Roadmap

(1) Determinant and Trace
(2) Eigenvalues and Eigenvectors
(3) Cholesky Decomposition
(4) Eigendecomposition and Diagonalization
(5) Singular Value Decomposition
(6) Matrix Approximation
(7) Matrix Phylogeny

Summary

- How to summarize matrices: determinants and eigenvalues
- How matrices can be decomposed: Cholesky decomposition, diagonalization, singular value decomposition
- How these decompositions can be used for matrix approximation

Roadmap

(1) Determinant and Trace
(2) Eigenvalues and Eigenvectors
(3) Cholesky Decomposition
(4) Eigendecomposition and Diagonalization
(5) Singular Value Decomposition
(6) Matrix Approximation
(7) Matrix Phylogeny

Determinant: Motivation (1)

- For $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, $\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$.
- $\mathbf{A}$ is invertible iff $a_{11}a_{22} - a_{12}a_{21} \neq 0$
- Let's define $\det(\mathbf{A}) = a_{11}a_{22} - a_{12}a_{21}$.
- Notation: $\det(\mathbf{A})$ or $|\text{whole matrix}|$
- What about 3×3 matrix? By doing some algebra (e.g., Gaussian elimination),

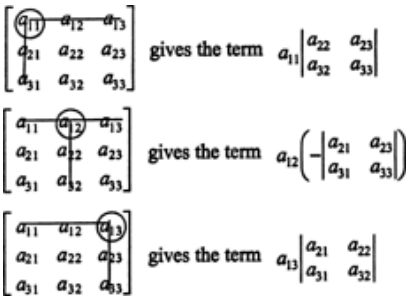
$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} - a_{31}a_{22}a_{13} - a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33}$$

Determinant: Motivation (2)

- Try to find some pattern ...

$$a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} - a_{31}a_{22}a_{13} - a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33} = a_{11}(-1)^{1+1}\det(\mathbf{A}_{1,1}) + a_{12}(-1)^{1+2}\det(\mathbf{A}_{1,2}) + a_{13}(-1)^{1+3}\det(\mathbf{A}_{1,3})$$

- $\mathbf{A}_{k,j}$ is the submatrix of $\mathbf{A}$ that we obtain when deleting row k and column j.



source: www.cliffsnotes.com

- This is called Laplace expansion.
- Now, we can generalize this and provide the formal definition of determinant.

Determinant: Formal Definition

Determinant

For a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, for all $j = 1, \ldots, n$,

Expansion along column j: $\det(\mathbf{A}) = \sum_{k=1}^{n} (-1)^{k+j} a_{kj} \det(\mathbf{A}_{k,j})$

Expansion along row j: $\det(\mathbf{A}) = \sum_{k=1}^{n} (-1)^{k+j} a_{jk} \det(\mathbf{A}_{j,k})$

- All expansions are equal, so no problem with the definition.
- **Theorem.** $\det(\mathbf{A}) \neq 0 \iff \operatorname{rank}(\mathbf{A}) = n \iff \mathbf{A}$ is invertible.

Determinant: Properties

(1) $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$

(2) $\det(\mathbf{A}) = \det(\mathbf{A}^\top)$

(3) For a regular $\mathbf{A}$, $\det(\mathbf{A}^{-1}) = 1/\det(\mathbf{A})$

(4) For two similar matrices $\mathbf{A}, \mathbf{A}'$ (i.e., $\mathbf{A}' = \mathbf{S}^{-1}\mathbf{AS}$ for some $\mathbf{S}$), $\det(\mathbf{A}) = \det(\mathbf{A}')$

(5) For a triangular matrix[1] $\mathbf{T}$, $\det(\mathbf{T}) = \prod_{i=1}^{n} T_{ii}$

(6) Adding a multiple of a column/row to another one does not change $\det(\mathbf{A})$

(7) Multiplication of a column/row with λ scales det by λ. Hence, $\det(\lambda\mathbf{A}) = \lambda^n \det(\mathbf{A})$

(8) Swapping two rows/columns changes the sign of $\det(\mathbf{A})$

- Using (5)-(8), Gaussian elimination (reaching a triangular matrix) enables computing the determinant.

[1] This includes diagonal matrices.

Trace

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

$$\mathrm{tr}(\mathbf{A}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} a_{ii}$$

- $\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$
- $\mathrm{tr}(\alpha \mathbf{A}) = \alpha \mathrm{tr}(\mathbf{A})$
- $\mathrm{tr}(\mathbf{I}_n) = n$

Invariant under Cyclic Permutations

- $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$ for $\mathbf{A} \in \mathbb{R}^{n \times k}$ and $\mathbf{B} \in \mathbb{R}^{k \times n}$
- $\mathrm{tr}(\mathbf{AKL}) = \mathrm{tr}(\mathbf{KLA})$, for $\mathbf{A} \in \mathbb{R}^{a \times k}$, $\mathbf{K} \in \mathbb{R}^{k \times l}$, $\mathbf{L} \in \mathbb{R}^{l \times a}$
- $\mathrm{tr}(\mathbf{x}\mathbf{y}^\top) = \mathrm{tr}(\mathbf{y}^\top \mathbf{x}) = \mathbf{y}^\top \mathbf{x} \in \mathbb{R}$
- A linear mapping $\Phi : V \to V$ can be represented by a matrix $\mathbf{A}$ or another matrix $\mathbf{B}$.
  - $\mathbf{A}$ and $\mathbf{B}$ use different bases, where $\mathbf{B} = \mathbf{S}^{-1}\mathbf{AS}$

$$\mathrm{tr}(\mathbf{B}) = \mathrm{tr}(\mathbf{S}^{-1}\mathbf{AS}) = \mathrm{tr}(\mathbf{ASS}^{-1}) = \mathrm{tr}(\mathbf{A})$$

  - Message. While matrix representations of linear mappings are basis dependent, their traces are not.

