# Lecture notes on matrix analysis

Mark W. Meckes

April 27, 2019

## Contents

# 1 Linear algebra background

If you need to brush up on linear algebra background, the best source is, of course,

> *Linear Algebra*, by Elizabeth S. Meckes and Mark W. Meckes, Cambridge University Press, 2018.

## 1.1 Fundamentals

We will use $\mathbb{F}$ to stand for either the set of real numbers $\mathbb{R}$ or the set of complex numbers $\mathbb{C}$. In this class we will deal only with *finite-dimensional* vector spaces over $\mathbb{R}$ or $\mathbb{C}$.

Basics terms which you should be comfortable with:

- **vector space** over $\mathbb{F}$
- **subspace**
- **span**
- **linearly (in)dependent**
- **basis**
- **standard basis of $\mathbb{F}^n$**
- **dimension**
- **linear transformation/map**
- **matrix**
- **identity matrix**
- **identity map**
- **invertible matrix**
- **invertible linear map**
- **singular matrix**
- **singular linear map**
- **inverse matrix**
- **inverse map**
- **kernel/null space**
- **image/range**
- **determinant**

- **eigenvector**
- **eigenvalue**
- **characteristic polynomial**
- **inner product**
- **norm** (associated with an inner product)
- **standard inner product** on $\mathbb{F}^n$
- **orthogonal**
- **orthonormal basis**
- **unitary map**
- **unitary matrix**
- **orthogonal matrix**

## 1.2 Matrices and linear maps

Give a matrix $u \in V$ and a basis $\mathcal{B} = (v_1, \ldots, v_n)$ of $V$, the **matrix representing** $u$ **with respect to** $\mathcal{B}$ is the column matrix $\mathbf{x} \in M_{n,1}$ such that

$$u = \sum_{i=1}^{n} x_i v_i.$$

Given a linear map $T : V \to W$ and bases $\mathcal{B}_1 = (v_1, \ldots, v_n)$ of $V$ and $\mathcal{B}_2 = (w_1, \ldots, w_m)$ of $W$, the **matrix representing** $T$ **with respect to** $\mathcal{B}_1$ **and** $\mathcal{B}_2$ is the unique matrix $\mathbf{A} \in M_{m,n}$ such that

$$T(v_j) = \sum_{i=1}^{m} a_{ij} w_i$$

for each $j$. Equivalently, the $j$th column of $\mathbf{A}$ is the matrix of $T(v_j)$ with respect to $\mathcal{B}_2$.

If $V = W$ and we consider the same basis $\mathcal{B} = \mathcal{B}_1 = \mathcal{B}_2$ in both cases, we speak simply of the matrix representing $T$ with respect to $\mathcal{B}$.

There are no universally agreed-upon notations for the above notions, and we will not introduce any since they inevitably lead to a lot of notational clutter. Writing out "the matrix of $T$ with respect to $\mathcal{B}_1$ and $\mathcal{B}_2$" is pretty cumbersome, too, but we won't need to write it that often after this section.

**Proposition 1.1.** *Let $\mathcal{B}_1$ be a basis of $V$ and $\mathcal{B}_2$ be a basis of $W$. Let $T : V \to W$ be a linear map, $v \in V$, and write $w = T(v)$. Let $\mathbf{A}$ be the matrix of $T$ with respect to $\mathcal{B}_1$ and $\mathcal{B}_2$, $\mathbf{x}$ the matrix of $v$ with respect to $\mathcal{B}_1$, and $\mathbf{y}$ the matrix of $w$ with respect to $\mathcal{B}_2$. Then $\mathbf{y} = \mathbf{A}\mathbf{x}$.*

**Proposition 1.2.** *Let $\mathcal{B}_1$ be a basis of $V$, $\mathcal{B}_2$ be a basis of $W$, and $\mathcal{B}_3$ be a basis of $X$. Let $S : V \to W$ and $T : W \to X$ be linear maps, let $\mathbf{A}$ be the matrix of $S$ with respect to $\mathcal{B}_1$ and $\mathcal{B}_2$, and let $\mathbf{B}$ be the matrix of $T$ with respect to $\mathcal{B}_2$ and $\mathcal{B}_3$. Then the matrix of the linear map $TS : V \to X$ with respect to $\mathcal{B}_1$ and $\mathcal{B}_3$ is $\mathbf{BA}$.*

**Corollary 1.3** (Change of basis formula)**.** *Let $\mathcal{B}_1 = (v_1, \dots, v_n)$ and $\mathcal{B}_2 = (w_1, \dots, w_n)$ be bases of $V$, and let $\mathbf{S} \in M_n$ be the matrix representing the identity map with respect to $\mathcal{B}_1$ and $\mathcal{B}_2$. That is, $\mathbf{S}$ is the unique matrix such that*

$$v_j = \sum_{i=1}^{m} s_{ij} w_i$$

*for each $j$. Then $\mathbf{S}$ is invertible; it is called the **change of basis matrix**.*

*Let $T : V \to V$ be a linear map, and let $\mathbf{A}$ be the matrix representing $T$ with respect to $\mathcal{B}_1$ and let $\mathbf{B}$ be the matrix representing $T$ with respect to $\mathcal{B}_2$. Then $\mathbf{B} = \mathbf{S}^{-1}\mathbf{AS}$.*

**Definition 1.4.** Two matrices $\mathbf{A}, \mathbf{B} \in M_n$ are **similar** if there exists an invertible matrix $\mathbf{S} \in M_n$ such that

$$\mathbf{A} = \mathbf{SBS}^{-1}.$$

In that case $\mathbf{S}$ is called a **similarity transformation** between $\mathbf{A}$ and $\mathbf{B}$.

A minor linguistic clarification may be in order: we may more precisely say that $\mathbf{A}$ and $\mathbf{B}$ are similar *to each other*, or that $\mathbf{A}$ is *similar to* $\mathbf{B}$. It doesn't mean anything to say that a single matrix $\mathbf{A}$ is similar.

Matrices which are similar (in this technical sense) to each other share many properties. Two simple but important examples are contained in the following lemma.

**Lemma 1.5.** *If $\mathbf{A}, \mathbf{B} \in M_n$ are similar, then $\operatorname{tr} \mathbf{A} = \operatorname{tr} \mathbf{B}$ and $\det \mathbf{A} = \det \mathbf{B}$.*

A deeper explanation of the relationship between similar matrices is contained in the following result, which says, informally, that similar matrices are just different representations of the same linear map. The proof is a direct application of the change of basis formula.

**Theorem 1.6.** *Two matrices $\mathbf{A}, \mathbf{B} \in M_n$ are similar if and only if there exist a linear map $T \in \mathcal{L}(\mathbb{F}^n)$ and two bases $\mathcal{B}_1$ and $\mathcal{B}_2$ of $\mathbb{F}^n$ such that $\mathbf{A}$ is the matrix of $T$ with respect to $\mathcal{B}_1$ and $\mathbf{B}$ is the matrix of $T$ with respect to $\mathcal{B}_2$.*

Both of the implications in this "if and only if" theorem are important. On the one hand, it says that any property of matrices which is preserved by similarity is really a coordinate-independent property of the underlying linear maps. For example, it implies that the following definitions make sense.

**Definition 1.7.** Given a linear map $T \in \mathcal{L}(V)$, let $\mathbf{A}$ be the matrix of $T$ with respect to any basis of $V$. Then the **trace** and **determinant** of $T$ are defined by

$$\operatorname{tr} T = \operatorname{tr} \mathbf{A} \quad \text{and} \quad \det T = \det \mathbf{A}.$$

The point is that it doesn't matter which basis of $V$ is used here — even though different bases give different matrices $\mathbf{A}$, all the possible matrices are similar, and therefore have the same trace and determinant. It is possible to define the trace and determinant of an operator without making reference to bases or matrices (and there are advantages to doing so), but it is much more complicated.

**Proposition 1.8.** *Let $V$ be an inner product space, let $\mathcal{B} = (e_1, \ldots, e_n)$ be an orthonormal basis of $V$, and let $v \in V$. Then the matrix $\mathbf{x}$ of $v$ with respect to $\mathcal{B}$ is given by*

$$x_i = \langle v, e_i \rangle .$$

**Proposition 1.9.** *Let $V$ and $W$ be inner product spaces, $\mathcal{B}_1 = (e_1, \ldots, e_n)$ be an orthonormal basis of $V$, $\mathcal{B}_2 = (f_1, \ldots, f_m)$ be an orthonormal basis of $W$, and let $T : V \to W$ be a linear map. Then the matrix $\mathbf{A}$ of $T$ with respect to $\mathcal{B}_1$ and $\mathcal{B}_2$ is given by*

$$a_{ij} = \langle Te_j, f_i \rangle .$$

**Proposition 1.10.** *Let $V$ be an inner product space, $\mathcal{B}$ an orthonormal basis of $V$, and $T : V \to V$ a linear map. Then $T$ is a unitary map if and only if the matrix of $T$ with respect to $\mathcal{B}$ is a unitary matrix (in the real case, an orthogonal matrix).*

## 1.3 Rank and eigenvalues

There are several approaches to defining the rank of a linear map or matrix. We will say that the **rank** of a linear map is the dimension of its image.

**Proposition 1.11.** *Let $\mathbf{A}$ be a matrix. The number of linearly independent columns of $\mathbf{A}$ is equal to the number of linearly independent rows of $\mathbf{A}$.*

**Corollary 1.12.** *The **rank** of a matrix $\mathbf{A}$ may be equivalently defined as any of:*

- *the number of linearly independent columns of $\mathbf{A}$,*
- *the dimension of the span of the columns of $\mathbf{A}$,*
- *the number of linearly independent rows of $\mathbf{A}$.*

*Furthermore* $\operatorname{rank} \mathbf{A} = \operatorname{rank} \mathbf{A}^{\mathrm{T}}$, *and* $\operatorname{rank} \mathbf{A}$ *is equal to the rank of any linear map represented by $\mathbf{A}$.*

**Proposition 1.13** (Rank-Nullity Theorem)**.** *If $\mathbf{A} \in M_{m,n}$, then* $\operatorname{rank} \mathbf{A} + \dim \ker \mathbf{A} = n$. *If $T : V \to W$ is a linear map then* $\operatorname{rank} T + \dim \ker T = \dim V$.

**Corollary 1.14.** *The following are equivalent for a square matrix $\mathbf{A} \in Mn$.*

1. *$\mathbf{A}$ is invertible.*
2. $\operatorname{rank} \mathbf{A} = n$.
3. $\dim \ker \mathbf{A} = 0$.

*The following are equivalent for a linear map $T : V \to V$.*

1. *$T$ is invertible.*
2. *$T$ is surjective.*
3. *$T$ is injective.*

One important application of the last corollary is that it gives a way to talk about eigenvalues without dealing with the (in general harder) issue of identifying eigenvectors.

**Corollary 1.15.** *Let $\mathbf{A} \in M_n$ be a square matrix and $\lambda \in \mathbb{F}$. The following are equivalent.*

1. *$\lambda$ is an eigenvalue of $\mathbf{A}$.*
2. *$\dim \ker(\mathbf{A} - \lambda \mathbf{I}) > 0$.*
3. *$\operatorname{rank}(\mathbf{A} - \lambda \mathbf{I}) < n$.*
4. *$\mathbf{A} - \lambda \mathbf{I}$ is singular.*
5. *$p_{\mathbf{A}}(\lambda) = 0$, where $p_{\mathbf{A}}(z) = \det(\mathbf{A} - z\mathbf{I})$ is the characteristic polynomial of $\mathbf{A}$.*

# 2 Matrix factorizations

## 2.1 SVD: The fundamental theorem of matrix analysis

**Theorem 2.1** (Singular value decomposition)**.** *1. Let $A \in M_{m,n}$, and denote $p = \min\{m, n\}$. Then there exist $U \in \mathcal{U}_m$, $V \in \mathcal{U}_n$, and real numbers $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$ such that*

$$A = U \Sigma V^*, \tag{1}$$

*where $\Sigma \in M_{m,n}$ has entries $[\Sigma]_{jj} = \sigma_j$ for $1 \leq j \leq p$, and all other entries are $0$. If $A \in M_{m,n}(\mathbb{R})$, then $U$ and $V$ can be taken to be orthogonal.*

*We also have*

$$A = \sum_{j=1}^{p} \sigma_j u_j v_j^*, \tag{2}$$

*where $u_j$ and $v_j$ denote the columns of $U$ and $V$, respectively.*

*The numbers $\sigma_1, \ldots, \sigma_p$ are called the **singular values** of $A$, and are uniquely defined by $A$.*

*2. Suppose that $V$ and $W$ are finite dimensional inner product spaces and that $T : V \to W$ is a linear map. Write $n = \dim V$, $m = \dim W$, and $p = \min\{m, n\}$. Then there exist orthonormal bases $e_1, \ldots, e_n$ of $V$ and $f_1, \ldots, f_m$ of $W$, and real numbers $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$ such that*

$$Te_j = \begin{cases} \sigma_j f_j & \text{if } 1 \leq j \leq p, \\ 0 & \text{if } j > p. \end{cases} \tag{3}$$

*The numbers $\sigma_1, \ldots, \sigma_p$ are called the **singular values** of $T$, and are uniquely defined by $T$.*

*Proof.* The first step is to observe that the two parts of the theorem are exactly equivalent to each other.

We will prove the existence statement by induction on $p$, switching freely between the two viewpoints. The basis case $p = 1$ is trivial. Suppose now that the theorem is known for all smaller values of $p$.

The function $f : \mathbb{F}^n \to \mathbb{R}$ defined by $f(x) = \|Ax\|$ is continuous. By a theorem from multivariable calculus, it achieves its maximum value on the closed, bounded set $S = \{x \in \mathbb{F}^n \mid \|x\| = 1\}$, say at $x_0$. That is, $\|x_0\| = 1$, and $\|Ax\| \leq \|Ax_0\|$ whenever $\|x\| = 1$. Note that if $\|Ax_0\| = 0$, then $A = 0$ and the result is trivial, so we may assume that $\sigma = \|Ax_0\| > 0$.

I claim that if $\langle u, x_0 \rangle = 0$, then $\langle Au, Ax_0 \rangle = 0$ as well. For $t \in \mathbb{C}$, define

$$g(t) = \|A(x_0 + tu)\|^2 = \langle A(x_0 + tu), A(x_0 + tu) \rangle = \sigma^2 + 2 \operatorname{Re} t \langle Ax_0, Au \rangle + |t|^2 \|Au\|^2 .$$

On the one hand,

$$g(t) \geq \sigma^2 + 2 \operatorname{Re} t \langle Ax_0, Au \rangle .$$

On the other hand,

$$g(t) = \|x_0 + tu\|^2 \left\| A \left( \frac{x_0 + tu}{\|x_0 + tu\|} \right) \right\|^2 \leq \sigma^2 \|x_0 + tu\|^2 = \sigma^2 \langle x_0 + tu, x_0 + tu \rangle = \sigma^2 (1 + |t|^2 \|u\|^2).$$

Together, these inequalities imply that

$$2 \operatorname{Re} t \langle Ax_0, Au \rangle \leq |t|^2 \sigma^2 \|u\|^2$$

for every $t \in \mathbb{C}$. Setting $t = \langle Au, Ax_0 \rangle \varepsilon$ for $\varepsilon > 0$, we obtain that

$$2 |\langle Ax_0, Au \rangle|^2 \leq |\langle Ax_0, Au \rangle|^2 \sigma^2 \|u\|^2 \varepsilon$$

for every $\varepsilon > 0$, which is only possible if $\langle Ax_0, Au \rangle = 0$.

Switching now to the linear map perspective, it follows that $T(x_0^\perp) \subseteq (Tx_0)^\perp$. Note that $V_1 := x_0^\perp \subsetneq V$ and $W_1 := (Tx_0)^\perp \subsetneq W$. We can therefore apply the induction hypothesis to the linear map $T|_{V_1} : V_1 \to W_1$. The induction hypothesis states that there exist orthonormal bases $e_1, \ldots, e_{n-1}$ of $V_1$ and $f_1, \ldots, f_m$ of $W_1$ such that

$$Te_j = \begin{cases} \sigma_j f_j & \text{if } 1 \leq j \leq p-1, \\ 0 & \text{if } j > p = 1. \end{cases}$$

Now define $e_0 = x_0$, $\sigma_0 = \sigma = \|Ax_0\|$, and $f_0 = \frac{Ax_0}{\sigma_0}$. Then $e_0$ is orthonormal to $e_1, \ldots, e_{n-1}$ and the claim above implies that $f_0$ is also orthonormal to $f_1, \ldots, f_{m-1}$. After reindexing, the claim follows.

For uniqueness, note that if $A = U\Sigma V^*$, then $A^*A = V\Sigma^2 V^*$, and that

$$\Sigma^2 = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_p^2, 0, \ldots 0) \in M_n.$$

Therefore $\sigma_1, \ldots, \sigma_p$ are the square roots of the first $p$ eigenvalues of $A^*A$, and are therefore uniquely determined by $A$. □

**Corollary 2.2.** *If $A \in M_{m,n}$, then* $\operatorname{rank} A$ *is equal to the number of nonzero singular values of $A$.*

*Proof.* Again switching to the linear map perspective, the range of $T$ is spanned by those $f_j$ for which $\sigma_j > 0$. □

## 2.2 A first application: the Moore–Penrose inverse

Suppose that $m = n$, and that $A \in M_n$ has SVD

$$A = U\Sigma V^* = U \operatorname{diag}(\sigma_1, \ldots, \sigma_n) V^*.$$

By Corollary 2.2, $A$ is invertible iff $\sigma_n > 0$. In that case,

$$A^{-1} = V\Sigma^{-1}U^* = V \operatorname{diag}(\sigma_1^{-1}, \ldots, \sigma_n^{-1}) U^*.$$

More generally, for $A \in M_{m,n}$ we have an SVD

$$A = U \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} V^*, \tag{4}$$

where $D \in M_r$ has positive diagonal entries and $r = \operatorname{rank} A$. We define

$$A^\dagger = V \begin{bmatrix} D^{-1} & 0 \\ 0 & 0 \end{bmatrix} V^*. \tag{5}$$

We call $A^\dagger$ the **Moore–Penrose inverse** or **pseudoinverse** of $A$. (Note that if $m \neq n$ then the 0 blocks in (4) and in (5) have different sizes. The block matrix on the right side of (4) is $m \times n$, but the one in (5) is $n \times m$; this determines the sizes of all the 0 blocks.)

**Lemma 2.3.** *The Moore–Penrose inverse of $A \in M_{m,n}$ satisfies the following.*

1. *If $m = n$ and $A$ is invertible, then $A^\dagger = A^{-1}$.*
2. *$AA^\dagger A = A$.*
3. *$A^\dagger A A^\dagger = A^\dagger$.*
4. *$(AA^\dagger)^* = AA^\dagger$.*
5. *$(A^\dagger A)^* = A^\dagger A$.*

Consider an $m \times n$ linear system $Ax = b$. When $m = n = \operatorname{rank} A$, this has the unique solution $x = A^{-1}b$. Now consider the underdetermined but full-rank case, when $\operatorname{rank} A = m < n$. The system then always has a solution, but not a unique one.

**Proposition 2.4.** *If $A \in M_{m,n}$ and $\operatorname{rank} A = m < n$, then $A(A^\dagger b) = b$ for each $b \in \mathbb{F}^m$. Moreover, if $x \in \mathbb{F}^m$ and $Ax = b$, then $\left\|A^\dagger b\right\| \leq \|x\|$.*

We therefore say that $A^\dagger b$ is the **least-squares solution** of $Ax = b$.

*Proof.* Since $\operatorname{rank} A = m$, given $b \in \mathbb{F}^m$, there exists some $x \in \mathbb{F}^n$ such that $Ax = b$. Therefore

$$AA^\dagger b = AA^\dagger Ax = Ax = b.$$

Furthermore,

$$\left\|A^\dagger b\right\| = \left\|A^\dagger Ax\right\| = \left\|V \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix} V^* x\right\|.$$

Writing $y = V^*x$,

$$\left\|A^\dagger b\right\| = \left\|\begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix} y\right\| \leq \|y\| = \|x\|. \qquad \square$$

If the system is overdetermined, there will usually be no solution at all. In that case we may wish to find the closest possible thing to a solution: an $x$ that minimizes $\|Ax - b\|$.

**Proposition 2.5.** *If $A \in M_{m,n}$ and $\operatorname{rank} A = n < m$, then for any $x \in \mathbb{F}^n$ and $b \in \mathbb{F}^m$,*

$$\left\| A(A^\dagger b) - b \right\| \leq \|Ax - b\| .$$

We again call $A^\dagger b$ a **least-squares solution** of $Ax = b$, but note that in this case $A^\dagger b$ may not actually be a solution of $Ax = b$.

*Proof.* In this case we can write the SVD of $A$ as $A = U \begin{bmatrix} D \\ 0 \end{bmatrix} V^*$, where $D \in M_n$ is diagonal with positive diagonal entries $\sigma_1, \ldots, \sigma_n$. If we write $c = U^* b$ and $y = V^* x$, then

$$\|Ax - b\| = \left\| U \begin{bmatrix} D \\ 0 \end{bmatrix} V^* x - b \right\| = \left\| \begin{bmatrix} D \\ 0 \end{bmatrix} y - c \right\| = \left\| \begin{bmatrix} Dy \\ 0 \end{bmatrix} - c \right\| = \sqrt{\sum_{j=1}^{n} (\sigma_j y_j - c_j)^2 + \sum_{j=m+1}^{n} c_j^2}.$$

For a given $b$ (hence given $c$), this is clearly smallest when $y_j = \sigma_j^{-1} c_j$ for $j = 1, \ldots, n$. That is,

$$y = \begin{bmatrix} D^{-1} \\ 0 \end{bmatrix} c,$$

and so

$$x = Vy = V \begin{bmatrix} D^{-1} \\ 0 \end{bmatrix} U^* c = A^\dagger b. \qquad \square$$

Propositions 2.4 and 2.5 can be combined into a single result, which also covers the non-full rank case, as you will see in homework.

## 2.3 The spectral theorems and polar decomposition

Recall that a matrix $A \in M_n$ is called **Hermitian** if $A^* = A$. Note that a real matrix $A$ is Hermitian if and only if it is **symmetric**, that is, if $A^T = A$.

Recall also that the **adjoint** of a linear map $T : V \to W$ between inner product spaces is the unique linear map $T^* : W \to V$ such that

$$\langle Tv, w \rangle = \langle v, T^* w \rangle$$

for every $v \in V$ and $w \in W$. This corresponds to conjugate transpose of matrices: if $A \in M_{m,n}$, then for every $x \in \mathbb{C}^n$ and $y \in \mathbb{C}^m$, we have

$$\langle Av, w \rangle = w^*(Av) = w^*(A^*)^* v = (A^* w)^* v = \langle v, A^* w \rangle .$$

A linear map $T : V \to V$ is called **self-adjoint** or **Hermitian** if $T^* = T$.

The next two results have corresponding versions for self-adjoint linear maps (which we will use in the proof of Theorem 2.8 below), but for simplicity we will state and prove them only in the matrix versions.

**Lemma 2.6.** *If $A$ is Hermitian, then every eigenvalue of $A$ is real.*

*Proof.* Suppose that $Ax = \lambda x$. Then

$$\langle Ax, x\rangle = x^*Ax = x^*(\lambda x) = \lambda x^*x = \lambda \|x\|^2,$$

but also

$$\langle Ax, x\rangle = x^*Ax = x^*A^*x = (Ax)^*x = (\lambda x)^*x = \overline{\lambda}\|x\|^2.$$

Therefore if $x$ is an eigenvector of $A$ with eigenvalue $\lambda$, then $\lambda\|x\|^2 = \overline{\lambda}\|x\|^2$. Dividing by $\|x\|^2$, we see that $\lambda = \overline{\lambda}$, and so $\lambda \in \mathbb{R}$. □

Lemma 2.6 is a first hint of a broad general analogy in which certain types of matrices are like certain types of complex numbers. One aspect of this analogy is that the conjugate transpose operation on matrices corresponds to complex conjugation of numbers; another is that the eigenvalues of a class of matrices belong to the corresponding class of complex numbers. So we see in two different ways that Hermitian matrices are analogous to real numbers: both because the equation $A^* = A$ is analogous to $\overline{\lambda} = \lambda$, and because eigenvalues of Hermitian matrices are automatically real numbers. More manifestations of this analogy will come up below.

Recall that eigenvectors of a matrix corresponding to distinct eigenvalues are necessarily linearly independent. It turns out that more is true for Hermitian matrices.

**Lemma 2.7.** *If $A \in M_n(\mathbb{F})$ is Hermitian, then $A$ has an eigenvector in $\mathbb{F}^n$.*

*Proof.* Let $A = U\Sigma V^*$ be an SVD of $A$. Then

$$A^2 = A^*A = V\Sigma^2V^*.$$

Each column $v_j$ of $V$ is therefore an eigenvector of $A^2$ with corresponding eigenvalue $\sigma_j^2$. Therefore

$$0 = (A^2 - \sigma_j^2 I_n)v_j = (A + \sigma_j I_n)(A - \sigma_j I_n)v_j.$$

Now if $(A - \sigma_j I_n)v_j = 0$, then $v_j$ is an eigenvector of $A$ with eigenvalue $\sigma_j$. On the other hand, if $(A-\sigma_j I_n)v_j \neq 0$, then $(A-\sigma_j I_n)v_j$ is an eigenvector of $A$ with eigenvalue $-\sigma_j$. □

**Theorem 2.8** (The spectral theorem for Hermitian matrices)**.** *1. If $A \in M_n(\mathbb{F})$ is Hermitian, then there exists $U \in \mathcal{U}_n$ and a diagonal matrix $\Lambda \in M_n(\mathbb{R})$ such that $A = U\Lambda U^*$. If $\mathbb{F} = \mathbb{R}$ then $U$ can be taken to be real orthogonal.*

*2. If $T : V \to V$ is self-adjoint, then there exists an orthonormal basis of $V$ consisting of eigenvectors of $T$.*

*Proof.* As with Theorem 2.1, the first step is to observe that the two parts are equivalent to one another.

Again we will proceed by induction, this time on $n = \dim V$. The case $n = 1$ is trivial. Suppose now that the theorem is known for spaces of dimension $< n$.

By Lemma 2.7, $T$ has an eigenpair $v$ and $\lambda$. I claim that if $u \in V$ and $\langle u, v\rangle = 0$, then $\langle Tu, v\rangle = 0$ as well. Since $T$ is self-adjoint,

$$\langle Tu, v\rangle = \langle u, T^*v\rangle = \langle u, Tv\rangle = \langle u, \lambda v\rangle = \overline{\lambda}\langle u, v\rangle = 0.$$

So if we define $V_1 = v^\perp$, it follows that $T(V_1) = V_1$.

So the linear map $T|_{V_1}$ maps $V_1 \to V_1$, and is furthermore still self-adjoint. The induction hypothesis implies that there is an orthonormal basis $(e_1, \ldots, e_{n-1})$ of $V_1$ consisting of eigenvectors of $T$. If we define $e_n = \frac{v}{\|v\|}$, then $(e_1, \ldots, e_n)$ will now be an orthonormal basis of $V$ consisting of eigenvectors of $T$. □

Continuing the analogy between matrices and complex numbers described above, we might ask which matrices correspond to nonnegative real numbers. Three possibilities to consider are:

- matrices with nonnegative eigenvalues,
- matrices of the form $A = B^2$ for some Hermitian matrix $B$ (analogous to the fact that $\lambda \geq 0$ iff $\lambda = x^2$ for some $x \in \mathbb{R}$), or
- matrices of the form $A = B^*B$ for some matrix $B$ (analogous to the fact that $\lambda \geq 0$ iff $\lambda = |z|^2 = \overline{z}z$ for some $z \in \mathbb{C}$).

Using Theorem 2.8 we can show that these three possibilities are all equivalent, and are equivalent to a fourth positivity condition.

**Theorem 2.9.** *Suppose that $A \in M_n(\mathbb{F})$ is Hermitian. Then the following are equivalent:*

1. *Each eigenvalue of $A$ is nonnegative.*
2. *There exists a Hermitian matrix $B \in M_n(\mathbb{F})$ such that $A = B^2$.*
3. *There exists a matrix $B \in M_{m,n}(\mathbb{F})$ for some $m$ such that $A = B^*B$.*
4. *For every $x \in \mathbb{F}^n$, $\langle Ax, x\rangle \geq 0$.*

A Hermitian matrix satisfying the conditions of Theorem 2.9 is called a **positive semidefinite** matrix. (The closely related notion of a **positive definite** matrix will appear in homework.)

*Proof.* **(1) ⇒ (2):** By the spectral theorem (Theorem 2.8), we can write $A = U\Lambda U^*$ for a unitary matrix $U \in M_n(\mathbb{F})$ and a diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, where the $\lambda_j$ are all eigenvalues of $A$. By assumption, $\lambda_j \geq 0$, so it has a real square root $\sqrt{\lambda_j}$. Define

$$B = U \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})U^*.$$

Then $B^* = B$ and

$$B^2 = U \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})U^*U \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})U^* = A.$$

**(2) ⇒ (3):** This follows immediately.

**(3) ⇒ (4):** If $A = B^*B$, then for each $x \in \mathbb{F}^n$, $\langle Ax, x\rangle = \langle B^*Bx, Bx\rangle = \langle Bx, Bx\rangle = \|Bx\|^2 \geq 0$.

**(4) ⇒ (1):** Suppose that $\lambda$ is an eigenvalue of $A$ with corresponding eigenvector $v$. Then

$$0 \leq \langle Av, v\rangle = \langle \lambda v, v\rangle = \lambda \langle v, v\rangle = \lambda \|v\|^2 .$$

Since $\|v\| > 0$, this implies that $\lambda \geq 0$. □

The first part of the proof of Theorem 2.9 illustrates the basic philosophy of how the spectral theorem is usually applied: many things are simple to do for diagonal matrices, and the spectral theorem lets us pass from diagonal matrices to Hermitian matrices for many purposes.

Theorem 2.8 raises the question of which matrices can be factorized as $A = U\Lambda U^*$ for $U$ unitary and $\Lambda$ diagonal. If $\Lambda$ has real entries, then $A$ has to be Hermitian, since in that case

$$A^* = (U\Lambda U^*)^* = U\Lambda^* U^* = U\Lambda U^* = A.$$

If we don't insist that $\Lambda$ have real entries, this is no longer the case, since $\Lambda^* \neq \Lambda$ for a complex diagonal matrix. But it is still true that $\Lambda\Lambda^* = \operatorname{diag}(|\lambda_1|^2, \ldots, |\lambda_n|^2) = \Lambda^*\Lambda$. So if $A = U\Lambda U^*$, then we must have that

$$AA^* = U\Lambda U^* U\Lambda^* U^* = U\Lambda\Lambda^* U^* = U\Lambda^*\Lambda U^* = A^*A.$$

We call a matrix $A \in M_n(\mathbb{C})$ a **normal matrix** if $AA^* = A^*A$. It turns out that this is also a sufficient condition for a decomposition as in Theorem 2.8, as we will see in Theorem 2.11 below.

Recall next that if $z = x + iy$ is a complex number with $x, y \in \mathbb{R}$, then the real and imaginary parts of $z$ are

$$\operatorname{Re} z = x = \frac{z + \overline{z}}{2}, \qquad \operatorname{Im} z = y = \frac{z - \overline{z}}{2i}.$$

(Note that the imaginary part is actually a *real* number.) In analogy with this, we define for $A \in M_n(\mathbb{C})$,

$$\operatorname{Re} A = \frac{A + A^*}{2}, \operatorname{Im} A = \frac{A - A^*}{2i}.$$

We define $\operatorname{Re} T$ and $\operatorname{Im} T$ for a linear map $T : V \to V$ on an inner product space similarly.

**Lemma 2.10.** *Suppose that $A \in M_n(\mathbb{C})$. Then:*

1. $\operatorname{Re} A$ *and* $\operatorname{Im} A$ *are Hermitian.*
2. $A = (\operatorname{Re} A) + i(\operatorname{Im} A)$.
3. *$A$ is normal if and only if* $(\operatorname{Re} A)(\operatorname{Im} A) = (\operatorname{Im} A)(\operatorname{Re} A)$.

Note that even if $A \in M_n(\mathbb{R})$, $\operatorname{Im} A$ will typically have nonreal entries.

**Theorem 2.11** (The spectral theorem for normal matrices)**.** *1. If $A \in M_n(\mathbb{C})$ is normal, then there exists $U \in \mathcal{U}_n$ and a diagonal matrix $\Lambda \in M_n(\mathbb{C})$ such that $A = U\Lambda U^*$.*

*2. If $T : V \to V$ is normal and $V$ is a complex inner product space, then there exists an orthonormal basis of $V$ consisting of eigenvectors of $T$.*

*Proof.* We will prove the second part of the theorem, which as usual is equivalent to the first. For brevity we write $T_r = \operatorname{Re} T$ and $T_i = \operatorname{Im} T$.

First note that if $v$ is an eigenvector of $T_r$ with eigenvalue $\lambda$, then by Lemma 2.10,

$$T_r T_i v = T_i T_r v = \lambda T_i v.$$

It follows that $T_i v$ is also an eigenvector of $T_r$ with eigenvalue $\lambda$. Thus $T_i$ maps the eigenspace $\ker(T_r - \lambda I)$ to itself.

Now by the spectral theorem for Hermitian matrices (Theorem 2.8), for each eigenvalue $\lambda$ of $T_r$, there exists an orthonormal basis of $\ker(T_r - \lambda I)$ consisting of eigenvectors of $T_i$, which must also be eigenvectors of $T_r$. Moreover, Theorem 2.8 implies that the eigenvectors of $T_r$ span all of $V$, and that the distinct eigenspaces of $T_r$ are orthogonal to each other, so that combining these orthonormal bases of the eigenspaces yields an orthonormal basis of all of $V$. □

The last factorization in this section is the matrix analogue of the fact that a complex number can be written in the form $z = r\omega$ where $r = |z| \geq 0$ and $|\omega| = 1$. Since we can furthermore write $\omega = e^{i\theta} = \cos\theta + i \sin\theta$ this amounts to polar coordinates in the complex plane, which explains the name of the following result. (The fact that unitary matrices are analogous to complex numbers with absolute value 1 will be justified by various results we'll see, both in homework and in class.)

