Principles of Machine Learning

Lecture 2: Matrix Decomposition

Sharif University of Technology
Dept. of Aerospace Engineering

February 13, 2025

Table of Contents

Outline

Introduction Overview

- In this Season, we explore several key matrix decompositions that are essential in numerical linear algebra.
- These tools are crucial in solving systems of equations, data compression, dimensionality reduction, and various applications in machine learning.
- Today, we begin by discussing the motivation and background before delving into determinants and traces.

What are Matrix Decompositions?

- **Definition:** Matrix decompositions involve expressing a given matrix as a product (or sum) of matrices with special properties.
- They simplify many operations in linear algebra by revealing structure hidden within the matrix.
- Common decompositions include:
  - Cholesky Decomposition
  - Eigendecomposition (and Diagonalization)
  - Singular Value Decomposition (SVD)

Why Are Matrix Decompositions Important?

- They provide insight into the underlying structure of matrices.
- They allow efficient numerical computations and stability in solving linear systems.
- Applications include:
  - Signal processing and image compression
  - Principal Component Analysis (PCA) for dimensionality reduction
  - Probabilistic models and statistical inference

Applications in Machine Learning and Data Analysis

- **Dimensionality Reduction:** SVD and eigendecomposition are used in PCA.
- **Optimization:** Decompositions help solve large linear systems efficiently.
- **Probabilistic Models:** Determinants appear in the normalization constants of probability distributions.
- **Data Compression:** Low-rank approximations help compress high-dimensional data.

Outline

Determinant: Definition and Importance

- The **determinant** of a square matrix A (denoted as $\det(A)$) is a scalar summarizing key properties of A.
- It determines:
  - **Invertibility:** A is invertible if and only if $\det(A) \neq 0$.
  - **Volume Scaling:** $|\det(A)|$ scales the volume when transforming space.
  - **Orientation:** The sign of $\det(A)$ indicates if the transformation preserves or reverses orientation.

Geometric Interpretation: Volume Scaling

- For a 3×3 matrix A with column vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$, the absolute value $|\det(A)|$ represents the volume of the parallelepiped spanned by these vectors.
- This interpretation extends to higher dimensions as the hypervolume.

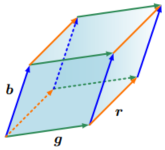

Figure: Volume of a parallelepiped formed by three column vectors

Example: Volume Formed by Column Vectors (Advanced)

- Consider the following three vectors in $\mathbb{R}^3$:

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 4 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 2 \\ -1 \\ 5 \end{bmatrix}.$$

- Form the matrix $A = [\mathbf{v}_1 \ \mathbf{v}_2 \ \mathbf{v}_3]$:

$$A = \begin{bmatrix} 1 & 4 & 2 \\ 2 & 0 & -1 \\ 3 & 1 & 5 \end{bmatrix}.$$

- Compute the determinant using the formula:

$$\det(A) = 1\Big[(0)(5)-(-1)(1)\Big]-4\Big[(2)(5)-(-1)(3)\Big]+2\Big[(2)(1)-(0)(3)\Big].$$

- Thus, the volume of the parallelepiped is $|\det(A)| = 47$.

Laplace Expansion: The Method

- Laplace expansion (or cofactor expansion) is a method to compute the determinant by expanding along a row or column.
- For a matrix A expanding along row i:

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij}),$$

  where A_{ij} is the submatrix obtained by deleting row i and column j.
- This recursive formula reduces the computation to determinants of smaller matrices.

