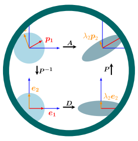
4

Matrix Decompositions



In Chapters 2 and 3, we studied ways to manipulate and measure vectors, projections of vectors, and linear mappings. Mappings and transformations of vectors can be conveniently described as operations performed by matrices. Moreover, data is often represented in matrix form as well, e.g., where the rows of the matrix represent different people and the columns describe different features of the people, such as weight, height, and socioeconomic status. In this chapter, we present three aspects of matrices: how to summarize matrices, how matrices can be decomposed, and how these decompositions can be used for matrix approximations.

We first consider methods that allow us to describe matrices with just a few numbers that characterize the overall properties of matrices. We will do this in the sections on determinants (Section 4.1) and eigenvalues (Section 4.2) for the important special case of square matrices. These characteristic numbers have important mathematical consequences and allow us to quickly grasp what useful properties a matrix has. From here we will proceed to matrix decomposition methods: An analogy for matrix decomposition is the factoring of numbers, such as the factoring of 21 into prime numbers $7 \cdot 3$. For this reason matrix decomposition is also often referred to as *matrix factorization*. Matrix decompositions are used to describe a matrix by means of a different representation using factors of interpretable matrices.

matrix factorization

We will first cover a square-root-like operation for symmetric, positive definite matrices, the Cholesky decomposition (Section 4.3). From here we will look at two related methods for factorizing matrices into canonical forms. The first one is known as matrix diagonalization (Section 4.4), which allows us to represent the linear mapping using a diagonal transformation matrix if we choose an appropriate basis. The second method, singular value decomposition (Section 4.5), extends this factorization to non-square matrices, and it is considered one of the fundamental concepts in linear algebra. These decompositions are helpful, as matrices representing numerical data are often very large and hard to analyze. We conclude the chapter with a systematic overview of the types of matrices and the characteristic properties that distinguish them in the form of a matrix taxonomy (Section 4.7).

The methods that we cover in this chapter will become important in

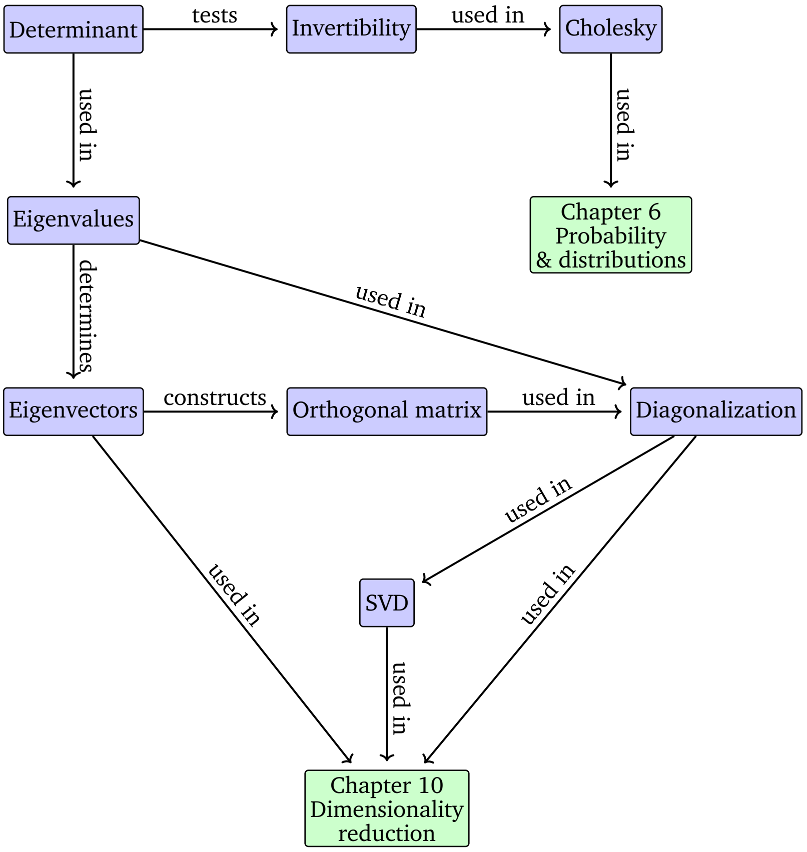


Figure 4.1 A mind map of the concepts introduced in this chapter, along with where they are used in other parts of the book.

both subsequent mathematical chapters, such as Chapter 6, but also in applied chapters, such as dimensionality reduction in Chapters 10 or density estimation in Chapter 11. This chapter's overall structure is depicted in the mind map of Figure 4.1.

4.1 Determinant and Trace

Determinants are important concepts in linear algebra. A determinant is a mathematical object in the analysis and solution of systems of linear equations. Determinants are only defined for square matrices $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, i.e., matrices with the same number of rows and columns. In this book, we write the determinant as $\det(\boldsymbol{A})$ or sometimes as $|\boldsymbol{A}|$ so that

The determinant notation $|\boldsymbol{A}|$ must not be confused with the absolute value.

$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} . \tag{4.1}$$

The *determinant* of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a function that maps $\boldsymbol{A}$

determinant

onto a real number. Before providing a definition of the determinant for general $n \times n$ matrices, let us have a look at some motivating examples, and define determinants for some special matrices.

Example 4.1 (Testing for Matrix Invertibility)
Let us begin with exploring if a square matrix $\boldsymbol{A}$ is invertible (see Section 2.2.2). For the smallest cases, we already know when a matrix is invertible. If $\boldsymbol{A}$ is a 1×1 matrix, i.e., it is a scalar number, then $\boldsymbol{A} = a \implies \boldsymbol{A}^{-1} = \frac{1}{a}$. Thus $a\,\frac{1}{a} = 1$ holds, if and only if $a \neq 0$.

For 2×2 matrices, by the definition of the inverse (Definition 2.3), we know that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$. Then, with (2.24), the inverse of $\boldsymbol{A}$ is

$$\boldsymbol{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}. \tag{4.2}$$

Hence, $\boldsymbol{A}$ is invertible if and only if

$$a_{11}a_{22} - a_{12}a_{21} \neq 0\,. \tag{4.3}$$

This quantity is the determinant of $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$, i.e.,

$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}\,. \tag{4.4}$$

Example 4.1 points already at the relationship between determinants and the existence of inverse matrices. The next theorem states the same result for $n \times n$ matrices.

Theorem 4.1. *For any square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ it holds that $\boldsymbol{A}$ is invertible if and only if* $\det(\boldsymbol{A}) \neq 0$.

We have explicit (closed-form) expressions for determinants of small matrices in terms of the elements of the matrix. For $n = 1$,

$$\det(\boldsymbol{A}) = \det(a_{11}) = a_{11}\,. \tag{4.5}$$

For $n = 2$,

$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}\,, \tag{4.6}$$

which we have observed in the preceding example.