Background: Characteristic Polynomial

- **Definition.** For $\lambda \in \mathbb{R}$ and a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, the characteristic polynomial of $\mathbf{A}$ is defined as:

$$\begin{aligned} p_{\mathbf{A}}(\lambda) &\stackrel{\text{def}}{=} \det(\mathbf{A} - \lambda \mathbf{I}) \\ &= c_0 + c_1\lambda + c_2\lambda^2 + \cdots + c_{n-1}\lambda^{n-1} + (-1)^n\lambda^n, \end{aligned}$$

  where $c_0 = \det(\mathbf{A})$ and $c_{n-1} = (-1)^{n-1}\mathrm{tr}(\mathbf{A})$.

- **Example.** For $\mathbf{A} = \begin{pmatrix} 4 & 2 \\ 1 & 3 \end{pmatrix}$,

$$p_{\mathbf{A}}(\lambda) = \begin{vmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{vmatrix} = (4-\lambda)(3-\lambda) - 2 \cdot 1$$

Quick Recap: Determinant & Trace

What we just learned:

- Determinant = one number that tells you if a matrix is invertible ($\det(\mathbf{A}) \neq 0$)
- Trace = sum of diagonal entries
- Both are invariant under basis change — they describe the matrix itself, not a particular representation
- **Next up:** Eigenvalues and eigenvectors — the "DNA" of a matrix

Roadmap

(1) Determinant and Trace
(2) Eigenvalues and Eigenvectors
(3) Cholesky Decomposition
(4) Eigendecomposition and Diagonalization
(5) Singular Value Decomposition
(6) Matrix Approximation
(7) Matrix Phylogeny

Why Should I Care About Eigenvalues?

- Google's PageRank: Ranks billions of web pages using the dominant eigenvector of the web link matrix
- Vibration analysis: Engineers find eigenvalues to predict when a bridge might resonate and collapse
- Facial recognition: "Eigenfaces" — the most important patterns in face images are eigenvectors
- Machine Learning: PCA (Principal Component Analysis) uses eigenvalues to reduce data from 1000s of features to just a few

In one sentence: Eigenvalues reveal the most important directions hidden inside data.

Eigenvalue and Eigenvector

- **Definition.** Consider a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$. Then, $\lambda \in \mathbb{R}$ is an eigenvalue of $\mathbf{A}$ and $\mathbf{x} \in \mathbb{R}^n \setminus \{0\}$ is the corresponding eigenvector of $\mathbf{A}$ if

$$\mathbf{A}\,\mathbf{x} = \lambda\,\mathbf{x}$$

- Equivalent statements
  - λ is an eigenvalue.
  - $(\mathbf{A} - \lambda \mathbf{I}_n)\mathbf{x} = 0$ can be solved non-trivially, i.e., $\mathbf{x} \neq \mathbf{0}$.
  - $\mathrm{rank}(\mathbf{A} - \lambda \mathbf{I}_n) < n$.
  - $\det(\mathbf{A} - \lambda \mathbf{I}_n) = 0 \iff$ The characteristic polynomial $p_{\mathbf{A}}(\lambda) = 0$.

Example

- For $\mathbf{A} = \left(\begin{smallmatrix} 4 & 2 \\ 1 & 3 \end{smallmatrix}\right)$, $p_{\mathbf{A}}(\lambda) = \begin{vmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{vmatrix} = (4-\lambda)(3-\lambda) - 2 \cdot 1 = \lambda^2 - 7\lambda + 10$
- Eigenvalues $\lambda = 2$ or $\lambda = 5$.
- Eigenspace E_5 for $\lambda = 5$

$$\begin{pmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{pmatrix} \mathbf{x} = 0 \implies \begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0 \implies E_5 = \text{span} \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

- Eigenspace E_2 for $\lambda = 2$. Similarly, we get $E_2 = \text{span} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$
- Message. Eigenvectors are not unique.

Properties (1)

- If $\mathbf{x}$ is an eigenvector of $\mathbf{A}$, so are all vectors that are collinear[2].
- E_λ: the set of all eigenvectors for eigenvalue λ, spanning a subspace of $\mathbb{R}^n$. We call this the eigenspace of $\mathbf{A}$ for λ.
- E_λ is the solution space of $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$, thus $E_\lambda = \ker(\mathbf{A} - \lambda\mathbf{I})$
- Geometric interpretation
  - The eigenvector corresponding to a nonzero eigenvalue points in a direction stretched by the linear mapping.
  - The eigenvalue is the factor of stretching.
- Identity matrix $\mathbf{I}$: one eigenvalue $\lambda = 1$ and all vectors $\mathbf{x} \neq \mathbf{0}$ are eigenvectors.

[2] Two vectors are collinear if they point in the same or the opposite direction.

Properties (2)

- $\mathbf{A}$ and $\mathbf{A}^\top$ share the eigenvalues, but not necessarily eigenvectors.
- For two similar matrices $\mathbf{A}, \mathbf{A}'$ (i.e., $\mathbf{A}' = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$ for some $\mathbf{S}$), they possess the same eigenvalues.
  - Meaning: A linear mapping Φ has eigenvalues that are independent of the choice of basis of its transformation matrix.
  - Symmetric, positive definite matrices always have positive, real eigenvalues.

determinant, trace, eigenvalues: all invariant under basis change

Examples for Geometric Interpretation (1)

$\mathbf{A} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{pmatrix}$, $\det(\mathbf{A}) = 1$

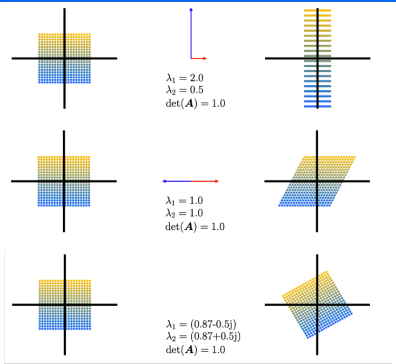
- $\lambda_1 = \frac{1}{2}, \lambda_2 = 2$
- eigenvectors: canonical basis vectors
- area preserving, just vertical and horizontal stretching.