Worked Example: Laplace Expansion

- Consider the 3×3 matrix:

$$A = \begin{bmatrix} 2 & 0 & 1 \\ 3 & 1 & 0 \\ 1 & 4 & 5 \end{bmatrix}.$$

- Expanding along the first row:

$$\det(A) = 2 \cdot \det \begin{bmatrix} 1 & 0 \\ 4 & 5 \end{bmatrix} - 0 \cdot (\cdots) + 1 \cdot \det \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix}.$$

- Compute the 2×2 determinants:

$$\det \begin{bmatrix} 1 & 0 \\ 4 & 5 \end{bmatrix} = 1 \cdot 5 - 0 \cdot 4 = 5,$$

$$\det \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix} = 3 \cdot 4 - 1 \cdot 1 = 11.$$

- Thus, $\det(A) = 2(5) + 1(11) = 10 + 11 = 21.$

Key Properties of Determinants

- **Multiplicative:** $\det(AB) = \det(A)\det(B)$.
- **Transpose:** $\det(A^\top) = \det(A)$.
- **Row Operations:**
  - Swapping two rows multiplies the determinant by -1.
  - Multiplying a row by a scalar k multiplies $\det(A)$ by k.
  - Adding a multiple of one row to another does not change the determinant.
- **Invertibility:** A is invertible if and only if $\det(A) \neq 0$.

Additional Determinant Properties

- **Block Diagonal Matrices:** If A is block-diagonal,

$$A = \begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix},$$

  then $\det(A) = \det(B)\det(C)$.

- **Scaling:** For an $n \times n$ matrix A, if every element is scaled by k, then

$$\det(kA) = k^n \det(A).$$

- These properties are useful in both theoretical derivations and computational algorithms.

Trace: Definition

- The **trace** of a square matrix $A = [a_{ij}]$ is the sum of its diagonal elements:
$$\operatorname{tr}(A) = \sum_{i=1}^{n} a_{ii}.$$
- The trace is a simple invariant that often appears in theory and applications.
- It is particularly useful because it is:
  - **Linear:** $\operatorname{tr}(A + B) = \operatorname{tr}(A) + \operatorname{tr}(B)$ and $\operatorname{tr}(cA) = c\operatorname{tr}(A)$.
  - **Cyclic Invariant:** $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ (when the products are defined).

Properties of the Trace

- **Linearity:** For matrices A and B,

$$\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B).$$

- **Cyclic Property:** For any matrices A and B,

$$\mathrm{tr}(AB) = \mathrm{tr}(BA).$$

- **Similarity Invariance:** If A and B are similar (i.e. $B = P^{-1}AP$), then

$$\mathrm{tr}(A) = \mathrm{tr}(B).$$

- **Eigenvalue Relation:** The trace equals the sum of the eigenvalues of A (with multiplicities).

Trace: An Example

- Consider the matrix:

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

- The trace is:

$$\mathrm{tr}(A) = 4 + 3 = 7.$$

- If the eigenvalues of A are λ_1 and λ_2, then $\lambda_1 + \lambda_2 = 7$.

- This diagram serves as a roadmap to later sections.

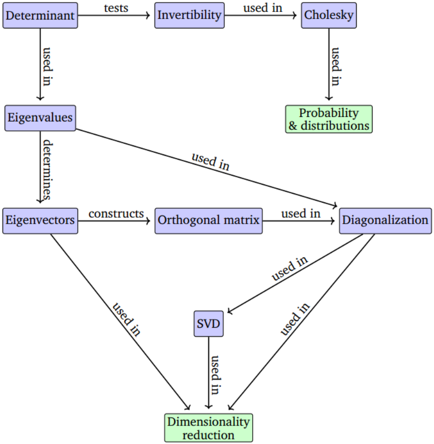

Figure: A mind map of the concepts introduced in this chapter

Outline

Definition of Eigenvalues and Eigenvectors

- Let A be an $n \times n$ matrix.
- A scalar λ is an **eigenvalue** of A if there exists a nonzero vector $\mathbf{x}$ such that
$$A\mathbf{x} = \lambda\mathbf{x}.$$
- The vector $\mathbf{x}$ is called an **eigenvector** corresponding to λ.

