Matrix Analysis and Spectral Theory

Anna Skripka and Maxim Zinchenko

UNM, MCTP, summer 2013

1 Review of linear algebra.

Definition 1.1. A **field** is a set $\mathbb{F}$ with operations of addition and multiplication on elements of $\mathbb{F}$ (called scalars) satisfying the following axioms:

(*i*) Closure of $\mathbb{F}$ under addition and multiplication.

(*ii*) Associativity of addition and multiplication.

(*iii*) Commutativity of addition and multiplication.

(*iv*) Existence of additive and multiplicative identity elements (i.e., 0 and 1).

(*v*) Existence of additive inverses and multiplicative inverses.

(*vi*) Distributivity of multiplication over addition.

Examples of fields include real numbers $\mathbb{R}$, complex numbers $\mathbb{C}$, rational numbers $\mathbb{Q}$, and finite fields $\mathbb{Z}_p$, where addition and multiplication are defined modulo a prime number p.

Definition 1.2. A **vector space** over a field $\mathbb{F}$ is a set V with two operations: 1) addition of elements of V (called vectors) and 2) multiplication of elements of V by elements of $\mathbb{F}$ (called scalars) satisfying the following axioms:

(*i*) Closure of V under addition and multiplication.

(*ii*) Associativity of addition and multiplication.

(*iii*) Commutativity of addition and multiplication.

(*iv*) Existence of an additive identity element (i.e., $\vec{0}$).

(*v*) Existence of an additive inverse.

(*vi*) Multiplicative identity $1\vec{v} = \vec{v}$ for all $\vec{v} \in V$.

(*vii*) Distributivity of multiplication over the field and vector addition.

Examples of vector spaces include spaces of n-tuples $\mathbb{F}^n$, $n \times m$ matrices $\mathbb{F}^{n\times m}$, polynomials, and functions with a common domain.

Definition 1.3. Transposition of a matrix or a vector is the operation of interchanging the rows and columns and is denoted by a superscript T. For example,

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

Hermitian adjoint (or simply adjoint) of a matrix or a vector is the transposition followed by the complex conjugation and is denoted by a superscript $*$. For example,

$$\begin{bmatrix} 1+i \\ i-3 \\ 5 \end{bmatrix}^* = \begin{bmatrix} 1-i, -i-3, 5 \end{bmatrix}.$$

Definition 1.4. An **inner product space** V is a vector space with an additional product structure called **inner product** (or scalar product) denoted by $\langle \cdot, \cdot \rangle$ which satisfies the following axioms:

(i) $\langle \vec{x}, \vec{x} \rangle \geq 0$ for all $\vec{x} \in V$ with $\langle \vec{x}, \vec{x} \rangle = 0$ if and only if $\vec{x} = \vec{0}$.

(ii) $\langle \alpha\vec{x} + \beta\vec{y}, \vec{z} \rangle = \alpha\langle \vec{x}, \vec{z} \rangle + \beta\langle \vec{y}, \vec{z} \rangle$ for all $\alpha, \beta \in \mathbb{F}$ and $\vec{x}, \vec{y}, \vec{z} \in V$.

(iii) $\langle \vec{x}, \vec{y} \rangle = \overline{\langle \vec{y}, \vec{x} \rangle}$ for all $\vec{x}, \vec{y} \in V$.

Examples of inner product spaces include $\mathbb{R}^n$, $\mathbb{C}^n$, $\mathbb{Z}_p^n$, $\mathbb{F}^n$. The **canonical inner product** in these spaces is defined by

$$\langle \vec{x}, \vec{y} \rangle = \sum_{j=1}^{n} x_j \, \overline{y_j}, \quad \text{where } \vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \ \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}. \tag{1.1}$$

Exercises 1.5. (i) Make sure that (1.1) satisfies the properties of the inner product and explain why the complex conjugation is crucial in the inner product (1.1) on $\mathbb{C}^n$, as distinct from the canonical inner product on $\mathbb{R}^n$.

(ii) Verify that

$$\langle A, B \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} \overline{b_{ij}} \tag{1.2}$$

defines an inner product on the space of $m \times n$ matrices. Conclude that $\sqrt{\langle A, A \rangle}$ defines a norm on the space on $m \times n$ matrices. The latter norm is denoted by $\|A\|_2$ and called the **Hilbert-Schmidt** (or Frobenius) **norm** of A.

Definition 1.6. A **norm** on a vector space V is a real-valued function on V denoted by $\| \cdot \|$ which satisfies the following axioms:

(i) $\|\vec{x}\| \geq 0$ for all $\vec{x} \in V$ with $\|\vec{x}\| = 0$ if and only if $\vec{x} = \vec{0}$.

(ii) $\|\alpha\vec{x}\| = |\alpha|\|\vec{x}\|$ for all $\alpha \in \mathbb{F}$ and $\vec{x} \in V$.

(iii) $\|\vec{x}+\vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$ for all $\vec{x}, \vec{y} \in V$.

Theorem 1.7 (**The Cauchy-Schwarz inequality**). *For any* $\vec{f}, \vec{g} \in V$,

$$|\langle \vec{f}, \vec{g} \rangle| \leq \|\vec{f}\| \cdot \|\vec{g}\|.$$

Theorem 1.8. *Every inner product space has the canonical norm* $\|\vec{x}\| = \sqrt{\langle \vec{x}, \vec{x} \rangle}$.

Exercises 1.9. (i) Prove that $\sqrt{\langle \vec{x}, \vec{x} \rangle}$ indeed defines a norm.

(ii) Show that the canonical norm satisfies the Parallelogram Law:

$$2\big(\|\vec{x}\|^2 + \|\vec{y}\|^2\big) = \|\vec{x}-\vec{y}\|^2 + \|\vec{x}+\vec{y}\|^2 \text{ for all } \vec{x}, \vec{y} \in V.$$

(iii) Prove the Polarization Identity for the canonical norm:

$$\langle \vec{x}, \vec{y} \rangle = \frac{1}{4}\big(\|\vec{x}+\vec{y}\|^2 - \|\vec{x}-\vec{y}\|^2 + i\|\vec{x}+i\vec{y}\|^2 - i\|\vec{x}-i\vec{y}\|^2\big) \text{ for all } \vec{x}, \vec{y} \in V.$$

Definition 1.10. (i) Vectors $\vec{v}_1, ..., \vec{v}_n$ are called **linearly independent** if and only if the only linear combination $c_1\vec{v}_1 + \cdots + c_n\vec{v}_n$ that equals $\vec{0}$ is the trivial one, that is, $c_1 = 0, ..., c_n = 0$. Otherwise, the vectors are called **linearly dependent**.

(ii) The set of all linear combinations of vectors $\vec{v}_1, ..., \vec{v}_n$ is called **span** and is denoted $\text{span}\{\vec{v}_1, ..., \vec{v}_n\}$. A collection of vectors whose span equals V is called **spanning**.

(iii) An ordered collection of vectors is called a **basis** of V if an only if the collection is linearly independent and spanning. A basis is called **orthogonal** if the basis vectors are mutually orthogonal. If, in addition, the basis vectors are all of norm 1 then the basis is called **orthonormal**. The **canonical basis** in the space $\mathbb{F}^n$ is given by

$$\vec{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ \vec{e}_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \ \vec{e}_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

(iv) The **dimension** of a vector space V is the number of elements in a basis of V. The dimension is independent of the choice of a basis.

Theorem 1.11. *An ordered collection of vectors* $B = \{\vec{v}_1, ..., \vec{v}_n\} \subset V$ *is a basis of* V *if and only if for every* $\vec{x} \in V$ *there is a unique vector* $\left[c_1, ..., c_n\right]^T \in \mathbb{F}^n$ *such that* $\vec{x} = c_1\vec{v}_1 + \cdots + c_n\vec{v}_n$*. The vector* $\left[c_1, ..., c_n\right]^T \in \mathbb{F}^n$ *is called the representation of* $\vec{x}$ *in the basis* B *and is denoted by* $[\vec{x}]_B$*. The map* $\mathrm{Rep} : V \to \mathbb{F}^n$ *given by* $\mathrm{Rep}(\vec{x}) = [\vec{x}]_B$ *is a linear invertible map. In addition, if the basis is orthonormal, then* $\langle \vec{x}, \vec{y} \rangle_V = \langle [\vec{x}]_B, [\vec{y}]_B \rangle_{\mathbb{F}^n}$ *for all* $\vec{x}, \vec{y} \in V$*.*

Definition 1.12. A subset S of a vector space V is called a **subspace** of V if and only if S is closed under addition and multiplication. Each subspace can, itself, be viewed as a vector space. The **orthogonal complement** of a subspace S in an inner product space V is the subspace $S^\perp$ given by

$$S^\perp = \{\vec{x} \in V \ : \ \langle \vec{x}, \vec{y} \rangle = 0 \text{ for all } \vec{y} \in S\}.$$

Definition 1.13. A map $L : V \to W$ from a vector space V to a vector space W is called a **linear operator** if and only if it satisfies

$$L(\alpha\vec{x} + \beta\vec{y}) = \alpha L(\vec{x}) + \beta L(\vec{y}) \text{ for all } \alpha, \beta \in \mathbb{F} \text{ and all } \vec{x}, \vec{y} \in V.$$

Each linear operator L from $\mathbb{C}^n$ to $\mathbb{C}^m$ is given as a multiplication by a matrix, that is, there is an $m \times n$ matrix A such that $L(\vec{x}) = A\vec{x}$ for every $\vec{x} \in \mathbb{C}^n$. The matrix A is called the representation of the operator L in the canonical bases of $\mathbb{C}^n$ and $\mathbb{C}^m$ and is often identified with L. The columns of the matrix A are the coordinates of the vectors $L(\vec{e}_j)$, $j = 1, ..., n$, in the canonical basis $\{\vec{e}_k\}_{k=1}^m$.

Application 1.14 (Rigid Motions in $\mathbb{R}^n$). A map $M : \mathbb{R}^n \to \mathbb{R}^n$ is called a rigid motion if and only if $\|M(\vec{x}) - M(\vec{y})\| = \|\vec{x} - \vec{y}\|$ for all $\vec{x}, \vec{y} \in \mathbb{R}^n$.

Theorem 1.15. *If* M *is a rigid motion of* $\mathbb{R}^n$ *then there is a vector* $\vec{s} \in \mathbb{R}^n$ *and a liner transformation* L *such that* $M(\vec{x}) = \vec{s} + L(\vec{x})$*.*

Proof. Let $\vec{s} = M(\vec{0})$ and $L(\vec{x}) = M(\vec{x}) - \vec{s}$. Then for all $\vec{x}, \vec{y} \in \mathbb{R}^n$ we have

$$\|L(\vec{x}) - L(\vec{y})\| = \|M(\vec{x}) - \vec{s} - M(\vec{y}) + \vec{s}\| = \|M(\vec{x}) - M(\vec{y})\| = \|\vec{x} - \vec{y}\|.$$

Hence,

$$\|L(\vec{x})\| = \|L(\vec{x}) - \vec{0}\| = \|L(\vec{x}) - L(\vec{0})\| = \|M(\vec{x}) - M(\vec{0})\| = \|\vec{x} - \vec{0}\| = \|\vec{x}\|$$

and

$$\|\vec{x}\|^2 - 2\langle \vec{x}, \vec{y} \rangle + \|\vec{y}\|^2 = \|\vec{x} - \vec{y}\|^2 = \|L(\vec{x}) - L(\vec{y})\|^2 = \|L(\vec{x})\|^2 - 2\langle L(\vec{x}), L(\vec{y}) \rangle + \|L(\vec{y})\|^2.$$

It follows that $\langle L(\vec{x}), L(\vec{y}) \rangle = \langle \vec{x}, \vec{y} \rangle$. Now let $\vec{x}, \vec{y} \in \mathbb{R}^n$ and $\vec{z} = \vec{x} + \vec{y}$, then

$$\begin{aligned}
&\|L(\vec{z}) - L(\vec{x}) - L(\vec{y})\|^2 \\
&\quad = \|L(\vec{z})\|^2 + \|L(\vec{x})\|^2 + \|L(\vec{y})\|^2 - 2\langle L(\vec{z}), L(\vec{x}) \rangle - 2\langle L(\vec{z}), L(\vec{y}) \rangle + 2\langle L(\vec{x}), L(\vec{y}) \rangle \\
&\quad = \|\vec{z}\|^2 + \|\vec{x}\|^2 + \|\vec{y}\|^2 - 2\langle \vec{z}, \vec{x} \rangle - 2\langle \vec{z}, \vec{y} \rangle + 2\langle \vec{x}, \vec{y} \rangle \\
&\quad = \|\vec{z} - \vec{x} - \vec{y}\|^2 = 0.
\end{aligned}$$

Thus, $L(\vec{x}+\vec{y}) = L(\vec{z}) = L(\vec{x}) + L(\vec{y})$. Next, let $\alpha \in \mathbb{R}$, $\vec{x} \in \mathbb{R}^n$ and $z = \alpha\vec{x}$, then

$$\begin{aligned}\|L(\vec{z}) - \alpha L(\vec{x})\|^2 &= \|L(\vec{z})\|^2 + \|\alpha L(\vec{x})\|^2 - 2\langle L(\vec{z}), \alpha L(\vec{x})\rangle \\ &= \|L(\vec{z})\|^2 + |\alpha|^2\|L(\vec{x})\|^2 - 2\alpha\langle L(\vec{z}), L(\vec{x})\rangle \\ &= \|\vec{z}\|^2 + |\alpha|^2\|\vec{x}\|^2 - 2\alpha\langle \vec{z}, \vec{x}\rangle \\ &= \|\vec{z}\|^2 + \|\alpha\vec{x}\|^2 - 2\langle \vec{z}, \alpha\vec{x}\rangle = \|\vec{z} - \alpha\vec{x}\|^2 = 0.\end{aligned}$$

Thus, $L(\alpha\vec{x}) = L(\vec{z}) = \alpha L(\vec{x})$ and it follows that L is linear. □

Definition 1.16. Each $m \times n$ matrix A is associated with three important subspaces:

(i) **Null space** $\mathrm{Nul}(A)$ is the space of all solutions of $A\vec{x} = \vec{0}$. The dimension of the null space is called **nullity** of A.

(ii) **Row space** $\mathrm{Row}(A)$ is the span of rows of A.

(iii) **Column space** $\mathrm{Col}(A)$ is the span of columns of A. The dimensions of the row and column spaces are equal and are called **rank** of A.

Similarly, for a linear operator $L : V \to W$ there are two important subspaces:

(iv) **Kernel** $\mathrm{Ker}(L) = \{\vec{x} \in V \ : \ L(\vec{x}) = \vec{0}\}$.

(v) **Range** $\mathrm{Ran}(L) = \{\vec{y} \in W \ : \ \vec{y} = L(\vec{x}) \text{ for some } \vec{x} \in V\}$.

For a linear operator L given by a multiplication by a matrix A it follows that $\mathrm{Ker}(L) = \mathrm{Nul}(A)$ and $\mathrm{Ran}(L) = \mathrm{Col}(A)$.

