1\,1\,0 \end{pmatrix} = D - A.$$

The natural generalization of the notion of graph Laplacian to weighted graphs is this:

Definition 18.16. Given any weighted graph $G = (V, W)$ with $V = \{v_1, \ldots, v_m\}$, the *(unnormalized) graph Laplacian* $L(G)$ *of* G is defined by

$$L(G) = D(G) - W,$$

where $D(G) = \mathrm{diag}(d_1, \ldots, d_m)$ is the degree matrix of G (a diagonal matrix), with

$$d_i = \sum_{j=1}^{m} w_{i\,j}.$$

As usual, unless confusion arises, we write D instead of $D(G)$ and L instead of $L(G)$.

The graph Laplacian can be interpreted as a linear map from $\mathbb{R}^V$ to itself. For all $x \in \mathbb{R}^V$, we have

$$(Lx)_i = \sum_{j \sim i} w_{ij}(x_i - x_j).$$

It is clear from the equation $L = D - W$ that each row of L sums to 0, so the vector $\mathbf{1}$ is the nullspace of L, but it is less obvious that L is positive semidefinite. One way to prove it is to generalize slightly the notion of incidence matrix.

Definition 18.17. Given a weighted graph $G = (V, W)$, with $V = \{v_1, \ldots, v_m\}$, if $\{e_1, \ldots, e_n\}$ are the edges of the underlying graph of G

(recall that $\{v_i, v_j\}$ is an edge of this graph iff $w_{ij} > 0$), for any oriented graph G^σ obtained by giving an orientation to the underlying graph of G, the *incidence matrix* B^σ of G^σ is the $m \times n$ matrix whose entries $b_{i\,j}$ are given by

$$b_{i\,j} = \begin{cases} +\sqrt{w_{ij}} & \text{if } s(e_j) = v_i \\ -\sqrt{w_{ij}} & \text{if } t(e_j) = v_i \\ 0 & \text{otherwise.} \end{cases}$$

For example, given the weight matrix

$$W = \begin{pmatrix} 0 & 3 & 6 & 3 \\ 3 & 0 & 0 & 3 \\ 6 & 0 & 0 & 3 \\ 3 & 3 & 3 & 0 \end{pmatrix},$$

the incidence matrix B corresponding to the orientation of the underlying graph of W where an edge (i, j) is oriented positively iff $i < j$ is

$$B = \begin{pmatrix} 1.7321 & 2.4495 & 1.7321 & 0 & 0 \\ -1.7321 & 0 & 0 & 1.7321 & 0 \\ 0 & -2.4495 & 0 & 0 & 1.7321 \\ 0 & 0 & -1.7321 & -1.7321 & -1.7321 \end{pmatrix}.$$

The reader should verify that $BB^\top = D - W$. This is true in general, see Proposition 18.3.

It is easy to see that Proposition 18.1 applies to the underlying graph of G. For any oriented graph G^σ obtained from the underlying graph of G, the rank of the incidence matrix B^σ is equal to $m - c$, where c is the number of connected components of the underlying graph of G, and we have $(B^\sigma)^\top \mathbf{1} = 0$. We also have the following version of Proposition 18.2 whose proof is immediately adapted.

Proposition 18.3. *Given any weighted graph $G = (V, W)$ with $V = \{v_1, \ldots, v_m\}$, if B^σ is the incidence matrix of any oriented graph G^σ obtained from the underlying graph of G and D is the degree matrix of G, then*

$$B^\sigma (B^\sigma)^\top = D - W = L.$$

Consequently, $B^\sigma (B^\sigma)^\top$ is independent of the orientation of the underlying graph of G and $L = D - W$ is symmetric and positive semidefinite; that is, the eigenvalues of $L = D - W$ are real and nonnegative.

Another way to prove that L is positive semidefinite is to evaluate the quadratic form $x^\top Lx$.

Proposition 18.4. *For any $m \times m$ symmetric matrix $W = (w_{ij})$, if we let $L = D - W$ where D is the degree matrix associated with W (that is, $d_i = \sum_{j=1}^m w_{ij}$), then we have*

$$x^\top Lx = \frac{1}{2}\sum_{i,j=1}^{m} w_{ij}(x_i - x_j)^2 \quad \text{for all } x \in \mathbb{R}^m.$$

Consequently, $x^\top Lx$ does not depend on the diagonal entries in W, and if $w_{ij} \geq 0$ for all $i, j \in \{1, \ldots, m\}$, then L is positive semidefinite.

Proof. We have

$$\begin{aligned} x^\top Lx &= x^\top Dx - x^\top Wx \\ &= \sum_{i=1}^{m} d_i x_i^2 - \sum_{i,j=1}^{m} w_{ij} x_i x_j \\ &= \frac{1}{2}\left(\sum_{i=1}^{m} d_i x_i^2 - 2\sum_{i,j=1}^{m} w_{ij} x_i x_j + \sum_{i=1}^{m} d_i x_i^2\right) \\ &= \frac{1}{2}\sum_{i,j=1}^{m} w_{ij}(x_i - x_j)^2. \end{aligned}$$

Obviously, the quantity on the right-hand side does not depend on the diagonal entries in W, and if $w_{ij} \geq 0$ for all i, j, then this quantity is nonnegative. □

Proposition 18.4 immediately implies the following facts: For any weighted graph $G = (V, W)$,

(1) The eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m$ of L are real and nonnegative, and there is an orthonormal basis of eigenvectors of L.
(2) The smallest eigenvalue λ_1 of L is equal to 0, and $\mathbf{1}$ is a corresponding eigenvector.

It turns out that the dimension of the nullspace of L (the eigenspace of 0) is equal to the number of connected components of the underlying graph of G.

Proposition 18.5. *Let $G = (V, W)$ be a weighted graph. The number c of connected components $K_1, \ldots, K_c$ of the underlying graph of G is equal*

to the dimension of the nullspace of L, which is equal to the multiplicity of the eigenvalue 0. *Furthermore, the nullspace of L has a basis consisting of indicator vectors of the connected components of G, that is, vectors $(f_1, \ldots, f_m)$ such that $f_j = 1$ iff $v_j \in K_i$ and $f_j = 0$ otherwise.*

Proof. Since $L = BB^\top$ for the incidence matrix B associated with any oriented graph obtained from G, and since L and $B^\top$ have the same nullspace, by Proposition 18.1, the dimension of the nullspace of L is equal to the number c of connected components of G and the indicator vectors of the connected components of G form a basis of $\mathrm{Ker}\,(L)$. □

Proposition 18.5 implies that if the underlying graph of G is connected, then the second eigenvalue λ_2 of L is strictly positive.

Remarkably, the eigenvalue λ_2 contains a lot of information about the graph G (assuming that $G = (V, E)$ is an undirected graph). This was first discovered by Fiedler in 1973, and for this reason, λ_2 is often referred to as the *Fiedler number*. For more on the properties of the Fiedler number, see Godsil and Royle [Godsil and Royle (2001)] (Chapter 13) and Chung [Chung (1997)]. More generally, the spectrum $(0, \lambda_2, \ldots, \lambda_m)$ of L contains a lot of information about the combinatorial structure of the graph G. Leverage of this information is the object of *spectral graph theory*.

18.3 Normalized Laplacian Matrices of Graphs

It turns out that normalized variants of the graph Laplacian are needed, especially in applications to graph clustering. These variants make sense only if G has no isolated vertices.

Definition 18.18. Given a weighted graph $G = (V, W)$, a vertex $u \in V$ is *isolated* if it is not incident to any other vertex. This means that every row of W contains some strictly positive entry.

If G has no isolated vertices, then the degree matrix D contains positive entries, so it is invertible and $D^{-1/2}$ makes sense; namely

$$D^{-1/2} = \mathrm{diag}(d_1^{-1/2}, \ldots, d_m^{-1/2}),$$

and similarly for any real exponent α.

Definition 18.19. Given any weighted directed graph $G = (V, W)$ with no isolated vertex and with $V = \{v_1, \ldots, v_m\}$, the *(normalized) graph Lapla-*

cians L_{sym} *and* L_{rw} *of* G are defined by

$$L_{\mathrm{sym}} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$$
$$L_{\mathrm{rw}} = D^{-1} L = I - D^{-1} W.$$

Observe that the Laplacian $L_{\mathrm{sym}} = D^{-1/2} L D^{-1/2}$ is a symmetric matrix (because L and $D^{-1/2}$ are symmetric) and that

$$L_{\mathrm{rw}} = D^{-1/2} L_{\mathrm{sym}} D^{1/2}.$$

The reason for the notation L_{rw} is that this matrix is closely related to a random walk on the graph G.

Example 18.1. As an example, the matrices L_{sym} and L_{rw} associated with the graph G_1 are

$$L_{\mathrm{sym}} = \begin{pmatrix} 1.0000 & -0.3536 & -0.4082 & 0 & 0 \\ -0.3536 & 1.0000 & -0.2887 & -0.2887 & -0.3536 \\ -0.4082 & -0.2887 & 1.0000 & -0.3333 & 0 \\ 0 & -0.2887 & -0.3333 & 1.0000 & -0.4082 \\ 0 & -0.3536 & 0 & -0.4082 & 1.0000 \end{pmatrix}$$

and

$$L_{\mathrm{rw}} = \begin{pmatrix} 1.0000 & -0.5000 & -0.5000 & 0 & 0 \\ -0.2500 & 1.0000 & -0.2500 & -0.2500 & -0.2500 \\ -0.3333 & -0.3333 & 1.0000 & -0.3333 & 0 \\ 0 & -0.3333 & -0.3333 & 1.0000 & -0.3333 \\ 0 & -0.5000 & 0 & -0.5000 & 1.0000 \end{pmatrix}.$$

Since the unnormalized Laplacian L can be written as $L = BB^\top$, where B is the incidence matrix of any oriented graph obtained from the underlying graph of $G = (V, W)$, if we let

$$B_{\mathrm{sym}} = D^{-1/2} B,$$

we get

$$L_{\mathrm{sym}} = B_{\mathrm{sym}} B_{\mathrm{sym}}^\top.$$

In particular, for any singular decomposition $B_{\mathrm{sym}} = U \Sigma V^\top$ of B_{sym} (with U an $m \times m$ orthogonal matrix, Σ a "diagonal" $m \times n$ matrix of singular values, and V an $n \times n$ orthogonal matrix), the eigenvalues of L_{sym} are the squares of the top m singular values of B_{sym}, and the vectors in U are orthonormal eigenvectors of L_{sym} with respect to these eigenvalues (the squares of the top m diagonal entries of Σ). Computing the SVD of B_{sym}

generally yields more accurate results than diagonalizing L_{sym}, especially when L_{sym} has eigenvalues with high multiplicity.

There are simple relationships between the eigenvalues and the eigenvectors of L_{sym}, and L_{rw}. There is also a simple relationship with the generalized eigenvalue problem $Lx = \lambda Dx$.

Proposition 18.6. *Let $G = (V, W)$ be a weighted graph without isolated vertices. The graph Laplacians, L, L_{sym}, and L_{rw} satisfy the following properties:*

(1) The matrix L_{sym} is symmetric and positive semidefinite. In fact,

$$x^\top L_{\mathrm{sym}} x = \frac{1}{2}\sum_{i,j=1}^{m} w_{i\,j}\left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}}\right)^2 \quad \text{for all } x \in \mathbb{R}^m.$$

(2) The normalized graph Laplacians L_{sym} and L_{rw} have the same spectrum $(0 = \nu_1 \leq \nu_2 \leq \ldots \leq \nu_m)$, and a vector $u \neq 0$ is an eigenvector of L_{rw} for λ iff $D^{1/2}u$ is an eigenvector of L_{sym} for λ.

(3) The graph Laplacians L and L_{sym} are symmetric and positive semidefinite.

(4) A vector $u \neq 0$ is a solution of the generalized eigenvalue problem $Lu = \lambda Du$ iff $D^{1/2}u$ is an eigenvector of L_{sym} for the eigenvalue λ iff u is an eigenvector of L_{rw} for the eigenvalue λ.

(5) The graph Laplacians, L and L_{rw} have the same nullspace. For any vector u, we have $u \in \mathrm{Ker}\,(L)$ iff $D^{1/2}u \in \mathrm{Ker}\,(L_{\mathrm{sym}})$.

(6) The vector $\mathbf{1}$ is in the nullspace of L_{rw}, and $D^{1/2}\mathbf{1}$ is in the nullspace of L_{sym}.

(7) For every eigenvalue ν_i of the normalized graph Laplacian L_{sym}, we have $0 \leq \nu_i \leq 2$. Furthermore, $\nu_m = 2$ iff the underlying graph of G contains a nontrivial connected bipartite component.

(8) If $m \geq 2$ and if the underlying graph of G is not a complete graph,[1] *then $\nu_2 \leq 1$. Furthermore the underlying graph of G is a complete graph iff $\nu_2 = \frac{m}{m-1}$.*

(9) If $m \geq 2$ and if the underlying graph of G is connected, then $\nu_2 > 0$.

(10) If $m \geq 2$ and if the underlying graph of G has no isolated vertices, then $\nu_m \geq \frac{m}{m-1}$.

Proof. (1) We have $L_{\mathrm{sym}} = D^{-1/2}LD^{-1/2}$, and $D^{-1/2}$ is a symmetric invertible matrix (since it is an invertible diagonal matrix). It is a

[1] Recall that an undirected graph is complete if for any two distinct nodes u, v, there is an edge $\{u, v\}$.

well-known fact of linear algebra that if B is an invertible matrix, then a matrix S is symmetric, positive semidefinite iff $BSB^\top$ is symmetric, positive semidefinite. Since L is symmetric, positive semidefinite, so is $L_{\mathrm{sym}} = D^{-1/2}LD^{-1/2}$. The formula

$$x^\top L_{\mathrm{sym}} x = \frac{1}{2}\sum_{i,j=1}^{m} w_{i\,j}\left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}}\right)^2 \quad \text{for all } x \in \mathbb{R}^m$$

follows immediately from Proposition 18.4 by replacing x by $D^{-1/2}x$, and also shows that L_{sym} is positive semidefinite.

(2) Since

$$L_{\mathrm{rw}} = D^{-1/2}L_{\mathrm{sym}}D^{1/2},$$

the matrices L_{sym} and L_{rw} are similar, which implies that they have the same spectrum. In fact, since $D^{1/2}$ is invertible,

$$L_{\mathrm{rw}}u = D^{-1}Lu = \lambda u$$

iff

$$D^{-1/2}Lu = \lambda D^{1/2}u$$

iff

$$D^{-1/2}LD^{-1/2}D^{1/2}u = L_{\mathrm{sym}}D^{1/2}u = \lambda D^{1/2}u,$$

which shows that a vector $u \neq 0$ is an eigenvector of L_{rw} for λ iff $D^{1/2}u$ is an eigenvector of L_{sym} for λ.

(3) We already know that L and L_{sym} are positive semidefinite.

(4) Since $D^{-1/2}$ is invertible, we have

$$Lu = \lambda Du$$

iff

$$D^{-1/2}Lu = \lambda D^{1/2}u$$

iff

$$D^{-1/2}LD^{-1/2}D^{1/2}u = L_{\mathrm{sym}}D^{1/2}u = \lambda D^{1/2}u,$$

which shows that a vector $u \neq 0$ is a solution of the generalized eigenvalue problem $Lu = \lambda Du$ iff $D^{1/2}u$ is an eigenvector of L_{sym} for the eigenvalue λ. The second part of the statement follows from (2).

(5) Since D^{-1} is invertible, we have $Lu = 0$ iff $D^{-1}Lu = L_{\mathrm{rw}}u = 0$. Similarly, since $D^{-1/2}$ is invertible, we have $Lu = 0$ iff $D^{-1/2}LD^{-1/2}D^{1/2}u = 0$ iff $D^{1/2}u \in \mathrm{Ker}\,(L_{\mathrm{sym}})$.

(6) Since $L\mathbf{1} = 0$, we get $L_{\mathrm{rw}}\mathbf{1} = D^{-1}L\mathbf{1} = 0$. That $D^{1/2}\mathbf{1}$ is in the nullspace of L_{sym} follows from (2). Properties (7)–(10) are proven in Chung [Chung (1997)] (Chapter 1). □

The eigenvalues the matrices L_{sym} and L_{rw} from Example 18.1 are

$$0,\ 7257,\ 1.1667,\ 1.5,\ 1.6076.$$

On the other hand, the eigenvalues of the unnormalized Laplacian for G_1 are

$$0,\ 1.5858,\ 3,\ 4.4142,\ 5.$$

Remark: Observe that although the matrices L_{sym} and L_{rw} have the same spectrum, the matrix L_{rw} is generally not symmetric, whereas L_{sym} is symmetric.

A version of Proposition 18.5 also holds for the graph Laplacians L_{sym} and L_{rw}. This follows easily from the fact that Proposition 18.1 applies to the underlying graph of a weighted graph. The proof is left as an exercise.

Proposition 18.7. *Let $G = (V, W)$ be a weighted graph. The number c of connected components $K_1, \ldots, K_c$ of the underlying graph of G is equal to the dimension of the nullspace of both L_{sym} and L_{rw}, which is equal to the multiplicity of the eigenvalue 0. Furthermore, the nullspace of L_{rw} has a basis consisting of indicator vectors of the connected components of G, that is, vectors $(f_1, \ldots, f_m)$ such that $f_j = 1$ iff $v_j \in K_i$ and $f_j = 0$ otherwise. For L_{sym}, a basis of the nullpace is obtained by multiplying the above basis of the nullspace of L_{rw} by $D^{1/2}$.*

A particularly interesting application of graph Laplacians is graph clustering.

18.4 Graph Clustering Using Normalized Cuts

In order to explain this problem we need some definitions.

Definition 18.20. Given any subset of nodes $A \subseteq V$, we define the *volume* $\mathrm{vol}(A)$ of A as the sum of the weights of all edges adjacent to nodes in A:

$$\mathrm{vol}(A) = \sum_{v_i \in A} \sum_{j=1}^{m} w_{i\,j}.$$

Given any two subsets $A, B \subseteq V$ (not necessarily distinct), we define $\mathrm{links}(A, B)$ by

$$\mathrm{links}(A, B) = \sum_{v_i \in A, v_j \in B} w_{i\,j}.$$

The quantity $\text{links}(A, \overline{A}) = \text{links}(\overline{A}, A)$ (where $\overline{A} = V - A$ denotes the complement of A in V) measures how many links escape from A (and $\overline{A}$). We define the *cut* of A as

$$\text{cut}(A) = \text{links}(A, \overline{A}).$$

The notion of volume is illustrated in Figure 18.5 and the notions of cut is illustrated in Figure 18.6.

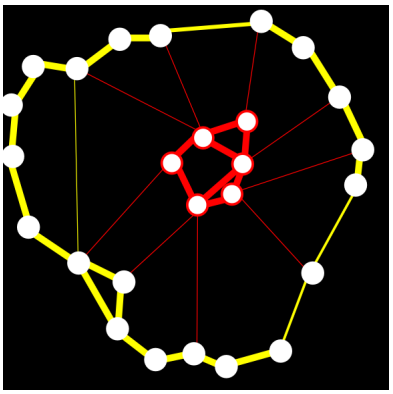

Fig. 18.5 Volume of a set of nodes.

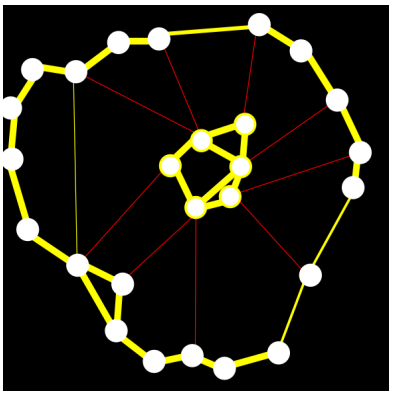

Fig. 18.6 A cut involving the set of nodes in the center and the nodes on the perimeter.

The above concepts play a crucial role in the theory of normalized cuts.

This beautiful and deeply original method first published in Shi and Malik [Shi and Malik (2000)], has now come to be a "textbook chapter" of computer vision and machine learning. It was invented by Jianbo Shi and Jitendra Malik and was the main topic of Shi's dissertation. This method was extended to $K \geq 3$ clusters by Stella Yu in her dissertation [Yu (2003)] and is also the subject of Yu and Shi [Yu and Shi (2003)].

Given a set of data, the goal of clustering is to partition the data into different groups according to their similarities. When the data is given in terms of a similarity graph G, where the weight w_{ij} between two nodes v_i and v_j is a measure of similarity of v_i and v_j, the problem can be stated as follows: Find a partition $(A_1, \ldots, A_K)$ of the set of nodes V into different groups such that the edges between different groups have very low weight (which indicates that the points in different clusters are dissimilar), and the edges within a group have high weight (which indicates that points within the same cluster are similar).

The above graph clustering problem can be formalized as an optimization problem, using the notion of cut mentioned earlier. If we want to partition V into K clusters, we can do so by finding a partition $(A_1, \ldots, A_K)$ that minimizes the quantity

$$\mathrm{cut}(A_1, \ldots, A_K) = \frac{1}{2}\sum_{i=1}^{K} \mathrm{cut}(A_i) = \frac{1}{2}\sum_{i=1}^{K} \mathrm{links}(A_i, \overline{A}_i).$$

For $K = 2$, the mincut problem is a classical problem that can be solved efficiently, but in practice, it does not yield satisfactory partitions. Indeed, in many cases, the mincut solution separates one vertex from the rest of the graph. What we need is to design our cost function in such a way that it keeps the subsets A_i "reasonably large" (reasonably balanced).

An example of a weighted graph and a partition of its nodes into two clusters is shown in Figure 18.7.

A way to get around this problem is to normalize the cuts by dividing by some measure of each subset A_i. A solution using the volume $\mathrm{vol}(A_i)$ of A_i (for $K = 2$) was proposed and investigated in a seminal paper of Shi and Malik [Shi and Malik (2000)]. Subsequently, Yu (in her dissertation [Yu (2003)]) and Yu and Shi [Yu and Shi (2003)] extended the method to $K > 2$ clusters. The idea is to minimize the cost function

$$\mathrm{Ncut}(A_1, \ldots, A_K) = \sum_{i=1}^{K} \frac{\mathrm{links}(A_i, \overline{A_i})}{\mathrm{vol}(A_i)} = \sum_{i=1}^{K} \frac{\mathrm{cut}(A_i)}{\mathrm{vol}(A_i)}.$$

The next step is to express our optimization problem in matrix form, and this can be done in terms of Rayleigh ratios involving the graph Laplacian

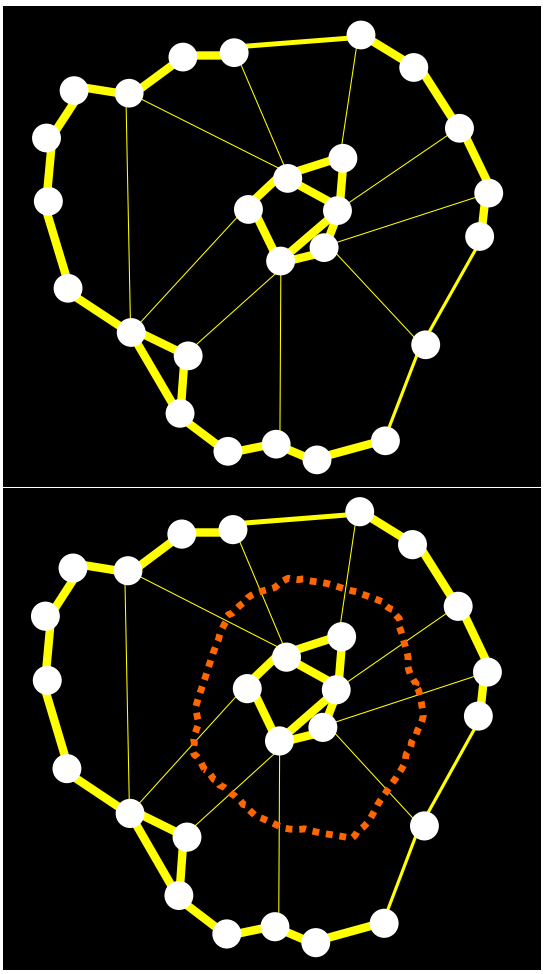

Fig. 18.7 A weighted graph and its partition into two clusters.

in the numerators. This theory is very beautiful, but we do not have the space to present it here. The interested reader is referred to Gallier [Gallier (2019)].

18.5 Summary

The main concepts and results of this chapter are listed below:

- Directed graphs, undirected graphs.
- Incidence matrices, adjacency matrices.
- Weighted graphs.
- Degree matrix.
- Graph Laplacian (unnormalized).
- Normalized graph Laplacian.
- Spectral graph theory.
- Graph clustering using normalized cuts.

18.6 Problems

Problem 18.1. Find the unnormalized Laplacian of the graph representing a triangle and of the graph representing a square.

Problem 18.2. Consider the complete graph K_m on $m \geq 2$ nodes.

(1) Prove that the normalized Laplacian L_{sym} of K is

$$L_{\mathrm{sym}} = \begin{pmatrix} 1 & -1/(m-1) & \dots & -1/(m-1) & -1/(m-1) \\ -1/(m-1) & 1 & \dots & -1/(m-1) & -1/(m-1) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -1/(m-1) & -1/(m-1) & \dots & 1 & -1/(m-1) \\ -1/(m-1) & -1/(m-1) & \dots & -1/(m-1) & 1 \end{pmatrix}.$$

(2) Prove that the characteristic polynomial of L_{sym} is

$$\begin{vmatrix} \lambda - 1 & 1/(m-1) & \dots & 1/(m-1) & 1/(m-1) \\ 1/(m-1) & \lambda - 1 & \dots & 1/(m-1) & 1/(m-1) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 1/(m-1) & 1/(m-1) & \dots & \lambda - 1 & 1/(m-1) \\ 1/(m-1) & 1/(m-1) & \dots & 1/(m-1) & \lambda - 1 \end{vmatrix} = \lambda \left(\lambda - \frac{m}{m-1} \right)^{m-1}.$$

Hint. First subtract the second column from the first, factor $\lambda - m/(m-1)$, and then add the first row to the second. Repeat this process. You will end up with the determinant

$$\begin{vmatrix} \lambda - 1/(m-1) & 1 \\ 1/(m-1) & \lambda - 1 \end{vmatrix}.$$

Problem 18.3. Consider the complete bipartite graph $K_{m,n}$ on $m+n \geq 3$ nodes, with edges between each of the first $m \geq 1$ nodes to each of the last $n \geq 1$ nodes. Prove that the eigenvalues of the normalized Laplacian L_{sym} of $K_{m,n}$ are 0, 1 with multiplicity $m+n-2$, and 2.

Problem 18.4. Let G be a graph with a set of nodes V with $m \geq 2$ elements, without isolated nodes, and let $L_{\mathrm{sym}} = D^{-1/2}LD^{-1/2}$ be its normalized Laplacian (with L its unnormalized Laplacian).

(1) For any $y \in \mathbb{R}^V$, consider the Rayleigh ratio

$$R = \frac{y^\top L_{\mathrm{sym}}\, y}{y^\top y}.$$

Prove that if $x = D^{-1/2}y$, then

$$R = \frac{x^\top Lx}{(D^{1/2}x)^\top D^{1/2}x} = \frac{\displaystyle\sum_{u\sim v}(x(u) - x(v))^2}{\displaystyle\sum_v d_v x(v)^2}.$$

(2) Prove that the second eigenvalue ν_2 of L_{sym} is given by

$$\nu_2 = \min_{\mathbf{1}^\top Dx=0, x\neq 0} \frac{\displaystyle\sum_{u\sim v}(x(u) - x(v))^2}{\displaystyle\sum_v d_v x(v)^2}.$$

(3) Prove that the largest eigenvalue ν_m of L_{sym} is given by

$$\nu_m = \max_{x\neq 0} \frac{\displaystyle\sum_{u\sim v}(x(u) - x(v))^2}{\displaystyle\sum_v d_v x(v)^2}.$$

Problem 18.5. Let G be a graph with a set of nodes V with $m \geq 2$ elements, without isolated nodes. If $0 = \nu_1 \leq \nu_1 \leq \ldots \leq \nu_m$ are the eigenvalues of L_{sym}, prove the following properties:

(1) We have $\nu_1 + \nu_2 + \cdots + \nu_m = m$.
(2) We have $\nu_2 \leq m/(m-1)$, with equality holding iff $G = K_m$, the complete graph on m nodes.
(3) We have $\nu_m \geq m/(m-1)$.
(4) If G is not a complete graph, then $\nu_2 \leq 1$
Hint. If a and b are nonadjacent nodes, consider the function x given by

$$x(v) = \begin{cases} d_b & \text{if } v = a \\ -d_a & \text{if } v = b \\ 0 & \text{if } v \neq a, b, \end{cases}$$

and use Problem 18.4(2).

(5) Prove that $\nu_m \leq 2$. Prove that $\nu_m = 2$ iff the underlying graph of G contains a nontrivial connected bipartite component.
Hint. Use Problem 18.4(3).
(6) Prove that if G is connected, then $\nu_2 > 0$.

Problem 18.6. Let G be a graph with a set of nodes V with $m \geq 2$ elements, without isolated nodes. Let $\mathrm{vol}(G) = \sum_{v \in V} d_v$ and let

$$\overline{x} = \frac{\sum_v d_v x(v)}{\mathrm{vol}(G)}.$$

Prove that

$$\nu_2 = \min_{x \neq 0} \frac{\sum\limits_{u \sim v} (x(u) - x(v))^2}{\sum\limits_v d_v (x(v) - \overline{x})^2}.$$

Problem 18.7. Let G be a connected bipartite graph. Prove that if ν is an eigenvalue of L_{sym}, then $2 - \nu$ is also an eigenvalue of L_{sym}.

Problem 18.8. Prove Proposition 18.7.

Chapter 19

Spectral Graph Drawing

19.1 Graph Drawing and Energy Minimization

Let $G = (V, E)$ be some undirected graph. It is often desirable to draw a graph, usually in the plane but possibly in 3D, and it turns out that the graph Laplacian can be used to design surprisingly good methods. Say $|V| = m$. The idea is to assign a point $\rho(v_i)$ in $\mathbb{R}^n$ to the vertex $v_i \in V$, for every $v_i \in V$, and to draw a line segment between the points $\rho(v_i)$ and $\rho(v_j)$ iff there is an edge $\{v_i, v_j\}$.

Definition 19.1. Let $G = (V, E)$ be some undirected graph with m vertices. A *graph drawing* is a function $\rho\colon V \to \mathbb{R}^n$, for some $n \geq 1$. The *matrix of a graph drawing ρ (in $\mathbb{R}^n$)* is a $m \times n$ matrix R whose ith row consists of the row vector $\rho(v_i)$ corresponding to the point representing v_i in $\mathbb{R}^n$.

For a graph drawing to be useful we want $n \leq m$; in fact n should be much smaller than m, typically $n = 2$ or $n = 3$.

Definition 19.2. A graph drawing is *balanced* iff the sum of the entries of every column of the matrix of the graph drawing is zero, that is,

$$\mathbf{1}^\top R = 0.$$

If a graph drawing is not balanced, it can be made balanced by a suitable translation. We may also assume that the columns of R are linearly independent, since any basis of the column space also determines the drawing. Thus, from now on, we may assume that $n \leq m$.

Remark: A graph drawing $\rho\colon V \to \mathbb{R}^n$ is not required to be injective, which may result in degenerate drawings where distinct vertices are drawn

as the same point. For this reason, we prefer not to use the terminology *graph embedding*, which is often used in the literature. This is because in differential geometry, an embedding always refers to an injective map. The term *graph immersion* would be more appropriate.