$\mathbf{A} = \begin{pmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{pmatrix}$, $\det(\mathbf{A}) = 1$

- $\lambda_1 = \lambda_2 = 1$
- eigenvectors: collinear over the horizontal line
- area preserving, shearing

$\mathbf{A} = \begin{pmatrix} \cos(\frac{\pi}{6}) & -\sin(\frac{\pi}{6}) \\ \sin(\frac{\pi}{6}) & \cos(\frac{\pi}{6}) \end{pmatrix}$, $\det(\mathbf{A}) = 1$

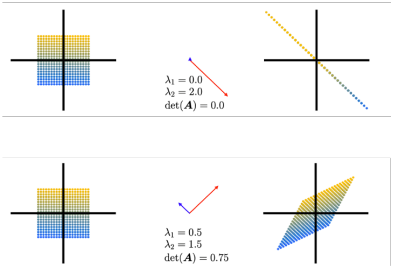
- Rotation by $\pi/6$ counter-clockwise
- only complex eigenvalues (no real eigenvectors)
- area preserving



Examples for Geometric Interpretation (2)

4. $\mathbf{A} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$, $\det(\mathbf{A}) = 0$
   - $\lambda_1 = 0, \lambda_2 = 2$
   - Mapping that collapses a 2D onto 1D
   - area collapses

5. $\mathbf{A} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$, $\det(\mathbf{A}) = 3/4$
   - $\lambda_1 = 0.5, \lambda_2 = 1.5$
   - area scales by 75%, shearing and stretching



Properties (3)

- For $\mathbf{A} \in \mathbb{R}^{n \times n}$, n distinct eigenvalues $\implies$ eigenvectors are linearly independent, which form a basis of $\mathbb{R}^n$.
  - Converse is not true.
  - Example of n linearly independent eigenvectors for fewer than n distinct eigenvalues?
- Determinant. For (possibly repeated) eigenvalues λ_i of $\mathbf{A} \in \mathbb{R}^{n \times n}$,

$$\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i$$

- Trace. For (possibly repeated) eigenvalues λ_i of $\mathbf{A} \in \mathbb{R}^{n \times n}$,

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$$

- **Message.** $\det(\mathbf{A})$ is the area scaling and $\mathrm{tr}(\mathbf{A})$ is the circumference scaling

Quick Recap: Eigenvalues & Eigenvectors

What we just learned:

- An eigenvector is a direction that a matrix only stretches (not rotates)
- The eigenvalue tells you how much it stretches
- $\mathbf{Ax} = \lambda\mathbf{x}$ — the matrix acts like simple multiplication along special directions
- $\det(\mathbf{A}) = \prod \lambda_i$ and $\mathrm{tr}(\mathbf{A}) = \sum \lambda_i$
- **Next up:** Cholesky decomposition — a matrix "square root"

Roadmap

(1) Determinant and Trace
(2) Eigenvalues and Eigenvectors
(3) Cholesky Decomposition
(4) Eigendecomposition and Diagonalization
(5) Singular Value Decomposition
(6) Matrix Approximation
(7) Matrix Phylogeny

Why Should I Care About Cholesky?

- Speed: Solving $\mathbf{Ax} = \mathbf{b}$ with Cholesky is twice as fast as general methods
- Machine Learning: Every Gaussian distribution has a covariance matrix — Cholesky decomposes it efficiently
- Numerical stability: More stable than LU for symmetric positive definite matrices
- Sampling: To generate random data that follows a specific correlation pattern, you need Cholesky

In one sentence: Cholesky is the fast, stable way to work with symmetric positive definite matrices.

LU Decomposition

Upper Triangular Matrix

Lower Triangular Matrix

Source: `http://mathonline.wikidot.com/`

- Gaussian elimination transforms $\mathbf{A}$ into an upper triangular matrix $\mathbf{U}$.
- Each elimination step is an elementary matrix $\mathbf{E}_i$ acting on the left:

$$\mathbf{E}_k \mathbf{E}_{k-1} \cdots \mathbf{E}_1 \mathbf{A} = \mathbf{U}.$$

- Hence,

$$\mathbf{A} = (\mathbf{E}_1^{-1} \mathbf{E}_2^{-1} \cdots \mathbf{E}_k^{-1})\, \mathbf{U} = \mathbf{L}\mathbf{U}.$$

- The matrices $\mathbf{E}_i$ are lower triangular; so are their inverses and their product $\mathbf{L}$.

Cholesky Decomposition

- Analogy for real numbers: a square root, e.g., $9 = 3 \times 3$
- **Theorem.** For a symmetric, positive definite matrix $\mathbf{A}$, $\mathbf{A} = \mathbf{L}\mathbf{L}^\top$, where
  - $\mathbf{L}$ is a lower-triangular matrix with positive diagonals
  - Such a $\mathbf{L}$ is unique, called Cholesky factor of $\mathbf{A}$.
- Applications
  (a) factorization of covariance matrix of a multivariate Gaussian variable
  (b) linear transformation of random variables
  (c) fast determinant computation: $\det(\mathbf{A}) = \det(\mathbf{L})\det(\mathbf{L}^\top) = \det(\mathbf{L})^2$, where $\det(\mathbf{L}) = \prod_i l_{ii}$. Thus, $\det(\mathbf{A}) = \prod_i l_{ii}^2$.

Quick Recap: Cholesky Decomposition

What we just learned:

- LU: Any square matrix = lower $\times$ upper triangular
- Cholesky: For symmetric positive definite matrices: $\mathbf{A} = \mathbf{L}\mathbf{L}^\top$
- Think of it as the "square root" of a matrix
- **Next up:** Eigendecomposition — breaking a matrix into its fundamental pieces

Roadmap

(1) Determinant and Trace
(2) Eigenvalues and Eigenvectors
(3) Cholesky Decomposition
(4) Eigendecomposition and Diagonalization
(5) Singular Value Decomposition
(6) Matrix Approximation
(7) Matrix Phylogeny

Why Should I Care About Eigendecomposition?

- Powers of matrices: Instead of multiplying $\mathbf{A} \times \mathbf{A} \times \cdots \times \mathbf{A}$, just compute $\mathbf{P}\mathbf{D}^k\mathbf{P}^{-1}$ — exponentially faster!
- Understanding transformations: Eigendecomposition shows you exactly *what* a matrix does: stretch along specific directions
- Stability analysis: Are eigenvalues < 1? System is stable. > 1? System explodes.
- Quantum mechanics: Observable quantities (energy, momentum) are eigenvalues of operators

In one sentence: Eigendecomposition breaks a matrix into its simplest building blocks.