For $n = 3$ (known as Sarrus' rule),

$$\begin{aligned} \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} &= a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} \\ &\quad - a_{31}a_{22}a_{13} - a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33}\,. \end{aligned} \tag{4.7}$$

For a memory aid of the product terms in Sarrus' rule, try tracing the elements of the triple products in the matrix.

We call a square matrix $\boldsymbol{T}$ an *upper-triangular matrix* if $T_{ij} = 0$ for $i > j$, i.e., the matrix is zero below its diagonal. Analogously, we define a *lower-triangular matrix* as a matrix with zeros above its diagonal. For a triangular matrix $\boldsymbol{T} \in \mathbb{R}^{n \times n}$, the determinant is the product of the diagonal elements, i.e.,

upper-triangular matrix

lower-triangular matrix

$$\det(\boldsymbol{T}) = \prod_{i=1}^{n} T_{ii} \,. \tag{4.8}$$

The determinant is the signed volume of the parallelepiped formed by the columns of the matrix.

Example 4.2 (Determinants as Measures of Volume)
The notion of a determinant is natural when we consider it as a mapping from a set of n vectors spanning an object in $\mathbb{R}^n$. It turns out that the determinant $\det(\boldsymbol{A})$ is the signed volume of an n-dimensional parallelepiped formed by columns of the matrix $\boldsymbol{A}$.

For $n = 2$, the columns of the matrix form a parallelogram; see Figure 4.2. As the angle between vectors gets smaller, the area of a parallelogram shrinks, too. Consider two vectors $\boldsymbol{b}, \boldsymbol{g}$ that form the columns of a matrix $\boldsymbol{A} = [\boldsymbol{b}, \boldsymbol{g}]$. Then, the absolute value of the determinant of $\boldsymbol{A}$ is the area of the parallelogram with vertices $\boldsymbol{0}, \boldsymbol{b}, \boldsymbol{g}, \boldsymbol{b} + \boldsymbol{g}$. In particular, if $\boldsymbol{b}, \boldsymbol{g}$ are linearly dependent so that $\boldsymbol{b} = \lambda \boldsymbol{g}$ for some $\lambda \in \mathbb{R}$, they no longer form a two-dimensional parallelogram. Therefore, the corresponding area is 0. On the contrary, if $\boldsymbol{b}, \boldsymbol{g}$ are linearly independent and are multiples of the canonical basis vectors $\boldsymbol{e}_1, \boldsymbol{e}_2$ then they can be written as $\boldsymbol{b} = \begin{bmatrix} b \\ 0 \end{bmatrix}$ and $\boldsymbol{g} = \begin{bmatrix} 0 \\ g \end{bmatrix}$, and the determinant is $\begin{vmatrix} b & 0 \\ 0 & g \end{vmatrix} = bg - 0 = bg$.

Figure 4.2 The area of the parallelogram (shaded region) spanned by the vectors $\boldsymbol{b}$ and $\boldsymbol{g}$ is $|\det([\boldsymbol{b}, \ \boldsymbol{g}])|$.

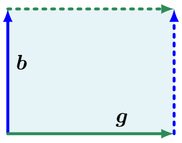
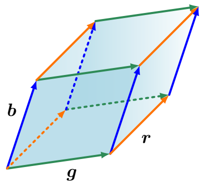


Figure 4.3 The volume of the parallelepiped (shaded volume) spanned by vectors $\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{g}$ is $|\det([\boldsymbol{r}, \ \boldsymbol{b}, \ \boldsymbol{g}])|$.



The sign of the determinant indicates the orientation of the spanning vectors.

The sign of the determinant indicates the orientation of the spanning vectors $\boldsymbol{b}, \boldsymbol{g}$ with respect to the standard basis $(\boldsymbol{e}_1, \boldsymbol{e}_2)$. In our figure, flipping the order to $\boldsymbol{g}, \boldsymbol{b}$ swaps the columns of $\boldsymbol{A}$ and reverses the orientation of the shaded area. This becomes the familiar formula: area = height × length. This intuition extends to higher dimensions. In $\mathbb{R}^3$, we consider three vectors $\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{g} \in \mathbb{R}^3$ spanning the edges of a parallelepiped, i.e., a solid with faces that are parallel parallelograms (see Figure 4.3). The absolute value of the determinant of the 3×3 matrix $[\boldsymbol{r}, \ \boldsymbol{b}, \ \boldsymbol{g}]$ is the volume of the solid. Thus, the determinant acts as a function that measures the signed volume formed by column vectors composed in a matrix.

Consider the three linearly independent vectors $\boldsymbol{r}, \boldsymbol{g}, \boldsymbol{b} \in \mathbb{R}^3$ given as

$$\boldsymbol{r} = \begin{bmatrix} 2 \\ 0 \\ -8 \end{bmatrix}, \quad \boldsymbol{g} = \begin{bmatrix} 6 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 4 \\ -1 \end{bmatrix}. \tag{4.9}$$

Writing these vectors as the columns of a matrix

$$\boldsymbol{A} = [\boldsymbol{r},\ \boldsymbol{g},\ \boldsymbol{b}] = \begin{bmatrix} 2 & 6 & 1 \\ 0 & 1 & 4 \\ -8 & 0 & -1 \end{bmatrix} \tag{4.10}$$

allows us to compute the desired volume as

$$V = |\det(\boldsymbol{A})| = 186\,. \tag{4.11}$$

Computing the determinant of an $n \times n$ matrix requires a general algorithm to solve the cases for $n > 3$, which we are going to explore in the following. Theorem 4.2 below reduces the problem of computing the determinant of an $n \times n$ matrix to computing the determinant of $(n-1) \times (n-1)$ matrices. By recursively applying the Laplace expansion (Theorem 4.2), we can therefore compute determinants of $n \times n$ matrices by ultimately computing determinants of 2×2 matrices.

Laplace expansion

Theorem 4.2 (Laplace Expansion)**.** *Consider a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. Then, for all $j = 1, \ldots, n$:*

$\det(\boldsymbol{A}_{k,j})$ is called a *minor* and $(-1)^{k+j}\det(\boldsymbol{A}_{k,j})$ a *cofactor*.

1. Expansion along column j

$$\det(\boldsymbol{A}) = \sum_{k=1}^{n} (-1)^{k+j} a_{kj} \det(\boldsymbol{A}_{k,j})\,. \tag{4.12}$$

2. Expansion along row j

$$\det(\boldsymbol{A}) = \sum_{k=1}^{n} (-1)^{k+j} a_{jk} \det(\boldsymbol{A}_{j,k})\,. \tag{4.13}$$

Here $\boldsymbol{A}_{k,j} \in \mathbb{R}^{(n-1)\times(n-1)}$ is the submatrix of $\boldsymbol{A}$ that we obtain when deleting row k and column j.