As explained in Godsil and Royle [Godsil and Royle (2001)], we can imagine building a physical model of G by connecting adjacent vertices (in $\mathbb{R}^n$) by identical springs. Then it is natural to consider a representation to be better if it requires the springs to be less extended. We can formalize this by defining the *energy* of a drawing R by

$$\mathcal{E}(R) = \sum_{\{v_i, v_j\} \in E} \|\rho(v_i) - \rho(v_j)\|^2,$$

where $\rho(v_i)$ is the ith row of R and $\|\rho(v_i) - \rho(v_j)\|^2$ is the square of the Euclidean length of the line segment joining $\rho(v_i)$ and $\rho(v_j)$.

Then, "good drawings" are drawings that minimize the energy function $\mathcal{E}$. Of course, the trivial representation corresponding to the zero matrix is optimum, so we need to impose extra constraints to rule out the trivial solution.

We can consider the more general situation where the springs are not necessarily identical. This can be modeled by a symmetric weight (or stiffness) matrix $W = (w_{ij})$, with $w_{ij} \geq 0$. Then our energy function becomes

$$\mathcal{E}(R) = \sum_{\{v_i, v_j\} \in E} w_{ij} \|\rho(v_i) - \rho(v_j)\|^2.$$

It turns out that this function can be expressed in terms of the Laplacian $L = D - W$. The following proposition is shown in Godsil and Royle [Godsil and Royle (2001)]. We give a slightly more direct proof.

Proposition 19.1. *Let $G = (V, W)$ be a weighted graph, with $|V| = m$ and W an $m \times m$ symmetric matrix, and let R be the matrix of a graph drawing ρ of G in $\mathbb{R}^n$ (a $m \times n$ matrix). If $L = D - W$ is the unnormalized Laplacian matrix associated with W, then*

$$\mathcal{E}(R) = \mathrm{tr}(R^\top L R).$$

Proof. Since $\rho(v_i)$ is the ith row of R (and $\rho(v_j)$ is the jth row of R), if

we denote the kth column of R by R^k, using Proposition 18.4, we have

$$\begin{aligned}\mathcal{E}(R) &= \sum_{\{v_i,v_j\}\in E} w_{ij}\,\|\rho(v_i)-\rho(v_j)\|^2\\ &= \sum_{k=1}^{n}\sum_{\{v_i,v_j\}\in E} w_{ij}(R_{ik}-R_{jk})^2\\ &= \sum_{k=1}^{n}\frac{1}{2}\sum_{i,j=1}^{m} w_{ij}(R_{ik}-R_{jk})^2\\ &= \sum_{k=1}^{n}(R^k)^\top LR^k = \mathrm{tr}(R^\top LR),\end{aligned}$$

as claimed. □

Since the matrix $R^\top LR$ is symmetric, it has real eigenvalues. Actually, since L is positive semidefinite, so is $R^\top LR$. Then the trace of $R^\top LR$ is equal to the sum of its positive eigenvalues, and this is the energy $\mathcal{E}(R)$ of the graph drawing.

If R is the matrix of a graph drawing in $\mathbb{R}^n$, then for any $n\times n$ invertible matrix M, the map that assigns $\rho(v_i)M$ to v_i is another graph drawing of G, and these two drawings convey the same amount of information. From this point of view, *a graph drawing is determined by the column space of R.* Therefore, it is reasonable to assume that the columns of R are pairwise orthogonal and that they have unit length. Such a matrix satisfies the equation $R^\top R = I$.

Definition 19.3. If the matrix R of a graph drawing satisfies the equation $R^\top R = I$, then the corresponding drawing is called an *orthogonal graph drawing*.

This above condition also rules out trivial drawings. The following result tells us how to find minimum energy orthogonal balanced graph drawings, provided the graph is connected. Recall that

$$L\mathbf{1} = 0,$$

as we already observed.

Theorem 19.1. *Let $G = (V, W)$ be a weighted graph with $|V| = m$. If $L = D - W$ is the (unnormalized) Laplacian of G, and if the eigenvalues of L are $0 = \lambda_1 < \lambda_2 \leq \lambda_3 \leq \ldots \leq \lambda_m$, then the minimal energy of any balanced orthogonal graph drawing of G in $\mathbb{R}^n$ is equal to $\lambda_2 + \cdots + \lambda_{n+1}$*

(in particular, this implies that $n < m$). The $m \times n$ matrix R consisting of any unit eigenvectors $u_2, \ldots, u_{n+1}$ associated with $\lambda_2 \leq \ldots \leq \lambda_{n+1}$ yields a balanced orthogonal graph drawing of minimal energy; it satisfies the condition $R^\top R = I$.

Proof. We present the proof given in Godsil and Royle [Godsil and Royle (2001)] (Section 13.4, Theorem 13.4.1). The key point is that the sum of the n smallest eigenvalues of L is a lower bound for $\mathrm{tr}(R^\top LR)$. This can be shown using a Rayleigh ratio argument; see Proposition 16.13 (the Poincaré separation theorem). Then any n eigenvectors $(u_1, \ldots, u_n)$ associated with $\lambda_1, \ldots, \lambda_n$ achieve this bound. Because the first eigenvalue of L is $\lambda_1 = 0$ and because we are assuming that $\lambda_2 > 0$, we have $u_1 = \mathbf{1}/\sqrt{m}$. Since the u_j are pairwise orthogonal for $i = 2, \ldots, n$ and since u_i is orthogonal to $u_1 = \mathbf{1}/\sqrt{m}$, the entries in u_i add up to 0. Consequently, for any ℓ with $2 \leq \ell \leq n$, by deleting u_1 and using $(u_2, \ldots, u_\ell)$, we obtain a balanced orthogonal graph drawing in $\mathbb{R}^{\ell-1}$ with the same energy as the orthogonal graph drawing in $\mathbb{R}^\ell$ using $(u_1, u_2, \ldots, u_\ell)$. Conversely, from any balanced orthogonal drawing in $\mathbb{R}^{\ell-1}$ using $(u_2, \ldots, u_\ell)$, we obtain an orthogonal graph drawing in $\mathbb{R}^\ell$ using $(u_1, u_2, \ldots, u_\ell)$ with the same energy. Therefore, the minimum energy of a balanced orthogonal graph drawing in $\mathbb{R}^n$ is equal to the minimum energy of an orthogonal graph drawing in $\mathbb{R}^{n+1}$, and this minimum is $\lambda_2 + \cdots + \lambda_{n+1}$. □

Since $\mathbf{1}$ spans the nullspace of L, using u_1 (which belongs to $\mathrm{Ker}\, L$) as one of the vectors in R would have the effect that all points representing vertices of G would have the same first coordinate. This would mean that the drawing lives in a hyperplane in $\mathbb{R}^n$, which is undesirable, especially when $n = 2$, where all vertices would be collinear. This is why we omit the first eigenvector u_1.

Observe that for any orthogonal $n \times n$ matrix Q, since

$$\mathrm{tr}(R^\top LR) = \mathrm{tr}(Q^\top R^\top LRQ),$$

the matrix RQ also yields a minimum orthogonal graph drawing. This amounts to applying the rigid motion $Q^\top$ to the rows of R.

In summary, if $\lambda_2 > 0$, an automatic method for drawing a graph in $\mathbb{R}^2$ is this:

(1) Compute the two smallest nonzero eigenvalues $\lambda_2 \leq \lambda_3$ of the graph Laplacian L (it is possible that $\lambda_3 = \lambda_2$ if λ_2 is a multiple eigenvalue);
(2) Compute two unit eigenvectors u_2, u_3 associated with λ_2 and λ_3, and let $R = [u_2\ u_3]$ be the $m \times 2$ matrix having u_2 and u_3 as columns.

(3) Place vertex v_i at the point whose coordinates is the ith row of R, that is, (R_{i1}, R_{i2}).

This method generally gives pleasing results, but beware that there is no guarantee that distinct nodes are assigned distinct images since R can have identical rows. This does not seem to happen often in practice.

19.2 Examples of Graph Drawings

We now give a number of examples using `Matlab`. Some of these are borrowed or adapted from Spielman [Spielman (2012)].

Example 1. Consider the graph with four nodes whose adjacency matrix is

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

We use the following program to compute u_2 and u_3:

```
A = [0 1 1 0; 1 0 0 1; 1 0 0 1; 0 1 1 0];
D = diag(sum(A));
L = D - A;
[v, e] = eigs(L);
gplot(A, v(:,[3 2]))
hold on;
gplot(A, v(:,[3 2]),'o')
```

The graph of Example 1 is shown in Figure 19.1. The function `eigs(L)` computes the six largest eigenvalues of L in decreasing order, and corresponding eigenvectors. It turns out that $\lambda_2 = \lambda_3 = 2$ is a double eigenvalue.

Example 2. Consider the graph G_2 shown in Figure 18.3 given by the adjacency matrix

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

We use the following program to compute u_2 and u_3:

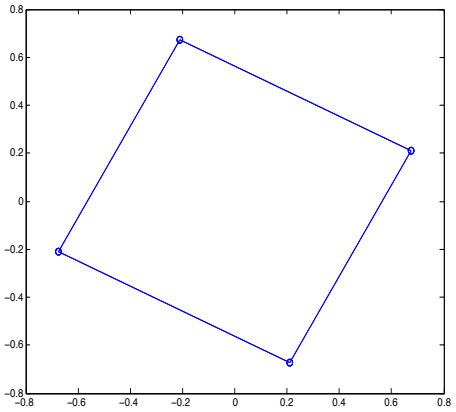

Fig. 19.1 Drawing of the graph from Example 1.

```
A = [0 1 1 0 0; 1 0 1 1 1; 1 1 0 1 0; 0 1 1 0 1; 0 1 0 1 0];
D = diag(sum(A));
L = D - A;
[v, e] = eig(L);
gplot(A, v(:, [2 3]))
hold on
gplot(A, v(:, [2 3]),'o')
```

The function `eig(L)` (with no `s` at the end) computes the eigenvalues of L in increasing order. The result of drawing the graph is shown in Figure 19.2. Note that node v_2 is assigned to the point $(0, 0)$, so the difference between this drawing and the drawing in Figure 18.3 is that the drawing of Figure 19.2 is not convex.

Example 3. Consider the ring graph defined by the adjacency matrix A given in the `Matlab` program shown below:

```
A = diag(ones(1, 11),1);
A = A + A';
A(1, 12) = 1; A(12, 1) = 1;
D = diag(sum(A));
L = D - A;
[v, e] = eig(L);
gplot(A, v(:, [2 3]))
hold on
```

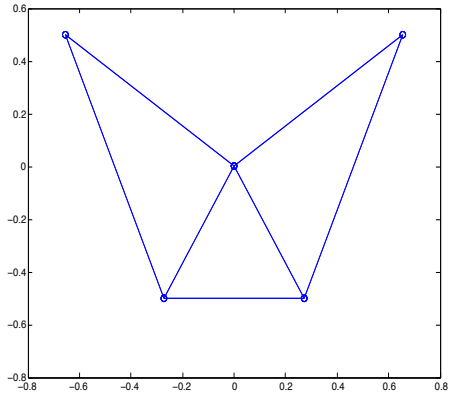

Fig. 19.2 Drawing of the graph from Example 2.

```
gplot(A, v(:, [2 3]),'o')
```

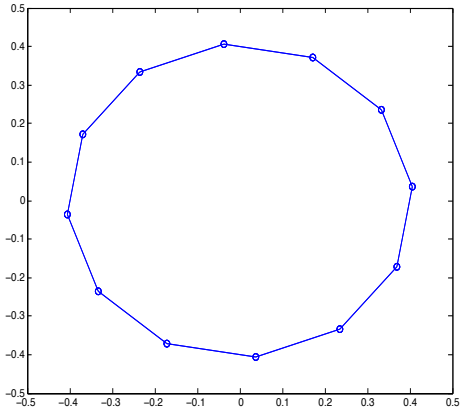

Fig. 19.3 Drawing of the graph from Example 3.

Observe that we get a very nice ring; see Figure 19.3. Again $\lambda_2 = 0.2679$ is a double eigenvalue (and so are the next pairs of eigenvalues, except the last, $\lambda_{12} = 4$).

Example 4. In this example adapted from Spielman, we generate 20 randomly chosen points in the unit square, compute their Delaunay trian-

gulation, then the adjacency matrix of the corresponding graph, and finally draw the graph using the second and third eigenvalues of the Laplacian.

```
A = zeros(20,20);
xy = rand(20, 2);
trigs = delaunay(xy(:,1), xy(:,2));
elemtrig = ones(3) - eye(3);
for i = 1:length(trigs),
 A(trigs(i,:),trigs(i,:)) = elemtrig;
end
A = double(A >0);
gplot(A,xy)
D = diag(sum(A));
L = D - A;
[v, e] = eigs(L, 3, 'sm');
figure(2)
gplot(A, v(:, [2 1]))
hold on
gplot(A, v(:, [2 1]),'o')
```

The Delaunay triangulation of the set of 20 points and the drawing of the corresponding graph are shown in Figure 19.4. The graph drawing on the right looks nicer than the graph on the left but is is no longer planar.

Example 5. Our last example, also borrowed from Spielman [Spielman (2012)], corresponds to the skeleton of the "Buckyball," a geodesic dome invented by the architect Richard Buckminster Fuller (1895–1983). The Montréal Biosphère is an example of a geodesic dome designed by Buckminster Fuller.

```
A = full(bucky);
D = diag(sum(A));
L = D - A;
[v, e] = eig(L);
gplot(A, v(:, [2 3]))
hold on;
gplot(A,v(:, [2 3]), 'o')
```

Figure 19.5 shows a graph drawing of the Buckyball. This picture seems a bit squashed for two reasons. First, it is really a 3-dimensional graph; second, $\lambda_2 = 0.2434$ is a triple eigenvalue. (Actually, the Laplacian of L

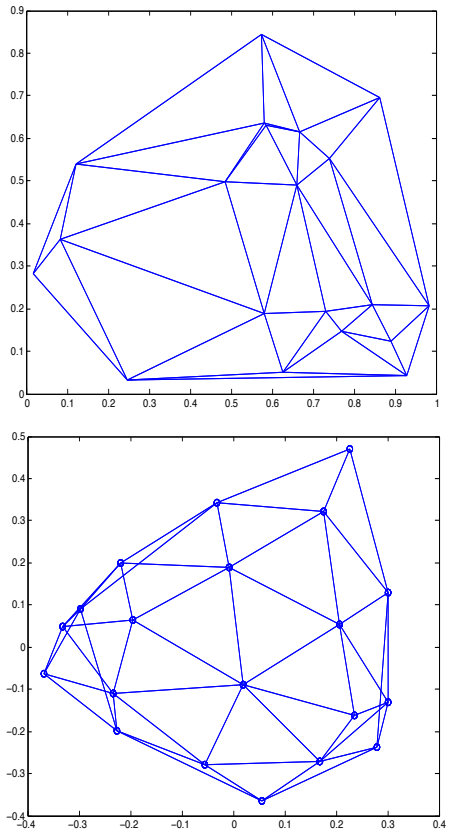

Fig. 19.4 Delaunay triangulation (left) and drawing of the graph from Example 4 (right).

has many multiple eigenvalues.) What we should really do is to plot this graph in $\mathbb{R}^3$ using three orthonormal eigenvectors associated with λ_2.

A 3D picture of the graph of the Buckyball is produced by the following `Matlab` program, and its image is shown in Figure 19.6. It looks better!

```
[x, y] = gplot(A, v(:, [2 3]));
[x, z] = gplot(A, v(:, [2 4]));
plot3(x,y,z)
```

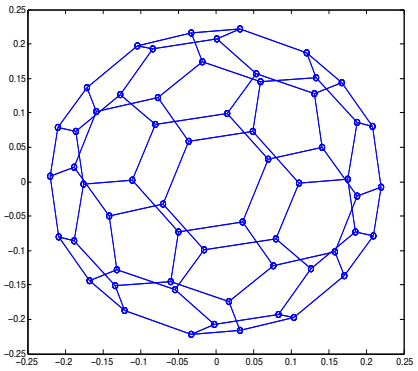

Fig. 19.5 Drawing of the graph of the Buckyball.

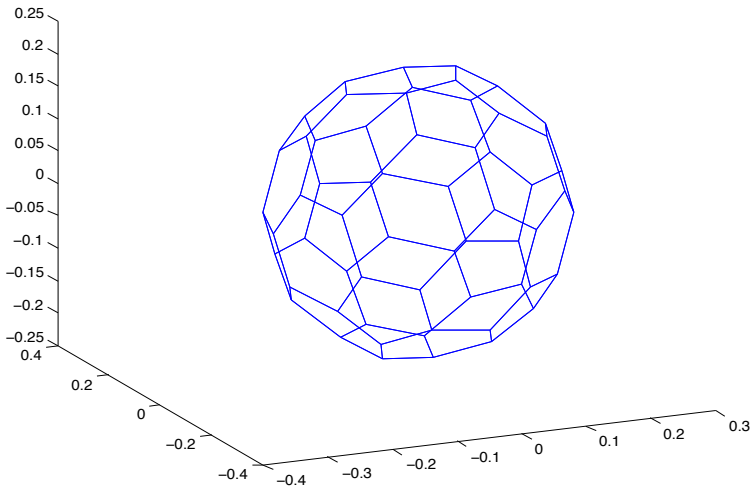

Fig. 19.6 Drawing of the graph of the Buckyball in $\mathbb{R}^3$.

19.3 Summary

The main concepts and results of this chapter are listed below:

- Graph drawing.
- Matrix of a graph drawing.

- Balanced graph drawing.
- Energy $\mathcal{E}(R)$ of a graph drawing.
- Orthogonal graph drawing.
- Delaunay triangulation.
- Buckyball.

Chapter 20

Singular Value Decomposition and Polar Form

20.1 Properties of $f^* \circ f$

In this section we assume that we are dealing with real Euclidean spaces. Let $f\colon E \to E$ be any linear map. In general, it may not be possible to diagonalize f. We show that every linear map can be diagonalized if we are willing to use *two* orthonormal bases. This is the celebrated *singular value decomposition (SVD)*. A close cousin of the SVD is the *polar form* of a linear map, which shows how a linear map can be decomposed into its purely rotational component (perhaps with a flip) and its purely stretching part.

The key observation is that $f^* \circ f$ is self-adjoint since

$$\langle (f^* \circ f)(u), v\rangle = \langle f(u), f(v)\rangle = \langle u, (f^* \circ f)(v)\rangle.$$

Similarly, $f \circ f^*$ is self-adjoint.

The fact that $f^* \circ f$ and $f \circ f^*$ are self-adjoint is very important, because by Theorem 16.1, it implies that *$f^* \circ f$ and $f \circ f^*$ can be diagonalized and that they have real eigenvalues.* In fact, these *eigenvalues are all nonnegative* as shown in the following proposition.

Proposition 20.1. *The eigenvalues of $f^* \circ f$ and $f \circ f^*$ are nonnegative.*

Proof. If u is an eigenvector of $f^* \circ f$ for the eigenvalue λ, then

$$\langle (f^* \circ f)(u), u\rangle = \langle f(u), f(u)\rangle$$

and

$$\langle (f^* \circ f)(u), u\rangle = \lambda \langle u, u\rangle,$$

and thus

$$\lambda \langle u, u\rangle = \langle f(u), f(u)\rangle,$$

which implies that $\lambda \geq 0$, since $\langle -, -\rangle$ is positive definite. A similar proof applies to $f \circ f^*$. □

Thus, the eigenvalues of $f^* \circ f$ are of the form $\sigma_1^2, \ldots, \sigma_r^2$ or 0, where $\sigma_i > 0$, and similarly for $f \circ f^*$.

The above considerations also apply to any linear map $f\colon E \to F$ between two Euclidean spaces $(E, \langle -, - \rangle_1)$ and $(F, \langle -, - \rangle_2)$. Recall that the adjoint $f^*\colon F \to E$ of f is the unique linear map f^* such that

$$\langle f(u), v \rangle_2 = \langle u, f^*(v) \rangle_1, \quad \text{for all } u \in E \text{ and all } v \in F.$$

Then $f^* \circ f$ and $f \circ f^*$ are self-adjoint (the proof is the same as in the previous case), and the eigenvalues of $f^* \circ f$ and $f \circ f^*$ are nonnegative.

Proof. If λ is an eigenvalue of $f^* \circ f$ and u $(\neq 0)$ is a corresponding eigenvector, we have

$$\langle (f^* \circ f)(u), u \rangle_1 = \langle f(u), f(u) \rangle_2,$$

and also

$$\langle (f^* \circ f)(u), u \rangle_1 = \lambda \langle u, u \rangle_1,$$

so

$$\lambda \langle u, u \rangle_1, = \langle f(u), f(u) \rangle_2,$$

which implies that $\lambda \geq 0$. A similar proof applies to $f \circ f^*$. □

The situation is even better, since we will show shortly that $f^* \circ f$ and $f \circ f^*$ have the same nonzero eigenvalues.

Remark: Given any two linear maps $f\colon E \to F$ and $g\colon F \to E$, where $\dim(E) = n$ and $\dim(F) = m$, it can be shown that

$$\lambda^m \det(\lambda I_n - g \circ f) = \lambda^n \det(\lambda I_m - f \circ g),$$

and thus $g \circ f$ and $f \circ g$ always have the same nonzero eigenvalues; see Problem 14.14.

Definition 20.1. Given any linear map $f\colon E \to F$, the square roots $\sigma_i > 0$ of the positive eigenvalues of $f^* \circ f$ (and $f \circ f^*$) are called the *singular values of* f.

Definition 20.2. A self-adjoint linear map $f\colon E \to E$ whose eigenvalues are nonnegative is called *positive semidefinite* (or *positive*), and if f is also invertible, f is said to be *positive definite*. In the latter case, every eigenvalue of f is strictly positive.

If $f\colon E \to F$ is any linear map, we just showed that $f^* \circ f$ and $f \circ f^*$ are positive semidefinite self-adjoint linear maps. This fact has the remarkable consequence that every linear map has two important decompositions:

(1) The polar form.
(2) The singular value decomposition (SVD).

The wonderful thing about the singular value decomposition is that there exist two orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ such that, with respect to these bases, f is a diagonal matrix consisting of the singular values of f or 0. Thus, in some sense, f can always be diagonalized with respect to *two* orthonormal bases. The SVD is also a useful tool for solving overdetermined linear systems in the least squares sense and for data analysis, as we show later on.

First we show some useful relationships between the kernels and the images of f, f^*, $f^* \circ f$, and $f \circ f^*$. Recall that if $f\colon E \to F$ is a linear map, the *image* $\operatorname{Im} f$ *of* f is the subspace $f(E)$ of F, and the *rank of* f is the dimension $\dim(\operatorname{Im} f)$ of its image. Also recall that (Theorem 5.1)

$$\dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f) = \dim(E),$$

and that (Propositions 11.9 and 13.12) for every subspace W of E,

$$\dim(W) + \dim(W^{\perp}) = \dim(E).$$

Proposition 20.2. *Given any two Euclidean spaces E and F, where E has dimension n and F has dimension m, for any linear map $f\colon E \to F$, we have*

$$\begin{aligned}
\operatorname{Ker} f &= \operatorname{Ker}(f^* \circ f),\\
\operatorname{Ker} f^* &= \operatorname{Ker}(f \circ f^*),\\
\operatorname{Ker} f &= (\operatorname{Im} f^*)^{\perp},\\
\operatorname{Ker} f^* &= (\operatorname{Im} f)^{\perp},\\
\dim(\operatorname{Im} f) &= \dim(\operatorname{Im} f^*),
\end{aligned}$$

and f, f^*, $f^* \circ f$, and $f \circ f^*$ have the same rank.

Proof. To simplify the notation, we will denote the inner products on E and F by the same symbol $\langle -, - \rangle$ (to avoid subscripts). If $f(u) = 0$, then $(f^* \circ f)(u) = f^*(f(u)) = f^*(0) = 0$, and so $\operatorname{Ker} f \subseteq \operatorname{Ker}(f^* \circ f)$. By definition of f^*, we have

$$\langle f(u), f(u) \rangle = \langle (f^* \circ f)(u), u \rangle$$

for all $u \in E$. If $(f^* \circ f)(u) = 0$, since $\langle -, - \rangle$ is positive definite, we must have $f(u) = 0$, and so $\operatorname{Ker}(f^* \circ f) \subseteq \operatorname{Ker} f$. Therefore,

$$\operatorname{Ker} f = \operatorname{Ker}(f^* \circ f).$$

The proof that $\operatorname{Ker} f^* = \operatorname{Ker}(f \circ f^*)$ is similar.

By definition of f^*, we have

$$\langle f(u), v \rangle = \langle u, f^*(v) \rangle \quad \text{for all } u \in E \text{ and all } v \in F. \tag{$*$}$$

This immediately implies that

$$\operatorname{Ker} f = (\operatorname{Im} f^*)^\perp \quad \text{and} \quad \operatorname{Ker} f^* = (\operatorname{Im} f)^\perp.$$

Let us explain why $\operatorname{Ker} f = (\operatorname{Im} f^*)^\perp$, the proof of the other equation being similar.

Because the inner product is positive definite, for every $u \in E$, we have

- $u \in \operatorname{Ker} f$
- iff $f(u) = 0$
- iff $\langle f(u), v \rangle = 0$ for all v,
- by $(*)$ iff $\langle u, f^*(v) \rangle = 0$ for all v,
- iff $u \in (\operatorname{Im} f^*)^\perp$.

Since

$$\dim(\operatorname{Im} f) = n - \dim(\operatorname{Ker} f)$$

and

$$\dim(\operatorname{Im} f^*) = n - \dim((\operatorname{Im} f^*)^\perp),$$

from

$$\operatorname{Ker} f = (\operatorname{Im} f^*)^\perp$$

we also have

$$\dim(\operatorname{Ker} f) = \dim((\operatorname{Im} f^*)^\perp),$$

from which we obtain

$$\dim(\operatorname{Im} f) = \dim(\operatorname{Im} f^*).$$

Since

$$\dim(\operatorname{Ker}(f^* \circ f)) + \dim(\operatorname{Im}(f^* \circ f)) = \dim(E),$$

$\operatorname{Ker}(f^* \circ f) = \operatorname{Ker} f$ and $\operatorname{Ker} f = (\operatorname{Im} f^*)^\perp$, we get

$$\dim((\operatorname{Im} f^*)^\perp) + \dim(\operatorname{Im}(f^* \circ f)) = \dim(E).$$

Since

$$\dim((\operatorname{Im} f^*)^\perp) + \dim(\operatorname{Im} f^*) = \dim(E),$$

we deduce that

$$\dim(\operatorname{Im} f^*) = \dim(\operatorname{Im}(f^* \circ f)).$$

A similar proof shows that

$$\dim(\operatorname{Im} f) = \dim(\operatorname{Im}(f \circ f^*)).$$

Consequently, f, f^*, $f^* \circ f$, and $f \circ f^*$ have the same rank. □

20.2 Singular Value Decomposition for Square Matrices

We will now prove that every square matrix has an SVD. Stronger results can be obtained if we first consider the polar form and then derive the SVD from it (there are uniqueness properties of the polar decomposition). For our purposes, uniqueness results are not as important so we content ourselves with existence results, whose proofs are simpler. Readers interested in a more general treatment are referred to Gallier [Gallier (2011b)].

The early history of the singular value decomposition is described in a fascinating paper by Stewart [Stewart (1993)]. The SVD is due to Beltrami and Camille Jordan independently (1873, 1874). Gauss is the grandfather of all this, for his work on least squares (1809, 1823) (but Legendre also published a paper on least squares!). Then come Sylvester, Schmidt, and Hermann Weyl. Sylvester's work was apparently "opaque." He gave a computational method to find an SVD. Schmidt's work really has to do with integral equations and symmetric and asymmetric kernels (1907). Weyl's work has to do with perturbation theory (1912). Autonne came up with the polar decomposition (1902, 1915). Eckart and Young extended SVD to rectangular matrices (1936, 1939).

Theorem 20.1. *(Singular value decomposition) For every real $n \times n$ matrix A there are two orthogonal matrices U and V and a diagonal matrix D such that $A = VDU^\top$, where D is of the form*

$$D = \begin{pmatrix} \sigma_1 & \cdots & & \\ & \sigma_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \sigma_n \end{pmatrix},$$

where $\sigma_1, \ldots, \sigma_r$ are the singular values of f, i.e., the (positive) square roots of the nonzero eigenvalues of $A^\top A$ and $AA^\top$, and $\sigma_{r+1} = \cdots = \sigma_n = 0$. The columns of U are eigenvectors of $A^\top A$, and the columns of V are eigenvectors of $AA^\top$.

Proof. Since $A^\top A$ is a symmetric matrix, in fact, a positive semidefinite matrix, there exists an orthogonal matrix U such that

$$A^\top A = UD^2U^\top,$$

with $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$, where $\sigma_1^2, \ldots, \sigma_r^2$ are the nonzero eigenvalues of $A^\top A$, and where r is the rank of A; that is, $\sigma_1, \ldots, \sigma_r$ are the

singular values of A. It follows that

$$U^\top A^\top AU = (AU)^\top AU = D^2,$$

and if we let f_j be the jth column of AU for $j = 1, \ldots, n$, then we have

$$\langle f_i, f_j \rangle = \sigma_i^2 \delta_{ij}, \quad 1 \leq i, j \leq r$$

and

$$f_j = 0, \quad r+1 \leq j \leq n.$$

If we define $(v_1, \ldots, v_r)$ by

$$v_j = \sigma_j^{-1} f_j, \quad 1 \leq j \leq r,$$

then we have

$$\langle v_i, v_j \rangle = \delta_{ij}, \quad 1 \leq i, j \leq r,$$

so complete $(v_1, \ldots, v_r)$ into an orthonormal basis $(v_1, \ldots, v_r, v_{r+1}, \ldots, v_n)$ (for example, using Gram–Schmidt). Now since $f_j = \sigma_j v_j$ for $j = 1 \ldots, r$, we have

$$\langle v_i, f_j \rangle = \sigma_j \langle v_i, v_j \rangle = \sigma_j \delta_{i,j}, \quad 1 \leq i \leq n,\ 1 \leq j \leq r$$

and since $f_j = 0$ for $j = r+1, \ldots, n$,

$$\langle v_i, f_j \rangle = 0 \quad 1 \leq i \leq n,\ r+1 \leq j \leq n.$$

If V is the matrix whose columns are $v_1, \ldots, v_n$, then V is orthogonal and the above equations prove that

$$V^\top AU = D,$$

which yields $A = VDU^\top$, as required.

The equation $A = VDU^\top$ implies that

$$A^\top A = UD^2U^\top, \quad AA^\top = VD^2V^\top,$$

which shows that $A^\top A$ and $AA^\top$ have the same eigenvalues, that the columns of U are eigenvectors of $A^\top A$, and that the columns of V are eigenvectors of $AA^\top$. □

Example 20.1. Here is a simple example of how to use the proof of Theorem 20.1 to obtain an SVD decomposition. Let $A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$. Then $A^\top = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$, $A^\top A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, and $AA^\top = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$. A simple calculation shows that the eigenvalues of $A^\top A$ are 2 and 0, and for the eigenvalue 2,

a unit eigenvector is $\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$, while a unit eigenvector for the eigenvalue 0 is $\begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$. Observe that the singular values are $\sigma_1 = \sqrt{2}$ and $\sigma_2 = 0$. Furthermore, $U = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} = U^\top$. To determine V, the proof of Theorem 20.1 tells us to first calculate

$$AU = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix},$$

and then set

$$v_1 = (1/\sqrt{2}) \begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Once v_1 is determined, since $\sigma_2 = 0$, we have the freedom to choose v_2 such that (v_1, v_2) forms an orthonormal basis for $\mathbb{R}^2$. Naturally, we chose $v_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and set $V = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. The columns of V are unit eigenvectors of $AA^\top$, but finding V by computing unit eigenvectors of $AA^\top$ does not guarantee that these vectors are consistent with U so that $A = V\Sigma U^\top$. Thus one has to use AU instead. We leave it to the reader to check that

$$A = V \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix} U^\top.$$

Theorem 20.1 suggests the following definition.