Theorem 1.17 (Rank-Nullity)**.** *For any $m \times n$ matrix A,*

$$\mathrm{rank}(A) + \mathrm{nullity}(A) = n.$$

Theorem 1.18 (Fundamental Subspaces)**.** *For any linear operator $A : \mathbb{C}^n \to \mathbb{C}^m$,*

$$\mathrm{Ran}(A)^\perp = \mathrm{Ker}(A^*) \quad \textit{and} \quad \mathrm{Ran}(A^*)^\perp = \mathrm{Ker}(A).$$

Definition 1.19. A matrix A is called **left invertible** if and only if there is a matrix B such that $BA = I$. A matrix A is called **right invertible** if and only if there is a matrix B such that $AB = I$. A matrix A is called **invertible** if and only if it is both left and right invertible.

Theorem 1.20. *A matrix A is left invertible if and only if the columns of A are linearly independent. A matrix A is right invertible if and only if the rows of A are linearly independent. A matrix A is invertible if and only if A is a square matrix with $det(A) \neq 0$. For square matrices A and B it follows from the Rank-Nullity Theorem that B is the left inverse of A if and only if B is the right inverse of A.*

2 Application: error correcting codes.

In real world applications, transmission and storage of data is always subject to noise. It is important, therefore, to be able to encode data beforehand in such a way that it can be decoded to its original form after noise scrambles it. One way to achieve this is by repeating a message two or three times, something very common in human speech. However, transmitting/storing multiple copies of data requires extra space and time. In this application, we will examine more efficient ways of coding. A code that detects errors in a scrambled message is called error-detecting. If, in addition, it can correct the error it is called error-correcting. It is much harder to find error-correcting than error-detecting codes. In the following we will consider messages represented by digital-sequences of 0's and 1's, for example $[1, 0, 0, 1]$. In addition, we will assume that errors do not occur too often. For simplicity, we will assume that noise can produce at most one error per message.

One way to achieve error-detection is to encode a message, say $[1, 0, 1, 1]$, by repeating it twice, such as $[1, 0, 1, 1|1, 0, 1, 1]$. Then if $[0, 0, 1, 1|1, 0, 1, 1]$ were received, we know that one of the two halves was distorted and the error is in position 1. A more efficient approach to the error-detection is the code called parity check. In this code one attaches to each message a binary tail which is 1 or 0, depending on whether we have an odd or an even number of 1's in the message (i.e., the sum of all bits in the message modulo 2). This way all encoded messages will have an even number of 1's (i.e. the sum of all bits in the encoded message is 0 modulo 2). For example, $[1, 0, 1, 1]$ will be encoded as $[1, 0, 1, 1, 1]$. Now if this is distorted to $[0, 0, 1, 1, 1]$ we know that an error has occurred, because we only received an odd number of 1's. This error-detecting code is very simple and efficient but it fails to detect multiple errors and it is not error-correcting.

Next, we consider error-correcting codes. A naive error-correcting code consists of encoding a message by repeating it three times. Then since we assumed that the transmitted message could have at most one error, we can recover the original message using the majority. For example, if $[0, 0, 1, 0|0, 1, 1, 0|0, 1, 1, 0]$ were received, we know that the first copy is distorted and the two equal copies contain the original message $[0, 1, 1, 0]$. A more efficient error-correcting code may be obtained from a combination of the previous ideas. From now on we will use addition modulo 2 and denote it by "+". Suppose $[x_1, x_2, x_3]$ is a message we would like to transmit and let $[y_1, \ldots, y_6] = [x_1, x_2, x_3, x_1 + x_2, x_2 + x_3, x_1 + x_3]$ be it's encoding. To check for errors after the transmission we look at $[z_1, z_2, z_3] = [y_1 + y_2 + y_4, y_2 + y_3 + y_5, y_1 + y_3 + y_6]$ which equal $[0, 0, 0]$ if no error, $[1, 0, 1]$ if y_1 is distorted, $[1, 1, 0]$ if y_2 is distorted, etc. The situation becomes much more transparent in the matrix notation,

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = G\vec{x} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \vec{z} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = P\vec{y} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix}$$

Since $PG = \mathrm{O} \mod 2$ we have $\vec{z} = \vec{0}$ if $\vec{y}$ has no errors. By linearity of the matrix multiplication we also have $\vec{z} = \vec{P}_j$ (the j-th column of P) if $\vec{y}$ is distorted at the j-th position (i.e., y_j is changed to $y_j + 1$). Now, since the columns of P are all distinct we conclude that $\vec{z} = \vec{P}_j$ if and only if y_j is distorted. This allows us to correct a possible error in the transmitted message $\vec{y}$ and then recover the original message $\vec{x}$.

Exercises 2.1. Find a $(4, 2)$ code that encodes 2 bits messages into 4 bits messages and allows error correction.

In the 1950, Richard Hamming introduced even more efficient error-correcting codes that became known as the Hamming codes. These codes take $2^n - 1 - n$ bits of data and encode them as $2^n - 1$ bits allowing error-correction. Next, we describe a construction of the $(7, 4)$ Hamming code. Start with a parity matrix P of height $n = 3$ containing all possible distinct nonzero columns modulo 2:

$$P = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Now we look for a matrix G with as many linearly independent columns as possible and such that $PG = 0$. It follows that the columns of G must be in the null space of P hence we can take the columns of G to be basis vectors of the null space of P,

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

The rest is as before. To encode a message $\vec{x}$ we multiply it by G, $\vec{y} = G\vec{x}$. To check for errors after the transmission we compute $P\vec{y}$ which becomes equal to the j-th column of P if and only if $\vec{y}$ is distorted at the j-th position. After the error is corrected we recover $\vec{x}$ from $\vec{y}$ via $x_k = y_k$, $k = 1, ..., 4$. (In more general situations we can recover $\vec{x}$ from $\vec{y}$ via $\vec{x} = (G^T G)^{-1} G^T \vec{y}$.)

Exercises 2.2. Find the $(15, 11)$ Hamming code.

3 Spectrum of a matrix and diagonalization.

Application 3.1 (Differential equations). Let A be an $n \times n$ and $\vec{x}_0$ an $n \times 1$ matrices with given scalar entries; let $\vec{x}(t)$ be a vector-valued function of the variable t whose entries are unknown continuously differentiable functions, such that for every t the dimension of the matrix $\vec{x}(t)$ is $n \times 1$. The solution of the matrix differential equation with initial condition

$$\begin{cases} \frac{d}{dt}\vec{x}(t) = A\vec{x}(t) \\ \vec{x}(0) = \vec{x}_0 \end{cases} \tag{3.1}$$

can be found by the formula mimicking the one for the solution of the scalar linear first order initial value problem $x'(t) = a \cdot x(t)$, $x(0) = x_0$. That is, the solution of (3.1) can be found by

$$\vec{x}(t) = e^{tA} \cdot \vec{x}_0. \tag{3.2}$$

To verify that the solution of (3.1) is given by (3.2) and to compute the matrix exponential function e^{tA}, we need to know eigenvalues, eigenvectors, and how to diagonalize matrices.

Definition 3.2. Let A be an $n \times n$ matrix. A complex number λ is called an **eigenvalue** of A if there exists a nonzero vector $\vec{x}$ (called **eigenvector** corresponding to λ) such that

$$A\,\vec{x} = \lambda\,\vec{x}. \tag{3.3}$$

The set of all eigenvalues of A is called the **spectrum** of A and is denoted by $\sigma(A)$.

To find the eigenvalues of A, we need to solve the equation $A\vec{x} = \lambda\,\vec{x}$ for λ, but the vector $\vec{x}$ is also unknown. To simplify the task, we use the standard linear algebra:

$$\begin{aligned} A\,\vec{x} = \lambda\,\vec{x},\ \vec{x} \neq \vec{0} \quad &\Leftrightarrow \quad \big(A - \lambda I\big)\,\vec{x} = \vec{0} \ \text{ has a nontrivial solution} \\ &\Leftrightarrow \quad \det(A - \lambda I) = 0. \end{aligned}$$

Thus, the eigenvalues of A are exactly the solutions of the **characteristic equation**

$$\det(A - \lambda I) = 0 \tag{3.4}$$

of the matrix A. The eigenvectors of A corresponding to an eigenvalue λ are nontrivial solutions of (3.3).

Exercises 3.3. (i) Prove that every $n \times n$ matrix has n (not necessarily distinct) eigenvalues.

(ii) Prove that the eigenvectors corresponding to an eigenvalue λ of an $n \times n$ matrix A along with $\vec{0}$ form a subspace of $\mathbb{C}^n$. (This subspace is called the **eigenspace** associated with the eigenvalue λ.)

(iii) Conclude that given an eigenvalue of a matrix, there are infinitely many corresponding eigenvectors.

(iv) Find the eigenvalues, eigenspaces, and bases of eigenspaces for each of the following matrices: (a) $\begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix}$, (b) $\begin{pmatrix} 1 & \mathrm{i} & 0 \\ -\mathrm{i} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, (c) $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$.

(v) Conclude how to find the eigenvalues of a triangular matrix quickly.

(vi) Prove that $\sigma(A) = \sigma(A^T)$. (*Hint:* use properties of the determinant.)

(vii) A matrix A is called **nilpotent** if there is a natural number n such that the product of n copies of A equals the zero matrix. Which of the matrices in 3.3(iv) is nilpotent? Prove that $\sigma(A) = \{0\}$ for any nilpotent matrix A.

Diagonal matrices are the most convenient matrices to work with. We have seen this in problems on multiplication of matrices and on computation of determinants. There are many other problems in which it is desirable to work with diagonal matrices. When a matrix is similar to a diagonal matrix (such matrix is also called diagonalizable), many problems involving this matrix can be significantly simplified via reduction to a diagonal matrix.

Definition 3.4. A square matrix A is called **diagonalizable** if there is a diagonal matrix D and an invertible matrix S such that

$$A = SDS^{-1}. \tag{3.5}$$

Theorem 3.5. *An $n \times n$ matrix A is diagonalizable if and only if A has n linearly independent eigenvectors. In this case, the diagonal elements of the matrix D in the decomposition* (3.5) *are the eigenvalues of A and the columns of S are the respective eigenvectors of A.*

Proof. For every diagonalizable matrix A, there exist a diagonal matrix D with entries $\lambda_1, \dots, \lambda_n$ and a matrix S whose columns $\vec{x_1}, \dots, \vec{x_n}$ are linearly independent and such that $AS = SD$. This implies $A\vec{x_j} = \lambda_j \vec{x_j}$, $j = 1, \dots, n$, that is, $\vec{x_j}$ is an eigenvector associated with the eigenvalue λ_j.

Suppose now that $\lambda_1, \dots, \lambda_n$ are eigenvalues of A and $\vec{x_1}, \dots, \vec{x_n}$ are the respective linearly independent eigenvectors of A. Define the matrices $D = \begin{pmatrix} \lambda_1 & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & \lambda_n \end{pmatrix}$ and $S = (\vec{x_1}, \dots, \vec{x_n})$. Multiplying out the matrices gives

$$AS = (A\vec{x_1}, \dots, A\vec{x_n}) = (\lambda_1\vec{x_1}, \dots, \lambda_n\vec{x_n}) = SD.$$

□

Exercises 3.6. (i) Diagonalize those matrices in 3.3(iv) that are diagonalizable.

(ii) Prove that no nilpotent matrix other than the zero matrix is diagonalizable.

Now we establish useful sufficient conditions for diagonalizability of a matrix.

Corollary 3.7. *If an $n \times n$ matrix A has n distinct eigenvalues, then A is diagonalizable, that is, there exists a basis of eigenvectors of A.*

Proof. We prove by contradiction that A has n linearly independent eigenvectors. Suppose that the maximal number of linearly independent eigenvectors of A is $k < n$. Enumerate those eigenvectors from 1 to k: $\vec{x_1}, \ldots, \vec{x_k}$. Let $\vec{x}_{k+1}$ be another eigenvector. Due to linear dependence, there exist complex numbers $c_1, \ldots, c_{k+1}$, not all zero, such that

$$c_1\vec{x_1} + \ldots + c_{k+1}\vec{x}_{k+1} = \vec{0}. \tag{3.6}$$

Note that the constant c_{k+1} cannot be zero because otherwise we would have had $c_i \neq 0$ for some $1 \leq i \leq k$ and $c_1\vec{x_1} + \ldots + c_k\vec{x}_k = \vec{0}$, contradicting the linear independence of $\vec{x_1}, \ldots, \vec{x_k}$. Since $c_{k+1}\vec{x}_{k+1} \neq \vec{0}$, we also have $c_1\vec{x_1} + \ldots + c_k\vec{x}_k \neq \vec{0}$ and, hence, there exists $1 \leq j \leq k$ such that $c_j \neq 0$. By applying A to (3.6) and also by multiplying (3.6) by λ_{k+1}, we obtain

$$\begin{cases} c_1\lambda_1\vec{x_1} + \ldots + c_k\lambda_k\vec{x_k} + c_{k+1}\lambda_{k+1}\vec{x}_{k+1} = \vec{0} \\ c_1\lambda_{k+1}\vec{x_1} + \ldots + c_k\lambda_{k+1}\vec{x_k} + c_{k+1}\lambda_{k+1}\vec{x}_{k+1} = \vec{0}. \end{cases} \tag{3.7}$$

Subtracting the second equation from the first one in (3.7) gives

$$c_1(\lambda_1 - \lambda_{k+1})\vec{x_1} + \ldots + c_k(\lambda_k - \lambda_{k+1})\vec{x_k} = \vec{0},$$

with $c_j(\lambda_j - \lambda_{k+1}) \neq 0$, contradicting the linear independence of $\vec{x_1}, \ldots, \vec{x_k}$. □

Remark 3.8. Let A be a linear operator on $\mathbb{C}^n$ given by a diagonalizable matrix (also denoted by A) in the canonical base of $\mathbb{C}^n$. The matrix S in (3.5) is the transition (also called the change of basis) matrix from the basis of eigenvectors of A to the canonical basis of $\mathbb{C}^n$, so $D = S^{-1}AS$ is the matrix representation of the operator A in the basis of eigenvectors of A.

Definition 3.9. The matrix obtained from a matrix A by applying complex conjugation entriwise and then taking the transpose is called **adjoint** to a matrix A and is denoted by A^*. That is, if the ij-th entry of A is a_{ij}, then the ij-th entry of A^* is $\overline{a_{ji}}$.

Definition 3.10. (i) A matrix A is called **normal** if and only if $AA^* = A^*A$ (in this case, we say "A commutes with its adjoint").
(ii) A matrix is called **self-adjoint** (or **Hermitian**) if and only if $A = A^*$.
(iii) A matrix is called **unitary** if and only if $U^* = U^{-1}$, that is, $U^*U = I = UU^*$.

Exercises 3.11. (i) Show that self-adjoint and unitary matrices are normal.

(ii) Determine which of the matrices in Exercise 3.3(iv) is normal, which is self-adjoint.