How to Obtain Eigenvalues and Eigenvectors

- To find the eigenvalues, solve the **characteristic equation**:

$$\det(A - \lambda I) = 0,$$

  where I is the identity matrix.

- For each eigenvalue λ, the corresponding eigenvectors are found by solving:

$$(A - \lambda I)\,\mathbf{x} = \mathbf{0}.$$

- The nonzero solutions $\mathbf{x}$ form the **eigenspace** associated with λ.

Example: A 2×2 Matrix

- Consider the matrix:

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

- The characteristic equation is:

$$\det\left(\begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix}\right) = (2-\lambda)^2 - 1 = 0.$$

- Solve:

$$(2-\lambda)^2 = 1 \quad \Longrightarrow \quad 2-\lambda = \pm 1,$$

which gives the eigenvalues:

$$\lambda_1 = 1 \quad \text{and} \quad \lambda_2 = 3.$$

Finding the Eigenvectors (Example)

- For $\lambda_1 = 1$:

$$(A - I)\mathbf{x} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \mathbf{x} = \mathbf{0}.$$

The equation $x_1 + x_2 = 0$ yields $x_2 = -x_1$. One valid eigenvector is

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

- For $\lambda_2 = 3$:

$$(A - 3I)\mathbf{x} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{x} = \mathbf{0}.$$

The equation $-x_1 + x_2 = 0$ yields $x_2 = x_1$. One valid eigenvector is

$$\mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Independent Eigenvectors and New Spaces

- If A has n distinct eigenvalues, then its eigenvectors are linearly independent.
- These independent eigenvectors form a basis for $\mathbb{R}^n$ (or $\mathbb{C}^n$) and enable the diagonalization of A:

$$A = PDP^{-1},$$

  where D is a diagonal matrix containing the eigenvalues.
- In cases with repeated eigenvalues, the number of independent eigenvectors (the geometric multiplicity) may be less than the algebraic multiplicity.

Eigenvalues of a Diagonal (Jordan) Matrix

- For a diagonal matrix

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix},$$

  the eigenvalues are simply the diagonal entries.
- In Jordan form, even if A is not diagonalizable, the Jordan blocks reveal the eigenvalues.

Overview of Eigenvalue-Based Transformation Matrices

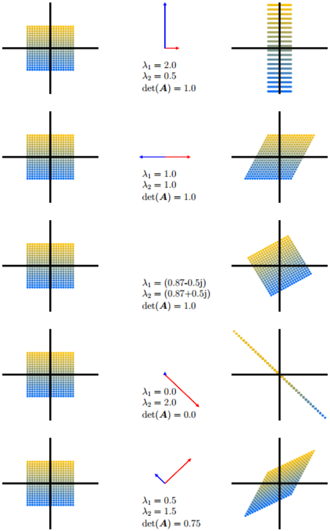

Figure: Overview of five eigenvalue-based transformation matrices

Determinant and Trace via Eigenvalues

- If A has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, then:

$$\det(A) = \prod_{i=1}^{n} \lambda_i \quad \text{and} \quad \operatorname{tr}(A) = \sum_{i=1}^{n} \lambda_i.$$

- These relationships simplify the calculation of these invariants when the eigenvalues are known.

- **Positive Definite:** A symmetric matrix A is *positive definite* if for every nonzero vector $\mathbf{x}$,
$$\mathbf{x}^\top A\mathbf{x} > 0.$$
  Properties:
  - All eigenvalues of A are positive.
  - A is invertible.
  - A admits a unique Cholesky decomposition.
-
- **Negative Definite:** A symmetric matrix A is *negative definite* if for every nonzero vector $\mathbf{x}$,
$$\mathbf{x}^\top A\mathbf{x} < 0.$$
  Properties:
  - All eigenvalues of A are negative.
  - A is invertible.
  - Often used in stability analysis and optimization (e.g., to characterize concavity).

Outline

Why Cholesky Decomposition?