Definition 20.3. A triple (U, D, V) such that $A = VDU^\top$, where U and V are orthogonal and D is a diagonal matrix whose entries are nonnegative (it is positive semidefinite) is called a *singular value decomposition (SVD) of* A. If $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$, it is customary to assume that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$.

The `Matlab` command for computing an SVD $A = VDU^\top$ of a matrix A is `[V, D, U] = svd(A)`.

The proof of Theorem 20.1 shows that there are two orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$, where $(u_1, \ldots, u_n)$ are eigenvectors of $A^\top A$ and $(v_1, \ldots, v_n)$ are eigenvectors of $AA^\top$. Furthermore, $(u_1, \ldots, u_r)$ is an orthonormal basis of $\mathrm{Im}\, A^\top$, $(u_{r+1}, \ldots, u_n)$ is an orthonormal basis of $\mathrm{Ker}\, A$, $(v_1, \ldots, v_r)$ is an orthonormal basis of $\mathrm{Im}\, A$, and $(v_{r+1}, \ldots, v_n)$ is an orthonormal basis of $\mathrm{Ker}\, A^\top$.

Using a remark made in Chapter 3, if we denote the columns of U by $u_1, \ldots, u_n$ and the columns of V by $v_1, \ldots, v_n$, then we can write

$$A = VDU^\top = \sigma_1 v_1 u_1^\top + \cdots + \sigma_r v_r u_r^\top,$$

with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. As a consequence, if r is a lot smaller than n (we write $r \ll n$), we see that A can be reconstructed from U and V using a much smaller number of elements. This idea will be used to provide "low-rank" approximations of a matrix. The idea is to keep only the k top singular values for some suitable $k \ll r$ for which $\sigma_{k+1}, \ldots \sigma_r$ are very small.

Remarks:

(1) In Strang [Strang (1988)] the matrices U, V, D are denoted by $U = Q_2$, $V = Q_1$, and $D = \Sigma$, and an SVD is written as $A = Q_1 \Sigma Q_2^\top$. This has the advantage that Q_1 comes before Q_2 in $A = Q_1 \Sigma Q_2^\top$. This has the disadvantage that A maps the columns of Q_2 (eigenvectors of $A^\top A$) to multiples of the columns of Q_1 (eigenvectors of $A\,A^\top$).

(2) Algorithms for actually computing the SVD of a matrix are presented in Golub and Van Loan [Golub and Van Loan (1996)], Demmel [Demmel (1997)], and Trefethen and Bau [Trefethen and Bau III (1997)], where the SVD and its applications are also discussed quite extensively.

(3) If A is a symmetric matrix, then in general, there is no SVD $V \Sigma U^\top$ of A with $V = U$. However, if A is positive semidefinite, then the eigenvalues of A are nonnegative, and so the nonzero eigenvalues of A are equal to the singular values of A and SVDs of A are of the form

$$A = V \Sigma V^\top.$$

(4) The SVD also applies to complex matrices. In this case, for every complex $n \times n$ matrix A, there are two unitary matrices U and V and a diagonal matrix D such that

$$A = V D U^*,$$

where D is a diagonal matrix consisting of real entries $\sigma_1, \ldots, \sigma_n$, where $\sigma_1 \geq \cdots \geq \sigma_r$ are the singular values of A, i.e., the positive square roots of the nonzero eigenvalues of A^*A and $A\,A^*$, and $\sigma_{r+1} = \ldots = \sigma_n = 0$.

20.3 Polar Form for Square Matrices

A notion closely related to the SVD is the polar form of a matrix.

Definition 20.4. A pair (R, S) such that $A = RS$ with R orthogonal and S symmetric positive semidefinite is called a *polar decomposition of* A.

Theorem 20.1 implies that for every real $n \times n$ matrix A, there is some orthogonal matrix R and some positive semidefinite symmetric matrix S such that

$$A = RS.$$

This is easy to show and we will prove it below. Furthermore, R, S are unique if A is invertible, but this is harder to prove; see Problem 20.9.

For example, the matrix

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

is both orthogonal and symmetric, and $A = RS$ with $R = A$ and $S = I$, which implies that some of the eigenvalues of A are negative.

Remark: In the complex case, the polar decomposition states that for every complex $n \times n$ matrix A, there is some unitary matrix U and some positive semidefinite Hermitian matrix H such that

$$A = UH.$$

It is easy to go from the polar form to the SVD, and conversely.

Given an SVD decomposition $A = VDU^\top$, let $R = VU^\top$ and $S = UDU^\top$. It is clear that R is orthogonal and that S is positive semidefinite symmetric, and

$$RS = VU^\top UDU^\top = VDU^\top = A.$$

Example 20.2. Recall from Example 20.1 that $A = VDU^\top$ where $V = I_2$ and

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \qquad U = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}, \qquad D = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix}.$$

Set $R = VU^\top = U$ and

$$S = UDU^\top = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

Since $S = \frac{1}{\sqrt{2}}A^\top A$, S has eigenvalues $\sqrt{2}$ and 0. We leave it to the reader to check that $A = RS$.

Going the other way, given a polar decomposition $A = R_1 S$, where R_1 is orthogonal and S is positive semidefinite symmetric, there is an orthogonal matrix R_2 and a positive semidefinite diagonal matrix D such that $S = R_2 D\, R_2^\top$, and thus

$$A = R_1 R_2 D\, R_2^\top = V D\, U^\top,$$

where $V = R_1 R_2$ and $U = R_2$ are orthogonal.

Example 20.3. Let $A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$ and $A = R_1 S$, where $R_1 = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$ and $S = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$. This is the polar decomposition of Example 20.2. Observe that

$$S = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} = R_2 D R_2^\top.$$

Set $U = R_2$ and $V = R_1 R_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ to obtain the SVD decomposition of Example 20.1.

The eigenvalues and the singular values of a matrix are typically not related in any obvious way. For example, the $n \times n$ matrix

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 2 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 1 & 2 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & 2 \\ 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{pmatrix}$$

has the eigenvalue 1 with multiplicity n, but its singular values, $\sigma_1 \geq \cdots \geq \sigma_n$, which are the positive square roots of the eigenvalues of the matrix $B = A^\top A$ with

$$B = \begin{pmatrix} 1 & 2 & 0 & 0 & \dots & 0 & 0 \\ 2 & 5 & 2 & 0 & \dots & 0 & 0 \\ 0 & 2 & 5 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 2 & 5 & 2 & 0 \\ 0 & 0 & \dots & 0 & 2 & 5 & 2 \\ 0 & 0 & \dots & 0 & 0 & 2 & 5 \end{pmatrix}$$

have a wide spread, since

$$\frac{\sigma_1}{\sigma_n} = \mathrm{cond}_2(A) \geq 2^{n-1}.$$

If A is a complex $n \times n$ matrix, the eigenvalues $\lambda_1, \ldots, \lambda_n$ and the singular values
$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ of A are not unrelated, since

$$\sigma_1^2 \cdots \sigma_n^2 = \det(A^* A) = |\det(A)|^2$$

and

$$|\lambda_1| \cdots |\lambda_n| = |\det(A)|,$$

so we have

$$|\lambda_1| \cdots |\lambda_n| = \sigma_1 \cdots \sigma_n.$$

More generally, Hermann Weyl proved the following remarkable theorem:

Theorem 20.2. *(Weyl's inequalities, 1949) For any complex $n \times n$ matrix, A, if $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ are the eigenvalues of A and $\sigma_1, \ldots, \sigma_n \in \mathbb{R}_+$ are the singular values of A, listed so that $|\lambda_1| \geq \cdots \geq |\lambda_n|$ and $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$, then*

$$|\lambda_1| \cdots |\lambda_n| = \sigma_1 \cdots \sigma_n \quad \textit{and}$$
$$|\lambda_1| \cdots |\lambda_k| \leq \sigma_1 \cdots \sigma_k, \quad \textit{for} \quad k = 1, \ldots, n-1.$$

A proof of Theorem 20.2 can be found in Horn and Johnson [Horn and Johnson (1994)], Chapter 3, Section 3.3, where more inequalities relating the eigenvalues and the singular values of a matrix are given.

Theorem 20.1 can be easily extended to rectangular $m \times n$ matrices, as we show in the next section. For various versions of the SVD for rectangular matrices, see Strang [Strang (1988)] Golub and Van Loan [Golub and Van Loan (1996)], Demmel [Demmel (1997)], and Trefethen and Bau [Trefethen and Bau III (1997)].

20.4 Singular Value Decomposition for Rectangular Matrices

Here is the generalization of Theorem 20.1 to rectangular matrices.

Theorem 20.3. *(Singular value decomposition) For every real $m \times n$ matrix A, there are two orthogonal matrices U $(n \times n)$ and V $(m \times m)$ and a*

diagonal $m \times n$ matrix D such that $A = VDU^\top$, where D is of the form

$$D = \begin{pmatrix} \sigma_1 & & \dots & \\ & \sigma_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & \sigma_n \\ 0 & \vdots & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \vdots & \dots & 0 \end{pmatrix} \text{ or } D = \begin{pmatrix} \sigma_1 & & \dots & & 0 \dots 0 \\ & \sigma_2 & \dots & & 0 \dots 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \ \vdots \ 0 \\ & & \dots & \sigma_m & 0 \dots 0 \end{pmatrix},$$

where $\sigma_1, \dots, \sigma_r$ are the singular values of f, i.e. the (positive) square roots of the nonzero eigenvalues of $A^\top A$ and $AA^\top$, and $\sigma_{r+1} = \dots = \sigma_p = 0$, where $p = \min(m, n)$. The columns of U are eigenvectors of $A^\top A$, and the columns of V are eigenvectors of $AA^\top$.

Proof. As in the proof of Theorem 20.1, since $A^\top A$ is symmetric positive semidefinite, there exists an $n \times n$ orthogonal matrix U such that

$$A^\top A = U\Sigma^2 U^\top,$$

with $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0)$, where $\sigma_1^2, \dots, \sigma_r^2$ are the nonzero eigenvalues of $A^\top A$, and where r is the rank of A. Observe that $r \leq \min\{m, n\}$, and AU is an $m \times n$ matrix. It follows that

$$U^\top A^\top AU = (AU)^\top AU = \Sigma^2,$$

and if we let $f_j \in \mathbb{R}^m$ be the jth column of AU for $j = 1, \dots, n$, then we have

$$\langle f_i, f_j \rangle = \sigma_i^2 \delta_{ij}, \quad 1 \leq i, j \leq r$$

and

$$f_j = 0, \quad r + 1 \leq j \leq n.$$

If we define $(v_1, \dots, v_r)$ by

$$v_j = \sigma_j^{-1} f_j, \quad 1 \leq j \leq r,$$

then we have

$$\langle v_i, v_j \rangle = \delta_{ij}, \quad 1 \leq i, j \leq r,$$

so complete $(v_1, \dots, v_r)$ into an orthonormal basis $(v_1, \dots, v_r, v_{r+1}, \dots, v_m)$ (for example, using Gram–Schmidt).

Now since $f_j = \sigma_j v_j$ for $j = 1 \ldots, r$, we have

$$\langle v_i, f_j \rangle = \sigma_j \langle v_i, v_j \rangle = \sigma_j \delta_{i,j}, \quad 1 \leq i \leq m,\ 1 \leq j \leq r$$

and since $f_j = 0$ for $j = r+1, \ldots, n$, we have

$$\langle v_i, f_j \rangle = 0 \quad 1 \leq i \leq m,\ r+1 \leq j \leq n.$$

If V is the matrix whose columns are $v_1, \ldots, v_m$, then V is an $m \times m$ orthogonal matrix and if $m \geq n$, we let

$$D = \begin{pmatrix} \Sigma \\ 0_{m-n} \end{pmatrix} = \begin{pmatrix} \sigma_1 & \cdots & \\ & \sigma_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \sigma_n \\ 0 & \vdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \vdots & \cdots & 0 \end{pmatrix},$$

else if $n \geq m$, then we let

$$D = \begin{pmatrix} \sigma_1 & \cdots & 0 \ldots 0 \\ & \sigma_2 & \cdots & 0 \ldots 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \vdots & 0 \\ & & \cdots & \sigma_m & 0 \ldots 0 \end{pmatrix}.$$

In either case, the above equations prove that

$$V^\top A U = D,$$

which yields $A = V D U^\top$, as required.

The equation $A = V D U^\top$ implies that

$$A^\top A = U D^\top D U^\top = U \mathrm{diag}(\sigma_1^2, \ldots, \sigma_r^2, \underbrace{0, \ldots, 0}_{n-r}) U^\top$$

and

$$A A^\top = V D D^\top V^\top = V \mathrm{diag}(\sigma_1^2, \ldots, \sigma_r^2, \underbrace{0, \ldots, 0}_{m-r}) V^\top,$$

which shows that $A^\top A$ and $A A^\top$ have the same nonzero eigenvalues, that the columns of U are eigenvectors of $A^\top A$, and that the columns of V are eigenvectors of $A A^\top$. □

A triple (U, D, V) such that $A = VDU^\top$ is called a *singular value decomposition (SVD) of* A. If $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ (with $p = \min(m, n)$), it is customary to assume that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$.

Example 20.4. Let $A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$. Then $A^\top = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$ $A^\top A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, and $AA^\top = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. The reader should verify that $A^\top A = U\Sigma^2 U^\top$ where $\Sigma^2 = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$ and $U = U^\top = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$. Since $AU = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$, set $v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} \sqrt{2} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, and complete an orthonormal basis for $\mathbb{R}^3$ by assigning $v_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, and $v_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$. Thus $V = I_3$, and the reader should verify that $A = VDU^\top$, where $D = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$.

Even though the matrix D is an $m \times n$ rectangular matrix, since its only nonzero entries are on the descending diagonal, we still say that D is a diagonal matrix.

The `Matlab` command for computing an SVD $A = VDU^\top$ of a matrix A is also `[V, D, U] = svd(A)`.

If we view A as the representation of a linear map $f\colon E \to F$, where $\dim(E) = n$ and $\dim(F) = m$, the proof of Theorem 20.3 shows that there are two orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ for E and F, respectively, where $(u_1, \ldots, u_n)$ are eigenvectors of $f^* \circ f$ and $(v_1, \ldots, v_m)$ are eigenvectors of $f \circ f^*$. Furthermore, $(u_1, \ldots, u_r)$ is an orthonormal basis of $\mathrm{Im}\, f^*$, $(u_{r+1}, \ldots, u_n)$ is an orthonormal basis of $\mathrm{Ker}\, f$, $(v_1, \ldots, v_r)$ is an orthonormal basis of $\mathrm{Im}\, f$, and $(v_{r+1}, \ldots, v_m)$ is an orthonormal basis of $\mathrm{Ker}\, f^*$.

The SVD of matrices can be used to define the pseudo-inverse of a rectangular matrix; we will do so in Chapter 21. The reader may also consult Strang [Strang (1988)], Demmel [Demmel (1997)], Trefethen and

Bau [Trefethen and Bau III (1997)], and Golub and Van Loan [Golub and Van Loan (1996)].

One of the spectral theorems states that a symmetric matrix can be diagonalized by an orthogonal matrix. There are several numerical methods to compute the eigenvalues of a symmetric matrix A. One method consists in *tridiagonalizing* A, which means that there exists some orthogonal matrix P and some symmetric tridiagonal matrix T such that $A = PTP^\top$. In fact, this can be done using Householder transformations; see Theorem 17.2. It is then possible to compute the eigenvalues of T using a bisection method based on Sturm sequences. One can also use Jacobi's method. For details, see Golub and Van Loan [Golub and Van Loan (1996)], Chapter 8, Demmel [Demmel (1997)], Trefethen and Bau [Trefethen and Bau III (1997)], Lecture 26, Ciarlet [Ciarlet (1989)], and Chapter 17. Computing the SVD of a matrix A is more involved. Most methods begin by finding orthogonal matrices U and V and a *bidiagonal* matrix B such that $A = VBU^\top$; see Problem 12.8 and Problem 20.3. This can also be done using Householder transformations. Observe that $B^\top B$ is symmetric tridiagonal. Thus, in principle, the previous method to diagonalize a symmetric tridiagonal matrix can be applied. However, it is unwise to compute $B^\top B$ explicitly, and more subtle methods are used for this last step; the matrix of Problem 20.1 can be used, and see Problem 20.3. Again, see Golub and Van Loan [Golub and Van Loan (1996)], Chapter 8, Demmel [Demmel (1997)], and Trefethen and Bau [Trefethen and Bau III (1997)], Lecture 31.

The polar form has applications in continuum mechanics. Indeed, in any deformation it is important to separate stretching from rotation. This is exactly what QS achieves. The orthogonal part Q corresponds to rotation (perhaps with an additional reflection), and the symmetric matrix S to stretching (or compression). The real eigenvalues $\sigma_1, \ldots, \sigma_r$ of S are the stretch factors (or compression factors) (see Marsden and Hughes [Marsden and Hughes (1994)]). The fact that S can be diagonalized by an orthogonal matrix corresponds to a natural choice of axes, the principal axes.

The SVD has applications to data compression, for instance in image processing. The idea is to retain only singular values whose magnitudes are significant enough. The SVD can also be used to determine the rank of a matrix when other methods such as Gaussian elimination produce very small pivots. One of the main applications of the SVD is the computation of the pseudo-inverse. Pseudo-inverses are the key to the solution of various optimization problems, in particular the method of least squares. This topic is discussed in the next chapter (Chapter 21). Applications of the material

of this chapter can be found in Strang [Strang (1988, 1986)]; Ciarlet [Ciarlet (1989)]; Golub and Van Loan [Golub and Van Loan (1996)], which contains many other references; Demmel [Demmel (1997)]; and Trefethen and Bau [Trefethen and Bau III (1997)].

20.5 Ky Fan Norms and Schatten Norms

The singular values of a matrix can be used to define various norms on matrices which have found recent applications in quantum information theory and in spectral graph theory. Following Horn and Johnson [Horn and Johnson (1994)] (Section 3.4) we can make the following definitions:

Definition 20.5. For any matrix $A \in \mathrm{M}_{m,n}(\mathbb{C})$, let $q = \min\{m, n\}$, and if $\sigma_1 \geq \cdots \geq \sigma_q$ are the singular values of A, for any k with $1 \leq k \leq q$, let

$$N_k(A) = \sigma_1 + \cdots + \sigma_k,$$

called the *Ky Fan k-norm* of A.

More generally, for any $p \geq 1$ and any k with $1 \leq k \leq q$, let

$$N_{k;p}(A) = (\sigma_1^p + \cdots + \sigma_k^p)^{1/p},$$

called the *Ky Fan p-k-norm* of A. When $k = q$, $N_{q;p}$ is also called the *Schatten p-norm*.

Observe that when $k = 1$, $N_1(A) = \sigma_1$, and the Ky Fan norm N_1 is simply the *spectral norm* from Chapter 8, which is the subordinate matrix norm associated with the Euclidean norm. When $k = q$, the Ky Fan norm N_q is given by

$$N_q(A) = \sigma_1 + \cdots + \sigma_q = \mathrm{tr}((A^*A)^{1/2})$$

and is called the *trace norm* or *nuclear norm*. When $p = 2$ and $k = q$, the Ky Fan $N_{q;2}$ norm is given by

$$N_{k;2}(A) = (\sigma_1^2 + \cdots + \sigma_q^2)^{1/2} = \sqrt{\mathrm{tr}(A^*A)} = \|A\|_F,$$

which is the *Frobenius norm* of A.

It can be shown that N_k and $N_{k;p}$ are unitarily invariant norms, and that when $m = n$, they are matrix norms; see Horn and Johnson [Horn and Johnson (1994)] (Section 3.4, Corollary 3.4.4 and Problem 3).

20.6 Summary

The main concepts and results of this chapter are listed below:

- For any linear map $f\colon E \to E$ on a Euclidean space E, the maps $f^* \circ f$ and $f \circ f^*$ are self-adjoint and positive semidefinite.
- The *singular values* of a linear map.
- *Positive semidefinite* and *positive definite* self-adjoint maps.
- Relationships between $\operatorname{Im} f$, $\operatorname{Ker} f$, $\operatorname{Im} f^*$, and $\operatorname{Ker} f^*$.
- The *singular value decomposition theorem* for square matrices (Theorem 20.1).
- The *SVD* of matrix.
- The *polar decomposition* of a matrix.
- The *Weyl inequalities*.
- The *singular value decomposition theorem* for $m \times n$ matrices (Theorem 20.3).
- Ky Fan k-norms, Ky Fan p-k-norms, Schatten p-norms.

20.7 Problems

Problem 20.1. (1) Let A be a real $n \times n$ matrix and consider the $(2n) \times (2n)$ real symmetric matrix

$$S = \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}.$$

Suppose that A has rank r. If $A = V\Sigma U^\top$ is an SVD for A, with $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ and $\sigma_1 \geq \cdots \geq \sigma_r > 0$, denoting the columns of U by u_k and the columns of V by v_k, prove that σ_k is an eigenvalue of S with corresponding eigenvector $\begin{pmatrix} v_k \\ u_k \end{pmatrix}$ for $k = 1, \ldots, n$, and that $-\sigma_k$ is an eigenvalue of S with corresponding eigenvector $\begin{pmatrix} v_k \\ -u_k \end{pmatrix}$ for $k = 1, \ldots, n$.

Hint. We have $Au_k = \sigma_k v_k$ for $k = 1, \ldots, n$. Show that $A^\top v_k = \sigma_k u_k$ for $k = 1, \ldots, r$, and that $A^\top v_k = 0$ for $k = r+1, \ldots, n$. Recall that $\operatorname{Ker}(A^\top) = \operatorname{Ker}(AA^\top)$.

(2) Prove that the $2n$ eigenvectors of S in (1) are pairwise orthogonal. Check that if A has rank r, then S has rank $2r$.

(3) Now assume that A is a real $m \times n$ matrix and consider the $(m +$

$n) \times (m+n)$ real symmetric matrix

$$S = \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}.$$

Suppose that A has rank r. If $A = V\Sigma U^\top$ is an SVD for A, prove that σ_k is an eigenvalue of S with corresponding eigenvector $\begin{pmatrix} v_k \\ u_k \end{pmatrix}$ for $k = 1, \ldots, r$, and that $-\sigma_k$ is an eigenvalue of S with corresponding eigenvector $\begin{pmatrix} v_k \\ -u_k \end{pmatrix}$ for $k = 1, \ldots, r$.

Find the remaining $m + n - 2r$ eigenvectors of S associated with the eigenvalue 0.

(4) Prove that these $m + n$ eigenvectors of S are pairwise orthogonal.

Problem 20.2. Let A be a real $m \times n$ matrix of rank r and let $q = \min(m, n)$.

(1) Consider the $(m+n) \times (m+n)$ real symmetric matrix

$$S = \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}$$

and prove that

$$\begin{pmatrix} I_m & z^{-1}A \\ 0 & I_n \end{pmatrix} \begin{pmatrix} zI_m & -A \\ -A^\top & zI_n \end{pmatrix} = \begin{pmatrix} zI_m - z^{-1}AA^\top & 0 \\ -A^\top & zI_n \end{pmatrix}.$$

Use the above equation to prove that

$$\det(zI_{m+n} - S) = t^{n-m} \det(t^2 I_m - AA^\top).$$

(2) Prove that the eigenvalues of S are $\pm\sigma_1, \ldots, \pm\sigma_q$, with $|m - n|$ additional zeros.

Problem 20.3. Let B be a real bidiagonal matrix of the form

$$B = \begin{pmatrix} a_1 & b_1 & 0 & \cdots & 0 \\ 0 & a_2 & b_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{n-1} & b_{n-1} \\ 0 & 0 & \cdots & 0 & a_n \end{pmatrix}.$$

Let A be the $(2n) \times (2n)$ symmetric matrix

$$A = \begin{pmatrix} 0 & B^\top \\ B & 0 \end{pmatrix},$$

and let P be the permutation matrix given by $P = [e_1, e_{n+1}, e_2, e_{n+2}, \cdots, e_n, e_{2n}]$.

(1) Prove that $T = P^\top AP$ is a symmetric tridiagonal $(2n) \times (2n)$ matrix with zero main diagonal of the form

$$T = \begin{pmatrix} 0 & a_1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ a_1 & 0 & b_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & b_1 & 0 & a_2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & a_2 & 0 & b_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{n-1} & 0 & b_{n-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & b_{n-1} & 0 & a_n \\ 0 & 0 & 0 & \cdots & 0 & 0 & a_n & 0 \end{pmatrix}.$$

(2) Prove that if x_i is a unit eigenvector for an eigenvalue λ_i of T, then $\lambda_i = \pm\sigma_i$ where σ_i is a singular value of B, and that

$$Px_i = \frac{1}{\sqrt{2}} \begin{pmatrix} u_i \\ \pm v_i \end{pmatrix},$$

where the u_i are unit eigenvectors of $B^\top B$ and the v_i are unit eigenvectors of $BB^\top$.

Problem 20.4. Find the SVD of the matrix

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

Problem 20.5. Let $u, v \in \mathbb{R}^n$ be two nonzero vectors, and let $A = uv^\top$ be the corresponding rank 1 matrix. Prove that the nonzero singular value of A is $\|u\|_2 \|v\|_2$.

Problem 20.6. Let A be a $n \times n$ real matrix. Prove that if $\sigma_1, \ldots, \sigma_n$ are the singular values of A, then $\sigma_1^3, \ldots, \sigma_n^3$ are the singular values of $AA^\top A$.

Problem 20.7. Let A be a real $n \times n$ matrix.

(1) Prove that the largest singular value σ_1 of A is given by

$$\sigma_1 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2},$$

and that this supremum is achieved at $x = u_1$, the first column in U in an SVD $A = V\Sigma U^\top$.

(2) Extend the above result to real $m \times n$ matrices.

Problem 20.8. Let A be a real $m \times n$ matrix. Prove that if B is any submatrix of A (by keeping $M \leq m$ rows and $N \leq n$ columns of A), then $(\sigma_1)_B \leq (\sigma_1)_A$ (where $(\sigma_1)_A$ is the largest singular value of A and similarly for $(\sigma_1)_B$).

Problem 20.9. Let A be a real $n \times n$ matrix.

(1) Assume A is invertible. Prove that if $A = Q_1S_1 = Q_2S_2$ are two polar decompositions of A, then $Q_1 = Q_2$ and $S_1 = S_2$.
Hint. $A^\top A = S_1^2 = S_2^2$, with S_1 and S_2 symmetric positive definite. Then use Problem 16.7.

(2) Now assume that A is singular. Prove that if $A = Q_1S_1 = Q_2S_2$ are two polar decompositions of A, then $S_1 = S_2$, but Q_1 may not be equal to Q_2.

Problem 20.10. (1) Let A be any invertible (real) $n \times n$ matrix. Prove that for every SVD, $A = VDU^\top$ of A, the product $VU^\top$ is the same (i.e., if $V_1DU_1^\top = V_2DU_2^\top$, then $V_1U_1^\top = V_2U_2^\top$). What does $VU^\top$ have to do with the polar form of A?

(2) Given any invertible (real) $n \times n$ matrix, A, prove that there is a unique orthogonal matrix, $Q \in \mathbf{O}(n)$, such that $\|A - Q\|_F$ is minimal (under the Frobenius norm). In fact, prove that $Q = VU^\top$, where $A = VDU^\top$ is an SVD of A. Moreover, if $\det(A) > 0$, show that $Q \in \mathbf{SO}(n)$.

What can you say if A is singular (i.e., non-invertible)?

Problem 20.11. (1) Prove that for any $n \times n$ matrix A and any orthogonal matrix Q, we have

$$\max\{\mathrm{tr}(QA) \mid Q \in \mathbf{O}(n)\} = \sigma_1 + \cdots + \sigma_n,$$

where $\sigma_1 \geq \cdots \geq \sigma_n$ are the singular values of A. Furthermore, this maximum is achieved by $Q = UV^\top$, where $A = V\Sigma U^\top$ is any SVD for A.

(2) By applying the above result with $A = Z^\top X$ and $Q = R^\top$, deduce the following result : For any two fixed $n \times k$ matrices X and Z, the minimum of the set

$$\{\|X - ZR\|_F \mid R \in \mathbf{O}(k)\}$$

is achieved by $R = VU^\top$ for any SVD decomposition $V\Sigma U^\top = Z^\top X$ of $Z^\top X$.

Remark: The problem of finding an orthogonal matrix R such that ZR comes as close as possible to X is called the *orthogonal Procrustes problem*; see Strang [Strang (2019)] (Section IV.9) for the history of this problem.

Chapter 21

Applications of SVD and Pseudo-Inverses

> De tous les principes qu'on peut proposer pour cet objet, je pense qu'il n'en est pas de plus général, de plus exact, ni d'une application plus facile, que celui dont nous avons fait usage dans les recherches précédentes, et qui consiste à rendre *minimum* la somme des carrés des erreurs. Par ce moyen il s'établit entre les erreurs une sorte d'équilibre qui, empêchant les extrêmes de prévaloir, est très propre às faire connaitre l'état du système le plus proche de la vérité.
>
> —**Legendre, 1805,** *Nouvelles Méthodes pour la détermination des Orbites des Comètes*

21.1 Least Squares Problems and the Pseudo-Inverse

This chapter presents several applications of SVD. The first one is the pseudo-inverse, which plays a crucial role in solving linear systems by the method of least squares. The second application is data compression. The third application is principal component analysis (PCA), whose purpose is to identify patterns in data and understand the variance–covariance structure of the data. The fourth application is the best affine approximation of a set of data, a problem closely related to PCA.

The method of least squares is a way of "solving" an overdetermined system of linear equations

$$Ax = b,$$

i.e., a system in which A is a rectangular $m \times n$ matrix with more equations than unknowns (when $m > n$). Historically, the method of least squares was used by Gauss and Legendre to solve problems in astronomy and geodesy. The method was first published by Legendre in 1805 in a paper on methods

for determining the orbits of comets. However, Gauss had already used the method of least squares as early as 1801 to determine the orbit of the asteroid Ceres, and he published a paper about it in 1810 after the discovery of the asteroid Pallas. Incidentally, it is in that same paper that Gaussian elimination using pivots is introduced.

The reason why more equations than unknowns arise in such problems is that repeated measurements are taken to minimize errors. This produces an overdetermined and often inconsistent system of linear equations. For example, Gauss solved a system of eleven equations in six unknowns to determine the orbit of the asteroid Pallas.

Example 21.1. As a concrete illustration, suppose that we observe the motion of a small object, assimilated to a point, in the plane. From our observations, we suspect that this point moves along a straight line, say of equation $y = cx + d$. Suppose that we observed the moving point at three different locations (x_1, y_1), (x_2, y_2), and (x_3, y_3). Then we should have

$$\begin{aligned} d + cx_1 &= y_1, \\ d + cx_2 &= y_2, \\ d + cx_3 &= y_3. \end{aligned}$$

If there were no errors in our measurements, these equations would be compatible, and c and d would be determined by only two of the equations. However, in the presence of errors, the system may be inconsistent. Yet we would like to find c and d!

The idea of the method of least squares is to determine (c, d) such that it minimizes the sum of the squares of the errors, namely,

$$(d + cx_1 - y_1)^2 + (d + cx_2 - y_2)^2 + (d + cx_3 - y_3)^2.$$

See Figure 21.1.