(iii) Prove that a matrix U on $\mathbb{C}^n$ is unitary if and only if $U^*U = I$ or $UU^* = I$. (*Hint:* Analyze properties of the kernel of U or U^*.)

(iv) Prove that a matrix U on $\mathbb{C}^n$ is unitary if and only if its column vectors form an orthonormal basis in $\mathbb{C}^n$.

Theorem 3.12 (Spectral theorem). *An $n \times n$ matrix A is normal if and only if A is diagonalizable and the matrix S in* (3.6) *is unitary.*

Exercises 3.13. (i) Conclude that for any normal $n \times n$ matrix A, there exists an orthonormal basis in $\mathbb{C}^n$ consisting of eigenvectors of A.

(ii) Find a nonnormal diagonalizable matrix in Exercise 3.3(iv) and explain why there is no contradiction with the spectral theorem.

(iii) Prove that if $A = A^*$, then the spectrum of A is a subset of $\mathbb{R}$. (*Hint:* determine a type of the matrix D in (3.5).)

(iv) Is it true that $\sigma(A) \subseteq \mathbb{R}$ always implies that the matrix A is self-adjoint?

(v) Let $\langle \vec{x}, \vec{y} \rangle$ denote the canonical inner product of $\vec{x} = (x_1, \ldots, x_n)^T$ and $\vec{y} = (y_1, \ldots, y_n)^T$. Verify that for any $n \times n$ matrix A, and any vectors $\vec{f}$ and $\vec{g}$ in $\mathbb{C}^n$, we have

$$\langle A\vec{f}, \vec{g} \rangle = \langle \vec{f}, A^*\vec{g} \rangle. \tag{3.8}$$

(vi) Prove that if $\lambda \neq \mu$ are eigenvalues of a self-adjoint matrix A and $\vec{f}, \vec{g} \in \mathbb{C}^n$ are such that $A\vec{f} = \lambda\vec{f}$ and $A\vec{g} = \mu\vec{g}$, then $\vec{f}$ is orthogonal to $\vec{g}$. (*Hint:* apply (3.8).)

(vii) Prove that if U is unitary, then the spectrum of U is a subset of the unit circle. (*Hint:* work with $\langle U\vec{f}, U\vec{f} \rangle$, where $\vec{f}$ is an eigenvector.)

(viii) Verify that any unitary operator maps an orthonormal system to an orthonormal system. Prove that the matrix implementing the change of any orthonormal basis to the canonical basis is unitary.

Along with the notion of the adjoint of a matrix, we also have notion of the adjoint of a linear operator.

Definition 3.14. Let L be a linear operator from $\mathbb{C}^n$ to $\mathbb{C}^n$. A linear operator L^* is called **adjoint of the operator** L if for all $\vec{f}, \vec{g} \in \mathbb{C}^n$,

$$\langle L(\vec{f}), \vec{g} \rangle = \langle \vec{f}, L^*(\vec{g}) \rangle. \tag{3.9}$$

Exercises 3.15. Let L be a linear operator.

(i) Verify that the operator L^* defined by (3.9) is linear.

(ii) Justify that if A is the matrix of L in the canonical basis, then A^* is the matrix of L^*.

Answers to selected exercises.

3.3(iv)**:** (a) $\lambda_1 = \lambda_2 = 1$, span $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \right\}$; $\lambda_3 = 2$, span $\left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right\}$.

(b) $\lambda_1 = 0$, span $\left\{ \begin{pmatrix} 1 \\ \mathrm{i} \\ 0 \end{pmatrix} \right\}$; $\lambda_2 = 1$, span $\left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$; $\lambda_3 = 2$, span $\left\{ \begin{pmatrix} 1 \\ -\mathrm{i} \\ 0 \end{pmatrix} \right\}$.

(c) $\lambda_1 = \lambda_2 = \lambda_3 = 0$, span $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right\}$.

3.3(v)**:** The eigenvalues of a triangular matrix are its diagonal entries.

3.13(ii)**:** For A in (a), the matrix S is not unitary.

4 Functions of matrices.

We start with defining the exponential of a matrix and solving the initial value problem (3.1).

Having in mind the power x^k and exponential e^x scalar functions, we define the power of a matrix A via

$$A^k = \underbrace{A \cdot \ldots \cdot A}_{k \text{ times}}, \tag{4.1}$$

and the exponential of a matrix A via

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k = I + tA + \frac{1}{2!} t^2 A^2 + \frac{1}{3!} t^3 A^3 + \ldots . \tag{4.2}$$

The series in (4.2) converges, in particular, entriwise to a matrix which we denoted by e^A. Note that we preserve the property of the exponential function $e^{\mathrm{O}} = I$.

Exercises 4.1. (i) Calculate e^A for $A = \begin{pmatrix} 0 & 3 & 0 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{pmatrix}$.

(ii) Calculate e^A for $A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 6 \end{pmatrix}$. Calculate e^D, where D is an arbitrary diagonal matrix.

(iii) Find convenient formulas for A^k, with $k \in \mathbb{N}$, and e^A, where A is a diagonalizable matrix. Apply these formulas to calculate e^A for $A = \begin{pmatrix} 1 & \mathrm{i} & 0 \\ -\mathrm{i} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

Theorem 4.2. *The function $\vec{x}(t) = e^{tA}\vec{x}_0$ is the solution to the IVP* (3.1).

Proof. We differentiate entriwise the vector-valued function $e^{tA}\vec{x}_0$ and obtain

$$\begin{aligned} \frac{d}{dt}\left(e^{tA}\vec{x}_0\right) &= \frac{d}{dt}\left(\sum_{k=0}^{\infty} \frac{1}{k!} t^k A^k \vec{x}_0\right) = \sum_{k=1}^{\infty} \frac{1}{(k-1)!} t^{k-1} A^k \vec{x}_0 = A \sum_{k=0}^{\infty} \frac{1}{k!} t^k A^k \vec{x}_0 \\ &= A e^{tA} \vec{x}_0, \end{aligned}$$

so $\vec{x}(t) = e^{tA}\vec{x}_0$ is a solution to (3.1), whose uniqueness follows from the uniqueness theorem for systems of ordinary differential equations. □

Exercises 4.3. (i) Solve the IVP (3.1), with $A = \begin{pmatrix} 1 & \mathrm{i} & 0 \\ -\mathrm{i} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $\vec{x}_0 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$.

(ii) For a self-adjoint invertible matrix A, solve the IVP

$$\begin{cases} \vec{x}\,''(t) = -A^2 \vec{x}(t) \\ \vec{x}(0) = \vec{x}_0 \\ \vec{x}\,'(0) = \vec{x}_1. \end{cases}$$

Every matrix $A = (a_{ij})_{i,j=1}^n$ can be written as a linear combination of n^2 elementary matrices $A = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} E_{ij}$. Recall that the only nonzero entry of the elementary matrix E_{ij} is the ij-th entry, which is equal to 1. A diagonal matrix $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ can be written as a linear combination of n elementary matrices

$$D = \lambda_1 E_{11} + \ldots + \lambda_n E_{nn}. \tag{4.3}$$

Every normal matrix can be written as a linear combination of n orthogonal projections. As we will see below, the latter fact is an immediate consequence of the spectral theorem.

Recall that the orthogonal projection $\vec{v}$ of a vector $\vec{u} \in \mathbb{R}^2$ onto a unit vector $\vec{x}_1 \in \mathbb{R}^2$ can be computed by the formula $\vec{v} = \langle \vec{u}, \vec{x}_1 \rangle \vec{x}_1$. Let $\vec{x}_1, \vec{x}_2, \vec{x}_3$ be an orthonormal basis in $\mathbb{R}^3$. The orthogonal projection $\vec{v}$ of $\vec{u} \in \mathbb{R}^3$ onto the plane spanned by $\vec{x}_1$ and $\vec{x}_2$ equals

$\vec{v} = \vec{u} - \langle \vec{u}, \vec{x}_3 \rangle \vec{x}_3 = \langle \vec{u}, \vec{x}_1 \rangle \vec{x}_1 + \langle \vec{u}, \vec{x}_2 \rangle \vec{x}_2$. The projection $\vec{v}$ of $\vec{u}$ onto $S = \text{span}\{\vec{x}_1, \vec{x}_2\}$ is called orthogonal because $\vec{u} - \vec{v} \in S^\perp$. Often the name "orthogonal projection" refers to a linear operator mapping $\vec{u}$ to $\vec{v}$.

Definition 4.4. Let S be a subspace in $\mathbb{C}^n$ spanned by orthonormal vectors $\vec{x}_1, \ldots, \vec{x}_k$. The linear operator P defined by

$$P(\vec{u}) = \sum_{j=1}^{k} \langle \vec{u}, \vec{x}_j \rangle \vec{x}_j, \quad \vec{u} \in \mathbb{C}^n,$$

is called the **orthogonal projection** onto S.

Exercises 4.5. (i) (Independence of the choice of a basis of S.) Verify that P is an orthogonal projection onto a subspace S of $\mathbb{C}^n$ if and only if $P(\vec{u}) \in S$ and $\vec{u} - P(\vec{u}) \in S^\perp$ for every $\vec{u} \in \mathbb{C}^n$.

(ii) Verify that any every orthogonal projection satisfies properties $P^2 = P$ and $P^* = P$.

(iii) Assume the notation of Definition 4.4 and find the matrix of an orthogonal projection P in an orthonormal basis $\{\vec{x}_1, \ldots, \vec{x}_n\}$ of $\mathbb{C}^n$.

Proposition 4.6. *Let P be a linear operator on $\mathbb{C}^n$ satisfying the properties $P^2 = P$ and $P^* = P$ and let S denote its image space $P(\mathbb{C}^n)$. Then,*

$$P(\vec{x}) = \vec{x}, \quad \textit{for every } \vec{x} \in S, \tag{4.4}$$

and

$$P(\vec{y}) = \vec{0}, \quad \textit{for every } \vec{y} \in S^\perp. \tag{4.5}$$

Proof. Let $\vec{x} \in S$. Then, there exist $\vec{y} \in \mathbb{C}^n$ such that $\vec{x} = P(\vec{y})$. Since $P^2 = P$, we have $P(\vec{x}) = P^2(\vec{y}) = P(\vec{y}) = \vec{x}$.

Since $P^* = P$ and (3.9) holds, we have that for every $\vec{x} \in S$ and every $\vec{y} \in S^\perp$, $\langle P(\vec{y}), \vec{x} \rangle = \langle \vec{y}, P(\vec{x}) \rangle = \langle \vec{y}, \vec{x} \rangle = 0$ and, hence, $P(\vec{y}) \in S^\perp$. However, $P(\mathbb{C}^n) = S$, so we also have $P(\vec{y}) \in S$. Thus, we obtain (4.5). □

Corollary 4.7. *If a linear operator P on $\mathbb{C}^n$ satisfies the properties $P^2 = P$ and $P^* = P$, then P is an orthogonal projection onto its image space $S = P(\mathbb{C}^n)$.*

Proof. We appeal to an equivalent definition of an orthogonal projection given in Exercise 4.5(i). Let $\vec{x}$ be an arbitrary vector in S. Then, by linearity of the scalar product and by (3.9),

$$\langle \vec{u} - P(\vec{u}), \vec{x} \rangle = \langle \vec{u}, \vec{x} \rangle - \langle P(\vec{u}), \vec{x} \rangle = \langle \vec{u}, \vec{x} \rangle - \langle \vec{u}, P^*(\vec{x}) \rangle,$$

which equals zero because $P^* = P$ and (4.4) holds. □

Combining Exercise 4.5(ii) and Corollary 4.7, we arrive at the following theorem.

Theorem 4.8. *A linear operator P on $\mathbb{C}^n$ is an orthogonal projection if and only if $P^2 = P$ and $P^* = P$.*

Immediately from (4.3), we derive

Proposition 4.9. *If A is a normal matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$, then for S and D from the decomposition* (3.5)*,*

$$A = \lambda_1 S E_{11} S^{-1} + \ldots + \lambda_n S E_{nn} S^{-1}. \tag{4.6}$$

Exercises 4.10. (i) Prove that the operators $P_j = SE_{jj}S^{-1}$, for S and E_{jj} as in (4.6), $1 \leq j \leq n$, are orthogonal projections. These projections are called **spectral projections** of A.

(ii) Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of a normal $n \times n$ matrix A and let $\{\vec{x}_1, \ldots, \vec{x}_n\}$ be an orthonormal basis of eigenvectors of A such that $A\vec{x}_j = \lambda_j \vec{x}_j$, $1 \leq j \leq n$. Assuming the notation of 4.10(i), prove that P_j is the orthogonal projection onto span$\{\vec{x}_j\}$, $1 \leq j \leq n$. (*Hint:* Write the matrix P_j as a linear combination of matrices whose only nonzero rows are equal to a row of S^{-1} and apply $S^{-1}S = I$.) Thus, (4.6) rewrites as

$$A = \lambda_1 P_1 + \ldots + \lambda_n P_n. \tag{4.7}$$

The representation (4.7) is a consequence of Theorem 3.12; it is often called "**spectral theorem**".

Definition 4.11. (i) Let $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Given a scalar function f defined on $\sigma(D)$, we define the function f of a matrix D as the matrix

$$f(D) = \text{diag}(f(\lambda_1), \ldots, f(\lambda_n)).$$

(ii) Let A be a normal matrix, for which we have the decomposition (3.5). Given a scalar function f defined on $\sigma(A)$, we define the function f of a matrix A by

$$f(A) = Sf(D)S^{-1}.$$

Remark 4.12. Definition 4.11 is consistent with definitions (4.1) and (4.2) (see Exercise 4.1(iii)).

Exercises 4.13. (i) Assuming the notation of Definition 4.11 and Exercise 4.10(ii), let $P_j = SE_{jj}S^{-1}$ and verify that

$$f(A) = \sum_{j=1}^{n} f(\lambda_j) P_j. \tag{4.8}$$

(ii) Verify on an example of e^A for $A = \begin{pmatrix} 1 & \mathrm{i} & 0 \\ -\mathrm{i} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ that the formula (4.8) along with the result of Exercise 4.10(ii) provides a quick way of finding a function of a normal matrix.

(iii) Verify that if A is a normal matrix, then for all functions f and g defined on $\sigma(A)$, $(f+g)(A) = f(A) + g(A)$ and $(fg)(A) = f(A)g(A)$.

(iv) (**Spectral mapping theorem** for normal matrices.) Let A be a normal matrix and f a function defined on $\sigma(A)$. Find the spectrum of $f(A)$. (*Hint:* Apply the property of the determinant $\det(BC) = \det(CB)$ or Theorem 3.5.)

(v) Let χ_R denote the **characteristic** (or **indicator**) **function of a set** $R \subset \mathbb{C}$, that is,

$$\chi_R(x) = \begin{cases} 1 & \text{if } x \in R \\ 0 & \text{if } x \notin R. \end{cases}$$

Let $R = (0.5, 3)$ and calculate $\chi_R(A)$ for $A = \begin{pmatrix} 1 & \mathrm{i} & 0 \\ -\mathrm{i} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

For an arbitrary subset R of $\mathbb{C}$ and an arbitrary normal matrix A, prove that $\chi_R(A)$ is an orthogonal projection and describe its image set.