Diagonal Matrix and Diagonalization

- Diagonal matrix. zero on all off-diagonal elements, $\mathbf{D} = \begin{pmatrix} d_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & d_n \end{pmatrix}$

$$\mathbf{D}^k = \begin{pmatrix} d_1^k & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & d_n^k \end{pmatrix}, \quad \mathbf{D}^{-1} = \begin{pmatrix} 1/d_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1/d_n \end{pmatrix}, \quad \det(\mathbf{D}) = d_1 d_2 \cdots d_n$$

- **Definition.** $\mathbf{A} \in \mathbb{R}^{n \times n}$ is diagonalizable if it is similar to a diagonal matrix $\mathbf{D}$, i.e., $\exists$ an invertible $\mathbf{P} \in \mathbb{R}^{n \times n}$, such that $\mathbf{D} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$.
- **Definition.** $\mathbf{A} \in \mathbb{R}^{n \times n}$ is orthogonally diagonalizable if it is similar to a diagonal matrix $\mathbf{D}$, i.e., $\exists$ an orthogonal $\mathbf{P} \in \mathbb{R}^{n \times n}$, such that $\mathbf{D} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{P}^{\top}\mathbf{A}\mathbf{P}$.

Power of Diagonalization

- $\mathbf{A}^k = \mathbf{P}\mathbf{D}^k\mathbf{P}^{-1}$
- $\det(\mathbf{A}) = \det(\mathbf{P})\det(\mathbf{D})\det(\mathbf{P}^{-1}) = \det(\mathbf{D}) = \prod_i d_{ii}$
- Many other things ...
- **Question.** Under what condition is $\mathbf{A}$ diagonalizable (or orthogonally diagonalizable) and how can we find $\mathbf{P}$ (thus $\mathbf{D}$)?

Diagonalizability, Algebraic/Geometric Multiplicity

- **Definition.** For a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with an eigenvalue λ_i,
  - the algebraic multiplicity α_i of λ_i is the number of times the root appears in the characteristic polynomial.
  - the geometric multiplicity ζ_i of λ_i is the number of linearly independent eigenvectors associated with λ_i (i.e., the dimension of the eigenspace spanned by the eigenvectors of λ_i)
- **Example.** The matrix $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$ has two repeated eigenvalues $\lambda_1 = \lambda_2 = 2$, thus $\alpha_1 = 2$. However, it has only one distinct unit eigenvector $\mathbf{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, thus $\zeta_1 = 1$.
- **Theorem.** $\mathbf{A} \in \mathbb{R}^{n \times n}$ is diagonalizable $\iff \sum_i \alpha_i = \sum_i \zeta_i = n$.

Orthogonally Diagonalizable and Symmetric Matrix

Theorem. $\mathbf{A} \in \mathbb{R}^{n \times n}$ is orthogonally diagonalizable $\iff$ $\mathbf{A}$ is symmetric.

- **Question.** How to find $\mathbf{P}$ (thus $\mathbf{D}$)?
- Spectral Theorem. If $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric,
  (a) the eigenvalues are all real
  (b) the eigenvectors to different eigenvalues are perpendicular.
  (c) there exists an orthogonal eigenbasis
- For (c), from each set of eigenvectors, say $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ associated with a particular eigenvalue, say λ_j, we can construct another set of eigenvectors $\{\mathbf{x}'_1, \ldots, \mathbf{x}'_k\}$ that are orthonormal, using the Gram–Schmidt process.
- Then, all eigenvectors can form an orthonormal basis.

Example

- **Example.** $\mathbf{A} = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{pmatrix}$. $p_{\mathbf{A}}(\lambda) = -(\lambda - 1)^2(\lambda - 7)$, thus $\lambda_1 = 1, \lambda_2 = 7$

$$E_1 = \text{span}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \quad E_7 = \text{span}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

  - $(111)^\top$ is perpendicular to $(-110)^\top$ and $(-101)^\top$
  - $\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} -1/2 \\ -1/2 \\ 1 \end{pmatrix}$ (for $\lambda = 1$) and $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ (for $\lambda = 7$) are the orthogonal basis in $\mathbb{R}^3$.
  - After normalization, we can make the orthonormal basis.

Eigendecomposition

- **Theorem.** The following are equivalent.
  - (a) A square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ can be factorized into $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$, where $\mathbf{P} \in \mathbb{R}^{n \times n}$ and $\mathbf{D}$ is the diagonal matrix whose diagonal entries are eigenvalues of $\mathbf{A}$.
  - (b) The eigenvectors of $\mathbf{A}$ form a basis of $\mathbb{R}^n$ (i.e., the n eigenvectors of $\mathbf{A}$ are linearly independent)
- The above implies the columns of $\mathbf{P}$ are the n eigenvectors of $\mathbf{A}$ (because $\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{D}$)
- When $\mathbf{A}$ is symmetric, $\mathbf{P}$ is an orthogonal matrix, so $\mathbf{P}^\top = \mathbf{P}^{-1}$
- If $\mathbf{A}$ is symmetric, then (b) holds (Spectral Theorem).

Example of Orthogonal Diagonalization (1)

- Eigendecomposition for $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$
- Eigenvalues: $\lambda_1 = 1, \lambda_2 = 3$
- (normalized) eigenvectors: $\mathbf{p}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, $\mathbf{p}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.
- $\mathbf{p}_1$ and $\mathbf{p}_2$ linearly independent, so $\mathbf{A}$ is diagonalizable.
- $\mathbf{P} = \begin{pmatrix} \mathbf{p}_1 & \mathbf{p}_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$
- $\mathbf{D} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$. Finally, we get $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$

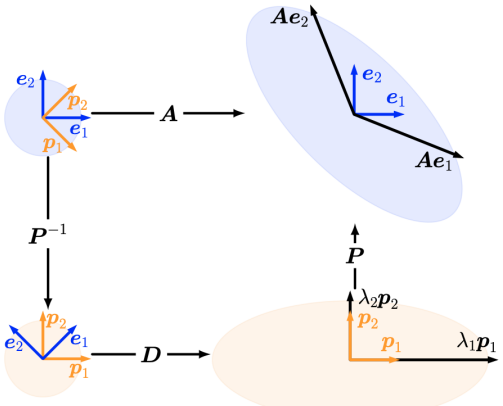
Example of Orthogonal Diagonalization (2)