Example 4.3 (Laplace Expansion)
Let us compute the determinant of

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.14}$$

using the Laplace expansion along the first row. Applying (4.13) yields

$$\begin{aligned} \begin{vmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 0 & 0 & 1 \end{vmatrix} &= (-1)^{1+1} \cdot 1 \begin{vmatrix} 1 & 2 \\ 0 & 1 \end{vmatrix} \\ &\quad + (-1)^{1+2} \cdot 2 \begin{vmatrix} 3 & 2 \\ 0 & 1 \end{vmatrix} + (-1)^{1+3} \cdot 3 \begin{vmatrix} 3 & 1 \\ 0 & 0 \end{vmatrix}\,. \end{aligned} \tag{4.15}$$

We use (4.6) to compute the determinants of all 2×2 matrices and obtain

$$\det(\boldsymbol{A}) = 1(1-0) - 2(3-0) + 3(0-0) = -5 \, . \tag{4.16}$$

For completeness we can compare this result to computing the determinant using Sarrus' rule (4.7):

$$\det(\boldsymbol{A}) = 1\cdot1\cdot1+3\cdot0\cdot3+0\cdot2\cdot2-0\cdot1\cdot3-1\cdot0\cdot2-3\cdot2\cdot1 = 1-6 = -5 \, . \tag{4.17}$$

For $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ the determinant exhibits the following properties:

- The determinant of a matrix product is the product of the corresponding determinants, $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$.
- Determinants are invariant to transposition, i.e., $\det(\boldsymbol{A}) = \det(\boldsymbol{A}^\top)$.
- If $\boldsymbol{A}$ is regular (invertible), then $\det(\boldsymbol{A}^{-1}) = \frac{1}{\det(\boldsymbol{A})}$.
- Similar matrices (Definition 2.22) possess the same determinant. Therefore, for a linear mapping $\Phi : V \to V$ all transformation matrices $\boldsymbol{A}_\Phi$ of Φ have the same determinant. Thus, the determinant is invariant to the choice of basis of a linear mapping.
- Adding a multiple of a column/row to another one does not change $\det(\boldsymbol{A})$.
- Multiplication of a column/row with $\lambda \in \mathbb{R}$ scales $\det(\boldsymbol{A})$ by λ. In particular, $\det(\lambda\boldsymbol{A}) = \lambda^n \det(\boldsymbol{A})$.
- Swapping two rows/columns changes the sign of $\det(\boldsymbol{A})$.

Because of the last three properties, we can use Gaussian elimination (see Section 2.1) to compute $\det(\boldsymbol{A})$ by bringing $\boldsymbol{A}$ into row-echelon form. We can stop Gaussian elimination when we have $\boldsymbol{A}$ in a triangular form where the elements below the diagonal are all 0. Recall from (4.8) that the determinant of a triangular matrix is the product of the diagonal elements.

Theorem 4.3. *A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ has $\det(\boldsymbol{A}) \neq 0$ if and only if $\mathrm{rk}(\boldsymbol{A}) = n$. In other words, $\boldsymbol{A}$ is invertible if and only if it is full rank.*

When mathematics was mainly performed by hand, the determinant calculation was considered an essential way to analyze matrix invertibility. However, contemporary approaches in machine learning use direct numerical methods that superseded the explicit calculation of the determinant. For example, in Chapter 2, we learned that inverse matrices can be computed by Gaussian elimination. Gaussian elimination can thus be used to compute the determinant of a matrix.

Determinants will play an important theoretical role for the following sections, especially when we learn about eigenvalues and eigenvectors (Section 4.2) through the characteristic polynomial.

Definition 4.4. The *trace* of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is defined as

trace

$$\mathrm{tr}(\boldsymbol{A}) := \sum_{i=1}^{n} a_{ii} \,, \tag{4.18}$$

i.e. , the trace is the sum of the diagonal elements of $\boldsymbol{A}$.

The trace satisfies the following properties:

- $\mathrm{tr}(\boldsymbol{A}+\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B})$ for $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$
- $\mathrm{tr}(\alpha\boldsymbol{A}) = \alpha\mathrm{tr}(\boldsymbol{A})\,, \alpha \in \mathbb{R}$ for $\boldsymbol{A} \in \mathbb{R}^{n\times n}$
- $\mathrm{tr}(\boldsymbol{I}_n) = n$
- $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{A})$ for $\boldsymbol{A} \in \mathbb{R}^{n\times k}, \boldsymbol{B} \in \mathbb{R}^{k\times n}$

It can be shown that only one function satisfies these four properties together – the trace (Gohberg et al., 2012).

The properties of the trace of matrix products are more general. Specifically, the trace is invariant under cyclic permutations, i.e.,

The trace is invariant under cyclic permutations.

$$\mathrm{tr}(\boldsymbol{AKL}) = \mathrm{tr}(\boldsymbol{KLA}) \tag{4.19}$$

for matrices $\boldsymbol{A} \in \mathbb{R}^{a\times k}, \boldsymbol{K} \in \mathbb{R}^{k\times l}, \boldsymbol{L} \in \mathbb{R}^{l\times a}$. This property generalizes to products of an arbitrary number of matrices. As a special case of (4.19), it follows that for two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$\mathrm{tr}(\boldsymbol{x}\boldsymbol{y}^\top) = \mathrm{tr}(\boldsymbol{y}^\top\boldsymbol{x}) = \boldsymbol{y}^\top\boldsymbol{x} \in \mathbb{R}\,. \tag{4.20}$$

Given a linear mapping $\Phi : V \to V$, where V is a vector space, we define the trace of this map by using the trace of matrix representation of Φ. For a given basis of V, we can describe Φ by means of the transformation matrix $\boldsymbol{A}$. Then the trace of Φ is the trace of $\boldsymbol{A}$. For a different basis of V, it holds that the corresponding transformation matrix $\boldsymbol{B}$ of Φ can be obtained by a basis change of the form $\boldsymbol{S}^{-1}\boldsymbol{AS}$ for suitable $\boldsymbol{S}$ (see Section 2.7.2). For the corresponding trace of Φ, this means

$$\mathrm{tr}(\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{S}^{-1}\boldsymbol{AS}) \overset{(4.19)}{=} \mathrm{tr}(\boldsymbol{ASS}^{-1}) = \mathrm{tr}(\boldsymbol{A})\,. \tag{4.21}$$

Hence, while matrix representations of linear mappings are basis dependent the trace of a linear mapping Φ is independent of the basis.

In this section, we covered determinants and traces as functions characterizing a square matrix. Taking together our understanding of determinants and traces we can now define an important equation describing a matrix $\boldsymbol{A}$ in terms of a polynomial, which we will use extensively in the following sections.