**Theorem 2.12** (The polar decomposition)**.** *Let $A \in M_n(\mathbb{F})$. Then there exists a $U \in \mathcal{U}_n$ (which can be taken to be orthogonal if $\mathbb{F} = \mathbb{R}$) and positive semidefinite matrices $P, Q \in M_n(\mathbb{F})$ such that $A = PU = UQ$.*

*Proof.* Let $A = W\Sigma V^*$ be an SVD for $A$. Then

$$A = W\Sigma W^* W V^* = W V^* V \Sigma V^*.$$

Define $U = WV^*$, $P = W\Sigma W^*$, and $Q = V\Sigma V^*$. □

Observe that in this case $P^2 = AA^*$ and $Q^2 = A^*A$. This means that both $P$ and $Q$ have a claim to be analogous to the "absolute value" of $A$.

## 2.4 Factorizations involving triangular matrices

Recall the following fundamental result about orthonormalization:

**Proposition 2.13** (Gram–Schmidt process)**.** *Let $(v_1, \ldots, v_n)$ be a linearly independent list of vectors in an inner product space $V$. Define $e_1 = v_1 / \|v_1\|$, and for each $j = 2, \ldots, n$, define recursively*

$$u_j = v_j - \sum_{k=1}^{j-1} \langle v_j, e_k \rangle, \qquad e_j = u_j / \|u_j\|.$$

*Then $(e_1, \ldots, e_n)$ is orthonormal, and for each $j$,*

$$\operatorname{span}(e_1, \ldots, e_j) = \operatorname{span}(v_1, \ldots, v_j).$$

The Gram–Schmidt process can be basically restated as a matrix factorization result:

**Theorem 2.14** (The QR decomposition)**.** *If $A \in M_n(\mathbb{F})$, there exist a unitary matrix $Q \in \mathcal{U}_n$ (orthogonal if $\mathbb{F} = \mathbb{R}$) and an upper triangular matrix $R \in M_n(\mathbb{F})$ such that $A = QR$.*

*Proof of Theorem 2.14 when $A$ is nonsingular.* Let $a_1, \ldots, a_n \in \mathbb{F}^n$ be the columns of $A$. If $A$ is nonsingular, then $(a_1, \ldots, a_n)$ forms a basis of $\mathbb{F}^n$. We apply the Gram–Schmidt process to obtain an orthonormal basis $(q_1, \ldots, q_n)$ of $\mathbb{F}^n$. The matrix $Q$ with these columns is unitary. If we define $R = Q^*A$, then $R$ has entries

$$r_{jk} = q_j^* a_k = \langle a_k, q_j \rangle .$$

Since $a_k$ is in the span of $(q_1, \ldots, q_k)$, and the $q_j$ are orthonormal, this implies that $r_{jk} = 0$ if $j > k$. Thus $R$ is upper triangular. □

We can extend Theorem 2.14 to singular matrices with a little more algebraic work, but because this is a class on matrix *analysis*, we will instead use an analytic approach, for which we need the following analytic fact:

**Proposition 2.15.** *If $\Omega$ is a closed, bounded subset of $\mathbb{R}^N$, and $\{\omega_k \mid k \in \mathbb{N}\}$ is a sequence of points in $\Omega$, then there is a subsequence of $\{\omega_k \mid k \in \mathbb{N}\}$ that converges to a point in $\Omega$.*

Observe also that $\mathcal{U}_n$ and $\mathcal{O}_n$ are closed and bounded sets: closed since if $\{U_k \mid k \in \mathbb{N}\}$ is a sequence in $\mathcal{U}_n$ or $\mathcal{O}_n$ with $U_k \to U$, then

$$U^*U = \lim_{k\to\infty} U_k^* U_k = \lim_{k\to\infty} I_n = I_n;$$

and bounded since each column of $U$ is a unit vector, and so $|u_{ij}| \leq 1$ for every $i, j$.

*Proof of Theorem 2.14 when $A$ is singular.* Given $A$, then matrix $A + \varepsilon I_n$ is nonsingular as long as $-\varepsilon$ is not an eigenvalue of $A$. We define $A_k = A + \frac{1}{k} I_n$. Then if $-\frac{1}{k}$ is greater than the largest negative eigenvalue of $A$ (if there are any at all), then $A_k$ is nonsingular.

For all such $k$, let $A_k = Q_k R_k$ be a QR decomposition, which we have already seen exists since $A_k$ is nonsingular. Then there is a subsequence of the sequence $\{Q_k \mid k \in \mathbb{N}\}$ which converges to a matrix $Q \in \mathcal{U}_n$ (in $\mathcal{O}_n$ if $\mathbb{F} = \mathbb{R}$). Writing this subsequence as $Q_{k_m}$ for $m \in \mathbb{N}$, we have that

$$R := \lim_{m\to\infty} R_{k_m} = \lim_{m\to\infty} Q_{k_m}^* A_{k_m} = Q^* A$$

exists. Furthermore, the limit of a sequence of upper triangular matrices is upper triangular. From this we have that $A = QR$. □

The next theorem is one of the most powerful tools for studying eigenvalues of arbitrary (especially non-normal) matrices. Be careful to notice that it requires working over $\mathbb{C}$, even if the original matrix $A$ has real entries.

**Theorem 2.16** (Schur factorization)**.** *If $A \in M_n(\mathbb{C})$, then there exist $U \in \mathcal{U}_n$ and an upper triangular $T \in M_n(\mathbb{C})$ such that $A = UTU^*$.*

*Proof.* We proceed by induction on $n$, the base case $n = 1$ being trivial.

By the Fundamental Theorem of Algebra, every polynomial with complex coefficients has a complex root. Applied to the characteristic polynomial of $A$, this implies that $A$ has an eigenvalue $\lambda_1$ with a corresponding eigenvector $v_1$. Extend $v_1$ to a basis $(v_1, \ldots, v_n)$ of $\mathbb{C}^n$, and let $S \in M_n$ be the invertible matrix with columns $v_1, \ldots, v_n$. Since $Av_1 = \lambda v_1$, it follows that

$$S^{-1}AS = \begin{bmatrix} \lambda_1 & w^T \\ 0 & B \end{bmatrix}$$

for some $w \in \mathbb{C}^{n-1}$ and $B \in M_{n-1}(\mathbb{C})$. By the induction hypothesis, $B = S_1 T_1 S_1^{-1}$ for some invertible $S_1 \in M_{n-1}(\mathbb{C})$ and upper triangular $T_1 \in M_{n-1}(\mathbb{C})$. This implies that

$$S^{-1}AS = \begin{bmatrix} \lambda_1 & w^T \\ 0 & S_1 T_1 S_1^{-1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & S_1 \end{bmatrix} \begin{bmatrix} \lambda_1 & x^T \\ 0 & T_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & S_1 \end{bmatrix}^{-1},$$

where $x^T = w^T S_1$. Thus

$$A = S \begin{bmatrix} 1 & 0 \\ 0 & S_1 \end{bmatrix} \begin{bmatrix} \lambda_1 & x^T \\ 0 & T_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & S_1 \end{bmatrix}^{-1} S^{-1} = S_2 T_2 S_2^{-1}$$

for an invertible $S_2 \in M_n(\mathbb{C})$ and upper triangular $T_2 \in M_n(\mathbb{C})$. Now let $S_2 = QR$ be a QR decomposition for $S_2$. Then $A = Q(RT_2R^{-1})Q^*$, where $Q$ is unitary and $RT_2R^{-1}$ is upper triangular, since the inverse of an upper triangular matrix is upper triangular. Letting $U = Q$ and $T = RT_2R^{-1}$ proves the theorem. □

Observe also that in the first step, we could pick any eigenvalue of $A$ to be $\lambda_1$. It follows that there exist Schur decompositions of $A$ in which the eigenvalues of $A$ appear on the diagonal of $T$ in any chosen order.

Recall that the **Frobenius norm** of a matrix $A \in M_{m,n}(\mathbb{C})$ is given by

$$\|A\|_F = \sqrt{\operatorname{tr} A^*A} = \sqrt{\sum_{j=1}^{m} \sum_{k=1}^{n} |a_{jk}|^2}.$$

That is, it is the norm of $A$ when we identify $A$ with a vector in $\mathbb{C}^{m \times n}$. The following basic property will be used many times in this course.

**Lemma 2.17.** *Suppose that $A \in M_{m,n}(\mathbb{C})$, $U \in \mathcal{U}_m$, and $V \in \mathcal{U}_n$. Then $\|UAV\|_F = \|A\|$.*

*Proof.* $\|UAV\|_F^2 = \operatorname{tr}(UAV)^*(UAV) = \operatorname{tr} V^*A^*U^*UAV = \operatorname{tr} V^*A^*AV = \operatorname{tr} VV^*A^*A = \operatorname{tr} A^*A = \|A\|_F^2$. □

**Corollary 2.18.** *For each $A \in M_n(\mathbb{C})$ and $\varepsilon > 0$, there exists a diagonalizable matrix $B \in M_n(\mathbb{C})$ such that $\|A - B\| < \varepsilon$.*

In analytic language, Corollary 2.18 says that the set of diagonalizable matrices is **dense** in $M_n(\mathbb{C})$.

*Proof.* Let $A = UTU^*$ be a Schur decomposition for $A$. Pick numbers $\lambda_1, \ldots, \lambda_n$ which are distinct from each other, such that $|\lambda_j - t_{jj}| < \frac{\varepsilon}{\sqrt{n}}$ for each $j$. Define $T_1$ to be the upper triangular matrix whose $jj$ entry is $\lambda_j$, and all other entries are the same as $T$, and then define $B = UT_1U^*$. Then $B$ is diagonalizable, because it has distinct eigenvalues $\lambda_1, \ldots, \lambda_n$, and

$$\|A - B\|_F = \|U(T - T_1)U^*\|_F = \|T - T_1\|_F = \sqrt{\sum_{j=1}^{n} |\lambda_j - t_{jj}|^2} < \varepsilon. \qquad \square$$

The following result can also be proved in a more algebraic way, but Corollary 2.18 allows for a simple analytic proof.

**Corollary 2.19** (Cayley–Hamilton theorem)**.** *Let $p$ be the characteristic polynomial of $A \in M_n(\mathbb{C})$. Then $p(A) = 0$.*

*Proof.* Suppose first that $A$ is diagonalizable, and that $A = SDS^{-1}$ for a diagonal matrix $D$. Then $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_j$ are the eigenvalues of $A$, and we can factor $p(x) = (x - \lambda_1) \cdots (x - \lambda_n)$. It follows that

$$p(D) = (D - \lambda_1 I_n) \cdots (D - \lambda_n I_n) = 0,$$

since the $j^{\text{th}}$ factor here is a diagonal matrix whose $j^{\text{th}}$ entry is 0, and then $p(A) = Sp(D)S^{-1} = 0$.

For the general case, by Corollary 2.18 we can find a sequence $\{A_k \mid k \in \mathbb{N}\}$ of diagonalizable matrices such that $A_k \to A$. If $p_k$ denotes the characteristic polynomial of $A_k$, then coefficient of $p_k$ converges to the corresponding coefficient of $p$. Therefore

$$p(A) = \lim_{k \to \infty} p_k(A_k) = 0$$

by the above argument. $\square$

**Corollary 2.20.** *The eigenvalues of a matrix $A \in M_n(\mathbb{C})$ depend continuously on $A$. That is, given $A \in M_n(\mathbb{C})$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ and $\varepsilon > 0$, there exists a $\delta > 0$ such that whenever $B \in M_n(\mathbb{C})$ and $\|A - B\|_F < \delta$, it follows that we can write the eigenvalues $\mu_1, \ldots, \mu_n$ of $B$ in some order so that $|\lambda_j - \mu_j| < \varepsilon$ for each $j$.*

*Proof.* Suppose the claim is not true. Then there exists an $\varepsilon > 0$ such that for every $k \in \mathbb{N}$ we can find a $B_k \in M_n(\mathbb{C})$ such that $\|A - B_k\|_F \leq \frac{1}{k}$ but for every ordering of the eigenvalues $\mu_j^{(m)}$ of $B_k$, we have $\max_j \left|\lambda_j - \mu_j^{(k)}\right| \geq \varepsilon$.

Let $B_k = U_k T_k U_k^*$ be Schur factorizations, in which the diagonal entries of $T_k$ are ordered in this way. There exists a subsequence of $\{U_k \mid k \in \mathbb{N}\}$ that converges to some $U \in \mathcal{U}_n$; say $U_{k_m} \to U$. Then

$$T_{k_m} = U_{k_m}^* B_{k_m} U_{k_m} \xrightarrow{m \to \infty} U^* A U,$$

which implies that $T := U^*AU$ is upper triangular, with diagonal entries equal to the eigenvalues of $A$ in some order. But now this implies that, perhaps after reordering the eigenvalues of $A$, $\max_j \left|\lambda_j - \mu_j^{(k_m)}\right| \xrightarrow{m \to \infty} 0$, which is a contradiction. $\square$

## 2.5 Simultaneous factorizations

**Theorem 2.21.** *Let $\mathcal{A} \subseteq M_n(\mathbb{F})$ be a family of diagonalizable matrices. The following are equivalent:*

1. *$\mathcal{A}$ is **simultaneously diagonalizable**. That is, there exists a single invertible matrix $S \in M_n(\mathbb{F})$ such that $S^{-1}AS$ is diagonal for every $A \in \mathcal{A}$.*
2. *$\mathcal{A}$ is **commuting**. That is, for every $A, B \in \mathcal{A}$, we have $AB = BA$.*

**Lemma 2.22.** *If $A = A_1 \oplus \cdots \oplus A_k$ is diagonalizable, then each block $A_j$ is diagonalizable.*

*Proof.* It suffices to assume that $k = 2$. Suppose that $C = A \oplus B$ for $A \in M_m$ and $B \in M_n$, and that $S^{-1}CS = \operatorname{diag}(\lambda_1, \ldots, \lambda_{m+n})$. We need to show that $A$ and $B$ are each diagonalizable.

Write

$$S = \begin{bmatrix} x_1 & \cdots & x_{m+n} \\ y_1 & \cdots & y_{m+n} \end{bmatrix}$$

for $x_j \in \mathbb{F}^m$ and $y_j \in \mathbb{F}^n$. Then $CS = S \operatorname{diag}(\lambda_1, \ldots, \lambda_{m+n})$ implies that $Ax_j = \lambda_j x_j$ and $By_j = \lambda_j y_j$ for $j = 1, \ldots, m+n$. Now since $S$ has rank $m+n$, it follows that

$$\operatorname{rank} \begin{bmatrix} x_1 & \cdots & x_{m+n} \end{bmatrix} = m \qquad \text{and} \qquad \operatorname{rank} \begin{bmatrix} y_1 & \cdots & y_{m+n} \end{bmatrix} = n$$

(since otherwise the number of linearly independent rows of $S$ would be smaller than $m+n$). Therefore there exist $m$ linearly independent $x_j$, which form a basis of $\mathbb{F}^m$ consisting of eigenvectors of $A$, so $A$ is diagonalizable, and similarly $B$ is diagonalizable. □

*Proof of Theorem 2.21.* $1 \Rightarrow 2$ is trivial. We prove $2 \Rightarrow 1$ by induction on $n$, the case $n = 1$ being trivial.

The result is moreover trivial if every matrix in $\mathcal{A}$ is a scalar matrix (that is, of the form $\lambda I_n$ for some $\lambda \in \mathbb{F}$). So suppose that $A \in \mathcal{A}$ is one fixed nonscalar matrix. Since $B$ is diagonalizable, there exists a nonsingular $S$ such that $SAS^{-1}$ is diagonal. By permuting the columns of $S$ if necessary, we may assume that

$$S^{-1}AS = (\lambda_1 I_{n_1}) \oplus \cdots \oplus (\lambda_k I_{n_k})$$

for distinct $\lambda_1, \ldots, \lambda_k$, with $k \geq 2$ and $n_1 + \cdots + n_k = n$.

Since $\mathcal{A}$ is commuting, for each $B, C \in \mathcal{A}$, $(S^{-1}BS)(S^{-1}CS) = (S^{-1}CS)(S^{-1}BS)$. If we write $SBS^{-1}$ in block form as

$$S^{-1}BS = \begin{bmatrix} B_{11} & \cdots & B_{1k} \\ \vdots & \ddots & \vdots \\ B_{k1} & \cdots & B_{kk} \end{bmatrix}$$

with $B_{ij} \in M_{n_i, n_j}(\mathbb{F})$, then $AB = BA$ implies that $\lambda_i B_{ij} = \lambda_j B_{ij}$ for each $i, j$. Since the $\lambda_i$ are distinct, this implies that $B_{ij} = 0$ for $i \neq j$, and so $SBS^{-1} = B_{11} \oplus \cdots \oplus B_{kk}$ is block-diagonal.

By Lemma 2.22, each $B_{jj}$ is diagonalizable for every $B \in \mathcal{A}$, and furthermore $\mathcal{A}_j = \{B_{jj} \mid B \in \mathcal{A}\}$ is a commuting family. By the induction hypothesis, there exist invertible matrices $S_j \in M_{n_j}(\mathbb{F})$ such that $S_j^{-1} B_{jj} S_j$ is diagonal for each $j$ and every $B \in \mathcal{A}$. Letting

$$\widetilde{S} = S(S_1 \oplus \cdots \oplus S_k),$$

it follows that $\widetilde{S}^{-1} B \widetilde{S}$ is diagonal for every $B \in \mathcal{A}$. □

**Theorem 2.23.** *If $\mathcal{A} \subseteq M_n(\mathbb{C})$ is commuting, then there exists a single matrix $U \in \mathcal{U}_n$ such that $U^*AU$ is upper triangular for every $A \in \mathcal{A}$.*

*Proof.* As in the previous proof, we proceed by induction on $n$, assuming without loss of generality that there exists a nonscalar $A \in \mathcal{A}$.

Let $\lambda$ be an eigenvalue of $A$, let $v_1, \ldots, v_m$ be an orthonormal basis of the range of $A - \lambda I_n$, extend it to an orthonormal basis $v_1, \ldots, v_n$ of all of $\mathbb{C}^n$, and let $V \in \mathcal{U}_n$ be the matrix with columns $v_j$. For each $B \in \mathcal{A}$ and $x \in \mathbb{C}^n$,

$$B(A - \lambda I_n)x = (A - \lambda I_n)Bx.$$

It follows that $V^*BV$ has the block triangular form

$$V^*BV = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}$$

with $B_{11} \in M_m$ and $B_{22} \in M_{n-m}$. Since $\mathcal{A}$ is commuting, if $B, C \in \mathcal{A}$, then

$$\begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} \\ 0 & C_{22} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ 0 & C_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix},$$

which implies that $B_{11}C_{11} = C_{11}B_{11}$ and $B_{22}C_{22} = C_{22}B_{22}$. Therefore, the families

$$\mathcal{A}_1 = \{B_{11} \mid B \in \mathcal{A}\} \qquad \text{and} \qquad \mathcal{A}_2 = \{B_{22} \mid B \in \mathcal{A}\}$$

are commuting.

By the induction hypothesis, there exist $V_1 \in \mathcal{U}_m$ and $V_2 \in \mathcal{U}_{n-m}$ such that $V_j^* B_{jj} V_j$ is upper triangular for every $B \in \mathcal{A}$ and $j = 1, 2$. If we now define $U = V(V_1 \oplus V_2)$, it follows that for each $B \in \mathcal{A}$,

$$U^*BU = \begin{bmatrix} V_1^* & 0 \\ 0 & V_2^* \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix} \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} = \begin{bmatrix} V_1^* B_{11} V_1 & V_1^* B_{12} V_2 \\ 0 & V_2^* B_{22} V_2 \end{bmatrix}$$

is upper triangular. □

**Corollary 2.24.** *Let $\mathcal{A} \subseteq M_n(\mathbb{C})$ be a family of normal matrices. The following are equivalent:*

1. *There exists a single matrix $U \in \mathcal{U}_n$ such that $U^*AU$ is diagonal for every $A \in \mathcal{A}$.*
2. *$\mathcal{A}$ is commuting.*

*Proof.* As you saw in homework, an upper triangular normal matrix is diagonal, so that this follows immediately from Theorem 2.23. □

This is the finite-dimensional case of the commutative Gelfand–Naimark theorem from the theory of operator algebras.

# 3 Eigenvalues of Hermitian matrices

If $A \in M_n(\mathbb{C})$ is Hermitian, then we know that the $n$ eigenvalues of $A$ (counted with multiplicity) are all real. By default we will put them in nondecreasing order and write them as $\lambda_1(A) \leq \lambda_2(A) \leq \cdots \leq \lambda_n(A)$ (omitting the $A$ when there is no ambiguity). We will also sometimes write $\lambda_{\min}$ and $\lambda_{\max}$ for the smallest and largest eigenvalues. Observe that, as with the Schur decomposition, we can always arrange to have the eigenvalues appear on the diagonal part of any spectral decomposition of $A$ in any chosen order.

## 3.1 Variational formulas

**Theorem 3.1** (Rayleigh–Ritz theorem)**.** *Suppose that $A \in M_n(\mathbb{F})$ is Hermitian. Then for every $x \in \mathbb{F}^n$,*

$$\lambda_{\min} \|x\|^2 \leq \langle Ax, x\rangle \leq \lambda_{\max} \|x\|^2 .$$

*Moreover,*

$$\lambda_{\min} = \min_{\substack{x\in\mathbb{F}^n \\ x\neq 0}} \frac{\langle Ax, x\rangle}{\|x\|^2} = \min_{\substack{x\in\mathbb{F}^n \\ \|x\|=1}} \langle Ax, x\rangle$$

*and*

$$\lambda_{\max} = \max_{\substack{x\in\mathbb{F}^n \\ x\neq 0}} \frac{\langle Ax, x\rangle}{\|x\|^2} = \max_{\substack{x\in\mathbb{F}^n \\ \|x\|=1}} \langle Ax, x\rangle .$$

*Proof.* Let $A = U\Lambda U^*$ be a spectral decomposition of $A$. Then for each $x \in \mathbb{F}^n$, writing $y = U^*x$ we have

$$\langle Ax, x\rangle = \langle U\Lambda U^*x, x\rangle = \langle \Lambda y, y\rangle = \sum_{j=1}^{n} \lambda_j |y_j|^2 \leq \lambda_{\max} \sum_{j=1}^{n} |y_j|^2 = \lambda_{\max} \|y\|^2 = \lambda_{\max} \|x\|^2 .$$

This immediately implies that $\lambda_{\max} \geq \frac{\langle Ax,x\rangle}{\|x\|^2}$ for every nonzero $x$, and in particular $\lambda_{\max} \geq \langle Ax, x\rangle$ whenever $\|x\| = 1$. Thus $\lambda_{\max}$ is greater than or equal to both of the max expressions. Moreover, if $x = Ue_n$ (so that $y = e_n$), then

$$\langle Ax, x\rangle = \langle \Lambda e_n, e_n\rangle = \lambda_n = \lambda_{\max}.$$

Thus $\lambda_{\max}$ is also less than or equal to both of the max expressions.

The proofs for $\lambda_{\min}$ are similar. □

The following corollary is immediate.

**Corollary 3.2.** *Suppose that $A \in M_n(\mathbb{F})$ is Hermitian, and that $\alpha = \langle Ax, x\rangle \|x\|^2$ for some $x \neq 0$. Then $A$ has at least one eigenvalue in $[\alpha, \infty)$ and at least one eigenvalue in $(-\infty, \alpha]$.*

**Theorem 3.3** (Courant–Fischer theorem)**.** *If $A \in M_n(\mathbb{F})$ is Hermitian, then*

$$\lambda_k(A) = \min_{\dim S = k} \max_{0\neq x\in S} \frac{\langle Ax, x\rangle}{\|x\|^2} = \max_{\dim S = n-k+1} \min_{0\neq x\in S} \frac{\langle Ax, x\rangle}{\|x\|^2},$$

*where in both cases $S$ varies over subspaces of $\mathbb{F}^n$ of the stated dimension.*

**Lemma 3.4.** *If $U$ and $V$ are subspaces of an $n$-dimensional vector space, then*

$$\dim(U \cap V) \geq \dim U + \dim V - n.$$

*In particular, if $\dim U + \dim V > n$, then $U \cap V$ contains nonzero vectors.*

*Proof.* Writing $k = \dim U \cap V$, $\ell = \dim U$, and $m = \dim V$, let $w_1, \ldots, w_k$ be a basis of $U \cap V$, and extend it to bases $w_1, \ldots, w_k, u_1, \ldots, u_{\ell-k}$ of $U$ and $w_1, \ldots, w_k, v_1, \ldots, v_{m-k}$ of $V$. I claim that the combined list

$$w_1, \ldots, w_k, u_1, \ldots, u_{\ell-k}, v_1, \ldots, v_{m-k}$$

is linearly independent. Indeed, if

$$a_1 w_1 + \cdots + a_k w_k + b_1 u_1 + \cdots + b_{\ell-k} u_{\ell-k} + c_1 v_1 + \cdots + c_{m-k} v_{m-k} = 0,$$

then the vector

$$c_1 v_1 + \cdots + c_{m-k} v_{m-k} = -(a_1 w_1 + \cdots + a_k w_k + b_1 u_1 + \cdots + b_{\ell-k} u_{\ell-k})$$

is in $U \cap V$ (considering the two sides of the equation separately). Therefore

$$c_1 v_1 + \cdots + c_{m-k} v_{m-k} = d_1 w_1 + \cdots d_k w_k$$

for some $d_1, \ldots, d_k \in \mathbb{F}$, and so

$$(a_1 + d_1) w_1 + \cdots + (a_k + d_k) w_k + b_1 u_1 + \cdots + b_{\ell-k} u_{\ell-k} = 0.$$

Since $w_1, \ldots, w_k, u_1, \ldots, u_{\ell-k}$ is linearly independent, it follows that $b_1 = \cdots = b_{\ell-k} = 0$. A similar argument shows that $c_1 = \cdots = c_{m-k} = 0$. It then follows that $a_1 w_1 + \cdots + a_k w_k = 0$, and therefore $a_1 = \cdots = a_k = 0$ as well. Thus the combined list of $k + (\ell - k) + (m - k) = \ell + m - k$ vectors is linearly independent.

It follows that $\ell + m - k \leq n$, which is equivalent to the claim. □

*Proof of Theorem 3.3.* Let $A = U \Lambda U^*$ be a spectral decomposition of $A$. Then $\lambda_k(A) = \lambda_k(\Lambda)$, and making the substitution $y = U^* x$,

$$\min_{\dim S = k} \max_{0 \neq x \in S} \frac{\langle Ax, x \rangle}{\|x\|^2} = \min_{\dim S = k} \max_{0 \neq y \in S} \frac{\langle \Lambda y, y \rangle}{\|y\|^2}$$

since the subspace $\{Uy \mid y \in S\}$ has the same dimension as $S$.

Given a $k$-dimensional subspace $S$, Lemma 3.4 implies there exists a nonzero $z \in S \cap \operatorname{span}(e_k, \ldots, e_n)$. It follows that

$$\max_{0 \neq y \in S} \frac{\langle \Lambda y, y \rangle}{\|y\|^2} \geq \frac{\langle \Lambda z, z \rangle}{\|z\|^2} = \frac{\sum_{j=k}^n \lambda_j |z_j|^2}{\sum_{j=k}^n |z_j|^2} \geq \lambda_k \frac{\sum_{j=k}^n |z_j|^2}{\sum_{j=k}^n |z_j|^2} \geq \lambda_k,$$

and therefore

$$\lambda_k \leq \min_{\dim S = k} \max_{0 \neq y \in S} \frac{\langle \Lambda y, y \rangle}{\|y\|^2}.$$

On the other hand, if $S = \operatorname{span}(e_1, \ldots, e_k)$, then for any $0 \neq y \in S$, we have

$$\frac{\langle \Lambda y, y \rangle}{\|y\|^2} = \frac{\sum_{j=1}^k \lambda_j |y_j|^2}{\sum_{j=1}^k |y_j|^2} \leq \frac{\lambda_k \sum_{j=1}^k |y_j|^2}{\sum_{j=1}^k |y_j|^2} = \lambda_k,$$

with equality if $y = e_k$. Thus

$$\lambda_k = \max_{0 \neq y \in \operatorname{span}(e_1, \ldots, e_k)} \frac{\langle \Lambda y, y \rangle}{\|y\|^2},$$

and so

$$\lambda_k \geq \min_{\dim S = k} \max_{0 \neq y \in S} \frac{\langle \Lambda y, y \rangle}{\|y\|^2}.$$

The max-min expression can be proved similarly, or can be deduced by replacing $A$ with $-A$. □

The following version of the Courant–Fischer theorem for singular values can be proved in a similar way, but it can also be deduced directly from Theorem 3.3.

**Corollary 3.5.** *If $A \in M_{m,n}(\mathbb{F})$, then*

$$\sigma_k(A) = \min_{\dim S = n-k+1} \max_{0 \neq x \in S} \frac{\|Ax\|}{\|x\|} = \max_{\dim S = k} \min_{0 \neq x \in S} \frac{\|Ax\|}{\|x\|},$$

*where in both cases $S$ varies over subspaces of $\mathbb{F}^n$ of the stated dimension.*

*Proof.* Recall that $\sigma_k^2 = \lambda_{n-k+1}(A^*A)$ (noting that we have chosen to list singular values in nonincreasing order, but eigenvalues in nondecreasing order). The result now follows immediately from the Courant–Fischer theorem applied to the Hermitian matrix $A^*A$. □

## 3.2 Inequalities for eigenvalues of two Hermitian matrices

**Theorem 3.6** (Weyl's inequalities)**.** *If $A, B \in M_n$ are Hermitian, then for each $1 \leq k \leq n$,*

$$\lambda_k(A + B) \leq \lambda_{k+j}(A) + \lambda_{n-j}(B)$$

*for $j = 0, \ldots, n-k$, and*

$$\lambda_{k-j+1}(A) + \lambda_j(B) \leq \lambda_k(A + B)$$

*for $j = 1, \ldots, k$.*

*Proof.* By the Courant–Fischer theorem (Theorem 3.3), there exist subspaces $S_A$ and $S_B$ of $\mathbb{F}^n$ such that $\dim S_A = k + j$, $\dim S_B = n - j$, and

$$\begin{aligned} \lambda_{k+j}(A) &= \max_{\substack{x \in S_A, \\ \|x\|=1}} \langle Ax, x \rangle, \\ \lambda_{n-j}(B) &= \max_{\substack{x \in S_B, \\ \|x\|=1}} \langle Bx, x \rangle. \end{aligned} \tag{6}$$

By Lemma 3.4, we have $\dim(S_A \cap S_B) \geq k$. Applying the Courant–Fischer theorem to $A + B$, we now have that

$$\begin{aligned}
\lambda_k(A+B) &= \min_{\dim S=k} \max_{\substack{x\in S\\ \|x\|=1}} \langle (A+B)x, x\rangle \\
&\leq \min_{\substack{\dim S=k,\\ S\subseteq S_A\cap S_B}} \max_{\substack{x\in S,\\ \|x\|=1}} \left(\langle Ax, x\rangle + \langle Bx, x\rangle\right) \\
&\leq \min_{\substack{\dim S=k,\\ S\subseteq S_A\cap S_B}} \left( \max_{\substack{x\in S,\\ \|x\|=1}} \langle Ax, x\rangle + \max_{\substack{x\in S,\\ \|x\|=1}} \langle Bx, x\rangle \right) \\
&= \lambda_{k+j}(A) + \lambda_{n-j}(B),
\end{aligned}$$

where the last equality follows from (6).

The second inequality in the theorem follows similarly from the max-min version of the Courant–Fischer theorem. □

Weyl's inequalities are useful for bounding the effect on eigenvalues of various types of perturbations of Hermitian matrices. An immediate consequence is the following.

**Corollary 3.7.** *If $A, B \in M_n$ are Hermitian, then*

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

Corollary 3.7 says that if $B$ is small in the sense that all its eigenvalues are small, then the eigenvalues of $A + B$ are close to those of $A$. A further consequence of Corollary 3.7 is the following.

**Corollary 3.8** (Weyl's monotonicity theorem)**.** *If $A \in M_n$ is Hermitian and $B \in M_n$ is positive semidefinite, then*

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

*for each $1 \leq k \leq n$.*

Another natural question is what the effect is on the eigenvalues of perturbing a fixed matrix $A$ by a matrix $B$ which is small in the sense of having small rank.

**Corollary 3.9.** *If $A, B \in M_n$ and $\operatorname{rank} B = r$, then*

$$\lambda_k(A+B) \leq \lambda_{k+r}(A)$$

*for $1 \leq k \leq n - r$ and*

$$\lambda_k(A+B) \geq \lambda_{k-r}(A)$$

*for $r + 1 \leq k \leq n$.*

*Proof.* Since $\operatorname{rank} B = r$, $B$ has exactly $r$ nonzero eigenvalues. Therefore $\lambda_{n-r}(B) \leq 0 \leq \lambda_{r+1}(B)$. Weyl's inequalities then imply that

$$\lambda_k(A+B) \leq \lambda_{k+r}(A) + \lambda_{n-r}(B) \leq \lambda_{k+r}(A)$$

for $1 \leq k \leq n - r$ and

$$\lambda_k(A+B) \geq \lambda_{k-r}(A) + \lambda_{r+1}(B) \geq \lambda_{k-r}(A)$$

for $r + 1 \leq k \leq n$. □

Our last consequence of Weyl's inequalities, which follows immediately from the last two results is an example of an *interlacing* theorem.