Answers to selected exercises.

4.1(i): $\begin{pmatrix} 1 & 3 & 4.5 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}$.

4.1(ii): $\begin{pmatrix} 1 & 0 & 0 \\ 0 & e^3 & 0 \\ 0 & 0 & e^6 \end{pmatrix}$. e^D is the diagonal matrix whose diagonal entries are the exponentiated respective diagonal entries of D.

4.1(iii): For $A = SDS^{-1}$, $A^k = SD^kS^{-1}$ and $e^A = Se^DS^{-1}$.

4.3(ii): $X(t) = \cos(tA)X_0 + \sin(tA)A^{-1}X_1$.

4.5(ii): $P = \mathrm{diag}(\underbrace{1, \ldots, 1}_{k}, \underbrace{0, \ldots, 0}_{n-k})$.

4.13(iv): $f(\sigma(A))$.

4.13(v): $\chi_{(0.5,3)}(A)$ is the orthogonal projection onto span $\left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -\mathrm{i} \\ 0 \end{pmatrix} \right\}$. $\chi_R(A)$ is the sum of the orthogonal projections onto the eigenspaces corresponding to those eigenvalues of A that belong to the set R.

5 Applications to mechanics

Application 5.1 (Coupled Oscillators). First, we consider the simple case of a single harmonic oscillator consisting of a spring with the spring constant k and one end attached to a wall and the other end attached to a mass m. If $x(t)$ denotes the displacement of the mass from the equilibrium position at time t, then Newton's second law of motion together with Hooke's law give

$$m\ddot{x}(t) = -kx(t).$$

Introducing the velocity $v(t) = \dot{x}(t)$, we may rewrite the above second order differential equation as a system of first order coupled differential equations

$$\begin{bmatrix} \dot{x}(t) \\ \dot{v}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -k/m & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ v(t) \end{bmatrix}.$$

Letting $\omega_0 = \sqrt{\frac{k}{m}}$ observe that the matrix $A = \begin{bmatrix} 0 & 1 \\ -k/m & 0 \end{bmatrix}$ is diagonalized by

$$\begin{bmatrix} 0 & 1 \\ -k/m & 0 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ i\omega_0 & i\omega_0 \end{bmatrix} \begin{bmatrix} i\omega_0 & 0 \\ 0 & -i\omega_0 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ i\omega_0 & i\omega_0 \end{bmatrix}^{-1}.$$

Then changing variables to $\begin{bmatrix} y(t) \\ u(t) \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ i\omega_0 & i\omega_0 \end{bmatrix}^{-1} \begin{bmatrix} x(t) \\ v(t) \end{bmatrix}$ leads to a system of decoupled first order differential equations,

$$\begin{bmatrix} \dot{y}(t) \\ \dot{u}(t) \end{bmatrix} = \begin{bmatrix} i\omega_0 & 0 \\ 0 & -i\omega_0 \end{bmatrix} \begin{bmatrix} y(t) \\ u(t) \end{bmatrix} = \begin{bmatrix} i\omega_0 y(t) \\ -i\omega_0 u(t) \end{bmatrix},$$

whose solution is readily given by

$$\begin{bmatrix} y(t) \\ u(t) \end{bmatrix} = \begin{bmatrix} y_0 e^{i\omega_0 t} \\ u_0 e^{-i\omega_0 t} \end{bmatrix}, \text{ where } \begin{bmatrix} y_0 \\ u_0 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ i\omega_0 & i\omega_0 \end{bmatrix}^{-1} \begin{bmatrix} x_0 \\ v_0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} x_0 + \frac{v_0}{i\omega_0} \\ -x_0 + \frac{v_0}{i\omega_0} \end{bmatrix}.$$

Returning to the original variables $\begin{bmatrix} x(t) \\ v(t) \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ i\omega_0 & i\omega_0 \end{bmatrix} \begin{bmatrix} y(t) \\ u(t) \end{bmatrix}$ gives

$$\begin{bmatrix} x(t) \\ v(t) \end{bmatrix} = \begin{bmatrix} y_0 e^{i\omega_0 t} - u_0 e^{-i\omega_0 t} \\ i\omega_0 (y_0 e^{i\omega_0 t} + u_0 e^{-i\omega_0 t}) \end{bmatrix} = \begin{bmatrix} x_0 \cos(\omega_0 t) + \frac{v_0}{\omega_0} \sin(\omega_0 t) \\ -x_0 \omega_0 \sin(\omega_0 t) + v_0 \cos(\omega_0 t) \end{bmatrix}.$$

Alternatively, the solution could be obtained using the formula (3.2)

$$\begin{aligned}\begin{bmatrix} x(t) \\ v(t) \end{bmatrix} = \exp(tA) \begin{bmatrix} x_0 \\ v_0 \end{bmatrix} &= \begin{bmatrix} 1 & -1 \\ i\omega_0 & i\omega_0 \end{bmatrix} \begin{bmatrix} e^{i\omega_0 t} & 0 \\ 0 & e^{-i\omega_0 t} \end{bmatrix} \begin{bmatrix} 1 & -1 \\ i\omega_0 & i\omega_0 \end{bmatrix}^{-1} \begin{bmatrix} x_0 \\ v_0 \end{bmatrix} \\ &= \begin{bmatrix} \cos(\omega_0 t) & \frac{1}{\omega_0}\sin(\omega_0 t) \\ -\omega_0 \sin(\omega_0 t) & \cos(\omega_0 t) \end{bmatrix} \begin{bmatrix} x_0 \\ v_0 \end{bmatrix}.\end{aligned}$$

It follows that $x(t) = x_0 \cos(\omega_0 t) + \frac{v_0}{\omega_0} \sin(\omega_0 t) = \alpha \cos(\omega_0 t + \beta)$ where α, β are constants determined by the initial conditions x_0 and v_0.

Next, we consider the case of N masses $m_1, ..., m_N$ and $N+1$ springs with spring constants $k_0, ..., k_N$ connecting the masses to each other and two walls. Denoting by $x_1(t), ..., x_N(t)$ the displacements of the masses from their equilibrium positions at time t we get,

$$\begin{bmatrix} m_1 \ddot{x}_1(t) \\ m_2 \ddot{x}_2(t) \\ m_3 \ddot{x}_3(t) \\ \vdots \\ m_N \ddot{x}_N(t) \end{bmatrix} = \begin{bmatrix} -(k_0+k_1) & k_1 & 0 & 0 & \cdots & 0 \\ k_1 & -(k_1+k_2) & k_2 & 0 & \cdots & 0 \\ 0 & k_2 & -(k_2+k_3) & k_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -(k_{N-1}+k_N) \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ \vdots \\ x_N(t) \end{bmatrix}.$$

Changing variables to $y_k(t) = \sqrt{m_k} x_k(t)$, $k = 1, ..., n$, leads to

$$\begin{bmatrix} \ddot{y}_1(t) \\ \ddot{y}_2(t) \\ \ddot{y}_3(t) \\ \vdots \\ \ddot{y}_N(t) \end{bmatrix} = \begin{bmatrix} -\frac{k_0+k_1}{m_1} & \frac{k_1}{\sqrt{m_1 m_2}} & 0 & 0 & \cdots & 0 \\ \frac{k_1}{\sqrt{m_1 m_2}} & -\frac{k_1+k_2}{m_2} & \frac{k_2}{\sqrt{m_2 m_3}} & 0 & \cdots & 0 \\ 0 & \frac{k_2}{\sqrt{m_2 m_3}} & -\frac{k_2+k_3}{m_3} & \frac{k_3}{\sqrt{m_3 m_4}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\frac{k_{N-1}+k_N}{m_N} \end{bmatrix} \begin{bmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \\ \vdots \\ y_N(t) \end{bmatrix}. \qquad (5.1)$$

Next, for simplicity we assume that all masses are equal to m and all spring constants are equal to k. Then

$$\begin{bmatrix} \ddot{x}_1(t) \\ \ddot{x}_2(t) \\ \ddot{x}_3(t) \\ \vdots \\ \ddot{x}_N(t) \end{bmatrix} = \frac{k}{m} \begin{bmatrix} -2 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ \vdots \\ x_N(t) \end{bmatrix}.$$

The eigenvalues λ_k and the respective eigenvectors $\vec{u}^k$ of the above matrix are given by $\lambda_k = -\omega_k^2$, where $\omega_k = 2\omega_0 \sin\left(\frac{\pi k}{2(N+1)}\right)$, $\omega_0 = \sqrt{\frac{k}{m}}$, and $\vec{u}^k_n = \sqrt{\frac{2}{N+1}} \sin\left(\frac{\pi k n}{N+1}\right)$,

$k, n = 1, ..., N$. Setting $U = [\vec{u}^1, ..., \vec{u}^N]$ and $\vec{z}(t) = U^{-1}\vec{x}(t)$ we get

$$\begin{bmatrix} \ddot{z}_1(t) \\ \ddot{z}_2(t) \\ \ddot{z}_3(t) \\ \vdots \\ \ddot{z}_N(t) \end{bmatrix} = \begin{bmatrix} -\omega_1^2 & 0 & 0 & \cdots & 0 \\ 0 & -\omega_2^2 & 0 & \cdots & 0 \\ 0 & 0 & -\omega_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & -\omega_N^2 \end{bmatrix} \begin{bmatrix} z_1(t) \\ z_2(t) \\ z_3(t) \\ \vdots \\ z_N(t) \end{bmatrix}$$

which has the solutions $z_k(t) = \alpha_k \cos(\omega_k t + \beta_k)$, $k = 1, ..., N$, and hence

$$\vec{x}(t) = U\vec{z}(t) = \sum_{k=1}^{N} \vec{u}^k \alpha_k \cos(\omega_k t + \beta_k),$$

equivalently,

$$x_n(t) = \sum_{k=1}^{N} \alpha_k \sin\left(\frac{\pi k n}{N+1}\right) \cos(\omega_k t + \beta_k).$$

Exercises 5.2. (i) Show that the matrix in (5.1) is not changed if all the masses and spring constants are multiplied by a constant c.

(ii) Show that given a matrix as in (5.1) and if $k_0 = 0$, then it is possible to recover from the matrix entries the values of the masses $m_1, ..., m_N$ and the spring constants $k_0, ..., k_N$ up to an overall constant c.

(iii) Give an alternative solution to the system (5.1) using the Exercise 4.3(ii).

(iv) The system of coupled oscillators can be used to model the response of a tall building due to horizontal motion at the foundation generated by an earthquake. In this model assume that a building consists of N floors of mass m connected by stiff but flexible vertical walls. Restrict motion to the horizontal direction and assuming that the walls exert a flexural restoring force on the adjacent floors proportional to the relative displacement of the floors with the constant of proportionality k. Denoting by $x_1(t), ..., x_N(t)$ the displacement of the floors, relative to a fixed frame of reference at equilibrium and by $f(t)$ the horizontal displacement of the foundation due to the earthquake, write down the system of differential equations for this model. Diagonalize the matrix that you obtain for the system.

Answers to selected exercises.

5.2(iv)**:**

$$\begin{bmatrix} \ddot{x}_1(t) \\ \ddot{x}_2(t) \\ \ddot{x}_3(t) \\ \vdots \\ \ddot{x}_N(t) \end{bmatrix} = \frac{k}{m} \begin{bmatrix} -2 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ \vdots \\ x_N(t) \end{bmatrix} + \frac{k}{m} \begin{bmatrix} f(t) \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

The eigenvalues λ_k and the corresponding orthonormal eigenvectors $\vec{u}^k$ of the above matrix are given by $\lambda_k = -\omega_k^2$, where $\omega_k = 2\omega_0 \sin\left(\frac{\pi(2k-1)}{2(2N+1)}\right)$, $\omega_0 = \sqrt{\frac{k}{m}}$, and $\vec{u}_n^k = \frac{2}{\sqrt{2N+1}} \sin\left(\frac{\pi(2k-1)n}{2N+1}\right)$, $k, n = 1, ..., N$.

6 Application: Markov process.

Application 6.1 (Prediction). Consider a simple model of population migration given by the matrix $A = \begin{pmatrix} 0.95 & 0.03 \\ 0.05 & 0.97 \end{pmatrix}$, where the value of a_{11} equals the portion of the population that moves within a city every year, a_{22} the portion of the population that moves within a suburb, a_{12} the portion that moves from the suburb to city, and a_{21} the portion that moves from the city to suburb. Suppose that in 2013, the portion of the city residents was x_1 and the portion of the suburb residents was x_2. Note that $x_1 + x_2 = 1$ and let $\vec{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. Then, the distribution of the population in 2014 is given by the column $A\vec{x}$, whose first entry is the portion of the city residents and second entry is the portion of the suburb residents. In n years, the distribution of the population is given by $\underbrace{A \cdot \ldots \cdot A}_{n} \vec{x} = A^n \vec{x}$. We will find the population distribution in the long run (mathematically, as $n \to \infty$) by analyzing eigenvalues and eigenvectors of the matrix A.

The list of eigenvalues and respective eigenvectors of A is as follows: $\lambda = 1$, $\vec{u} = \begin{pmatrix} 3/8 \\ 5/8 \end{pmatrix}$; $\mu = 0.92$, $\vec{v} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. It is straightforward to see that the vectors $\vec{u}$ and $\vec{v}$ form a basis in $\mathbb{R}^2$. (This is also a consequence of Corollary 3.7.) Hence, there are $\alpha, \beta \in \mathbb{R}$ such that $\vec{x} = \alpha\vec{u} + \beta\vec{v}$. Next,

$$A^n\vec{x} = \alpha A^n\vec{u} + \beta A^n\vec{v} = \alpha\lambda^n\vec{u} + \beta\mu^n\vec{v} = \alpha\vec{u} + \beta(0.92)^n\vec{v}.$$

Since

$$\|\beta(0.92)^n\vec{v}\| = |\beta|(0.92)^n\sqrt{2} \to 0, \quad \text{as } n \to \infty,$$

we have

$$\|A^n\vec{x} - \alpha\vec{u}\| \to 0, \quad \text{as } n \to \infty.$$

Recalling that $x_1 + x_2 = 1$, we obtain $\left(\alpha \cdot \frac{3}{8} + \beta \cdot 1\right) + \left(\alpha \cdot \frac{5}{8} + \beta \cdot (-1)\right) = 1$, from what we derive that $\alpha = 1$. Therefore,

$$A^n\vec{x} \to \vec{u}, \quad \text{as } n \to \infty.$$

The latter means that, regardless of the initial distribution of the population, in the long run the distribution of the population becomes $\vec{u}$, the eigenvector of the migration matrix corresponding to the eigenvalue 1 and such that the sum of its entries equals 1. For the given $A = \begin{pmatrix} 0.95 & 0.03 \\ 0.05 & 0.97 \end{pmatrix}$, in the long run, 3/8 of the population concentrates in the city and 5/8 in the suburb.