Definition 4.5 (Characteristic Polynomial)**.** For $\lambda \in \mathbb{R}$ and a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$

$$p_{\boldsymbol{A}}(\lambda) := \det(\boldsymbol{A} - \lambda\boldsymbol{I}) \tag{4.22a}$$

$$= c_0 + c_1\lambda + c_2\lambda^2 + \cdots + c_{n-1}\lambda^{n-1} + (-1)^n\lambda^n\,, \tag{4.22b}$$

characteristic polynomial

$c_0, \ldots, c_{n-1} \in \mathbb{R}$, is the *characteristic polynomial* of $\boldsymbol{A}$. In particular,

$$c_0 = \det(\boldsymbol{A}), \tag{4.23}$$

$$c_{n-1} = (-1)^{n-1}\mathrm{tr}(\boldsymbol{A}). \tag{4.24}$$

The characteristic polynomial (4.22a) will allow us to compute eigenvalues and eigenvectors, covered in the next section.

4.2 Eigenvalues and Eigenvectors

We will now get to know a new way to characterize a matrix and its associated linear mapping. Recall from Section 2.7.1 that every linear mapping has a unique transformation matrix given an ordered basis. We can interpret linear mappings and their associated transformation matrices by performing an "eigen" analysis. As we will see, the eigenvalues of a linear mapping will tell us how a special set of vectors, the eigenvectors, is transformed by the linear mapping.

Eigen is a German word meaning "characteristic", "self", or "own".

Definition 4.6. Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a square matrix. Then $\lambda \in \mathbb{R}$ is an *eigenvalue* of $\boldsymbol{A}$ and $\boldsymbol{x} \in \mathbb{R}^n\backslash\{\boldsymbol{0}\}$ is the corresponding *eigenvector* of $\boldsymbol{A}$ if

eigenvalue

eigenvector

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}. \tag{4.25}$$

We call (4.25) the *eigenvalue equation*.

eigenvalue equation

Remark. In the linear algebra literature and software, it is often a convention that eigenvalues are sorted in descending order, so that the largest eigenvalue and associated eigenvector are called the first eigenvalue and its associated eigenvector, and the second largest called the second eigenvalue and its associated eigenvector, and so on. However, textbooks and publications may have different or no notion of orderings. We do not want to presume an ordering in this book if not stated explicitly. ♢

The following statements are equivalent:

- λ is an eigenvalue of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.
- There exists an $\boldsymbol{x} \in \mathbb{R}^n\backslash\{\boldsymbol{0}\}$ with $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$, or equivalently, $(\boldsymbol{A} - \lambda\boldsymbol{I}_n)\boldsymbol{x} = \boldsymbol{0}$ can be solved non-trivially, i.e., $\boldsymbol{x} \neq \boldsymbol{0}$.
- $\mathrm{rk}(\boldsymbol{A} - \lambda\boldsymbol{I}_n) < n$.
- $\det(\boldsymbol{A} - \lambda\boldsymbol{I}_n) = 0$.

Definition 4.7 (Collinearity and Codirection)**.** Two vectors that point in the same direction are called *codirected*. Two vectors are *collinear* if they point in the same or the opposite direction.

codirected

collinear

Remark (Non-uniqueness of eigenvectors). If $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$ associated with eigenvalue λ, then for any $c \in \mathbb{R}\backslash\{0\}$ it holds that $c\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$ with the same eigenvalue since

$$\boldsymbol{A}(c\boldsymbol{x}) = c\boldsymbol{A}\boldsymbol{x} = c\lambda\boldsymbol{x} = \lambda(c\boldsymbol{x}). \tag{4.26}$$

Thus, all vectors that are collinear to $\boldsymbol{x}$ are also eigenvectors of $\boldsymbol{A}$.

♢

Theorem 4.8. *$\lambda \in \mathbb{R}$ is an eigenvalue of $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ if and only if λ is a root of the characteristic polynomial $p_{\boldsymbol{A}}(\lambda)$ of $\boldsymbol{A}$.*

algebraic multiplicity

Definition 4.9. Let a square matrix $\boldsymbol{A}$ have an eigenvalue λ_i. The *algebraic multiplicity* of λ_i is the number of times the root appears in the characteristic polynomial.

eigenspace

eigenspectrum

spectrum

Definition 4.10 (Eigenspace and Eigenspectrum)**.** For $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, the set of all eigenvectors of $\boldsymbol{A}$ associated with an eigenvalue λ spans a subspace of $\mathbb{R}^n$, which is called the *eigenspace* of $\boldsymbol{A}$ with respect to λ and is denoted by E_λ. The set of all eigenvalues of $\boldsymbol{A}$ is called the *eigenspectrum*, or just *spectrum*, of $\boldsymbol{A}$.

If λ is an eigenvalue of $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, then the corresponding eigenspace E_λ is the solution space of the homogeneous system of linear equations $(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$. Geometrically, the eigenvector corresponding to a nonzero eigenvalue points in a direction that is stretched by the linear mapping. The eigenvalue is the factor by which it is stretched. If the eigenvalue is negative, the direction of the stretching is flipped.

Example 4.4 (The Case of the Identity Matrix)
The identity matrix $\boldsymbol{I} \in \mathbb{R}^{n \times n}$ has characteristic polynomial $p_{\boldsymbol{I}}(\lambda) = \det(\boldsymbol{I} - \lambda \boldsymbol{I}) = (1 - \lambda)^n = 0$, which has only one eigenvalue $\lambda = 1$ that occurs n times. Moreover, $\boldsymbol{I}\boldsymbol{x} = \lambda \boldsymbol{x} = 1\boldsymbol{x}$ holds for all vectors $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$. Because of this, the sole eigenspace E_1 of the identity matrix spans n dimensions, and all n standard basis vectors of $\mathbb{R}^n$ are eigenvectors of $\boldsymbol{I}$.

Useful properties regarding eigenvalues and eigenvectors include the following:

- A matrix $\boldsymbol{A}$ and its transpose $\boldsymbol{A}^\top$ possess the same eigenvalues, but not necessarily the same eigenvectors.
- The eigenspace E_λ is the null space of $\boldsymbol{A} - \lambda \boldsymbol{I}$ since

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x} &\iff \boldsymbol{A}\boldsymbol{x} - \lambda \boldsymbol{x} = \boldsymbol{0} && (4.27a) \\ &\iff (\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{x} = \boldsymbol{0} \iff \boldsymbol{x} \in \ker(\boldsymbol{A} - \lambda \boldsymbol{I}). && (4.27b) \end{aligned}$$

- Similar matrices (see Definition 2.22) possess the same eigenvalues. Therefore, a linear mapping Φ has eigenvalues that are independent of the choice of basis of its transformation matrix. This makes eigenvalues, together with the determinant and the trace, key characteristic parameters of a linear mapping as they are all invariant under basis change.
- Symmetric, positive definite matrices always have positive, real eigenvalues.