- $\mathbf{A} = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{pmatrix}$
- Eigenvalues: $\lambda_1 = -1,\ \lambda_2 = 5$ (algebraic mult.: $2, 1$)
- $E_{-1} = \text{span} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$
- Gram–Schmidt $\Rightarrow$

$$\text{span} \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \frac{1}{\sqrt{6}} \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}$$

- $E_5 = \text{span} \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$
- $$\mathbf{P} = \begin{pmatrix} -1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 0 & 2/\sqrt{6} & 1/\sqrt{3} \end{pmatrix}$$
- $$\mathbf{D} = \mathbf{P}^\top \mathbf{A} \mathbf{P} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

Eigendecomposition: Geometric Interpretation



Question. Can we generalize this beautiful result to a general matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$?

Quick Recap: Eigendecomposition

What we just learned:

- Eigendecomposition: $\mathbf{A} = \mathbf{PDP}^{-1}$ — change basis, scale, change back
- Works when $\mathbf{A}$ is square and has enough independent eigenvectors
- For symmetric matrices: $\mathbf{P}$ is orthogonal ($\mathbf{P}^{-1} = \mathbf{P}^{\top}$)
- **Problem:** What if $\mathbf{A}$ is not square? We need something more general...
- **Next up:** SVD — works for *any* matrix!

Roadmap

(1) Determinant and Trace
(2) Eigenvalues and Eigenvectors
(3) Cholesky Decomposition
(4) Eigendecomposition and Diagonalization
(5) Singular Value Decomposition
(6) Matrix Approximation
(7) Matrix Phylogeny

Why Should I Care About SVD?

- Image compression: A 1000×1000 image can be stored using only 50 singular values with almost no visible quality loss
- Recommendation systems: Netflix and Spotify use SVD to predict what you might like
- Natural Language Processing: SVD powers Latent Semantic Analysis — understanding meaning from text
- The universal tool: Unlike eigendecomposition, SVD works for any matrix (any shape, any type)

In one sentence: SVD is the Swiss Army knife of linear algebra — it works everywhere.

Storyline

- Eigendecomposition (also called EVD: EigenValue Decomposition): (Orthogonal) Diagonalization for symmetric matrices $\mathbf{A} \in \mathbb{R}^{n \times n}$.
- Extensions: Singular Value Decomposition (SVD)

  First extension: diagonalization for non-symmetric, but still square matrices $\mathbf{A} \in \mathbb{R}^{n \times n}$

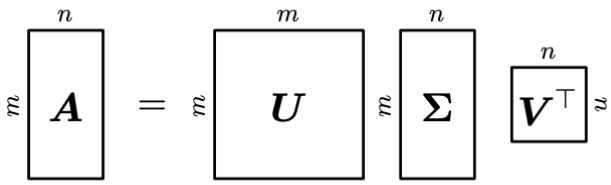
  Second extension: diagonalization for non-symmetric, and non-square matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$

- Background. For $\mathbf{A} \in \mathbb{R}^{m \times n}$, a matrix $\mathbf{S} \stackrel{\text{def}}{=} \mathbf{A}^\top \mathbf{A} \in \mathbb{R}^{n \times n}$ is always symmetric, positive semidefinite.
  - Symmetric, because $\mathbf{S}^\top = (\mathbf{A}^\top \mathbf{A})^\top = \mathbf{A}^\top \mathbf{A} = \mathbf{S}$.
  - Positive semidefinite, because $\mathbf{x}^\top \mathbf{S} \mathbf{x} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{A} \mathbf{x} = (\mathbf{A}\mathbf{x})^\top (\mathbf{A}\mathbf{x}) \geq 0$.
  - If $\mathrm{rank}(\mathbf{A}) = n$, then $\mathbf{S}$ is symmetric and positive definite.

Singular Value Decomposition

- **Theorem.** $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank $r \in [0, \min(m, n)]$. The SVD of $\mathbf{A}$ is a decomposition of the form

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top,$$



with an orthogonal matrix $\mathbf{U} = \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{pmatrix} \in \mathbb{R}^{m \times m}$ and an orthogonal matrix $\mathbf{V} = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix} \in \mathbb{R}^{n \times n}$. Moreover, Σ is an $m \times n$ matrix with $\Sigma_{ii} = \sigma_i \geq 0$ and $\Sigma_{ij} = 0,\ i \neq j$, which is uniquely determined for $\mathbf{A}$.

- Note
  - The diagonal entries $\sigma_i,\ i = 1, \ldots, r$ are called singular values.
  - $\mathbf{u}_i$ and $\mathbf{v}_j$ are called left and right singular vectors, respectively.

SVD: The Geometric Picture

- Every matrix transformation can be broken into three simple steps:

$$\mathbf{A} = \underbrace{\mathbf{U}}_{\text{Rotate}} \underbrace{\mathbf{\Sigma}}_{\text{Stretch}} \underbrace{\mathbf{V}^\top}_{\text{Rotate}}$$

- **Step 1** ($\mathbf{V}^\top$): Rotate/align the input to special directions
- **Step 2** ($\mathbf{\Sigma}$): Stretch each direction by its singular value σ_i
- **Step 3** ($\mathbf{U}$): Rotate the result into the output space

- Analogy: Think of reshaping clay:
  - First, orient the clay along its natural axes
  - Then, squeeze or stretch along each axis
  - Finally, rotate to the desired position

SVD: How It Works (for $\mathbf{A} \in \mathbb{R}^{n \times n}$)