Terminology: the vector $\vec{u}$ is called a **steady-state vector** for the Markov chain $\vec{x}, A\vec{x}, A^2\vec{x}, \ldots, A^n\vec{x}, \ldots$. The matrix A is called a transition matrix. The vector $\vec{x}$ is called a **probability vector** because its entries are nonnegative and add up to 1. Note that $\vec{u}$ and $A^n\vec{x}$, for every natural n, are also probability vectors.

To handle more general models, we need general theory of Markov processes.

Definition 6.2. (i) We say that a matrix A is **entriwise positive** if all of its entries are positive real numbers.

(ii) We say that a matrix with nonnegative entries is **stochastic** if the entries of every column add up to 1.

Definition 6.3. An eigenvalue λ of a matrix A is called **dominant** if every other eigenvalue of A is less than λ in absolute value.

Theorem 6.4 (Perron's theorem). *Every entriwise positive matrix A has a dominant eigenvalue λ with the following additional properties.*

(i) λ is positive and there is an associated eigenvector $\vec{u}$ whose all entries are positive.

(ii) The dimension of the eigenspace associated with λ is 1, that is, λ is a simple eigenvalue.

(iii) Every eigenvector corresponding to an eigenvalue other than λ has nonpositive entries.

Corollary 6.5. *(i) The dominant eigenvalue of every entriwise positive stochastic matrix is 1.*

(ii) If A is an entriwise positive stochastic matrix and $\vec{u}$ is its dominant probability eigenvector, then, for every probability vector $\vec{x}$,

$$\lim_{m\to\infty} A^m\vec{x} = \vec{u}.$$

Proof. (i) Let A be an entriwise positive stochastic matrix. By Perron's theorem, each of A and A^T has the dominant eigenvalue.

Let $\vec{h}$ denote the vector whose all entries are 1. Since every column of A, and hence, every row $\vec{r}_j$ of A^T is a probability vector, we have

$$A^T \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \langle \vec{r}_1, \vec{h} \rangle \\ \vdots \\ \langle \vec{r}_n, \vec{h} \rangle \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Thus, $\lambda = 1$ satisfies Theorem 6.4(i) and it is the dominant eigenvalue of A^T. Since $\sigma(A) = \sigma(A^T)$, $\lambda = 1$ is the eigenvalue of A and all the other eigenvalues of A are less than 1 in absolute value. Thus, $\lambda = 1$ is the dominant eigenvalue of A.

(ii) We prove this statement only for a diagonalizable matrix A. Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues and $\{\vec{v}_1, \dots, \vec{v}_n\}$ a basis of eigenvectors of A such that $A = \lambda_j \vec{v}_j$, $1 \le j \le n$, $\lambda_1 = 1$, and $\vec{v}_1 = \vec{u}$. There are constants $\alpha_1, \dots, \alpha_n$ such that

$$\vec{x} = \alpha_1 \vec{v}_1 + \alpha_2 \vec{v}_2 + \ldots + \alpha_n \vec{v}_n, \quad A^m \vec{x} = \alpha_1 \vec{u} + \alpha_2 \lambda_2^m \vec{v}_2 + \ldots + \alpha_n \lambda_n^m \vec{v}_n.$$

Since $|\lambda_j| < 1$, $2 \le j \le n$,

$$\|A^m \vec{x} - \alpha_1 \vec{u}\| \to 0, \quad \text{as } m \to \infty. \tag{6.1}$$

Now we prove that $\alpha_1 = 1$. Let $\vec{h}$ denote the vector whose all entries are 1 and recall from part (i) that $A^T \vec{h} = \vec{h}$. Consider the scalar products

$$\langle A^m \vec{x}, \vec{h} \rangle = \langle \vec{x}, (A^m)^T \vec{h} \rangle = \langle \vec{x}, (A^T)^m \vec{h} \rangle = \langle \vec{x}, \vec{h} \rangle. \tag{6.2}$$

Since $\vec{x}$ and $\vec{u}$ are probability vectors, we have $\langle \vec{x}, \vec{h} \rangle = 1$ and $\langle \vec{u}, \vec{h} \rangle = 1$. Taking limit in (6.2) gives

$$\alpha_1 = \alpha_1 \langle \vec{u}, \vec{h} \rangle = \lim_{m \to \infty} \langle A^m \vec{x}, \vec{h} \rangle = 1. \tag{6.3}$$

□

Exercises 6.6. (i) Explain where in the proof of Corollary 6.5(ii) we used diagonalizability of A.

(ii) Give an alternative proof that α_1 in the proof of Corollary 6.5(ii) equals 1 by analyzing the sum of entries of the vector $A^m \vec{x}$, for every natural m.

(iii) Prove that if $\vec{v}$ is an eigenvector of a stochastic matrix and all entries of $\vec{v}$ are nonnegative, then the eigenvalue to which $\vec{v}$ is associated equals 1.

(iv) Prove that the components of every eigenvector of a stochastic matrix A associated with an eigenvalue $\mu \ne 1$ add up to zero.

(v) In (6.3) we used the fact that (6.1) implies $\lim_{n \to \infty} \langle A^n \vec{x}, \vec{h} \rangle = \langle \lim_{n \to \infty} A^n \vec{x}, \vec{h} \rangle$. This is a consequence of the continuity of the scalar product. Prove that $\lim_{n \to \infty} \langle \vec{f}_n, \vec{g} \rangle = \langle \vec{f}, \vec{g} \rangle$, where $\vec{g}$ is an arbitrary vector and $\vec{f}_n \to \vec{f}$. (*Hint:* apply the Cauchy-Schwarz inequality.)

Application 6.7 (**Google page rank**)**.** Google search algorithm is a gigantic Markov process. If n is the number of sites in the network, then the transition matrix of the Markov process is the $n \times n$ matrix A, whose ij-th entry represents a probability that a random surfer will link from web site j to web site i. This probability is computed as follows. Suppose that a surfer follows one of the links on the current web page with probability $p \in (0, 1)$ and randomly opens any other page (including linked ones) with probability $1 - p$. Suppose that the surfer following only links on the current web page selects each of the linked pages with equal probability; thus, if k_j is the number of links from page j to other pages, then each of these linked pages is selected with probability $\dfrac{1}{k_j}$. Suppose that the probability of opening any particular random page is $\dfrac{1}{n}$; in particular, every surfer moves from a dangling page (that is, a page which has no hyperlinks to other pages) to any other page with equal probability $\dfrac{1}{n}$. Then, we obtain

$$a_{ij} = p\, m_{ij} + (1 - p)\frac{1}{n}, \tag{6.4}$$

where

$$m_{ij} = \begin{cases} \frac{1}{k_j} & \text{if there is a link from page } j \text{ to page } i \\ 0 & \text{if there is a link from page } j, \text{ but not to page } i \\ \frac{1}{n} & \text{if page } j \text{ is dangling.} \end{cases}$$

Exercises 6.8. (i) Verify that the matrix A is entriwise positive.

(ii) Verify that the matrix A with entries given by (6.4) is stochastic.

From Exercises 6.8 and Corollary 6.5, we conclude that the Markov chain with the transition matrix $A = (a_{ij})_{i,j=1}^{n}$ has a steady-state vector $\vec{u}$, which is the probability eigenvector associated with the dominant eigenvalue 1. The entries of the steady-state vector provide the page rankings. If the k-th component of the vector $\vec{u}$ is larger than its j-th component, then page k is ranked higher than page j. When a web search is conducted, the search engine first finds all sites that match all of the key words. It then lists them in decreasing order of their page ranks.

Exercises 6.9. (i) Consider the above primitive Google search algorithm for 4 sites in the network, if the first site references all the other sites, the second site references the third and fourth sites, the third site references only the first site, and the fourth site references the first and the third sites. Assume that a web surfer will follow a link on the current page 85% of the time.

Find the transition matrix and the steady-state vector.

Let $\vec{x} = (0.25, 0.25, 0.25, 0.25)^T$ and find the smallest number k for which $A^k\vec{x}$ coincides with the steady state vector in 9 decimal places.

To simplify computations, use MATLAB or Sage

```
http://aleph.sagemath.org/?lang=octave
```

(ii) Consider the two-sites network connected by the diagram $1 \longrightarrow 2$. Verify that a matrix $B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ is a natural candidate for a transition matrix judging by the probabilities involved, but that for every initial probability vector $\vec{x}$, $\lim_{m\to\infty} B^m\vec{x}$ gives no information about the relative importance of the pages 1 and 2. Find the transition matrix given by (6.4) and the steady-vector of the Markov chain.

(iii) Consider the network consisting of three sites connected by the cycle $1 \to 2 \to 3 \to 1$. Verify that $B = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ is a natural candidate for a transition matrix, but, for $\vec{x} = (1,0,0)^T$, the Markov chain $\vec{x}, B\vec{x}, B^2\vec{x}, B^3\vec{x}, \ldots$ does not converge. Find the transition matrix given by (6.4) and the steady-state vector.

(iv) The computation of $A^k\vec{x}$, where $A = (a_{ij})_{i,j=1}^n$ is given by (6.4) and $\vec{x}$ is a probability vector, can be simplified if we take advantage of the fact that many of the elements m_{ij} are equal to 0. Let $\vec{h}$ be the vector with all entries equal to 1, let $\vec{x}_1 = A\vec{x}$, and $\vec{x}_{k+1} = A\vec{x}_k$, $k \in \mathbb{N}$. Prove that

$$\vec{x}_{k+1} = pM\vec{x}_k + \frac{1-p}{n}\vec{h}, \quad \text{where } M = (m_{ij})_{i,j=1}^n.$$

Answers to selected exercises.

6.9(i): $A = \begin{pmatrix} 0.15/4 & 0.15/4 & 0.85+0.15/4 & 0.85/2+0.15/4 \\ 0.85/3+0.15/4 & 0.15/4 & 0.15/4 & 0.15/4 \\ 0.85/3+0.15/4 & 0.85/2+0.15/4 & 0.15/4 & 0.85/2+0.15/4 \\ 0.85/3+0.15/4 & 0.85/2+0.15/4 & 0.15/4 & 0.15/4 \end{pmatrix}.$

Code:

```
A=[0.0375,0.0375,0.8875,0.4625;0.3208333333,0.0375,0.0375,0.0375;
0.3208333333,0.4625,0.0375,0.4625;0.3208333333,0.4625,0.0375,0.0375];
[V,D]=eig(A);   % diagonalization of matrix A
y=V(:,[1]);   % first column from matrix V, a dominant eigenvector
y=y/norm(y,1)   % normalized vector: sum of moduli of coordinates equals 1
x=[1/4;1/4;1/4;1/4];
for n=1:26   % 26 iterations to the steady-state vector
x=A*x;
end
x
y-x
```

Steady-state vector:

```
y =
 0.368151
 0.141809
 0.287962
 0.202078
```

7 Inverse spectral problems

In this section we will discuss ways of recovering a matrix from its spectral data. We will consider two different types of spectral data. In the first problem we will take the spectral data to be the eigenvalues and the norming constants which are the first entries of each orthonormal eigenvector. In the second problem the spectral data will consist of two sets of eigenvalues - the eigenvalues of a matrix and the eigenvalues of its perturbation at a single entry.

To simplify the notation in the following, we recall the alternative form of the canonical inner product on $\mathbb{C}^N$,

$$\langle \vec{v}, \vec{u} \rangle = \sum_{n=1}^{N} \bar{u}_n v_n = \vec{u}^{\,*}\vec{v}, \quad \vec{u}^{\,*} = \begin{bmatrix} \bar{u}_1, ..., \bar{u}_N \end{bmatrix}, \ \vec{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}.$$

Application 7.1 (Direct and Inverse Spectral Problems). Consider a three diagonal self-adjoint matrix with coefficients $a_1, ..., a_{N-1} > 0$ and $b_1, ..., b_N \in \mathbb{R}$,

$$J = \begin{bmatrix} b_1 & a_1 & & & & \\ a_1 & b_2 & a_2 & & & \\ & a_2 & b_3 & a_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{N-2} & b_{N-1} & a_{N-1} \\ & & & & a_{N-1} & b_N \end{bmatrix} \tag{7.1}$$

Such matrices are called Jacobi matrices.

Exercises 7.2. Suppose A is a $N \times N$ self-adjoint matrix such that for some $\vec{x}$ the vectors $\{\vec{x}, A\vec{x}, ..., A^{N-1}\vec{x}\}$ form a basis of $\mathbb{C}^N$. Let $\{\vec{v}_1, ..., \vec{v}_N\}$ be the Gram–Schmidt orthonormalization of these vectors. Show that A has a three diagonal form (7.1) in the basis $\{\vec{v}_1, ..., \vec{v}_N\}$.

Exercises 7.3. (i) Let $P_{\vec{v}}$ be a projection in $\mathbb{C}^n$ on the subspace spanned by a vector $\vec{v} \in \mathbb{C}^n$. Show that $R_{\vec{v}} = I - 2P_{\vec{v}}$ is a unitary map that corresponds to a reflection across the plane with the normal vector $\vec{v}$.

(ii) Show that for any vector $\vec{a}$ there is a vector $\vec{v}$ such that $R_{\vec{v}}\vec{a} = \vec{e}_1$.

(iii) Derive Householder tri-diagonalization algorithm for a self-adjoint matrix A by finding a reflection R_1 such that

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & R_1 & \\ 0 & & & \end{bmatrix} A \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & R_1 & \\ 0 & & & \end{bmatrix}^* = \begin{bmatrix} b_1 & a_1 & 0 & \cdots & 0 \\ a_1 & & & & \\ 0 & & & & \\ \vdots & & & A_1 & \\ 0 & & & & \end{bmatrix}$$

and then repeating the procedure for A_1, A_2, etc.

Theorem 7.4. *Every Jacobi matrix J of the form (7.1) has N distinct real eigenvalues $\lambda_1, ..., \lambda_N$ corresponding to eigenvectors $\vec{u}^1, ..., \vec{u}^N$ with $\|\vec{u}^k\| = 1$ and $u_1^k > 0$ for all $k = 1, ..., N$. In addition, $\mu_k = u_1^k$, called norming constants, satisfy $\sum_{k=1}^N \mu_k^2 = 1$.*

Proof. Every eigenvector $\vec{u}^k$ of J corresponding to an eigenvalue λ_k satisfies,

$$\begin{aligned} b_1 u_1^k + a_1 u_2^k &= \lambda_k u_1^k \\ a_1 u_1^k + b_2 u_2^k + a_2 u_3^k &= \lambda_k u_2^k \\ &\cdots \\ a_{N-2} u_{N-2}^k + b_{N-1} u_{N-1}^k + a_{N-1} u_N^k &= \lambda_k u_{N-1}^k \\ a_{N-1} u_{N-1}^k + b_N u_N^k &= \lambda_k u_N^k. \end{aligned}$$

equivalently,

$$\begin{aligned} a_1 u_2^k &= \lambda_k u_1^k - b_1 u_1^k \\ a_2 u_3^k &= \lambda_k u_2^k - a_1 u_1^k - b_2 u_2^k \\ &\cdots \\ a_{N-1} u_N^k &= \lambda_k u_{N-1}^k - a_{N-2} u_{N-2}^k - b_{N-1} u_{N-1}^k \\ 0 &= \lambda_k u_N^k - a_{N-1} u_{N-1}^k - b_N u_N^k. \end{aligned} \tag{7.2}$$

Suppose by contradiction that $u_1^k = 0$ then it follows from the first equation that $u_2^k = 0$, then from the second that $u_3^k = 0$, etc., hence $\vec{u}^k = \vec{0}$ which is a contradiction. Thus, $u_1^k \neq 0$ for all $k = 1, ..., N$, so replacing, if necessary, the eigenvector $\vec{u}^k$ by the eigenvector $\frac{|u_1^k|}{u_1^k \|\vec{u}^k\|}\vec{u}^k$ we obtain $u_1^k > 0$ and $\|\vec{u}^k\| = 1$.