Example 4.5 (Computing Eigenvalues, Eigenvectors, and Eigenspaces)

Let us find the eigenvalues and eigenvectors of the 2×2 matrix

$$\boldsymbol{A} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} . \tag{4.28}$$

Step 1: Characteristic Polynomial. From our definition of the eigenvector $\boldsymbol{x} \neq \boldsymbol{0}$ and eigenvalue λ of $\boldsymbol{A}$, there will be a vector such that $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$, i.e., $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$. Since $\boldsymbol{x} \neq \boldsymbol{0}$, this requires that the kernel (null space) of $\boldsymbol{A} - \lambda\boldsymbol{I}$ contains more elements than just $\boldsymbol{0}$. This means that $\boldsymbol{A} - \lambda\boldsymbol{I}$ is not invertible and therefore $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$. Hence, we need to compute the roots of the characteristic polynomial (4.22a) to find the eigenvalues.

Step 2: Eigenvalues. The characteristic polynomial is

$$\begin{aligned} p_{\boldsymbol{A}}(\lambda) &= \det(\boldsymbol{A} - \lambda\boldsymbol{I}) && \text{(4.29a)} \\ &= \det\left(\begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) = \begin{vmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{vmatrix} && \text{(4.29b)} \\ &= (4-\lambda)(3-\lambda) - 2 \cdot 1 . && \text{(4.29c)} \end{aligned}$$

We factorize the characteristic polynomial and obtain

$$p(\lambda) = (4-\lambda)(3-\lambda) - 2 \cdot 1 = 10 - 7\lambda + \lambda^2 = (2-\lambda)(5-\lambda) \tag{4.30}$$

giving the roots $\lambda_1 = 2$ and $\lambda_2 = 5$.

Step 3: Eigenvectors and Eigenspaces. We find the eigenvectors that correspond to these eigenvalues by looking at vectors $\boldsymbol{x}$ such that

$$\begin{bmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{bmatrix} \boldsymbol{x} = \boldsymbol{0} . \tag{4.31}$$

For $\lambda = 5$ we obtain

$$\begin{bmatrix} 4-5 & 2 \\ 1 & 3-5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{0} . \tag{4.32}$$

We solve this homogeneous system and obtain a solution space

$$E_5 = \operatorname{span}[\begin{bmatrix} 2 \\ 1 \end{bmatrix}] . \tag{4.33}$$

This eigenspace is one-dimensional as it possesses a single basis vector.

Analogously, we find the eigenvector for $\lambda = 2$ by solving the homogeneous system of equations

$$\begin{bmatrix} 4-2 & 2 \\ 1 & 3-2 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix} \boldsymbol{x} = \boldsymbol{0} . \tag{4.34}$$

This means any vector $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, where $x_2 = -x_1$, such as $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$, is an eigenvector with eigenvalue 2. The corresponding eigenspace is given as

$$E_2 = \text{span}[\begin{bmatrix} 1 \\ -1 \end{bmatrix}]\,. \tag{4.35}$$

The two eigenspaces E_5 and E_2 in Example 4.5 are one-dimensional as they are each spanned by a single vector. However, in other cases we may have multiple identical eigenvalues (see Definition 4.9) and the eigenspace may have more than one dimension.

geometric multiplicity

Definition 4.11. Let λ_i be an eigenvalue of a square matrix $\boldsymbol{A}$. Then the *geometric multiplicity* of λ_i is the number of linearly independent eigenvectors associated with λ_i. In other words, it is the dimensionality of the eigenspace spanned by the eigenvectors associated with λ_i.

Remark. A specific eigenvalue's geometric multiplicity must be at least one because every eigenvalue has at least one associated eigenvector. An eigenvalue's geometric multiplicity cannot exceed its algebraic multiplicity, but it may be lower. ♢

Example 4.6

The matrix $\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$ has two repeated eigenvalues $\lambda_1 = \lambda_2 = 2$ and an algebraic multiplicity of 2. The eigenvalue has, however, only one distinct unit eigenvector $\boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and, thus, geometric multiplicity 1.

Graphical Intuition in Two Dimensions

Let us gain some intuition for determinants, eigenvectors, and eigenvalues using different linear mappings. Figure 4.4 depicts five transformation matrices $\boldsymbol{A}_1, \dots, \boldsymbol{A}_5$ and their impact on a square grid of points, centered at the origin:

In geometry, the area-preserving properties of this type of shearing parallel to an axis is also known as Cavalieri's principle of equal areas for parallelograms (Katz, 2004).

- $\boldsymbol{A}_1 = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix}$. The direction of the two eigenvectors correspond to the canonical basis vectors in $\mathbb{R}^2$, i.e., to two cardinal axes. The vertical axis is extended by a factor of 2 (eigenvalue $\lambda_1 = 2$), and the horizontal axis is compressed by factor $\frac{1}{2}$ (eigenvalue $\lambda_2 = \frac{1}{2}$). The mapping is area preserving ($\det(\boldsymbol{A}_1) = 1 = 2 \cdot \frac{1}{2}$).
- $\boldsymbol{A}_2 = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{bmatrix}$ corresponds to a shearing mapping , i.e., it shears the points along the horizontal axis to the right if they are on the positive

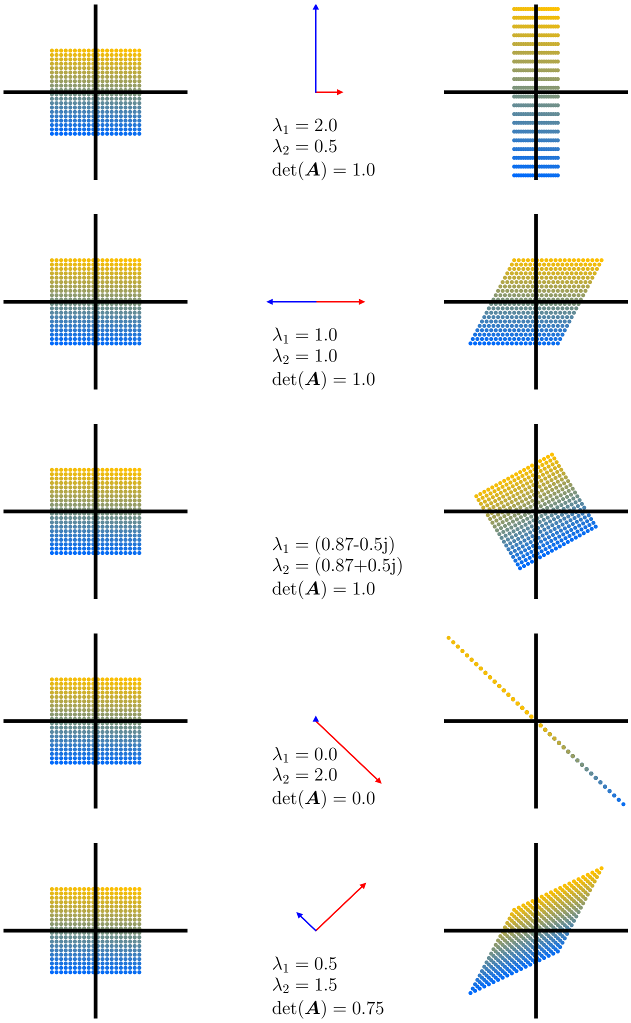


Figure 4.4 Determinants and eigenspaces. Overview of five linear mappings and their associated transformation matrices $\boldsymbol{A}_i \in \mathbb{R}^{2\times 2}$ projecting 400 color-coded points $\boldsymbol{x} \in \mathbb{R}^2$ (left column) onto target points $\boldsymbol{A}_i\boldsymbol{x}$ (right column). The central column depicts the first eigenvector, stretched by its associated eigenvalue λ_1, and the second eigenvector stretched by its eigenvalue λ_2. Each row depicts the effect of one of five transformation matrices $\boldsymbol{A}_i$ with respect to the standard basis.