- Many problems in numerical linear algebra involve solving systems of equations of the form $A\mathbf{x} = \mathbf{b}$, where A is a symmetric positive definite (SPD) matrix.
- Cholesky Decomposition is especially efficient for SPD matrices.
- It allows us to decompose A into a product of a lower triangular matrix and its transpose, reducing computational complexity compared to a full LU decomposition.
- This decomposition also improves numerical stability.

Definition: The Cholesky Factorization

- For a symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$, there exists a unique lower-triangular matrix L with positive diagonal entries such that
$$A = L L^\top.$$
- L is called the **Cholesky factor** of A.
- This factorization is valid only when A is SPD.

Computing L: A 3×3 Example

- Consider a 3×3 symmetric positive definite matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}.$$

- We seek a lower-triangular matrix

$$L = \begin{bmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{bmatrix}$$

such that $A = L\,L^\top$. This leads to the following relationships:

  - $a_{11} = \ell_{11}^2 \implies \ell_{11} = \sqrt{a_{11}}$.
  - $a_{12} = \ell_{11}\,\ell_{21} \implies \ell_{21} = \dfrac{a_{12}}{\ell_{11}}$.
  - $a_{13} = \ell_{11}\,\ell_{31} \implies \ell_{31} = \dfrac{a_{13}}{\ell_{11}}$.
  - $a_{22} = \ell_{21}^2 + \ell_{22}^2 \implies \ell_{22} = \sqrt{a_{22} - \ell_{21}^2}$.
  - $a_{23} = \ell_{21}\,\ell_{31} + \ell_{22}\,\ell_{32} \implies \ell_{32} = \dfrac{a_{23} - \ell_{21}\,\ell_{31}}{\ell_{22}}$.

Determinant via Cholesky Decomposition

- Since $A = L\,L^\top$, we have

$$\det(A) = \det(L)\,\det(L^\top) = (\det(L))^2.$$

- As L is lower triangular, its determinant is the product of its diagonal entries:

$$\det(L) = \ell_{11}\,\ell_{22}\,\ell_{33}.$$

- Therefore, the determinant of A can be computed as:

$$\det(A) = \left(\ell_{11}\,\ell_{22}\,\ell_{33}\right)^2.$$

Benefits of Using Cholesky Decomposition

- **Efficiency:** Requires roughly half the number of operations compared to LU decomposition for SPD matrices.
- **Numerical Stability:** Offers a more stable approach when solving systems of linear equations.
- **Simplicity:** The lower-triangular structure of L simplifies both the computation and storage.
- **Determinant and Inversion:** Once L is computed, the determinant and the inverse of A can be easily obtained.
- **Applicability:** Widely used in optimization, machine learning (e.g., Gaussian processes), and statistics (e.g., covariance matrix factorization).

Outline

Diagonal Matrices: D, Its Transpose and Inverse

- A diagonal matrix D is of the form

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

- Since D is diagonal, it is symmetric:

$$D^\top = D.$$

- If all $\lambda_i \neq 0$, its inverse is

$$D^{-1} = \begin{bmatrix} 1/\lambda_1 & 0 & \cdots & 0 \\ 0 & 1/\lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\lambda_n \end{bmatrix},$$

and note that $(D^{-1})^\top = D^{-1}$.

Relationship Between A, P, and D

- If A is diagonalizable, there exists an invertible matrix P and a diagonal matrix D such that

$$A = P\,D\,P^{-1}.$$

- The columns of P are the eigenvectors of A and the diagonal entries of D are the corresponding eigenvalues.

Matrix P Formed from Eigenvectors

- Explicitly, if the eigenvectors of A are

$$v_1,\ v_2,\ \dots,\ v_n,$$

then the matrix P is constructed as

$$P = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & & | \end{bmatrix}.$$

- This forms a basis for $\mathbb{R}^n$ (or $\mathbb{C}^n$) if the eigenvectors are linearly independent.