In general, for an overdetermined $m \times n$ system $Ax = b$, what Gauss and Legendre discovered is that there are solutions x minimizing

$$\|Ax - b\|_2^2$$

(where $\|u\|_2^2 = u_1^2 + \cdots + u_n^2$, the square of the Euclidean norm of the vector $u = (u_1, \ldots, u_n)$), and that these solutions are given by the square $n \times n$ system

$$A^\top Ax = A^\top b,$$

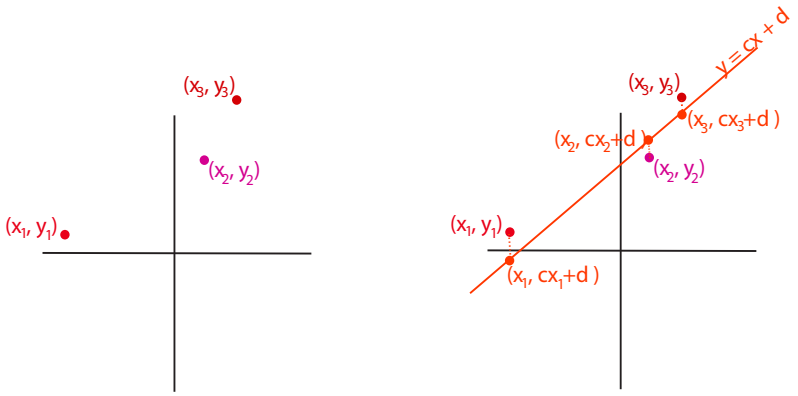

Fig. 21.1 Given three points (x_1, y_1), (x_2, y_2), (x_3, y_3), we want to determine the line $y = cx + d$ which minimizes the lengths of the dashed vertical lines.

called the *normal equations*. Furthermore, when the columns of A are linearly independent, it turns out that $A^\top A$ is invertible, and so x is unique and given by

$$x = (A^\top A)^{-1} A^\top b.$$

Note that $A^\top A$ is a symmetric matrix, one of the nice features of the normal equations of a least squares problem. For instance, since the above problem in matrix form is represented as

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix} \begin{pmatrix} d \\ c \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix},$$

the normal equations are

$$\begin{pmatrix} 3 & x_1 + x_2 + x_3 \\ x_1 + x_2 + x_3 & x_1^2 + x_2^2 + x_3^2 \end{pmatrix} \begin{pmatrix} d \\ c \end{pmatrix} = \begin{pmatrix} y_1 + y_2 + y_3 \\ x_1 y_1 + x_2 y_2 + x_3 y_3 \end{pmatrix}.$$

In fact, given any real $m \times n$ matrix A, there is always a unique x^+ of minimum norm that minimizes $\|Ax - b\|_2^2$, even when the columns of A are linearly dependent. How do we prove this, and how do we find x^+?

Theorem 21.1. *Every linear system $Ax = b$, where A is an $m \times n$ matrix, has a unique least squares solution x^+ of smallest norm.*

Proof. Geometry offers a nice proof of the existence and uniqueness of x^+. Indeed, we can interpret b as a point in the Euclidean (affine) space $\mathbb{R}^m$, and the image subspace of A (also called the column space of A) as a subspace U of $\mathbb{R}^m$ (passing through the origin). Then it is clear that

$$\inf_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 = \inf_{y \in U} \|y - b\|_2^2,$$

with $U = \operatorname{Im} A$, and we claim that x minimizes $\|Ax - b\|_2^2$ iff $Ax = p$, where p the orthogonal projection of b onto the subspace U.

Recall from Section 12.1 that the orthogonal projection $p_U \colon U \oplus U^\perp \to U$ is the linear map given by

$$p_U(u + v) = u,$$

with $u \in U$ and $v \in U^\perp$. If we let $p = p_U(b) \in U$, then for any point $y \in U$, the vectors $\overrightarrow{py} = y - p \in U$ and $\overrightarrow{bp} = p - b \in U^\perp$ are orthogonal, which implies that

$$\|\overrightarrow{by}\|_2^2 = \|\overrightarrow{bp}\|_2^2 + \|\overrightarrow{py}\|_2^2,$$

where $\overrightarrow{by} = y - b$. Thus, p is indeed the unique point in U that minimizes the distance from b to any point in U. See Figure 21.2.

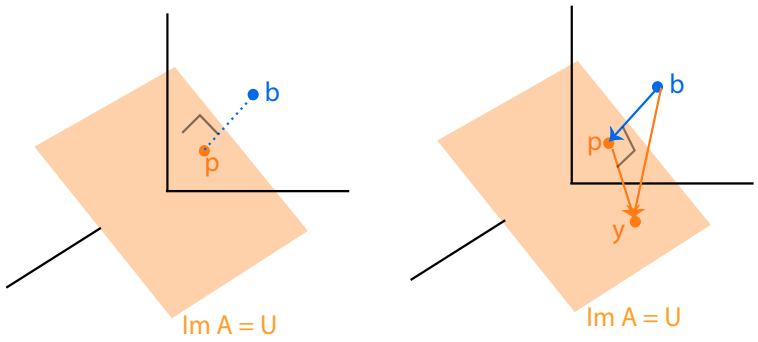

Fig. 21.2 Given a 3×2 matrix A, $U = \operatorname{Im} A$ is the peach plane in $\mathbb{R}^3$ and p is the orthogonal projection of b onto U. Furthermore, given $y \in U$, the points b, y, and p are the vertices of a right triangle.

Thus the problem has been reduced to proving that there is a unique x^+ of minimum norm such that $Ax^+ = p$, with $p = p_U(b) \in U$, the orthogonal

projection of b onto U. We use the fact that

$$\mathbb{R}^n = \operatorname{Ker} A \oplus (\operatorname{Ker} A)^\perp.$$

Consequently, every $x \in \mathbb{R}^n$ can be written uniquely as $x = u + v$, where $u \in \operatorname{Ker} A$ and $v \in (\operatorname{Ker} A)^\perp$, and since u and v are orthogonal,

$$\|x\|_2^2 = \|u\|_2^2 + \|v\|_2^2.$$

Furthermore, since $u \in \operatorname{Ker} A$, we have $Au = 0$, and thus $Ax = p$ iff $Av = p$, which shows that the solutions of $Ax = p$ for which x has minimum norm must belong to $(\operatorname{Ker} A)^\perp$. However, the restriction of A to $(\operatorname{Ker} A)^\perp$ is injective. This is because if $Av_1 = Av_2$, where $v_1, v_2 \in (\operatorname{Ker} A)^\perp$, then $A(v_2 - v_2) = 0$, which implies $v_2 - v_1 \in \operatorname{Ker} A$, and since $v_1, v_2 \in (\operatorname{Ker} A)^\perp$, we also have $v_2 - v_1 \in (\operatorname{Ker} A)^\perp$, and consequently, $v_2 - v_1 = 0$. This shows that there is a unique x^+ of minimum norm such that $Ax^+ = p$, and that x^+ must belong to $(\operatorname{Ker} A)^\perp$. By our previous reasoning, x^+ is the unique vector of minimum norm minimizing $\|Ax - b\|_2^2$. □

The proof also shows that x minimizes $\|Ax - b\|_2^2$ iff $\overrightarrow{pb} = b - Ax$ is orthogonal to U, which can be expressed by saying that $b - Ax$ is orthogonal to every column of A. However, this is equivalent to

$$A^\top(b - Ax) = 0, \quad \text{i.e.,} \quad A^\top Ax = A^\top b.$$

Finally, it turns out that the minimum norm least squares solution x^+ can be found in terms of the pseudo-inverse A^+ of A, which is itself obtained from any SVD of A.

Definition 21.1. Given any nonzero $m \times n$ matrix A of rank r, if $A = VDU^\top$ is an SVD of A such that

$$D = \begin{pmatrix} \Lambda & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix},$$

with

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$$

an $r \times r$ diagonal matrix consisting of the nonzero singular values of A, then if we let D^+ be the $n \times m$ matrix

$$D^+ = \begin{pmatrix} \Lambda^{-1} & 0_{r,m-r} \\ 0_{n-r,r} & 0_{n-r,m-r} \end{pmatrix},$$

with

$$\Lambda^{-1} = \operatorname{diag}(1/\lambda_1, \ldots, 1/\lambda_r),$$

the *pseudo-inverse* of A is defined by

$$A^+ = UD^+V^\top.$$

If $A = 0_{m,n}$ is the zero matrix, we set $A^+ = 0_{n,m}$. Observe that D^+ is obtained from D by inverting the nonzero diagonal entries of D, leaving all zeros in place, and then transposing the matrix. For example, given the matrix

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

its pseudo-inverse is

$$D^+ = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The pseudo-inverse of a matrix is also known as the *Moore–Penrose pseudo-inverse*.

Actually, it seems that A^+ depends on the specific choice of U and V in an SVD (U, D, V) for A, but the next theorem shows that this is not so.

Theorem 21.2. *The least squares solution of smallest norm of the linear system $Ax = b$, where A is an $m \times n$ matrix, is given by*

$$x^+ = A^+ b = UD^+ V^\top b.$$

Proof. First assume that A is a (rectangular) diagonal matrix D, as above. Then since x minimizes $\|Dx - b\|_2^2$ iff Dx is the projection of b onto the image subspace F of D, it is fairly obvious that $x^+ = D^+ b$. Otherwise, we can write

$$A = VDU^\top,$$

where U and V are orthogonal. However, since V is an isometry,

$$\|Ax - b\|_2 = \|VDU^\top x - b\|_2 = \|DU^\top x - V^\top b\|_2.$$

Letting $y = U^\top x$, we have $\|x\|_2 = \|y\|_2$, since U is an isometry, and since U is surjective, $\|Ax - b\|_2$ is minimized iff $\|Dy - V^\top b\|_2$ is minimized, and we have shown that the least solution is

$$y^+ = D^+ V^\top b.$$

Since $y = U^\top x$, with $\|x\|_2 = \|y\|_2$, we get

$$x^+ = UD^+ V^\top b = A^+ b.$$

Thus, the pseudo-inverse provides the optimal solution to the least squares problem. □

By Theorem 21.2 and Theorem 21.1, A^+b is uniquely defined by every b, and thus A^+ depends only on A.

The `Matlab` command for computing the pseudo-inverse B of the matrix A is
`B = pinv(A)`.

Example 21.2. If A is the rank 2 matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{pmatrix}$$

whose eigenvalues are $-1.1652, 0, 0, 17.1652$, using `Matlab` we obtain the SVD $A = VDU^\top$ with

$$U = \begin{pmatrix} -0.3147 & 0.7752 & 0.2630 & -0.4805 \\ -0.4275 & 0.3424 & 0.0075 & 0.8366 \\ -0.5402 & -0.0903 & -0.8039 & -0.2319 \\ -0.6530 & -0.5231 & 0.5334 & -0.1243 \end{pmatrix},$$

$$V = \begin{pmatrix} -0.3147 & -0.7752 & 0.5452 & 0.0520 \\ -0.4275 & -0.3424 & -0.7658 & 0.3371 \\ -0.5402 & 0.0903 & -0.1042 & -0.8301 \\ -0.6530 & 0.5231 & 0.3247 & 0.4411 \end{pmatrix}, \quad D = \begin{pmatrix} 17.1652 & 0 & 0 & 0 \\ 0 & 1.1652 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then

$$D^+ = \begin{pmatrix} 0.0583 & 0 & 0 & 0 \\ 0 & 0.8583 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and

$$A^+ = UD^+V^\top = \begin{pmatrix} -0.5100 & -0.2200 & 0.0700 & 0.3600 \\ -0.2200 & -0.0900 & 0.0400 & 0.1700 \\ 0.0700 & 0.0400 & 0.0100 & -0.0200 \\ 0.3600 & 0.1700 & -0.0200 & -0.2100 \end{pmatrix},$$

which is also the result obtained by calling `pinv(A)`.

If A is an $m \times n$ matrix of rank n (and so $m \geq n$), it is immediately shown that the QR-decomposition in terms of Householder transformations applies as follows:

There are n $m \times m$ matrices $H_1, \ldots, H_n$, Householder matrices or the identity, and an upper triangular $m \times n$ matrix R of rank n such that

$$A = H_1 \cdots H_n R.$$

Then because each H_i is an isometry,

$$\|Ax - b\|_2 = \|Rx - H_n \cdots H_1 b\|_2,$$

and the least squares problem $Ax = b$ is equivalent to the system

$$Rx = H_n \cdots H_1 b.$$

Now the system

$$Rx = H_n \cdots H_1 b$$

is of the form

$$\begin{pmatrix} R_1 \\ 0_{m-n} \end{pmatrix} x = \begin{pmatrix} c \\ d \end{pmatrix},$$

where R_1 is an invertible $n \times n$ matrix (since A has rank n), $c \in \mathbb{R}^n$, and $d \in \mathbb{R}^{m-n}$, and the least squares solution of smallest norm is

$$x^+ = R_1^{-1} c.$$

Since R_1 is a triangular matrix, it is very easy to invert R_1.

The method of least squares is one of the most effective tools of the mathematical sciences. There are entire books devoted to it. Readers are advised to consult Strang [Strang (1988)], Golub and Van Loan [Golub and Van Loan (1996)], Demmel [Demmel (1997)], and Trefethen and Bau [Trefethen and Bau III (1997)], where extensions and applications of least squares (such as weighted least squares and recursive least squares) are described. Golub and Van Loan [Golub and Van Loan (1996)] also contains a very extensive bibliography, including a list of books on least squares.

21.2 Properties of the Pseudo-Inverse

We begin this section with a proposition which provides a way to calculate the pseudo-inverse of an $m \times n$ matrix A without first determining an SVD factorization.

Proposition 21.1. *When A has full rank, the pseudo-inverse A^+ can be expressed as $A^+ = (A^\top A)^{-1} A^\top$ when $m \geq n$, and as $A^+ = A^\top (AA^\top)^{-1}$ when $n \geq m$. In the first case ($m \geq n$), observe that $A^+ A = I$, so A^+ is a left inverse of A; in the second case ($n \geq m$), we have $AA^+ = I$, so A^+ is a right inverse of A.*

Proof. If $m \geq n$ and A has full rank n, we have

$$A = V \begin{pmatrix} \Lambda \\ 0_{m-n,n} \end{pmatrix} U^\top$$

with Λ an $n \times n$ diagonal invertible matrix (with positive entries), so

$$A^+ = U \begin{pmatrix} \Lambda^{-1} & 0_{n,m-n} \end{pmatrix} V^\top.$$

We find that

$$A^\top A = U \begin{pmatrix} \Lambda & 0_{n,m-n} \end{pmatrix} V^\top V \begin{pmatrix} \Lambda \\ 0_{m-n,n} \end{pmatrix} U^\top = U \Lambda^2 U^\top,$$

which yields

$$(A^\top A)^{-1} A^\top = U \Lambda^{-2} U^\top U \begin{pmatrix} \Lambda & 0_{n,m-n} \end{pmatrix} V^\top = U \begin{pmatrix} \Lambda^{-1} & 0_{n,m-n} \end{pmatrix} V^\top = A^+.$$

Therefore, if $m \geq n$ and A has full rank n, then

$$A^+ = (A^\top A)^{-1} A^\top.$$

If $n \geq m$ and A has full rank m, then

$$A = V \begin{pmatrix} \Lambda & 0_{m,n-m} \end{pmatrix} U^\top$$

with Λ an $m \times m$ diagonal invertible matrix (with positive entries), so

$$A^+ = U \begin{pmatrix} \Lambda^{-1} \\ 0_{n-m,m} \end{pmatrix} V^\top.$$

We find that

$$A A^\top = V \begin{pmatrix} \Lambda & 0_{m,n-m} \end{pmatrix} U^\top U \begin{pmatrix} \Lambda \\ 0_{n-m,m} \end{pmatrix} V^\top = V \Lambda^2 V^\top,$$

which yields

$$A^\top (A A^\top)^{-1} = U \begin{pmatrix} \Lambda \\ 0_{n-m,m} \end{pmatrix} V^\top V \Lambda^{-2} V^\top = U \begin{pmatrix} \Lambda^{-1} \\ 0_{n-m,m} \end{pmatrix} V^\top = A^+.$$

Therefore, if $n \geq m$ and A has full rank m, then $A^+ = A^\top (A A^\top)^{-1}$. □

For example, if $A = \begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 0 & 1 \end{pmatrix}$, then A has rank 2 and since $m \geq n$, $A^+ = (A^\top A)^{-1} A^\top$ where

$$A^+ = \begin{pmatrix} 5 & 8 \\ 8 & 14 \end{pmatrix}^{-1} A^\top = \begin{pmatrix} 7/3 & -4/3 \\ 4/3 & 5/6 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 2 & 3 & 1 \end{pmatrix} = \begin{pmatrix} -1/3 & 2/3 & -4/3 \\ 1/3 & -1/6 & 5/6 \end{pmatrix}.$$

If $A = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 0 & 1 & 1 & -1 \end{pmatrix}$, since A has rank 2 and $n \geq m$, then $A^+ = A^\top (AA^\top)^{-1}$ where

$$A^+ = A^\top \begin{pmatrix} 14 & 5 \\ 5 & 3 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \\ 3 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 3/17 & -5/17 \\ -5/17 & 14/17 \end{pmatrix} = \begin{pmatrix} 3/17 & -5/17 \\ 1/17 & 4/17 \\ 4/17 & -1/17 \\ 5/17 & -14/17 \end{pmatrix}.$$

Let $A = V\Sigma U^\top$ be an SVD for any $m \times n$ matrix A. It is easy to check that both AA^+ and A^+A are symmetric matrices. In fact,

$$AA^+ = V\Sigma U^\top U\Sigma^+ V^\top = V\Sigma\Sigma^+ V^\top = V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} V^\top$$

and

$$A^+A = U\Sigma^+ V^\top V\Sigma U^\top = U\Sigma^+\Sigma U^\top = U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} U^\top.$$

From the above expressions we immediately deduce that

$$AA^+A = A,$$
$$A^+AA^+ = A^+,$$

and that

$$(AA^+)^2 = AA^+,$$
$$(A^+A)^2 = A^+A,$$

so both AA^+ and A^+A are orthogonal projections (since they are both symmetric).

Proposition 21.2. *The matrix AA^+ is the orthogonal projection onto the range of A and A^+A is the orthogonal projection onto* $\mathrm{Ker}(A)^\perp = \mathrm{Im}(A^\top)$*, the range of $A^\top$.*

Proof. Obviously, we have $\mathrm{range}(AA^+) \subseteq \mathrm{range}(A)$, and for any $y = Ax \in \mathrm{range}(A)$, since $AA^+A = A$, we have

$$AA^+y = AA^+Ax = Ax = y,$$

so the image of AA^+ is indeed the range of A. It is also clear that $\mathrm{Ker}(A) \subseteq \mathrm{Ker}(A^+A)$, and since $AA^+A = A$, we also have $\mathrm{Ker}(A^+A) \subseteq \mathrm{Ker}(A)$, and so

$$\mathrm{Ker}(A^+A) = \mathrm{Ker}(A).$$

Since A^+A is symmetric, $\mathrm{range}(A^+A) = \mathrm{range}((A^+A)^\top) = \mathrm{Ker}(A^+A)^\perp = \mathrm{Ker}(A)^\perp$, as claimed. □

Proposition 21.3. *The set* $\mathrm{range}(A) = \mathrm{range}(AA^+)$ *consists of all vectors* $y \in \mathbb{R}^m$ *such that*

$$V^\top y = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

with $z \in \mathbb{R}^r$.

Proof. Indeed, if $y = Ax$, then

$$V^\top y = V^\top Ax = V^\top V \Sigma U^\top x = \Sigma U^\top x = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} U^\top x = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

where Σ_r is the $r \times r$ diagonal matrix $\mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. Conversely, if $V^\top y = \left(\begin{smallmatrix} z \\ 0 \end{smallmatrix}\right)$, then $y = V \left(\begin{smallmatrix} z \\ 0 \end{smallmatrix}\right)$, and

$$\begin{aligned} AA^+ y &= V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} V^\top y \\ &= V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} V^\top V \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= V \begin{pmatrix} z \\ 0 \end{pmatrix} = y, \end{aligned}$$

which shows that y belongs to the range of A. □

Similarly, we have the following result.

Proposition 21.4. *The set* $\mathrm{range}(A^+A) = \mathrm{Ker}(A)^\perp$ *consists of all vectors* $y \in \mathbb{R}^n$ *such that*

$$U^\top y = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

with $z \in \mathbb{R}^r$.

Proof. If $y = A^+Au$, then

$$y = A^+Au = U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} U^\top u = U \begin{pmatrix} z \\ 0 \end{pmatrix},$$

for some $z \in \mathbb{R}^r$. Conversely, if $U^\top y = \left(\begin{smallmatrix} z \\ 0 \end{smallmatrix}\right)$, then $y = U \left(\begin{smallmatrix} z \\ 0 \end{smallmatrix}\right)$, and so

$$\begin{aligned} A^+AU \begin{pmatrix} z \\ 0 \end{pmatrix} &= U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} U^\top U \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= U \begin{pmatrix} z \\ 0 \end{pmatrix} = y, \end{aligned}$$

which shows that $y \in \text{range}(A^+A)$. □

Analogous results hold for complex matrices, but in this case, V and U are unitary matrices and AA^+ and A^+A are Hermitian orthogonal projections.

If A is a normal matrix, which means that $AA^\top = A^\top A$, then there is an intimate relationship between SVD's of A and block diagonalizations of A. As a consequence, the pseudo-inverse of a normal matrix A can be obtained directly from a block diagonalization of A.

If A is a (real) normal matrix, then we know from Theorem 16.8 that A can be block diagonalized with respect to an orthogonal matrix U as

$$A = U\Lambda U^\top,$$

where Λ is the (real) block diagonal matrix

$$\Lambda = \text{diag}(B_1, \ldots, B_n),$$

consisting either of 2×2 blocks of the form

$$B_j = \begin{pmatrix} \lambda_j & -\mu_j \\ \mu_j & \lambda_j \end{pmatrix}$$

with $\mu_j \neq 0$, or of one-dimensional blocks $B_k = (\lambda_k)$. Then we have the following proposition:

Proposition 21.5. *For any (real) normal matrix A and any block diagonalization $A = U\Lambda U^\top$ of A as above, the pseudo-inverse of A is given by*

$$A^+ = U\Lambda^+ U^\top,$$

where Λ^+ is the pseudo-inverse of Λ. Furthermore, if

$$\Lambda = \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix},$$

where Λ_r has rank r, then

$$\Lambda^+ = \begin{pmatrix} \Lambda_r^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Proof. Assume that $B_1, \ldots, B_p$ are 2×2 blocks and that $\lambda_{2p+1}, \ldots, \lambda_n$ are the scalar entries. We know that the numbers $\lambda_j \pm i\mu_j$, and the λ_{2p+k} are the eigenvalues of A. Let $\rho_{2j-1} = \rho_{2j} = \sqrt{\lambda_j^2 + \mu_j^2} = \sqrt{\det(B_i)}$ for $j = 1, \ldots, p$, $\rho_j = |\lambda_j|$ for $j = 2p+1, \ldots, r$. Multiplying U by a suitable

permutation matrix, we may assume that the blocks of Λ are ordered so that $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_r > 0$. Then it is easy to see that

$$AA^\top = A^\top A = U\Lambda U^\top U\Lambda^\top U^\top = U\Lambda\Lambda^\top U^\top,$$

with

$$\Lambda\Lambda^\top = \mathrm{diag}(\rho_1^2, \ldots, \rho_r^2, 0, \ldots, 0),$$

so $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_r > 0$ are the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ of A. Define the diagonal matrix

$$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0),$$

where $r = \mathrm{rank}(A)$, $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and the block diagonal matrix Θ defined such that the block B_i in Λ is replaced by the block $\sigma^{-1}B_i$ where $\sigma = \sqrt{\det(B_i)}$, the nonzero scalar λ_j is replaced $\lambda_j/|\lambda_j|$, and a diagonal zero is replaced by 1. Observe that Θ is an orthogonal matrix and

$$\Lambda = \Theta\Sigma.$$

But then we can write

$$A = U\Lambda U^\top = U\Theta\Sigma U^\top,$$

and we if let $V = U\Theta$, since U is orthogonal and Θ is also orthogonal, V is also orthogonal and $A = V\Sigma U^\top$ *is an SVD for* A. Now we get

$$A^+ = U\Sigma^+ V^\top = U\Sigma^+\Theta^\top U^\top.$$

However, since Θ is an orthogonal matrix, $\Theta^\top = \Theta^{-1}$, and a simple calculation shows that

$$\Sigma^+\Theta^\top = \Sigma^+\Theta^{-1} = \Lambda^+,$$

which yields the formula

$$A^+ = U\Lambda^+ U^\top.$$

Also observe that Λ_r is invertible and

$$\Lambda^+ = \begin{pmatrix} \Lambda_r^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Therefore, the pseudo-inverse of a normal matrix can be computed directly from any block diagonalization of A, as claimed. □

Example 21.3. Consider the following real diagonal form of the normal matrix

$$A = \begin{pmatrix} -2.7500 & 2.1651 & -0.8660 & 0.5000 \\ 2.1651 & -0.2500 & -1.5000 & 0.8660 \\ 0.8660 & 1.5000 & 0.7500 & -0.4330 \\ -0.5000 & -0.8660 & -0.4330 & 0.2500 \end{pmatrix} = U\Lambda U^\top,$$

with

$$U = \begin{pmatrix} \cos(\pi/3) & 0 & \sin(\pi/3) & 0 \\ \sin(\pi/3) & 0 & -\cos(\pi/3) & 0 \\ 0 & \cos(\pi/6) & 0 & \sin(\pi/6) \\ 0 & -\cos(\pi/6) & 0 & \sin(\pi/6) \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 1 & -2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

We obtain

$$\Lambda^+ = \begin{pmatrix} 1/5 & 2/5 & 0 & 0 \\ -2/5 & 1/5 & 0 & 0 \\ 0 & 0 & -1/4 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and the pseudo-inverse of A is

$$A^+ = U\Lambda^+ U^\top = \begin{pmatrix} -0.1375 & 0.1949 & 0.1732 & -0.1000 \\ 0.1949 & 0.0875 & 0.3000 & -0.1732 \\ -0.1732 & -0.3000 & 0.1500 & -0.0866 \\ 0.1000 & 0.1732 & -0.0866 & 0.0500 \end{pmatrix},$$

which agrees with `pinv(A)`.

The following properties, due to Penrose, characterize the pseudo-inverse of a matrix. We have already proved that the pseudo-inverse satisfies these equations. For a proof of the converse, see Kincaid and Cheney [Kincaid and Cheney (1996)].

Proposition 21.6. *Given any $m \times n$ matrix A (real or complex), the pseudo-inverse A^+ of A is the unique $n \times m$ matrix satisfying the following properties:*

$$\begin{aligned} AA^+A &= A, \\ A^+AA^+ &= A^+, \\ (AA^+)^\top &= AA^+, \\ (A^+A)^\top &= A^+A. \end{aligned}$$

21.3 Data Compression and SVD

Among the many applications of SVD, a very useful one is *data compression*, notably for images. In order to make precise the notion of closeness of matrices, we use the notion of *matrix norm*. This concept is defined in Chapter 8, and the reader may want to review it before reading any further.

Given an $m \times n$ matrix of rank r, we would like to find a best approximation of A by a matrix B of rank $k \leq r$ (actually, $k < r$) such that $\|A - B\|_2$ (or $\|A - B\|_F$) is minimized. The following proposition is known as the *Eckart–Young theorem*.

Proposition 21.7. *Let A be an $m \times n$ matrix of rank r and let $V D U^\top = A$ be an SVD for A. Write u_i for the columns of U, v_i for the columns of V, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$ for the singular values of A ($p = \min(m, n)$). Then a matrix of rank $k < r$ closest to A (in the $\| \ \|_2$ norm) is given by*

$$A_k = \sum_{i=1}^{k} \sigma_i v_i u_i^\top = V \operatorname{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) U^\top$$

and $\|A - A_k\|_2 = \sigma_{k+1}$.

Proof. By construction, A_k has rank k, and we have

$$\begin{aligned} \|A - A_k\|_2 &= \Big\| \sum_{i=k+1}^{p} \sigma_i v_i u_i^\top \Big\|_2 \\ &= \big\| V \operatorname{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p) U^\top \big\|_2 = \sigma_{k+1}. \end{aligned}$$

It remains to show that $\|A - B\|_2 \geq \sigma_{k+1}$ for all rank k matrices B. Let B be any rank k matrix, so its kernel has dimension $n - k$. The subspace U_{k+1} spanned by $(u_1, \ldots, u_{k+1})$ has dimension $k + 1$, and because the sum of the dimensions of the kernel of B and of U_{k+1} is $(n - k) + k + 1 = n + 1$, these two subspaces must intersect in a subspace of dimension at least 1. Pick any unit vector h in $\operatorname{Ker}(B) \cap U_{k+1}$. Then since $Bh = 0$, and since U and V are isometries, we have

$$\begin{aligned} \|A - B\|_2^2 \geq \|(A - B)h\|_2^2 = \|Ah\|_2^2 = \big\|V D U^\top h\big\|_2^2 &= \big\|D U^\top h\big\|_2^2 \\ &\geq \sigma_{k+1}^2 \big\|U^\top h\big\|_2^2 = \sigma_{k+1}^2, \end{aligned}$$

which proves our claim. □

Note that A_k can be stored using $(m+n)k$ entries, as opposed to mn entries. When $k \ll m$, this is a substantial gain.

Example 21.4. Consider the badly conditioned symmetric matrix

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}$$

from Section 8.5. Since A is SPD, we have the SVD

$$A = UDU^\top,$$

with

$$U = \begin{pmatrix} -0.5286 & -0.6149 & 0.3017 & -0.5016 \\ -0.3803 & -0.3963 & -0.0933 & 0.8304 \\ -0.5520 & 0.2716 & -0.7603 & -0.2086 \\ -0.5209 & 0.6254 & 0.5676 & 0.1237 \end{pmatrix},$$

$$D = \begin{pmatrix} 30.2887 & 0 & 0 & 0 \\ 0 & 3.8581 & 0 & 0 \\ 0 & 0 & 0.8431 & 0 \\ 0 & 0 & 0 & 0.0102 \end{pmatrix}.$$

If we set $\sigma_3 = \sigma_4 = 0$, we obtain the best rank 2 approximation

$$A_2 = U(:, 1:2) * D(:, 1:2) * U(:, 1:2)' = \begin{pmatrix} 9.9207 & 7.0280 & 8.1923 & 6.8563 \\ 7.0280 & 4.9857 & 5.9419 & 5.0436 \\ 8.1923 & 5.9419 & 9.5122 & 9.3641 \\ 6.8563 & 5.0436 & 9.3641 & 9.7282 \end{pmatrix}.$$

A nice example of the use of Proposition 21.7 in image compression is given in Demmel [Demmel (1997)], Chapter 3, Section 3.2.3, pages 113–115; see the Matlab demo.

Proposition 21.7 also holds for the Frobenius norm; see Problem 21.4.

An interesting topic that we have not addressed is the actual computation of an SVD. This is a very interesting but tricky subject. Most methods reduce the computation of an SVD to the diagonalization of a well-chosen symmetric matrix which is not $A^\top A$; see Problem 20.1 and Problem 20.3. Interested readers should read Section 5.4 of Demmel's excellent book [Demmel (1997)], which contains an overview of most known methods and an extensive list of references.

21.4 Principal Components Analysis (PCA)

Suppose we have a set of data consisting of n points $X_1, \ldots, X_n$, with each $X_i \in \mathbb{R}^d$ *viewed as a row vector*. Think of the X_i's as persons, and if $X_i = (x_{i\,1}, \ldots, x_{i\,d})$, each $x_{i\,j}$ is the value of some *feature* (or *attribute*) of that person.

Example 21.5. For example, the X_i's could be mathematicians, $d = 2$, and the first component, $x_{i\,1}$, of X_i could be the year that X_i was born, and the second component, $x_{i\,2}$, the length of the beard of X_i in centimeters. Here is a small data set.