Since the Jacobi matrix J is self-adjoint, its eigenvalues are all real. Suppose by contradiction that there are less than N distinct eigenvalues. Then some eigenvalue λ_k must have multiplicity more than 1, so there are at least two linearly independent eigenvectors $\vec{u}^k$ and $\vec{v}^k$ corresponding to the eigenvalue λ_k. Then $\vec{w}^k = v_1^k \vec{u}^k - u_1^k \vec{v}^k$ is also an eigenvector corresponding to the eigenvalue λ_k, but $w_1^k = 0$ which is impossible by the considerations in the previous paragraph.

Since eigenvectors corresponding to different eigenvalues of any self-adjoint matrix are necessarily orthogonal, we have that the square matrix $U = \left[\vec{u}^1, ..., \vec{u}^N\right]$ has orthonormal columns and, hence, satisfies $U^*U = I$. Since a square matrix is left invertible if and only if it is right invertible, we also have $UU^* = I$ (see Exercise 3.11(iv)), that is, the rows of U are also orthonormal so, in particular, $\sum_{k=1}^N \mu_k^2 = \sum_{k=1}^N |u_1^k|^2 = 1$. □

Theorem 7.5. *Every Jacobi matrix J of the form (7.1) can be uniquely recovered from the spectral data consisting of distinct eigenvalues $\lambda_1, ..., \lambda_N$ and norming constants $\mu_1, ..., \mu_N > 0$ satisfying $\sum_{k=1}^N \mu_k^2 = 1$.*

Proof. We will proceed by recovering the matrix $U = \left[\vec{u}^1, ..., \vec{u}^N\right]$ row by row. The norming constants give us the first row $\vec{r}_1 = [u_1^1, ..., u_1^N] = [\mu_1, ..., \mu_N]$ of the matrix U. It follows from the first line of (7.2) that the second row $\vec{r}_2 = [u_2^1, ..., u_2^N]$ of U is given by a linear combination of $\vec{\rho}_1 = [\lambda_1 u_1^1, ..., \lambda_N u_1^N]$ and the first row $\vec{r}_1$,

$$\vec{r}_2 = \frac{1}{a_1}(\vec{\rho}_1 - b_1\vec{r}_1).$$

Since the rows of U should be orthonormal, we take the inner product with $\vec{r}_1$ to find

$$b_1 = \vec{r}_1{}^*\vec{\rho}_1$$

and then we evaluate the norm on both sides to get

$$a_1 = \|\vec{\rho}_1 - b_1\vec{r}_1\|.$$

The second line of (7.2) shows that the third row $\vec{r}_3 = [u_3^1, ..., u_3^N]$ of U is given by a linear combination of $\vec{\rho}_2 = [\lambda_1 u_2^1, ..., \lambda_N u_2^N]$ and the first and second rows $\vec{r}_1$, $\vec{r}_2$,

$$\vec{r}_3 = \frac{1}{a_2}(\vec{\rho}_2 - b_2\vec{r}_2 - a_1\vec{r}_1).$$

Again using the orthonormality of the rows of U, we get

$$b_2 = \vec{r}_2{}^*\vec{\rho}_2$$
$$a_2 = \|\vec{\rho}_2 - b_2\vec{r}_2 - a_1\vec{r}_1\|.$$

Proceeding similarly, we recover all the coefficients $a_1, ..., a_{N-1}$ and $b_1, ..., b_N$. □

Exercises 7.6. Verify that $\left[\mu_1, ..., \mu_N\right]$, $\left[\lambda_1\mu_1, ..., \lambda_N\mu_N\right]$, ..., $\left[\lambda_1^{N-1}\mu_1, ..., \lambda_N^{N-1}\mu_N\right]$ are linearly independent. Conclude that in the above construction all the coefficients a_n are going to be strictly positive.

Application 7.7 (Two Spectra Inverse Problem). Let c be a fixed nonzero real number and suppose J is a Jacobi matrix as in (7.1) and $\tilde{J}$ is a perturbed Jacobi matrix given by $\tilde{J} = J + c\vec{e}_1\vec{e}_1{}^*$, that is, $\tilde{J}$ has the same coefficients as J except $\tilde{b}_1 = b_1 + c$.

Theorem 7.8. *Given the constant c and the spectra of J and $\tilde{J}$ one can uniquely recover the coefficients of the Jacobi matrices J and $\tilde{J}$.*

Proof. We start by introducing the function $m(z) = \vec{e_1}^*(J-z)^{-1}\vec{e}_1$, called the Weyl–Titchmarsh m-function. As before let $U = \left[\vec{u}^1, ..., \vec{u}^N\right]$ be the unitary matrix whose columns are the eigenvectors of J corresponding to the eigenvalues $\lambda_1, ..., \lambda_N$ and $\mu_k = u_1^k > 0$, $k = 1, ..., N$, be the norming constants. Then $(J-z)^{-1}$ is the function $(x-z)^{-1}$ of the matrix J, so

$$(J-z)^{-1} = U\mathrm{diag}\Big(\frac{1}{\lambda_1 - z}, ..., \frac{1}{\lambda_N - z}\Big)U^*$$

and, hence,

$$m(z) = (U^*\vec{e}_1)^*\mathrm{diag}\Big(\frac{1}{\lambda_1 - z}, ..., \frac{1}{\lambda_N - z}\Big)(U^*\vec{e}_1) = \sum_{k=1}^{N} \frac{|u_1^k|^2}{\lambda_k - z} = \sum_{k=1}^{N} \frac{\mu_k^2}{\lambda_k - z}. \qquad (7.3)$$

It follows that $m(z)$ is a rational function with poles precisely at the eigenvalues of J and such that $m(z) \to 0$ as $z \to \infty$. Similarly, let $\tilde{m}(z)$ be the Weyl–Titchmarsh m-function associated with $\tilde{J}$ that has poles at the eigenvalues of $\tilde{J}$. Next, we seek a formula connecting $m(z)$ and $\tilde{m}(z)$. Consider the identity

$$(\tilde{J} - z)^{-1} = (J + c\vec{e}_1\vec{e_1}^* - z)^{-1} = (I + (J-z)^{-1}c\vec{e}_1\vec{e_1}^*)^{-1}(J-z)^{-1}$$

which is equivalent to

$$(I + c(J-z)^{-1}\vec{e}_1\vec{e_1}^*)(\tilde{J} - z)^{-1} = (J-z)^{-1}.$$

Multiplying both sides by $\vec{e}_1$ on the right and by $\vec{e_1}^*$ on the left gives,

$$(1 + cm(z))\tilde{m}(z) = m(z).$$

Solving for $m(z)$, we get $m(z) = (1/\tilde{m}(z) - c)^{-1}$, so at the eigenvalues of $\tilde{J}$ the function $m(z)$ takes the value $-c^{-1}$ since $\tilde{m}(z)$ blows up at these points. Thus, the function $m(z) + c^{-1}$ is a rational function with poles at $\lambda_1, ..., \lambda_N$ and zeros at $\tilde{\lambda}_1, ..., \tilde{\lambda}_N$,

$$m(z) + \frac{1}{c} = d\frac{(\tilde{\lambda}_1 - z)\cdots(\tilde{\lambda}_N - z)}{(\lambda_1 - z)\cdots(\lambda_N - z)}.$$

Taking limit as $z \to \infty$, we get $d = c^{-1}$, so the function $m(z)$ is uniquely determined by the constant c and the eigenvalues of J and $\tilde{J}$,

$$m(z) = \frac{1}{c}\left[\frac{(\tilde{\lambda}_1 - z)\cdots(\tilde{\lambda}_N - z)}{(\lambda_1 - z)\cdots(\lambda_N - z)} - 1\right].$$

Finally, performing partial fraction decomposition and comparing with (7.3) gives,

$$\mu_k^2 = \frac{1}{c}\frac{(\tilde{\lambda}_1 - \lambda_k)\cdots(\tilde{\lambda}_N - \lambda_k)}{(\lambda_1 - \lambda_k)\cdots(\lambda_{k-1} - \lambda_k)(\lambda_{k+1} - \lambda_k)\cdots(\lambda_N - \lambda_k)}.$$

Thus, the norming constants $\mu_1, ..., \mu_N$ are also uniquely determined by the constant c and the eigenvalues of J and $\tilde{J}$. The Inverse Spectral Problem discussed above now allows us to recover the coefficients of the matrix J from its eigenvalues $\lambda_1, ..., \lambda_N$ and norming constants $\mu_1, ..., \mu_N$. □

Exercises 7.9. By considering the limit $zm(z)$ as $z \to \infty$ show that the a priory knowledge of the constant c is not necessary in Theorem 7.8 because the value of c can be recovered from the eigenvalues by the formula

$$c = \sum_{k=1}^{N} (\tilde{\lambda}_k - \lambda_k). \tag{7.4}$$

Definition 7.10. The **trace** of a square matrix $A = (a_{ij})_{i,j=1}^n$ is the sum of the diagonal entries of A, that is,

$$\mathrm{tr}(A) = \sum_{j=1}^{n} a_{jj}.$$

Important properties of the trace are established in the following exercises.

Exercises 7.11. (i) Verify the linearity: $\mathrm{tr}(\alpha A + \beta B) = \alpha\,\mathrm{tr}(A) + \beta\,\mathrm{tr}(B)$ for any $\alpha, \beta \in \mathbb{C}$ and any square matrices A and B of the same dimension.

(ii) Verify the commutativity of the trace: $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for arbitrary $n \times m$ matrix A and $m \times n$ matrix B.

(iii) Verify the **cyclicity** of the trace: $\mathrm{tr}(ABC) = \mathrm{tr}(CAB)$ for arbitrary square matrices A, B, and C of the same dimension.

(iv) Prove that if the matrices A and B are similar, that is, there is an invertible matrix C such that $A = CBC^{-1}$, then $\mathrm{tr}(A) = \mathrm{tr}(B)$.

(v) (**Lidskii's theorem** for normal matrices.) Prove that for a normal matrix A with eigenvalues $\lambda_1, \ldots, \lambda_n$,

$$\mathrm{tr}(A) = \sum_{j=1}^{n} \lambda_j. \tag{7.5}$$

(vi) Given an orthonormal basis $\{\vec{x}_1, \ldots, \vec{x}_n\}$ in $\mathbb{C}^n$ and an $n \times n$ matrix A, prove

$$\mathrm{tr}(A) = \sum_{j=1}^{n} \langle A\vec{x}_j, \vec{x}_j \rangle.$$

(*Hint:* Change to the canonical basis.)

Exercises 7.12. Use the properties of the trace to derive the formula (7.4).

Exercises 7.13. Given a mass-spring system as in (5.1) suppose one performed an experiment and obtained the frequencies of vibration $\omega_1, ..., \omega_n$. Suppose, in addition, that one then managed to remove the spring k_0, measured the value of k_0, and performed a second experiment and determined the frequencies $\omega_1, ..., \omega_n$ for the mass spring system with the spring k_0 removed. Show that the above experimental data is enough to recover the values of the remaining spring constants $k_1, ..., k_N$ and the masses $m_1, ..., m_N$.

8 Toda lattice

In this section we consider a nonlinear system of coupled oscillators. Consider N unit masses connected to each other by nonlinear springs that exert the following force when stretched by x units,

$$F(x) = e^{-x} - 1 = -x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots .$$

Denoting by $x_1(t)$,...,$x_N(t)$ the displacements of the masses from the equilibrium at time t, we obtain the system of N nonlinear ordinary differential equations

$$\ddot{x}_n(t) = F\big(x_n(t) - x_{n-1}(t)\big) - F\big(x_{n+1}(t) - x_n(t)\big), \quad n = 1, ..., N,$$

equivalently,

$$\ddot{x}_n(t) = e^{x_{n-1}(t) - x_n(t)} - e^{x_n(t) - x_{n+1}(t)}, \quad n = 1, ..., N, \tag{8.1}$$

where we set $x_0(t) = -\infty$ and $x_{N+1} = +\infty$ for simplicity of notation. Changing variables to Flaschke's variables:

$$a_n(t) = \frac{1}{2} e^{(x_n(t) - x_{n+1}(t))/2} \quad \text{and} \quad b_n(t) = -\frac{1}{2} \dot{x}_n(t) \tag{8.2}$$

with $a_0(t) = a_N(t) = 0$, the system of differential equations (8.1) rewrites as

$$\begin{aligned} \dot{a}_n(t) &= a_n(t)[b_{n+1}(t) - b_n(t)], \quad n = 1, ..., N-1, \\ \dot{b}_n(t) &= 2[a_n^2(t) - a_{n-1}^2(t)], \quad n = 1, ..., N, \end{aligned}$$

equivalently,

$$\frac{d}{dt} J(t) = P(t)J(t) - J(t)P(t), \tag{8.3}$$

where

$$J(t) = \begin{bmatrix} b_1(t) & a_1(t) & & & & \\ a_1(t) & b_2(t) & a_2(t) & & & \\ & a_2(t) & b_3(t) & a_3(t) & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{N-2}(t) & b_{N-1}(t) & a_{N-1}(t) \\ & & & & a_{N-1}(t) & b_N(t) \end{bmatrix}, \tag{8.4}$$

$$P(t) = \begin{bmatrix} 0 & a_1(t) & & & & \\ -a_1(t) & 0 & a_2(t) & & & \\ & -a_2(t) & 0 & a_3(t) & & \\ & & \ddots & \ddots & \ddots & \\ & & & -a_{N-2}(t) & 0 & a_{N-1}(t) \\ & & & & -a_{N-1}(t) & 0 \end{bmatrix}. \tag{8.5}$$

Let $U(t)$ be the solution of the initial value problem,

$$\frac{d}{dt}U(t) = P(t)U(t), \quad U(0) = I. \tag{8.6}$$

Then, since $P(t)^* = -P(t)$, it follows that,

$$\frac{d}{dt}[U^*(t)U(t)] = [P(t)U(t)]^*U(t) + U^*(t)[P(t)U(t)] = U^*(t)[P^*(t) + P(t)]U(t) = 0,$$

so $U^*(t)U(t) = U^*(0)U(0) = I$, that is, $U(t)$ is unitary. Moreover, a straightforward substitution shows that

$$J(t) = U(t)J(0)U^{-1}(t) \tag{8.7}$$

solves (8.3), hence the eigenvalues $\lambda_1, ..., \lambda_N$ of $J(t)$ are time independent. Thus, to find the solution $J(t)$, it remains to find the time evolution of the norming constants. Let $\vec{u}^1(t), ..., \vec{u}^N(t)$ be the orthonormal eigenvectors of $J(t)$ corresponding to the eigenvalues $\lambda_1, ..., \lambda_N$. Then, by (8.7) and (8.6), we get

$$\vec{u}^k(t) = U(t)\vec{u}^k(0) \quad \Leftrightarrow \quad \frac{d}{dt}\vec{u}^k(t) = P(t)\vec{u}^k(t)$$

and, by (8.5) and (7.2),

$$\frac{d}{dt}u_1^k(t) = a_1(t)u_2^k = [\lambda_k - b_1(t)]u_1^k(t)$$

so

$$u_1^k(t) = u_1^k(0)e^{\lambda_k t}e^{-\int_0^t b_1(s)ds}, \quad k = 1, ..., N.$$

Let $f(t) = e^{-\int_0^t b_1(s)ds}$. Then, recalling that the norming constants of $J(t)$ are given by $\mu_k(t) = u_1^k(t) > 0$ and satisfy $\sum_{k=1}^N \mu_k^2(t) = 1$ as discussed in Theorem 7.4, we get

$$\mu_k(t) = \mu_k(0)e^{\lambda_k t} f(t), \quad k = 1, ..., N,$$

and

$$\sum_{k=1}^N \mu_k^2(0)e^{2t\lambda_k} f^2(t) = 1 \quad \Leftrightarrow \quad f(t) = \left(\sum_{k=1}^N \mu_k^2(0)e^{2t\lambda_k}\right)^{-1/2}.$$

Thus,

$$\mu_k(t) = \frac{\mu_k(0)e^{t\lambda_k}}{\sqrt{\sum_{k=1}^N \mu_k^2(0)e^{2t\lambda_k}}}, \quad k = 1, ..., N.$$

By solving the inverse spectral problem, we can reconstruct $J(t)$ from the eigenvalues $\lambda_1, ..., \lambda_N$ and the norming constants $\mu_1(t), ..., \mu_N(t)$ and, subsequently, recover the solution $x_1(t), ..., x_N(t)$ from (8.2).