**Corollary 3.10.** *If $A \in M_n$ is Hermitian, and $B \in M_n$ is positive semidefinite with rank $r$, then*

$$\lambda_{k-r}(A+B) \leq \lambda_k(A) \leq \lambda_k(A+B),$$

*with the first inequality valid for $r+1 \leq k \leq n$, and the second inequality valid for $1 \leq k \leq n$.*

When $r = 1$, Corollary 3.10 says in particular that

$$\lambda_1(A) \leq \lambda_1(A+B) \leq \lambda_2(A) \leq \lambda_2(A+B) \leq \cdots \leq \lambda_{n-1}(A+B) \leq \lambda_n(A) \leq \lambda_n(A+B).$$

The next result is the most famous interlacing theorem for eigenvalues.

**Theorem 3.11** (Cauchy interlacing theorem)**.** *Suppose that $A \in M_n$ is Hermitian and has the block form*

$$A = \begin{bmatrix} B & C \\ C^* & D \end{bmatrix},$$

*where $B \in M_m$, $C \in M_{m,n-m}$, and $D \in M_{n-m}$. Then for each $1 \leq k \leq m$,*

$$\lambda_k(A) \leq \lambda_k(B) \leq \lambda_{k+n-m}(A).$$

*Proof.* Observe first that if $y \in \mathbb{F}^m$ and $x = (y_1, \ldots, y_m, 0, \ldots, 0) \in \mathbb{F}^n$, then $\langle Ax, x\rangle = \langle By, y\rangle$, and moreover $\|x\| = \|y\|$. By the Courant–Fischer theorem,

$$\begin{aligned} \lambda_k(A) &= \min_{\substack{S \subseteq \mathbb{F}^n \\ \dim S = k}} \max_{\substack{x \in S \\ \|x\|=1}} \langle Ax, x\rangle \leq \min_{\substack{S \subseteq \operatorname{span}(e_1,\ldots,e_m) \subseteq \mathbb{F}^n \\ \dim S = k}} \max_{\substack{x \in S \\ \|x\|=1}} \langle Ax, x\rangle \\ &= \min_{\substack{S \subseteq \mathbb{F}^m \\ \dim S = k}} \max_{\substack{y \in S \\ \|y\|=1}} \langle By, y\rangle = \lambda_k(B) \end{aligned}$$

and

$$\begin{aligned} \lambda_{k+n-m}(A) &= \max_{\substack{S \subseteq \mathbb{F}^n \\ \dim S = m-k+1}} \min_{\substack{x \in S \\ \|x\|=1}} \langle Ax, x\rangle \geq \max_{\substack{S \subseteq \operatorname{span}(e_1,\ldots,e_m) \subseteq \mathbb{F}^n \\ \dim S = m-k+1}} \min_{\substack{x \in S \\ \|x\|=1}} \langle Ax, x\rangle \\ &= \max_{\substack{S \subseteq \mathbb{F}^m \\ \dim S = n-k+1}} \min_{\substack{y \in S \\ \|y\|=1}} \langle By, y\rangle = \lambda_k(B). \end{aligned}$$

□

When $m = n - 1$, Theorem 3.11 says in particular that

$$\lambda_1(A) \leq \lambda_1(B) \leq \lambda_2(A) \leq \lambda_2(B) \leq \cdots \leq \lambda_{n-1}(A) \leq \lambda_{n-1}(B) \leq \lambda_n(A).$$

This is the best possible description of the relationship between the eigenvalues of a Hermitian matrix and those of a principal $(n-1) \times (n-1)$ submatrix, in a sense made precise by the following result (stated here without proof).

**Theorem 3.12.** *Suppose that $\lambda_1, \ldots, \lambda_n$ and $\mu_1, \ldots, \mu_{n-1}$ are real numbers such that*

$$\lambda_1 \leq \mu_1 \leq \lambda_2 \leq \mu_2 \leq \cdots \leq \mu_{n-1} \leq \lambda_n.$$

*Then there exists a Hermitian matrix $A$ such that $\lambda_k = \lambda_k(A)$ and $\mu_k = \lambda_k(B)$, where $B$ is the $(n-1) \times (n-1)$ upper-left principle submatrix of $A$.*

A similar converse holds for Corollary 3.10.

One important application of Cauchy's interlacing theorem is the following variational formula for a sum of eigenvalues.

**Theorem 3.13** (Fan's maximal principle)**.** *If $A \in M_n(\mathbb{F})$ is Hermitian, then*

$$\sum_{k=1}^{m} \lambda_k(A) = \min_{\substack{U \in M_{n,m}(\mathbb{F}) \\ U^*U = I_m}} \operatorname{tr}(U^*AU)$$

*and*

$$\sum_{k=n-m+1}^{n} \lambda_k(A) = \max_{\substack{U \in M_{n,m}(\mathbb{F}) \\ U^*U = I_m}} \operatorname{tr}(U^*AU).$$

*Proof.* If $U \in M_{n,m}$ and $U^*U = I_m$, then the columns of $U$ are orthonormal. We extend that list of columns to an orthonormal basis of $\mathbb{F}^n$, and form the unitary matrix $V \in M_n(\mathbb{F})$ with those basis vectors as its columns. Then $U^*AU \in M_m$ is the $m \times m$ upper-left principal submatrix of $V^*AV$. By Cauchy's interlacing theorem (Theorem 3.11), $\lambda_k(A) = \lambda_k(V^*AV) \leq \lambda_k(U^*AU)$, and so for each such $U$ we obtain

$$\sum_{k=1}^{m} \lambda_k(A) \leq \sum_{k=1}^{m} \lambda_k(U^*AU) = \operatorname{tr}(U^*AU).$$

On the other hand, suppose that $A = V \operatorname{diag}(\lambda_1, \ldots, \lambda_n) V^*$ is a spectral decomposition, and let $U \in M_{n,m}$ consist of the first $m$ columns of $V$. Then $U^*U = I_m$, and

$$\operatorname{tr}(U^*AU) = \operatorname{tr} \operatorname{diag}(\lambda_1, \ldots, \lambda_n) = \sum_{k=1}^{m} \lambda_k(A).$$

The proof of the other half is similar. □

**Corollary 3.14.** *If $A, B \in M_n$ are Hermitian, then*

$$\sum_{k=m}^{n} \lambda_k(A+B) \leq \sum_{k=m}^{n} \lambda_k(A) + \sum_{k=m}^{n} \lambda_k(B)$$

*for each $1 \leq m \leq n$, with equality for $m = 1$.*

*Proof.* By Theorem 3.13,

$$\begin{aligned}
\sum_{k=m}^{n} \lambda_k(A+B) &= \max_{\substack{U \in M_{n,n-m+1} \\ U^*U = I_{n-m+1}}} \operatorname{tr}(U^*(A+B)U) \\
&= \max_{\substack{U \in M_{n,n-m+1} \\ U^*U = I_{n-m+1}}} \big[\operatorname{tr}(U^*AU) + \operatorname{tr}(U^*BU)\big] \\
&\leq \max_{\substack{U \in M_{n,n-m+1} \\ U^*U = I_{n-m+1}}} \operatorname{tr}(U^*AU) + \max_{\substack{U \in M_{n,n-m+1} \\ U^*U = I_{n-m+1}}} \operatorname{tr}(U^*BU) \\
&= \sum_{k=m}^{n} \lambda_k(A) + \sum_{k=m}^{n} \lambda_k(B).
\end{aligned}$$

If $m = 1$, the claim reduces to $\operatorname{tr}(A+B) = \operatorname{tr} A + \operatorname{tr} B$. □

### 3.3 Majorization

Given $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we write $x^\downarrow$ and $x^\uparrow$ for the nonincreasing and nondecreasing rearrangements of $x$. That is, $x_j^\downarrow$ is the $j^{\text{th}}$ largest entry of $x$, and $x_j^\uparrow$ is the $j^{\text{th}}$ smallest entry of $x$.

Let $x, y \in \mathbb{R}^n$. If

$$\sum_{i=1}^k x_i \le \sum_{i=1}^k y_i \qquad \text{for each } 1 \le k \le n,$$

then we say that $x$ is **weakly majorized** by $y$, written $x \prec_w y$. If moreover $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$, then we say that $x$ is **majorized** by $y$, written $x \prec y$.

For example, if $x_i \ge 0$ for each $i$ and $\sum_{i=1}^n x_i = 1$, then

$$\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) \prec x \prec (1, 0, \ldots, 0).$$

A more significant example follows from Corollary 3.14 above. We write $\lambda(A) \in \mathbb{R}^n$ for the vector of eigenvalues of a Hermitian matrix $A \in M_n$.

**Corollary 3.15** (Fan's majorization theorem)**.** *Let $A, B \in M_n$ be Hermitian. Then*

$$\lambda^\downarrow(A + B) \prec \lambda^\downarrow(A) + \lambda^\downarrow(B).$$

*Proof.* By Corollary 3.14, for each $1 \le k \le n$,

$$\sum_{i=1}^k \lambda_i^\downarrow(A + B) \le \sum_{i=1}^k \big[\lambda_i^\downarrow(A) + \lambda_i^\downarrow(B)\big] = \sum_{i=1}^n \big[\lambda^\downarrow(A) + \lambda^\downarrow(B)\big]_i^\downarrow,$$

with equality throughout if $k = n$. □

**Theorem 3.16** (Lidskii's majorization theorem)**.** *Let $A, B \in M_n$ be Hermitian. Then*

$$\lambda^\downarrow(A + B) - \lambda^\downarrow(A) \prec \lambda^\downarrow(B).$$

It is important to note that this does not follow from Corollary 3.15 simply by subtracting $\lambda^\downarrow(A)$ from both sides: majorization is not necessarily preserved by addition and subtraction of vectors.

*Proof.* Majorization *is* preserved by addition of a constant vector $(c, \ldots, c)$. We can therefore replace $B$ with $B - \lambda_k^\downarrow(B) I_n$, and thereby assume without loss of generality that $\lambda_k^\downarrow(B) = 0$.

Under this assumption (as you will show in homework), we can write $B = B_+ - B_-$, where $B_+$ and $B_-$ are both positive semidefinite, $\lambda_j^\downarrow(B_+) = \lambda_j^\downarrow(B)$ for $j \le k$, and $\lambda_j^\downarrow(B_+) = 0$ for $j \ge k$.

Then

$$\sum_{j=1}^k \lambda_j^\downarrow(B) = \sum_{j=1}^k \lambda_j^\downarrow(B_+) = \sum_{j=1}^n \lambda_j^\downarrow(B_+) = \operatorname{tr} B_+.$$

Therefore we need to show that

$$\sum_{j=1}^{k} \big[\lambda^{\downarrow}(A+B-B_-) - \lambda^{\downarrow}(A)\big]_j^{\downarrow} \le \operatorname{tr} B_+.$$

Since $B_+$ and $B_-$ are both positive semidefinite, Weyl's monotonicity theorem (Corollary 3.8) implies that $\lambda_j^{\downarrow}(A+B_+-B_-) \le \lambda_j^{\downarrow}(A+B_+)$ and $\lambda_j^{\downarrow}(A) \le \lambda^{\downarrow}(A+B_+)$. Therefore

$$\begin{aligned}\sum_{j=1}^{k} \big[\lambda^{\downarrow}(A+B-B_-) - \lambda^{\downarrow}(A)\big]_j^{\downarrow} &\le \sum_{j=1}^{k} \big[\lambda^{\downarrow}(A+B) - \lambda^{\downarrow}(A)\big]_j^{\downarrow} \\ &\le \sum_{j=1}^{n} \big[\lambda^{\downarrow}(A+B) - \lambda^{\downarrow}(A)\big]_j^{\downarrow} \\ &= \operatorname{tr}(A+B_+) - \operatorname{tr} A = \operatorname{tr} B_+.\end{aligned}$$

When $k = n$, we get equality since both sides are equal to $\operatorname{tr} B$. □

A different connection between majorization and matrix theory is provided by Proposition 3.17 below, for which we need another definition.

A matrix $A \in M_n(\mathbb{R})$ is called **doubly stochastic** if $a_{ij} \ge 0$ for each $i$ and $j$,

$$\sum_{j=1}^{n} a_{ij} = 1$$

for each $i$, and

$$\sum_{i=1}^{n} a_{ij} = 1$$

for each $j$.

The following examples will all be important:

- Every permutation matrix is doubly stochastic.
- More generally, if $P_1, \dots, P_N \in M_n$ are permutation matrices, $t_1, \dots, t_N \ge 0$, and $\sum_{k=1}^{N} t_k = 1$, then $\sum_{k=1}^{N} t_k P_k$ is doubly stochastic.
- Still more generally, if $A_1, \dots, A_N \in M_n$ are doubly stochastic, $t_1, \dots, t_N \ge 0$, and $\sum_{k=1}^{N} t_k = 1$, then $\sum_{k=1}^{N} t_k A_k$ is doubly stochastic.
- If $P$ and $Q$ are permutation matrices and $A$ is doubly stochastic, then $PAQ$ is also doubly stochastic.
- More generally, if $A$ and $B$ are both doubly stochastic, then $AB$ is also doubly stochastic. This can be proved by direct computation, but also follows easily from Birkhoff's theorem below.
- If $U \in M_n(\mathbb{R})$ is orthogonal and $A \in M_n(\mathbb{R})$ is defined by $a_{ij} = u_{ij}^2$, then $A$ is doubly stochastic. Such a doubly stochastic matrix is called **orthostochastic**.

- More generally, if $U \in M_n(\mathbb{C})$ is unitary and $A \in M_n(\mathbb{C})$ is defined by $a_{ij} = |u_{ij}|^2$, then $A$ is doubly stochastic. Such a doubly stochastic matrix is called **unitary-stochastic**.

**Proposition 3.17.** *Let $x, y \in \mathbb{R}^n$. Then $x \prec y$ if and only if there exists a doubly stochastic matrix $A$ such that $x = Ay$.*

*Proof.* Suppose first that $x = Ay$ for a doubly stochastic matrix $A$. There exist permutation matrices $P$ and $Q$ such that $x^\downarrow = Px$ and $y^\downarrow = Qy$. Then $x^\downarrow = (PAQ^T)y^\downarrow$, and $PAQ^T$ is also doubly stochastic. It therefore suffices to assume that $x = x^\downarrow$ and $y = y^\downarrow$.

Fix $k$. We have

$$\sum_{i=1}^{k} x_i = \sum_{i=1}^{k}\sum_{j=1}^{n} a_{ij}y_j - \sum_{j=1}^{n}\left(\sum_{i=1}^{k} a_{ij}\right) y_j.$$

Define $t_j = \sum_{i=1}^{k} a_{ij}$. Then $0 \le t_j \le 1$ and

$$\sum_{j=1}^{n} t_j = \sum_{i=1}^{k}\sum_{j=1}^{n} a_{ij} = k,$$

which implies that

$$\sum_{j=1}^{k}(t_j - 1) + \sum_{j=k+1}^{n} t_j = 0.$$

Therefore

$$\begin{aligned}\sum_{i=1}^{k} x_i - \sum_{i=1}^{k} y_i = \sum_{j=1}^{n} t_j y_j - \sum_{j=1}^{k} y_j &= \sum_{j=1}^{k}(t_j - 1)y_j + \sum_{j=k+1}^{n} t_j y_j \\ &= \sum_{j=1}^{k}(t_j - 1)(y_j - y_k) + \sum_{j=k+1}^{n} t_j(y_j - y_k) \le 0.\end{aligned}$$

If $k = n$, then we would have $t_j = 1$ for each $j$, and obtain equality above.

Now suppose that $x \prec y$. By a similar argument to above, we may assume without loss of generality that $x = x^\downarrow$ and $y = y^\downarrow$. We will proceed by induction on $n$, the case $n = 1$ being trivial.

The majorization $x \prec y$ implies that $y_n \le x_1 \le y_1$. Therefore there exists a $k \ge 1$ such that $y_k \le x_1 \le y_{k-1}$ (or $x_1 = y_1$ if $k = 1$). Let $t \in [0, 1]$ be such that

$$x_1 = ty_1 + (1 - t)y_k,$$

and define

$$x' = (x_2, \ldots, x_n) \in \mathbb{R}^{n-1}$$

and

$$y' = (y_2, \ldots, y_{k-1}, (1 - t)y_1 + ty_k, y_{k+1}, \ldots, y_n) =: (y'_2, \ldots, y'_n) \in \mathbb{R}^{n-1}.$$

Then $x'$ is in nonincreasing order,

$$\sum_{j=2}^{m} x_j \le \sum_{j=2}^{m} y_j$$

for $2 \leq m \leq k-1$ simply because each $x_j \leq y_j$ in this range, and for $k \leq m \leq n$ we have

$$\begin{aligned}\sum_{j=2}^{m} x_j &= \sum_{j=1}^{m} x_j - x_1 \leq \sum_{j=1}^{m} y_j - x_1 = \sum_{j=1}^{m} y_j - ty_1 - (1-t)y_k \\ &= \sum_{j=1}^{k-1} y_j + \big[(1-t)y_1 + ty_k\big] + \sum_{j=k+1}^{m} y_j = \sum_{j=2}^{m} y'_j,\end{aligned}$$

with equality if $m = n$. Now $y'$ need not be in nonincreasing order, but in any case $\sum_{j=2}^{m} y'_j$ is at most the sum of the $m-1$ largest entries of $y'$, and so it follows that $x' \prec y'$.

Now by the induction hypothesis there is a doubly stochastic $B \in M_{n-1}(\mathbb{R})$ such that $x' = By'$. Define

$$C = \begin{bmatrix} 1 & 0 \\ 0 & B \end{bmatrix} \in M_n(\mathbb{R})$$

and observe that $x = C(x_1, y')$. Now

$$(x_1, y') = (ty_1 + (1-t)y_k, y_2, \ldots, y_{k-1}, (1-t)y_1 + ty_k, y_{k+1}, \ldots, y_n) = \big[tI_n + (1-t)P\big]y,$$

where $P$ is the permutation matrix that transposes the first and $k^{\text{th}}$ entries. It follows that $x = Ay$, where $A = C\big[tI_n + (1-t)P\big]$ is doubly stochastic. □

**Theorem 3.18** (Schur's majorization theorem)**.** *If $A \in M_n$ is Hermitian, then $(a_{11}, \ldots, a_{nn}) \prec \lambda(A)$.*

*Proof.* Let $A = U\Lambda U^*$ be a spectral decomposition. Then

$$a_{jj} = \sum_{k=1}^{n} u_{jk}\lambda_k(A)\overline{u_{jk}} = \sum_{k=1}^{n} |u_{jk}|^2 \lambda_k(A).$$

That is, $(a_{11}, \ldots, a_{nn}) = B\lambda(A)$, where $B$ is the unitary-stochastic matrix $b_{jk} = |u_{jk}|^2$. The result now follows from Proposition 3.17. □

Similarly to the situation with Cauchy's interlacing theorem, Theorem 3.18 is an optimal result, as shown by the following converse (stated here without proof).

**Theorem 3.19** (Horn's theorem)**.** *Suppose that $(d_1, \ldots, d_n) \prec (\lambda_1, \ldots, \lambda_n)$. Then there exists a symmetric matrix $A \in M_n(\mathbb{R})$ with diagonal entries $d_1, \ldots, d_n$ and eigenvalues $\lambda_1, \ldots, \lambda_n$.*

For further applications of majorization theorems to matrix analysis, we will need the following additional characterization of majorization.

**Proposition 3.20.** *For $x, y \in \mathbb{R}^n$, $x \prec y$ if and only if there exist permutations $\sigma_1, \ldots, \sigma_N$ and numbers $t_1, \ldots, t_N \geq 0$ such that $\sum_{i=1}^{N} t_i = 1$ and $x = \sum_{i=1}^{N} t_i(y_{\sigma_i(1)}, \ldots, y_{\sigma_i(n)})$.*

*Proof.* We first note that the proposition can be equivalently stated as saying that $x \prec y$ iff there exist permutation matrices $P_1, \dots, P_N$ and $t_1, \dots, t_N \geq 0$ such that $\sum_{i=1}^N t_i = 1$ and $x = \sum_{i=1}^N t_i P_i y$. Since $\sum_{i=1}^N t_i P_i$ is doubly stochastic, the "if" direction follows from Proposition 3.17.

We can deduce the "only if" from a formally stronger statement. We will call a matrix of the form $T = tI_n + (1-t)P$ a $T$-matrix if $t \in [0,1]$ and $P$ is a permutation matrix. A close inspection of our proof of Proposition 3.17 shows that it actually proves the following statement: if $x \prec y$, then there exist some $T$-matrices $T_1, \dots, T_m$ such that $x = T_1 \cdots T_m y$. It is easy to check that a product of $T$-matrices is a permutation matrix, so the proposition then follows. □

Propsitions 3.17 and 3.20 immediately suggest the following result, which is indeed true:

**Theorem 3.21** (Birkhoff's theorem)**.** *A matrix $A \in M_n(\mathbb{R})$ is doubly stochastic if and only if there exist permutation matrices $P_1, \dots, P_N$ and numbers $t_1, \dots, t_N \geq 0$ such that $\sum_{i=1}^N t_i = 1$ and $A = \sum_{i=1}^N t_i P_i$.*

Indeed, Birkhoff's theorem implies that Propositions 3.17 and 3.20 are equivalent to each other; however, the proof of Birkhoff's theorem requires more overhead than the direct proof of Proposition 3.20 given above. We will leave Birkhoff's theorem unproved for the time being.

The next result is a substantial strengthening, for Hermitian matrices, of the fact that the eigenvalues of a matrix depend continuously on the matrix itself.

**Corollary 3.22** (Hoffman–Wielandt inequality for Hermitian matrices)**.** *If $A, B \in M_n$ are Hermitian, then*

$$\left\| \lambda^{\downarrow}(A) - \lambda^{\downarrow}(B) \right\| \leq \|A - B\|_F .$$

*Proof.* By Lidskii's majorization theorem (Theorem 3.16),

$$\lambda^{\downarrow}(A) - \lambda^{\downarrow}(B) \prec \lambda^{\downarrow}(A - B).$$

By Proposition 3.20, this implies that there exist permutation matrices $P_1, \dots, P_N$ and numbers $t_1, \dots, t_N \geq 0$ such that $\sum_{i=1}^N t_i = 1$ and

$$\lambda^{\downarrow}(A) - \lambda^{\downarrow}(B) = \sum_{i=1}^N t_i P_i (\lambda^{\downarrow}(A - B)).$$

It follows from the triangle inequality that

$$\begin{aligned} \left\| \lambda^{\downarrow}(A) - \lambda^{\downarrow}(B) \right\| &= \left\| \sum_{i=1}^N t_i P_i (\lambda^{\downarrow}(A - B)) \right\| \leq \sum_{i=1}^N t_i \left\| P_i (\lambda^{\downarrow}(A - B)) \right\| \\ &= \sum_{i=1}^N t_i \left\| \lambda^{\downarrow}(A - B) \right\| = \left\| \lambda^{\downarrow}(A - B) \right\| . \end{aligned}$$

□

# 4 Norms

## 4.1 Vector norms

A **norm** on a real or complex vector space $V$ is a function $\|\cdot\| : V \to \mathbb{R}$ such that:

- For each $v \in V$, $\|v\| \geq 0$.
- If $\|v\| = 0$, then $v = 0$.
- For each $v \in V$ and $c \in \mathbb{F}$, $\|cv\| = |c| \, \|v\|$.
- For each $v, w \in V$, $\|v + w\| \leq \|v\| + \|w\|$ (the triangle inequality).

If all but the second condition are satisfied, then we call $\|\cdot\|$ a **seminorm**.

The quantity which we have been accustomed to calling *the* norm on $\mathbb{F}^n$ is just one example of a norm. To avoid ambiguity, from now on we will denote it $\|\cdot\|_2$ and call it the $\ell^2$ norm.

The Frobenius norm $\|\cdot\|_F$ is of course also a norm on the vector space $M_{m,n}(\mathbb{F})$, since it amounts to the $\ell^2$ norm when we identify $M_{m,n}(\mathbb{F})$ with $\mathbb{F}^{mn}$.

The $\ell^2$ norm is part of a larger family: for $1 \leq p < \infty$, we define the $\ell^p$ norm on $\mathbb{F}^n$ by

$$\|x\|_p = \left( \sum_{j=1}^{n} |x_j|^p \right)^{1/p}.$$

For $p = \infty$, we have the limiting case

$$\|x\|_\infty = \max_{1 \leq j \leq n} |x_j| \, .$$

Except for the extreme cases $p = 1, \infty$, it is not obvious that these "norms" satisfy the triangle inequality. (Recall that it's not obvious for $p = 2$, either — the proof uses the Cauchy–Schwarz inequality.) To prove that we will need some preliminaries.

**Lemma 4.1** (The arithmetic–geometric mean inequality)**.** *If $a, b \geq 0$ and $0 < t < 1$, then $a^t b^{1-t} \leq ta + (1-t)b$.*

*Proof.* First observe that the function $f(t) = \log t$ satisfies $f''(t) < 0$. This implies that $f$ is **concave**, i.e., that

$$f(tx + (1-t)y) \geq tf(x) + (1-t)f(y)$$

for all $x, y > 0$ and $0 < t < 1$. The claim is trivial if either $a = 0$ or $b = 0$; if both are positive then we have

$$\log(ta + (1-t)b) \geq t \log a + (1-t) \log b,$$

and so

$$ta + (1-t)b = e^{\log(ta + (1-t)b)} \geq e^{t \log a + (1-t) \log b} = a^t b^{1-t}. \qquad \square$$

**Proposition 4.2** (Hölder's inequality)**.** *Suppose that* $1 \le p, q \le \infty$ *and* $\frac{1}{p} + \frac{1}{q} = 1$ *(where we interpret* $\frac{1}{\infty}$ *as* $0$*). Then for any* $x, y \in \mathbb{C}^n$*,*

$$|\langle x, y\rangle| \le \|x\|_p \|y\|_q .$$

*Proof.* The proof is easy in the case that $p = 1$ and $q = \infty$ (or vice-versa), so we assume that $1 < p, q < \infty$. We may also assume that $x, y \neq 0$, so that $\|x\|_p, \|y\|_q > 0$. Define

$$a_j = \frac{|x_j|^p}{\|x\|_p^p} \qquad \text{and} \qquad b_j = \frac{|y_j|^p}{\|y\|_p^p},$$

and write $t = \frac{1}{p}$, so $1 - t = \frac{1}{q}$. By the arithmetic geometric mean inequality,

$$|x_j \overline{y_j}| = \|x\|_p \|y\|_q a_j^t b_j^{1-t} \le \|x\|_p \|y\|_q \left(t a_j + (1-t) b_j\right).$$

Therefore

$$|\langle x, y\rangle| = \left|\sum_{j=1}^n x_j \overline{y_j}\right| \le \sum_{j=1}^n |x_j \overline{y_j}| \le \|x\|_p \|y\|_q \left(t \sum_{j=1}^n a_j + (1-t) \sum_{j=1}^n b_j\right) = \|x\|_p \|y\|_q . \quad \square$$

**Corollary 4.3** (Minkowski's inequality)**.** *Let* $1 \le p \le \infty$*. For* $x, y \in \mathbb{C}^n$*,*

$$\|x + y\|_p \le \|x\|_p + \|y\|_q .$$

*Proof.* As noted above, this result is easy if $p = 1, \infty$, so we will assume that $1 < p < \infty$ and let $q = \frac{p}{p-1}$ (so $\frac{1}{p} + \frac{1}{q} = 1$. For each $j$, by the triangle inequality for absolute values,

$$|x_j + y_j|^p = |x_j + y_j| \, |x_j + y_j|^{p-1} \le \left(|x_j| + |y_j|\right) |x_j + y_j|^{p-1} .$$

Now by Hölder's inequality,

$$\begin{aligned} \|x + y\|^p = \sum_{j=1}^n |x_j + y_j|^p &\le \sum_{j=1}^n |x_j| \, |x_j + y_j|^{p-1} + \sum_{j=1}^n |y_j| \, |x_j + y_j|^{p-1} \\ &\le \left(\|x\|_p + \|y\|_p\right) \left(\sum_{j=1}^n |x_j + y_j|^{(p-1)q}\right)^{1/q} = \left(\|x\|_p + \|y\|_p\right) \|x + y\|_p^{p/q} , \end{aligned}$$

where the last equality follows from the fact that $(p-1)q = p$. It follows that

$$\|x + y\|_p = \|x + y\|^{p - \frac{p}{q}} \le \|x\|_p + \|y\|_p . \qquad \square$$

Minkowski's inequality shows that the $\ell^p$ norms do satisfy the triangle inequality, and therefore really are norms.

We will not prove the following theorem, or use it below, but it provides an important piece of perspective.

**Theorem 4.4.** *Suppose that $V$ is a real or complex finite dimensional vector space, and that $\|\cdot\|$ and $\|\cdot\|'$ are both norms on $V$. Then there exist constants $c, C > 0$ such that for every $v \in V$ we have*

$$c\,\|v\| \leq \|v\|' \leq C\,\|v\|\,.$$

Theorem 4.4 roughly says that any two norms are almost the same as each other, up to a constant multiple. For certain purposes, this means that any one norm on a finite dimensional vector space is as good as any other. For example, a sequence of vectors $\{v_n\}$ in $V$ is said to converge to $v \in V$ with respect to a norm $\|\cdot\|$ if

$$\lim_{n\to\infty} \|v - v_n\| = 0.$$

Theorem 4.4 says that if $\{v_n\}$ converges to $v$ with respect to one norm, then it also converges to $v$ with respect to any other norm.

However, Theorem 4.4 is of rather limited importance in practical terms, since the constants $c$ and $C$ may be very different from each other. In particular, even for very nice, familiar norms (like the $\ell^p$ norms), these constants may be forced to be very far apart when the dimension of $V$ is large.

## 4.2 Special classes of norms

Give $x \in \mathbb{C}^n$, we define $|x| \in \mathbb{R}^n$ to be the coordinate-wise absolute value of $x$: $|x| = (|x_1|, \ldots, |x_n|)$. We also write $x \leq y$ for $x, y \in \mathbb{R}^n$ if $x_j \leq y_j$ for each $j$.

A norm $\|\cdot\|$ on $\mathbb{F}^n$ is called **monotone** if, whenever $|x| \leq |y|$, we have $\|x\| \leq \|y\|$. A norm $\|\cdot\|$ on $\mathbb{F}^n$ is called **absolute** if $\|x\| = \||x|\|$ for each $x \in \mathbb{F}^n$. For example, the $\ell^p$ norms are both monotone and absolute.

**Proposition 4.5.** *A norm on $\mathbb{F}^n$ is monotone if and only if it is absolute.*

*Proof.* Suppose that $\|\cdot\|$ is a monotone norm. Given $x \in \mathbb{F}^n$, let $y = |x|$. Then $|y| = |x|$, and therefore monotonicity implies that both $\|x\| \leq \|y\|$ and $\|y\| \leq \|x\|$. Therefore $\|\cdot\|$ is absolute.

Now suppose that $\|\cdot\|$ is an absolute norm. Let $x \in \mathbb{R}^n$. For each $j$ and $t \in [-1, 1]$, $tx_j = \frac{1+t}{2}x_j + \frac{1-t}{2}(-x_j)$. By the absolute norm property, it follows that

$$\begin{aligned}\|(x_1, \ldots, tx_j, \ldots, x_n)\| &= \left\|\frac{1+t}{2}x + \frac{1-t}{2}(x_1, \ldots, -x_j, \ldots, x_n)\right\| \\ &\leq \frac{1+t}{2}\,\|x\| + \frac{1-t}{2}\,\|(x_1, \ldots, -x_j, \ldots, x_n)\| = \|x\|\,.\end{aligned}$$

Iterating this implies that $\|\cdot\|$ satisfies the monotone norm property on $\mathbb{R}^n$. If $\mathbb{F} = \mathbb{C}$ the result follows since $\|\cdot\|$ is absolute, so it depends only on the absolute values of the components of a vector. □

A norm $\|\cdot\|$ on $\mathbb{F}^n$ is called a **symmetric gauge function** if it is an absolute norm and also $\|x\| = \|Px\|$ for each $x \in \mathbb{F}^n$ and permutation matrix $P$. Thus for example the $\ell^p$ norms are symmetric gauge functions.

**Proposition 4.6.** *Suppose that $\|\cdot\|$ is a symmetric gauge function on $\mathbb{F}^n$. If $x, y \in \mathbb{R}^n$ and $x \prec y$, then $\|x\| \leq \|y\|$.*

*Proof.* By Proposition 3.20, we can write $x = \sum_{i=1}^N t_i P_i y$ for some $t_1, \ldots, t_N \geq 0$ with $\sum_{i=1}^N t_i = 1$ and permutation matrices $P_1, \ldots, P_N$. Then since $\|\cdot\|$ is a symmetric gauge function,

$$\|x\| \leq \sum_{i=1}^N t_i \|P_i y\| = \sum_{i=1}^N t_i \|y\| = \|y\|. \qquad \square$$

As a first application of Proposition 4.6, we obtain the following result, which simultaneously generalizes Corollary 3.7 (the $\ell^\infty$ case) and the Hoffman–Wielandt inequality (Corollary 3.22, the $\ell^2$ case).

**Corollary 4.7.** *Suppose that $\|\cdot\|$ is a symmetric gauge function on $\mathbb{R}^n$. If $A, B \in M_n$ are Hermitian, then*

$$\left\|\lambda^\downarrow(A) - \lambda^\downarrow(B)\right\| \leq \|\lambda(A - B)\|.$$

*Proof.* This follows immediately from Lidskii's majorization theorem (3.16) and Proposition 4.6. $\square$

The majorization hypothesis in Proposition 4.6 can be weakened to weak majorization for nonnegative vectors.