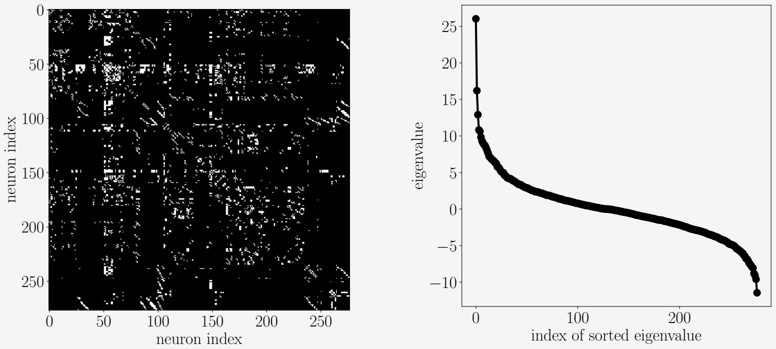
half of the vertical axis, and to the left vice versa. This mapping is area preserving ($\det(\boldsymbol{A}_2) = 1$). The eigenvalue $\lambda_1 = 1 = \lambda_2$ is repeated and the eigenvectors are collinear (drawn here for emphasis in two opposite directions). This indicates that the mapping acts only along one direction (the horizontal axis).

- $\boldsymbol{A}_3 = \begin{bmatrix}\cos(\frac{\pi}{6}) & -\sin(\frac{\pi}{6})\\ \sin(\frac{\pi}{6}) & \cos(\frac{\pi}{6})\end{bmatrix} = \frac{1}{2}\begin{bmatrix}\sqrt{3} & -1\\ 1 & \sqrt{3}\end{bmatrix}$ The matrix $\boldsymbol{A}_3$ rotates the points by $\frac{\pi}{6}$ rad $= 30°$ counter-clockwise and has only complex eigenvalues, reflecting that the mapping is a rotation (hence, no eigenvectors are drawn). A rotation has to be volume preserving, and so the determinant is 1. For more details on rotations, we refer to Section 3.9.
- $\boldsymbol{A}_4 = \begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}$ represents a mapping in the standard basis that collapses a two-dimensional domain onto one dimension. Since one eigen-

value is 0, the space in direction of the (blue) eigenvector corresponding to $\lambda_1 = 0$ collapses, while the orthogonal (red) eigenvector stretches space by a factor $\lambda_2 = 2$. Therefore, the area of the image is 0.

- $\boldsymbol{A}_5 = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$ is a shear-and-stretch mapping that scales space by 75% since $|\det(\boldsymbol{A}_5)| = \frac{3}{4}$. It stretches space along the (red) eigenvector of λ_2 by a factor 1.5 and compresses it along the orthogonal (blue) eigenvector by a factor 0.5.

Example 4.7 (Eigenspectrum of a Biological Neural Network)



(a) Connectivity matrix.

(b) Eigenspectrum.

Figure 4.5 Caenorhabditis elegans neural network (Kaiser and Hilgetag, 2006).(a) Symmetrized connectivity matrix; (b) Eigenspectrum.

Methods to analyze and learn from network data are an essential component of machine learning methods. The key to understanding networks is the connectivity between network nodes, especially if two nodes are connected to each other or not. In data science applications, it is often useful to study the matrix that captures this connectivity data.

We build a connectivity/adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{277\times277}$ of the complete neural network of the worm *C.Elegans*. Each row/column represents one of the 277 neurons of this worm's brain. The connectivity matrix $\boldsymbol{A}$ has a value of $a_{ij} = 1$ if neuron i talks to neuron j through a synapse, and $a_{ij} = 0$ otherwise. The connectivity matrix is not symmetric, which implies that eigenvalues may not be real valued. Therefore, we compute a symmetrized version of the connectivity matrix as $\boldsymbol{A}_{sym} := \boldsymbol{A} + \boldsymbol{A}^\top$. This new matrix $\boldsymbol{A}_{sym}$ is shown in Figure 4.5(a) and has a nonzero value a_{ij} if and only if two neurons are connected (white pixels), irrespective of the direction of the connection. In Figure 4.5(b), we show the corresponding eigenspectrum of $\boldsymbol{A}_{sym}$. The horizontal axis shows the index of the eigenvalues, sorted in descending order. The vertical axis shows the corresponding eigenvalue. The S-like shape of this eigenspectrum is typical for many biological neural networks. The underlying mechanism responsible for this is an area of active neuroscience research.

Theorem 4.12. *The eigenvectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with n distinct eigenvalues $\lambda_1, \dots, \lambda_n$ are linearly independent.*

This theorem states that eigenvectors of a matrix with n distinct eigenvalues form a basis of $\mathbb{R}^n$.

Definition 4.13. A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is *defective* if it possesses fewer than n linearly independent eigenvectors.

defective

A non-defective matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ does not necessarily require n distinct eigenvalues, but it does require that the eigenvectors form a basis of $\mathbb{R}^n$. Looking at the eigenspaces of a defective matrix, it follows that the sum of the dimensions of the eigenspaces is less than n. Specifically, a defective matrix has at least one eigenvalue λ_i with an algebraic multiplicity $m > 1$ and a geometric multiplicity of less than m.