9 Singular value decomposition

Application 9.1 (**Digital image compression).** A photograph can be digitized by breaking it up into a rectangular array of cells (pixels) and measuring the gray level of each cell. The gray levels of any cell is generally close to the gray level of its neighboring cells, so it is possible to reduce the amount of storage. One way to implement this is using a singular values decomposition of the matrix that models the photograph. Singular values provide a measure of how close A is to a matrix of lower rank.

Definition 9.2. Eigenvalues of $|A| = \sqrt{A^*A}$ are called the **singular values** of a matrix A. That is, if $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A^*A, then $\sigma_1 = \sqrt{\lambda_1}, \sigma_2 = \sqrt{\lambda_2}, \ldots, \sigma_n = \sqrt{\lambda_n}$ are the singular values of A.

Exercises 9.3. Prove that for any matrix A, the eigenvalues of A^*A are nonnegative.

Theorem 9.4. *Let r be the rank of an $m \times n$ matrix A and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$ the singular values of the matrix A. The following assertions hold.*

(i) The number of nonzero singular values of A equals r.

(ii) If $\{\vec{v}_1, \ldots, \vec{v}_n\}$ is an orthonormal basis consisting of the eigenvectors of A^*A such that $A^*A\vec{v}_j = \sigma_j^2\vec{v}_j$, $1 \leq j \leq n$, and let $\vec{u}_j = \dfrac{1}{\sigma_j}A\vec{v}_j$, $1 \leq j \leq r$, then

$$A = \sum_{j=1}^r (A\vec{v}_j)\vec{v}_j^* = \sum_{j=1}^r \sigma_j \vec{u}_j \vec{v}_j^*. \tag{9.1}$$

Proof. (i) Firstly, we prove that $\operatorname{rank}(A) = \operatorname{rank}(A^*A)$. Applying the rank-nullity theorem to matrices A and A^*A gives

$$\operatorname{rank}(A) = n - \operatorname{nullity}(A), \quad \operatorname{rank}(A^*A) = n - \operatorname{nullity}(A^*A).$$

Hence, it is enough to establish that $\operatorname{nullity}(A) = \operatorname{nullity}(A^*A)$. Clearly, the null space of the matrix A is a subspace of the null space of A^*A. Conversely, suppose that $A^*A\vec{x} = \vec{0}$. Then, $0 = \langle A^*A\vec{x}, \vec{x} \rangle = \langle A\vec{x}, A\vec{x} \rangle = \|A\vec{x}\|^2$, which implies $A\vec{x} = \vec{0}$. Thus, $\operatorname{Nul}(A) = \operatorname{Nul}(A^*A)$.

Since A^*A is self-adjoint, it can be factored into $A^*A = SDS^{-1}$, where $D = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ and S is a unitary matrix. The rank of D, which is also the dimension of the range of the linear operator D, equals r. By Exercise 3.13(viii), a unitary operator preserves the dimension of a subspace to which it is applied, so the dimension of the range of A^*A is also r.

(ii) Since

$$\sum_{j=1}^{r} \sigma_j \vec{u}_j \vec{v}_j^* \vec{v}_k = \begin{cases} \sigma_k \vec{u}_k = A\vec{v}_k & \text{if } k \le r \\ 0 & \text{if } k > r, \end{cases}$$

we obtain that the left and right hand sides of (9.1) coincide on every vector in $\mathbb{C}^n$.

□

Exercises 9.5. (i) In case A is a normal matrix, relate (9.1) to the representation (4.7) provided by the spectral theorem.

(ii) Prove that $\{\vec{u}_1, \ldots, \vec{u}_r\}$ is an orthonormal system.

Returning to Application 9.1, we can approximate the matrix A by truncating the sum in (9.1):

$$A_k = \sum_{j=1}^{k} \sigma_j \vec{u}_j \vec{v}_j^*, \tag{9.2}$$

which requires $k(m + n + 1)$ cells because each vector $\vec{u}_j$ has m entries and each vector $\vec{v}_j$. The value of k is determined by the needed accuracy, which, as we will see below, can be measured by $(\sigma_{k+1}^2 + \ldots + \sigma_r^2)^{1/2}$. (By noticing that $\vec{v}_j$ has n entries and $\|\vec{v}_j\| = 1$ and agreeing that certain, for instance, the first, component of each $\vec{v}_j$ is positive, we could reduce the number of cells to $k(m + n - 1 + 1) = km$. Furthermore, if we use the representation $A = \sum_{j=1}^{r} (A\vec{v}_j)\vec{v}_j^*$, we can reduce to $k(m + n - 1)$ cells.)

Exercises 9.6. (i) Verify that $\|A - A_k\|_2 = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij} - a_{ij,k}|^2 \right)^{\frac{1}{2}}$, where $a_{ij,k}$ is the ij-th entry of the matrix A_k.

(ii) Verify that $\operatorname{tr}(B^*A) = \langle A, B\rangle$ and that $\|A\|_2 = \left(\sum_{j=1}^{n} \sigma_j^2\right)^{1/2}$, where the inner product of matrices is given by (1.2).

(iii) Verify that the matrix A_k given by (9.2) satisfies

$$\|A_k\|_2 = \left(\sum_{j=1}^{k} \sigma_j^2\right)^{1/2}, \qquad \|A - A_k\|_2 = \left(\sum_{j=k+1}^{r} \sigma_j^2\right)^{1/2}.$$

(iv) Find the rank and singular values of A_k.

Definition 9.7. Let A be an $m \times n$ matrix. We say that the matrix A_k of rank k is closest to A in the Hilbert-Schmidt norm if it satisfies

$$\|A - A_k\|_2 = \min\left\{\|A - X\|_2 : X \text{ is an } m \times n \text{ matrix of rank } k\right\}.$$

Theorem 9.8. *Given an $m \times n$ matrix A of rank r with nonzero singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$, the matrix A_k of rank $k \leq r$ closest to A in the Hilbert-Schmidt norm is given by* (9.2)*.*

The decomposition (9.1) can be rewritten in the matrix form called a singular value decomposition of A.

Definition 9.9. Let A be an $m \times n$ matrix. If

$$A = U\Sigma V^*,$$

where U is an $m \times m$ and V is an $n \times n$ unitary matrices, and Σ is an $m \times n$ matrix whose off-diagonal entries are all zeros and whose "diagonal" entries are the singular values of A written in the decreasing order

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0,$$

where $r = \operatorname{rank}(A)$, then the factorization $U\Sigma V^*$ is called a **singular value decomposition of A.**

Corollary 9.10. *Every matrix has a singular value decomposition.*

The proof of Corollary 9.10 is based on the following exercise.

Exercises 9.11. Assume the notations of Theorem 9.4. Show that A admits the reduced singular value decomposition $A = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^*$, where $\widetilde{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$, $\widetilde{V}$ is the matrix with columns $\vec{v}_1, \ldots, \vec{v}_r$ and $\widetilde{U}$ is the matrix with columns $\vec{u}_1, \ldots, \vec{u}_r$.

Proof of Corollary 9.10. We assume the notations of Theorem 9.4 and complete the system $\{\vec{u}_1, \dots, \vec{u}_r\}$ to an orthonormal basis $\{\vec{u}_1, \dots, \vec{u}_m\}$. It is possible to do the latter due to the result of Exercise 9.5(ii). Let

$$\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_n) = \operatorname{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0),$$

V be the matrix with columns $\vec{v}_1, \dots, \vec{v}_n$, and U the matrix with columns $\vec{u}_1, \dots, \vec{u}_m$. Performing block multiplication of matrices and applying Exercise 9.11 completes the proof. □

Exercises 9.12. (i) Find the rank, singular values, the representation (9.1), and the singular value decomposition for $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{pmatrix}$.

(ii) Let $A = \begin{pmatrix} -0.02 & 0.08 & 0.2 \\ 0.14 & 0.19 & 0.1 \\ 0.02 & -0.02 & 0.01 \end{pmatrix}$ be factored in the singular value decomposition form

$$A = \begin{pmatrix} 3/5 & -4/5 & 0 \\ 4/5 & 3/5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0.3 & 0 & 0 \\ 0 & 0.15 & 0 \\ 0 & 0 & 0.03 \end{pmatrix} \begin{pmatrix} 1/3 & 2/3 & 2/3 \\ 2/3 & 1/3 & -2/3 \\ 2/3 & -2/3 & 1/3 \end{pmatrix}.$$

Find the rank and singular values of A as well as the closest (in the Hilbert-Schmidt norm) to A matrices A_1 of rank 1 and A_2 of rank 2. What is the total storage needed in each of the cases? How close are the matrices?

Answers to selected exercises.

9.12(ii): $A_1 = \begin{pmatrix} 0.06 & 0.12 & 0.12 \\ 0.08 & 0.16 & 0.16 \\ 0 & 0 & 0 \end{pmatrix}$, storage $= 5$, distance ≈ 0.153;

$A_2 = \begin{pmatrix} -0.02 & 0.08 & 0.2 \\ 0.14 & 0.19 & 0.1 \\ 0 & 0 & 0 \end{pmatrix}$, storage $= 10$, distance $= 0.03$.

10 Compact operators

In this section we consider linear operators on infinite dimensional Hilbert spaces.

Definition 10.1. A sequence of vectors $\{\vec{x}_n\}_{n=1}^{\infty}$ in a normed vector space V is called a **Cauchy sequence** if and only if $\|\vec{x}_n - \vec{x}_k\| \to 0$ as $n, k \to \infty$.

Exercises 10.2. Show that every convergent sequence is a Cauchy sequence.

Definition 10.3. An inner product space V in which every Cauchy sequence is convergent is called a **Hilbert space**.

An example of an infinite dimensional Hilbert space is the space of square summable sequences $\ell^2 = \left\{\vec{x} = (x_n)_{n=1}^{\infty} \ : \ x_n \in \mathbb{C} \text{ for all } n \text{ and } \sum_{n=1}^{\infty} |x_n|^2 < \infty\right\}$. The canonical inner product on this space is defined by the familiar formula

$$\langle \vec{x}, \vec{y} \rangle = \sum_{n=1}^{\infty} x_n \overline{y_n}, \quad \vec{x}, \vec{y} \in \ell^2.$$

Definition 10.4. An **orthonormal basis** $\{\vec{b}_n\}_{n=1}^{\infty}$ of a Hilbert space V is a sequence of mutually orthogonal vectors of norm 1 that satisfy

$$\sum_{n=1}^{\infty} |\langle \vec{x}, \vec{b}_n \rangle|^2 = \|\vec{x}\|^2 \ \text{ for all } \ \vec{x} \in V,$$

equivalently,

$$\sum_{n=1}^{\infty} \langle \vec{x}, \vec{b}_n \rangle \vec{b}_n = \vec{x} \ \text{ for all } \ \vec{x} \in V.$$

Definition 10.5. A linear operator A is called **bounded** if and only if the **operator norm** of A defined by

$$\|A\| = \sup_{\|\vec{v}\|=1} \|A\vec{v}\|$$

is finite. A linear operator A is called **compact** if and only if for every bounded sequence of vectors $\{\vec{v}_n\}_{n=1}^{\infty}$, the sequence $\{A\vec{v}_n\}_{n=1}^{\infty}$ has a convergent subsequence.

Exercises 10.6. (i) Show that every finite rank operator, that is, an operator with finite dimensional range is compact.

(ii) Show that every compact operator is a bounded operator.

(iii) Show that every bounded operator is a continuous map.

(iv) Show that a composition of a bounded operator and a compact one is compact.

(v) Show that a square of a compact operator is compact.

(vi) Show that self-adjoint operators can have only real eigenvalues.

(vii) Show that eigenvectors corresponding to different eigenvalues of a self-adjoint operator are necessarily orthogonal.

Theorem 10.7. *Suppose A is an operator and $\{A_n\}_{n=1}^{\infty}$ are compact operators such that $\|A - A_n\| \to 0$ as $n \to \infty$, then A is also a compact operator. In particular, if A can be approximated arbitrarily closely by finite rank operators in the operator norm, then A is compact.*

Proof. Let $\{\vec{v}_n\}_{n=1}^{\infty}$ be such that $\|\vec{v}_n\| \le C$ for all n. Since A_1 is compact there is a subsequence $\{\vec{v}_{1,n}\}_{n=1}^{\infty}$ of $\{\vec{v}_n\}_{n=1}^{\infty}$ such that $A_1\vec{v}_{1,n}$ is convergent as $n \to \infty$. Similarly, since A_2 is compact there is a subsequence $\{\vec{v}_{2,n}\}_{n=1}^{\infty}$ of $\{\vec{v}_{1,n}\}_{n=1}^{\infty}$ such that $A_2\vec{v}_{2,n}$ is convergent as $n \to \infty$. Proceeding this way we get for each $j = 1, 2...$, a subsequence $\{\vec{v}_{j,n}\}_{n=1}^{\infty}$ of $\{\vec{v}_{j-1,n}\}_{n=1}^{\infty}$ so that $A_j\vec{v}_{j,n}$ is convergent as $n \to \infty$.