Intuition Behind Eigendecomposition

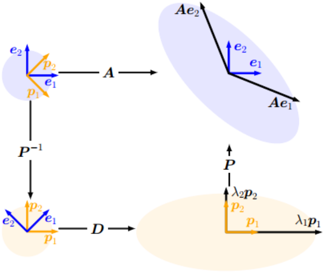

Figure: Intuition: Sequential transformations—projecting onto the eigenvector space, scaling by eigenvalues, and transforming back.

Example: Diagonalizing a Sample Matrix — Finding Eigenvalues

- Consider the matrix

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

- The characteristic equation is

$$\det(A - \lambda I) = \det \begin{bmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{bmatrix} = (4-\lambda)(3-\lambda) - 2 = \lambda^2 - 7\lambda + 10 = 0.$$

- Solving the quadratic equation yields the eigenvalues:

$$\lambda_1 = 5 \quad \text{and} \quad \lambda_2 = 2.$$

Example: Diagonalizing a Sample Matrix — Finding Eigenvectors

- **For $\lambda_1 = 5$:**

$$(A - 5I) = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix}.$$

The equation $-1 \cdot x_1 + 1 \cdot x_2 = 0$ implies $x_2 = x_1$. A valid eigenvector is

$$v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

- **For $\lambda_2 = 2$:**

$$(A - 2I) = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}.$$

The equation $2x_1 + x_2 = 0$ implies $x_2 = -2x_1$. A valid eigenvector is

$$v_2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

Example: Diagonalizing a Sample Matrix — Forming P and D

- Form the matrix P with eigenvectors as columns:

$$P = \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix}.$$

- Form the diagonal matrix D with the eigenvalues on the diagonal:

$$D = \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix}.$$

- Thus, the eigendecomposition of A is

$$A = P\,D\,P^{-1}.$$

Proof: $\det(A) = \det(D)$

- Since $A = PDP^{-1}$, taking determinants gives

$$\det(A) = \det(PDP^{-1}) = \det(P)\,\det(D)\,\det(P^{-1}).$$

- Because $\det(P^{-1}) = 1/\det(P)$, we conclude that

$$\det(A) = \det(D).$$

- This shows that the product of the eigenvalues (the diagonal entries of D) equals $\det(A)$.

Computing A^k via Eigendecomposition

- If $A = PDP^{-1}$, then for any positive integer k,

$$A^k = \left(PDP^{-1}\right)^k = PD^kP^{-1}.$$

- Since D is diagonal, its kth power is given by

$$D^k = \begin{bmatrix} \lambda_1^k & 0 & \cdots & 0 \\ 0 & \lambda_2^k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n^k \end{bmatrix}.$$

- This makes computing high powers of A very efficient.

Summary of Eigendecomposition and Diagonalization

- We introduced the diagonal matrix D and noted that $D^\top = D$ and (if invertible) $(D^{-1})^\top = D^{-1}$.
- We saw that if A is diagonalizable then $A = P\,D\,P^{-1}$ with $P = [v_1, v_2, \ldots, v_n]$ and D containing the eigenvalues.
- The intuition behind eigendecomposition was illustrated as sequential transformations.
- The example demonstrated step-by-step how to diagonalize a matrix, and we proved that $\det(A) = \det(D)$ and that $A^k = P\,D^k\,P^{-1}$.

Outline

Why Use Singular Value Decomposition?