**Proposition 4.8.** *Suppose that $\|\cdot\|$ is a symmetric gauge function on $\mathbb{R}^n$. If $x, y \in \mathbb{R}^n_+$ and $x \prec_w y$, then $\|x\| \leq \|y\|$.*

*Proof.* Without loss of generality we may assume that $x = x^\downarrow$ and $y = y^\downarrow$. Let $r = \min\{x_i \mid x_i > 0\}$, $s = \min\{y_i \mid y_i > 0\}$, and

$$u = \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \geq 0.$$

Pick $m \in \mathbb{N}$ such that $\frac{u}{m} \leq \min\{r, s\}$. Define $x', y' \in \mathbb{R}^{n+m}$ by

$$x' = \left(x_1, \ldots, x_n, \frac{u}{m}, \ldots, \frac{u}{m}\right) \qquad \text{and} \qquad y' = (y_1, \ldots, y_n, 0, \ldots, 0).$$

Then $x' \prec y'$, so by Proposition 3.20, there exist $t_1, \ldots, t_N \geq 0$ with $\sum_{i=1}^N t_i = 1$ and permutation matrices $P_1, \ldots, P_N \in M_{n+m}$ such that $x' = \sum_{i=1}^N t_i P_i y'$. It follows that $x = \sum_{i=1}^N t_i Q_i y$, where $Q_i$ is the $n \times n$ upper-left submatrix of $P_i$. The components of each $Q_i y$ are some subcollection of the components of $y$, possibly in a different order, together with some 0 components. Since $\|\cdot\|$ is both permutation-invariant and monotone, $\|Q_i y\| \leq \|y\|$ for each $i$, and so

$$\|x\| \leq \sum_{i=1}^N t_i \|Q_i y\| \leq \sum_{i=1}^N t_i \|y\| = \|y\|. \qquad \square$$

**Corollary 4.9.** *Let $p = \min\{m, n\}$, and suppose that $\|\cdot\|$ is a symmetric gauge function on $\mathbb{R}^p$. If $A, B \in M_{m,n}$, then*

$$\left\|s^{\downarrow}(A+B)\right\| \leq \left\|s^{\downarrow}(A) + s^{\downarrow}(B)\right\| \leq \left\|s^{\downarrow}(A)\right\| + \left\|s^{\downarrow}(B)\right\|.$$

*Proof.* This follows immediately from Proposition 4.8 and problem 2 from the February 18 homework. □

Another specific family of symmetric gauge functions will be important later. If $1 \leq k \leq n$, we define the $k$-norm

$$\|x\|_{(k)} = \sum_{i=1}^{k} |x|_i^{\downarrow}$$

for $x \in \mathbb{C}^n$, where $|x|_i^{\downarrow}$ denotes the $i^{\text{th}}$ largest component of $|x| \in \mathbb{R}_+^n$. It is not hard to check that $\|\cdot\|_{(k)}$ is indeed a symmetric gauge function. Note that $\|x\|_{(1)} = \|x\|_\infty$, and $\|x\|_{(n)} = \|x\|_1$; other than these special cases, these $k$-norms are different from the $\ell^p$ norms. The special role of the $k$-norms is due to the following result.

**Proposition 4.10.** *Let $x, y \in \mathbb{R}^n$. The following are equivalent.*

1. $|x| \prec_w |y|$.
2. $\|x\|_{(k)} \leq \|y\|_{(k)}$ *for each* $k = 1, \ldots, n$.
3. $\|x\| \leq \|y\|$ *for every symmetric gauge function* $\|\cdot\|$ *on* $\mathbb{R}^n$.

*Proof.* The equivalence of statements 1 and 2 follows directly from the definition of the $k$-norms. Statement 1 implies statement 3 by Proposition 4.8. Statement 3 implies statement 2 because each $k$-norm is a symmetric gauge function. □

## 4.3 Duality

Given a norm $\|\cdot\|$ on $\mathbb{F}^n$, its **dual norm** is defined by

$$\|x\|^* = \max_{\substack{y \in \mathbb{F}^n \\ \|y\| \leq 1}} |\langle x, y \rangle| = \max_{0 \neq y \in \mathbb{F}^n} \frac{|\langle x, y \rangle|}{\|y\|} = \max_{\substack{y \in \mathbb{F}^n \\ \|y\| \leq 1}} \operatorname{Re} \langle x, y \rangle .$$

The equality of the three maxima above is left as an exercise. The use of max here implicitly relies on the fact that a continuous function on a closed, bounded set achieves its maximum value. The fact that $\{y \in \mathbb{F}^n \mid \|y\| \leq 1\}$ is closed and bounded follows from Theorem 4.4.

**Proposition 4.11.** *If $\|\cdot\|$ is a norm on $\mathbb{F}^n$, then $\|\cdot\|^*$ is also a norm on $\mathbb{F}^n$.*

*Proof.* It is obvious that $\|x\|^* \geq 0$. If $\|x\|^* = 0$, then $|\langle x, y \rangle| = 0$ for every $y$ with $\|y\| \leq 1$. If $x \neq 0$, we could then let $y = \frac{x}{\|x\|}$ and obtain that $\frac{\|x\|_2^2}{\|x\|} = 0$, whence $\|x\|_2 = 0$, thus contradicting that $x \neq 0$.

The homogeneity property is obvious.

If $x, y \in \mathbb{F}^n$, then

$$\begin{aligned}\|x+y\|^* &= \max_{\|z\|\leq 1} |\langle x+y, z\rangle| \leq \max_{\|z\|\leq 1} (|\langle x, z\rangle| + |\langle y, z\rangle|) \\ &\leq \max_{\|z\|\leq 1} |\langle x, z\rangle| + \max_{\|z\|\leq 1} |y|\, z = \|x\|^* + \|y\|^* . \end{aligned}$$

□

The second maximum expression given for the dual norm gives the following interpretation: for every $x, y \in \mathbb{F}^n$

$$|\langle x, y\rangle| \leq \|x\|^* \|y\| .$$

Moreover, given $x \in \mathbb{F}^n$, $\|x\|^*$ is the smallest constant $C$ such that $|\langle x, y\rangle| \leq C \|y\|$ for every $y \in \mathbb{F}^n$.

**Proposition 4.12.** *Suppose that $1 \leq p, q \leq \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$ (where we interpret $\frac{1}{\infty}$ as 0). Then $\|\cdot\|_p^* = \|\cdot\|_q$.*

*Proof.* By Hölder's inequality, for every $x, y \in \mathbb{F}^n$ we have

$$|\langle x, y\rangle| \leq \|x\|_p \|y\|_q ;$$

this immediately implies that $\|\cdot\|_p^* \leq \|\cdot\|_q$.

For the remainder of the proof we assume that $1 < p < \infty$; the other cases are, as usual, easier and are left as an exercise.

Given $x \in \mathbb{F}^n$, define $y \in \mathbb{F}^n$ by

$$y_j = \begin{cases} \frac{|x_j|^p}{x_j} & \text{if } x_j \neq 0, \\ 0 & \text{if } x_j = 0. \end{cases}$$

Then

$$\|y\|_q = \left(\sum_{j=1}^n |y_j|^q\right)^{1/q} = \left(\sum_{j=1}^n |x_j|^{(p-1)q}\right)^{1/q} = \left(\sum_{j=1}^n |x_j|^p\right)^{\frac{1}{p}\cdot\frac{p}{q}} = \|x\|_p^{p-1}$$

and

$$|\langle x, y\rangle| = \sum_{j=1}^n |x_j|^p = \|x\|_p^p = \|x\|_p \|y\|_q .$$

From this it follows that $\|\cdot\|_p^* \geq \|\cdot\|_q$. □

The following basic properties are left as exercises:

- If $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are two norms such that $\|x\|_\alpha \leq \|x\|_\beta$ for every $x \in \mathbb{F}^n$, then $\|x\|_\beta^* \leq \|x\|_\alpha^*$ for every $x \in \mathbb{F}^n$.
- For $c > 0$, $(c\|\cdot\|)^* = \frac{1}{c}\|\cdot\|^*$.

The next result highlights the special role held by the $\ell_2$ norm.

**Proposition 4.13.** *If $\|\cdot\|$ is a norm on $\mathbb{F}^n$ and $\|\cdot\|^* = \|\cdot\|$, then $\|\cdot\| = \|\cdot\|_2$.*

*Proof.* For any $x \in \mathbb{F}^n$, $\|x\|_2^2 = |\langle x, x\rangle| \le \|x\|^* \|x\| = \|x\|^2$. Therefore $\|x\|_2 \le \|x\|$. It then follows that $\|x\| = \|x\|^* \le \|x\|_2^2 = \|x\|_2$. □

**Proposition 4.14.** *If $\|\cdot\|$ is an absolute norm on $\mathbb{F}^n$, then $\|\cdot\|^*$ is also an absolute norm. If $\|\cdot\|$ is a symmetric gauge function, then $\|\cdot\|^*$ is also a symmetric gauge function.*

*Proof.* For any $x, y \in \mathbb{F}^n$,

$$|\langle x, y\rangle| \le \sum_{j=1}^n |x_j|\,|y_j| = \langle |x|\,, |y|\rangle\,,$$

with equality when $x_j y_j \ge 0$ for each $j$. Therefore if $\|\cdot\|$ is an absolute norm, then

$$\|x\|^* = \max_{\|y\|\le 1} |\langle x, y\rangle| = \max_{\||y|\|\le 1} |\langle |x|\,, |y|\rangle|\,,$$

which implies that $\|\cdot\|^*$ is absolute.

Now suppose that $P \in M_n$ is a permutation matrix. Then, making the substitution $z = P^* y$,

$$\|Px\|^* = \max_{\|y\|\le 1} |\langle Px, y\rangle| = \max_{\|y\|\le 1} |\langle x, P^* y\rangle| = \max_{\|Pz\|\le 1} |\langle x, z\rangle| = \max_{\|z\|\le 1} |\langle x, z\rangle| = \|x\|^*\,. \qquad \square$$

**Theorem 4.15.** *If $\|\cdot\|$ is a norm on $\mathbb{F}^n$, then $\|\cdot\|^{**} = \|\cdot\|$.*

*Proof.* For any $x \in \mathbb{F}^n$,

$$\|x\|^{**} = \max_{\|y\|^*\le 1} |\langle x, y\rangle| \le \max_{\|y\|^*\le 1} \|x\|\,\|y\|^* = \|x\|\,.$$

For the opposite inequality, it suffices to assume that $\|x\|^{**} = 1$ and prove that $\|x\| \le 1$. The set $B = \{z \in \mathbb{F}^n \mid \|z\| \le 1\}$ is a closed, bounded, convex set. A basic result from convexity theory says that $B$ is equal to the intersection of the family of closed half-spaces $H_{y,t} = \{z \in \mathbb{F}^n \mid \operatorname{Re}\langle y, z\rangle \le t\}$ which contain $B$. Since $0 \in B$, we need only consider $t \ge 0$ here.

Now $B \subseteq H_{y,t}$ iff $\operatorname{Re}\langle y, z\rangle \le t$ whenever $\|z\| \le 1$, hence iff $\|y\|^* \le t$. Since $\|x\|^{**} \le 1$, we have $\operatorname{Re}\langle y, x\rangle \le 1$ whenever $\|y\|^* \le 1$; by homogeneity $\operatorname{Re}\langle y, x\rangle \le t$ whenever $\|y\|^* \le t$. Thus $x \in B$. □

**Corollary 4.16.** *For each $x \in \mathbb{F}^n$,*

$$\|x\| = \max_{\substack{y\in\mathbb{F}^n \\ \|y\|^*\le 1}} |\langle x, y\rangle| = \max_{0\ne y\in\mathbb{F}^n} \frac{|\langle x, y\rangle|}{\|y\|^*} = \max_{\substack{y\in\mathbb{F}^n \\ \|y\|^*\le 1}} \operatorname{Re}\langle x, y\rangle\,.$$

## 4.4 Matrix norms

The space of matrices $M_{m,n}(\mathbb{F})$ is a vector space, so we can consider norms on it. For example, we can consider $\ell^p$ norms via the obvious identification of $M_{m,n}(\mathbb{F})$ with $\mathbb{F}^{mn}$:

$$\|A\|_p = \left(\sum_{j=1}^m \sum_{k=1}^n |a_{jk}|^p\right)^{1/p} \qquad \text{or} \qquad \|A\|_\infty = \max_{\substack{1\le j\le m \\ 1\le k\le n}} |a_{jk}|\,.$$

When working with matrices we typically are interested in using norms that interact in some natural way with the action of matrices as linear maps on $\mathbb{F}^n$, or with the product structure on $M_n(\mathbb{F})$.

A norm $\|\cdot\|$ on $M_n(\mathbb{F})$ is called **submultiplicative** if $\|AB\| \leq \|A\| \, \|B\|$ for every $A, B \in M_n(\mathbb{F})$. A submultiplicative norm on $M_n(\mathbb{F})$ is often called a **matrix norm**, in contrast to a "vector norm", which is merely a norm on the vector space $M_n(\mathbb{F})$.[1]

For example, the $\ell^1$ norm on $M_n(\mathbb{F})$ is submultiplicative:

$$\|AB\|_1 = \sum_{j,k=1}^{n} \left| \sum_{\ell=1}^{n} a_{j\ell} b_{\ell k} \right| \leq \sum_{j,k,\ell=1}^{n} |a_{j\ell}| \, |b_{\ell k}| \leq \sum_{j,k,\ell,m=1}^{n} |a_{j\ell}| \, |b_{mk}| = \|A\|_1 \, \|B\|_1 ,$$

and so is the $\ell^2$ norm (same as the Frobenius norm):

$$\|AB\|_F^2 = \sum_{j,k=1}^{n} \left| \sum_{\ell=1}^{n} a_{j\ell} b_{\ell k} \right|^2 \leq \sum_{j,k=1}^{n} \left( \sum_{\ell=1}^{n} |a_{j\ell}|^2 \right) \left( \sum_{\ell=1}^{n} |b_{\ell k}|^2 \right) = \|A\|_F^2 \, \|B\|_F^2 ,$$

where the inequality follows from the Cauchy–Schwarz inequality. On the other hand, the $\ell^\infty$ norm on $M_n(\mathbb{F})$ is *not* submultiplicative (examples are easy to come by), although $n \, \|\cdot\|_\infty$ is.

A large class of submultiplicative norms arises from the following construction. Suppose that $\|\cdot\|_\alpha$ is a norm on $\mathbb{F}^n$ and $\|\cdot\|_\beta$ is a norm on $\mathbb{F}^m$. For $A \in M_{m,n}(\mathbb{F})$, we define[2]

$$\|A\|_{\alpha \to \beta} = \max_{\substack{x \in \mathbb{F}^n \\ \|x\|_\alpha \leq 1}} \|Ax\|_\beta = \max_{0 \neq x \in \mathbb{F}^n} \frac{\|Ax\|_\beta}{\|x\|_\alpha}.$$

That is, $\|A\|_{\alpha \to \beta}$ is the smallest constant $C > 0$ such that $\|Ax\|_\beta \leq C \, \|x\|_\alpha$ for every $x \in \mathbb{F}^n$.

**Proposition 4.17.** *Suppose that $\|\cdot\|_\alpha$ is a norm on $\mathbb{F}^n$ and $\|\cdot\|_\beta$ is a norm on $\mathbb{F}^m$. Then $\|\cdot\|_{\alpha \to \beta}$ is a norm on $M_{m,n}(\mathbb{F})$.*

The proof of Proposition 4.17 is essentially the same as the proof of Proposition 4.11 above, which it generalizes.

For example, if $A \in M_{m,n}(\mathbb{F})$, then

$$\|A\|_{2 \to 2} = \sigma_1(A).$$

This is easy to prove directly from the singular value decomposition, and is also a special case of Corollary 3.5; it also follows from the proof we gave of the singular value decomposition.

[1] Terminological warning: the term "matrix norm" is used by some authors to mean any norm on a space of matrices.

[2] Notational warning: If $\|\cdot\|_\alpha$ is a norm on $\mathbb{F}^n$, then the norm on $M_n(\mathbb{F})$ that we denote by $\|\cdot\|_{\alpha \to \alpha}$ here is often denoted $\|\cdot\|_\alpha$. Note that this conflicts with the common practice of writing, say $\|A\|_p$ for the $\ell^p$ norm of the entries of $A$. Later we will introduce yet another family of norms on matrices denoted by $\|A\|_p$. So it is vital, any time you read something about norms on matrices, to figure out what the author's notational conventions are!

It is seldom easy to express an induced norm in terms of the entries of a matrix. Two prominent exceptions are

$$\|A\|_{1\to 1} = \max_{1\le k\le n} \sum_{j=1}^{m} |a_{jk}|$$

and

$$\|A\|_{\infty\to\infty} = \max_{1\le j\le m} \sum_{k=1}^{n} |a_{jk}|\,,$$

which are also referred to as the maximum column sum and maximum row sum norms of $A$. To prove the first of these, for any $x \in \mathbb{F}^n$,

$$\begin{aligned}\|Ax\|_1 &= \sum_{j=1}^{m} \left| \sum_{k=1}^{n} a_{jk}x_k \right| \le \sum_{j=1}^{m}\sum_{k=1}^{n} |a_{jk}|\,|x_k| \\ &= \sum_{k=1}^{n} |x_k| \left( \sum_{j=1}^{m} |a_{jk}| \right) \le \|x\|_1 \max_{1\le k\le n} \sum_{j=1}^{m} |a_{jk}|\,.\end{aligned}$$

On the other hand, suppose that the $k^{\text{th}}$ column of $A$ has the largest $\ell^1$ norm of the columns of $A$. Then

$$\max_{1\le k\le n} \sum_{j=1}^{m} |a_{jk}| = \sum_{j=1}^{m} |a_{jk}| = \|Ae_k\|_1\,.$$

The expression for $\|A\|_{\infty\to\infty}$ can be proved similarly, but also follows from the the expression for $\|A\|_{1\to 1}$ using Corollary 4.21 below.

**Proposition 4.18.** *Suppose that $\|\cdot\|_\alpha$ is a norm on $\mathbb{F}^n$, $\|\cdot\|_\beta$ is a norm on $\mathbb{F}^m$, and $\|\cdot\|_\gamma$ is a norm on $\mathbb{F}^p$. Then*

$$\|AB\|_{\alpha\to\gamma} \le \|A\|_{\beta\to\gamma}\,\|B\|_{\alpha\to\beta}$$

*for every $A \in M_{p,m}(\mathbb{F})$ and $B \in M_{m,n}(\mathbb{F})$.*

*Proof.* For any $x \in \mathbb{F}^n$,

$$\|ABx\|_\gamma \le \|A\|_{\beta\to\gamma}\,\|Bx\|_\beta \le \|A\|_{\beta\to\gamma}\,\|B\|_{\alpha\to\beta}\,\|x\|_\alpha\,. \qquad \square$$

**Corollary 4.19.** *If $\|\cdot\|_\alpha$ is any norm on $\mathbb{F}^n$, then $\|\cdot\|_{\alpha\to\alpha}$ is a submultiplicative norm on $M_n(\mathbb{F})$.*

The following is an immediate reformulation of the definition of an induced norm, using the definition of a dual norm.

**Proposition 4.20.** *Suppose that $\|\cdot\|_\alpha$ is a norm on $\mathbb{F}^n$ and $\|\cdot\|_\beta$ is a norm on $\mathbb{F}^m$. Then*

$$\|A\|_{\alpha\to\beta} = \max_{\substack{\|x\|_\alpha\le 1\\ \|y\|_\beta^*\le 1}} |\langle Ax, y\rangle|\,.$$

Proposition 4.20 and Theorem 4.15 immediately imply the following.

**Corollary 4.21.** *Suppose that $\|\cdot\|_\alpha$ is a norm on $\mathbb{F}^n$ and $\|\cdot\|_\beta$ is a norm on $\mathbb{F}^m$. Then*

$$\|A\|_{\alpha\to\beta} = \|A^*\|_{\beta^*\to\alpha^*}\,,$$

*where $\|\cdot\|_{\beta^*\to\alpha^*}$ denotes the induced norm induced by $\|\cdot\|_\beta^*$ on $\mathbb{F}^m$ and $\|\cdot\|_\alpha^*$ on $\mathbb{F}^n$.*

## 4.5 The spectral radius

Given $A \in M_n(\mathbb{C})$, we call the set $\sigma(A)$ of eigenvalues (in $\mathbb{C}$) of $A$ the **spectrum** of $A$, and the number

$$\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$$

the **spectral radius** of $A$. Note that $\rho$ is *not* a norm on $M_n(\mathbb{C})$. For example, $\rho\left(\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\right) = 0$. However, it is closely related to matrix norms.

A first simple observation is that if $\|\cdot\|_\alpha$ is any norm on $\mathbb{C}^n$ and $Ax = \lambda x$ for $x \neq 0$, then

$$\|Ax\|_\alpha = \|\lambda x\|_\alpha = |\lambda| \, \|x\|_\alpha ,$$

which implies that $|\lambda| \leq \|A\|_{\alpha\to\alpha}$, and so $\rho(A) \leq \|A\|_{\alpha\to\alpha}$. This result can be generalized to arbitrary submultiplicative norms:

**Theorem 4.22.** *If $\|\cdot\|$ is a submultiplicative norm on $M_n(\mathbb{C})$, then $\rho(A) \leq \|A\|$ for every $A \in M_n(\mathbb{C})$.*

*Proof.* Suppose that $Ax = \lambda x$ for $x \neq 0$. Define $X \in M_n(\mathbb{C})$ to be the matrix whose columns are all equal to $x$. Then $AX = \lambda X$, and so

$$|\lambda| \, \|X\| = \|AX\| \leq \|A\| \, \|X\| .$$

Therefore $|\lambda| \leq \|A\|$. □

**Corollary 4.23.** *Suppose $A \in M_n(\mathbb{C})$ and that $\|\cdot\|$ is any submultiplicative norm on $M_n(\mathbb{C})$. If $\|A\| < 1$ then $I_n - A$ is nonsingular.*

*Proof.* By Theorem 4.22, $\rho(A) < 1$. The eigenvalues of $I_n - A$ are $1 - \lambda_j(A)$, so they are all nonzero, and thuse $I_n - A$ is nonsingular. □

As an application, we get the following easy-to-check sufficient condition for invertibility. We say that $A \in M_n(\mathbb{C})$ is **strictly diagonally dominant** if for each $j$,

$$|a_{jj}| > \sum_{k \neq j} |a_{kj}| .$$

**Corollary 4.24** (Levy–Desplanques theorem)**.** *If $A \in M_n(\mathbb{C})$ is strictly diagonally dominant, then $A$ is nonsingular.*

*Proof.* Let $D = \operatorname{diag}(a_{11}, \ldots, a_{nn})$. Then $D$ is nonsingular, and we define $B = I - D^{-1}A$. The entries of $B$ are

$$b_{jk} = \begin{cases} -\frac{a_{jk}}{a_{jj}} & \text{if } j \neq k, \\ 0 & \text{if } j = k. \end{cases}$$

Then $\|B\|_{\infty\to\infty} < 1$ since $A$ is strictly diagonally dominant, so $I_n - B = D^{-1}A$ is nonsingular, and hence $A$ is as well. □

**Lemma 4.25.** *Let $A \in M_n(\mathbb{C})$ and $\varepsilon > 0$ be given. Then there exists a submultiplicative norm (depending on both $A$ and $\varepsilon$) such that $\|A\| \leq \rho(A) + \varepsilon$. That is,*

$$\rho(A) = \inf \{\|A\| \mid \|\cdot\| \text{ is a submultiplicative norm on } M_n(\mathbb{C})\}.$$

*Proof.* You proved in homework (problem 3 from February 1) that there exists a nonsingular $S$ and upper triangular $T$ such that $A = STS^{-1}$ and $|t_{jk}| < \varepsilon/n$ for $j < k$. Define $\|\cdot\|$ by

$$\|B\| = \left\|S^{-1}BS\right\|_{1\to 1}.$$

Then $\|\cdot\|$ is submultiplicative, and

$$\|A\| = \|T\|_{1\to 1} = \max_{1\leq k\leq n} \sum_{j=1}^{n} |t_{jk}| \leq \max_{1\leq k\leq n} \left(|t_{kk}| + \varepsilon\right) = \rho(A) + \varepsilon. \qquad \square$$

**Theorem 4.26.** *Let $A \in M_n(\mathbb{C})$. Then $A^k \xrightarrow{k\to\infty} 0$ if and only if $\rho(A) < 1$.*

*Proof.* Suppose that $A^k \to 0$, and that $Ax = \lambda x$ for $x \neq 0$. Then $A^k x = \lambda^k x$, which implies that $\lambda^k x \to 0$, and therefore $\lambda^k \to 0$, so $|\lambda| < 1$.

Now suppose that $\rho(A) < 1$. By Lemma 4.25 there exists a submultiplicative norm $\|\cdot\|$ on $M_n(\mathbb{C})$ such that $\|A\| < 1$. Then $\left\|A^k\right\| \leq \|A\|^k$, and $\|A\|^k \to 0$, so $A^k \to 0$. $\square$

**Corollary 4.27** (The Gelfand formula)**.** *Let $\|\cdot\|$ be any submultiplicative norm on $M_n(\mathbb{C})$. Then $\rho(A) \leq \left\|A^k\right\|^{1/k}$ for each $k \in \mathbb{N}$, and*

$$\rho(A) = \lim_{k\to\infty} \left\|A^k\right\|^{1/k}.$$

*Proof.* By Theorem 4.22, $\rho(A)^k = \rho(A^k) \leq \left\|A^k\right\|$, and so $\rho(A) \leq \left\|A^k\right\|^{1/k}$.

Now given $\varepsilon > 0$, let $B = \frac{1}{\rho(A)+\varepsilon} A$. Then $\rho(B) = \frac{\rho(A)}{\rho(A)+\varepsilon} < 1$, so $B^k \xrightarrow{k\to\infty} 0$ by Theorem 4.26. Therefore there exists a $K$ such that for all $k \geq K$, $\left\|B^k\right\| < 1$. Equivalently, for all such $k$, $\left\|A^k\right\|^{1/k} \leq \rho(A) + \varepsilon$. $\square$

Since there exist submultiplicative norms that are straightforward to compute or estimate from the entries of a matrix (the Frobenius norm or the maximum row- or column-sum norms), Corollary 4.27 can be a very useful tool for estimate the spectral radius of a matrix.

A fundamental fact about infinite series of real or complex numbers is that every absolutely convergent series is convergent. That is, if $\sum_{k=0}^{\infty} |a_k|$ converges, then $\sum_{k=0}^{\infty} a_k$ converges as well. This fact extends to any finite dimensional normed space (and further to any complete normed space): if $\sum_{k=0}^{\infty} \|v_k\|$ converges, then $\sum_{k=0}^{\infty} v_k$ converges.

Another fundamental fact is that a power series $\sum_{k=0}^{\infty} a_k z^k$ with coefficients $a_k \in \mathbb{C}$ has a **radius of convergence** $R \in [0, \infty]$: $\sum_{k=0}^{\infty} a_k z^k$ converges whenever $|z| < R$ and diverges whenever $|z| > R$.

**Proposition 4.28.** *Suppose that the power series $\sum_{k=0}^{\infty} a_k z^k$ has radius of convergence $R$. Then the matrix-valued series*

$$\sum_{k=0}^{\infty} a_k A^k$$

*converges for every matrix $A \in M_n(\mathbb{C})$ with $\rho(A) < R$, and diverges for every $A$ with $\rho(A) > R$.*

*Proof.* If $\rho(A) < R$, then there exists a submultiplicative norm $\|\cdot\|$ on $M_n(\mathbb{C})$ such that $\|A\| < R$. It follows that

$$\sum_{k=0}^{\infty} \left\| a_k A^k \right\| \leq \sum_{k=0}^{\infty} |a_k| \, \|A\|^k$$

converges, so $\sum_{k=0}^{\infty} a_k A^k$ converges.

If $\rho(A) > R$, then there is a $\lambda \in \sigma(A)$ with $|\lambda| > R$. Let $x$ be a corresponding eigenvector. Then

$$\left( \sum_{k=0}^{N} a_k A^k \right) x = \left( \sum_{k=0}^{N} a_k \lambda^k \right) x,$$

and we know $\sum_{k=0}^{\infty} a_k \lambda^k$ diverges, so this series must also diverge, and hence $\sum_{k=0}^{\infty} a_k A^k$ diverges. □

The following is a matrix analogue of the formula $\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k$ for the sum of a geometric series.

**Corollary 4.29.** *Let $A \in M_n(\mathbb{C})$. If $\rho(A) < 1$, then $I_n - A$ is nonsingular, and*

$$(I_n - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

*Proof.* The eigenvalues of $I_n - A$ are $1 - \lambda_j(A)$, so if $\rho(A) < 1$ then all the eigenvalues of $I_n - A$ are nonzero, hence $I_n - A$ is nonsingular.

Now the radius of convergence of $\sum_{k=0}^{\infty} z^k$ is 1, so Proposition 4.28 implies that $\sum_{k=0}^{\infty} A^k$ converges, say to $B$. We have

$$(I_n - A) \sum_{k=0}^{N} A^k = I_n - A^{n+1},$$

and taking the limit $N \to \infty$ yields, by Theorem 4.26, that $(I_n - A)B = I_n$. □

Replacing $A$ with $I_n - A$ in Corollary 4.29 yields the following.

**Corollary 4.30.** *Let $A \in M_n(\mathbb{C})$. If $\sigma(A) \subseteq (0, 2)$, then*

$$A^{-1} = \sum_{k=0}^{\infty} (I_n - A)^k.$$

## 4.6 Unitarily invariant norms

A norm $\|\cdot\|$ on $M_{m,n}(\mathbb{C})$ is called **unitarily invariant** if $\|UAV\| = \|A\|$ for every $A \in M_{m,n}(\mathbb{C})$, $U \in \mathcal{U}_m$, and $V \in \mathcal{U}_n$.

We have already seen two examples of unitarily invariant norms: the Frobenius norm $\|\cdot\|_F$ and the standard operator norm $\|\cdot\|_{2\to 2}$.

If $A = U\Sigma V^*$ is a singular value decomposition and $\|\cdot\|$ is unitarily invariant, then $\|A\| = \|\Sigma\|$. Therefore a unitarily invariant norm depends only on the singular values of $A$. For example, as we have seen previously,

$$\|A\|_F = \sqrt{\sum_{j=1}^{p} \sigma_j(A)^2} \qquad \text{and} \qquad \|A\|_{2\to 2} = \sigma_1(A).$$

It turns out that any unitarily invariant norm on $M_{m,n}(\mathbb{C})$ can be described in terms of a norm of the sequence of singular values $s(A) \in \mathbb{R}^p$, where $p = \min\{m, n\}$. More precisely, unitarily invariant norms on $M_{m,n}(\mathbb{C})$ are in one-to-one correspondence with symmetric gauge functions on $\mathbb{R}^p$:[3]

**Theorem 4.31.** *Suppose that $\|\cdot\|$ is a unitarily invariant norm on $M_{m,n}(\mathbb{C})$. There is a symmetric gauge function on $\mathbb{R}^p$, again denoted $\|\cdot\|$, such that $\|A\| = \|s(A)\|$.*

*Conversely, given a symmetric gauge function on $\mathbb{R}^p$, the formula $\|A\| = \|s(A)\|$ defines a unitarily invariant norm on $M_{m,n}(\mathbb{C})$.*

*Proof.* Suppose that $\|\cdot\|$ is a unitarily invariant norm on $M_{m,n}(\mathbb{C})$. For simplicity, we assume that $m \le n$, the proof in the case $m > n$ being similar.

Given $x \in \mathbb{R}^m$, we define $\|x\|$ as follows: let

$$M(x) = \begin{bmatrix} x_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & x_m & 0 & \cdots & 0 \end{bmatrix} \in M_{m,n}(\mathbb{C}),$$

and set $\|x\| = \|M(x)\|$. All the properties of a norm follow easily (this essentially amounts to observing that the restriction of a norm to a subspace is still a norm).

Given $x \in \mathbb{R}^m$, define $D \in M_m$ to be the diagonal matrix $D = \operatorname{diag}(d_1, \ldots, d_m)$, where

$$d_j = \begin{cases} 1 & \text{if } x \ge 0, \\ -1 & \text{if } x < 0. \end{cases}$$

Then $D$ is unitary, and so

$$\||x|\| = \|M(|x|)\| = \|DM(x)\| = \|M(x)\| = \|x\|$$

by unitary invariance. Therefore $\|\cdot\|$ on $\mathbb{R}^m$ is an absolute norm.

[3] As you already saw in homework, symmetric gauge functions on $\mathbb{R}^p$ and $\mathbb{C}^p$ are essentially the same thing.

Now let $P \in M_m$ be a permutation matrix. Define $Q \in M_n$ to have the block decomposition

$$Q = \begin{bmatrix} P & 0 \\ 0 & I_{n-m} \end{bmatrix},$$

so $Q$ is also a permutation matrix, and $P$ and $Q$ are both unitary. Then

$$\|Px\| = \|M(Px)\| = \|PM(x)Q^*\| = \|M(x)\| = \|x\|.$$

Therefore $\|\cdot\|$ on $\mathbb{R}^m$ is a symmetric gauge function.

Now suppose that $\|\cdot\|$ is a symmetric gauge function on $\mathbb{R}^p$, and define $\|A\| = \|s(A)\|$ for $A \in M_{m,n}(\mathbb{C})$. It follows immediately that $\|A\| \geq 0$ and $\|A\| = 0$ iff $A = 0$. If $c \in \mathbb{C}$, then so $s(cA) = |c|\, s(A)$, and so $\|cA\| = |c|\, \|A\|$. The triangle inequality for $\|\cdot\|$ on $M_{m,n}(\mathbb{C})$ follows from Corollary 4.9.

Finally, if $A \in M_{m,n}(\mathbb{C})$, $U \in \mathcal{U}_m$, and $V \in \mathcal{U}_n$ are given, then $s(UAV) = s(A)$, and so $\|\cdot\|$ on $M_{m,n}(\mathbb{C})$ is unitarily invariant. □

Using Theorem 4.31, we get additional examples of unitarily invariant norms:

- For $1 \leq p < \infty$, the **Schatten $p$-norm** of $A \in M_{m,n}(\mathbb{C})$ is given by

$$\|A\|_p = \left( \sum_{j=1}^{\min\{m,n\}} \sigma_j(A)^p \right)^{1/p},$$

  and for $p = \infty$, the Schatten $\infty$-norm is

$$\|A\|_\infty = \max_{1 \leq j \leq \min\{m,n\}} \sigma_j(A) = \sigma_1(A).$$

  Note that $\|A\|_2 = \|A\|_F$ and $\|A\|_\infty = \|A\|_{2\to 2}$. The other Schatten norms are new to us. One other that is frequently singled out is the Schatten 1-norm, which is sometimes called the **trace norm** $\|\cdot\|_{\mathrm{tr}}$ (for reasons that will be explored in homework).