Remark. A defective matrix cannot have n distinct eigenvalues, as distinct eigenvalues have linearly independent eigenvectors (Theorem 4.12). ♢

Theorem 4.14. *Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, we can always obtain a symmetric, positive semidefinite matrix $\boldsymbol{S} \in \mathbb{R}^{n \times n}$ by defining*

$$\boldsymbol{S} := \boldsymbol{A}^\top \boldsymbol{A} . \tag{4.36}$$

Remark. If $\mathrm{rk}(\boldsymbol{A}) = n$, then $\boldsymbol{S} := \boldsymbol{A}^\top \boldsymbol{A}$ is symmetric, positive definite. ♢

Understanding why Theorem 4.14 holds is insightful for how we can use symmetrized matrices: Symmetry requires $\boldsymbol{S} = \boldsymbol{S}^\top$, and by inserting (4.36) we obtain $\boldsymbol{S} = \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{A}^\top (\boldsymbol{A}^\top)^\top = (\boldsymbol{A}^\top \boldsymbol{A})^\top = \boldsymbol{S}^\top$. Moreover, positive semidefiniteness (Section 3.2.3) requires that $\boldsymbol{x}^\top \boldsymbol{S} \boldsymbol{x} \geqslant 0$ and inserting (4.36) we obtain $\boldsymbol{x}^\top \boldsymbol{S} \boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{x} = (\boldsymbol{x}^\top \boldsymbol{A}^\top)(\boldsymbol{A}\boldsymbol{x}) = (\boldsymbol{A}\boldsymbol{x})^\top (\boldsymbol{A}\boldsymbol{x}) \geqslant 0$, because the dot product computes a sum of squares (which are themselves non-negative).

spectral theorem

Theorem 4.15 (Spectral Theorem)**.** *If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is symmetric, there exists an orthonormal basis of the corresponding vector space V consisting of eigenvectors of $\boldsymbol{A}$, and each eigenvalue is real.*

A direct implication of the spectral theorem is that the eigendecomposition of a symmetric matrix $\boldsymbol{A}$ exists (with real eigenvalues), and that we can find an ONB of eigenvectors so that $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top$, where $\boldsymbol{D}$ is diagonal and the columns of $\boldsymbol{P}$ contain the eigenvectors.

Example 4.8

Consider the matrix

$$\boldsymbol{A} = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix} . \tag{4.37}$$

The characteristic polynomial of $\boldsymbol{A}$ is

$$p_{\boldsymbol{A}}(\lambda) = -(\lambda - 1)^2(\lambda - 7)\,, \tag{4.38}$$

so that we obtain the eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 7$, where λ_1 is a repeated eigenvalue. Following our standard procedure for computing eigenvectors, we obtain the eigenspaces

$$E_1 = \text{span}[\underbrace{\begin{bmatrix}-1\\1\\0\end{bmatrix}}_{=:\boldsymbol{x}_1}, \underbrace{\begin{bmatrix}-1\\0\\1\end{bmatrix}}_{=:\boldsymbol{x}_2}], \quad E_7 = \text{span}[\underbrace{\begin{bmatrix}1\\1\\1\end{bmatrix}}_{=:\boldsymbol{x}_3}]\,. \tag{4.39}$$

We see that $\boldsymbol{x}_3$ is orthogonal to both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. However, since $\boldsymbol{x}_1^\top \boldsymbol{x}_2 = 1 \neq 0$, they are not orthogonal. The spectral theorem (Theorem 4.15) states that there exists an orthogonal basis, but the one we have is not orthogonal. However, we can construct one.

To construct such a basis, we exploit the fact that $\boldsymbol{x}_1, \boldsymbol{x}_2$ are eigenvectors associated with the same eigenvalue λ. Therefore, for any $\alpha, \beta \in \mathbb{R}$ it holds that

$$\boldsymbol{A}(\alpha\boldsymbol{x}_1 + \beta\boldsymbol{x}_2) = \boldsymbol{A}\boldsymbol{x}_1\alpha + \boldsymbol{A}\boldsymbol{x}_2\beta = \lambda(\alpha\boldsymbol{x}_1 + \beta\boldsymbol{x}_2)\,, \tag{4.40}$$

i.e., any linear combination of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is also an eigenvector of $\boldsymbol{A}$ associated with λ. The Gram-Schmidt algorithm (Section 3.8.3) is a method for iteratively constructing an orthogonal/orthonormal basis from a set of basis vectors using such linear combinations. Therefore, even if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are not orthogonal, we can apply the Gram-Schmidt algorithm and find eigenvectors associated with $\lambda_1 = 1$ that are orthogonal to each other (and to $\boldsymbol{x}_3$). In our example, we will obtain

$$\boldsymbol{x}_1' = \begin{bmatrix}-1\\1\\0\end{bmatrix}, \quad \boldsymbol{x}_2' = \frac{1}{2}\begin{bmatrix}-1\\-1\\2\end{bmatrix}, \tag{4.41}$$

which are orthogonal to each other, orthogonal to $\boldsymbol{x}_3$, and eigenvectors of $\boldsymbol{A}$ associated with $\lambda_1 = 1$.

Before we conclude our considerations of eigenvalues and eigenvectors it is useful to tie these matrix characteristics together with the concepts of the determinant and the trace.

Theorem 4.16. *The determinant of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is the product of its eigenvalues, i.e.,*

$$\det(\boldsymbol{A}) = \prod_{i=1}^{n} \lambda_i\,, \tag{4.42}$$

where $\lambda_i \in \mathbb{C}$ are (possibly repeated) eigenvalues of $\boldsymbol{A}$.

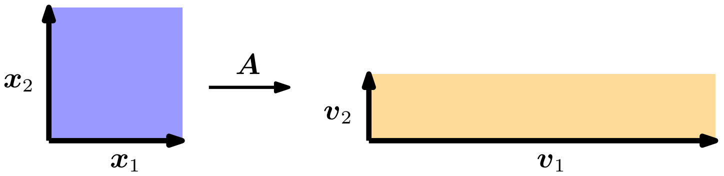


Figure 4.6 Geometric interpretation of eigenvalues. The eigenvectors of $\boldsymbol{A}$ get stretched by the corresponding eigenvalues. The area of the unit square changes by $|\lambda_1\lambda_2|$, the perimeter changes by a factor of $\frac{1}{2}(|\lambda_1| + |\lambda_2|)$.

Theorem 4.17. *The trace of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is the sum of its eigenvalues, i.e.,*

$$tr(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i\,, \tag{4.43}$$

where $\lambda_i \in \mathbb{C}$ are (possibly repeated) eigenvalues of $\boldsymbol{A}$.

Let us provide a geometric intuition of these two theorems. Consider a matrix $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ that possesses two linearly independent eigenvectors $\boldsymbol{x}_1, \boldsymbol{x}_2$. For this example, we assume $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ are an ONB of $\mathbb{R}^2$ so that they are orthogonal and the area of the square they span is 1; see Figure 4.6. From Section 4.1, we know that the determinant computes the change of area of unit square under the transformation $\boldsymbol{A}$. In this example, we can compute the change of area explicitly: Mapping the eigenvectors using $\boldsymbol{A}$ gives us vectors $\boldsymbol{v}_1 = \boldsymbol{A}\boldsymbol{x}_1 = \lambda_1\boldsymbol{x}_1$ and $\boldsymbol{v}_2 = \boldsymbol{A}\boldsymbol{x}_2 = \lambda_2\boldsymbol{x}_2$, i.e., the new vectors $\boldsymbol{v}_i$ are scaled versions of the eigenvectors $\boldsymbol{x}_i$, and the scaling factors are the corresponding eigenvalues λ_i. $\boldsymbol{v}_1, \boldsymbol{v}_2$ are still orthogonal, and the area of the rectangle they span is $|\lambda_1\lambda_2|$.