- **SVD** decomposes any $m \times n$ matrix (even if non-square) into a product of three matrices.
- It is widely used for:
  - Dimensionality reduction (e.g., Principal Component Analysis)
  - Data compression and noise reduction
  - Solving ill-posed problems and computing pseudoinverses
- **Definition:** For $A \in \mathbb{R}^{m \times n}$,

$$A = U \Sigma V^\top.$$

The SVD Factorization: $A = U\Sigma V^\top$

- U: An $m \times m$ orthogonal matrix (columns are left singular vectors).
- Σ: An $m \times n$ diagonal (or rectangular diagonal) matrix with nonnegative singular values arranged in nonincreasing order.
- V: An $n \times n$ orthogonal matrix (columns are right singular vectors).
- In the reduced (or "thin") SVD for rank r:

$$A = U_{(m\times r)}\,\Sigma_{(r\times r)}\,V^\top_{(n\times r)}.$$

Dimensions and Shapes of the SVD Matrices

- For $A \in \mathbb{R}^{m \times n}$:
  - U is $m \times m$ (or $m \times r$ in the reduced SVD).
  - Σ is $m \times n$ (or $r \times r$ in the reduced SVD).
  - V is $n \times n$ (or $n \times r$ in the reduced SVD).
- Σ contains the singular values $\sigma_1, \sigma_2, \ldots, \sigma_{\min(m,n)}$ along its diagonal.

What Does Each Matrix Do?

- $V^\top$: Rotates the input into the coordinate system defined by the right singular vectors.
- Σ: Scales the rotated coordinates by the singular values.
- U: Rotates the scaled data into the output space.
- Together, the SVD represents the action of A as a sequence of three transformations: rotation $\rightarrow$ scaling $\rightarrow$ rotation.

The Σ Matrix Structure in SVD

- Given a matrix $A \in \mathbb{R}^{m \times n}$, its SVD is written as:

$$A = U \Sigma V^\top.$$

- The matrix Σ is a rectangular diagonal matrix with dimensions $m \times n$.
- It is structured as follows:

$$\Sigma = \left[\begin{array}{cccc} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p \\ \hline 0 & 0 & \cdots & 0 \end{array}\right],$$

where $p = \min(m, n)$ and the singular values satisfy:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0.$$

- In a reduced (or "thin") SVD, one only uses the first p singular values, resulting in a $p \times p$ diagonal matrix.

How to Calculate Singular Values

- The singular values of A are defined as the nonnegative square roots of the eigenvalues of $A^\top A$ (or $AA^\top$). Since $A^\top A$ is symmetric and positive semidefinite, all its eigenvalues λ_i are nonnegative.
- The process is as follows:
  1. Compute $A^\top A$ (an $n \times n$ matrix).
  2. Solve the eigenvalue equation:
  $$A^\top A\, v_i = \lambda_i\, v_i.$$
  3. Define the singular values as:
  $$\sigma_i = \sqrt{\lambda_i}, \quad \text{for } i = 1, \ldots, \min(m, n).$$
  4. Arrange the singular values in nonincreasing order:
  $$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0.$$
- These singular values populate the diagonal of the Σ matrix.

- Consider a 3×2 matrix mapping $\mathbb{R}^2$ to $\mathbb{R}^3$.
- The SVD can be interpreted as:
  1. **Rotation:** $V^\top$ reorients the 2D input.
  2. **Scaling:** Σ stretches/compresses the rotated coordinates.
  3. **Embedding:** U then maps the scaled result into 3D space.

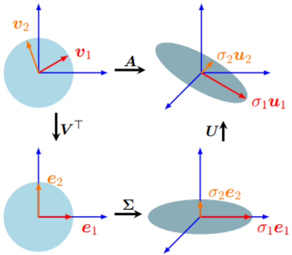

Figure: Sequential transformations in the SVD of a 3×2 matrix.

Example: Calculating the SVD (Part I)

- Consider the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \in \mathbb{R}^{3\times 2}.$$

- **Step 1:** Compute $A^\top A$ (a 2×2 matrix). Its eigenvalues are the squares of the singular values.
- (Assume the computed singular values are approximately)

$$\sigma_1 \approx 9.5255, \quad \sigma_2 \approx 0.5143.$$

Example: Calculating the SVD (Part II)

- **Step 2:** Form the matrix V using the eigenvectors of $A^\top A$. For example:

$$V \approx \begin{bmatrix} -0.6196 & -0.7849 \\ -0.7849 & 0.6196 \end{bmatrix}.$$

- **Step 3:** Compute U as:

$$U = A\,V\Sigma^{-1}.$$

- For the "thin" SVD, U is 3×2 and Σ is 2×2.