- For $1 \leq k \leq n$, the **Fan $k$-norm** of $A \in M_{m,n}(\mathbb{C})$ is given by

$$\|A\|_{(k)} = \sum_{j=1}^{k} \sigma_j(A).$$

  Note that $\|A\|_{(1)} = \|A\|_\infty = \|A\|_{2\to 2}$, and that $\|A\|_{\min\{m,n\}} = \|A\|_1 = \|A\|_{\mathrm{tr}}$.

The Fan $k$-norms play an important role in the general theory of unitarily invariance norms, thanks to the following result, which follows immediately from Theorem 4.31 and Proposition 4.10.

**Theorem 4.32** (Fan dominance principle)**.** *Let $A, B \in M_{m,n}(\mathbb{C})$. If $\|A\|_{(k)} \leq \|B\|_{(k)}$ foer each $1 \leq k \leq \min\{m,n\}$, then $\|A\| \leq \|B\|$ for every unitarily invariant norm $\|\cdot\|$ on $M_{m,n}(\mathbb{C})$.*

We next address the question of when which unitarily invariant norms fit into other special classes of norms we have considered.

**Theorem 4.33.** *A norm $\|\cdot\|$ on $M_{m,n}(\mathbb{C})$ is unitarily invariant if and only if*

$$\|ABC\| \leq \|A\|_{2\to 2}\, \|B\|\, \|C\|_{2\to 2}$$

*for every $A \in M_m$, $B \in M_{m,n}$, and $C \in M_n$.*

*Proof.* Suppose first that $\|\cdot\|$ is unitarily invariant. The Courant–Fischer theorem for singular values (Corollary 3.5) implies that $\sigma_j(ABC) \leq \|A\|_{2\to 2}\, \sigma_j(B)\, \|C\|_{2\to 2}$ for each $j$, and so $s(ABC) \leq \|A\|_{2\to 2}\, \|C\|_{2\to 2}\, s(B)$ coordinate-wise. The symmetric gauge function on $\mathbb{R}^{\min\{m,n\}}$ corresponding to $\|\cdot\|$ is an absolute, and hence monotone norm, and so

$$\begin{aligned}\|ABC\| = \|s(ABC)\| &\leq \|\|A\|_{2\to 2}\, \|C\|_{2\to 2}\, s(B)\| \\ &= \|A\|_{2\to 2}\, \|C\|_{2\to 2}\, \|s(B)\| = \|A\|_{2\to 2}\, \|C\|_{2\to 2}\, \|B\|\,.\end{aligned}$$

Now suppose that $\|\cdot\|$ has the stated property. Given $A \in M_{m,n}(\mathbb{C})$, $U \in \mathcal{U}_m$, and $V \in \mathcal{U}_n$, we have

$$\|UAV\| \leq \|U\|_{2\to 2}\, \|A\|\, \|V\|_{2\to 2} = \|A\|$$

and

$$\|A\| = \|U^*UAVV^*\| \leq \|U^*\|_{2\to 2}\, \|UAV\|\, \|V^*\|_{2\to 2} = \|UAV\|\,. \qquad \square.$$

**Corollary 4.34.** *Suppose that $\|\cdot\|$ is a unitarily invariant norm on $M_n(\mathbb{C})$, and let $E_{11} = \operatorname{diag}(1, 0, \ldots, 0) \in M_n(\mathbb{C})$. If $\|E_{11}\| = 1$, then $\|\cdot\|$ is submultiplicative.*

*Proof.* Let $\|\cdot\|$ also denote the corresponding symmetric gauge function on $\mathbb{R}^n$. The normalization assumption implies that $\|e_1\| = 1$, which further implies that for any $x \in \mathbb{R}^n$,

$$\|x\| = \left\| |x|^{\downarrow} \right\| \geq \|\|x\|_\infty\, e_1\| = \|x\|_\infty\,,$$

where the inequality above follows by monotonicity. It follows that $\|A\|_{2\to 2} \leq \|A\|$ for any $A \in M_n(\mathbb{C})$. The claim now follows immediately from Theorem 4.33. $\square$

**Lemma 4.35.** *If $\|\cdot\|_\alpha$ is a norm on $\mathbb{F}^n$, $\|\cdot\|_\beta$ is a norm on $\mathbb{F}^m$, and $x \in \mathbb{F}^m$, $y \in \mathbb{F}^n$, then*

$$\|xy^*\|_{\alpha\to\beta} = \|x\|_\beta\, \|y\|_\alpha^*\,.$$

The proof of Lemma 4.35 is left as an exercise.

**Theorem 4.36.** *Suppose that $\|\cdot\|_\alpha$ is a norm on $\mathbb{C}^n$ and $\|\cdot\|_\beta$ is a norm on $\mathbb{C}^m$, and that the induced norm $\|\cdot\|_{\alpha\to\beta}$ on $M_{m,n}(\mathbb{C})$ is unitarily invariant. Then there exist $a, b > 0$ such that*

$$\|\cdot\|_\alpha = a\,\|\cdot\|_2\,, \qquad \|\cdot\|_\beta = b\,\|\cdot\|_2\,, \qquad \text{and} \qquad \|\cdot\|_{\alpha\to\beta} = \frac{b}{a}\,\|\cdot\|_{2\to 2}\,.$$

*In particular, if an induced norm $\|\cdot\|_{\alpha\to\alpha}$ on $M_n(\mathbb{C})$ is unitarily invariant, then $\|\cdot\|_{\alpha\to\alpha} = \|\cdot\|_{2\to 2}$.*

*Proof.* Let $x \in \mathbb{C}^m$, $y \in \mathbb{C}^n$, $U \in \mathcal{U}_m$, and $V \in \mathcal{U}_n$. Then by unitary invariance and Lemma 4.35,

$$\|x\|_\beta \|y\|_\alpha^* = \|xy^*\|_{\alpha\to\beta} = \|Uxy^*V^*\|_{\alpha\to\beta} = \|Ux\|_\beta \|Vy\|_\alpha^*.$$

Therefore

$$\frac{\|Ux\|_\beta}{\|x\|_\beta} = \frac{\|Vy\|_\alpha^*}{\|y\|_\alpha^*}$$

for all nonzero $x$ and $y$. In particular, the value $C$ of this ratio is independent of $x$, $y$, $U$, and $V$. Setting $U = I_m$ or $V = I_n$ implies that $C = 1$.

So $\|Ux\|_\beta = \|x\|_\beta$ for every $x \in \mathbb{C}^m$ and $U \in \mathcal{U}_m$. Given $x \neq 0$, let $v = \frac{x}{\|x\|_2}$. There is a $U \in \mathcal{U}_m$ such that $Uv = e_1$. It follows that

$$\|x\|_\beta = \|v\|_\beta \|x\|_2 = \|Uv\|_\beta \|x\|_2 = \|e_1\|_\beta \|x\|_2$$

for every $x \in \mathbb{C}^m$.

Similarly, $\|y\|_\alpha^* = \|e_1\|_\alpha^* \|y\|_2$ for every $y \in \mathbb{C}^n$, and therefore $\|\cdot\|_\alpha = \|\cdot\|_\alpha^{**} = (\|e_1\|_\alpha^*)^{-1} \|\cdot\|_2$. □

**Theorem 4.37.** *Suppose that $\|\cdot\|$ is a unitarily invariant and absolute norm on $M_{m,n}(\mathbb{C})$. Then there is a constant $c > 0$ such that $\|\cdot\| = c\|\cdot\|_F$.*

*Proof.* Without loss of generality we may assume that $\|E_{11}\| = 1$. Let $A \in M_{m,n}(\mathbb{C})$. We will prove that $\|A\| = \|A\|_F$ by induction on $\operatorname{rank} A$. The case $\operatorname{rank} A = 0$ is trivial; if $\operatorname{rank} A = 1$ then the singular value decomposition of $A$ has the form $A = U(\sigma_1 E_{11})V^*$, so $\|A\| = \|\sigma_1 E_{11}\| = \sigma_1 = \|A\|_F$.

Now suppose that $r = \operatorname{rank} A \geq 2$, and that the result is known for matrices of rank smaller than $r$. Let $\sigma_j = \sigma_j(A)$, we define

$$a = \frac{\sqrt{\sigma_1^2 + \sigma_r^2} + \sigma_1 - \sigma_r}{2}, \qquad b = \sqrt{\frac{\sigma_1\sigma_r}{2}}, \qquad \text{and} \qquad c = \frac{\sqrt{\sigma_1^2 + \sigma_r^2} - \sigma_1 + \sigma_r}{2},$$

and then define

$$A_\pm = \begin{bmatrix} a & b & 0 & \cdots & 0 & 0 & \cdots & 0 \\ b & \pm c & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_{r-1} & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \cdots & 0 & \cdots & 0 \end{bmatrix} \in M_{m,n}(\mathbb{C}).$$

The matrices $\begin{bmatrix} a & b \\ b & \pm c \end{bmatrix}$ are real symmetric, with eigenvalues $\sigma_1$ and $-\sigma_r$ in the $-$ case, and $\sqrt{\sigma_1^2 + \sigma_r^2}$ and 0 in the $+$ case. It follows that $s(A_-) = s(A)$, and

$$s(A_+) = (\sqrt{\sigma_1^2 + \sigma_r^2}, \sigma_2, \ldots, \sigma_{r-1}, 0, \ldots, 0).$$

Therefore

$$\|A\| = \|A_-\| = \|A_+\| = \|A_+\|_F = \|A\|_F,$$

by (in order) unitary invariance, the fact that $\|\cdot\|$ is absolute, the induction hypothesis, and the expression for $\|\cdot\|_F$ in terms of singular values. □

Recall that Corollary 4.7 showed that if $A, B \in M_n$ are Hermitian, then

$$\left\|\lambda^{\downarrow}(A) - \lambda^{\downarrow}(B)\right\| \leq \|\lambda(A - B)\|$$

for each symmetric gauge function $\|\cdot\|$ on $\mathbb{R}^n$. Using Theorem 4.31, we can rephrase this as follows.

**Theorem 4.38** (Mirsky's inequality)**.** *Let $\|\cdot\|$ denote both a symmetric gauge function on $\mathbb{R}^n$ and the corresponding unitarily invariant norm on $M_n$. Then*

$$\left\|\lambda^{\downarrow}(A) - \lambda^{\downarrow}(B)\right\| \leq \|A - B\|$$

*for any Hermitian matrices $A, B \in M_n$.*

This states in a very precise, quantitative, and general way that the eigenvalues of a Hermitian matrix depend in a continuous fashion on the matrix itself. Note that this includes as special cases both Weyl's perturbation theorem (Corollary 3.7, when $\|\cdot\| = \|\cdot\|_\infty$) and the Hoffman–Wielandt inequality for Hermitian matrices (Corollary 3.22, when $\|\cdot\| = \|\cdot\|_2$).

For general (non-Hermitian, possibly non-square) matrices, there is a similar theorem for singular values.

**Corollary 4.39.** *Let $\|\cdot\|$ denote both a symmetric gauge function on $\mathbb{R}^p$ and the corresponding unitarily invariant norm on $M_{m,n}$, with $p = \min\{m, n\}$. Then*

$$\left\|s^{\downarrow}(A) - s^{\downarrow}(B)\right\| \leq \|A - B\|$$

*for any $A, B \in M_{m,n}$.*

*Proof.* Recall from homework that the eigenvalues of $\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}$ are $\pm$ the singular values of $A$, plus some 0s. Lidskii's majorization theorem (Theorem 3.16) implies that

$$\lambda^{\downarrow}\left(\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}\right) - \lambda^{\downarrow}\left(\begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix}\right) \prec \lambda^{\downarrow}\left(\begin{bmatrix} 0 & A - B \\ (A-B)^* & 0 \end{bmatrix}\right).$$

Note that the nonzero entries of the left hand side of the majorization above are $\pm|\sigma_j(A) - \sigma_j(B)|$. Then for $1 \leq k \leq p$,

$$\sum_{j=1}^{k} \sigma_j(A - B) \geq \sum_{j=1}^{k} \left[\lambda^{\downarrow}\left(\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}\right) - \lambda^{\downarrow}\left(\begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix}\right)\right]_j^{\downarrow} \geq \sum_{j=1}^{k} \left|s^{\downarrow}(A) - s^{\downarrow}(B)\right|_j^{\downarrow}.$$

If $A = U_A \Sigma_A V_A^*$ and $B = U_B \Sigma_B V_B^*$ are singular value decompositions, then the inequality above states that $\|A - B\|_{(k)} \geq \|\Sigma_A - \Sigma_B\|_{(k)}$ for every $1 \leq k \leq p$. By Fan's dominance principle (Theorem 4.32), this implies that $\|A - B\| \geq \|\Sigma_A - \Sigma_B\|$ for every unitarily invariant norm $\|\cdot\|$, which is equivalent to the claim. □

Corollary 4.39 is a powerful tool for determining how well a matrix of one kind can be approximated by another kind of matrix, with respect to unitarily invariant norms.

**Corollary 4.40.** *Let $\|\cdot\|$ denote both a symmetric gauge function on $\mathbb{R}^p$ and the corresponding unitarily invariant norm on $M_{m,n}$, with $p = \min\{m, n\}$. Suppose that $A \in M_{m,n}$ has singular value decomposition $A = \sum_{j=1}^{p} \sigma_j u_j v_j^*$ and $B \in M_{m,n}$ has rank $k$. Then*

$$\|A - B\| \geq \|(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p)\|,$$

*with equality when $B = \sum_{j=1}^{k} \sigma_j u_j v_j^*$.*

*Proof.* By Corollary 4.39, if $\operatorname{rank} B = k$ then

$$\|A - B\| \geq \|(\sigma_1(A) - \sigma_1(B), \ldots, \sigma_k(A) - \sigma_k(B), \sigma_{k+1}(A), \ldots, \sigma_p(A))\|.$$

The right-hand-side above is greater than the right-hand-side in the statement of the corollary since $\|\cdot\|$ is absolute and monotone on $\mathbb{R}^p$. The equality case is immediate. □

A special case worth noting is that if $A \in M_n$ is nonsingular with singular value decomposition $A = \sum_{j=1}^{n} \sigma_j u_j v_j^*$, then the closest singular matrix $B$ with respect to any unitarily invariant norm is $B = \sum_{j=1}^{n-1} \sigma_j u_j v_j^*$, and then $\|A - B\| = \sigma_n \|e_1\|$. (Compare this to problem 4 from the March 1 homework.)

## 4.7 Duality for matrix norms

Duality was defined above specifically for norms on $\mathbb{F}^n$, but the definition obviously extends to any space with a fixed inner product. Here we will extend the definition to $M_{m,n}(\mathbb{F})$ with the Frobenius inner product $\langle A, B \rangle_F = \operatorname{tr} AB^*$ (which we recall is the same as the standard inner product when we identify $M_{m,n}(\mathbb{F})$ with $\mathbb{F}^{mn}$ in the obvious way). The definition of the **dual norm** to a given norm $\|\cdot\|$ on $M_{m,n}(\mathbb{F})$ becomes

$$\|A\|^* = \max_{\substack{B \in M_{m,n}(\mathbb{F}) \\ \|B\| \leq 1}} |\operatorname{tr} AB^*| = \max_{0 \neq B \in M_{m,n}(\mathbb{F})} \frac{|\operatorname{tr} AB^*|}{\|B\|} = \max_{\substack{B \in M_{m,n}(\mathbb{F}) \\ \|B\| \leq 1}} \operatorname{Re} \operatorname{tr} AB^*.$$

Because of the appearance of traces here, this is sometimes referred to as **trace duality**.

**Theorem 4.41.** *Suppose that $\|\cdot\|$ is a submultiplicative norm on $M_n(\mathbb{F})$. Then*

$$\|AB\|^* \leq \min\left\{\|A^*\| \|B\|^*, \|A\|^* \|B^*\|\right\}$$

*for all $A, B \in M_n(\mathbb{F})$. In particular, if $\|A^*\| \leq \|A\|^*$ for every $A \in M_n(\mathbb{F})$, then $\|\cdot\|^*$ is submultiplicative.*

*Proof.* If $A, B, C \in M_n(\mathbb{F})$, then

$$\operatorname{tr}(AB)C^* = \operatorname{tr} A(CB^*)^* \leq \|A\|^* \|CB^*\| \leq \|A\|^* \|C\| \|B^*\|,$$

where the first inequality follows from the definition of the dual norm $\|\cdot\|^*$, and the second inequality from the submultiplicativity of $\|\cdot\|$. This implies that $\|AB\|^* \leq \|A\|^* \|B^*\|$. The inequality $\|AB\|^* \leq \|A^*\| \|B\|^*$ is proved similarly. □

**Corollary 4.42.** *Let $\|\cdot\|_\alpha$ be a norm on $\mathbb{F}^n$. Then $\|\cdot\|^*_{\alpha\to\alpha}$ is a submultiplicative norm on $M_n(\mathbb{F})$.*

*Proof.* By Corollary 4.21 and Proposition 4.20,

$$\begin{aligned}\|A^*\|_{\alpha\to\alpha} = \|A\|_{\alpha^*\to\alpha^*} &= \max_{\substack{\|x\|^*_\alpha\le 1\\ \|y\|_\alpha\le 1}} |\langle Ax, y\rangle| = \max_{\substack{\|x\|^*_\alpha\le 1\\ \|y\|_\alpha\le 1}} |y^*Ax| = \max_{\substack{\|x\|^*_\alpha\le 1\\ \|y\|_\alpha\le 1}} |\operatorname{tr} Axy^*| \\ &\le \max_{\substack{\|x\|^*_\alpha\le 1\\ \|y\|_\alpha\le 1}} \|A\|^*_{\alpha\to\alpha}\,\|yx^*\|_{\alpha\to\alpha} \le \|A\|^*_{\alpha\to\alpha}\,,\end{aligned}$$

where the last inequality follows from Lemma 4.35. The result now follows from Theorem 4.41. □

Recall that Proposition 4.14 shows that the dual norm of a symmetric gauge function is again a symmetric gauge function. It is therefore obvious to guess that the dual of a unitarily invariant norm is again unitarily invariant, and that the corresponding symmetric gauge functions are dual as well. This guess turns out to be correct; we will need a technical preliminary result.

A matrix $A \in M_n(\mathbb{R})$ is called **doubly substochastic** if $a_{ij} \ge 0$ for each $i, j$ and, $\sum_{i=1}^n a_{ij} \le 1$ for each $j$, and $\sum_{j=1}^n a_{ij} \le 1$ for each $i$.

**Lemma 4.43.** *If $A \in M_n(\mathbb{R})$ is doubly substochastic, then there exists a doubly stochastic matrix $B \in M_n(\mathbb{R})$ such that $A \le B$ entrywise.*

*Proof.* Note that the sum of the column sums of $A$ is equal to the sum of the row sums of $A$, so if row sum is $< 1$ then some column sum is $< 1$ as well. Pick the smallest $i$ such that the $i^{\text{th}}$ row sum is $< 1$ and the smallest $j$ such that the $i^{\text{th}}$ column sum is $< 1$. Increase $a_{ij}$ until one of these row/column sums is 1. This decreases the total number of row and column sums which are $< 1$. Iterate this procedure; the process must eventually terminate because there are only finitely many rows and columns. At the end we obtain a matrix whose entries are greater than or equal to the entries of $A$, and whose row and column sums are all equal to 1. □

**Theorem 4.44.** *Let $\|\cdot\|$ be a unitarily invariant norm on $M_{m,n}(\mathbb{C})$, corresponding to the symmetric gauge function $\|\cdot\|_v$ on $\mathbb{R}^p$, where $p = \min\{m, n\}$. Then $\|\cdot\|^*$ is the unitarily invariant norm corresponding to the symmetric gauge function $\|\cdot\|^*_v$.*

*Proof.* First observe that if $U \in \mathcal{U}_m$ and $V \in \mathcal{U}_n$, then

$$\|UAV\|^* = \max_{\|B\|\le 1} |\operatorname{tr} UAVB^*| = \max_{\|B\|\le 1} |\operatorname{tr} A(U^*BV^*)^*| = \max_{\|C\|\le 1} |\operatorname{tr} AC^*| = \|A\|^*\,,$$

where we have made the substitution $C = U^*BV^*$. Thus $\|\cdot\|^*$ is unitarily invariant.

Suppose for now that $m = n$. Let $A = U_A\Sigma_A V_A^*$ and $B = U_B\Sigma_B V_B^*$ be singular value decompositions, and define $U = U_B^*U_A$ and $V = V_A^*V_B$. Then

$$|\operatorname{tr} AB^*| = |\operatorname{tr} \Sigma_A V \Sigma_B U| = \left|\sum_{j,k=1}^n \sigma_j(A) v_{jk} \sigma_k(B) u_{kj}\right| \le \sum_{j=1}^n \sigma_j(A) \sum_{k=1}^n |v_{jk}u_{kj}|\,\sigma_k(B).$$

Let $w_{jk} = |v_{jk}u_{kj}|$. By the Cauchy–Schwarz inequality, for each $k$ we have

$$\sum_{j=1}^{n} w_{jk} \leq \sqrt{\sum_{j=1}^{n} |v_{jk}|^2} \sqrt{\sum_{j=1}^{n} |u_{kj}|^2} = 1$$

since $V$ and $U$ are both unitary. Similarly $\sum_{k=1}^{n} w_{jk} \leq 1$ for each $j$. Therefore $W$ is substochastic, and by Lemma 4.43 there exists a doubly stochastic matrix $C$ such that $W \leq C$ entrywise. We then have

$$|\operatorname{tr} AB^*| \leq \langle s(A), Ws(B)\rangle \leq \langle s(A), Cs(B)\rangle .$$

Now $Cs(B) \prec s(B)$ by Proposition 3.17, and so by Proposition 3.20, $Cs(B) = \sum_{i=1}^{N} t_i P_i s(B)$ for some $t_i \geq 0$ with $\sum_{i=1}^{N} t_i = 1$ and permutation matrices $P_i$. It follows that

$$|\operatorname{tr} AB^*| \leq \sum_{i=1}^{N} t_i \langle s(A), P_i s(B)\rangle .$$

Now if $m \neq n$, we can add rows or columns of 0s as necessary to make $A$ and $B$ square; this adds 0s to the vectors $s(A)$ and $s(B)$. If we call the resulting extended vectors $\widetilde{s}(A) = (s(A), 0, \ldots, 0)$ and $\widetilde{s}(B)$ similarly, then the argument above yields

$$|\operatorname{tr} AB^*| \leq \sum_{i=1}^{N} t_i \langle \widetilde{s}(A), P_i \widetilde{s}(B)\rangle .$$

For each $i$, we could further permute the entries of $P_i\widetilde{s}(B)$ so that all the nonzero entries appear among the first $p$; we obtain that there are permutation matrices $Q_i \in M_p$ such that $\langle \widetilde{s}(A), P_i\widetilde{s}(B)\rangle = \langle s(A), Q_i s(B)\rangle$. Therefore

$$\begin{aligned} |\operatorname{tr} AB^*| &\leq \sum_{i=1}^{N} t_i \langle \widetilde{s}(A), Q_i\widetilde{s}(B)\rangle \leq \sum_{i=1}^{N} t_i \|s(A)\|_v^* \|Q_i s(B)\|_v \\ &= \sum_{i=1}^{N} t_i \|s(A)\|_v^* \|s(B)\|_v = \|s(A)\|_v^* \|B\| , \end{aligned}$$

which implies that $\|A\|^* \leq \|s(A)\|_v^*$.

Finally, given $A$ with singular value decomposition $A = U\Sigma V^*$, pick $x \in \mathbb{R}^p$ such that $|\langle s(A), x\rangle| = \|s(A)\|_v^* \|x\|_v$, and define $M(x) \in M_{m,n}$ as before and $B = UM(x)V^*$. Then $\|M(x)\| = \|x\|_v$ and

$$\operatorname{tr} AB^* = \operatorname{tr} \Sigma M(x)^* = \langle s(A), x\rangle ,$$

which implies that $\|A\|^* \geq \|s(A)\|^*$. □

# 5 Some topics in solving linear systems

## 5.1 Condition numbers

Consider a linear system of equations, written in matrix form as $Ax = b$. We will restrict attention here to an $n \times n$ system, so $A \in M_n$ and $b \in \mathbb{F}^n$ are fixed, and we wish to find the

unknown vector $x \in \mathbb{F}^n$. Assuming that $A$ is nonsingular, we could do this by computing $A^{-1}$ and then $x = A^{-1}b$. (This isn't necessarily the best way to find $x$, depending on what constitutes "best" for our purposes, but we will assume that approach for now.)

Now suppose that the vector $b$ is not precisely known, due to measurement error, round-off error in an earlier calculation, or any number of other real-world issues. That is, in place of the true vector $b$, we are actually using $b + \Delta b$, where $\Delta b \in \mathbb{F}^n$ represents the error or uncertainty in $b$. Then our computed value of $x$ is not $A^{-1}b$ but instead

$$A^{-1}(b + \Delta b) = A^{-1}b + A^{-1}(\Delta b).$$

That is, we get an error $\Delta x = A^{-1}(\Delta b)$ in the computed value.

We can use norms to quantify the size of the error. Let $\|\cdot\|_\alpha$ be a norm on $\mathbb{F}^n$. Then

$$\|\Delta x\|_\alpha = \left\|A^{-1}(\Delta b)\right\|_\alpha \leq \left\|A^{-1}\right\|_{\alpha\to\alpha} \|\Delta b\|_\alpha .$$

More relevant for many purposes is the *relative* size of the error:

$$\frac{\|\Delta x\|_\alpha}{\|x\|_\alpha} \leq \frac{\left\|A^{-1}\right\|_{\alpha\to\alpha} \|\Delta b\|_\alpha}{\|A^{-1}b\|_\alpha} = \left\|A^{-1}\right\|_{\alpha\to\alpha} \frac{\|b\|_\alpha}{\|A^{-1}b\|_\alpha} \frac{\|\Delta b\|_\alpha}{\|b\|_\alpha} \leq \left\|A^{-1}\right\|_{\alpha\to\alpha} \|A\|_{\alpha\to\alpha} \frac{\|\Delta b\|_\alpha}{\|b\|_\alpha}.$$

We define the **condition number** of $A \in M_n$ with respect to the submultiplicative norm $\|\cdot\|$ to be

$$\kappa_{\|\cdot\|}(A) = \begin{cases} \|A\| \left\|A^{-1}\right\| & \text{if } A \text{ is nonsingular,} \\ \infty & \text{if } A \text{ is singular.} \end{cases}$$

Note that $\kappa_{\|\cdot\|}(A) \geq \|I_n\| \geq 1$ by submultiplicativity. When $\|\cdot\|$ is the operator norm $\|\cdot\|_{\alpha\to\alpha}$ induced by the norm $\|\cdot\|_\alpha$ on $\mathbb{F}^n$, we write $\kappa_\alpha$. The particular case $\kappa_2$ with respect to the $\ell^2$ norm is usually simply called the condition number of $A$; note that $\kappa_2(A) = \frac{\sigma_1(A)}{\sigma_n(A)}$.

We have just seen that in solving $Ax = b$ with an error $\Delta b$ in the right hand side, using the known value of $A^{-1}$,

$$\frac{\|\Delta x\|_\alpha}{\|x\|_\alpha} \leq \kappa_\alpha(A) \frac{\|\Delta b\|_\alpha}{\|b\|_\alpha}.$$

That is, the condition number bounds how much the size of the relative error (with respect to the norm $\|\cdot\|_\alpha$) is increased.

Now suppose that there is some uncertainty or error in the matrix $A$ itself. How will that error propagate to the computed inverse matrix $A^{-1}$? It turns out that condition numbers control the relative error here as well.

**Proposition 5.1.** *Let $\|\cdot\|$ be a submultiplicative norm on $M_n$ and write $\kappa = \kappa_{\|\cdot\|}$. Suppose that $A \in M_n$ is nonsingular, and that $\left\|A^{-1}\right\| \|\Delta A\| = \kappa(A)\frac{\|\Delta A\|}{\|A\|} < 1$. Then $A + \Delta A$ is nonsingular, and*

$$\frac{\left\|A^{-1} - (A + \Delta A)^{-1}\right\|}{\|A^{-1}\|} \leq \frac{\kappa(A)\frac{\|\Delta A\|}{\|A\|}}{1 - \kappa(A)\frac{\|\Delta A\|}{\|A\|}}.$$

Note that $\frac{x}{1-x} = x + x^2 + x^3 + \cdots \approx x$ for small $x$. Thus the upper bound in the inequality in Proposition 5.1 is about $\kappa(A)\frac{\|\Delta A\|}{\|A\|}$ when this quantity is small.

*Proof.* Let $B = A + \Delta A = A(I_n + A^{-1}\Delta A)$. By submultiplicativity and the hypothesis, $\left\|A^{-1}\Delta A\right\| \leq \left\|A^{-1}\right\| \|\Delta A\| < 1$, so by Corollary 4.29, $I_n + A^{-1}\Delta A$ is nonsingular, and hence $B$ is nonsingular as well. Now

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1} = A^{-1}(\Delta A)B^{-1},$$

so

$$\left\|A^{-1} - B^{-1}\right\| \leq \left\|A^{-1}\Delta A\right\| \left\|B^{-1}\right\|.$$

By the triangle inequality,

$$\left\|B^{-1}\right\| = \left\|A^{-1} - A^{-1}(\Delta A)B^{-1}\right\| \leq \left\|A^{-1}\right\| + \left\|A^{-1}\Delta A\right\| \left\|B^{-1}\right\|,$$

which implies that

$$\left\|B^{-1}\right\| \leq \frac{\left\|A^{-1}\right\|}{1 - \|A^{-1}\Delta A\|}.$$

It follows that

$$\frac{\left\|A^{-1} - B^{-1}\right\|}{\|A^{-1}\|} \leq \frac{\left\|A^{-1}\Delta A\right\|}{1 - \|A^{-1}\Delta A\|}.$$

Since $\left\|A^{-1}\Delta A\right\| \leq \left\|A^{-1}\right\| \|\Delta A\| = \kappa(A)\frac{\|\Delta A\|}{\|A\|}$, this proves the claim. □

## 5.2 Sparse signal recovery

Consider an $m \times n$ linear system $Ax = b$. If $\operatorname{rank} A < n$ (in particular, if $m > n$) and a solution $x$ exists, then there will be infinitely many solutions. We saw in section 2.2 that singular value decomposition, via the Moore–Penrose pseudoinverse, gives a way to pick out one particular solution: the least squares solution, which is the solution to the optimization problem:

$$\text{Minimize } \|x\|_2 \text{ among all } x \in \mathbb{R}^n \text{ such that } Ax = b. \tag{7}$$

For different purposes, we might wish to pick out a different solution. In many signal processing applications, it is particularly useful to find *sparse* solutions, that is, solutions such that $x_j = 0$ for most $j$. In particular, we would often like to solve the optimation problem:

$$\text{Minimize } \#\{j \mid x_j \neq 0\} \text{ among all } x \in \mathbb{R}^n \text{ such that } Ax = b. \tag{8}$$

Unfortunately, solving (8) is much more difficult computationally; it essentially requires searching through exponentially many subspaces, making it in general a computationally intractable problem for large matrices.

One way to attack this problem is to replace the combinatorial quantity $\#\{j \mid x_j \neq 0\}$ with something that has nicer analytic properties. For example, if $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, then there are good computational tools for solving optimation problems of the form

$$\text{Minimize } f(x) \text{ among all } x \in \mathbb{R}^n \text{ such that } Ax = b.$$

Notice in particular that (7) is of this form, using $f(x) = \|x\|_2$. Superficially it might appear that (7) could be a reasonable substitute for (8), since, after all, $\|\cdot\|_2$ is a monotone norm,

so less sparse matrices have larger norm. However, this fact doesn't interact well with the restriction $Ax = b$.

Suppose for example that the solution space of $Ax = b$ consists of the line $x + 2y = 1$. The two sparsest solutions are $(1/2, 0)$ and $(0, 1)$, but the solution with least $\ell^2$ norm is $(1/5, 2/5)$. In fact, in this two-dimensional setting the solution with least $\ell^2$ norm will never be a sparsest solution unless the solution space is either a horizontal or vertical line.

A next thought might be to replace $\|\cdot\|_2$ with a different norm that behaves more like $\#\{j \mid x_j \neq 0\}$. The most obvious candidate is the $\ell^1$ norm, so we consider the optimization problem

$$\text{Minimize } \|x\|_1 \text{ among all } x \in \mathbb{R}^n \text{ such that } Ax = b. \tag{9}$$

In the example above, for instance, $(1/2, 0)$ minimizes the $\ell^1$ norm on the line $x+2y = 1$. In fact, on any line in $\mathbb{R}^2$, the $\ell^1$ norm is minimized by a point lying on one of the coordinate axes, and uniquely minimized by such a point unless the line has slope $\pm 1$. (Note that those situations are as exceptional as the situations in which the $\ell^2$ norm is minimized by a point on one of the coordinate axes.)

It turns out that (9) is indeed a useful substitute for (8). To state a precise result along these lines we will need a little more terminology.

First, for $x \in \mathbb{R}^n$, we define $\|x\|_0 = \#\{j \mid x_j \neq 0\}$. Note that $\|\cdot\|_0$ is *not* a norm. Next, we say that $A \in M_{m,n}$ satisfies the **restricted isometry property (RIP)** with parameters $\alpha, \beta > 0$ and $s$ if

$$\alpha \|x\|_2 \leq \|Ax\|_2 \leq \beta \|x\|_2$$

whenever $\|x\|_0 \leq s$. That is, $A$ approximately preserves the $\ell^2$ norm, up to a scalar multiple, when we restrict it to acting on sufficiently sparse vectors.

**Theorem 5.2.** *Suppose that $A \in M_{m,n}$ satisfies the RIP with parameters $\alpha$, $\beta$, and $(1+\lambda)s$, with $\lambda > (\beta/\alpha)^2$. Then, whenever $\|x\|_0 \leq s$ and $Ax = b$, $x$ is the solution of the optimization problem* (9).