Name	year	length
Carl Friedrich Gauss	1777	0
Camille Jordan	1838	12
Adrien-Marie Legendre	1752	0
Bernhard Riemann	1826	15
David Hilbert	1862	2
Henri Poincaré	1854	5
Emmy Noether	1882	0
Karl Weierstrass	1815	0
Eugenio Beltrami	1835	2
Hermann Schwarz	1843	20

We usually form the $n \times d$ matrix X whose ith row is X_i, with $1 \leq i \leq n$. Then the jth column is denoted by C_j ($1 \leq j \leq d$). It is sometimes called a *feature vector*, but this terminology is far from being universally accepted. In fact, many people in computer vision call the data points X_i feature vectors!

The purpose of *principal components analysis*, for short *PCA*, is to identify patterns in data and understand the *variance–covariance* structure of the data. This is useful for the following tasks:

(1) Data reduction: Often much of the variability of the data can be accounted for by a smaller number of *principal components*.
(2) Interpretation: PCA can show relationships that were not previously suspected.

Given a vector (a *sample* of measurements) $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$,

recall that the *mean* (or *average*) $\overline{x}$ of x is given by

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

We let $x - \overline{x}$ denote the *centered data point*

$$x - \overline{x} = (x_1 - \overline{x}, \ldots, x_n - \overline{x}).$$

In order to *measure the spread* of the x_i's around the mean, we define the *sample variance* (for short, *variance*) $\mathrm{var}(x)$ (or s^2) of the sample x by

$$\mathrm{var}(x) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}.$$

Example 21.6. If $x = (1, 3, -1)$, $\overline{x} = \frac{1+3-1}{3} = 1$, $x - \overline{x} = (0, 2, -2)$, and $\mathrm{var}(x) = \frac{0^2+2^2+(-2)^2}{2} = 4$. If $y = (1, 2, 3)$, $\overline{y} = \frac{1+2+3}{3} = 2$, $y - \overline{y} = (-1, 0, 1)$, and $\mathrm{var}(y) = \frac{(-1)^2+0^2+1^2}{2} = 2$.

There is a reason for using $n-1$ instead of n. The above definition makes $\mathrm{var}(x)$ an unbiased estimator of the variance of the random variable being sampled. However, we don't need to worry about this. Curious readers will find an explanation of these peculiar definitions in Epstein [Epstein (2007)] (Chapter 14, Section 14.5) or in any decent statistics book.

Given two vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, the *sample covariance* (for short, *covariance*) of x and y is given by

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}.$$

Example 21.7. If we take $x = (1, 3, -1)$ and $y = (0, 2, -2)$, we know from Example 21.6 that $x - \overline{x} = (0, 2, -2)$ and $y - \overline{y} = (-1, 0, 1)$. Thus, $\mathrm{cov}(x, y) = \frac{0(-1)+2(0)+(-2)(1)}{2} = -1$.

The covariance of x and y measures how x and y vary from the mean with respect to each other. Obviously, $\mathrm{cov}(x, y) = \mathrm{cov}(y, x)$ and $\mathrm{cov}(x, x) = \mathrm{var}(x)$.

Note that

$$\mathrm{cov}(x, y) = \frac{(x - \overline{x})^\top (y - \overline{y})}{n-1}.$$

We say that x and y are *uncorrelated* iff $\mathrm{cov}(x, y) = 0$.

Finally, given an $n \times d$ matrix X of n points X_i, for PCA to be meaningful, it will be necessary to translate the origin to the *centroid* (or *center of gravity*) μ of the X_i's, defined by

$$\mu = \frac{1}{n}(X_1 + \cdots + X_n).$$

Observe that if $\mu = (\mu_1, \ldots, \mu_d)$, then μ_j is the mean of the vector C_j (the jth column of X).

We let $X - \mu$ denote the *matrix* whose ith row is the centered data point $X_i - \mu$ $(1 \leq i \leq n)$. Then the *sample covariance matrix* (for short, *covariance matrix*) of X is the $d \times d$ symmetric matrix

$$\Sigma = \frac{1}{n-1}(X - \mu)^\top (X - \mu) = (\mathrm{cov}(C_i, C_j)).$$

Example 21.8. Let $X = \begin{pmatrix} 1 & 1 \\ 3 & 2 \\ -1 & 3 \end{pmatrix}$, the 3×2 matrix whose columns are the vector x and y of Example 21.6. Then

$$\mu = \frac{1}{3}\left[(1,1) + (3,2) + (-1,3)\right] = (1,2),$$

$$X - \mu = \begin{pmatrix} 0 & -1 \\ 2 & 0 \\ -2 & 1 \end{pmatrix},$$

and

$$\Sigma = \frac{1}{2}\begin{pmatrix} 0 & 2 & -2 \\ -1 & 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 2 & 0 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix}.$$

Remark: The factor $\frac{1}{n-1}$ is irrelevant for our purposes and can be ignored.

Example 21.9. Here is the matrix $X - \mu$ in the case of our bearded mathematicians: since

$$\mu_1 = 1828.4, \quad \mu_2 = 5.6,$$

we get the following centered data set.

Name	year	length
Carl Friedrich Gauss	−51.4	−5.6
Camille Jordan	9.6	6.4
Adrien-Marie Legendre	−76.4	−5.6
Bernhard Riemann	−2.4	9.4
David Hilbert	33.6	−3.6
Henri Poincaré	25.6	−0.6
Emmy Noether	53.6	−5.6
Karl Weierstrass	13.4	−5.6
Eugenio Beltrami	6.6	−3.6
Hermann Schwarz	14.6	14.4

See Figure 21.3.

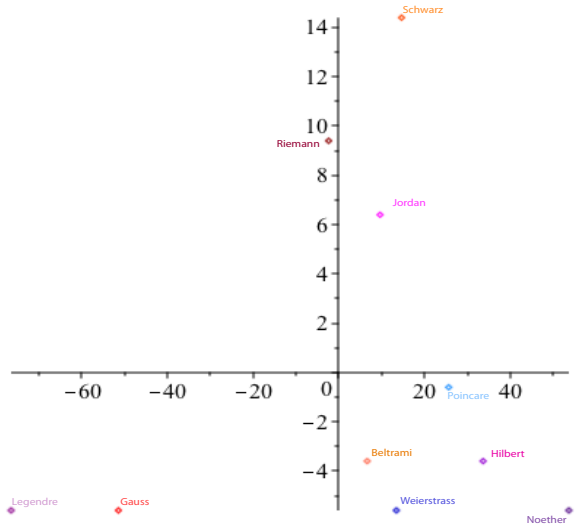

Fig. 21.3 The centered data points of Example 21.9.

We can think of the vector C_j as representing the features of X in the direction e_j (the jth canonical basis vector in $\mathbb{R}^d$, namely $e_j = (0, \ldots, 1, \ldots 0)$, with a 1 in the jth position).

If $v \in \mathbb{R}^d$ is a unit vector, we wish to consider the projection of the data points $X_1, \ldots, X_n$ onto the line spanned by v. Recall from Euclidean geometry that if $x \in \mathbb{R}^d$ is any vector and $v \in \mathbb{R}^d$ is a unit vector, the projection of x onto the line spanned by v is

$$\langle x, v\rangle v.$$

Thus, with respect to the basis v, the projection of x has coordinate $\langle x, v\rangle$. If x is represented by a row vector and v by a column vector, then

$$\langle x, v\rangle = xv.$$

Therefore, the vector $Y \in \mathbb{R}^n$ consisting of the coordinates of the projections of $X_1, \ldots, X_n$ onto the line spanned by v is given by $Y = Xv$, and this is the linear combination

$$Xv = v_1 C_1 + \cdots + v_d C_d$$

of the columns of X (with $v = (v_1, \ldots, v_d)$).

Observe that because μ_j is the mean of the vector C_j (the jth column of X), we get

$$\overline{Y} = \overline{Xv} = v_1\mu_1 + \cdots + v_d\mu_d,$$

and so the centered point $Y - \overline{Y}$ is given by

$$Y - \overline{Y} = v_1(C_1 - \mu_1) + \cdots + v_d(C_d - \mu_d) = (X - \mu)v.$$

Furthermore, if $Y = Xv$ and $Z = Xw$, then

$$\begin{aligned}\operatorname{cov}(Y, Z) &= \frac{((X - \mu)v)^\top (X - \mu)w}{n - 1} \\ &= v^\top \frac{1}{n - 1}(X - \mu)^\top (X - \mu)w \\ &= v^\top \Sigma w,\end{aligned}$$

where Σ is the covariance matrix of X. Since $Y - \overline{Y}$ has zero mean, we have

$$\operatorname{var}(Y) = \operatorname{var}(Y - \overline{Y}) = v^\top \frac{1}{n - 1}(X - \mu)^\top (X - \mu)v.$$

The above suggests that we should move the origin to the centroid μ of the X_i's and consider the matrix $X - \mu$ of the centered data points $X_i - \mu$.

From now on beware that we denote the columns of $X - \mu$ by $C_1, \ldots, C_d$ and that Y denotes the *centered* point $Y = (X - \mu)v = \sum_{j=1}^d v_j C_j$, where v is a unit vector.

Basic idea of PCA: The principal components of X are *uncorrelated* projections Y of the data points $X_1, \ldots, X_n$ onto some directions v (where the v's are unit vectors) such that $\mathrm{var}(Y)$ is maximal. This suggests the following definition:

Definition 21.2. Given an $n \times d$ matrix X of data points $X_1, \ldots, X_n$, if μ is the centroid of the X_i's, then a *first principal component of X (first PC)* is a centered point $Y_1 = (X - \mu)v_1$, the projection of $X_1, \ldots, X_n$ onto a direction v_1 such that $\mathrm{var}(Y_1)$ is maximized, where v_1 is a unit vector (recall that $Y_1 = (X - \mu)v_1$ is a linear combination of the C_j's, the columns of $X - \mu$).

More generally, if $Y_1, \ldots, Y_k$ are k principal components of X along some unit vectors $v_1, \ldots, v_k$, where $1 \leq k < d$, a *$(k+1)$th principal component of X ($(k+1)$th PC)* is a centered point $Y_{k+1} = (X - \mu)v_{k+1}$, the projection of $X_1, \ldots, X_n$ onto some direction v_{k+1} such that $\mathrm{var}(Y_{k+1})$ is maximized, subject to $\mathrm{cov}(Y_h, Y_{k+1}) = 0$ for all h with $1 \leq h \leq k$, and where v_{k+1} is a unit vector (recall that $Y_h = (X - \mu)v_h$ is a linear combination of the C_j's). The v_h are called *principal directions*.

The following proposition is the key to the main result about PCA. This result was already proven in Proposition 16.11 except that the eigenvalues were listed in increasing order. For the reader's convenience we prove it again.

Proposition 21.8. *If A is a symmetric $d \times d$ matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and if $(u_1, \ldots, u_d)$ is any orthonormal basis of eigenvectors of A, where u_i is a unit eigenvector associated with λ_i, then*

$$\max_{x \neq 0} \frac{x^\top A x}{x^\top x} = \lambda_1$$

(with the maximum attained for $x = u_1$) and

$$\max_{x \neq 0, x \in \{u_1, \ldots, u_k\}^\perp} \frac{x^\top A x}{x^\top x} = \lambda_{k+1}$$

(with the maximum attained for $x = u_{k+1}$), where $1 \leq k \leq d - 1$.

Proof. First observe that

$$\max_{x \neq 0} \frac{x^\top A x}{x^\top x} = \max_x \{x^\top A x \mid x^\top x = 1\},$$

and similarly,

$$\max_{x \neq 0, x \in \{u_1, \ldots, u_k\}^\perp} \frac{x^\top A x}{x^\top x} = \max_x \left\{x^\top A x \mid (x \in \{u_1, \ldots, u_k\}^\perp) \wedge (x^\top x = 1)\right\}.$$

Since A is a symmetric matrix, its eigenvalues are real and it can be diagonalized with respect to an orthonormal basis of eigenvectors, so let $(u_1, \ldots, u_d)$ be such a basis. If we write

$$x = \sum_{i=1}^{d} x_i u_i,$$

a simple computation shows that

$$x^\top A x = \sum_{i=1}^{d} \lambda_i x_i^2.$$

If $x^\top x = 1$, then $\sum_{i=1}^{d} x_i^2 = 1$, and since we assumed that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, we get

$$x^\top A x = \sum_{i=1}^{d} \lambda_i x_i^2 \leq \lambda_1 \left(\sum_{i=1}^{d} x_i^2 \right) = \lambda_1.$$

Thus,

$$\max_x \left\{ x^\top A x \mid x^\top x = 1 \right\} \leq \lambda_1,$$

and since this maximum is achieved for $e_1 = (1, 0, \ldots, 0)$, we conclude that

$$\max_x \left\{ x^\top A x \mid x^\top x = 1 \right\} = \lambda_1.$$

Next observe that $x \in \{u_1, \ldots, u_k\}^\perp$ and $x^\top x = 1$ iff $x_1 = \cdots = x_k = 0$ and $\sum_{i=1}^{d} x_i = 1$. Consequently, for such an x, we have

$$x^\top A x = \sum_{i=k+1}^{d} \lambda_i x_i^2 \leq \lambda_{k+1} \left(\sum_{i=k+1}^{d} x_i^2 \right) = \lambda_{k+1}.$$

Thus,

$$\max_x \left\{ x^\top A x \mid (x \in \{u_1, \ldots, u_k\}^\perp) \wedge (x^\top x = 1) \right\} \leq \lambda_{k+1},$$

and since this maximum is achieved for $e_{k+1} = (0, \ldots, 0, 1, 0, \ldots, 0)$ with a 1 in position $k + 1$, we conclude that

$$\max_x \left\{ x^\top A x \mid (x \in \{u_1, \ldots, u_k\}^\perp) \wedge (x^\top x = 1) \right\} = \lambda_{k+1},$$

as claimed. □

The quantity

$$\frac{x^\top Ax}{x^\top x}$$

is known as the *Rayleigh ratio* or *Rayleigh–Ritz ratio* (see Section 16.6) and Proposition 21.8 is often known as part of the *Rayleigh–Ritz theorem*.

Proposition 21.8 also holds if A is a Hermitian matrix and if we replace $x^\top Ax$ by x^*Ax and $x^\top x$ by x^*x. The proof is unchanged, since a Hermitian matrix has real eigenvalues and is diagonalized with respect to an orthonormal basis of eigenvectors (with respect to the Hermitian inner product).

We then have the following fundamental result showing how *the SVD of X yields the PCs*:

Theorem 21.3. *(SVD yields PCA) Let X be an $n\times d$ matrix of data points $X_1,\ldots,X_n$, and let μ be the centroid of the X_i's. If $X-\mu = VDU^\top$ is an SVD decomposition of $X-\mu$ and if the main diagonal of D consists of the singular values $\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_d$, then the centered points $Y_1,\ldots,Y_d$, where*

$$Y_k = (X-\mu)u_k = kth\ column\ of\ VD$$

and u_k is the kth column of U, are d principal components of X. Furthermore,

$$\mathrm{var}(Y_k) = \frac{\sigma_k^2}{n-1}$$

and $\mathrm{cov}(Y_h,Y_k) = 0$, whenever $h\neq k$ and $1\leq k,h\leq d$.

Proof. Recall that for any unit vector v, the centered projection of the points $X_1,\ldots,X_n$ onto the line of direction v is $Y=(X-\mu)v$ and that the variance of Y is given by

$$\mathrm{var}(Y) = v^\top\frac{1}{n-1}(X-\mu)^\top(X-\mu)v.$$

Since $X-\mu = VDU^\top$, we get

$$\begin{aligned}\mathrm{var}(Y) &= v^\top\frac{1}{(n-1)}(X-\mu)^\top(X-\mu)v\\ &= v^\top\frac{1}{(n-1)}UDV^\top VDU^\top v\\ &= v^\top U\frac{1}{(n-1)}D^2U^\top v.\end{aligned}$$

Similarly, if $Y = (X - \mu)v$ and $Z = (X - \mu)w$, then the covariance of Y and Z is given by

$$\mathrm{cov}(Y, Z) = v^\top U \frac{1}{(n-1)} D^2 U^\top w.$$

Obviously, $U\frac{1}{(n-1)}D^2U^\top$ is a symmetric matrix whose eigenvalues are $\frac{\sigma_1^2}{n-1} \geq \cdots \geq \frac{\sigma_d^2}{n-1}$, and the columns of U form an orthonormal basis of unit eigenvectors.

We proceed by induction on k. For the base case, $k = 1$, maximizing $\mathrm{var}(Y)$ is equivalent to maximizing

$$v^\top U \frac{1}{(n-1)} D^2 U^\top v,$$

where v is a unit vector. By Proposition 21.8, the maximum of the above quantity is the largest eigenvalue of $U\frac{1}{(n-1)}D^2U^\top$, namely $\frac{\sigma_1^2}{n-1}$, and it is achieved for u_1, the first columnn of U. Now we get

$$Y_1 = (X - \mu)u_1 = VDU^\top u_1,$$

and since the columns of U form an orthonormal basis, $U^\top u_1 = e_1 = (1, 0, \ldots, 0)$, and so Y_1 is indeed the first column of VD.

By the induction hypothesis, the centered points $Y_1, \ldots, Y_k$, where $Y_h = (X - \mu)u_h$ and $u_1, \ldots, u_k$ are the first k columns of U, are k principal components of X. Because

$$\mathrm{cov}(Y, Z) = v^\top U \frac{1}{(n-1)} D^2 U^\top w,$$

where $Y = (X - \mu)v$ and $Z = (X - \mu)w$, the condition $\mathrm{cov}(Y_h, Z) = 0$ for $h = 1, \ldots, k$ is equivalent to the fact that w belongs to the orthogonal complement of the subspace spanned by $\{u_1, \ldots, u_k\}$, and maximizing $\mathrm{var}(Z)$ subject to $\mathrm{cov}(Y_h, Z) = 0$ for $h = 1, \ldots, k$ is equivalent to maximizing

$$w^\top U \frac{1}{(n-1)} D^2 U^\top w,$$

where w is a unit vector orthogonal to the subspace spanned by $\{u_1, \ldots, u_k\}$. By Proposition 21.8, the maximum of the above quantity is the $(k+1)$th eigenvalue of $U\frac{1}{(n-1)}D^2U^\top$, namely $\frac{\sigma_{k+1}^2}{n-1}$, and it is achieved for u_{k+1}, the $(k+1)$th columnn of U. Now we get

$$Y_{k+1} = (X - \mu)u_{k+1} = VDU^\top u_{k+1},$$

and since the columns of U form an orthonormal basis, $U^\top u_{k+1} = e_{k+1}$, and Y_{k+1} is indeed the $(k+1)$th column of VD, which completes the proof of the induction step. □

The d columns $u_1, \ldots, u_d$ of U are usually called the *principal directions* of $X - \mu$ (and X). We note that not only do we have $\mathrm{cov}(Y_h, Y_k) = 0$ whenever $h \neq k$, but the directions $u_1, \ldots, u_d$ along which the data are projected are mutually orthogonal.

Example 21.10. For the centered data set of our bearded mathematicians (Example 21.9) we have $X - \mu = V\Sigma U^\top$, where Σ has two nonzero singular values, $\sigma_1 = 116.9803, \sigma_2 = 21.7812$, and with

$$U = \begin{pmatrix} 0.9995 & 0.0325 \\ 0.0325 & -0.9995 \end{pmatrix},$$

so the principal directions are $u_1 = (0.9995, 0.0325)$ and $u_2 = (0.0325, -0.9995)$. Observe that u_1 is almost the direction of the x-axis, and u_2 is almost the opposite direction of the y-axis. We also find that the projections Y_1 and Y_2 along the principal directions are

$$VD = \begin{pmatrix} -51.5550 & 3.9249 \\ 9.8031 & -6.0843 \\ -76.5417 & 3.1116 \\ -2.0929 & -9.4731 \\ 33.4651 & 4.6912 \\ 25.5669 & 1.4325 \\ 53.3894 & 7.3408 \\ 13.2107 & 6.0330 \\ 6.4794 & 3.8128 \\ 15.0607 & -13.9174 \end{pmatrix}, \quad \text{with} \quad X - \mu = \begin{pmatrix} -51.4000 & -5.6000 \\ 9.6000 & 6.4000 \\ -76.4000 & -5.6000 \\ -2.4000 & 9.4000 \\ 33.6000 & -3.6000 \\ 25.6000 & -0.6000 \\ 53.6000 & -5.6000 \\ 13.4000 & -5.6000 \\ 6.6000 & -3.6000 \\ 14.6000 & 14.4000 \end{pmatrix}.$$

See Figures 21.4, 21.5, and 21.6.

We know from our study of SVD that $\sigma_1^2, \ldots, \sigma_d^2$ are the eigenvalues of the symmetric positive semidefinite matrix $(X - \mu)^\top (X - \mu)$ and that $u_1, \ldots, u_d$ are corresponding eigenvectors. Numerically, it is preferable to use SVD on $X - \mu$ rather than to compute explicitly $(X - \mu)^\top (X - \mu)$ and then diagonalize it. Indeed, the explicit computation of $A^\top A$ from a matrix A can be numerically quite unstable, and good SVD algorithms avoid computing $A^\top A$ explicitly.

In general, since an SVD of X is not unique, *the principal directions $u_1, \ldots, u_d$ are not unique*. This can happen when a data set has some *rotational symmetries*, and in such a case, PCA is not a very good method for analyzing the data set.

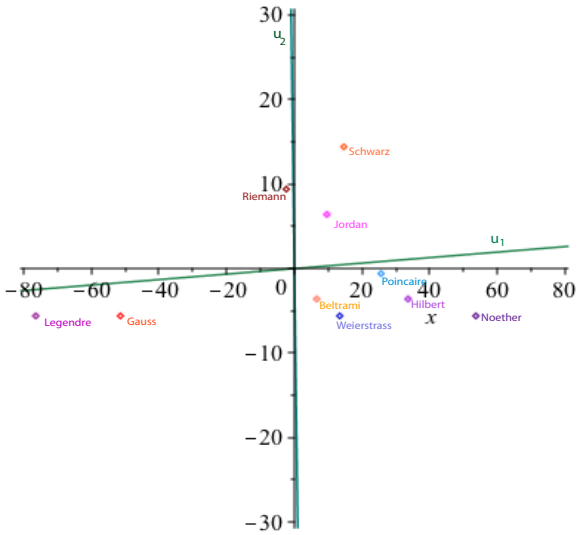

Fig. 21.4 The centered data points of Example 21.9 and the two principal directions of Example 21.10.

21.5 Best Affine Approximation

A problem very close to PCA (and based on least squares) is to *best approximate a data set of n points $X_1, \ldots, X_n$, with $X_i \in \mathbb{R}^d$, by a p-dimensional affine subspace A of $\mathbb{R}^d$, with $1 \leq p \leq d-1$* (the terminology rank $d-p$ is also used).

First consider $p = d-1$. Then $A = A_1$ is an affine hyperplane (in $\mathbb{R}^d$), and it is given by an equation of the form

$$a_1 x_1 + \cdots + a_d x_d + c = 0.$$

By *best approximation*, we mean that $(a_1, \ldots, a_d, c)$ solves the homogeneous linear system

$$\begin{pmatrix} x_{1\,1} & \cdots & x_{1\,d} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n\,1} & \cdots & x_{n\,d} & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_d \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

in the *least squares sense, subject to the condition that $a = (a_1, \ldots, a_d)$ is a unit vector*, that is, $a^\top a = 1$, where $X_i = (x_{i\,1}, \cdots, x_{i\,d})$.

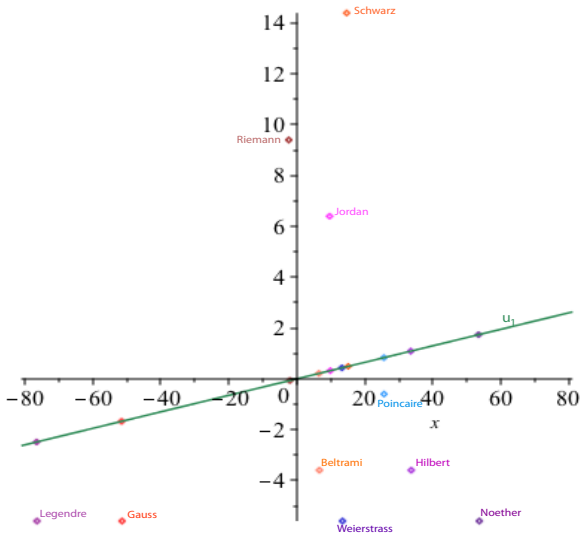

Fig. 21.5 The first principal components of Example 21.10, i.e. the projection of the centered data points onto the u_1 line.

If we form the symmetric matrix

$$\begin{pmatrix} x_{1\,1} & \cdots & x_{1\,d} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n\,1} & \cdots & x_{n\,d} & 1 \end{pmatrix}^\top \begin{pmatrix} x_{1\,1} & \cdots & x_{1\,d} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n\,1} & \cdots & x_{n\,d} & 1 \end{pmatrix}$$

involved in the normal equations, we see that the bottom row (and last column) of that matrix is

$$n\mu_1 \quad \cdots \quad n\mu_d \quad n,$$

where $n\mu_j = \sum_{i=1}^{n} x_{i\,j}$ is n times the mean of the column C_j of X.

Therefore, if $(a_1, \ldots, a_d, c)$ is a least squares solution, that is, a solution of the normal equations, we must have

$$n\mu_1 a_1 + \cdots + n\mu_d a_d + nc = 0,$$

that is,

$$a_1\mu_1 + \cdots + a_d\mu_d + c = 0,$$

which means that the *hyperplane* A_1 *must pass through the centroid* μ *of the data points* $X_1, \ldots, X_n$. Then we can rewrite the original system with

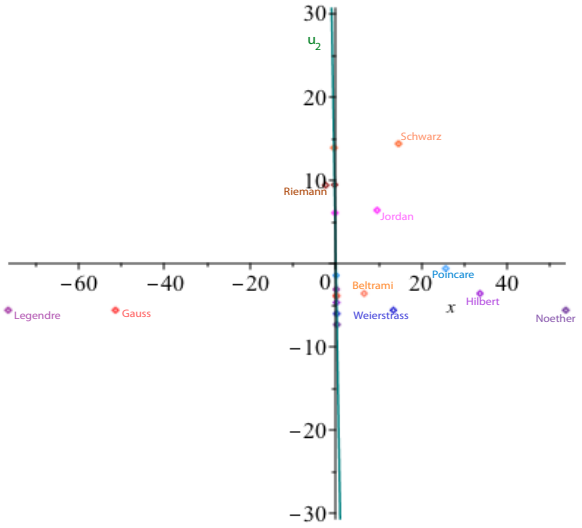

Fig. 21.6 The second principal components of Example 21.10, i.e. the projection of the centered data points onto the u_2 line.

respect to the centered data $X_i - \mu$, find that the variable c drops out, get the system

$$(X - \mu)a = 0,$$

where $a = (a_1, \ldots, a_d)$.

Thus, we are looking for a unit vector a solving $(X - \mu)a = 0$ in the least squares sense, that is, some a such that $a^\top a = 1$ minimizing

$$a^\top (X - \mu)^\top (X - \mu)a.$$

Compute some SVD $VDU^\top$ of $X - \mu$, where the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ of $X - \mu$ arranged in descending order. Then

$$a^\top (X - \mu)^\top (X - \mu)a = a^\top U D^2 U^\top a,$$

where $D^2 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is a diagonal matrix, so pick a to be *the last column in* U (corresponding to the smallest eigenvalue σ_d^2 of $(X - \mu)^\top (X - \mu)$). This is a solution to our best fit problem.

Therefore, if U_{d-1} is the linear hyperplane defined by a, that is,

$$U_{d-1} = \{u \in \mathbb{R}^d \mid \langle u, a \rangle = 0\},$$

where a is the last column in U for some SVD $VDU^\top$ of $X - \mu$, we have shown that the affine hyperplane $A_1 = \mu + U_{d-1}$ is a best approximation of the data set $X_1, \ldots, X_n$ in the least squares sense.

It is easy to show that this hyperplane $A_1 = \mu + U_{d-1}$ minimizes the sum of the square distances of each X_i to its orthogonal projection onto A_1. Also, since U_{d-1} is the orthogonal complement of a, the last column of U, we see that U_{d-1} is spanned by the first $d-1$ columns of U, that is, the first $d-1$ principal directions of $X - \mu$.

All this can be generalized to a *best* $(d-k)$*-dimensional affine subspace* A_k *approximating* $X_1, \ldots, X_n$ *in the least squares sense* $(1 \leq k \leq d-1)$. Such an affine subspace A_k is cut out by k independent hyperplanes H_i (with $1 \leq i \leq k$), each given by some equation

$$a_{i\,1}x_1 + \cdots + a_{i\,d}x_d + c_i = 0.$$

If we write $a_i = (a_{i\,1}, \cdots, a_{i\,d})$, to say that the H_i are independent means that $a_1, \ldots, a_k$ are linearly independent. In fact, we may assume that $a_1, \ldots, a_k$ form an *orthonormal system*.

Then finding a best $(d-k)$-dimensional affine subspace A_k amounts to solving the homogeneous linear system

$$\begin{pmatrix} X & \mathbf{1} & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & X & \mathbf{1} \end{pmatrix} \begin{pmatrix} a_1 \\ c_1 \\ \vdots \\ a_k \\ c_k \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

in the least squares sense, subject to the conditions $a_i^\top a_j = \delta_{i\,j}$, for all i, j with $1 \leq i, j \leq k$, where the matrix of the system is a block diagonal matrix consisting of k diagonal blocks $(X, \mathbf{1})$, where $\mathbf{1}$ denotes the column vector $(1, \ldots, 1) \in \mathbb{R}^n$.

Again it is easy to see that each hyperplane H_i *must pass through the centroid* μ of $X_1, \ldots, X_n$, and by switching to the centered data $X_i - \mu$ we get the system

$$\begin{pmatrix} X - \mu & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X - \mu \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

with $a_i^\top a_j = \delta_{i\,j}$ for all i, j with $1 \leq i, j \leq k$.

If $VDU^\top = X - \mu$ is an SVD decomposition, it is easy to see that a least squares solution of this system is given by the last k columns of

U, assuming that the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ of $X - \mu$ arranged in descending order. But now the $(d-k)$-dimensional subspace U_{d-k} cut out by the hyperplanes defined by $a_1, \ldots, a_k$ is simply the orthogonal complement of $(a_1, \ldots, a_k)$, which is the subspace spanned by the first $d-k$ columns of U.

So the best $(d-k)$-dimensional affine subpsace A_k approximating $X_1, \ldots, X_n$ in the least squares sense is $A_k = \mu + U_{d-k}$, where U_{d-k} is the linear subspace spanned by the first $d-k$ principal directions of $X - \mu$, that is, the first $d-k$ columns of U. Consequently, we get the following interesting interpretation of PCA (actually, principal directions):

Theorem 21.4. *Let X be an $n \times d$ matrix of data points $X_1, \ldots, X_n$, and let μ be the centroid of the X_i's. If $X - \mu = VDU^\top$ is an SVD decomposition of $X - \mu$ and if the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$, then a best $(d-k)$-dimensional affine approximation A_k of $X_1, \ldots, X_n$ in the least squares sense is given by*

$$A_k = \mu + U_{d-k},$$

where U_{d-k} is the linear subspace spanned by the first $d-k$ columns of U, the first $d-k$ principal directions of $X - \mu$ ($1 \leq k \leq d-1$).

Example 21.11. Going back to Example 21.10, a best 1-dimensional affine approximation A_1 is the affine line passing through $(\mu_1, \mu_2) = (1824.4, 5.6)$ of direction $u_1 = (0.9995, 0.0325)$.

Example 21.12. Suppose in the data set of Example 21.5 that we add the month of birth of every mathematician as a feature. We obtain the following data set.

Name	month	year	length
Carl Friedrich Gauss	4	1777	0
Camille Jordan	1	1838	12
Adrien-Marie Legendre	9	1752	0
Bernhard Riemann	9	1826	15
David Hilbert	1	1862	2
Henri Poincaré	4	1854	5
Emmy Noether	3	1882	0
Karl Weierstrass	10	1815	0
Eugenio Beltrami	10	1835	2
Hermann Schwarz	1	1843	20

The mean of the first column is 5.2, and the centered data set is given below.