- $\mathbf{A} \in \mathbb{R}^{n \times n}$ with rank $r \leq n$. Then, $\mathbf{A}^\top \mathbf{A}$ is symmetric.
- Orthogonal diagonalization of $\mathbf{A}^\top \mathbf{A}$:

$$\mathbf{A}^\top \mathbf{A} = \mathbf{V} \mathbf{D} \mathbf{V}^\top.$$

- $\mathbf{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$ and an orthogonal matrix $\mathbf{V} = \begin{pmatrix} \mathbf{v}_1 \cdots \mathbf{v}_n \end{pmatrix}$, where $\lambda_1 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$ are the eigenvalues of $\mathbf{A}^\top \mathbf{A}$ and $\{\mathbf{v}_i\}$ are orthonormal.
- All λ_i are non-negative: for eigenvector $\mathbf{v}_i$ with eigenvalue λ_i,

$$0 \leq \|\mathbf{A}\mathbf{v}_i\|^2 = \mathbf{v}_i^\top \mathbf{A}^\top \mathbf{A} \mathbf{v}_i = \lambda_i \|\mathbf{v}_i\|^2 \implies \lambda_i \geq 0$$

- $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}^\top \mathbf{A}) = \operatorname{rank}(\mathbf{D}) = r$
- Choose $\mathbf{U}' = \begin{pmatrix} \mathbf{u}_1 & \cdots \mathbf{u}_r \end{pmatrix}$, where

$$\mathbf{u}_i = \frac{\mathbf{A}\mathbf{v}_i}{\sqrt{\lambda_i}}, \ 1 \leq i \leq r.$$

- We can construct $\{\mathbf{u}_i\}$, $i = r+1, \cdots, n$, so that $\mathbf{U} = \begin{pmatrix} \mathbf{u}_1 & \cdots \mathbf{u}_n \end{pmatrix}$ is an orthonormal basis of $\mathbb{R}^n$.

Example

- $\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{pmatrix}$
- $\mathbf{A}^\top \mathbf{A} = \begin{pmatrix} 5 & -2 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} = \mathbf{V}\mathbf{D}\mathbf{V}^\top,$

$$\mathbf{D} = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \mathbf{V} = \begin{pmatrix} \frac{5}{\sqrt{30}} & \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{30}} \\ 0 & \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix}$$

- $\mathrm{rank}(\mathbf{A}) = 2$ because we have two singular values $\sigma_1 = \sqrt{6}$ and $\sigma_2 = 1$
- $\Sigma = \begin{pmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$

- $\mathbf{u}_1 = \mathbf{A}\mathbf{v}_1/\sigma_1 = \begin{pmatrix} \frac{1}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} \end{pmatrix}$
- $\mathbf{u}_2 = \mathbf{A}\mathbf{v}_2/\sigma_2 = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix}$
- $\mathbf{U} = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{pmatrix} = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$
- Then, we can see that $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$.

EVD ($\mathbf{A} = \mathbf{PDP}^{-1}$) vs. SVD ($\mathbf{A} = \mathbf{U\Sigma V}^\top$)

- SVD: always exists, EVD: square matrix and exists if we can find a basis of eigenvectors (such as symmetric matrices)
- $\mathbf{P}$ in EVD is not necessarily orthogonal (only true for symmetric $\mathbf{A}$), but $\mathbf{U}$ and $\mathbf{V}$ are orthogonal (so representing rotations)
- Both EVD and SVD: (i) basis change in the domain, (ii) independent scaling of each new basis vector and mapping from domain to codomain, (iii) basis change in the codomain. The difference: for SVD, different vector spaces of domain and codomain.
- SVD and EVD are closely related through their projections
  - The left-singular (resp. right-singular) vectors of $\mathbf{A}$ are eigenvectors of $\mathbf{AA}^\top$ (resp. $\mathbf{A}^\top\mathbf{A}$)
  - The singular values of $\mathbf{A}$ are the square roots of eigenvalues of $\mathbf{AA}^\top$ and $\mathbf{A}^\top\mathbf{A}$
  - When $\mathbf{A}$ is symmetric, EVD = SVD (from spectral theorem)

Different Forms of SVD

- When $\text{rank}(\mathbf{A}) = r$, we can construct SVD as the following with only non-zero diagonal entries in $\mathbf{\Sigma}$:

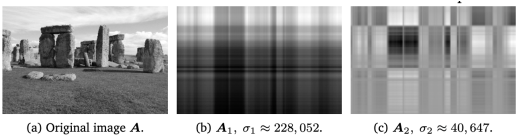
$$\mathbf{A} = \overbrace{\mathbf{U}}^{m \times r} \overbrace{\mathbf{\Sigma}}^{r \times r} \overbrace{\mathbf{V}^\top}^{r \times n}$$

- We can even truncate the decomposed matrices, which can be an approximation of $\mathbf{A}$: for $k < r$

$$\mathbf{A} \approx \overbrace{\mathbf{U}}^{m \times k} \overbrace{\mathbf{\Sigma}}^{k \times k} \overbrace{\mathbf{V}^\top}^{k \times n}$$

We will cover this in the next slides.

Matrix Approximation via SVD



(a) Original image $\boldsymbol{A}$. (b) $\boldsymbol{A}_1$, $\sigma_1 \approx 228,052$. (c) $\boldsymbol{A}_2$, $\sigma_2 \approx 40,647$.

(d) $\boldsymbol{A}_3$, $\sigma_3 \approx 26,125$. (e) $\boldsymbol{A}_4$, $\sigma_4 \approx 20,232$. (f) $\boldsymbol{A}_5$, $\sigma_5 \approx 15,436$.

- $\mathbf{A} = \sum_{i=1}^{r} \sigma_i \underbrace{\mathbf{u}_i \mathbf{v}_i^\top}_{\mathbf{A}_i}$, where $\mathbf{A}_i$ is the outer product[3] of $\mathbf{u}_i$ and $\mathbf{v}_i$
- Rank k-approximation: $\hat{\mathbf{A}}(k) = \sum_{i=1}^{k} \sigma_i \mathbf{A}_i$, $k < r$

[3] If $\mathbf{u}$ and $\mathbf{v}$ are both nonzero, then the outer product matrix $\mathbf{u}\,\mathbf{v}^\top$ always has matrix rank 1. Indeed, the columns of the outer product are all proportional to the first column.

How Close $\hat{\mathbf{A}}(k)$ is to $\mathbf{A}$?