Theorem 5.2 implies that the (computationally tractable) convex optimization problem (9) will find the solution of the hard combinatorial optimization problem (8), as long as the matrix $A$ satisfies the RIP with suitable parameters, relative to the sparsity of the solution of (8).

The difficulty now is how to tell whether $A$ satisfies the RIP. Unfortunately, for a given matrix $A$ this is not much easier than solving (8) directly. On the other hand, it is known that many natural ways of generating a large matrix *randomly* have a very high probability of producing a matrix that satisfies the RIP. Here we will not deal with these issues, and merely proof Theorem 5.2.

*Proof of Theorem 5.2.* Assume that $\|x\|_0 \leq s$ and that $Ax = b$. Let $\widehat{x}$ be a solution of the optimization problem (9). We wish to show that $\widehat{x} = x$. Let $h = \widehat{x} - x$, so that we need to show $h = 0$. Note that

$$Ah = A\widehat{x} - Ax = b - b = 0.$$

We start by decomposing the the set of indices, and decomposing $h$ in a corresponding way. For a subset $I \subseteq \{1, \ldots, n\}$ and $y \in \mathbb{R}^n$, we write $y_I \in \mathbb{R}^n$ for the vector with

components

$$(y_I)_j = \begin{cases} y_j & \text{if } j \in I, \\ 0 & \text{if } j \notin I. \end{cases}.$$

Let

$$I_0 = \{j \mid x_j \neq 0\},$$

so that $\#I_0 = \|x\|_0 \leq s$. Let $I_1$ denote the indices for the $\lambda s$ largest (in absolute value) entries of $h_{I_0^c}$, let $I_2$ denote the indices for the $\lambda s$ next largest entries, and so on; say $I_t$ is the last of these (which may have fewer than $\lambda s$ elements). (If there are entries of equal magnitude, we can order them arbitrarily.) We also write $I_{0,1} = I_0 \cup I_1$.

We first observe that $\|\widehat{x}\|_1 \geq \|x\|_1$ by definition of $\widehat{x}$. On the other hand,

$$\|\widehat{x}\|_1 = \|x + h\|_1 = \|x + h_{I_0}\|_1 + \left\|h_{I_0^c}\right\|_1 \geq \|x\|_1 - \|h_{I_0}\|_1 + \left\|h_{I_0}^c\right\|_1$$

by the triangle inequality. Therefore

$$\left\|h_{I_0}^c\right\|_1 \leq \|h_{I_0}\|_1.$$

Next, by the triangle inequality,

$$0 = \|Ah\|_2 = \left\|A(h_{I_{0,1}} + h_{I_{0,1}^c})\right\|_2 \geq \left\|Ah_{I_{0,1}}\right\|_2 - \left\|Ah_{I_{0,1}^c}\right\|_2.$$

Now $\left\|h_{I_{0,1}}\right\|_0 \leq \#I_0 + \#I_1 \leq (1+\lambda)s$, $\|h_{I_k}\|_0 \leq \lambda s$ for $k \geq 2$, and $h_{I_{0,1}^c} = \sum_{k=2}^t h_{I_k}$. Therefore by the RIP hypothesis and the triangle inequality

$$\alpha \left\|h_{I_{0,1}}\right\|_2 \leq \left\|Ah_{I_{0,1}}\right\|_2 \leq \left\|Ah_{I_{0,1}^c}\right\|_2 \leq \sum_{k=2}^t \|Ah_{I_k}\|_2 \leq \beta \sum_{k=2}^t \|h_{I_k}\|_2.$$

To further bound the right hand side of this, note that by the definition of $I_k$, for $k \geq 2$ we have

$$\|h_{I_k}\|_\infty \leq \min_{j \in I_{k-1}} |h_j| \leq \frac{1}{\lambda s} \sum_{j \in I_{k-1}} |h_j| = \frac{1}{\lambda s} \left\|h_{I_{k-1}}\right\|_1.$$

It follows that $\|h_{I_k}\|_2 \leq \sqrt{\lambda s}\, \|h_{I_k}\|_\infty \leq \frac{1}{\sqrt{\lambda s}} \|h_{I_k}\|_1$, and therefore

$$\sum_{k=2}^t \|h_{I_k}\|_2 \leq \frac{1}{\sqrt{\lambda s}} \sum_{k=2}^t \left\|h_{I_{k-1}}\right\|_1 \leq \frac{1}{\sqrt{\lambda s}} \sum_{k=1}^t \|h_{I_k}\|_1 = \frac{1}{\sqrt{\lambda s}} \left\|h_{I_0^c}\right\|_1 \leq \frac{1}{\sqrt{\lambda s}} \|h_{I_0}\|_1.$$

We now have that

$$\alpha \left\|h_{I_{0,1}}\right\|_2 \leq \frac{\beta}{\sqrt{\lambda s}} \|h_{I_0}\|_1 \leq \frac{\beta}{\sqrt{\lambda}} \|h_{I_0}\|_2 \leq \frac{\beta}{\sqrt{\lambda}} \left\|h_{I_{0,1}}\right\|_2.$$

Since $\lambda > (\beta/\alpha)^2$, this implies that $\left\|h_{I_{0,1}}\right\|_2 = 0$, so that $h_{I_{0,1}} = 0$, and therefore $h = 0$. $\square$

# 6 Positive (semi)definite matrices

## 6.1 Characterizations

Recall the notion of a positive semidefinite matrix, which can be characterized in several equivalent ways (Theorem 2.9). As you saw in homework, positive definite matrices have several analogous characterizations, including as nonsingular positive semidefinite matrices.

One immediate consequence of the characterizations in terms of inner products $\langle Ax, x\rangle$ is the following fact, which we will often use without comment.

**Lemma 6.1.** *If $A \in M_n$ is positive (semi)definite, then every principal submatrix of $A$ is positive (semi)definite.*

Another characterization of positive definite matrices is a special case of yet another homework problem.

**Theorem 6.2** (Sylvester's criterion)**.** *A Hermitian matrix $A$ is positive definite if and only if the determinant of each upper-left principal submatrix of $A$ is positive.*

*Proof.* As you showed in homework, the determinant criterion implies that all the eigenvalues of $A$ are positive, which implies that $A$ is positive semidefinite.

Conversely, if $A$ is positive definite then each of its principal submatrices is positive definite, and hence determinant (the product of the eigenvalues) of each of those submatrices is positive. □

**Theorem 6.3.** *If $A \in M_n(\mathbb{F})$ is positive (semi)definite, then $A$ has a unique positive (semi)definite $k^{\text{th}}$ root (that is, a matrix $B$ such that $B^k = A$) for each $k \in \mathbb{N}$.*

*Proof.* Existence is proved just as in Theorem 2.9, which included the case $k = 2$: let $A = U \operatorname{diag}(\lambda_1, \ldots, \lambda_n)U^*$ be a spectral decomposition, and define $B = U \operatorname{diag}(\lambda_1^{1/k}, \ldots, \lambda_n^{1/k})U^*$.

For uniqueness, given any distinct $x_1, \ldots, x_n$ and $y_1, \ldots, y_n \in \mathbb{R}$, there exists a polynomial $p(x)$ such that $p(x_j) = y_j$ for each $j$ (for example, by the Lagrange interpolation formula). Therefore, if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $A$, there is a polynomial $p$ such that $p(\lambda_j) = \lambda_j^{1/k}$ for each $j$. It follows that $p(A) = B$ for $B$ defined as above.

Now suppose that $C$ is positive semidefinite and $C^k = A$. Then $B = p(A) = p(C^k)$. This implies that $B$ and $C$ commute. Since $B$ and $C$ are both Hermitian and therefore diagonalizable, Theorem 2.21 implies that they are simultaneously diagonalizable:

$$B = SD_1S^{-1} \qquad \text{and} \qquad C = SD_2S^{-1}$$

for some nonsingular $S \in M_n(\mathbb{F})$ and diagonal $D_1, D_2 \in M_n(\mathbb{F})$. Then

$$SD_1^kS^{-1} = B^k = A = C^k = SD_2^kS^{-1},$$

which implies that $D_1^k = D_2^k$. The diagonal entries of $D_1$ and $D_2$ are the eigenvalues of $B$ and $C$, and therefore nonnegative, so this implies that $D_1 = D_2$, and therefore $B = C$. □

For an arbitrary $A \in M_n$, the **absolute value** of $A$ is the unique positive semidefinite square root $|A|$ of the positive semidefinite matrix $A^*A$. (We could alternatively define it using $AA^*$; this is not the same matrix when $A$ is non-normal, but the difference is only a matter of convention.) We can describe $|A|$ explicitly in terms of a singular value decomposition $A = U\Sigma V^*$ as $A = V\Sigma V^*$. Thus the singular values of $A$ are the same as the eigenvalues of $|A|$. This implies that we can write the Schatten 1-norm (or Ky Fan $(n)$-norm) of $A$ as

$$\|A\|_1 = \operatorname{tr} |A| .$$

For this reason, the Schatten 1-norm is sometimes called the **trace norm**.

We have also encountered the matrix absolute value before in the polar decomposition (Theorem 2.12), which we can now be state by saying that $A = U|A|$ for some unitary matrix $U$.

**Proposition 6.4.** *Suppose that $B \in M_{m,n}$. Then:*

1. $\ker B^*B = \ker B$.
2. $\operatorname{rank} B^*B = \operatorname{rank} B$.
3. *$B^*B$ is positive definite if and only if* $\operatorname{rank} B = n$.

*Proof.* If $Bx = 0$ then clearly $B^*Bx = 0$. If $B^*Bx = 0$, then

$$0 = \langle B^*Bx, x\rangle = \langle Bx, Bx\rangle ,$$

and therefore $Bx = 0$. This proves the first statement. The second statement follows from the rank–nullity theorem, and the third statement follows from the second statement and the fact that a positive semidefinite matrix is positive definite if and only if it is nonsingular. □

**Proposition 6.5** (Cholesky factorization)**.** *A Hermitian matrix $A \in M_n(\mathbb{F})$ is positive semidefinite if and only if there exists a lower triangular matrix $L \in M_n(\mathbb{F})$ such that $A = LL^*$.*

*Proof.* If $A = LL^*$ then $A$ is positive semidefinite. Conversely, if $A$ is positive semidefinite, then $A = B^*B$ for some $B \in M_n$. Let $B = QR$ be a QR decomposition. Then

$$A = B^*B = R^*Q^*QR = R^*R.$$

Thus we can let $L = R^*$. □

Suppose that $V$ is an inner product space. The **Gram matrix** of a list of vectors $v_1, \ldots, v_n$ is the matrix $A \in M_n$ with entries $a_{jk} = \langle v_k, v_j\rangle$.

**Theorem 6.6.** 1. *A matrix $A \in M_n$ is positive semidefinite if and only if it the the Gram matrix of some list of vectors in some inner product space.*

2. *A matrix $A \in M_n$ is positive definite if and only if it the the Gram matrix of some linearly independent list of vectors in some inner product space.*

*Proof.* Suppose that $A$ is the Gram matrix of $v_1, \ldots, v_n$. Given $x \in \mathbb{F}^n$,

$$\langle Ax, x\rangle = \sum_{j,k=1}^{n} \langle v_k, v_j\rangle\, x_k\overline{x_j} = \sum_{j,k=1}^{n} \langle x_k v_k, x_j v_j\rangle = \left\|\sum_{j=1}^{n} x_j v_j\right\|^2 \geq 0,$$

and thus $A$ is positive semidefinite. Furthermore, this shows that $\langle Ax, x\rangle = 0$ if and only if $\sum_{j=1}^{n} x_j v_j = 0$. If $v_1, \ldots, v_n$ is a linearly independent list, then this is the case if and only if $x = 0$.

Now suppose that $A$ is positive semidefinite. Then $A = B^*B$ for some $B \in M_n$. This implies that $a_{jk} = b_j^* b_k = \langle b_k, b_j\rangle$, where $b_j$ are the columns of $B$. Therefore $A$ is the Gram matrix of $b_1, \ldots, b_n \in \mathbb{F}^n$. If $A$ is positive definite then $\operatorname{rank} B = n$, and so $b_1, \ldots, b_n$ are linearly independent.

□

**Corollary 6.7.** *A list of vectors $v_1, \ldots, v_n$ in an inner product space is linearly independent if and only if their Gram matrix is nonsingular.*

## 6.2 Kronecker and Hadamard products

If $A \in M_{m_1,n_1}$ and $B \in M_{m_2,n_2}$, the **Kronecker product** $A \otimes B \in M_{m_1m_2,n_1n_2}$ is defined by the block presentation

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n_1}B \\ \vdots & \ddots & \vdots \\ a_{m_11}B & \cdots & a_{m_1n_1}B \end{bmatrix}.$$

Straightforward computations show that

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$$

when all of these products are defined, that $(A \otimes B)^* = A^* \otimes B^*$. If follows that if $U$ and $V$ are unitary matrices (not necessarily of the same size) then $U \otimes V$ is also unitary, and if $A$ and $B$ are Hermitian then so is $A \otimes B$.

**Proposition 6.8.** *1. Suppose that $A \in M_m$ has eigenvalues $\lambda_1, \ldots, \lambda_m$ and $B \in M_n$ has eigenvalues $\mu_1, \ldots, \mu_n$ (with multiplicity). Then $A \otimes B$ has eigenvalues $\lambda_j\mu_k$ for $1 \leq j \leq m$ and $1 \leq k \leq n$.*

*2. Suppose that $A \in M_{m_1,n_1}$ has singular values $\sigma_1, \ldots, \sigma_{p_1}$ and $B \in M_{m_2,n_2}$ has singular values $\tau_1, \ldots, \tau_{p_2}$, where $p_i = \min\{m_i, n_i\}$. Then $A \otimes B$ has singular values $\sigma_j\tau_k$ for $1 \leq j \leq p_1$ and $1 \leq k \leq p_2$, possibly with additional 0s.*

*Proof.* 1. Let $A = UT_1U^*$ and $B = VT_2V^*$ be Schur decompositions. Then

$$A \otimes B = (U \otimes V)(T_1 \otimes T_2)(U^* \otimes V^*) = (U \otimes V)(T_1 \otimes T_2)(U \otimes V)^*.$$

Since $U \otimes V$ is unitary, the eigenvalues of $A \otimes B$ are the eigenvalues of the triangular matrix $T_1 \otimes T_2$, which are precisely as given in the statement of the proposition.

2. Let $A = U_1\Sigma_1 V_1^*$ and $B = U_2\Sigma_2 V_2^*$ be singular value decompositions. Then

$$A \otimes B = (U_1 \otimes U_2)(\Sigma_1 \otimes \Sigma_2)(V_1 \otimes V_2)^*$$

is a singular value decomposition, in which the singular values are precisely as given in the statement of the proposition.

□

**Corollary 6.9.** *If $A \in M_m$ and $B \in M_n$ are both positive (semi)definite, then so is $A \otimes B$.*

If $A, B \in M_{m,n}$, the **Hadamard product** $A \circ B \in M_{m,n}$ is defined to have entries $[A \circ B]_{jk} = a_{jk}b_{jk}$. Note that $A \circ B$ is a submatrix of $A \otimes B$, and is a principal submatrix when $A, B \in M_n$. With this observation in mind, the next result follows immediately from Corollary 6.9.

**Corollary 6.10** (Schur product theorem)**.** *If $A, B \in M_n$ are both positive (semi)definite, then so is $A \circ B$.*

## 6.3 Inequalities for positive (semi)definite matrices

If $A, B \in M_n$ are Hermitian, we write $A \preceq B$ if $B - A$ is positive semidefinite, and $A \prec B$ if $B - A$ is positive definite.

**Proposition 6.11.** *The relation $\preceq$ is a partial order on the set of $n \times n$ Hermitian matrices. That is:*

1. *$A \prec A$ for every $A$.*
2. *If $A \prec B$ and $B \prec C$, then $A \prec C$.*
3. *If $A \prec B$ and $B \prec A$, then $A = B$.*

The proof of Proposition 6.11 is left as an exercise.

The following result follows immediately from the Weyl monotonicity theorem (Corollary 3.8).

**Proposition 6.12.** *If $A \preceq B$ then $\lambda_j^{\downarrow}(A) \leq \lambda_j^{\downarrow}(B)$ for every $1 \leq j \leq n$.*

The following corollary is immediate.

**Corollary 6.13.** *If $A \preceq B$ then $\operatorname{tr} A \leq \operatorname{tr} B$. If $0 \preceq A \preceq B$ then $\det A \leq \det B$.*

Recall the following result from the January 28 homework, sometimes known as Hadamard's inequality: if $A \in M_n$ has columns $a_1, \ldots, a_n$, then $|\det A| \leq \prod_{j=1}^n \|a_j\|_2$. This has the following consequence, which often goes by the same name.

**Theorem 6.14** (Hadamard's inequality for positive semidefinite matrices)**.** *If $A \in M_n$ is positive semidefinite, then $\det A \leq \prod_{j=1}^n a_{jj}$.*

*Proof.* Write $A = B^2$ for a Hermitian matrix $B$ with columns $b_j = Be_j$. Then

$$a_{jj} = \langle Ae_j, e_j \rangle = \langle B^2 e_j, e_j \rangle = \langle Be_j, Be_j \rangle = \|b_j\|_2^2 .$$

Therefore, by the earlier Hadamard's inequality applied to $B$,

$$\det A = (\det B)^2 \leq \prod_{j=1}^{n} \|b_j\|_2^2 = \prod_{j=1}^{n} a_{jj}. \qquad \square$$

**Theorem 6.15** (Fischer's inequality)**.** *Suppose that* $H = \begin{bmatrix} A & B \\ B^* & C \end{bmatrix}$ *is positive semidefinite with* $A$ *and* $C$ *square.*

*Proof.* Let $A = U\Lambda U^*$ and $C = V\Gamma V^*$ be spectral decompositions, and define $W = U \oplus V$. Then

$$W^*HW = \begin{bmatrix} \Lambda & U^*BV \\ V^*BU & \Gamma \end{bmatrix}$$

By Hadamard's inequality (Theorem 6.14),

$$\det H = \det(W^*HW) \leq \prod_{j=1}^{n} \lambda_j \gamma_j = (\det \Lambda)(\det \Gamma) = (\det A)(\det C). \qquad \square$$

**Theorem 6.16.** *If* $A \in M_n$ *is positive definite, then*

$$(\det A)^{1/n} = \min \left\{ \frac{1}{n} \operatorname{tr}(AB) \;\middle|\; B \in M_n \text{ is positive definite with } \det B = 1 \right\}.$$

*Proof.* Let $A = U\Lambda U^*$ be a spectral decomposition. Then $\det A = \det \Lambda$ and $\operatorname{tr}(AB) = \operatorname{tr}(\Lambda U^* B U)$. It therefore suffices to assume that $A = \Lambda$. By the arithmetic geometric mean inequality (Lemma 4.1) and Hadamard's inequality (Theorem 6.14),

$$\frac{1}{n} \operatorname{tr}(\Lambda B) = \frac{1}{n} \sum_{j=1}^{n} \lambda_j b_{jj} \geq \left( \prod_{j=1}^{n} \lambda_j \right)^{1/n} \left( \prod_{j=1}^{n} b_{jj} \right)^{1/n} \geq (\det A)^{1/n} (\det B)^{1/n}$$

for any positive definite $B$. In particular, if $\det B = 1$ then

$$(\det A)^{1/n} \leq \frac{1}{n} \operatorname{tr}(AB).$$

Moreover, we have equality here if $B = (\det A)^{1/n} A^{-1}$. $\square$

**Corollary 6.17** (Minkowski's determinant inequality)**.** *If* $A, B \in M_n$ *are positive definite then*

$$\big[\det(A+B)\big]^{1/n} \geq (\det A)^{1/n} + (\det B)^{1/n}.$$

*Proof.* By Theorem 6.16,

$$\begin{aligned}\left[\det(A+B)\right]^{1/n} &= \min\left\{\frac{1}{n}\operatorname{tr}(A+B)C \;\middle|\; C \succ 0,\ \det C = 1\right\} \\ &\geq \min\left\{\frac{1}{n}\operatorname{tr}(AC) \;\middle|\; C \succ 0,\ \det C = 1\right\} \\ &\qquad + \min\left\{\frac{1}{n}\operatorname{tr}(BC) \;\middle|\; C \succ 0,\ \det C = 1\right\} \\ &= (\det A)^{1/n} + (\det B)^{1/n}.\end{aligned}$$

□

# 7 Locations and perturbations of eigenvalues

## 7.1 The Geršgorin circle theorem

**Theorem 7.1** (Geršgorin's theorem)**.** *Let $A \in M_n$, and define $R_j(A) = \sum_{k\neq j} |a_{jk}|$ and*

$$D_j(A) = \{z \in \mathbb{C} \mid |z - a_{jj}| \leq R_j(A)\}.$$

*Then each eigenvalue of $A$ lies in at least one $D_j(A)$.*

The sets $D_j(A)$ are sometimes called the *Geršgorin discs* of $A$. Note that Theorem 7.1 does *not* say that each disc $D_j(A)$ contains an eigenvalue.

*Proof.* Suppose that $Ax = \lambda x$ for $x \neq 0$, and pick an index $p$ such that $|x_p| = \|x\|_\infty$. Considering the $p^{\text{th}}$ entry of $(Ax - \lambda x) = 0$, we have

$$(\lambda - a_{pp})x_p = \sum_{k\neq p} a_{pk}x_k,$$

and therefore

$$|\lambda - a_{pp}| \leq R_p(A)\,\|x\|_\infty = R_p(A)\,|x|_p. \qquad \square$$

Recall that $A$ is called **strictly diagonally dominant** if $|a_{jj}| > R_j(A)$ for each $j$. Geršgorin's theorem gives us a new proof of the Levy–Desplanques theorem (Corollary 4.24), which states that a strictly diagonally dominant matrix is nonsingular:

*Second proof of Corollary 4.24.* If $A$ is strictly diagonally dominant, then for each $j$, $0 \notin D_j(A)$. By Theorem 7.1, this implies that 0 is not an eigenvalue of $A$ and so $A$ is nonsingular. □

Conversely, the Levy–Desplanques theorem implies Geršgorin's theorem. Since we already have an independent proof of the Levy–Desplanques theorem, this gives a second approach to proving Geršgorin's theorem.

This is one manifestation of a general phenomenon: any result about locations of eigenvalues gives a sufficient condition for invertibility (whatever condition forces 0 not to be an

eigenvalue). And conversely, any sufficient condition for invertibility implies a result about locations of eigenvalues (since $\sigma(A) = \{z \in \mathbb{C} \mid A - zI_n \text{ is singular}\}$).[4]

Geršgorin's theorem can be generalized and extended in many ways. A first obvious observation is that columns can be used just as well as sums: if $C_k(A) = \sum_{j \neq k} |a_{jk}|$, then each eigenvalue of $A$ lies in some disc

$$\{z \in \mathbb{C} \mid |z - a_{kk}| \leq C_k(A)\}.$$

This can be proved analogously, or deduced directly from Geršgorin's theorem since $\sigma(A^T) = \sigma(A)$. More subtle analogues exist which consider both the rows and sums of $A$ at the same time.

Another easy way to extend Geršgorin's theorem is to note that $\sigma(S^{-1}AS) = \sigma(A)$ for any nonsingular $S$, and apply Geršgorin's theorem to $S^{-1}AS$. If we let $S = \operatorname{diag}(d_1, \ldots, d_n)$ with $d_j > 0$ for each $j$, we get the following.

**Corollary 7.2.** *Suppose that $d_1, \ldots, d_n > 0$ and $A \in M_n$. Then each eigenvalue of $A$ lies in one of the discs*

$$\left\{ z \in \mathbb{C} \;\middle|\; |z - a_{jj}| \leq \frac{1}{d_j} \sum_{k \neq j} d_k |a_{jk}| \right\}$$

*for $j = 1, \ldots, n$.*

## 7.2 Eigenvalue perturbations for non-Hermitian matrices

Recall Mirsky's inequality (Theorem 4.38), which states that

$$\left\| \lambda^{\downarrow}(A) - \lambda^{\downarrow}(B) \right\| \leq \|A - B\|$$

for any corresponding pair of a symmetric gauge function and unitarily invariant norm, and any Hermitian matrices $A, B \in M_n$. This result refines, in precise, quantitative form, the fact that eigenvalues depend continuously on the matrix (Corollary 2.20) — but only for Hermitian matrices.

For general matrices, the first obstacle is that since the eigenvalues need not be real, and it's not clear how to quantify how similar two sets of complex numbers are to each other absent a natural ordering. In fact there are many ways to do this, and different eigenvalue perturbation theorems involve different ones.

We will first prove two results for normal matrices. Despite the fact that eigenvalues of normal matrices can be any complex numbers, eigenvalues of normal matrices are still better behaved as functions of the matrix than in the completely general case.

Given two closed, bounded sets $X, Y \subseteq \mathbb{C}$, the **Hausdorff distance** between them is defined to be

$$d_H(X, Y) = \max \left\{ \max_{x \in X} \min_{y \in Y} |x - y|, \max_{y \in Y} \min_{x \in X} |x - y| \right\}.$$

[4] A very careful reader might note that our first proof of the Levy–Desplanques theorem was based on applying Corollary 4.23 with the maximum row sum norm, and that Corollary 4.23 itself was proved by this strategy: deduce invertibility of $I_n - A$ by considering where the eigenvalues of $A$ are. So this second suggested proof of the Levy–Desplanques theorem uses all the same ideas as the first one.

That is, $d_H(X,Y)$ is the farthest that a point from one of the two sets can be from the other set.

**Theorem 7.3** (Bauer–Fike theorem)**.** *Suppose that $A, B \in M_n$ are normal. Then*

$$d_H(\sigma(A), \sigma(B)) \leq \|A - B\|_{2\to 2}\,.$$

*Proof.* We will prove a stronger fact: if $A \in M_n$ is normal and $B \in M_n$ is arbitrary, then

$$\max_{\mu\in\sigma(B)} \min_{\lambda\in\sigma(A)} |\lambda - \mu| \leq \|A - B\|_{2\to 2}\,.$$

Suppose that $Bx = \mu x$ with $\|x\|_2 = 1$, and let $A = U\Lambda U^*$ be a spectral decomposition. Then, with the substitution $y = U^*x$,

$$\begin{aligned}\|A - B\|_{2\to 2} &\geq \|(A-B)x\|_2 = \|Ax - \mu x\|_2 = \|U(\Lambda - \mu I_n)U^*x\|_2 \\ &= \|(\Lambda - \mu I_n)y\|_2 = \sqrt{\sum_{j=1}^n |\lambda_j - \mu|^2 |y_j|^2} \geq \min_{1\leq j\leq n} |\lambda_j - \mu|\,,\end{aligned}$$

since $\sum_{j=1}^n |y_j|^2 = \|y\|_2^2 = 1$. □

The Hausdorff distance measures a kind of "worst case scenario" for comparing two sets of complex numbers. The following result, which generalizes Corollary 3.22, considers a kind of average comparison.

**Theorem 7.4** (Hoffman–Wielandt inequality for normal matrices)**.** *Suppose that $A, B \in M_n$ are both normal. The eigenvalues $\{\lambda_j\}$ of $A$ and $\{\mu_j\}$ of $B$ can be ordered so that*

$$\sqrt{\sum_{j=1}^n |\lambda_j - \mu_j|^2} \leq \|A - B\|_F\,.$$

To prove this, we need to deal with some leftovers from before. Suppose that $V$ is a finite-dimensional real vector space and that $K \subseteq V$ is a closed convex set. A point $x \in K$ is called an **extreme point** of $K$ if, whenever $x = ty + (1-t)z$ for $y, z \in K$ and $0 < t < 1$, we must have $x = y = z$.

**Proposition 7.5.** *Let $K_n \subseteq M_n$ be the set of $n \times n$ doubly stochastic matrices. If $A \in K_n$ is an extreme point of $K_n$, then $A$ is a permutation matrix.*

*Proof.* We will show that if $A \in K_n$ is not a permutation matrix, then $A$ is not an extreme point of $K_n$.

If $A$ is not a permutation matrix, it has some row with two positive entries; choose one such entry $a_{i_1,j_1} \in (0,1)$. Then there is an entry $a_{i_2,j_1} \in (0,1)$ for some $i_2 \neq i_1$, and then some $a_{i_2,j_2} \in (0,1)$ for some $j_2 \neq j_1$. We continue picking entries in $(0,1)$ in this way until the first time an entry $a_{ij}$ is picked twice.

Let $a$ be the value of the smallest entry picked from the first to the second time $a_{ij}$ occurs. Define $B \in M_n$ to have 1 in the position of the first entry in this sequence, $-1$ in the position of the second entry, and so on, with all other entries 0. Then the row and column sums of $B$ are all 0. We then have that $A_+ = A + aB$ and $A_- = A - aB$ are nonnegative with row and column sums all equal to 1, so $A_\pm \in K_n$, and $A = \frac{1}{2}A_+ + \frac{1}{2}A_-$. Therefore $A$ is not an extreme point of $K_n$. □

A basic fact from convexity is that a closed, bounded, convex set is the convex hull of its extreme points; thus Proposition 7.5 implies Birkhoff's theorem (Theorem 3.21).

*Proof of Theorem 7.4.* Let $A = U\Lambda U^*$ and $B = V\Gamma V^*$ be spectral decompositions, and let $W = U^*V$. Then

$$\|A - B\|_F^2 = \|\Lambda W - W\Gamma\|_F^2 = \sum_{j,k=1}^{n} |\lambda_j - \gamma_k|^2 \, |w_{jk}|^2 .$$

The matrix $\left[|w_{jk}|^2\right]$ is unitary stochastic, hence doubly stochastic. So by Birkhoff's theorem, it can be written as $\sum_{i=1}^{N} t_i P_i$ for permutation matrices $P_1, \ldots, P_N$ and $t_1, \ldots, t_N \geq 0$ with $\sum_{i=1}^{N} t_i = 1$. Writing $p_{jk}^i$ for the entries of $P_i$, this implies

$$\|A - B\|_F^2 = \sum_{i=1}^{N} t_i \sum_{j,k=1}^{n} |\lambda_j - \gamma_k|^2 \, p_{jk}^i \geq \min_{1 \leq i \leq N} \sum_{j,k=1}^{n} |\lambda_j - \gamma_k|^2 \, p_{jk}^i .$$

Let $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be the permutation corresponding to $P_i$, so that $p_{jk}^i = 1$ if $\pi(j) = k$, and other entries are 0. Then we have

$$\|A - B\|_F^2 \geq \sum_{j=1}^{n} \left|\lambda_j - \gamma_{\pi(j)}\right|^2 ,$$

so we can let $\mu_j = \gamma_{\pi(j)}$. □

For non-normal matrices, the dependence of eigenvalues on the matrix can be much more irregular. Consider the matrix

$$A_\varepsilon = \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & 0 & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 1 & 0 \\ 0 & & & & 0 & 1 \\ \varepsilon & 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix} \in M_n$$

for $\varepsilon \geq 0$. Since $A_0$ is triangular, we can tell immediately that its only eigenvalue is 0. It can be shown (in homework!) that the eigenvalues of $A_\varepsilon$ all have modulus $\varepsilon^{1/n}$. Thus $d_H(\sigma(A_\varepsilon), \sigma(A_0)) = \varepsilon^{1/n} = \|A_\varepsilon - A_0\|_{2\to 2}^{1/n}$. The same holds for the distances in the Hoffman–Wielandt inequality.

The following theorem shows roughly that the example above is the worst things can get.

**Theorem 7.6.** *If $A, B \in M_n$, then*

$$d_H\big(\sigma(A), \sigma(B)\big) \leq \left(\|A\|_{2\to 2} + \|B\|_{2\to 2}\right)^{1 - \frac{1}{n}} \|A - B\|_{2\to 2}^{1/n} .$$

*Proof.* Let $\mu$ be an eigenvalue of $B$ and let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $A$. Then

$$\min_{1 \le j \le n} |\lambda_j - \mu| \le \prod_{j=1}^{n} |\lambda_j - \mu|^{1/n} = |\det(A - \mu I_n)|^{1/n} .$$

Let $v_1, \ldots, v_n$ be an orthonormal basis of $\mathbb{C}^n$ such that $Bv_1 = \mu v_1$, and let $V$ be the unitary matrix with those columns. Then

$$|\det(A - \mu I_n)| = |\det(A - \mu I_n)V| \le \prod_{j=1}^{n} \|(A - \mu I_n)v_j\|_2$$

by Hadamard's inequality. Now

$$\|(A - \mu I_n)v_1\|_2 = \|Av_1 - \mu v_1\|_2 = \|(A - B)v_1\|_2 \le \|A - B\|_{2\to 2} ,$$

and for $j \ge 2$,

$$\|(A - \mu I_n)v_j\|_2 \le \|Av_j\|_2 + \|\mu v_j\|_2 \le \|A\|_{2\to 2} + \|B\|_{2\to 2} .$$

Combining the above estimates proves the claim. □

# 8 Nonnegative matrices

## 8.1 Inequalities for the spectral radius

We now turn to another special class of matrices: **nonnegative** matrices, by which we mean matrices which have only nonnegative real entries, and the subclass of **positive** matrices, those with only positive entries. It is important to be careful of the distinction between positive matrices and positive (semi)definite matrices (likewise between nonnegative matrices and nonnegative definite matrices, another term for positive semidefinite) — especially since some authors use the term positive matrix to mean a positive (semi)definite matrix.

Nevertheless, we will see that many of the results for positive (semi)definite matrices, or Hermitian matrices more generally, have analogues for nonnegative or positive matrices. However, the methods tend to be quite different. In particular, in working with Hermitian matrices we tend (with some notable exceptions) either to avoid thinking about the individual matrix entries, or only think about them after invoking the spectral theorem in order to reduce attention to diagonal entries. On the other hand, when working with nonnegative matrices we work with matrix entries quite a lot — unsurprisingly, given that the assumption of nonnegativity is entirely about matrix entries.