Name	month	year	length
Carl Friedrich Gauss	−1.2	−51.4	−5.6
Camille Jordan	−4.2	9.6	6.4
Adrien-Marie Legendre	3.8	−76.4	−5.6
Bernhard Riemann	3.8	−2.4	9.4
David Hilbert	−4.2	33.6	−3.6
Henri Poincaré	−1.2	25.6	−0.6
Emmy Noether	−2.2	53.6	−5.6
Karl Weierstrass	4.8	13.4	−5.6
Eugenio Beltrami	4.8	6.6	−3.6
Hermann Schwarz	−4.2	14.6	14.4

Running SVD on this data set we get

$$U = \begin{pmatrix} 0.0394 & 0.1717 & 0.9844 \\ -0.9987 & 0.0390 & 0.0332 \\ -0.0327 & -0.9844 & 0.1730 \end{pmatrix}, \quad D = \begin{pmatrix} 117.0706 & 0 & 0 \\ 0 & 22.0390 & 0 \\ 0 & 0 & 10.1571 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and

$$VD = \begin{pmatrix} 51.4683 & 3.3013 & -3.8569 \\ -9.9623 & -6.6467 & -2.7082 \\ 76.6327 & 3.1845 & 0.2348 \\ 2.2393 & -8.6943 & 5.2872 \\ -33.6038 & 4.1334 & -3.6415 \\ -25.5941 & 1.3833 & -0.4350 \\ -53.4333 & 7.2258 & -1.3547 \\ -13.0100 & 6.8594 & 4.2010 \\ -6.2843 & 4.6254 & 4.3212 \\ -15.2173 & -14.3266 & -1.1581 \end{pmatrix},$$

$$X - \mu = \begin{pmatrix} -1.2000 & -51.4000 & -5.6000 \\ -4.2000 & 9.6000 & 6.4000 \\ 3.8000 & -76.4000 & -5.6000 \\ 3.8000 & -2.4000 & 9.4000 \\ -4.2000 & 33.6000 & -3.6000 \\ -1.2000 & 25.6000 & -0.6000 \\ -2.2000 & 53.6000 & -5.6000 \\ 4.8000 & 13.4000 & -5.6000 \\ 4.8000 & 6.6000 & -3.6000 \\ -4.2000 & 14.6000 & 14.4000 \end{pmatrix}.$$

The first principal direction $u_1 = (0.0394, -0.9987, -0.0327)$ is basically the opposite of the y-axis, and the most significant feature is the year of birth. The second principal direction $u_2 = (0.1717, 0.0390, -0.9844)$ is close to the opposite of the z-axis, and the second most significant feature is the lenght of beards. A best affine plane is spanned by the vectors u_1 and u_2.

There are many applications of PCA to data compression, dimension reduction, and pattern analysis. The basic idea is that in many cases, given a data set $X_1, \ldots, X_n$, with $X_i \in \mathbb{R}^d$, only a "small" subset of $m < d$ of the features is needed to describe the data set accurately.

If $u_1, \ldots, u_d$ are the principal directions of $X - \mu$, then the first m projections of the data (the first m principal components, i.e., the first m columns of VD) onto the first m principal directions represent the data without much loss of information. Thus, instead of using the original data points $X_1, \ldots, X_n$, with $X_i \in \mathbb{R}^d$, we can use their projections onto the first m principal directions $Y_1, \ldots, Y_m$, where $Y_i \in \mathbb{R}^m$ and $m < d$, obtaining a compressed version of the original data set.

For example, PCA is used in computer vision for *face recognition*. Sirovitch and Kirby (1987) seem to be the first to have had the idea of using PCA to compress facial images. They introduced the term *eigenpicture* to refer to the principal directions, u_i. However, an explicit face recognition algorithm was given only later by Turk and Pentland (1991). They renamed eigenpictures as *eigenfaces*.

For details on the topic of eigenfaces, see Forsyth and Ponce [Forsyth and Ponce (2002)] (Chapter 22, Section 22.3.2), where you will also find exact references to Turk and Pentland's papers.

Another interesting application of PCA is to the *recognition of handwritten digits*. Such an application is described in Hastie, Tibshirani, and Friedman, [Hastie *et al.* (2009)] (Chapter 14, Section 14.5.1).

21.6 Summary

The main concepts and results of this chapter are listed below:

- *Least squares problems*.
- Existence of a least squares solution of smallest norm (Theorem 21.1).
- The *pseudo-inverse* A^+ of a matrix A.
- The least squares solution of smallest norm is given by the pseudo-inverse (Theorem 21.2)
- Projection properties of the pseudo-inverse.
- The pseudo-inverse of a normal matrix.
- The *Penrose characterization* of the pseudo-inverse.
- Data compression and SVD.
- Best approximation of rank $< r$ of a matrix.
- *Principal component analysis*.
- Review of basic statistical concepts: *mean, variance, covariance*, *covariance matrix*.
- Centered data, *centroid*.
- The *principal components* (*PCA*).
- The *Rayleigh–Ritz theorem* (Theorem 21.8).
- The main theorem: *SVD yields PCA* (Theorem 21.3).
- Best affine approximation.
- SVD yields a best affine approximation (Theorem 21.4).
- Face recognition, eigenfaces.

21.7 Problems

Problem 21.1. Consider the overdetermined system in the single variable x:

$$a_1 x = b_1, \ldots, a_m x = b_m,$$

with $a_1^2 + \cdots + a_m^2 \neq 0$. Prove that the least squares solution of smallest norm is given by

$$x^+ = \frac{a_1 b_1 + \cdots + a_m b_m}{a_1^2 + \cdots + a_m^2}.$$

Problem 21.2. Let X be an $m \times n$ real matrix. For any strictly positive constant $K > 0$, the matrix $X^\top X + K I_n$ is invertible. Prove that the limit of the matrix $(X^\top X + K I_n)^{-1} X^\top$ when K goes to zero is equal to the pseudo-inverse X^+ of X.

Problem 21.3. Use `Matlab` to find the pseudo-inverse of the 8×6 matrix

$$A = \begin{pmatrix} 64 & 2 & 3 & 61 & 60 & 6 \\ 9 & 55 & 54 & 12 & 13 & 51 \\ 17 & 47 & 46 & 20 & 21 & 43 \\ 40 & 26 & 27 & 37 & 36 & 30 \\ 32 & 34 & 35 & 29 & 28 & 38 \\ 41 & 23 & 22 & 44 & 45 & 19 \\ 49 & 15 & 14 & 52 & 53 & 11 \\ 8 & 58 & 59 & 5 & 4 & 62 \end{pmatrix}.$$

Observe that the sums of the columns are all equal to to 256. Let b be the vector of dimension 8 whose coordinates are all equal to 256. Find the solution x^+ of the system $Ax = b$.

Problem 21.4. The purpose of this problem is to show that Proposition 21.7 (the Eckart–Young theorem) also holds for the Frobenius norm. This problem is adapted from Strang [Strang (2019)], Section I.9.

Suppose the $m \times n$ matrix B of rank at most k minimizes $\|A - B\|_F$. Start with an SVD of B,

$$B = V \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} U^\top,$$

where D is a diagonal $k \times k$ matrix. We can write

$$A = V \begin{pmatrix} L + E + R & F \\ G & H \end{pmatrix} U^\top,$$

where L is strictly lower triangular in the first k rows, E is diagonal, and R is strictly upper triangular, and let

$$C = V \begin{pmatrix} L + D + R & F \\ 0 & 0 \end{pmatrix} U^\top,$$

which clearly has rank $\leq k$.

(1) Prove that

$$\|A - B\|_F^2 = \|A - C\|_F^2 + \|L\|_F^2 + \|R\|_F^2 + \|F\|_F^2.$$

Since $\|A - B\|_F$ is minimal, show that $L = R = F = 0$.

Similarly, show that $G = 0$.

(2) We have

$$V^\top A U = \begin{pmatrix} E & 0 \\ 0 & H \end{pmatrix}, \quad V^\top B U = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix},$$

where E is diagonal, so deduce that

(1) $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$.
(2) The singular values of H must be the smallest $n-k$ singular values of A.
(3) The minimum of $\|A - B\|_F$ must be $\|H\|_F = (\sigma_{k+1}^2 + \cdots + \sigma_r^2)^{1/2}$.

Problem 21.5. Prove that the closest rank 1 approximation (in $\| \ \|_2$) of the matrix

$$A = \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix}$$

is

$$A_1 = \frac{3}{2}\begin{pmatrix} 1 & 1 \\ 3 & 3 \end{pmatrix}.$$

Show that the Eckart–Young theorem fails for the operator norm $\| \ \|_\infty$ by finding a rank 1 matrix B such that $\|A - B\|_\infty < \|A - A_1\|_\infty$.

Problem 21.6. Find a closest rank 1 approximation (in $\| \ \|_2$) for the matrices

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

Problem 21.7. Find a closest rank 1 approximation (in $\| \ \|_2$) for the matrix

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

Problem 21.8. Let S be a real symmetric positive definite matrix and let $S = U\Sigma U^\top$ be a diagonalization of S. Prove that the closest rank 1 matrix (in the L^2-norm) to S is $u_1\sigma_1 u_1^\top$, where u_1 is the first column of U.

Chapter 22

Annihilating Polynomials and the Primary Decomposition

In this chapter all vector spaces are defined over an arbitrary field K.

In Section 6.7 we explained that if $f\colon E \to E$ is a linear map on a K-vector space E, then for any polynomial $p(X) = a_0X^d + a_1X^{d-1} + \cdots + a_d$ with coefficients in the field K, we can define the *linear map* $p(f)\colon E \to E$ by

$$p(f) = a_0f^d + a_1f^{d-1} + \cdots + a_d\mathrm{id},$$

where $f^k = f \circ \cdots \circ f$, the k-fold composition of f with itself. Note that

$$p(f)(u) = a_0f^d(u) + a_1f^{d-1}(u) + \cdots + a_du,$$

for every vector $u \in E$. Then we showed that if E is finite-dimensional and if $\chi_f(X) = \det(XI - f)$ is the characteristic polynomial of f, by the Cayley–Hamilton theorem, we have

$$\chi_f(f) = 0.$$

This fact suggests looking at the set of all polynomials $p(X)$ such that

$$p(f) = 0.$$

Such polynomials are called *annihilating polynomials* of f, the set of all these polynomials, denoted $\mathrm{Ann}(f)$, is called the *annihilator* of f, and the Cayley-Hamilton theorem shows that it is nontrivial since it contains a polynomial of positive degree. It turns out that $\mathrm{Ann}(f)$ contains a polynomial m_f of smallest degree that generates $\mathrm{Ann}(f)$, and this polynomial divides the characteristic polynomial. Furthermore, the polynomial m_f encapsulates a lot of information about f, in particular whether f can be diagonalized. One of the main reasons for this is that a scalar $\lambda \in K$ is a zero of the minimal polynomial m_f if and only if λ is an eigenvalue of f.

The first main result is Theorem 22.2 which states that if $f\colon E \to E$ is a linear map on a finite-dimensional space E, then f is diagonalizable iff its minimal polynomial m is of the form

$$m = (X - \lambda_1) \cdots (X - \lambda_k),$$

where $\lambda_1, \ldots, \lambda_k$ are distinct elements of K.

One of the technical tools used to prove this result is the notion of *f-conductor*; see Definition 22.7. As a corollary of Theorem 22.2 we obtain results about finite commuting families of diagonalizable or triangulable linear maps.

If $f\colon E \to E$ is a linear map and $\lambda \in K$ is an eigenvalue of f, recall that the eigenspace E_λ associated with λ is the kernel of the linear map $\lambda \mathrm{id} - f$. If all the eigenvalues $\lambda_1 \ldots, \lambda_k$ of f are in K and if f is diagonalizable, then

$$E = E_{\lambda_1} \oplus \cdots \oplus E_{\lambda_k},$$

but in general there are not enough eigenvectors to span E. A remedy is to generalize the notion of eigenvector and look for (nonzero) vectors u (called generalized eigenvectors) such that

$$(\lambda \mathrm{id} - f)^r(u) = 0, \quad \text{for some } r \geq 1.$$

Then it turns out that if the minimal polynomial of f is of the form

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

then $r = r_i$ does the job for λ_i; that is, if we let

$$W_i = \mathrm{Ker}\,(\lambda_i \mathrm{id} - f)^{r_i},$$

then

$$E = W_1 \oplus \cdots \oplus W_k.$$

The above facts are parts of the *primary decomposition theorem* (Theorem 22.4). It is a special case of a more general result involving the factorization of the minimal polynomial m into its irreducible monic factors; see Theorem 22.3.

Theorem 22.4 implies that every linear map f that has all its eigenvalues in K can be written as $f = D + N$, where D is diagonalizable and N is nilpotent (which means that $N^r = 0$ for some positive integer r). Furthermore D and N commute and are unique. This is the *Jordan decomposition*, Theorem 22.5.

The Jordan decomposition suggests taking a closer look at nilpotent maps. We prove that for any nilpotent linear map $f: E \to E$ on a finite-dimensional vector space E of dimension n over a field K, there is a basis of E such that the matrix N of f is of the form

$$N = \begin{pmatrix} 0 & \nu_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \nu_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \nu_n \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix},$$

where $\nu_i = 1$ or $\nu_i = 0$; see Theorem 22.6. As a corollary we obtain the *Jordan form*; which involves matrices of the form

$$J_r(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 1 \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix},$$

called *Jordan blocks*; see Theorem 22.7.

22.1 Basic Properties of Polynomials; Ideals, GCD's

In order to understand the structure of $\mathrm{Ann}(f)$, we need to review three basic properties of polynomials. We refer the reader to Hoffman and Kunze, [Hoffman and Kunze (1971)], Artin [Artin (1991)], Dummit and Foote [Dummit and Foote (1999)], and Godement [Godement (1963)] for comprehensive discussions of polynomials and their properties.

We begin by recalling some basic nomenclature. Given a field K, any nonzero polynomial $p(X) \in K[X]$ has some monomial of highest degree a_0X^n with $a_0 \neq 0$, and the integer $n = \deg(p) \geq 0$ is called the *degree* of p. It is convenient to set the degree of the zero polynomial (denoted by 0) to be

$$\deg(0) = -\infty.$$

A polynomial $p(X)$ such that the coefficient a_0 of its monomial of highest degree is 1 is called a *monic* polynomial. For example, let $K = \mathbb{R}$. The polynomial $p(X) = 4X^7 + 2X^5$ is of degree 7 but is not monic since $a_0 = 4$. On the other hand, the polynomial $p(X) = X^3 - 3X + 1$ is a monic polynomial of degree 3.

We now discuss three key concepts of polynomial algebra:

(1) Ideals
(2) Greatest common divisors and the Bezout identity.
(3) Irreducible polynomials and prime factorization.

Recall the definition a of ring (see Definition 2.2).

Definition 22.1. A *ring* is a set A equipped with two operations $+\colon A \times A \to A$ (called *addition*) and $*\colon A \times A \to A$ (called *multiplication*) having the following properties:

(R1) A is an abelian group w.r.t. $+$;
(R2) $*$ is associative and has an identity element $1 \in A$;
(R3) $*$ is distributive w.r.t. $+$.

The identity element for addition is denoted 0, and the additive inverse of $a \in A$ is denoted by $-a$. More explicitly, the axioms of a ring are the following equations which hold for all $a, b, c \in A$:

$$a + (b + c) = (a + b) + c \qquad \text{(associativity of } +) \tag{22.1}$$
$$a + b = b + a \qquad \text{(commutativity of } +) \tag{22.2}$$
$$a + 0 = 0 + a = a \qquad \text{(zero)} \tag{22.3}$$
$$a + (-a) = (-a) + a = 0 \qquad \text{(additive inverse)} \tag{22.4}$$
$$a * (b * c) = (a * b) * c \qquad \text{(associativity of } *) \tag{22.5}$$
$$a * 1 = 1 * a = a \qquad \text{(identity for } *) \tag{22.6}$$
$$(a + b) * c = (a * c) + (b * c) \qquad \text{(distributivity)} \tag{22.7}$$
$$a * (b + c) = (a * b) + (a * c) \qquad \text{(distributivity)} \tag{22.8}$$

The ring A is *commutative* if

$$a * b = b * a \quad \text{for all } a, b \in A.$$

From (22.7) and (22.8), we easily obtain

$$a * 0 = 0 * a = 0 \tag{22.9}$$
$$a * (-b) = (-a) * b = -(a * b). \tag{22.10}$$

The first crucial notion is that of an ideal.

Definition 22.2. Given a commutative ring A with unit 1, *an ideal of* A is a nonempty subset $\mathfrak{I}$ of A satisfying the following properties:

(ID1) If $a, b \in \mathfrak{I}$, then $b - a \in \mathfrak{I}$.
(ID2) If $a \in \mathfrak{I}$, then $ax \in \mathfrak{I}$ for all $x \in A$.

An ideal $\mathfrak{I}$ is a *principal ideal* if there is some $a \in \mathfrak{I}$, called a *generator*, such that

$$\mathfrak{I} = \{ax \mid x \in A\}.$$

In this case we usually write $\mathfrak{I} = aA$ or $\mathfrak{I} = (a)$. The ideal $\mathfrak{I} = (0) = \{0\}$ is called the *null ideal* (or *zero ideal*).

The following proposition is a fundamental result about polynomials over a field.

Proposition 22.1. *If K is a field, then every polynomial ideal $\mathfrak{I} \subseteq K[X]$ is a principal ideal. As a consequence, if $\mathfrak{I}$ is not the zero ideal, then there is a unique monic polynomial*

$$p(X) = X^n + a_1 X^{n-1} + \cdots + a_{n-1}X + a_n$$

in $\mathfrak{I}$ such that $\mathfrak{I} = (p)$.

Proof. This result is not hard to prove if we recall that polynomials can divided. Given any two nonzero polynomials $f, g \in K[X]$, there are unique polynomials q, r such that

$$f = qg + r, \quad \text{and} \quad \deg(r) < \deg(g). \tag{*}$$

If $\mathfrak{I}$ is not the zero ideal, there is some polynomial of smallest degree in $\mathfrak{I}$, and since K is a field, by suitable multiplication by a scalar, we can make sure that this polynomial is monic. Thus, let f be a monic polynomial of smallest degree in $\mathfrak{I}$. By (ID2), it is clear that $(f) \subseteq \mathfrak{I}$. Now let $g \in \mathfrak{I}$. Using $(*)$, there exist unique $q, r \in K[X]$ such that

$$g = qf + r \quad \text{and} \quad \deg(r) < \deg(f).$$

If $r \neq 0$, there is some $\lambda \neq 0$ in K such that λr is a monic polynomial, and since $\lambda r = \lambda g - \lambda q f$, with $f, g \in \mathfrak{I}$, by (ID1) and (ID2), we have $\lambda r \in \mathfrak{I}$, where $\deg(\lambda r) < \deg(f)$ and λr is a monic polynomial, contradicting the minimality of the degree of f. Thus, $r = 0$, and $g \in (f)$. The uniqueness of the monic polynomial f is left as an exercise. □

We will also need to know that the greatest common divisor of polynomials exist. Given any two nonzero polynomials $f, g \in K[X]$, recall that f divides g if $g = qf$ for some $q \in K[X]$.

Definition 22.3. Given any two nonzero polynomials $f, g \in K[X]$, a polynomial $d \in K[X]$ is a *greatest common divisor of f and g* (for short, a *gcd of f and g*) if d divides f and g and whenever $h \in K[X]$ divides f and g, then h divides d. We say that f and g are *relatively prime* if 1 is a gcd of f and g.

Note that f and g are relatively prime iff all of their gcd's are constants (scalars in K), or equivalently, if f, g have no common divisor q of degree $\deg(q) \geq 1$. For example, over $\mathbb{R}$, $\gcd(X^2 - 1, X^3 + X^2 - X - 1) = (X - 1)(X+1)$ since $X^3+X^2-X-1 = (X-1)(X+1)^2$, while $\gcd(X^3+1, X-1) = 1$.

We can characterize gcd's of polynomials as follows.

Proposition 22.2. *Let K be a field and let $f, g \in K[X]$ be any two nonzero polynomials. For every polynomial $d \in K[X]$, the following properties are equivalent:*

(1) The polynomial d is a gcd of f and g.
(2) The polynomial d divides f and g and there exist $u, v \in K[X]$ such that

$$d = uf + vg.$$

(3) The ideals $(f), (g)$, and (d) satisfy the equation

$$(d) = (f) + (g).$$

In addition, $d \neq 0$, and d is unique up to multiplication by a nonzero scalar in K.

As a consequence of Proposition 22.2, two nonzero polynomials $f, g \in K[X]$ are relatively prime iff there exist $u, v \in K[X]$ such that

$$uf + vg = 1.$$

The identity

$$d = uf + vg$$

of Part (2) of Proposition 22.2 is often called the *Bezout identity*. For an example of Bezout's identity, take $K = \mathbb{R}$. Since $X^3 + 1$ and $X - 1$ are relatively prime, we have
$1 = 1/2(X^3 + 1) - 1/2(X^2 + X + 1)(X - 1)$.

An important consequence of the Bezout identity is the following result.

Proposition 22.3. *(Euclid's proposition) Let K be a field and let $f, g, h \in K[X]$ be any nonzero polynomials. If f divides gh and f is relatively prime to g, then f divides h.*

Proposition 22.3 can be generalized to any number of polynomials.

Proposition 22.4. *Let K be a field and let $f, g_1, \ldots, g_m \in K[X]$ be some nonzero polynomials. If f and g_i are relatively prime for all i, $1 \leq i \leq m$, then f and $g_1 \cdots g_m$ are relatively prime.*

Definition 22.3 is generalized to any finite number of polynomials as follows.

Definition 22.4. Given any nonzero polynomials $f_1, \ldots, f_n \in K[X]$, where $n \geq 2$, a polynomial $d \in K[X]$ is a *greatest common divisor of* $f_1, \ldots, f_n$ (for short, a *gcd of* $f_1, \ldots, f_n$) if d divides each f_i and whenever $h \in K[X]$ divides each f_i, then h divides d. We say that $f_1, \ldots, f_n$ are *relatively prime* if 1 is a gcd of $f_1, \ldots, f_n$.

It is easily shown that Proposition 22.2 can be generalized to any finite number of polynomials.

Proposition 22.5. *Let K be a field and let $f_1, \ldots, f_n \in K[X]$ be any $n \geq 2$ nonzero polynomials. For every polynomial $d \in K[X]$, the following properties are equivalent:*

(1) The polynomial d is a gcd of $f_1, \ldots, f_n$.
(2) The polynomial d divides each f_i and there exist $u_1, \ldots, u_n \in K[X]$ such that

$$d = u_1 f_1 + \cdots + u_n f_n.$$

(3) The ideals (f_i), and (d) satisfy the equation

$$(d) = (f_1) + \cdots + (f_n).$$

In addition, $d \neq 0$, and d is unique up to multiplication by a nonzero scalar in K.

As a consequence of Proposition 22.5, any $n \geq 2$ nonzero polynomials $f_1, \ldots, f_n \in K[X]$ are relatively prime iff there exist $u_1, \ldots, u_n \in K[X]$ such that

$$u_1 f_1 + \cdots + u_n f_n = 1,$$

the *Bezout identity*.

We will also need to know that every nonzero polynomial (over a field) can be factored into irreducible polynomials, which is the generalization of the prime numbers to polynomials.

Definition 22.5. Given a field K, a polynomial $p \in K[X]$ is *irreducible or indecomposable or prime* if $\deg(p) \geq 1$ and if p is not divisible by any polynomial $q \in K[X]$ such that $1 \leq \deg(q) < \deg(p)$. Equivalently, p is irreducible if $\deg(p) \geq 1$ and if $p = q_1 q_2$, then either $q_1 \in K$ or $q_2 \in K$ (and of course, $q_1 \neq 0$, $q_2 \neq 0$).

Every polynomial $aX + b$ of degree 1 is irreducible. Over the field $\mathbb{R}$, the polynomial $X^2 + 1$ is irreducible (why?), but $X^3 + 1$ is not irreducible, since

$$X^3 + 1 = (X + 1)(X^2 - X + 1).$$

The polynomial $X^2 - X + 1$ is irreducible over $\mathbb{R}$ (why?). It would seem that $X^4 + 1$ is irreducible over $\mathbb{R}$, but in fact,

$$X^4 + 1 = (X^2 - \sqrt{2}X + 1)(X^2 + \sqrt{2}X + 1).$$

However, in view of the above factorization, $X^4 + 1$ is irreducible over $\mathbb{Q}$.

It can be shown that the irreducible polynomials over $\mathbb{R}$ are the polynomials of degree 1 or the polynomials of degree 2 of the form $aX^2 + bX + c$, for which $b^2 - 4ac < 0$ (i.e., those having no real roots). This is not easy to prove! Over the complex numbers $\mathbb{C}$, the only irreducible polynomials are those of degree 1. This is a version of a fact often referred to as the "Fundamental Theorem of Algebra."

Observe that the definition of irreducibilty implies that any finite number of distinct irreducible polynomials are relatively prime.

The following fundamental result can be shown

Theorem 22.1. *Given any field K, for every nonzero polynomial*

$$f = a_d X^d + a_{d-1} X^{d-1} + \cdots + a_0$$

of degree $d = \deg(f) \geq 1$ in $K[X]$, there exists a unique set $\{\langle p_1, k_1\rangle, \ldots, \langle p_m, k_m\rangle\}$ such that

$$f = a_d p_1^{k_1} \cdots p_m^{k_m},$$

where the $p_i \in K[X]$ are distinct irreducible monic polynomials, the k_i are (not necessarily distinct) integers, and with $m \geq 1$, $k_i \geq 1$.

We can now return to minimal polynomials.

22.2 Annihilating Polynomials and the Minimal Polynomial

Given a linear map $f\colon E \to E$, it is easy to check that the set $\mathrm{Ann}(f)$ of polynomials that annihilate f is an ideal. Furthermore, when E is finite-dimensional, the Cayley–Hamilton theorem implies that $\mathrm{Ann}(f)$ is not the zero ideal. Therefore, by Proposition 22.1, there is a unique monic polynomial m_f that generates $\mathrm{Ann}(f)$.

Definition 22.6. If $f\colon E \to E$ is a linear map on a finite-dimensional vector space E, the unique monic polynomial $m_f(X)$ that generates the ideal $\mathrm{Ann}(f)$ of polynomials which annihilate f (the *annihilator* of f) is called the *minimal polynomial* of f.

The minimal polynomial m_f of f is the monic polynomial of smallest degree that annihilates f. Thus, the minimal polynomial divides the characteristic polynomial χ_f, and $\deg(m_f) \geq 1$. For simplicity of notation, we often write m instead of m_f.

If A is any $n \times n$ matrix, the set $\mathrm{Ann}(A)$ of polynomials that annihilate A is the set of polynomials

$$p(X) = a_0 X^d + a_1 X^{d-1} + \cdots + a_{d-1} X + a_d$$

such that

$$a_0 A^d + a_1 A^{d-1} + \cdots + a_{d-1} A + a_d I = 0.$$

It is clear that $\mathrm{Ann}(A)$ is a nonzero ideal and its unique monic generator is called the *minimal polynomial* of A. We check immediately that if Q is an invertible matrix, then A and $Q^{-1}AQ$ have the same minimal polynomial. Also, if A is the matrix of f with respect to some basis, then f and A have the same minimal polynomial.

The zeros (in K) of the minimal polynomial of f and the eigenvalues of f (in K) are intimately related.

Proposition 22.6. *Let $f\colon E \to E$ be a linear map on some finite-dimensional vector space E. Then $\lambda \in K$ is a zero of the minimal polynomial $m_f(X)$ of f iff λ is an eigenvalue of f iff λ is a zero of $\chi_f(X)$. Therefore, the minimal and the characteristic polynomials have the same zeros (in K), except for multiplicities.*

Proof. First assume that $m(\lambda) = 0$ (with $\lambda \in K$, and writing m instead of m_f). If so, using polynomial division, m can be factored as

$$m = (X - \lambda)q,$$

with $\deg(q) < \deg(m)$. Since m is the minimal polynomial, $q(f) \neq 0$, so there is some nonzero vector $v \in E$ such that $u = q(f)(v) \neq 0$. But then, because m is the minimal polynomial,

$$\begin{aligned} 0 &= m(f)(v) \\ &= (f - \lambda \mathrm{id})(q(f)(v)) \\ &= (f - \lambda \mathrm{id})(u), \end{aligned}$$

which shows that λ is an eigenvalue of f.

Conversely, assume that $\lambda \in K$ is an eigenvalue of f. This means that for some $u \neq 0$, we have $f(u) = \lambda u$. Now it is easy to show that

$$m(f)(u) = m(\lambda)u,$$

and since m is the minimal polynomial of f, we have $m(f)(u) = 0$, so $m(\lambda)u = 0$, and since $u \neq 0$, we must have $m(\lambda) = 0$. □

Proposition 22.7. *Let $f: E \to E$ be a linear map on some finite-dimensional vector space E. If f diagonalizable, then its minimal polynomial is a product of distinct factors of degree* 1.

Proof. If we assume that f is diagonalizable, then its eigenvalues are all in K, and if $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of f, and then by Proposition 22.6, the minimal polynomial m of f must be a product of powers of the polynomials $(X - \lambda_i)$. Actually, we claim that

$$m = (X - \lambda_1) \cdots (X - \lambda_k).$$

For this we just have to show that m annihilates f. However, for any eigenvector u of f, one of the linear maps $f - \lambda_i \mathrm{id}$ sends u to 0, so

$$m(f)(u) = (f - \lambda_1 \mathrm{id}) \circ \cdots \circ (f - \lambda_k \mathrm{id})(u) = 0.$$

Since E is spanned by the eigenvectors of f, we conclude that

$$m(f) = 0. \qquad \square$$

It turns out that the converse of Proposition 22.7 is true, but this will take a little work to establish it.

22.3 Minimal Polynomials of Diagonalizable Linear Maps

In this section we prove that if the minimal polynomial m_f of a linear map f is of the form

$$m_f = (X - \lambda_1) \cdots (X - \lambda_k)$$

for distinct scalars $\lambda_1, \ldots, \lambda_k \in K$, then f is diagonalizable. This is a powerful result that has a number of implications. But first we need of few properties of invariant subspaces.

Given a linear map $f: E \to E$, recall that a subspace W of E is *invariant under* f if $f(u) \in W$ for all $u \in W$. For example, if $f: \mathbb{R}^2 \to \mathbb{R}^2$ is $f(x, y) = (-x, y)$, the y-axis is invariant under f.

Proposition 22.8. *Let W be a subspace of E invariant under the linear map $f: E \to E$ (where E is finite-dimensional). Then the minimal polynomial of the restriction $f \mid W$ of f to W divides the minimal polynomial of f, and the characteristic polynomial of $f \mid W$ divides the characteristic polynomial of f.*

Sketch of proof. The key ingredient is that we can pick a basis $(e_1, \ldots, e_n)$ of E in which $(e_1, \ldots, e_k)$ is a basis of W. The matrix of f over this basis is a block matrix of the form

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

where B is a $k \times k$ matrix, D is an $(n-k) \times (n-k)$ matrix, and C is a $k \times (n-k)$ matrix. Then

$$\det(XI - A) = \det(XI - B)\det(XI - D),$$

which implies the statement about the characteristic polynomials. Furthermore,

$$A^i = \begin{pmatrix} B^i & C_i \\ 0 & D^i \end{pmatrix},$$

for some $k \times (n-k)$ matrix C_i. It follows that any polynomial which annihilates A also annihilates B and D. So the minimal polynomial of B divides the minimal polynomial of A. □

For the next step, there are at least two ways to proceed. We can use an old-fashion argument using Lagrange interpolants, or we can use a slight generalization of the notion of annihilator. We pick the second method because it illustrates nicely the power of principal ideals.