Now consider the "diagonal" subsequence $\{\vec{v}_{n,n}\}_{n=1}^{\infty}$ which is a subsequence of the original sequence $\{\vec{v}_n\}_{n=1}^{\infty}$. By construction, $A_j\vec{v}_{n,n}$ is convergent as $n \to \infty$ and hence is Cauchy for every j. Therefore,

$$\begin{aligned}\|A\vec{v}_{n,n} - A\vec{v}_{k,k}\| &\le \|(A - A_j)\vec{v}_{n,n}\| + \|A_j\vec{v}_{n,n} - A_j\vec{v}_{k,k}\| + \|(A_j - A)\vec{v}_{k,k}\| \\ &\le \|A - A_j\|C + \|A_j\vec{v}_{n,n} - A_j\vec{v}_{k,k}\| + \|A_j - A\|C,\end{aligned}$$

implies $\lim_{n,k\to\infty} \|A\vec{v}_{n,n} - A\vec{v}_{k,k}\| \le 2C\|A - A_j\|$. Since j was arbitrary, we may also take $j \to \infty$ to obtain $\lim_{n,k\to\infty} \|A\vec{v}_{n,n} - A\vec{v}_{k,k}\| = 0$, that is, $\{A\vec{v}_{n,n}\}_{n=1}^{\infty}$ is Cauchy and hence is convergent. □

Definition 10.8. An operator A is called **Hilbert–Schmidt** if and only if there is an orthonormal basis $\{\vec{b}_n\}_{n=1}^{\infty}$ such that $\sum_{n=1}^{\infty} \|A\vec{b}_n\|^2 < \infty$.

Theorem 10.9. *Every Hilbert–Schmidt operator is compact.*

Proof. Let $\{\vec{b}_n\}_{n=1}^{\infty}$ be an orthonormal basis such that $\sum_{n=1}^{\infty} \|A\vec{b}_n\|^2 < \infty$. For each k define a finite rank operator A_k by

$$A_k\vec{x} := \sum_{n=1}^{k} \langle \vec{x}, \vec{b}_n \rangle A\vec{b}_n.$$

Then we have for all $\vec{x}$ with $\|\vec{x}\| = 1$,

$$\begin{aligned}\|(A - A_k)\vec{x}\| &= \Big\| \sum_{n=k+1}^{\infty} \langle \vec{x}, \vec{b}_n \rangle A\vec{b}_n \Big\| \le \sum_{n=k+1}^{\infty} |\langle \vec{x}, \vec{b}_n \rangle| \, \|A\vec{b}_n\| \\ &\le \Big(\sum_{n=k+1}^{\infty} |\langle \vec{x}, \vec{b}_n \rangle|^2 \Big)^{\frac{1}{2}} \Big(\sum_{n=k+1}^{\infty} \|A\vec{b}_n\|^2 \Big)^{\frac{1}{2}} \le \Big(\sum_{n=k+1}^{\infty} \|A\vec{b}_n\|^2 \Big)^{\frac{1}{2}},\end{aligned}$$

and hence,

$$\|(A - A_k)\| \le \Big(\sum_{n=k+1}^{\infty} \|A\vec{b}_n\|^2 \Big)^{\frac{1}{2}} \to 0 \ \text{ as } \ k \to \infty.$$

Thus, A is compact by Theorem 10.7. □

Theorem 10.10. *For every compact self-adjoint operator A either $\|A\|$ or $-\|A\|$ is an eigenvalue.*

Proof. The case $\|A\| = 0$ is trivial therefore we assume $\alpha = \|A\| > 0$. Since

$$\|A\|^2 = \sup_{\|\vec{v}\|=1} \|A\vec{v}\|^2 = \sup_{\|\vec{v}\|=1} \langle A\vec{v}, A\vec{v}\rangle = \sup_{\|\vec{v}\|=1} \langle \vec{v}, A^2\vec{v}\rangle$$

we have a sequence $\vec{v}_n$ with $\|\vec{v}_n\| = 1$ and such that $\langle \vec{v}_n, A^2\vec{v}_n\rangle \to \alpha^2$ as $n \to \infty$. Since A is compact, A^2 is compact as well and so we may assume without loss of generality that $A^2\vec{v}_n \to \vec{v} \neq \vec{0}$. Then,

$$\|A^2\vec{v}_n - \alpha^2\vec{v}_n\|^2 = \|A^2 v_n\|^2 - 2\alpha^2\langle \vec{v}_n, A^2\vec{v}_n\rangle + \alpha^4 \le 2\alpha^2\big(\alpha^2 - \langle \vec{v}_n, A^2\vec{v}_n\rangle\big) \to 0,$$

implies that $\vec{v}_n \to \alpha^2\vec{v}$ and hence $(A^2 - \alpha^2)\vec{v} = \lim_{n\to\infty}(A^2 - \alpha^2)\vec{v}_n = \vec{0}$. Factoring out $(A^2 - \alpha^2) = (A + \alpha)(A - \alpha)$ and letting $\vec{u} = (A - \alpha)\vec{v}$ so that $(A + \alpha)\vec{u} = \vec{0}$, we get that either $\vec{u} = \vec{0}$ in which case α is an eigenvalue of A or $\vec{u} \neq \vec{0}$ in which case $-\alpha$ is an eigenvalue of A. □

Theorem 10.11. *Let A be a compact self-adjoint operator. Then there is a sequence of real eigenvalues $\lambda_1, \lambda_2, \ldots$ converging to 0. The corresponding normalized eigenvectors $\vec{u}_k$ form an orthonormal set such that for every $\vec{f} \in V$,*

$$\vec{f} = \sum_{k=1}^{\infty} P_{\vec{u}_k}\vec{f} + \vec{h} = \sum_{k=1}^{\infty}\langle \vec{f}, \vec{u}_k\rangle\vec{u}_k + \vec{h},$$

where $\vec{h} \in \mathrm{Ker}(A)$. As a consequence, the following spectral decomposition of A holds,

$$A\vec{f} = \sum_{k=1}^{\infty} \lambda_k P_{\vec{u}_k}\vec{f} = \sum_{k=1}^{\infty} \lambda_k\langle \vec{f}, \vec{u}_k\rangle\vec{u}_k.$$

Proof. Let λ_1 be an eigenvalue of $A_1 = A$ on $V_1 = V$ satisfying $|\lambda_1| = \|A_1\|$ and $\vec{u}_1 \in V_1$ be the corresponding normalized eigenvector of A_1. Then A_1 maps $\{\vec{u}_1\}^\perp$ into itself since for any $\vec{x} \in \{\vec{u}_1\}^\perp$,

$$\langle A_1\vec{x}, \vec{u}_1\rangle = \langle \vec{x}, A_1\vec{u}_1\rangle = \lambda_1\langle \vec{x}, \vec{u}_1\rangle = 0.$$

Denoting by A_2 the restriction of A_1 to $V_2 = \{\vec{u}_1\}^\perp = \{\vec{v} \in V_1 \ : \ \vec{v} \perp \vec{u}_1\}$ and noting that A_2 is a compact self-adjoint operator on V_2, we obtain as above an eigenvalue λ_2 such that $|\lambda_2| = \|A_2\|$ and the corresponding normalized eigenvector $\vec{u}_2 \in V_2$ of A_2 and hence of A. Continuing this way we get a sequence of eigenvalues $\lambda_1, \lambda_2, \lambda_3, \ldots$ and the corresponding normalized eigenvectors $\vec{u}_1, \vec{u}_2, \vec{u}_3, \ldots$ of A which are mutually orthogonal by construction.

Next, proceeding by contradiction suppose that the eigenvalues do not converge to 0, that is, there are infinitely many eigenvalues $|\lambda_{k_m}| \ge \varepsilon > 0$. Then $\vec{v}_{k_m} = \lambda_{k_m}^{-1}\vec{u}_{k_m}$ are bounded $\|\vec{v}_{k_m}\| \le 1/\varepsilon$ by our assumption, yet $A\vec{v}_{k_m} = \vec{u}_{k_m}$ do not have a convergent subsequence since $\|\vec{u}_k - \vec{u}_m\| = \sqrt{2}$ for all k, m, a contradiction.

Finally, let $\vec{f}_n$ be the orthogonal projection of $\vec{f}$ onto the span$\{\vec{u}_1, ..., \vec{u}_{n-1}\}$,

$$\vec{f}_n = \sum_{k=1}^{n-1} \langle \vec{f}, \vec{u}_k \rangle \vec{u}_k.$$

Then $\vec{f} - \vec{f}_n \in V_n$ and since $A = A_n$ on V_n and $\|A_n\| = |\alpha_n|$ we get $\|A(\vec{f} - \vec{f}_n)\| \leq |\alpha_n| \, \|\vec{f} - \vec{f}_n\| \leq 2|\alpha_n| \, \|\vec{f}\|$. As the eigenvalues of A go to zero, we conclude that $\|A(\vec{f} - \vec{f}_n)\| \to 0$ as $n \to \infty$, and hence, $\vec{h} = \vec{f} - \lim_{n\to\infty} \vec{f}_n$ satisfies $A\vec{h} = \vec{0}$. □

Exercises 10.12. Show that every compact self-adjoint operator can be approximated arbitrarily closely by finite rank self-adjoint operators in the operator norm.

Corollary 10.13. *Every compact self-adjoint operator has an associated orthonormal basis of eigenvectors.*

Proof. If the sequence of eigenvectors obtained in Theorem 10.11 does not form a basis of V then we extend it to a basis by adding vectors from Ker(A), which are automatically eigenvectors of A corresponding to the eigenvalue 0. □

11 Trace perturbation theory

As we can see from Theorem 4.2, the solution of the IVP (3.1) is determined by the parameters of the problem reflected in the matrix of coefficients A. Change of the parameters of the problem entails change of the solution. To determine how the solution changes, it is enough to determine how the function of a matrix given by Definition 4.11 and its spectrum changes when the matrix is perturbed. Simpler problems, but still having physical meaning, ask how $\frac{1}{n}\mathrm{tr}(f(A))$, the average of the eigenvalues of the function of a matrix, changes when the matrix is perturbed.

Exercises 11.1. Let $A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & \mathrm{i} & 0 \\ -\mathrm{i} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

(i) Compare the sets $\sigma(B - A)$ and $\{\mu - \lambda : \lambda \in \sigma(A), \mu \in \sigma(B)\}$.

(ii) Compare $\mathrm{tr}(f(B) - f(A))$ and $\sum_{j=1}^{3}(f(\mu_j) - f(\lambda_j))$, where μ_j are eigenvalues of B and λ_j are eigenvalues of A.

Definition 11.2. Let A and B be two self-adjoint $n \times n$ matrices. Denote by $\lambda_1, \ldots, \lambda_n$ the eigenvalues of A and by $\mu_1, \ldots, \mu_n$ the eigenvalues of B. The function

$$\xi(t) = \#\{j : \lambda_j < t\} - \#\{k : \mu_k < t\} \tag{11.1}$$

is called the **spectral shift function** for the pair of matrices A and B. (Here # denotes the cardinality of a set, that is, the number of the elements in the set.)

Remark 11.3. The function ξ admits the following interpretation: the value $\xi(t)$ equals the net number of the eigenvalues of a matrix that crossed t in the positive direction as A is perturbed to B. (If the eigenvalues crossed t in the negative direction, then $\xi(t) < 0$.)

Exercises 11.4. (i) Calculate the spectral shift function for the pair of matrices with the following spectra:

(a) eigenvalues of A are $-1, 0, 1$, eigenvalues of B are $0, 1, 2$;

(b) eigenvalues of A are $-1, 1, 2$, eigenvalues of B are $-2, 1, 3$;

(c) eigenvalues of A are $-4, -3, -3, -3, 2, 6$, eigenvalues of B are $-5, -2, -2, 4, 7, 8$.

(ii) For f a continuously differentiable function on $\mathbb{R}$ and the pairs of matrices from part (i), derive the formula $\operatorname{tr}(f(B) - f(A)) = \int_{-\infty}^{\infty} f'(t)\xi(t)\,dt$.

Theorem 11.5. *Let A, B be self-adjoint matrices of the same dimension. Then, for every $f \in C^1(\mathbb{R})$,*

$$\operatorname{tr}(f(B) - f(A)) = \int_{-\infty}^{\infty} f'(t)\,\xi(t)\,dt. \tag{11.2}$$

Note that (11.3) provides an analog of the fundamental theorem of calculus for functions of matrices. Theorem 11.5 was designed for operators on infinite dimensional spaces, whose spectra can be uncountable sets. It can be that the trace of $f(B) - f(A)$ is not defined, but the trace of the remainder of the higher order Taylor approximation of the operator function $f(B)$ is defined. For such cases, we have a more general version of Theorem 11.5, which we state only for matrices.

Theorem 11.6. *Let A, B be self-adjoint matrices of the same dimension and let k be a natural number. Then, there exists an integrable function $\eta_k(t)$ determined by k, A, B such that for every $f \in C^1(\mathbb{R})$,*

$$\operatorname{tr}\left(f(B) - f(A) - \sum_{j=1}^{k-1} \frac{1}{j!} \frac{d^j}{dt^j}\bigg|_{t=0} f\big(A + t(B-A)\big)\right) = \int_{-\infty}^{\infty} f^{(k)}(t)\,\eta_k(t)\,dt. \tag{11.3}$$

Project 11.7. Existence of the functions η_k in Theorem 11.6 is established implicitly. Complicated explicit formulas for η_k are known only in very simple cases. The project consists in understanding η_k better, in finding formulas for η_k and establishing properties of η_k.

References

[1] D. Austin, *How google finds your needle in the web's haystack.* `http://www.ams.org/samplings/feature-column/fcarc-pagerank`.

[2] J. Jauregui, *Error-correcting codes with linear algebra.* `http://www.math.upenn.edu/~jjau/ECC.pdf`.

[3] P. D. Lax, *Linear algebra and its applications*, 2nd ed., Pure and Applied Mathematics (Hoboken), Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2007. MR2356919 (2008j:15002)

[4] S. J. Leon, *Linear algebra with applications*, Macmillan Inc., New York, 1980. MR566769 (81c:15002)

[5] B. D. MacCluer, *Elementary functional analysis*, Graduate Texts in Mathematics, vol. 253, Springer, New York, 2009. MR2462971 (2010b:46001)

[6] G. Teschl, *Jacobi Operators and Completely Integrable Nonlinear Lattices*, Mathematical Surveys and Monographs, vol. 72, American Mathematical Society, Providence, RI, 2000. `http://www.mat.univie.ac.at/~gerald/ftp/book-jac/jacop.pdf`. MR1711536 (2001b:39019)

[7] G. Teschl, *Topics in Real and Functional Analysis*, web draft, 2013. `http://www.mat.univie.ac.at/~gerald/ftp/book-fa/fa.pdf`.

[8] S. Treil, *Linear Algebra Done Wrong*, web draft, 2011. `http://www.math.brown.edu/~treil/papers/LADW/book.pdf`.