One major difference is that whereas Hermitian matrices have real eigenvalues, the eigenvalues of a nonnegative matrix need not be real. Since many of the important results about Hermitian matrices take the form of inequalities involving eigenvalues, this limits how closely results about nonnegative matrices can resemble results for Hermitian matrices. The spectral radius $\rho(A)$, which of course is always a nonnegative real number, turns out to play a central role in the theory of nonnegative matrices, similar to eigenvalues themselves in the theory of Hermitian matrices.

Throughout this section $|A|$ will refer to the entrywise absolute value of a matrix, as opposed to the postive semidefinite absolute value introduced in section 6.1 above.

**Theorem 8.1.** *Suppose that $A \in M_n(\mathbb{C})$ and $B \in M_n(\mathbb{R})$. If $|A| \le B$, then $\rho(A) \le \rho(|A|) \le \rho(B)$.*

*In particular, if $0 \le A \le B$, then $\rho(A) \le \rho(B)$.*

Compare this to the fact that if $A, B \in M_n(\mathbb{C})$ are Hermitian and $A \preceq B$, then $\lambda_j^{\downarrow}(A) \le \lambda_j^{\downarrow}(B)$ for each $j$. In particular, if $0 \preceq A \preceq B$, then $\rho(A) \le \rho(B)$.

*Proof.* For any $m \in \mathbb{N}$ and $i, j$ we have

$$\begin{aligned} |[A^m]_{ij}| &= \left| \sum_{k_1,\dots,k_{m-1}=1}^{n} a_{ik_1} a_{k_1k_2} \cdots a_{k_{m-2}k_{m-1}} a_{k_{m-1}j} \right| \\ &\le \sum_{k_1,\dots,k_{m-1}=1}^{n} |a_{ik_1}|\,|a_{k_1k_2}| \cdots \left|a_{k_{m-2}k_{m-1}}\right| \left|a_{k_{m-1}j}\right| = [|A^m|]_{ij} \\ &\le \sum_{k_1,\dots,k_{m-1}=1}^{n} b_{ik_1} b_{k_1k_2} \cdots b_{k_{m-2}k_{m-1}} b_{k_{m-1}j} = [B^m]_{ij}\,. \end{aligned}$$

Now let $\|\cdot\|$ be any absolute submultiplicative norm on $M_n(\mathbb{C})$ (for example, the Frobenius norm or the maximum column sum norm). It follows that

$$\|A^m\| \le \||A|^m\| \le \|B^m\|\,.$$

The claim now follows from the Gelfand formula (Corollary 4.27) $\rho(A) = \lim_{m\to\infty} \|A^m\|^{1/m}$. □

From now on, unless otherwise specified, matrices are in $M_n(\mathbb{R})$.

**Corollary 8.2.** *If $A \ge 0$ and $B$ is a principal submatrix of $A$, then $\rho(B) \le \rho(A)$. In particular, $a_{jj} \le \rho(A)$ for each $j$.*

Compare this to the fact (a consequence of the Rayleigh–Ritz theorem) that if $A$ is Hermitian and $B$ is a principal submatrix of $A$, then $\rho(B) \le \rho(A)$, and in particular $|a_{jj}| \le \rho(A)$ for each $j$. (From now on we will mostly not comment on these analogies.)

*Proof.* Let $\widetilde{A}$ be the matrix obtained by replacing those entries of $A$ not in $B$ with 0. Then $0 \le \widetilde{A} \le A$, and $\rho(\widetilde{A}) = \rho(B)$. The result now follows from Theorem 8.1. □

We will frequently find it convenient to refer to $e = \sum_{i=1}^{n} e_i = (1, \dots, 1)$.

**Lemma 8.3.** *If $A \ge 0$ and all the rows of $A$ have the same sum $r$, then $\rho(A) = r = \|A\|_{\infty\to\infty}$.*

*If $A \ge 0$ and all the columns of $A$ have the same sum $c$, then $\rho(A) = c = \|A\|_{1\to 1}$.*

*Proof.* By Theorem 4.22, $\rho(A) \le \|A\|_{\infty\to\infty}$ for any matrix. If all the rows of $A \ge 0$ have the same sum $r$, this implies that $\rho(A) \le r$. Moreover, $Ae = re$, which implies that $\rho(A) \ge r$.

The second statement follows by applying the first statement to $A^T$. □

**Theorem 8.4.** *If $A \geq 0$, then*

$$\min_{1\leq i\leq n} \sum_{j=1}^{n} a_{ij} \leq \rho(A) \leq \max_{1\leq i\leq n} \sum_{j=1}^{n} a_{ij}$$

*and*

$$\min_{1\leq j\leq n} \sum_{i=1}^{n} a_{ij} \leq \rho(A) \leq \max_{1\leq j\leq n} \sum_{i=1}^{n} a_{ij}.$$

*Proof.* Let $r = \min_{1\leq i\leq n} \sum_{j=1}^{n} a_{ij}$ and $R = \max_{1\leq i\leq n} \sum_{j=1}^{n} a_{ij}$. Define $B \in M_n(\mathbb{R})$ as follows:

$$\begin{aligned} b_{i1} &= \min\{a_{i1}, r\}, \\ b_{i2} &= \min\{a_{i2}, \max\{r - a_{i1}, 0\}\}, \\ &\vdots \\ b_{in} &= \min\left\{a_{in}, \max\left\{r - \sum_{j=1}^{n-1} a_{ij}, 0\right\}\right\}. \end{aligned}$$

That is, in each row, the entries of $B$ match those of $A$ until the row sum of $A$ exceeds $r$ (if ever); at that point $B$ has whatever is necessary to make the row sum of $B$ equal to $r$ and then 0 for the rest of the row. Then $0 \leq B \leq A$, and $B$ has all row sums equal to $r$. By Theorem 8.1 and Lemma 8.3, this implies that $r = \rho(B) \leq \rho(A)$.

Now define $C \in M_n(\mathbb{R})$ by $c_{ij} = a_{ij}$ for $1 \leq j \leq n-1$ and

$$c_{in} = R - \sum_{j=1}^{n-1} a_{ij}.$$

Then $A \leq C$ and $C$ has all row sums equal to $R$. By Theorem 8.1 and Lemma 8.3, this implies that $\rho(A) \leq \rho(C) = R$.

The second claim can be proved similarly, or follows by applying the first to $A^T$. □

**Corollary 8.5.** *If $A \geq 0$ and each row of $A$ contains at least one nonzero entry, or each column of $A$ contains at least one nonzero entry, then $\rho(A) > 0$.*

*In particular, if $A > 0$ entrywise, then $\rho(A) > 0$.*

We can extend Theorem 8.4 by conjugating $A$ by a diagonal matrix, similar to the way Geršgorin's theorem was extended to Corollary 7.2.

**Corollary 8.6.** *If $A \geq 0$ and $x_1, \ldots, x_n > 0$, then*

$$\min_{1\leq i\leq n} \frac{1}{x_i} \sum_{j=1}^{n} a_{ij}x_j \leq \rho(A) \leq \max_{1\leq i\leq n} \frac{1}{x_i} \sum_{j=1}^{n} a_{ij}x_j.$$

**Corollary 8.7.** *If $A \geq 0$, $x > 0$, and $\alpha x \leq Ax \leq \beta x$ for some $\alpha, \beta \geq 0$, then $\alpha \leq \rho(A) \leq \beta$. If $\alpha x < Ax$, then $\alpha < \rho(A)$, and if $Ax < \beta x$, then $\rho(A) < \beta$.*

*Proof.* The first claim follows immediately from Corollary 8.6. For the second, if $\alpha x < Ax$, then there exists some $\alpha' > \alpha$ such that $\alpha x < \alpha' x < Ax$, and so $\rho(A) \geq \alpha' > \alpha$; the other part follows similarly. □

**Corollary 8.8.** *Let $A \geq 0$. If $A$ has a positive eigenvector $x$, then $Ax = \rho(A)x$.*

*Proof.* Suppose that $Ax = \lambda x$. Since $x > 0$ and $A \geq 0$, this implies that $\lambda \geq 0$. Now since $\lambda x \leq Ax \leq \lambda x$, Corollary 8.7 implies that $\rho(A) = \lambda$. □

The final result of this section can be thought of as a counterpart of the Rayleigh–Ritz theorem for the spectral radius of a nonnegative matrix.

**Corollary 8.9.** *If $A \geq 0$ has a positive eigenvector, then*

$$\rho(A) = \max_{x>0} \min_{1\leq i\leq n} \frac{1}{x_i} \sum_{j=1}^{n} a_{ij}x_j = \min_{x>0} \max_{1\leq j\leq n} \frac{1}{x_i} \sum_{j=1}^{n} a_{ij}x_j.$$

In the following sections we will see sufficient conditions for a nonnegative matrix to have a positive eigenvector.

*Proof.* Suppose that $y$ is a positive eigenvector of $A$. By Corollary 8.8, $Ay = \rho(A)y$, and so

$$\rho(A) = \frac{1}{y_i} \sum_{j=1}^{n} a_{ij}y_j$$

for each $i$. Together with Corollary 8.6 this implies the claim. □

## 8.2 Perron's theorem

This section is devoted to proving the following theorem, which is the fundamental result about the spectral radius of a positive matrix.

**Theorem 8.10** (Perron's theorem)**.** *Suppose that $A > 0$. Then:*

1. *$\rho(A) > 0$.*
2. *$\rho(A)$ is an eigenvalue of $A$ with multiplicity 1.*
3. *There is a positive eigenvector $x > 0$ of $A$ with eigenvalue $\rho(A)$, which is unique up to scalar multiples. With the normalization $\|x\|_1 = 1$, this is called the **Perron eigenvector** of $A$.*
4. *If $\lambda \in \sigma(A)$ and $\lambda \neq \rho(A)$, then $|\lambda| < \rho(A)$.*
5. *Let $x$ be the Perron eigenvector of $A$, and let $y > 0$ be an eigenvector of $A^T$ with eigenvalue $\rho(A)$ normalized so that $\langle x, y\rangle = 1$. Then*

$$\lim_{m\to\infty} \left(\frac{1}{\rho(A)}A\right)^m = xy^T.$$

We will prove Theorem 8.10 a bit at a time. The first part has already been proved in Corollary 8.5.

**Lemma 8.11.** *Suppose that $A > 0$, $Ax = \lambda x$ for $x \neq 0$, and that $|\lambda| = \rho(A)$. Then $A|x| = \rho(A)|x|$, and $|x| > 0$.*

*Proof.* Note first that

$$\rho(A)|x| = |\lambda||x| = |Ax| \leq |A||x| = A|x|,$$

and that $A|x| > 0$ since $A > 0$ and $|x| \geq 0$. Now if $A|x| - \rho(A)|x| \neq 0$, we would have

$$0 < A(A|x| - \rho(A)|x|) = A(A|x|) - \rho(A)(A|x|),$$

which implies that $\rho(A)(A|x|) < A(A|x|)$. Since $A|x| > 0$, Corollary 8.7 would then imply that $\rho(A) > \rho(A)$, which is impossible. Therefore we must have $A|x| - \rho(A)|x| = 0$.

Finally, since $\rho(A)|x| = A|x| > 0$, we must have $|x| > 0$. □

**Proposition 8.12.** *If $A > 0$ then $\rho(A)$ is an eigenvalue of $A$ with a positive eigenvector.*

*Proof.* This follows immediately from Lemma 8.11. □

Proposition 8.12 implies the first half of part 2 (without the multiplicity 1 part) and the first half of part 3 (without the uniqueness) of Theorem 8.10.

**Lemma 8.13.** *Suppose that $A > 0$, $Ax = \lambda x$ for $x \neq 0$, and that $|\lambda| = \rho(A)$. Then $x = e^{i\theta}|x|$ for some $\theta \in \mathbb{R}$.*

*Proof.* By Lemma 8.11,

$$|Ax| = |\lambda x| = \rho(A)|x| = A|x|.$$

Therefore for each $i$, $\left|\sum_{j=1}^n a_{ij}x_j\right| = \sum_{j=1}^n a_{ij}|x_j|$, which implies that the numbers $a_{ij}x_j$ all have the same argument. Since $a_{ij} > 0$, this implies that the $x_j$ all have the same argument, which is equivalent to the claim. □

**Proposition 8.14.** *Suppose that $A > 0$, $\lambda \in \sigma(A)$, and that $\lambda \neq \rho(A)$. Then $|\lambda| < \rho(A)$.*

*Proof.* By Lemma 8.13, if $x$ is an eigenvector associated to an eigenvalue $\lambda$ with $|\lambda| = \rho(A)$, then some scalar multiple $w$ of $x$ is positive. Corollary 8.8 then implies that $\lambda = \rho(A)$. □

Proposition 8.14 proves part 4 of Theorem 8.10.

**Proposition 8.15.** *If $A > 0$, then $\rho(A)$ has geometric multiplicity one as an eigenvalue of $A$. That is, the eigenspace $\ker(A - \rho(A)I_n)$ is one-dimensional.*

*Proof.* Suppose that $x, y \in \mathbb{C}^n$ are both eigenvectors of $A$ with eigenvalue $\rho(A)$. By Lemma 8.13 we may assume that $x, y \geq 0$, and by Lemma 8.11 we then have that $x, y > 0$. Let $\beta = \min_{1 \leq j \leq n} \frac{y_j}{x_j}$. Then $z := y - \beta x \geq 0$ has at least one zero entry, and $Az = \rho(A)z$. Therefore $Az$ has at least one zero entry, which implies $Az = 0$, and so $z = 0$. □

Proposition 8.15 completes the proof of part 3 of Theorem 8.10. Note that part 2 of Theorem 8.10 is actually a stronger statement than Proposition 8.15 (since the (algebraic) multiplicity of an eigenvalue is at least as large as its geometric multiplicity). However, we will use Proposition 8.15 in part to prove that statement.

**Lemma 8.16.** *Suppose that $A \in M_n$, $x, y \in \mathbb{C}^n$, $Ax = \lambda x$, $A^T y = \lambda y$ and $x^T y = 1$. Let $L = xy^T$. Then:*

1. *$Lx = x$ and $L^T y = y$.*
2. *$L^m = L$ for all $m \in \mathbb{N}$.*
3. *$A^m L = LA^m = \lambda^n A$ for all $m \in \mathbb{N}$.*
4. *$L(A - \lambda L) = 0$.*
5. *$(A - \lambda L)^m = A^m - \lambda^m L$ for all $m \in \mathbb{N}$.*
6. *If $0 \neq \mu \in \sigma(A - \lambda L)$ then $\mu \in \sigma(A)$.*
7. *If $0 \neq \lambda \in \sigma(A)$ has geometric multiplicity 1, then $\lambda \notin \sigma(A - \lambda L)$, so $\lambda I_n - (A - \lambda L)$ is nonsingular.*

*Moreover, if $\lambda$ has geometric multiplicty 1 and is the only eigenvalue of $A$ with absolute value $\rho(A)$, then:*

8. *$\rho(A - \lambda L) \leq |\lambda_{n-1}| < \rho(A)$.*
9. *$(\lambda^{-1}A)^m = L + (\lambda^{-1}A - L)^m \xrightarrow{m\to\infty} L$.*

*Proof.* The first three parts are immediate, the fourth follows from the second and third, and the fifth follows from the fourth by an easy induction.

6. Suppose $w$ is an eigenvector of $A - \lambda L$ with nonzero eigenvalue $\mu$. Then $\mu L w = L(A - \lambda L)w = 0$ by part 4, so $Lw = 0$, and therefore $\mu w = (A - \lambda L)w = Aw$.

7. Suppose $0 \neq \lambda \in \sigma(A)$ has geometric multiplicity 1. Then every eigenvector of $A$ with eigenvalue $\lambda$ is a scalar multiple of $x$. Suppose now that $\lambda$ is an eigenvalue of $A - \lambda L$ with eigenvector $w$. The proof above of part 6 showns that $Aw = \lambda w$, and so $w = \alpha x$ for some $\alpha \in C$. Then

$$\lambda w = (A - \lambda L)w = \alpha(A - \lambda L)x = 0,$$

   and so $w = 0$, which is a contradiction.

8. By part 6, either $\rho(A - \lambda L) = |\mu|$ for some $\mu \in \sigma(A)$, or else $\rho(A - \lambda L) = 0$. By part 7, in the former case $\mu \neq \lambda$, and so $|\mu| < \rho(A)$.

9. By parts 5 and 2,

$$(\lambda^{-1}A - L)^m = (\lambda^{-1}A)^m - L^m = (\lambda^{-1}A)^m - L.$$

   By part 8, $\rho(\lambda^{-1}A - L) = |\lambda|^{-1}\rho(A - \lambda L) < 1$, which implies that $(\lambda^{-1}A - L)^m \xrightarrow{m\to\infty} 0$ by Theorem 4.26. □

Propositions 8.14 and 8.15 and Lemma 8.16 imply part 5 of Theorem 8.10.

It remains to prove part 2 of Theorem 8.10:

**Proposition 8.17.** *If $A > 0$ then $\rho(A)$ has (algebraic) multiplicity 1 as an eigenvalue of $A$.*

*Proof.* Let $A = UTU^*$ be a Schur decomposition of $A$, with $\rho = \rho(A)$ as the first $k$ diagonal entries of $T$. By Proposition 8.14, $|t_{jj}| < \rho$ for $j > k$. Then $U^*(\rho^{-1}A)^m U^m = (\rho^{-1}T)^m$ has the first $k$ entries equal to 1 for each $m$. By part 5 of Theorem 8.10, $U^*(\rho^{-1}A)^m U$ converges as $m \to \infty$ to a matrix with rank 1. It follows that $k = 1$. □

Only one part of Perron's theorem (Theorem 8.10) extends to nonnegative matrices without adding additional assumptions:

**Theorem 8.18.** *Suppose that $A \geq 0$. Then $\rho(A)$ is an eigenvalue of $A$ with a nonnegative eigenvector.*

We will give two proofs of Theorem 8.18. The first fits neatly with recurring themes of this course.

*First proof of Theorem 8.18.* Let $A_k$ be a sequence of positive matrices such that $A_k \xrightarrow{k\to\infty} A$ (for example, $[A_k]_{ij} = \max\{a_{ij}, \frac{1}{k}\}$). Let $x_k > 0$ be the Perron eigenvector of $A_k$, so that $\|x_k\|_1 = 1$ and $A_k x_k = \rho(A_k)x_k$ for each $k$. The set $\{x \in \mathbb{R}^n \mid x \geq 0,\ \|x\|_1 = 1\}$ is closed and bounded, so there is a subsequence $x_{k_m}$ which converges to some $x \geq 0$ with $\|x\|_1 = 1$. Moreover, we have $A_{k_m} \xrightarrow{m\to\infty} A$ and, by the continuity of eigenvalues, $\rho(A_{k_m}) \to \rho(A)$. It follows that

$$Ax = \lim_{m\to\infty} A_{k_m}x_{k_m} = \lim_{m\to\infty} \rho(A_{k_m})x_{k_m} = \rho(A)x. \qquad \square$$

The second proof of Theorem 8.18 illustrates a new idea: using nontrivial topological theorems. We will need the following result, which is typically proved using homology theory.

**Theorem 8.19** (Brouwer's fixed point theorem)**.** *Suppose that $C \subseteq \mathbb{R}^n$ is closed, bounded, and convex. Then each continuous function $f : C \to C$ has a fixed point. That is, there exists an $x \in C$ such that $f(x) = x$.*

*Second proof of Theorem 8.18.* The set

$$C = \{x \in \mathbb{R}^n \mid x \geq 0,\ \|x\|_1 = 1,\ Ax \geq \rho(A)x\}$$

is closed, bounded, and convex. If $\lambda \in \sigma(A)$ satisfies $|\lambda| = \rho(A)$ and $Av = \lambda v$, then

$$A|v| \geq |Av| = |\lambda v| = \rho(A)|v|.$$

It follows that $\frac{|v|}{\|v\|_1} \in C$, and therefore $C \neq \emptyset$.

Define $f : C \to \mathbb{R}^n$ by

$$f(x) = \frac{Ax}{\|Ax\|_1}.$$

Clearly $f(x) \geq 0$ and $\|f(x)\|_1 = 1$. Furthermore,

$$Af(x) = \frac{1}{\|Ax\|_1} AAx \geq \frac{1}{\|Ax\|_1} A\rho(A)x = \rho(A)f(x).$$

Therefore $f$ maps $C$ into $C$. By Theorem 8.19, there exists a $y \in C$ such that $f(y) = y$. We then have

$$Ay = \|Ay\|_1 \, y \geq \rho(A)y,$$

where the inequality follows since $y \in C$. Thus $y \geq 0$ is an eigenvector of $A$ with positive eigenvalue $\|Ay\|_1 \geq \rho(A)$; it follows that in fact the eigenvalue is $\rho(A)$. □

## 8.3 Irreducible nonnegative matrices

To state a generalization of Perron's theorem, we will need another piece of terminology.

A matrix $A \in M_n$ is called **reducible** if there exists a permutation matrix $P \in M_n$ such that $P^{-1}AP$ has the block form

$$P^{-1}AP = \begin{bmatrix} B & C \\ 0_{n-r,r} & D \end{bmatrix}$$

where $0_{n-r,r}$ denotes an $(n-r) \times r$ block of 0's with $1 \leq r \leq n-1$. Equivalently, some proper subset of the standard basis vectors spans an invariant subspace for $A$. If $A$ is not reducible, then $A$ is called **irreducible**. Note that any strictly positive matrix is clearly irreducible.

Theorem 8.4 immediately implies:

**Proposition 8.20.** *If $A \geq 0$ is irreducible, then $\rho(A) > 0$.*

We will need the following reformulation of irreducibility for nonnegative matrices.

**Proposition 8.21.** *Suppose that $A \in M_n$, $A \geq 0$. Then $A$ is irreducible if and only if $(I_n + A)^{n-1} > 0$.*

*Proof.* We clearly have $(I_n + A)^{n-1} \geq 0$. We therefore need to show that $A$ is reducible if and only if some entry of $(I_n + A)^{n-1}$ is 0.

Suppose that $A$ is irreducible. Without loss of generality, we may assume that $A = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix}$ with square blocks $B$ and $D$. It follows that

$$(I_n + A)^{n-1} = \sum_{k=0}^{n-1} \binom{n-1}{k} A^k = \sum_{k=0}^{n-1} \binom{n-1}{k} \begin{bmatrix} B^k & * \\ 0 & D^k \end{bmatrix}$$

for some value of $*$, and so $(I_n + A)^{n-1}$ has a 0 entry.

Now suppose that $[(I_n + A)^{n-1}]_{pq} = 0$. Then

$$\sum_{i_1,\ldots,i_{n-1}=1}^{n} [I_n + A]_{p,i_1}[I_n + A]_{i_1,i_2} \cdots [I_n + A]_{i_{n-1},q} = 0.$$

Each factor of each of the summands of the left hand side above is nonnegative, and so each summand is 0, so for each choice of $i_1, \ldots, i_{n-1}$ at least one of the factors $[I_n + A]_{p,i_1}$, $[I_n + A]_{i_1,i_2}$, $\ldots$, $[I_n + A]_{i_{n-1},q}$ is 0.

Let $J_1 = \{k \mid [I_n + A]_{p,k} \neq 0\}$ and iteratively define

$$J_i = \{k \mid [I_n + A]_{m,k} \neq 0 \text{ for some } m \in J_{i-1}\}$$

for $2 \leq i \leq n$. Note that since $[I_n + A]_{jj} \geq 1$ for each $j$, $J_{i-1} \subseteq J_i$ for each $i$. The argument above shows that $q \notin J_n$. Note that $[I_n + A]_{jk} = 0$, and hence $a_{jk} = 0$, whenever $j \in J_n$ and $k \notin J_n$. Therefore, if $P$ is a permutation matrix that reorders $\{1, \ldots, n\}$ to put the indices in $J$ after all the indices in $J^c$, then $PAP^{-1}$ has the block form $\begin{bmatrix} B & C \\ 0 & D \end{bmatrix}$, and so $A$ is reducible. □

**Theorem 8.22** (Perron–Frobenius theorem)**.** *Suppose that $A \geq 0$ is irreducible. Then $\rho(A) > 0$ is an eigenvalue of $A$ with multiplicity* 1*, and there is a corresponding positive eigenvector.*

*Proof.* We have already seen that $\rho(A) > 0$ (in Proposition 8.20) and that $\rho(A)$ is an eigenvalue of $A$ with a nonnegative corresponding eigenvector $x$ (Theorem 8.18). Then

$$(I_n + A)x = (1 + \rho(A))x = \rho(I_n + A)x$$

since $\rho(A)$ is an eigenvalue of $A$. By Proposition 8.21, it follows that

$$0 < (I_n + A)^{n-1}x = (1 + \rho(A))^{n+1}x,$$

and so $x > 0$.

Finally, by Perron's theorem, $\rho(I_n + A)^{n-1}$ has multiplicity 1 as an eigenvalue of the positive (by Proposition 8.21) matrix $(I_n+A)^{n-1}$. This implies that $1+\rho(A) = \rho(I_n+A)$ has multiplicity 1 as an eigenvalue of $I_n + A$, and thus $\rho(A)$ has multiplicity 1 as an eigenvalue of $A$. □

To generalize the last part of Perron's theorem, we need yet another condition. A matrix $A \geq 0$ is called **primitive** if $A$ is irreducible and $\rho(A)$ is the unique eigenvalue of $A$ with modulus $\rho(A)$. The same argument as in the proof of Lemma 8.16 proves the following:

**Theorem 8.23.** *Suppose that $A \geq 0$ is primitive. Let $x > 0$ and $y > 0$ satisfy $Ax = \rho(A)x$ and $A^T y = \rho(A)y$ be normalized so that $\|x\|_1 = 1$ and $\langle x, y\rangle = 1$. Then*

$$\lim_{m\to\infty} \left(\frac{1}{\rho(A)}A\right)^m = xy^T.$$

Note that the existence of $x$ and $y$ as in the statement above is guaranteed by the Perron–Frobenius theorem.

Since we have essentially made a hypothesis of one of the key conclusions of Perron's theorem here, it is desirable to have another characterization of primitivity. The following is useful.

**Proposition 8.24.** *A matrix $A \geq 0$ is primitive if and only if $A^m > 0$ for some $m \in \mathbb{N}$.*

*Proof.* Suppose that $A$ is primitive. By Theorem 8.23, $\left(\frac{1}{\rho(A)}A\right)^m \to xy^T$, which is strictly positive. Thus for some $m$, $A^m > 0$.

Now suppose that $A^m > 0$. If we could write $PAP^{-1} = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix}$ for some permutation matrix $P$, we would have $PA^mP^{-1} = \begin{bmatrix} B^m & * \\ 0 & D^m \end{bmatrix}$, which is false; thus $A$ is irreducible. By Perron's theorem (Theorem 8.10), $\rho(A^m) = \rho(A)^m$ is the unique eigenvalue of $A^m$ with modulus $\rho(A)^m$, with multiplicity 1. It follows that $\rho(A)$ is the unique eigenvalue of $A$ with modulus $\rho(A)$. □

## 8.4 Stochastic matrices and Markov chains

A matrix $P \in M_n(\mathbb{R})$ is called **stochastic** if $P \geq 0$ and every row of $P$ adds up to 1. (Note that $P$ is doubly stochastic if and only if both $P$ and $P^T$ are stochastic.) We can state the latter condition as $Pe = e$, where $e = (1, \ldots, 1)$; this makes it easy to check that every power of a stochastic matrix is again stochastic.

Stochastic matrices have a natural interpretation in terms of probability. Let $\Omega = \{x_1, \ldots, x_n\}$ be some set with $n$ elements, and fix a stochastic matrix $P \in M_n(\mathbb{R})$. Then $P$ can be used to describe a **Markov chain** on $\Omega$, that is, a sequence $X_0, X_1, \ldots$ of random points in $\Omega$ with the property that if $X_t = x_i$, then $X_{t+1}$ will be $x_j$ with probability $p_{ij}$. The fact that $\sum_{j=1}^n p_{ij} = 1$ means that this completely describes the probability distribution of $X_{t+1}$ given $X_t$. (The initial point $X_0$ might be given explicitly, or it might be chosen at random according to some probability distribution on $\Omega$; we will return to this below.) The matrix $P$ is called the **transition matrix** of the Markov chain. A Markov chain is typically thought of as a "random walk without memory".

More generally, if we write $\mathbb{P}[A|B]$ for the probability that $A$ is true, given that we know $B$, then

$$\begin{aligned}
&\mathbb{P}[X_{t+s} = x_j | X_t = x_i] \\
&= \sum_{k_1,\ldots,k_{s-1}=1}^{n} \mathbb{P}[X_{t+1} = x_{k_1} | X_t = x_i]\mathbb{P}[X_{t+2} = x_{k_2} | X_{t+1} = x_{k_1}] \cdots \mathbb{P}[X_{t+s} = x_j | X_{t+s-1} = x_{k_{s-1}}] \\
&= \sum_{k_1,\ldots,k_{s-1}=1}^{n} p_{i,k_1} p_{k_1,k_2} \cdots p_{k_{s-1},j} = [P^s]_{i,j}.
\end{aligned}$$

Suppose that $\pi \geq 0$ is an $n$-dimensional *row* vector with $\sum_{i=1}^n \pi_i = 1$. Again, we can state the latter condition as $\pi e = 1$. We can interpret $\pi$ as a probability distribution on $\Omega$: $\pi_i$ is the probability of picking $x_i$. Below we will always consider probability distributions to be row vectors.

If $P$ is a stochastic matrix, then $\pi P \geq 0$, and $(\pi P)e = \pi(Pe) = \pi e = 1$, so $\pi P$ is again a probability distribution. Specifically, it describes the distribution of $X_1$, if $X_0$ is distributed according to $\pi$. We call $\pi$ a **stationary distribution** for $P$ is $\pi P = \pi$.

**Lemma 8.25.** *If $P$ is a stochastic matrix, then $P$ has a stationary distribution $\pi$.*

*Proof.* By Lemma 8.3, $\rho(P^T) = 1$. Theorem 8.18 implies that $P^T$ has a nonnegative eigenvalue $y$ with $P^T y = y$. By renormalizing if necessary we can set $\|y\|_1 = 1$. Then $\pi = y^T$ is a stationary distribution for $P$. □

The following is the fundamental result about the convergence of Markov chains.

**Theorem 8.26.** *If $P$ is an irreducible stochastic matrix, then $P$ has a unique stationary distribution $\pi$, and $\pi > 0$. If $P$ is moreover primitive, then for any probability distribution $\mu$,*

$$\lim_{t\to\infty} \mu P^t = \pi.$$

The probabilistic interpretation of the convergence statement is the following: no matter how the initial step $X_0$ of a Markov chain with a primitive transition matrix is chosen, for large $t$ the distribution of $X_t$ is approximately given by the stationary distribution $\pi$.

*Proof.* Note first that $P$ is irreducible or primitive if and only if $P$ is. By the Perron–Frobenius theorem 8.22, $P^T$ has a unique eigenvector with eigenvalue 1 (up to scalar multiples), which is moreover positive; its transpose is therefore the unique stationary distribution $\pi$.

If $P$ is also primitive, then Theorem 8.23 applies to $A = P^T$ with $x = \pi^T$ and $y = e^T$, and implies that $\lim_{t\to\infty} P^t = e\pi$. It follows that if $\mu$ is any probability distribution, then

$$\lim_{t\to\infty} \mu P^t = \mu(e\pi) = (\mu e)\pi = \pi.$$ □

As a first example, consider *simple random walk on the discrete circle*: given $n$, let $\Omega = \{x_1, \dots, x_n\}$ consist of $n$ equally spaced points on the unit circle. We consider the stochastic matrix $P$ with

$$p_{ij} = \begin{cases} 1/2 & \text{if } j = i \pm 1, \text{ or if } \{i,j\} = \{1,n\}, \\ 0 & \text{otherwise.} \end{cases}$$

That is, the random walk moves either clockwise or counterclockwise one space, with equal probability. It is easy to check that the uniform probability distribution $\pi = \left(\frac{1}{n}, \dots, \frac{1}{n}\right)$ is a stationary distribution for $P$; in fact $P$ is irreducible, and so the stationary distribution is unique.

However, if $n$ is even then $P$ is not primitive, and the convergence in Theorem 8.26 does not hold. For this reason, it is often convenient to consider the *lazy version* of the random walk:

$$p_{ij} = \begin{cases} 1/2 & \text{if } i = j, \\ 1/4 & \text{if } j = i \pm 1, \text{ or if } \{i,j\} = \{1,n\}, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

That is, the random walk stays put with probability 1/2; if it moves, it moves one step in either direction with equal probability. The lazy random walk is irreducible and primitive, so that the uniform distribution is its unique stationary distribution.

For a second example (which generalizes the first), consider a graph $G = (V, E)$ consisting of a set of vertices $V = \{x_1, \dots, x_n\}$ and edges connecting pairs of vertices. We write

$x_i \sim x_j$ if there is an edge connecting $x_i$ and $x_j$. The **degree** of a vertex $x_i$, written $\deg(x_i)$ is the number of vertices connected to $x_i$. The *simple random walk on* $G$ has transition matrix

$$p_{ij} = \begin{cases} \frac{1}{\deg(x_i)} & \text{if } x_i \sim x_j, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Then

$$\pi_i = \frac{\deg(x_i)}{2\#E} \tag{12}$$

defines a stationary distribution for $P$. Again $P$ is not necessarily primitive, but can be modified to be primitive by making it "lazy".

An important aspect of Markov chains is their "local" behavior: given $X_t$, picking $X_{t+1}$ only requires knowing $p_{x_t,j}$ for $1 \leq j \leq n$. That is, you never need to use the entire matrix $P$, and don't need to know the entire matrix a priori. In the context of random walk on a graph, for example, given a node $X_t$, to pick $X_{t+1}$ at random you only need to be able to see which vertices are connected to $X_t$. Now if $P$ is primitive, then by Theorem 8.26 we know that for large $t$, the probability of being at $x_i \in \Omega$ at a given step is approximately $\pi_i$ — regardless of how the Markov chain was started.