What we need is the notion of conductor (also called transporter).

Definition 22.7. Let $f\colon E \to E$ be a linear map on a finite-dimensional vector space E, let W be an invariant subspace of f, and let u be any vector in E. The set $S_f(u, W)$ consisting of all polynomials $q \in K[X]$ such that $q(f)(u) \in W$ is called the *f-conductor of u into W*.

Observe that the minimal polynomial m_f of f always belongs to $S_f(u, W)$, so this is a nontrivial set. Also, if $W = (0)$, then $S_f(u, (0))$ is just the annihilator of f. The crucial property of $S_f(u, W)$ is that it is an ideal.

Proposition 22.9. *If W is an invariant subspace for f, then for each $u \in E$, the f-conductor $S_f(u, W)$ is an ideal in $K[X]$.*

We leave the proof as a simple exercise, using the fact that if W invariant under f, then W is invariant under every polynomial $q(f)$ in $S_f(u, W)$.

Since $S_f(u, W)$ is an ideal, it is generated by a unique monic polynomial q of smallest degree, and because the minimal polynomial m_f of f is in $S_f(u, W)$, the polynomial q divides m.

Definition 22.8. The unique monic polynomial which generates $S_f(u, W)$ is called the *conductor of u into W*.

Example 22.1. For example, suppose $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ where $f(x, y) = (x, 0)$. Observe that $W = \{(x, 0) \in \mathbb{R}^2\}$ is invariant under f. By representing f as $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, we see that $m_f(X) = \chi_f(X) = X^2 - X$. Let $u = (0, y)$. Then $S_f(u, W) = (X)$, and we say X is the conductor of u into W.

Proposition 22.10. *Let $f\colon E \to E$ be a linear map on a finite-dimensional space E and assume that the minimal polynomial m of f is of the form*

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

where the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f belong to K. If W is a proper subspace of E which is invariant under f, then there is a vector $u \in E$ with the following properties:

(a) $u \notin W$;
(b) $(f - \lambda \mathrm{id})(u) \in W$, for some eigenvalue λ of f.

Proof. Observe that (a) and (b) together assert that the conductor of u into W is a polynomial of the form $X - \lambda_i$. Pick any vector $v \in E$ not in W, and let g be the conductor of v into W, i.e. $g(f)(v) \in W$. Since g divides m and $v \notin W$, the polynomial g is not a constant, and thus it is of the form

$$g = (X - \lambda_1)^{s_1} \cdots (X - \lambda_k)^{s_k},$$

with at least some $s_i > 0$. Choose some index j such that $s_j > 0$. Then $X - \lambda_j$ is a factor of g, so we can write

$$g = (X - \lambda_j)q. \tag{$*$}$$

By definition of g, the vector $u = q(f)(v)$ cannot be in W, since otherwise g would not be of minimal degree. However, $(*)$ implies that

$$\begin{aligned}(f - \lambda_j \mathrm{id})(u) &= (f - \lambda_j \mathrm{id})(q(f)(v)) \\ &= g(f)(v)\end{aligned}$$

is in W, which concludes the proof. □

We can now prove the main result of this section.

Theorem 22.2. *Let $f: E \to E$ be a linear map on a finite-dimensional space E. Then f is diagonalizable iff its minimal polynomial m is of the form*

$$m = (X - \lambda_1) \cdots (X - \lambda_k),$$

where $\lambda_1, \ldots, \lambda_k$ are distinct elements of K.

Proof. We already showed in Proposition 22.7 that if f is diagonalizable, then its minimal polynomial is of the above form (where $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of f).

For the converse, let W be the subspace spanned by all the eigenvectors of f. If $W \neq E$, since W is invariant under f, by Proposition 22.10, there is some vector $u \notin W$ such that for some λ_j, we have

$$(f - \lambda_j \mathrm{id})(u) \in W.$$

Let $v = (f - \lambda_j \mathrm{id})(u) \in W$. Since $v \in W$, we can write

$$v = w_1 + \cdots + w_k$$

where $f(w_i) = \lambda_i w_i$ (either $w_i = 0$ or w_i is an eigenvector for λ_i), and so for every polynomial h, we have

$$h(f)(v) = h(\lambda_1)w_1 + \cdots + h(\lambda_k)w_k,$$

which shows that $h(f)(v) \in W$ for every polynomial h. We can write

$$m = (X - \lambda_j)q$$

for some polynomial q, and also

$$q - q(\lambda_j) = p(X - \lambda_j)$$

for some polynomial p. We know that $p(f)(v) \in W$, and since m is the minimal polynomial of f, we have

$$0 = m(f)(u) = (f - \lambda_j \mathrm{id})(q(f)(u)),$$

which implies that $q(f)(u) \in W$ (either $q(f)(u) = 0$, or it is an eigenvector associated with λ_j). However,

$$q(f)(u) - q(\lambda_j)u = p(f)((f - \lambda_j \mathrm{id})(u)) = p(f)(v),$$

and since $p(f)(v) \in W$ and $q(f)(u) \in W$, we conclude that $q(\lambda_j)u \in W$. But, $u \notin W$, which implies that $q(\lambda_j) = 0$, so λ_j is a double root of m, a contradiction. Therefore, we must have $W = E$. □

Remark: Proposition 22.10 can be used to give a quick proof of Theorem 14.1.

22.4 Commuting Families of Diagonalizable and Triangulable Maps

Using Theorem 22.2, we can give a short proof about commuting diagonalizable linear maps.

Definition 22.9. If $\mathcal{F}$ is a family of linear maps on a vector space E, we say that $\mathcal{F}$ is a *commuting family* iff $f \circ g = g \circ f$ for all $f, g \in \mathcal{F}$.

Proposition 22.11. *Let $\mathcal{F}$ be a commuting family of diagonalizable linear maps on a vector space E. There exists a basis of E such that every linear map in $\mathcal{F}$ is represented in that basis by a diagonal matrix.*

Proof. We proceed by induction on $n = \dim(E)$. If $n = 1$, there is nothing to prove. If $n > 1$, there are two cases. If all linear maps in $\mathcal{F}$ are of the form $\lambda \mathrm{id}$ for some $\lambda \in K$, then the proposition holds trivially. In the second case, let $f \in \mathcal{F}$ be some linear map in $\mathcal{F}$ which is not a scalar multiple of the identity. In this case, f has at least two distinct eigenvalues $\lambda_1, \ldots, \lambda_k$, and because f is diagonalizable, E is the direct sum of the corresponding eigenspaces $E_{\lambda_1}, \ldots, E_{\lambda_k}$. For every index i, the eigenspace E_{λ_i} is invariant under f and under every other linear map g in $\mathcal{F}$, since for any $g \in \mathcal{F}$ and any $u \in E_{\lambda_i}$, because f and g commute, we have

$$f(g(u)) = g(f(u)) = g(\lambda_i u) = \lambda_i g(u)$$

so $g(u) \in E_{\lambda_i}$. Let $\mathcal{F}_i$ be the family obtained by restricting each $f \in \mathcal{F}$ to E_{λ_i}. By Proposition 22.8, the minimal polynomial of every linear map $f \mid E_{\lambda_i}$ in $\mathcal{F}_i$ divides the minimal polynomial m_f of f, and since f is diagonalizable, m_f is a product of distinct linear factors, so the minimal polynomial of $f \mid E_{\lambda_i}$ is also a product of distinct linear factors. By Theorem 22.2, the linear map $f \mid E_{\lambda_i}$ is diagonalizable. Since $k > 1$, we have $\dim(E_{\lambda_i}) < \dim(E)$ for $i = 1, \ldots, k$, and by the induction hypothesis, for each i there is a basis of E_{λ_i} over which $f \mid E_{\lambda_i}$ is represented by a diagonal matrix. Since the above argument holds for all i, by combining the bases of the E_{λ_i}, we obtain a basis of E such that the matrix of every linear map $f \in \mathcal{F}$ is represented by a diagonal matrix. □

There is also an analogous result for commuting families of linear maps represented by upper triangular matrices. To prove this we need the following proposition.

Proposition 22.12. *Let $\mathcal{F}$ be a nonempty commuting family of triangulable linear maps on a finite-dimensional vector space E. Let W be a proper*

subspace of E which is invariant under $\mathcal{F}$. Then there exists a vector $u \in E$ such that:

(1) $u \notin W$.
(2) For every $f \in \mathcal{F}$, the vector $f(u)$ belongs to the subspace $W \oplus Ku$ spanned by W and u.

Proof. By renaming the elements of $\mathcal{F}$ if necessary, we may assume that $(f_1, \ldots, f_r)$ is a basis of the subspace of $\mathrm{End}(E)$ spanned by $\mathcal{F}$. We prove by induction on r that there exists some vector $u \in E$ such that

(1) $u \notin W$.
(2) $(f_i - \alpha_i \mathrm{id})(u) \in W$ for $i = 1, \ldots, r$, for some scalars $\alpha_i \in K$.

Consider the base case $r = 1$. Since f_1 is triangulable, its eigenvalues all belong to K since they are the diagonal entries of the triangular matrix associated with f_1 (this is the easy direction of Theorem 14.1), so the minimal polynomial of f_1 is of the form

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

where the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f_1 belong to K. We conclude by applying Proposition 22.10.

Next assume that $r \geq 2$ and that the induction hypothesis holds for $f_1, \ldots, f_{r-1}$. Thus, there is a vector $u_{r-1} \in E$ such that

(1) $u_{r-1} \notin W$.
(2) $(f_i - \alpha_i \mathrm{id})(u_{r-1}) \in W$ for $i = 1, \ldots, r-1$, for some scalars $\alpha_i \in K$.

Let

$$V_{r-1} = \{w \in E \mid (f_i - \alpha_i \mathrm{id})(w) \in W,\ i = 1, \ldots, r-1\}.$$

Clearly, $W \subseteq V_{r-1}$ and $u_{r-1} \in V_{r-1}$. We claim that V_{r-1} is invariant under $\mathcal{F}$. This is because, for any $v \in V_{r-1}$ and any $f \in \mathcal{F}$, since f and f_i commute, we have

$$(f_i - \alpha_i \mathrm{id})(f(v)) = f((f_i - \alpha_i \mathrm{id})(v)), \quad 1 \leq i \leq r-1.$$

Now $(f_i - \alpha_i \mathrm{id})(v) \in W$ because $v \in V_{r-1}$, and W is invariant under $\mathcal{F}$, so $f(f_i - \alpha_i \mathrm{id})(v)) \in W$, that is, $(f_i - \alpha_i \mathrm{id})(f(v)) \in W$.

Consider the restriction g_r of f_r to V_{r-1}. The minimal polynomial of g_r divides the minimal polynomial of f_r, and since f_r is triangulable, just as we saw for f_1, the minimal polynomial of f_r is of the form

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

where the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f_r belong to K, so the minimal polynomial of g_r is of the same form. By Proposition 22.10, there is some vector $u_r \in V_{r-1}$ such that

(1) $u_r \notin W$.
(2) $(g_r - \alpha_r \mathrm{id})(u_r) \in W$ for some scalars $\alpha_r \in K$.

Now since $u_r \in V_{r-1}$, we have $(f_i - \alpha_i \mathrm{id})(u_r) \in W$ for $i = 1, \ldots, r-1$, so $(f_i - \alpha_i \mathrm{id})(u_r) \in W$ for $i = 1, \ldots, r$ (since g_r is the restriction of f_r), which concludes the proof of the induction step. Finally, since every $f \in \mathcal{F}$ is the linear combination of $(f_1, \ldots, f_r)$, Condition (2) of the inductive claim implies Condition (2) of the proposition. □

We can now prove the following result.

Proposition 22.13. *Let $\mathcal{F}$ be a nonempty commuting family of triangulable linear maps on a finite-dimensional vector space E. There exists a basis of E such that every linear map in $\mathcal{F}$ is represented in that basis by an upper triangular matrix.*

Proof. Let $n = \dim(E)$. We construct inductively a basis $(u_1, \ldots, u_n)$ of E such that if W_i is the subspace spanned by $(u_1 \ldots, u_i)$, then for every $f \in \mathcal{F}$,

$$f(u_i) = a^f_{1i} u_1 + \cdots + a^f_{ii} u_i,$$

for some $a^f_{ij} \in K$; that is, $f(u_i)$ belongs to the subspace W_i.

We begin by applying Proposition 22.12 to the subspace $W_0 = (0)$ to get u_1 so that for all $f \in \mathcal{F}$,

$$f(u_1) = \alpha^f_1 u_1.$$

For the induction step, since W_i invariant under $\mathcal{F}$, we apply Proposition 22.12 to the subspace W_i, to get $u_{i+1} \in E$ such that

(1) $u_{i+1} \notin W_i$.
(2) For every $f \in \mathcal{F}$, the vector $f(u_{i+1})$ belong to the subspace spanned by W_i and u_{i+1}.

Condition (1) implies that $(u_1, \ldots, u_i, u_{i+1})$ is linearly independent, and Condition (2) means that for every $f \in \mathcal{F}$,

$$f(u_{i+1}) = a^f_{1i+1} u_1 + \cdots + a^f_{i+1i+1} u_{i+1},$$

for some $a^f_{i+1j} \in K$, establishing the induction step. After n steps, each $f \in \mathcal{F}$ is represented by an upper triangular matrix. □

Observe that if $\mathcal{F}$ consists of a single linear map f and if the minimal polynomial of f is of the form

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

with all $\lambda_i \in K$, using Proposition 22.10 instead of Proposition 22.12, the proof of Proposition 22.13 yields another proof of Theorem 14.1.

22.5 The Primary Decomposition Theorem

If $f\colon E \to E$ is a linear map and $\lambda \in K$ is an eigenvalue of f, recall that the eigenspace E_λ associated with λ is the kernel of the linear map $\lambda \mathrm{id} - f$. If all the eigenvalues $\lambda_1 \ldots, \lambda_k$ of f are in K, it may happen that

$$E = E_{\lambda_1} \oplus \cdots \oplus E_{\lambda_k},$$

but in general there are not enough eigenvectors to span E. What if we generalize the notion of eigenvector and look for (nonzero) vectors u such that

$$(\lambda \mathrm{id} - f)^r(u) = 0, \quad \text{for some } r \geq 1?$$

It turns out that if the minimal polynomial of f is of the form

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

then $r = r_i$ does the job for λ_i; that is, if we let

$$W_i = \mathrm{Ker}\,(\lambda_i \mathrm{id} - f)^{r_i},$$

then

$$E = W_1 \oplus \cdots \oplus W_k.$$

This result is very nice but seems to require that the eigenvalues of f all belong to K. Actually, it is a special case of a more general result involving the factorization of the minimal polynomial m into its irreducible monic factors (see Theorem 22.1),

$$m = p_1^{r_1} \cdots p_k^{r_k},$$

where the p_i are distinct irreducible monic polynomials over K.

Theorem 22.3. *(Primary Decomposition Theorem) Let $f\colon E \to E$ be a linear map on the finite-dimensional vector space E over the field K. Write the minimal polynomial m of f as*

$$m = p_1^{r_1} \cdots p_k^{r_k},$$

where the p_i are distinct irreducible monic polynomials over K, and the r_i are positive integers. Let

$$W_i = \mathrm{Ker}\,(p_i^{r_i}(f)), \quad i = 1, \ldots, k.$$

Then

(a) $E = W_1 \oplus \cdots \oplus W_k$.
(b) Each W_i *is invariant under* f.
(c) The minimal polynomial of the restriction $f \mid W_i$ *of* f *to* W_i *is* $p_i^{r_i}$.

Proof. The trick is to construct projections π_i using the polynomials $p_j^{r_j}$ so that the range of π_i is equal to W_i. Let

$$g_i = m/p_i^{r_i} = \prod_{j \neq i} p_j^{r_j}.$$

Note that

$$p_i^{r_i} g_i = m.$$

Since $p_1, \ldots, p_k$ are irreducible and distinct, they are relatively prime. Then using Proposition 22.4, it is easy to show that $g_1, \ldots, g_k$ are relatively prime. Otherwise, some irreducible polynomial p would divide all of $g_1, \ldots, g_k$, so by Proposition 22.4 it would be equal to one of the irreducible factors p_i. But that p_i is missing from g_i, a contradiction. Therefore, by Proposition 22.5, there exist some polynomials $h_1, \ldots, h_k$ such that

$$g_1 h_1 + \cdots + g_k h_k = 1.$$

Let $q_i = g_i h_i$ and let $\pi_i = q_i(f) = g_i(f) h_i(f)$. We have

$$q_1 + \cdots + q_k = 1,$$

and since m divides $q_i q_j$ for $i \neq j$, we get

$$\pi_1 + \cdots + \pi_k = \mathrm{id}$$
$$\pi_i \pi_j = 0, \quad i \neq j.$$

(We implicitly used the fact that if p, q are two polynomials, the linear maps $p(f) \circ q(f)$ and $q(f) \circ p(f)$ are the same since $p(f)$ and $q(f)$ are polynomials in the powers of f, which commute.) Composing the first equation with π_i and using the second equation, we get

$$\pi_i^2 = \pi_i.$$

Therefore, the π_i are projections, and E is the direct sum of the images of the π_i. Indeed, every $u \in E$ can be expressed as

$$u = \pi_1(u) + \cdots + \pi_k(u).$$

Also, if

$$\pi_1(u) + \cdots + \pi_k(u) = 0,$$

then by applying π_i we get

$$0 = \pi_i^2(u) = \pi_i(u), \quad i = 1, \dots k.$$

To finish proving (a), we need to show that

$$W_i = \operatorname{Ker}(p_i^{r_i}(f)) = \pi_i(E).$$

If $v \in \pi_i(E)$, then $v = \pi_i(u)$ for some $u \in E$, so

$$\begin{aligned} p_i^{r_i}(f)(v) &= p_i^{r_i}(f)(\pi_i(u)) \\ &= p_i^{r_i}(f)g_i(f)h_i(f)(u) \\ &= h_i(f)p_i^{r_i}(f)g_i(f)(u) \\ &= h_i(f)m(f)(u) = 0, \end{aligned}$$

because m is the minimal polynomial of f. Therefore, $v \in W_i$.

Conversely, assume that $v \in W_i = \operatorname{Ker}(p_i^{r_i}(f))$. If $j \neq i$, then $g_j h_j$ is divisible by $p_i^{r_i}$, so

$$g_j(f)h_j(f)(v) = \pi_j(v) = 0, \quad j \neq i.$$

Then since $\pi_1 + \cdots + \pi_k = \mathrm{id}$, we have $v = \pi_i v$, which shows that v is in the range of π_i. Therefore, $W_i = \operatorname{Im}(\pi_i)$, and this finishes the proof of (a).

If $p_i^{r_i}(f)(u) = 0$, then $p_i^{r_i}(f)(f(u)) = f(p_i^{r_i}(f)(u)) = 0$, so (b) holds.

If we write $f_i = f \mid W_i$, then $p_i^{r_i}(f_i) = 0$, because $p_i^{r_i}(f) = 0$ on W_i (its kernel). Therefore, the minimal polynomial of f_i divides $p_i^{r_i}$. Conversely, let q be any polynomial such that $q(f_i) = 0$ (on W_i). Since $m = p_i^{r_i} g_i$, the fact that $m(f)(u) = 0$ for all $u \in E$ shows that

$$p_i^{r_i}(f)(g_i(f)(u)) = 0, \quad u \in E,$$

and thus $\operatorname{Im}(g_i(f)) \subseteq \operatorname{Ker}(p_i^{r_i}(f)) = W_i$. Consequently, since $q(f)$ is zero on W_i,

$$q(f)g_i(f) = 0 \quad \text{for all } u \in E.$$

But then qg_i is divisible by the minimal polynomial $m = p_i^{r_i} g_i$ of f, and since $p_i^{r_i}$ and g_i are relatively prime, by Euclid's proposition, $p_i^{r_i}$ must divide q. This finishes the proof that the minimal polynomial of f_i is $p_i^{r_i}$, which is (c). □

To best understand the projection constructions of Theorem 22.3, we provide the following two explicit examples of the primary decomposition theorem.

Example 22.2. First let $f\colon \mathbb{R}^3 \to \mathbb{R}^3$ be defined as $f(x, y, z) = (y, -x, z)$. In terms of the standard basis f is represented by the 3×3 matrix

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

Then a simple calculation shows that $m_f(x) = \chi_f(x) = (x^2 + 1)(x - 1)$. Using the notation of the preceding proof set

$$m = p_1 p_2, \qquad p_1 = x^2 + 1, \qquad p_2 = x - 1.$$

Then

$$g_1 = \frac{m}{p_1} = x - 1, \qquad g_2 = \frac{m}{p_2} = x^2 + 1.$$

We must find $h_1, h_2 \in \mathbb{R}[x]$ such that $g_1 h_1 + g_2 h_2 = 1$. In general this is the hard part of the projection construction. But since we are only working with two relatively prime polynomials g_1, g_2, we may apply the Euclidean algorithm to discover that

$$-\frac{x+1}{2}(x - 1) + \frac{1}{2}(x^2 + 1) = 1,$$

where $h_1 = -\frac{x+1}{2}$ while $h_2 = \frac{1}{2}$. By definition

$$\pi_1 = g_1(f)h_1(f) = -\frac{1}{2}(X_f - \mathrm{id})(X_f + \mathrm{id}) = -\frac{1}{2}(X_f^2 - \mathrm{id}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and

$$\pi_2 = g_2(f)h_2(f) = \frac{1}{2}(X_f^2 + \mathrm{id}) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then $\mathbb{R}^3 = W_1 \oplus W_2$, where

$$W_1 = \pi_1(\mathbb{R}^3) = \mathrm{Ker}\,(p_1(X_f)) = \mathrm{Ker}\,(X_f^2 + \mathrm{id}) = \mathrm{Ker} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \{(x, y, 0) \in \mathbb{R}^3\},$$

$$W_2 = \pi_2(\mathbb{R}^3) = \mathrm{Ker}\,(p_2(X_f)) = \mathrm{Ker}\,(X_f - \mathrm{id}) = \mathrm{Ker} \begin{pmatrix} -1 & -1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \{(0, 0, z) \in \mathbb{R}^3\}.$$

Example 22.3. For our second example of the primary decomposition theorem let $f\colon \mathbb{R}^3 \to \mathbb{R}^3$ be defined as $f(x, y, z) = (y, -x + z, -y)$, with standard matrix representation $X_f = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$. A simple calculation shows that $m_f(x) = \chi_f(x) = x(x^2 + 2)$. Set

$$p_1 = x^2 + 2, \qquad p_2 = x, \qquad g_1 = \frac{m_f}{p_1} = x, \qquad g_2 = \frac{m_f}{p_2} = x^2 + 2.$$

Since $\gcd(g_1, g_2) = 1$, we use the Euclidean algorithm to find

$$h_1 = -\frac{1}{2}x, \qquad h_2 = \frac{1}{2},$$

such that $g_1h_1 + g_2h_2 = 1$. Then

$$\pi_1 = g_1(f)h_1(f) = -\frac{1}{2}X_f^2 = \begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 1 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix},$$

while

$$\pi_2 = g_2(f)h_2(f) = \frac{1}{2}(X_f^2 + 2\mathrm{id}) = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}.$$

Although it is not entirely obvious, π_1 and π_2 are indeed projections since

$$\pi_1^2 = \begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 1 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 1 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 1 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} = \pi_1,$$

and

$$\pi_2^2 = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} = \pi_2.$$

Furthermore observe that $\pi_1 + \pi_2 = \mathrm{id}$. The primary decomposition theorem implies that $\mathbb{R}^3 = W_1 \oplus W_2$ where

$$\begin{aligned} W_1 = \pi_1(\mathbb{R}^3) = \operatorname{Ker}(p_1(f)) &= \operatorname{Ker}(X^2 + 2) \\ &= \operatorname{Ker} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix} = \operatorname{span}\{(0, 1, 0), (1, 0, -1)\}, \end{aligned}$$

$$W_2 = \pi_2(\mathbb{R}^3) = \operatorname{Ker}(p_2(f)) = \operatorname{Ker}(X) = \operatorname{span}\{(1, 0, 1)\}.$$

See Figure 22.1.

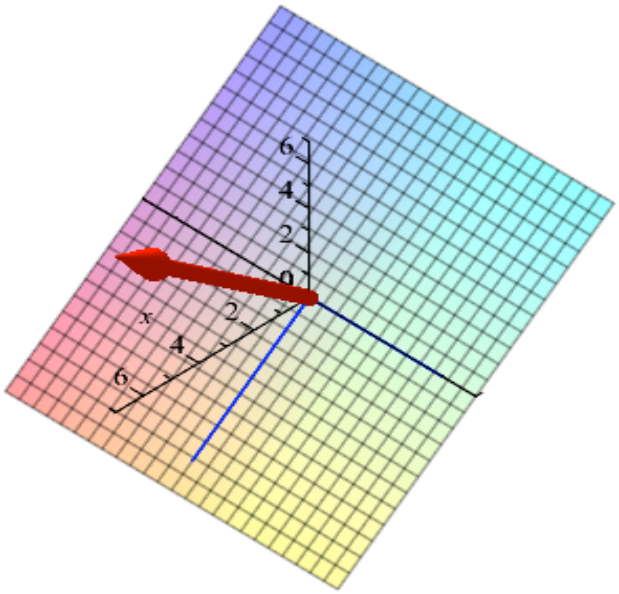

Fig. 22.1 The direct sum decomposition of $\mathbb{R}^3 = W_1 \oplus W_2$ where W_1 is the plane $x + z = 0$ and W_2 is line $t(1, 0, 1)$. The spanning vectors of W_1 are in blue.

If all the eigenvalues of f belong to the field K, we obtain the following result.

Theorem 22.4. *(Primary Decomposition Theorem, Version 2) Let $f\colon E \to E$ be a linear map on the finite-dimensional vector space E over the field K. If all the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f belong to K, write*

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k}$$

for the minimal polynomial of f,

$$\chi_f = (X - \lambda_1)^{n_1} \cdots (X - \lambda_k)^{n_k}$$

for the characteristic polynomial of f, with $1 \leq r_i \leq n_i$, and let

$$W_i = \operatorname{Ker}(\lambda_i \mathrm{id} - f)^{r_i}, \quad i = 1, \ldots, k.$$

Then

(a) $E = W_1 \oplus \cdots \oplus W_k$.
(b) Each W_i is invariant under f.
(c) $\dim(W_i) = n_i$.

(d) The minimal polynomial of the restriction $f \mid W_i$ of f to W_i is $(X - \lambda_i)^{r_i}$.

Proof. Parts (a), (b) and (d) have already been proven in Theorem 22.3, so it remains to prove (c). Since W_i is invariant under f, let f_i be the restriction of f to W_i. The characteristic polynomial χ_{f_i} of f_i divides $\chi(f)$, and since $\chi(f)$ has all its roots in K, so does $\chi_i(f)$. By Theorem 14.1, there is a basis of W_i in which f_i is represented by an upper triangular matrix, and since $(\lambda_i \mathrm{id} - f)^{r_i} = 0$, the diagonal entries of this matrix are equal to λ_i. Consequently,

$$\chi_{f_i} = (X - \lambda_i)^{\dim(W_i)},$$

and since χ_{f_i} divides $\chi(f)$, we conclude hat

$$\dim(W_i) \leq n_i, \quad i = 1, \ldots, k.$$

Because E is the direct sum of the W_i, we have $\dim(W_1) + \cdots + \dim(W_k) = n$, and since $n_1 + \cdots + n_k = n$, we must have

$$\dim(W_i) = n_i, \quad i = 1, \ldots, k,$$

proving (c). □

Definition 22.10. If $\lambda \in K$ is an eigenvalue of f, we define a *generalized eigenvector* of f as a nonzero vector $u \in E$ such that

$$(\lambda \mathrm{id} - f)^r(u) = 0, \quad \text{for some } r \geq 1.$$

The *index* of λ is defined as the smallest $r \geq 1$ such that

$$\mathrm{Ker}\,(\lambda \mathrm{id} - f)^r = \mathrm{Ker}\,(\lambda \mathrm{id} - f)^{r+1}.$$

It is clear that $\mathrm{Ker}\,(\lambda \mathrm{id} - f)^i \subseteq \mathrm{Ker}\,(\lambda \mathrm{id} - f)^{i+1}$ for all $i \geq 1$. By Theorem 22.4(d), if $\lambda = \lambda_i$, the index of λ_i is equal to r_i.

22.6 Jordan Decomposition

Recall that a linear map $g\colon E \to E$ is said to be *nilpotent* if there is some positive integer r such that $g^r = 0$. Another important consequence of Theorem 22.4 is that f can be written as the sum of a diagonalizable and a nilpotent linear map (which commute). For example $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ be the $\mathbb{R}$-linear map $f(x, y) = (x, x + y)$ with standard matrix representation $X_f = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. A basic calculation shows that $m_f(x) = \chi_f(x) = (x - 1)^2$. By Theorem 22.2 we know that f is not diagonalizable over $\mathbb{R}$. But since

the eigenvalue $\lambda_1 = 1$ of f does belong to $\mathbb{R}$, we may use the projection construction inherent within Theorem 22.4 to write $f = D + N$, where D is a diagonalizable linear map and N is a nilpotent linear map. The proof of Theorem 22.3 implies that

$$p_1^{r_1} = (x-1)^2, \qquad g_1 = 1 = h_1, \qquad \pi_1 = g_1(f)h_1(f) = \mathrm{id}.$$

Then

$$D = \lambda_1\pi_1 = \mathrm{id},$$
$$N = f - D = f(x,y) - \mathrm{id}(x,y) = (x, x+y) - (x,y) = (0,y),$$

which is equivalent to the matrix decomposition

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

This example suggests that the diagonal summand of f is related to the projection constructions associated with the proof of the primary decomposition theorem. If we write

$$D = \lambda_1\pi_1 + \cdots + \lambda_k\pi_k,$$

where π_i is the projection from E onto the subspace W_i defined in the proof of Theorem 22.3, since

$$\pi_1 + \cdots + \pi_k = \mathrm{id},$$

we have

$$f = f\pi_1 + \cdots + f\pi_k,$$

and so we get

$$N = f - D = (f - \lambda_1\mathrm{id})\pi_1 + \cdots + (f - \lambda_k\mathrm{id})\pi_k.$$

We claim that $N = f - D$ is a nilpotent operator. Since by construction the π_i are polynomials in f, they commute with f, using the properties of the π_i, we get

$$N^r = (f - \lambda_1\mathrm{id})^r\pi_1 + \cdots + (f - \lambda_k\mathrm{id})^r\pi_k.$$

Therefore, if $r = \max\{r_i\}$, we have $(f - \lambda_k\mathrm{id})^r = 0$ for $i = 1, \ldots, k$, which implies that

$$N^r = 0.$$

It remains to show that D is diagonalizable. Since N is a polynomial in f, it commutes with f, and thus with D. From

$$D = \lambda_1\pi_1 + \cdots + \lambda_k\pi_k,$$

and

$$\pi_1 + \cdots + \pi_k = \mathrm{id},$$

we see that

$$\begin{aligned} D - \lambda_i \mathrm{id} &= \lambda_1\pi_1 + \cdots + \lambda_k\pi_k - \lambda_i(\pi_1 + \cdots + \pi_k) \\ &= (\lambda_1 - \lambda_i)\pi_1 + \cdots + (\lambda_{i-1} - \lambda_i)\pi_{i-1} + (\lambda_{i+1} - \lambda_i)\pi_{i+1} \\ &\quad + \cdots + (\lambda_k - \lambda_i)\pi_k. \end{aligned}$$

Since the projections π_j with $j \neq i$ vanish on W_i, the above equation implies that $D - \lambda_i \mathrm{id}$ vanishes on W_i and that $(D - \lambda_j \mathrm{id})(W_i) \subseteq W_i$, and thus that the minimal polynomial of D is

$$(X - \lambda_1) \cdots (X - \lambda_k).$$

Since the λ_i are distinct, by Theorem 22.2, the linear map D is diagonalizable.