Eigenvalue Decomposition vs. Singular Value Decomposition

- **Eigenvalue Decomposition (EVD):**
  - Applicable only to square matrices.
  - May yield complex eigenvalues for non-symmetric matrices.
- **Singular Value Decomposition (SVD):**
  - Applies to any $m \times n$ matrix.
  - Yields nonnegative singular values.
  - Ideal for low-rank approximations and dimensionality reduction.

Derivation of SVD via $A^\top A$

- For $A \in \mathbb{R}^{m\times n}$, consider the symmetric matrix

$$A^\top A \in \mathbb{R}^{n\times n}.$$

- Its eigenvalue decomposition is

$$A^\top A = V\Lambda V^\top,$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $\lambda_i \geq 0$.

- Define the singular values as $\sigma_i = \sqrt{\lambda_i}$. Form the diagonal matrix

$$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n),$$

(with appropriate dimensions).

- Then, the left singular vectors can be obtained by

$$U = A V \Sigma^{-1}.$$

Low-Rank Approximation and the Eckart–Young Theorem

- SVD enables the best rank-k approximation of A.
- Let

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

  and let $A_k = U_k \Sigma_k V_k^\top$ where:
  - U_k contains the first k columns of U.
  - Σ_k is the $k \times k$ diagonal matrix with the largest k singular values.
  - V_k contains the first k columns of V.
- **Eckart–Young Theorem:** A_k minimizes the Frobenius norm $\|A - B\|_F$ among all matrices B of rank k.

Applications of SVD

- **Principal Component Analysis (PCA):** SVD is used to compute principal components.
- **Data Compression:** Low-rank approximations reduce storage requirements.
- **Noise Reduction:** Truncating small singular values can eliminate noise.
- **Pseudo-Inversion:** SVD is used to compute the Moore–Penrose pseudoinverse.
- **Latent Semantic Analysis (LSA):** SVD helps uncover latent relationships in text data.

Computational Aspects of SVD

- SVD is computationally more intensive than eigenvalue decompositions for symmetric matrices.
- Algorithms such as Golub–Kahan bidiagonalization have a time complexity of $O(mn \min(m, n))$.
- Highly optimized libraries (e.g., LAPACK) are used in practice to compute the SVD efficiently.
- Despite its cost, SVD's robustness and generality make it indispensable in many applications.

Detailed Summary of SVD

- For any $A \in \mathbb{R}^{m \times n}$, the SVD is given by:

$$A = U \Sigma V^{\top}.$$

- U and V are orthogonal matrices that rotate the input and output spaces.
- Σ is a diagonal (or rectangular diagonal) matrix that scales the coordinates.
- SVD provides an optimal low-rank approximation and is applicable to both square and rectangular matrices.

Eigenvalue Decomposition vs. Singular Value Decomposition (Revisited)

- **Eigenvalue Decomposition (EVD):**
  - Limited to square matrices.
  - Can produce complex eigenvalues for non-symmetric matrices.
- **Singular Value Decomposition (SVD):**
  - Applicable to any $m \times n$ matrix.
  - Singular values are always nonnegative.
  - Provides robust low-rank approximations and is ideal for data analysis.
- **Conclusion:** SVD is more general and better suited for many real-world applications.

Final Summary of SVD

- SVD expresses A as $A = U\Sigma V^\top$, decomposing its action into rotations and scaling.
- It reveals the intrinsic geometric structure of data.
- SVD is key for low-rank approximations, noise reduction, and many applications in statistics and machine learning.
- Although computationally demanding, its generality and robustness make it an indispensable tool.