This observation is the starting point of *Markov chain Monte Carlo* techniques: in order to choose a random element from $\Omega$ according to a probability distribution $\pi$, you can run a primitive Markov chain with stationary distribution $\pi$. Moreover, you can do this even without complete knowledge of the transition matrix or of $\pi$ itself.

To return to random walk on a graph, suppose we wish to estimate the number of edges in $G$. Assuming the transition matrix is primitive, if we run random walk for a long time, then

$$\mathbb{P}[X_t = x_i] \approx \frac{\deg(x_i)}{2\#E}.$$

Thus we can estimate

$$\#E \approx \frac{\deg(x_i)}{2\mathbb{P}[X_t = x_i]};$$

the right hand side of this can be estimated by observing just how often $X_t$ turns out to be a give $x_i$.

A last example is furnished by the PageRank algorithm, made famous by its use as a key component of Google's search algorithm for the World Wide Web. We start with the *directed* graph $G = (V, E)$ whose vertices are all the web pages containing a given search term, and with edges representing links from one page to another. The goal is to rank these pages in some useful way. The basic idea is that useful pages are likely to be ones that are linked to by many other useful pages. (Superficially this sounds circular, but it is really no more so than an eigenvector problem — which, in fact, it is an example of.) This suggests considering a random walk on $G$: the random walk is most likely to end up at these useful pages. So the invariant measure $\pi$ for the random walk should give a good measure of usefulness.

The trouble with this basic idea is that there may be many dead ends in the directed graph $G$, resulting in a reducible transition matrix. One simple way around this is to

add edges from each dead end to every other vertex. That is, we start with the modified adjacency matrix $A$ given by

$$a_{ij} = \begin{cases} 1 & \text{if } x_i \to x_j, \\ 1 & \text{if } x_i \text{ is a dead end, .} \\ 0 & \text{otherwise.} \end{cases}$$

We then normalize the rows to get a transition matrix $P$:

$$p_{ij} = \frac{a_{ij}}{\sum_{k=1}^{n} a_{ik}}.$$

There is still no guarantee that this transition matrix is primitive, so we modify it slightly by setting $Q = (1-\varepsilon)P + \frac{\varepsilon}{n}J$ for some small $\varepsilon > 0$, where $J$ is the matrix whose entries are all 1. Probabilistically, this represents a Markov chain where, given $X_t$, the random walker goes to a completely random page with probability $\varepsilon$; and otherwise choose a page that $X_t$ links to at random, assuming there are any; and otherwise again picks a completely random page. Since $Q > 0$, $Q$ is primitive. It therefore has a unique stationary distribution $\pi$, which can be approximated by, say, $e_1^T Q^m$ for large $m$. We rank web pages $x_i$ according to the size of $\pi_i$.

The last key observation is that $e_1^T Q^m$ can be computed quickly in practice, even though the size $n = \#V$ of the matrix is often huge. The point is that a given page typically has a small number of links, but is also unlikely to have no links. Therefore $P$ is a very sparse matrix. On the other hand, $J = ee^T$, so if $\mu$ is any probability distribution, then $\mu J = e^T$. It follows that

$$\mu Q = (1-\varepsilon)\mu P + \frac{\varepsilon}{n}e^T$$

can be computed very quickly, allowing $e_1^T Q^m \approx \pi$ to be quickly computed for large $m$.

## 8.5 Reversible Markov chains

A stochastic matrix $P$ is called **reversible** with respect to a probability distribution $\pi$ if

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{13}$$

for every $i, j$. The equation (13) is sometimes called the **detailed balance equation**. Probabilistically, it says that if $X_0$ is distributed according to $\pi$, then

$$\mathbb{P}[X_0 = x_i \text{ and } X_1 = x_j] = \mathbb{P}[X_0 = x_j \text{ and } X_1 = x_i].$$

Since $P$ is stochastic, (13) implies that

$$[\pi P]_j = \sum_{i=1}^{n} \pi_i p_{ij} = \sum_{i=1}^{n} \pi_j p_{ji} = \pi_j;$$

that is, $\pi P = \pi$, and so $\pi$ is a stationary distribution for $P$. Note that if $P$ is also irreducible, then it has a unique stationary distribution by Theorem 8.26, and therefore it can be reversible with respect to at most one probability distribution.

For example, if $P$ is a symmetric stochastic matrix, then $P$ is reversible with respect to $\pi = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$. Also, for simple random walk on a graph, from (11) and (12) we have

$$\pi_i p_{ij} = \begin{cases} \frac{1}{2\#E} & \text{if } x_i \sim x_j, \\ 0 & \text{otherwise,} \end{cases}$$

which implies that (13) holds.

On the other hand, not every stochastic matrix is reversible with respect to its stationary distribution. If $\varepsilon \in (0,1)$, the *biased random walk on the discrete circle* has the transition matrix

$$p_{ij} = \begin{cases} \varepsilon & \text{if } j = i+1 \text{ or } i = n \text{ and } j = 1, \\ 1-\varepsilon & \text{if } j = i-1 \text{ or } i = 1 \text{ and } j = n, \\ 0 & \text{otherwise.} \end{cases}$$

Then $P$ is irreducible and has $\pi = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ as its unique stationary distribution, but if $\varepsilon \neq 1/2$ then $P$ is not symmetric, and so $P$ is not reversible with respect to $\pi$.

Assuming that $\pi > 0$ (which, by Theorem 8.26, holds if $P$ is irreducible), (13) equivalently says that if we define $S = \operatorname{diag}(\sqrt{\pi_1}, \ldots, \sqrt{\pi_n})$, then $SPS^{-1} = \left[\sqrt{\frac{\pi_i}{\pi_j}} p_{ij}\right]$ is a symmetric matrix. This observation implies the following.

**Proposition 8.27.** *If $P$ is the transition matrix for an irreducible Markov chain, then all the eigenvalues of $P$ are real.*

The symmetry of $SPS^{-1}$ of course implies more: $SPS^{-1}$ is orthogonally diagonalizable. We will see one way to exploit that fact in this context in the next section.

## 8.6 Convergence rates for Markov chains

Theorem 8.26 assures us that if $P$ is a primitive stochastic matrix with stationary distribution $\pi$, then $\mu P^t \xrightarrow{m\to\infty} \pi$ for any probability distribution $\mu$. A potentially important question for applications is: how fast does this convergence happen?

That is, we would like to bound $\left\|\mu P^t - \pi\right\|$, as a function of $t$, for some norm $\|\cdot\|$. We note first that it suffices to consider $\mu = e_i^T$ for $i = 1, \ldots, n$ (probabilistically: to assume that $X_0 = x_i$ with probability 1 for some $i$):

$$\left\|\mu P^t - \pi\right\| = \left\|\sum_{i=1}^n \mu_i \left(e_i^T P^t - \pi\right)\right\| \leq \sum_{i=1}^n \mu_i \left\|e_i^T P^t - \pi\right\| \leq \max_{1\leq i\leq n} \left\|e_i^T P^t - \pi\right\|.$$

Note that $[e_i^T P^t]_j = \mathbb{P}[X_t = x_j | X_0 = x_i]$.

For the purposes of probability theory, one particularly natural choice is the $\ell^1$ norm (for reasons that will be explored in the homework). But as we know well, it is frequently easier to work with quantities related to $\ell^2$ norms. Given two probability distributions $\mu$ and $\pi$ with $\pi > 0$, we define the $\chi^2$ distance by

$$\chi^2(\mu, \pi) = \sum_{i=1}^n \pi_i \left(\frac{\mu_i}{\pi_i} - 1\right)^2 = \sum_{i=1}^n \frac{1}{\pi_i}(\mu_i - \pi_i)^2.$$

If $S = \operatorname{diag}(\sqrt{\pi_1}, \dots, \sqrt{\pi_n})$ as in the previous section, then

$$\chi^2(\mu, \pi) = \left\| (\mu_i - \pi_i) S^{-1} \right\|_2^2. \tag{14}$$

Note that $\chi^2(\mu, \pi) \neq \chi^2(\pi, \mu)$ in general.

**Lemma 8.28.** *If $\mu$ and $\pi$ are probability distributions with $\pi > 0$, then*

$$\|\mu - \pi\|_1 \leq \sqrt{\chi^2(\mu, \pi)}.$$

*Proof.* By the Cauchy–Schwarz inequality,

$$\begin{aligned} \|\mu - \pi\|_1 &= \sum_{i=1}^n |\mu_i - \pi_i| = \sum_{i=1}^n \sqrt{\frac{1}{\pi_i}} \, |\mu_i - \pi_i| \, \sqrt{\pi_i} \\ &\leq \sqrt{\sum_{i=1}^n \frac{1}{\pi_i} (\mu_i - \pi_i)^2} \sqrt{\sum_{i=1}^n \pi_i} = \sqrt{\chi^2(\mu, \pi)}. \end{aligned}$$

□

The following is a basic example of how the rate of convergence of a Markov chain can be bounded using spectral information.

**Theorem 8.29.** *Suppose that $P$ is a primitive stochastic matrix which is reversible with respect to $\pi$. Let $C = \max_{\lambda \in \sigma(P) \setminus \{1\}} |\lambda|$ and $\kappa = \min_{1 \leq i \leq n} \pi_i$. Then $C < 1$, and for each $i = 1, \dots, n$ and each $t \in \mathbb{N}$, we have*

$$\chi^2(e_i^T P^t, \pi) \leq \frac{1}{\kappa} C^{2t} \qquad \text{and} \qquad \left\| e_i^T P^t - \pi \right\|_1 \leq \frac{1}{\sqrt{\kappa}} C^t.$$

Theorem 8.29 shows that, under these hypotheses, $e_i^T P^t$ converges exponentially quickly to $\pi$. One commonly used (but rather arbitrary) way state such a result is in terms of the *mixing time*, defined as the smallest $\tau$ such that $\left\| e_i^T P^t - \mu \right\|_1 \leq 1/2$ for all $t \geq \tau$. Theorem 8.29 implies in particular that the mixing time is at most $\frac{\log(\sqrt{\kappa}/2)}{\log C}$.

*Proof.* Since $P$ is primitive and stochastic, $\rho(P) = 1$ and $C < 1$ by the definition of primitive.

Since $P$ is reversible with respect to $\pi$, if $S = \operatorname{diag}(\sqrt{\pi_1}, \dots, \sqrt{\pi_n})$, then $A = SPS^{-1}$ is real symmetric. We also have $\pi P = \pi$, so that $A(\pi S^{-1})^T = (\pi S^{-1})^T$; note that $\pi S^{-1} = (\sqrt{\pi_1}, \dots, \sqrt{\pi_n})$ is a unit vector. Then by the spectral theorem $A = U \Lambda U^T$ for an orthogonal $U \in M_n(\mathbb{R})$ and $\Lambda = \operatorname{diag}(1, \lambda_2, \dots, \lambda_n)$, with $|\lambda_j| \leq C$ for $2 \leq j \leq n$, and we can take the first column of $U$ to be $(\pi S^{-1})^T$. That is, $U e_1 = (\pi S^{-1})^T$, so $e_1^T U^T = \pi S^{-1}$, and thus $\pi S^{-1} U = e_1^T$.

We now have

$$\begin{aligned} \chi^2(e_i^T P^t, \pi) &= \left\| (e_i^T P^t - \pi) S^{-1} \right\|_2^2 = \left\| (e_i^T - \pi) P^t S^{-1} \right\|_2^2 = \left\| (e_i^T - \pi) S^{-1} A^t \right\|_2^2 \\ &= \left\| (e_i^T - \pi) S^{-1} U \Lambda^t U^T \right\|_2^2 = \left\| (e_i^T - \pi) S^{-1} U \Lambda^t \right\|_2^2 \\ &= \sum_{j=1}^n \left[ (e_i^T S^{-1} U - e_1^T) \lambda_j^t e_j \right]^2 \\ &= \left( e_i^T S^{-1} U e_1 - 1 \right) + \sum_{j=2}^n \left[ e_i^T S^{-1} U e_j \lambda_j^t \right]^2 \end{aligned}$$

Now $S^{-1}Ue_1 = S^{-2}\pi^T = e$, so $e_i^T S^{-1} U e_1 = 1$. Using this and the definition of $S$,

$$\chi^2(e_i^T P^t, \pi) = \frac{1}{\pi_i} \sum_{j=2}^{n} \left[e_i^T U e_j\right]^2 \lambda_j^{2t} \leq \frac{C^{2t}}{\kappa} \left\|e_i^T U\right\|_2^2 = \frac{C^{2t}}{\kappa}.$$

Finally, the $\ell^1$ bound follows from the $\chi^2$ bound via Lemma 8.28. □

We illustrate this result with the example of the lazy random walk on the discrete circle, with transition matrix given by (10).

This transition matrix is an example of a circulant matrix, so we can compute its eigenvalues and eigenvectors explicitly: for $k = 0, 1, \ldots, n-1$, the vector $v_k \in \mathbb{C}^n$ with entries $[v_k]_j = e^{2\pi i j k/n}$ is an eigenvector with eigenvalue

$$\lambda_k = \frac{1}{2} + \frac{1}{4} e^{2\pi i k/n} + \frac{1}{4} e^{-2\pi i k/n} = \frac{1}{2}\left(1 + \cos\frac{2\pi k}{n}\right).$$

Then $\lambda_0 = 1$ is the trivial eigenvalue that we already know about. The others are all between 0 and 1 with the largest being when $k = 1$ or $n-1$. So here we can apply Theorem 8.29 with $\kappa = \frac{1}{n}$ and

$$C = \frac{1}{2}\left(1 + \cos\frac{2\pi}{n}\right).$$

To be a bit more explicit, by Taylor's theorem with a remainder,

$$\cos x \leq 1 - \frac{x^2}{2} + \frac{1}{6}x^3,$$

so if, say, $n \geq 7$ (so that $\frac{2\pi}{n} < 1$), then

$$C \leq 1 - \frac{4\pi^2}{6n^2}.$$

Theorem 8.29 now implies that

$$\left\|e_i^T P^t - \pi\right\|_1 \leq \sqrt{n}\left(1 - \frac{4\pi^2}{6n^2}\right)^t.$$

As you will see in homework, it's possible to do much better in this example, in particular because we can identify all the eigenvalues explicitly.

# 9 Spectral graph theory

## 9.1 Eigenvalues of the adjacency matrix

In this final section we will see some examples of how properties of graphs are related to eigenvalues of matrices related to the graph. We will focus first on the adjacency matrix. Given a (simple, undirected) graph $G = (V, E)$ with $V = \{x_1, \ldots, x_n\}$, the **adjacency matrix** $A \in M_n(\mathbb{R})$ is given by

$$a_{ij} = \begin{cases} 1 & \text{if } x_i \sim x_j, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $A$ is both symmetric and nonnegative.

Before considering eigenvalues explicitly, let's see how some basic information about $G$ can be extracted from $A$. Of course, $A$ contains all the information about $G$; the point here is that we want to relate graph-theoretic information about $G$ to matrix-theoretic information about $A$.

We first need a little terminology. A graph $G = (V, E)$ is **disconnected** if $V = V_1 \cup V_2$ with $V_1, V_2 \neq \emptyset$ and whenever $v_1 \in V_1$ and $v_2 \in V_2$, we have $v_1 \not\sim v_2$. A graph is **connected** if it is not disconnected.

**Lemma 9.1.** *A graph is connected if and only if its adjacency matrix is irreducible.*

The proof of Lemma 9.1 is an exercise in remembering the definition of irreducibility.

A **triangle** in $G$ is a set of three distinct vertices $v_1, v_2, v_3$ such that $v_1 \sim v_2$, $v_2 \sim v_3$, and $v_3 \sim v_1$.

**Proposition 9.2.** *Let $G = (V, E)$ be a graph with adjacency matrix $A$. Then $\operatorname{tr} A^2 = 2\#E$ and $\operatorname{tr} A^3$ is 6 times the number of triangles in $G$.*

*Proof.* Since $A$ is real symmetric,

$$\operatorname{tr} A^2 = \sum_{i,j=1}^{n} a_{ij}^2 = \#\left\{(i,j) \mid x_i \sim x_j\right\} = 2\#E,$$

since the set above counts each edge twice. Similarly,

$$\operatorname{tr} A^3 = \sum_{i,j,k=1}^{n} a_{ij}a_{jk}a_{ki} = \#\left\{(i,j,k) \mid x_i \sim x_j,\ x_j \sim x_k,\ x_k \sim x_i\right\}.$$

Each triangle is counted in this set 6 times. □

Note that although Proposition 9.2 does not mention eigenvalues explicitly, $\operatorname{tr} A^k = \sum_{i=1}^{n} \lambda_i(A)^k$ for each $k$.

We next observe which facts can be deduced with little or no extra effort from results we know about nonnegative matrices or about real symmetric matrices.

The degree of a vertex in $G$ is the corresponding row sum of $A$:

$$\deg(x_i) = \sum_{j=1}^{n} a_{ij}.$$

We denote by

$$\delta(G) = \min_{1 \leq i \leq n} \deg(x_i) \qquad \text{and} \qquad \Delta(G) = \max_{1 \leq i \leq n} \deg(x_i)$$

the minimal and maximal degrees of $G$.

Theorem 8.4 immediately implies the following.

**Proposition 9.3.** *If $A$ is the adjacency matrix of a graph $G$, then $\delta(G) \leq \rho(A) \leq \Delta(G)$.*

From the Rayleigh–Ritz theorem (Theorem 3.1) we can further deduce the following refinement of Proposition 9.3.

**Proposition 9.4.** *If $A$ is the adjacency matrix of a graph $G$ with vertices $V = \{x_1, \dots, x_n\}$, then $\frac{1}{n}\sum_{i=1}^n \deg(x_i) \le \lambda_{\max}(A) \le \Delta(G)$.*

*Proof.* The upper bound follows from Proposition 9.3. The lower bound follows by applying the Theorem 3.1 with $x = e$:

$$\lambda_{\max}(A) \ge \frac{\langle Ae, e\rangle}{\|e\|^2} = \frac{\sum_{i,j=1}^n a_{ij}}{n} = \frac{1}{n}\sum_{i=1}^n \deg(x_i). \qquad \square$$

A graph $G$ is called **regular** if $\delta(G) = \Delta(G)$.

**Proposition 9.5.** *Suppose $G$ is connected. Then $\Delta = \Delta(G)$ is an eigenvalue of $A$ if and only if $G$ is regular.*

*Proof.* If $G$ is regular with degree $\Delta$, then $Ae = \Delta e$. (Note that this implication does not need connectedness.)

Now suppose that $\Delta$ is an eigenvalue of $A$. By Proposition 9.3, $\rho(A) = \Delta$. Since $G$ is connected, $A$ is irreducible by Lemma 9.1, so the Perron–Frobenius theorem applies. If $y$ is the Perron eigenvector of $A$, then we have $Ay = \Delta y$. Adding the components of this vector, we obtain

$$\Delta = \sum_{i=1}^n \Delta y_i = \sum_{i=1}^n \sum_{j=1}^n a_{ij} y_j = \sum_{j=1}^n \left( \sum_{i=1}^n a_{ij} \right) y_j = \sum_{j=1}^n \deg(x_j) y_j,$$

and so $\sum_{j=1}^n (\Delta - \deg(x_j)) y_j = 0$. Since each term of this sum is nonnegative, we must have $\Delta = \deg(x_j)$ for each $j$. $\square$

We will present one more substantial example result about the eigenvalues of the adjacency matrix. The **chromatic** number of a graph $G = (V, E)$, denoted $\chi(G)$, is the minimal number $k$ of subsets in a partition $V = V_1 \cup \cdots \cup V_k$ such that each edge in $G$ connects vertices in two different subsets $V_i$. Such a partition is called a **$k$-coloring** of $G$, which accounts for the term "chromatic number".

In the following, note that $\operatorname{tr} A = 0$, so $\lambda_{\min}(A) < 0$.

**Theorem 9.6.** *If $G$ is a graph with at least one edge and adjacency matrix $A$, then*

$$1 + \frac{\lambda_{\max}(A)}{-\lambda_{\min}(A)} \le \chi(G) \le 1 + \lambda_{\max}(A).$$

*Proof.* We begin with the upper bound. We first put the vertices of $G$ in a convenient order.

By Proposition 9.4, $G$ has at least one vertex with degree at most $\lambda_{\max}(A)$; we choose one and designate it $v_n$. We now consider the graph $G_{n-1}$ with vertex set $V \setminus \{v_n\}$ and every edge in $G$ which does not connect to $v_n$. Its adjacency matrix $A_{n-1}$ is a submatrix of $A$, and therefore $\lambda_{\max}(A_{n-1}) \le \lambda_{\max}(A)$. Proposition 9.4 now implies that $G_{n-1}$ contains a vertex $v_{n-1}$ of degree (in $G_{n-1}$) at most $\lambda_{\max}(A)$.

Continuing in this way, we get an ordering $v_1, \ldots, v_n$ of the vertices of $G$ such that each $v_j$ is connected to at most $\lambda_{\max}(A)$ of the vertices $v_i$ with $i < j$. We can then assign colors to the vertices in order, making sure always to give each vertex a color shared by none of its neighbors, and using at most $1 + \lambda_{\max}(A)$ colors.

Now suppose that $G$ has a $k$-coloring with color classes $V_1, \ldots, V_k$. For each $1 \leq m \leq k$, let $U_m$ be the subspace of $\mathbb{R}^n$ spanned by $\{e_i \mid x_i \in V_m\}$. By assumption, if $x_i, x_j \in V_m$, we have $x_i \not\sim x_j$, and therefore $a_{ij} = \langle Ae_j, e_i \rangle = 0$. It follows that $\langle Au, u \rangle = 0$ for any $u \in U_m$ and any $m$.

Now let $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ be an eigenvector of $A$ with eigenvalue $\lambda_{\max} = \lambda_{\max}(A)$. Let $y_m = \sum_{x_i \in V_m} y_i e_i$, so that $y_m \in U_m$ and $y = \sum_{m=1}^{k} y_m$. Write $y_m = c_m u_m$ for $c_m = \|y_m\|_2$ and $u_m \in U_m$ a unit vector (note $y_m$ may be 0). Extend $u_1, \ldots, u_k$ to an orthonormal basis of $\mathbb{R}^n$ and let $U$ be the orthogonal matrix with columns $u_i$. Then the $k \times k$ upper left submatrix of $U^*AU$ is $S^*AS$, where $S \in M_{n,k}$ has columns $u_1, \ldots, u_k$. By the Cauchy interlacing principle (Theorem 3.11), $\lambda_{\min}(S^*AS) \geq \lambda_{\min}(U^*AU) = \lambda_{\min}(A)$, and $\lambda_{\max}(S^*AS) \leq \lambda_{\max}(A)$ similarly. Furthermore

$$S^*AS \begin{bmatrix} c_1 \\ \vdots \\ c_k \end{bmatrix} = S^*Ay = \lambda_{\max}(A) S^*y = \lambda_{\max}(A) \begin{bmatrix} c_1 \\ \vdots \\ c_k \end{bmatrix},$$

so in fact $\lambda_{\max}(S^*AS) = \lambda_{\max}(A)$.

Also, for each $1 \leq i \leq k$, $\langle S^*ASe_i, e_i \rangle = \langle Au_i, u_i \rangle = 0$, so $\operatorname{tr} S^*AS = 0$. It follows that

$$0 = \sum_{i=1}^{k} \lambda_i(S^*AS) = \lambda_{\max}(A) + \sum_{i=2}^{k} \lambda_i^{\downarrow}(S^*AS) \geq \lambda_{\max}(A) + (k-1)\lambda_{\min}(A),$$

which implies (since $\lambda_{\min}(A) < 0$ as noted above) that $k \geq 1 + \frac{\lambda_{\max}(A)}{-\lambda_{\min}(A)}$. □

We end this section by remarking that the practical significance of a result like Theorem 9.6 is that the chromatic number of a graph is computationally expensive to compute exactly, but there are efficient algorithms to approximate extremal eigenvalues of a symmetric matrix. Theorem 9.6 furnishes relatively easy-to-compute upper and lower bounds on an important quantity which is hard to find directly.

## 9.2 The graph Laplacian

Let $G = (V, E)$ be a graph with $V = \{x_1, \ldots, x_n\}$ as above, and let $A$ be the adjacency matrix of $G$. Define $D = \operatorname{diag}(\deg(x_1), \ldots, \deg(x_n))$. The (**nonnormalized** or **combinatorial**) **Laplacian matrix** of $G$ is the matrix $L = D - A$.

If $G$ is regular, then $D = \Delta I_n$, and the eigenvalues of $L$ are just $\Delta - \lambda_j(A)$; in that case the Laplacian is essentially an equivalent tool to the adjacency matrix. For non-regular graphs, however, the Laplacian and its variations turn out to play a more important role in applications than the adjacency matrix.

We can see that $L$ is positive semidefinite using Geršgorin's theorem (as in Exercise 3 from the March 27 homework). Moreover, $Le = De - Ae = 0$, so $L$ is not positive definite. Alternatively, we can observe these facts from the following perspective on the quadratic

form associated with $L$. Here we choose to write a vector in $\mathbb{R}^n$ as $f$, associating it with a function $f : G \to \mathbb{R}$ given by $f(x_i) = f_i$. We have

$$\begin{aligned}\langle Lf, f\rangle &= \sum_{i,j=1}^{n} (\delta_{ij}\deg(x_i) - a_{ij}) f_i f_j = \sum_{i=1}^{n} \deg(x_i) f_i^2 - \sum_{i,j:x_i\sim x_j} f_i f_j \\ &= \frac{1}{2}\sum_{i,j:x_i\sim x_j} (f_i^2 - 2f_i f_j + f_j)^2 = \sum_{\{x_i,x_j\}\in E} (f_i - f_j)^2.\end{aligned} \tag{15}$$

It follows immediately from (15) that $\langle Lf, f\rangle \geq 0$, and that $\langle Lf, f\rangle = 0$ whenever $f$ is constant. In fact, (15) implies that $\langle Lf, f\rangle = 0$ if and only if $f$ is constant on each connected component of $G$, and thus:

**Proposition 9.7.** *The multiplicity of* 0 *as an eigenvalue of* $L$ *is equal to the number of connected components of* $G$*.*

The Laplacian gets its name because it is analogous in certain ways to the Laplacian operator $\Delta = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}$ for smooth functions on $\mathbb{R}^n$, and its generalizations to Riemannian manifolds. For instance, multivariable integration by parts implies that if $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function which decays sufficiently quickly, then

$$\int_{\mathbb{R}^n} \big(\Delta f(x)\big) f(x)\ dx = -\int_{\mathbb{R}^n} \|\nabla f(x)\|_2^2\ dx,$$

which (up to the minus sign — $L$ is actually analogous to $-\Delta$) is formally similar to (15).

For the remainder of this section we will write $\lambda_1 \leq \cdots \leq \lambda_n$ for the eigenvalues of $L$. Proposition 9.7 implies that $\lambda_1 = 0$ always, and that $\lambda_2 > 0$ if and only if $G$ is connected. When $G$ is connected, $\lambda_2$ can be used to quantify how difficult it is to cut $G$ into pieces. We will need the following lemma.

**Lemma 9.8.** *If* $L$ *is the Laplacian of a graph, then*

$$\lambda_2 = \min\left\{ \frac{\langle Lf, f\rangle}{\|f\|_2^2} \;\middle|\; f \neq 0 \text{ and } \langle f, e\rangle = 0 \right\}.$$

This can be proved either using the Courant–Fischer theorem, or directly from the spectral theorem by the same method as the Courant–Fischer theorem; the key point is that $e$ is an eigenvector associated to $\lambda_1 = 0$.

Given a graph $G = (V, E)$ and a subset $S \subseteq V$, the **cut** associated to $S$ is the set $\partial S \subseteq E$ of all edges in $G$ with one end in $S$ and the other end in $V \setminus S$. In the following result, we use $|X|$ to denote the cardinality of a set $X$.

**Theorem 9.9.** *Let* $G = (V, E)$ *be a graph. For any* $S \subset V$*, we have*

$$|\partial S| \geq \lambda_2 \frac{|S|\,|V \setminus S|}{|V|}.$$

*Proof.* Let $k = |S|$, $n = |V|$, and define $f \in \mathbb{R}^n$ by

$$f_i = \begin{cases} n-k & \text{if } x_i \in S, \\ -k & \text{if } x_i \in V \setminus S. \end{cases}$$

Then $\langle f, e\rangle = 0$ and $\|f\|_2^2 = kn(n-k)$. By (15) we have

$$\langle Lf, f\rangle = n^2 \left|\partial S\right|.$$

Lemma 9.8 then implies that

$$\lambda_2 \leq \frac{\langle Lf, f\rangle}{\|f\|_2^2} = \frac{n\left|\partial U\right|}{k(n-k)}. \qquad \square$$

Many algorithmic problems can be put in the framework of finding cuts which are "efficient" in some sense. Theorem 9.9 says that $\lambda_2$ bounds how well this can be done. In particular, given any way of dividing $V$ into two subsets of roughly comparable size, there are on the order of $\lambda_2 |V|$ edges between them.

Moreover, the proof of Theorem 9.9 suggests a strategy for finding cuts with few edges: divide $V$ into two subsets according to the signs of entries of a vector which achieves (or comes close to achieving) the minimum in Lemma 9.8. That is, using an eigenvector of $L$ with eigenvalue $\lambda_2$. This turns out to be almost but not quite the right idea (except in the case when $G$ is regular). In general, we first need to replace $L$ with a slightly different notion of Laplacian.

The matrix

$$\mathcal{L} = D^{-1/2} L D^{-1/2} = I_n - D^{-1/2} A D^{-1/2}$$

is called the **normalized** or **analytic Laplacian** of $G$. (We will assume for convenience that $G$ has no isolated vertices, that is, no vertices with degree 0, but the definition can be modified easily to handle that case.) We will write $\mu_1 \leq \cdots \leq \mu_n$ for the eigenvalues of $\mathcal{L}$. Note that $D^{1/2}e \in \ker \mathcal{L}$, so $\mu_1 = 0$.

**Lemma 9.10.** *If $\mathcal{L}$ is the normalized Laplacian of a graph $G = (V, E)$, then*

$$\mu_2 = \min\left\{ \frac{\sum_{\{x_1,x_2\}\in E}(f_i - f_j)^2}{\sum_{i=1}^n f_i^2 \deg(x_i)} \;\middle|\; f \neq 0 \text{ and } \langle f, De\rangle = 0 \right\}.$$

*Proof.* As with Lemma 9.8, we have

$$\mu_2 = \min\left\{ \frac{\langle \mathcal{L}g, g\rangle}{\|g\|_2^2} \;\middle|\; g \neq 0 \text{ and } \left\langle g, D^{1/2}e\right\rangle = 0 \right\}.$$

By the definition of $\mathcal{L}$ and (15), if we substitute $f = D^{-1/2}g$,

$$\frac{\langle \mathcal{L}g, g\rangle}{\|g\|_2^2} = \frac{\left\langle LD^{-1/2}g, D^{-1/2}g\right\rangle}{\|g\|_2^2} = \frac{\langle Lf, f\rangle}{\left\|D^{1/2}f\right\|_2^2} = \frac{\sum_{\{x_1,x_2\}\in E}(f_i - f_j)^2}{\sum_{i=1}^n f_i^2 \deg(x_i)}.$$

Finally, $\left\langle g, D^{1/2}e\right\rangle = \langle f, De\rangle$. $\square$

The **volume** of a subset $S \subseteq V$ of the set of vertices of a graph is

$$\mathrm{vol}(S) = \sum_{x_i \in S} \deg(x_i).$$

Using this notion of the size of a set of vertices, we have the following analogue of Theorem 9.9.

**Theorem 9.11.** *Let $G = (V, E)$ be a graph. For any $S \subset V$, we have*

$$|\partial S| \geq \mu_2 \frac{\min\{\mathrm{vol}(S), \mathrm{vol}(V \setminus S)\}}{2}.$$

*Proof.* Define $f \in \mathbb{R}^n$ by

$$f_i = \begin{cases} \frac{1}{\mathrm{vol}(S)} & \text{if } x_i \in S, \\ -\frac{1}{\mathrm{vol}(V \setminus S)} & \text{if } x_i \in V \setminus S. \end{cases}$$

Then

$$\langle f, De \rangle = \sum_{i=1}^{n} f_i \deg(x_i) = 0,$$

$$\sum_{i=1}^{n} f_i^2 \deg(x_i) = \frac{1}{\mathrm{vol}(S)} + \frac{1}{\mathrm{vol}(V \setminus S)},$$

and

$$\sum_{\{x_1, x_2\} \in E} (f_i - f_j)^2 = \left( \frac{1}{\mathrm{vol}(S)} + \frac{1}{\mathrm{vol}(V \setminus S)} \right)^2 |\partial S|.$$

Lemma 9.10 then implies that

$$\mu_2 \leq \left( \frac{1}{\mathrm{vol}(S)} + \frac{1}{\mathrm{vol}(V \setminus S)} \right) |\partial S|$$

which proves the claim. □

The following can be viewed as a partial converse to Theorem 9.11.

**Theorem 9.12** (Cheeger's inequality)**.** *There exists a $S \subseteq V$ such that*

$$\frac{|\partial S|}{\mathrm{vol}(S)\,\mathrm{vol}(V \setminus S)} \leq \sqrt{2\mu_2}.$$

In fact one proof of Theorem 9.11 (omitted here) gives a construction for a good $S$ which is much more computationally feasible than an exhaustive search through subsets of $V$. Namely, let $f$ be an eigenvector of $\mathcal{L}$ with eigenvalue $\mu_2$. Order $V = \{v_1, \ldots, v_n\}$ so that $f(v_i)$ is nondecreasing in $i$, and let $S_i = \{v_1, \ldots, v_i\}$. Then it can be shown that

$$\min_{1 \leq i \leq n} \frac{|\partial S_i|}{\mathrm{vol}(S_i)\,\mathrm{vol}(V \setminus S_i)} \leq \sqrt{2\mu_2}.$$

Thus a near-optimal $S$ can be found by solving an eigenvector problem for $\mathcal{L}$ and then checking $n$ subsets of $V$, as opposed to the obvious brute-force approach of trying $2^n$ subsets of $V$.