In summary we have shown that when all the eigenvalues of f belong to K, there exist a diagonalizable linear map D and a nilpotent linear map N such that

$$\begin{aligned} f &= D + N \\ DN &= ND, \end{aligned}$$

and N and D are polynomials in f.

Definition 22.11. A decomposition of f as $f = D + N$ as above is called a *Jordan decomposition*.

In fact, we can prove more: the maps D and N are uniquely determined by f.

Theorem 22.5. *(Jordan Decomposition) Let $f : E \to E$ be a linear map on the finite-dimensional vector space E over the field K. If all the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f belong to K, then there exist a diagonalizable linear map D and a nilpotent linear map N such that*

$$\begin{aligned} f &= D + N \\ DN &= ND. \end{aligned}$$

Furthermore, D and N are uniquely determined by the above equations and they are polynomials in f.

Proof. We already proved the existence part. Suppose we also have $f = D' + N'$, with $D'N' = N'D'$, where D' is diagonalizable, N' is nilpotent, and both are polynomials in f. We need to prove that $D = D'$ and $N = N'$.

Since D' and N' commute with one another and $f = D' + N'$, we see that D' and N' commute with f. Then D' and N' commute with any polynomial in f; hence they commute with D and N. From

$$D + N = D' + N',$$

we get

$$D - D' = N' - N,$$

and D, D', N, N' commute with one another. Since D and D' are both diagonalizable and commute, by Proposition 22.11, they are simultaneousy diagonalizable, so $D - D'$ is diagonalizable. Since N and N' commute, by the binomial formula, for any $r \geq 1$,

$$(N' - N)^r = \sum_{j=0}^{r} (-1)^j \binom{r}{j} (N')^{r-j} N^j.$$

Since both N and N' are nilpotent, we have $N^{r_1} = 0$ and $(N')^{r_2} = 0$, for some $r_1, r_2 > 0$, so for $r \geq r_1 + r_2$, the right-hand side of the above expression is zero, which shows that $N' - N$ is nilpotent. (In fact, it is easy that $r_1 = r_2 = n$ works). It follows that $D - D' = N' - N$ is both diagonalizable and nilpotent. Clearly, the minimal polynomial of a nilpotent linear map is of the form X^r for some $r > 0$ (and $r \leq \dim(E)$). But $D - D'$ is diagonalizable, so its minimal polynomial has simple roots, which means that $r = 1$. Therefore, the minimal polynomial of $D - D'$ is X, which says that $D - D' = 0$, and then $N = N'$. □

If K is an algebraically closed field, then Theorem 22.5 holds. This is the case when $K = \mathbb{C}$. This theorem reduces the study of linear maps (from E to itself) to the study of nilpotent operators. There is a special normal form for such operators which is discussed in the next section.

22.7 Nilpotent Linear Maps and Jordan Form

This section is devoted to a normal form for nilpotent maps. We follow Godement's exposition [Godement (1963)]. Let $f\colon E \to E$ be a nilpotent linear map on a finite-dimensional vector space over a field K, and assume that f is not the zero map. There is a smallest positive integer $r \geq 1$ such $f^r \neq 0$ and $f^{r+1} = 0$. Clearly, the polynomial X^{r+1} annihilates

f, and it is the minimal polynomial of f since $f^r \neq 0$. It follows that $r+1 \leq n = \dim(E)$. Let us define the subspaces N_i by

$$N_i = \mathrm{Ker}\,(f^i), \quad i \geq 0.$$

Note that $N_0 = (0)$, $N_1 = \mathrm{Ker}\,(f)$, and $N_{r+1} = E$. Also, it is obvious that

$$N_i \subseteq N_{i+1}, \quad i \geq 0.$$

Proposition 22.14. *Given a nilpotent linear map f with $f^r \neq 0$ and $f^{r+1} = 0$ as above, the inclusions in the following sequence are strict:*

$$(0) = N_0 \subset N_1 \subset \cdots \subset N_r \subset N_{r+1} = E.$$

Proof. We proceed by contradiction. Assume that $N_i = N_{i+1}$ for some i with $0 \leq i \leq r$. Since $f^{r+1} = 0$, for every $u \in E$, we have

$$0 = f^{r+1}(u) = f^{i+1}(f^{r-i}(u)),$$

which shows that $f^{r-i}(u) \in N_{i+1}$. Since $N_i = N_{i+1}$, we get $f^{r-i}(u) \in N_i$, and thus $f^r(u) = 0$. Since this holds for all $u \in E$, we see that $f^r = 0$, a contradiction. □

Proposition 22.15. *Given a nilpotent linear map f with $f^r \neq 0$ and $f^{r+1} = 0$, for any integer i with $1 \leq i \leq r$, for any subspace U of E, if $U \cap N_i = (0)$, then $f(U) \cap N_{i-1} = (0)$, and the restriction of f to U is an isomorphism onto $f(U)$.*

Proof. Pick $v \in f(U) \cap N_{i-1}$. We have $v = f(u)$ for some $u \in U$ and $f^{i-1}(v) = 0$, which means that $f^i(u) = 0$. Then $u \in U \cap N_i$, so $u = 0$ since $U \cap N_i = (0)$, and $v = f(u) = 0$. Therefore, $f(U) \cap N_{i-1} = (0)$. The restriction of f to U is obviously surjective on $f(U)$. Suppose that $f(u) = 0$ for some $u \in U$. Then $u \in U \cap N_1 \subseteq U \cap N_i = (0)$ (since $i \geq 1$), so $u = 0$, which proves that f is also injective on U. □

Proposition 22.16. *Given a nilpotent linear map f with $f^r \neq 0$ and $f^{r+1} = 0$, there exists a sequence of subspace $U_1, \ldots, U_{r+1}$ of E with the following properties:*

(1) $N_i = N_{i-1} \oplus U_i$, for $i = 1, \ldots, r+1$.
(2) We have $f(U_i) \subseteq U_{i-1}$, and the restriction of f to U_i is an injection, for $i = 2, \ldots, r+1$.

See Figure 22.2.

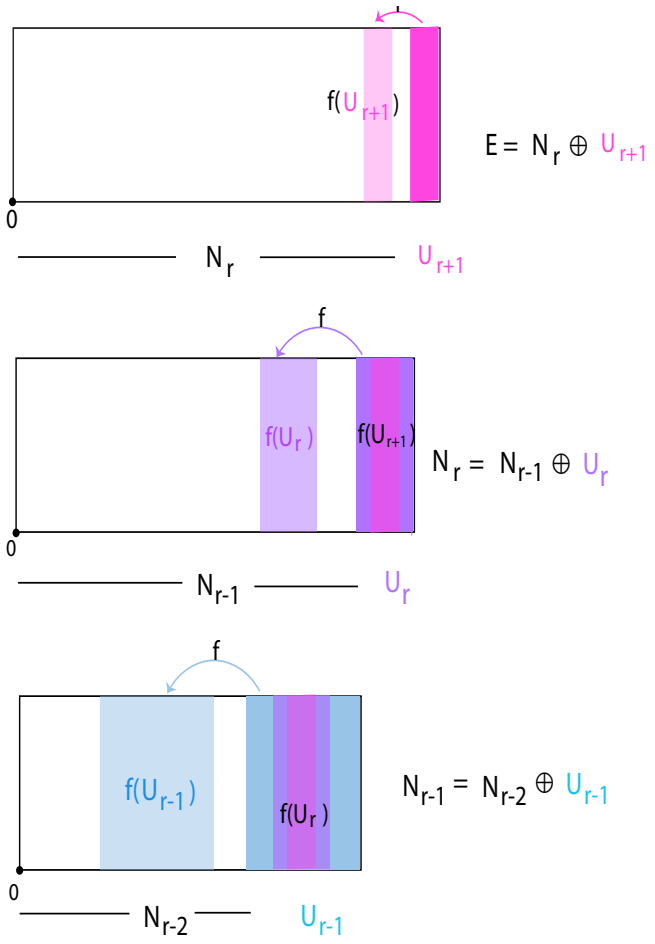

Fig. 22.2 A schematic illustration of $N_i = N_{i-1} \oplus U_i$ with $f(U_i) \subseteq U_{i-1}$ for $i = r+1, r, r-1$.

Proof. We proceed inductively, by defining the sequence $U_{r+1}, U_r, \ldots, U_1$. We pick U_{r+1} to be any supplement of N_r in $N_{r+1} = E$, so that

$$E = N_{r+1} = N_r \oplus U_{r+1}.$$

Since $f^{r+1} = 0$ and $N_r = \mathrm{Ker}\,(f^r)$, we have $f(U_{r+1}) \subseteq N_r$, and by Proposition 22.15, as $U_{r+1} \cap N_r = (0)$, we have $f(U_{r+1}) \cap N_{r-1} = (0)$. As a consequence, we can pick a supplement U_r of N_{r-1} in N_r so that $f(U_{r+1}) \subseteq U_r$.

We have

$$N_r = N_{r-1} \oplus U_r \quad \text{and} \quad f(U_{r+1}) \subseteq U_r.$$

By Proposition 22.15, f is an injection from U_{r+1} to U_r. Assume inductively that $U_{r+1}, \ldots, U_i$ have been defined for $i \geq 2$ and that they satisfy (1) and (2). Since

$$N_i = N_{i-1} \oplus U_i,$$

we have $U_i \subseteq N_i$, so $f^{i-1}(f(U_i)) = f^i(U_i) = (0)$, which implies that $f(U_i) \subseteq N_{i-1}$. Also, since $U_i \cap N_{i-1} = (0)$, by Proposition 22.15, we have $f(U_i) \cap N_{i-2} = (0)$. It follows that there is a supplement U_{i-1} of N_{i-2} in N_{i-1} that contains $f(U_i)$. We have

$$N_{i-1} = N_{i-2} \oplus U_{i-1} \quad \text{and} \quad f(U_i) \subseteq U_{i-1}.$$

The fact that f is an injection from U_i into U_{i-1} follows from Proposition 22.15. Therefore, the induction step is proven. The construction stops when $i = 1$. □

Because $N_0 = (0)$ and $N_{r+1} = E$, we see that E is the direct sum of the U_i:

$$E = U_1 \oplus \cdots \oplus U_{r+1},$$

with $f(U_i) \subseteq U_{i-1}$, and f an injection from U_i to U_{i-1}, for $i = r+1, \ldots, 2$. By a clever choice of bases in the U_i, we obtain the following nice theorem.

Theorem 22.6. *For any nilpotent linear map $f\colon E \to E$ on a finite-dimensional vector space E of dimension n over a field K, there is a basis of E such that the matrix N of f is of the form*

$$N = \begin{pmatrix} 0 & \nu_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \nu_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \nu_n \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix},$$

where $\nu_i = 1$ or $\nu_i = 0$.

Proof. First apply Proposition 22.16 to obtain a direct sum $E = \bigoplus_{i=1}^{r+1} U_i$. Then we define a basis of E inductively as follows. First we choose a basis

$$e_1^{r+1}, \ldots, e_{n_{r+1}}^{r+1}$$

of U_{r+1}. Next, for $i = r+1, \ldots, 2$, given the basis

$$e_1^i, \ldots, e_{n_i}^i$$

of U_i, since f is injective on U_i and $f(U_i) \subseteq U_{i-1}$, the vectors $f(e_1^i), \ldots, f(e_{n_i}^i)$ are linearly independent, so we define a basis of U_{i-1} by completing $f(e_1^i), \ldots, f(e_{n_i}^i)$ to a basis in U_{i-1}:

$$e_1^{i-1}, \ldots, e_{n_i}^{i-1}, e_{n_i+1}^{i-1}, \ldots, e_{n_{i-1}}^{i-1}$$

with

$$e_j^{i-1} = f(e_j^i), \quad j = 1 \ldots, n_i.$$

Since $U_1 = N_1 = \mathrm{Ker}\,(f)$, we have

$$f(e_j^1) = 0, \quad j = 1, \ldots, n_1.$$

These basis vectors can be arranged as the rows of the following matrix:

$$\begin{pmatrix}
e_1^{r+1} & \cdots & e_{n_{r+1}}^{r+1} & & & & & & & & \\
\vdots & & \vdots & & & & & & & & \\
e_1^r & \cdots & e_{n_{r+1}}^r & e_{n_{r+1}+1}^r & \cdots & e_{n_r}^r & & & & & \\
\vdots & & \vdots & \vdots & & \vdots & & & & & \\
e_1^{r-1} & \cdots & e_{n_{r+1}}^{r-1} & e_{n_{r+1}+1}^{r-1} & \cdots & e_{n_r}^{r-1} & e_{n_r+1}^{r-1} & \cdots & e_{n_{r-1}}^{r-1} & & \\
\vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & & \\
\vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & & \\
e_1^1 & \cdots & e_{n_{r+1}}^1 & e_{n_{r+1}+1}^1 & \cdots & e_{n_r}^1 & e_{n_r+1}^1 & \cdots & e_{n_{r-1}}^1 & \cdots \cdots & e_{n_1}^1
\end{pmatrix}$$

Finally, we define the basis $(e_1, \ldots, e_n)$ by listing each column of the above matrix from the bottom-up, starting with column one, then column two, *etc.* This means that we list the vectors e_j^i in the following order:

For $j = 1, \ldots, n_{r+1}$, list $e_j^1, \ldots, e_j^{r+1}$;

In general, for $i = r, \ldots, 1$,

for $j = n_{i+1} + 1, \ldots, n_i$, list $e_j^1, \ldots, e_j^i$.

Then because $f(e_j^1) = 0$ and $e_j^{i-1} = f(e_j^i)$ for $i \geq 2$, either

$$f(e_i) = 0 \quad \text{or} \quad f(e_i) = e_{i-1},$$

which proves the theorem. □

As an application of Theorem 22.6, we obtain the *Jordan form* of a linear map.

Definition 22.12. A *Jordan block* is an $r \times r$ matrix $J_r(\lambda)$, of the form

$$J_r(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 1 \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix},$$

where $\lambda \in K$, with $J_1(\lambda) = (\lambda)$ if $r = 1$. A *Jordan matrix*, J, is an $n \times n$ block diagonal matrix of the form

$$J = \begin{pmatrix} J_{r_1}(\lambda_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & J_{r_m}(\lambda_m) \end{pmatrix},$$

where each $J_{r_k}(\lambda_k)$ is a Jordan block associated with some $\lambda_k \in K$, and with $r_1 + \cdots + r_m = n$.

To simplify notation, we often write $J(\lambda)$ for $J_r(\lambda)$. Here is an example of a Jordan matrix with four blocks:

$$J = \begin{pmatrix} \lambda & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mu & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix}.$$

Theorem 22.7. *(Jordan form) Let E be a vector space of dimension n over a field K and let $f : E \to E$ be a linear map. The following properties are equivalent:*

(1) The eigenvalues of f all belong to K (i.e. the roots of the characteristic polynomial χ_f all belong to K).
(2) There is a basis of E in which the matrix of f is a Jordan matrix.

Proof. Assume (1). First we apply Theorem 22.4, and we get a direct sum $E = \bigoplus_{j=1}^{k} W_k$, such that the restriction of $g_i = f - \lambda_j \mathrm{id}$ to W_i is

nilpotent. By Theorem 22.6, there is a basis of W_i such that the matrix of the restriction of g_i is of the form

$$G_i = \begin{pmatrix} 0 & \nu_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \nu_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \nu_{n_i} \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix},$$

where $\nu_i = 1$ or $\nu_i = 0$. Furthermore, over any basis, $\lambda_i \mathrm{id}$ is represented by the diagonal matrix D_i with λ_i on the diagonal. Then it is clear that we can split $D_i + G_i$ into Jordan blocks by forming a Jordan block for every uninterrupted chain of 1s. By putting the bases of the W_i together, we obtain a matrix in Jordan form for f.

Now assume (2). If f can be represented by a Jordan matrix, it is obvious that the diagonal entries are the eigenvalues of f, so they all belong to K. □

Observe that Theorem 22.7 applies if $K = \mathbb{C}$. It turns out that there are uniqueness properties of the Jordan blocks but more machinery is needed to prove this result.

If a complex $n \times n$ matrix A is expressed in terms of its Jordan decomposition as $A = D + N$, since D and N commute, by Proposition 8.16, the exponential of A is given by

$$e^A = e^D e^N,$$

and since N is an $n \times n$ nilpotent matrix, $N^{n-1} = 0$, so we obtain

$$e^A = e^D \left(I + \frac{N}{1!} + \frac{N^2}{2!} + \cdots + \frac{N^{n-1}}{(n-1)!} \right).$$

In particular, the above applies if A is a Jordan matrix. This fact can be used to solve (at least in theory) systems of first-order linear differential equations. Such systems are of the form

$$\frac{dX}{dt} = AX, \tag{$*$}$$

where A is an $n \times n$ matrix and X is an n-dimensional vector of functions of the parameter t.

It can be shown that the columns of the matrix e^{tA} form a basis of the vector space of solutions of the system of linear differential equations ($*$); see Artin [Artin (1991)] (Chapter 4). Furthermore, for any matrix B and

any invertible matrix P, if $A = PBP^{-1}$, then the system $(*)$ is equivalent to

$$P^{-1}\frac{dX}{dt} = BP^{-1}X,$$

so if we make the change of variable $Y = P^{-1}X$, we obtain the system

$$\frac{dY}{dt} = BY. \tag{$**$}$$

Consequently, if B is such that the exponential e^{tB} can be easily computed, we obtain an explicit solution Y of $(**)$, and $X = PY$ is an explicit solution of $(*)$. This is the case when B is a Jordan form of A. In this case, it suffices to consider the Jordan blocks of B. Then we have

$$J_r(\lambda) = \lambda I_r + \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} = \lambda I_r + N,$$

and the powers N^k are easily computed.

For example, if

$$B = \begin{pmatrix} 3 & 1 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{pmatrix} = 3I_3 + \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

we obtain

$$tB = t\begin{pmatrix} 3 & 1 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{pmatrix} = 3tI_3 + \begin{pmatrix} 0 & t & 0 \\ 0 & 0 & t \\ 0 & 0 & 0 \end{pmatrix}$$

and so

$$e^{tB} = \begin{pmatrix} e^{3t} & 0 & 0 \\ 0 & e^{3t} & 0 \\ 0 & 0 & e^{3t} \end{pmatrix} \begin{pmatrix} 1 & t & (1/2)t^2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} e^{3t} & te^{3t} & (1/2)t^2e^{3t} \\ 0 & e^{3t} & te^{3t} \\ 0 & 0 & e^{3t} \end{pmatrix}.$$

The columns of e^{tB} form a basis of the space of solutions of the system of linear differential equations

$$\begin{aligned} \frac{dY_1}{dt} &= 3Y_1 + Y_2 \\ \frac{dY_2}{dt} &= 3Y_2 + Y_3 \\ \frac{dY_3}{dt} &= 3Y_3, \end{aligned}$$

in matrix form,

$$\begin{pmatrix} \frac{dY_1}{dt} \\ \frac{dY_2}{dt} \\ \frac{dY_3}{dt} \end{pmatrix} = \begin{pmatrix} 3 & 1 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}.$$

Explicitly, the general solution of the above system is

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = c_1 \begin{pmatrix} e^{3t} \\ 0 \\ 0 \end{pmatrix} + c_2 \begin{pmatrix} te^{3t} \\ e^{3t} \\ 0 \end{pmatrix} + c_3 \begin{pmatrix} (1/2)t^2 e^{3t} \\ te^{3t} \\ e^{3t} \end{pmatrix},$$

with $c_1, c_2, c_3 \in \mathbb{R}$.

Solving systems of first-order linear differential equations is discussed in Artin [Artin (1991)] and more extensively in Hirsh and Smale [Hirsh and Smale (1974)].

22.8 Summary

The main concepts and results of this chapter are listed below:

- Ideals, principal ideals, greatest common divisors.
- Monic polynomial, irreducible polynomial, relatively prime polynomials.
- Annihilator of a linear map.
- Minimal polynomial of a linear map.
- Invariant subspace.
- f-conductor of u into W; conductor of u into W.
- Diagonalizable linear maps.
- Commuting families of linear maps.
- Primary decomposition.
- Generalized eigenvectors.
- Nilpotent linear map.
- Normal form of a nilpotent linear map.
- Jordan decomposition.
- Jordan block.
- Jordan matrix.
- Jordan normal form.
- Systems of first-order linear differential equations.

22.9 Problems

Problem 22.1. Prove that the minimal monic polynomial of Proposition 22.1 is unique.

Problem 22.2. Given a linear map $f: E \to E$, prove that the set $\mathrm{Ann}(f)$ of polynomials that annihilate f is an ideal.

Problem 22.3. Provide the details of Proposition 22.8.

Problem 22.4. Prove that the f-conductor $S_f(u, W)$ is an ideal in $K[X]$ (Proposition 22.9).

Problem 22.5. Prove that the polynomials $g_1, \ldots, g_k$ used in the proof of Theorem 22.3 are relatively prime.

Problem 22.6. Find the minimal polynomial of the matrix

$$A = \begin{pmatrix} 6 & -3 & -2 \\ 4 & -1 & -2 \\ 10 & -5 & -3 \end{pmatrix}.$$

Problem 22.7. Find the Jordan decomposition of the matrix

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

Problem 22.8. Let $f: E \to E$ be a linear map on a finite-dimensional vector space. Prove that if f has rank 1, then either f is diagonalizable or f is nilpotent but not both.

Problem 22.9. Find the Jordan form of the matrix

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

Problem 22.10. Let N be a 3×3 nilpotent matrix over $\mathbb{C}$. Prove that the matrix
$A = I + (1/2)N - (1/8)N^2$ satisfies the equation

$$A^2 = I + N.$$

In other words, A is a square root of $I + N$.

Generalize the above fact to any $n \times n$ nilpotent matrix N over $\mathbb{C}$ using the binomial series for $(1 + t)^{1/2}$.

Problem 22.11. Let K be an algebraically closed field (for example, $K = \mathbb{C}$). Prove that every 4×4 matrix is similar to a Jordan matrix of the following form:

$$\begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix}, \quad \begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix}, \quad \begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix},$$

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

Problem 22.12. In this problem the field K is of characteristic 0. Consider an $(r \times r)$ Jordan block

$$J_r(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 1 \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix}.$$

Prove that for any polynomial $f(X)$, we have

$$f(J_r(\lambda)) = \begin{pmatrix} f(\lambda) & f_1(\lambda) & f_2(\lambda) & \cdots & f_{r-1}(\lambda) \\ 0 & f(\lambda) & f_1(\lambda) & \cdots & f_{r-2}(\lambda) \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & f_1(\lambda) \\ 0 & 0 & 0 & \cdots & f(\lambda) \end{pmatrix},$$

where

$$f_k(X) = \frac{f^{(k)}(X)}{k!},$$

and $f^{(k)}(X)$ is the kth derivative of $f(X)$.

Bibliography

Andrews, G. E., Askey, R., and Roy, R. (2000). *Special Functions*, 1st edn. (Cambridge University Press).
Artin, E. (1957). *Geometric Algebra*, 1st edn. (Wiley Interscience).
Artin, M. (1991). *Algebra*, 1st edn. (Prentice Hall).
Axler, S. (2004). *Linear Algebra Done Right*, 2nd edn., Undergraduate Texts in Mathematics (Springer Verlag).
Berger, M. (1990a). *Géométrie 1* (Nathan), english edition: Geometry 1, Universitext, Springer Verlag.
Berger, M. (1990b). *Géométrie 2* (Nathan), english edition: Geometry 2, Universitext, Springer Verlag.
Bertin, J. (1981). *Algèbre linéaire et géométrie classique*, 1st edn. (Masson).
Bourbaki, N. (1970). *Algèbre, Chapitres 1-3*, Eléments de Mathématiques (Hermann).
Bourbaki, N. (1981a). *Algèbre, Chapitres 4-7*, Eléments de Mathématiques (Masson).
Bourbaki, N. (1981b). *Espaces Vectoriels Topologiques*, Eléments de Mathématiques (Masson).
Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*, 1st edn. (Cambridge University Press).
Cagnac, G., Ramis, E., and Commeau, J. (1965). *Mathématiques Spéciales, Vol. 3, Géométrie* (Masson).
Chung, F. R. K. (1997). *Spectral Graph Theory*, *Regional Conference Series in Mathematics*, Vol. 92, 1st edn. (AMS).
Ciarlet, P. (1989). *Introduction to Numerical Matrix Analysis and Optimization*, 1st edn. (Cambridge University Press), french edition: Masson, 1994.
Coxeter, H. (1989). *Introduction to Geometry*, 2nd edn. (Wiley).
Demmel, J. W. (1997). *Applied Numerical Linear Algebra*, 1st edn. (SIAM Publications).
Dieudonné, J. (1965). *Algèbre Linéaire et Géométrie Elémentaire*, 2nd edn. (Hermann).
Dixmier, J. (1984). *General Topology*, 1st edn., UTM (Springer Verlag).
Dummit, D. S. and Foote, R. M. (1999). *Abstract Algebra*, 2nd edn. (Wiley).

Epstein, C. L. (2007). *Introduction to the Mathematics of Medical Imaging*, 2nd edn. (SIAM).

Forsyth, D. A. and Ponce, J. (2002). *Computer Vision: A Modern Approach*, 1st edn. (Prentice Hall).

Fresnel, J. (1998). *Méthodes Modernes En Géométrie*, 1st edn. (Hermann).

Gallier, J. H. (2011a). *Discrete Mathematics*, 1st edn., Universitext (Springer Verlag).

Gallier, J. H. (2011b). *Geometric Methods and Applications, For Computer Science and Engineering*, 2nd edn., TAM, Vol. 38 (Springer).

Gallier, J. H. (2019). Spectral Graph Theory of Unsigned and Signed Graphs. Applications to Graph Clustering: A survey, Tech. rep., University of Pennsylvania, http://www.cis.upenn.edu/ jean/spectral-graph-notes.pdf.

Godement, R. (1963). *Cours d'Algèbre*, 1st edn. (Hermann).

Godsil, C. and Royle, G. (2001). *Algebraic Graph Theory*, 1st edn., GTM No. 207 (Springer Verlag).

Golub, G. H. and Uhlig, F. (2009). The QR algorithm: 50 years later its genesis by john francis and vera kublanovskaya and subsequent developments, *IMA Journal of Numerical Analysis* **29**, pp. 467–485.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*, 3rd edn. (The Johns Hopkins University Press).

Hadamard, J. (1947). *Leçons de Géométrie Elémentaire. I Géométrie Plane*, thirteenth edn. (Armand Colin).

Hadamard, J. (1949). *Leçons de Géométrie Elémentaire. II Géométrie dans l'Espace*, eighth edn. (Armand Colin).

Halko, N., Martinsson, P., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, *SIAM Review* **53(2)**, pp. 217–288.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. (Springer).

Hirsh, M. W. and Smale, S. (1974). *Differential Equations, Dynamical Systems and Linear Algebra*, 1st edn. (Academic Press).

Hoffman, K. and Kunze, R. (1971). *Linear Algebra*, 2nd edn. (Prentice Hall).

Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis*, 1st edn. (Cambridge University Press).

Horn, R. A. and Johnson, C. R. (1994). *Topics in Matrix Analysis*, 1st edn. (Cambridge University Press).

Kincaid, D. and Cheney, W. (1996). *Numerical Analysis*, 2nd edn. (Brooks/Cole Publishing).

Kumpel, P. G. and Thorpe, J. A. (1983). *Linear Algebra*, 1st edn. (W. B. Saunders).

Lang, S. (1993). *Algebra*, 3rd edn. (Addison Wesley).

Lang, S. (1996). *Real and Functional Analysis*, 3rd edn., GTM 142 (Springer Verlag).

Lang, S. (1997). *Undergraduate Analysis*, 2nd edn., UTM (Springer Verlag).

Lax, P. (2007). *Linear Algebra and Its Applications*, 2nd edn. (Wiley).

Lebedev, N. N. (1972). *Special Functions and Their Applications*, 1st edn.

(Dover).
Mac Lane, S. and Birkhoff, G. (1967). *Algebra*, 1st edn. (Macmillan).
Marsden, J. E. and Hughes, T. J. (1994). *Mathematical Foundations of Elasticity*, 1st edn. (Dover).
Meyer, C. D. (2000). *Matrix Analysis and Applied Linear Algebra*, 1st edn. (SIAM).
O'Rourke, J. (1998). *Computational Geometry in C*, 2nd edn. (Cambridge University Press).
Parlett, B. N. (1997). *The Symmetric Eigenvalue Problem*, 1st edn. (SIAM Publications).
Pedoe, D. (1988). *Geometry, A comprehensive Course*, 1st edn. (Dover).
Rouché, E. and de Comberousse, C. (1900). *Traité de Géométrie*, seventh edn. (Gauthier-Villars).
Sansone, G. (1991). *Orthogonal Functions*, 1st edn. (Dover).
Schwartz, L. (1991). *Analyse I. Théorie des Ensembles et Topologie*, Collection Enseignement des Sciences (Hermann).
Schwartz, L. (1992). *Analyse II. Calcul Différentiel et Equations Différentielles*, Collection Enseignement des Sciences (Hermann).
Seberry, J., Wysocki, B. J., and Wysocki, T. A. (2005). On some applications of Hadamard matrices, *Metrika* **62**, pp. 221–239.
Serre, D. (2010). *Matrices, Theory and Applications*, 2nd edn., GTM No. 216 (Springer Verlag).
Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation, *Transactions on Pattern Analysis and Machine Intelligence* **22(8)**, pp. 888–905.
Snapper, E. and Troyer, R. J. (1989). *Metric Affine Geometry*, 1st edn. (Dover).
Spielman, D. (2012). Spectral graph theory, in U. Naumannn and O. Schenk (eds.), *Combinatorial Scientific Computing* (CRC Press).
Stewart, G. (1993). On the early history of the singular value decomposition, *SIAM review* **35(4)**, pp. 551–566.
Stollnitz, E. J., DeRose, T. D., and Salesin, D. H. (1996). *Wavelets for Computer Graphics Theory and Applications*, 1st edn. (Morgan Kaufmann).
Strang, G. (1986). *Introduction to Applied Mathematics*, 1st edn. (Wellesley-Cambridge Press).
Strang, G. (1988). *Linear Algebra and its Applications*, 3rd edn. (Saunders HBJ).
Strang, G. (2019). *Linear Algebra and Learning from Data*, 1st edn. (Wellesley-Cambridge Press).
Strang, G. and Truong, N. (1997). *Wavelets and Filter Banks*, 2nd edn. (Wellesley-Cambridge Press).
Tisseron, C. (1994). *Géométries affines, projectives, et euclidiennes*, 1st edn. (Hermann).
Trefethen, L. and Bau III, D. (1997). *Numerical Linear Algebra*, 1st edn. (SIAM Publications).
Tropp, J. A. (2011). Improved analysis of the subsampled Hadamard transform, *Advances in Adaptive Data Analysis* **3**, pp. 115–126.
Van Der Waerden, B. (1973). *Algebra, Vol. 1*, seventh edn. (Ungar).
van Lint, J. and Wilson, R. (2001). *A Course in Combinatorics*, 2nd edn. (Cam-

bridge University Press).
Veblen, O. and Young, J. W. (1946). *Projective Geometry, Vol. 2*, 1st edn. (Ginn).
Watkins, D. S. (1982). Understanding the QR algorithm, *SIAM Review* **24(4)**, pp. 447–440.
Watkins, D. S. (2008). The QR algorithm revisited, *SIAM Review* **50(1)**, pp. 133–145.
Yu, S. X. (2003). *Computational Models of Perceptual Organization*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA 15213, USA, dissertation.
Yu, S. X. and Shi, J. (2003). Multiclass spectral clustering, in *9th International Conference on Computer Vision, Nice, France, October 13-16* (IEEE).

Index