Given that $\boldsymbol{x}_1, \boldsymbol{x}_2$ (in our example) are orthonormal, we can directly compute the perimeter of the unit square as $2(1+1)$. Mapping the eigenvectors using $\boldsymbol{A}$ creates a rectangle whose perimeter is $2(|\lambda_1| + |\lambda_2|)$. Therefore, the sum of the absolute values of the eigenvalues tells us how the perimeter of the unit square changes under the transformation matrix $\boldsymbol{A}$.

Example 4.9 (Google's PageRank – Webpages as Eigenvectors)

Google uses the eigenvector corresponding to the maximal eigenvalue of a matrix $\boldsymbol{A}$ to determine the rank of a page for search. The idea for the PageRank algorithm, developed at Stanford University by Larry Page and Sergey Brin in 1996, was that the importance of any web page can be approximated by the importance of pages that link to it. For this, they write down all web sites as a huge directed graph that shows which page links to which. PageRank computes the weight (importance) $x_i \geqslant 0$ of a web site a_i by counting the number of pages pointing to a_i. Moreover, PageRank takes into account the importance of the web sites that link to a_i. The navigation behavior of a user is then modeled by a transition matrix $\boldsymbol{A}$ of this graph that tells us with what (click) probability somebody will end up on a different web site. The matrix $\boldsymbol{A}$ has the property that for any initial rank/importance vector $\boldsymbol{x}$ of a web site the sequence $\boldsymbol{x}, \boldsymbol{Ax}, \boldsymbol{A}^2\boldsymbol{x}, \ldots$ converges to a vector $\boldsymbol{x}^*$. This vector is called the *PageRank* and satisfies $\boldsymbol{Ax}^* = \boldsymbol{x}^*$, i.e., it is an eigenvector (with corresponding eigenvalue 1) of $\boldsymbol{A}$. After normalizing $\boldsymbol{x}^*$, such that $\|\boldsymbol{x}^*\| = 1$, we can interpret the entries as probabilities. More details and different perspectives on PageRank can be found in the original technical report (Page et al., 1999).

PageRank

4.3 Cholesky Decomposition

There are many ways to factorize special types of matrices that we encounter often in machine learning. In the positive real numbers, we have the square-root operation that gives us a decomposition of the number into identical components, e.g., $9 = 3 \cdot 3$. For matrices, we need to be careful that we compute a square-root-like operation on positive quantities. For symmetric, positive definite matrices (see Section 3.2.3), we can choose from a number of square-root equivalent operations. The *Cholesky decomposition/Cholesky factorization* provides a square-root equivalent operation on symmetric, positive definite matrices that is useful in practice.

Cholesky decomposition

Cholesky factorization

Theorem 4.18 (Cholesky Decomposition)**.** *A symmetric, positive definite matrix $\boldsymbol{A}$ can be factorized into a product $\boldsymbol{A} = \boldsymbol{LL}^\top$, where $\boldsymbol{L}$ is a lower-triangular matrix with positive diagonal elements:*

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ l_{n1} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & \cdots & l_{n1} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_{nn} \end{bmatrix} . \qquad (4.44)$$

$\boldsymbol{L}$ is called the Cholesky factor of $\boldsymbol{A}$, and $\boldsymbol{L}$ is unique.

Cholesky factor

Example 4.10 (Cholesky Factorization)
Consider a symmetric, positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{3\times 3}$. We are interested in finding its Cholesky factorization $\boldsymbol{A} = \boldsymbol{LL}^\top$, i.e.,

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \boldsymbol{LL}^\top = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix} . \qquad (4.45)$$

Multiplying out the right-hand side yields

$$\boldsymbol{A} = \begin{bmatrix} l_{11}^2 & l_{21}l_{11} & l_{31}l_{11} \\ l_{21}l_{11} & l_{21}^2 + l_{22}^2 & l_{31}l_{21} + l_{32}l_{22} \\ l_{31}l_{11} & l_{31}l_{21} + l_{32}l_{22} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{bmatrix} . \qquad (4.46)$$

Comparing the left-hand side of (4.45) and the right-hand side of (4.46) shows that there is a simple pattern in the diagonal elements l_{ii}:

$$l_{11} = \sqrt{a_{11}}\,, \quad l_{22} = \sqrt{a_{22} - l_{21}^2}\,, \quad l_{33} = \sqrt{a_{33} - (l_{31}^2 + l_{32}^2)}\,. \qquad (4.47)$$

Similarly for the elements below the diagonal (l_{ij}, where $i > j$), there is also a repeating pattern:

$$l_{21} = \frac{1}{l_{11}} a_{21}\,, \quad l_{31} = \frac{1}{l_{11}} a_{31}\,, \quad l_{32} = \frac{1}{l_{22}} (a_{32} - l_{31} l_{21})\,. \qquad (4.48)$$

Thus, we constructed the Cholesky decomposition for any symmetric, positive definite 3×3 matrix. The key realization is that we can backward calculate what the components l_{ij} for the $\boldsymbol{L}$ should be, given the values a_{ij} for $\boldsymbol{A}$ and previously computed values of l_{ij}.

The Cholesky decomposition is an important tool for the numerical computations underlying machine learning. Here, symmetric positive definite matrices require frequent manipulation, e.g., the covariance matrix of a multivariate Gaussian variable (see Section 6.5) is symmetric, positive definite. The Cholesky factorization of this covariance matrix allows us to generate samples from a Gaussian distribution. It also allows us to perform a linear transformation of random variables, which is heavily exploited when computing gradients in deep stochastic models, such as the variational auto-encoder (Jimenez Rezende et al., 2014; Kingma and Welling, 2014). The Cholesky decomposition also allows us to compute determinants very efficiently. Given the Cholesky decomposition $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$, we know that $\det(\boldsymbol{A}) = \det(\boldsymbol{L})\det(\boldsymbol{L}^\top) = \det(\boldsymbol{L})^2$. Since $\boldsymbol{L}$ is a triangular matrix, the determinant is simply the product of its diagonal entries so that $\det(\boldsymbol{A}) = \prod_i l_{ii}^2$. Thus, many numerical software packages use the Cholesky decomposition to make computations more efficient.

4.4 Eigendecomposition and Diagonalization

A *diagonal matrix* is a matrix that has value zero on all off-diagonal elements, i.e., they are of the form

diagonal matrix

$$\boldsymbol{D} = \begin{bmatrix} c_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c_n \end{bmatrix}. \qquad (4.49)$$

They allow fast computation of determinants, powers, and inverses. The determinant is the product of its diagonal entries, a matrix power $\boldsymbol{D}^k$ is given by each diagonal element raised to the power k, and the inverse $\boldsymbol{D}^{-1}$ is the reciprocal of its diagonal elements if all of them are nonzero.

In this section, we will discuss how to transform matrices into diagonal