- **Definition.** Spectral Norm of a Matrix. For $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\|\mathbf{A}\|_2 \stackrel{\text{def}}{=} \max_{\mathbf{x}} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$
  - As a concept of length of $\mathbf{A}$, it measures how long any vector $\mathbf{x}$ can at most become, when multiplied by $\mathbf{A}$
- **Theorem.** Eckart–Young. For $\mathbf{A} \in \mathbb{R}^{m \times n}$ of rank r and $\mathbf{B} \in \mathbb{R}^{m \times n}$ of rank k, for any $k \leq r$, we have:

$$\hat{\mathbf{A}}(k) = \arg \min_{\text{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_2, \quad \text{and} \quad \left\|\mathbf{A} - \hat{\mathbf{A}}(k)\right\|_2 = \sigma_{k+1}$$

  - Quantifies how much error is introduced by the SVD-based approximation
  - $\hat{\mathbf{A}}(k)$ is optimal in the sense that such SVD-based approximation is the best one among all rank-k approximations.
  - In other words, it is a projection of the full-rank matrix $\mathbf{A}$ onto a lower-dimensional space of rank-at-most-k matrices.

Quick Recap: SVD & Matrix Approximation

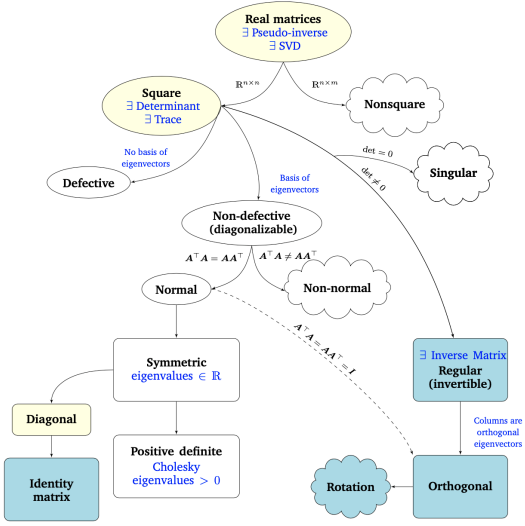
What we just learned:

- SVD: $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ — rotate, stretch, rotate (works for *any* matrix)
- Singular values tell you how important each direction is
- Truncated SVD gives the best rank-k approximation (Eckart–Young theorem)
- **Applications:** image compression, dimensionality reduction, noise removal

Roadmap

(1) Determinant and Trace
(2) Eigenvalues and Eigenvectors
(3) Cholesky Decomposition
(4) Eigendecomposition and Diagonalization
(5) Singular Value Decomposition
(6) Matrix Approximation
(7) Matrix Phylogeny

Phylogenetic Tree of Matrices



Which Decomposition Should I Use?

	Requirement	Form	Best For
LU	Any square matrix	$\mathbf{A} = \mathbf{LU}$	Solving $\mathbf{Ax} = \mathbf{b}$
Cholesky	Symmetric + PD	$\mathbf{A} = \mathbf{LL}^\top$	Covariance, fast solving
Eigendecomp.	Square + basis of eigvecs	$\mathbf{A} = \mathbf{PDP}^{-1}$	Powers $\mathbf{A}^k$, stability
SVD	Any matrix!	$\mathbf{A} = \mathbf{U\Sigma V}^\top$	Approximation, compression

Rule of thumb: Start with SVD — it always works.
Use Cholesky or Eigendecomposition when you know your matrix is special (symmetric, PD, etc.)

Common Mistakes to Avoid

(1) Applying Eigendecomposition to non-square matrices
Eigendecomposition requires $\mathbf{A} \in \mathbb{R}^{n \times n}$. For rectangular matrices, use SVD.

(2) Confusing eigenvalues with singular values
Eigenvalues can be negative or complex. Singular values are always ≥ 0.

(3) Forgetting to check symmetry before Cholesky
Cholesky requires $\mathbf{A}$ to be symmetric *and* positive definite.

(4) Assuming eigenvectors are unique
Any scalar multiple of an eigenvector is also an eigenvector.

(5) Mixing up algebraic and geometric multiplicity
Algebraic = root count in $p_{\mathbf{A}}(\lambda)$. Geometric = dimension of eigenspace.

Key Takeaways

(1) Determinant & Trace summarize a matrix: $\det(\mathbf{A}) = \prod_i \lambda_i$, $\mathrm{tr}(\mathbf{A}) = \sum_i \lambda_i$

(2) Eigendecomposition: $\mathbf{A} = \mathbf{PDP}^{-1}$ reveals how a matrix stretches space along its eigenvectors

(3) Cholesky: $\mathbf{A} = \mathbf{LL}^\top$ for SPD matrices — fast, stable factorization

(4) SVD: $\mathbf{A} = \mathbf{U\Sigma V}^\top$ works for *any* matrix — the most general decomposition

(5) Matrix Approximation: Truncated SVD gives the optimal rank-k approximation (Eckart–Young)

(6) Concept Chain:

$$\text{Eigenvalues} \rightarrow \text{Diagonalization} \rightarrow \text{SVD} \rightarrow \text{Low-Rank Approximation}$$

Questions?

Review Questions (1): Determinant Basics

1) Compute $\det(\mathbf{A})$ for $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ 4 & 3 \end{pmatrix}$ and decide if $\mathbf{A}$ is invertible.
2) Let $\mathbf{A}$ be upper triangular with diagonal entries $(3, -2, 5, 1)$. Compute $\det(\mathbf{A})$.
3) True/False (justify): If $\det(\mathbf{A}) = 0$ then the columns of $\mathbf{A}$ are linearly dependent.
4) Use Laplace expansion along the *first row* to write $\det(\mathbf{A})$ for a 3×3 matrix symbolically.

Review Questions (2): Determinant Properties + Elimination

1) Suppose $\mathbf{B}$ is obtained from $\mathbf{A}$ by swapping two rows. How does $\det(\mathbf{B})$ relate to $\det(\mathbf{A})$?
2) Suppose $\mathbf{B}$ is obtained from $\mathbf{A}$ by multiplying one row by α. How does $\det(\mathbf{B})$ change?
3) Given $\det(\mathbf{A}) = 5$ and $\det(\mathbf{B}) = -2$, compute:
   - $\det(\mathbf{AB})$
   - $\det(\mathbf{A}^\top)$
   - $\det(\mathbf{A}^{-1})$
4) Explain (in one sentence) why Gaussian elimination can be used to compute determinants efficiently.

Review Questions (3): Trace + Cyclic Invariance

1) Compute $\mathrm{tr}(\mathbf{A})$ for $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -1 & 4 \\ 5 & 0 & 2 \end{pmatrix}$.
2) True/False (justify): $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{A})\mathrm{tr}(\mathbf{B})$.
3) Prove using cyclic property:

$$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA}) \quad (\mathbf{A} \in \mathbb{R}^{n \times k},\ \mathbf{B} \in \mathbb{R}^{k \times n}).$$

4) Let $\mathbf{B} = \mathbf{S}^{-1}\mathbf{AS}$ (similar matrices). Show $\mathrm{tr}(\mathbf{B}) = \mathrm{tr}(\mathbf{A})$.

Review Questions (4): Characteristic Polynomial & Eigenvalues

1) For $\mathbf{A} = \begin{pmatrix} 4 & 2 \\ 1 & 3 \end{pmatrix}$ compute $p_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I})$.
2) Solve $p_{\mathbf{A}}(\lambda) = 0$ to obtain eigenvalues.
3) For each eigenvalue, find a basis of its eigenspace E_λ.
4) Explain the equivalence:

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0 \iff (\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0} \text{ has a nontrivial solution.}$$

Review Questions (5): Eigen Geometry + Invariants

1) Give a geometric meaning of an eigenvector/eigenvalue pair $(\lambda, \mathbf{x})$.
2) For $\mathbf{A} = \begin{pmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{pmatrix}$:
   - Find eigenvalues and eigenvectors.
   - Describe the mapping (rotation/shear/stretch/collapse).
3) If $\mathbf{A}' = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$, show $\mathbf{A}$ and $\mathbf{A}'$ have the same eigenvalues.
4) Using eigenvalues $\{\lambda_i\}_{i=1}^{n}$, state formulas for $\det(\mathbf{A})$ and $\mathrm{tr}(\mathbf{A})$.

Review Questions (6): LU & Cholesky

1) Conceptual: In Gaussian elimination, why do elimination matrices $\mathbf{E}_i$ tend to be lower triangular?
2) If $\mathbf{A} = \mathbf{LU}$ (LU), what are the shapes/properties of $\mathbf{L}$ and $\mathbf{U}$?
3) State the condition under which Cholesky exists, and write the factorization form.
4) Let $\mathbf{A}$ be SPD and $\mathbf{A} = \mathbf{L}\mathbf{L}^\top$. Show:

$$\det(\mathbf{A}) = \det(\mathbf{L})^2 = \prod_i \ell_{ii}^2.$$

Review Questions (7): Diagonalization & Multiplicities

1) Define: diagonalizable matrix. What does it mean that $\mathbf{A}$ is similar to a diagonal matrix?
2) For $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$:
   - Find eigenvalues and eigenvectors.
   - Compute algebraic and geometric multiplicities.
   - Decide if $\mathbf{A}$ is diagonalizable.
3) State the diagonalizability criterion in terms of $\sum_i \alpha_i$ and $\sum_i \zeta_i$.
4) If $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$, show $\mathbf{A}^k = \mathbf{P}\mathbf{D}^k\mathbf{P}^{-1}$.

Review Questions (8): Spectral Theorem & Orthogonal Diagonalization

1) State the Spectral Theorem for real symmetric matrices.
2) Prove/justify: $\mathbf{A}$ is orthogonally diagonalizable $\iff$ $\mathbf{A}$ is symmetric.
3) Given eigenvectors for a repeated eigenvalue, explain how Gram–Schmidt is used to build an orthonormal eigenbasis.
4) For a symmetric $\mathbf{A}$, explain why $\mathbf{P}^{-1} = \mathbf{P}^{\top}$ in $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{\top}$.

Review Questions (9): SVD Core Relationships

1) State the SVD theorem: $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$ and define singular values/vectors.
2) Show that $\mathbf{A}^\top\mathbf{A}$ is symmetric PSD and explain why its eigenvalues are ≥ 0.
3) Show the relationship:
   - right singular vectors $\mathbf{v}_i$ are eigenvectors of $\mathbf{A}^\top\mathbf{A}$
   - singular values satisfy $\sigma_i = \sqrt{\lambda_i(\mathbf{A}^\top\mathbf{A})}$
4) When does EVD coincide with SVD? (state the condition and reason)

Review Questions (10): Low-Rank Approximation (Eckart–Young)

1) Write the rank-1 expansion:

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$$

Explain why each term has rank 1.

2) Define the spectral norm:

$$\|\mathbf{A}\|_2 = \max_{\mathbf{x} \neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}.$$

3) State Eckart–Young and interpret:

$$\left\| \mathbf{A} - \hat{\mathbf{A}}(k) \right\|_2 = \sigma_{k+1}.$$

4) Practical: If singular values are $(10, 3, 1, 0.2, 0)$, what is the best rank-2 approximation error in spectral norm?

Theory Link (1): Data as an Operator in Hilbert Space

- Let data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$.
- Covariance operator:

$$\mathbf{C} = \frac{1}{n}\mathbf{X}^\top \mathbf{X}$$

- Spectral theorem:

$$\mathbf{C} = \mathbf{V}\Lambda\mathbf{V}^\top$$

- Eigenvectors = principal directions Eigenvalues = energy along directions.
- Functional view:

$$\mathbf{C}\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

- Interpretation:
  - Data lies near a low-dimensional subspace.
  - Learning = discovering dominant eigenspaces.
- Deep networks implicitly approximate:

  Projection onto dominant spectral components

Theory Link (2): Low-Rank Hypothesis in Deep Models

- Layer weight:

  $$\mathbf{W} \in \mathbb{R}^{m \times n}$$

- Empirical phenomenon:

  $$\mathrm{rank}(\mathbf{W}) \ll \min(m, n)$$

- Optimal rank-k approximation:

  $$\mathbf{W}_k = \arg \min_{\mathrm{rank}(\mathbf{B})=k} \|\mathbf{W} - \mathbf{B}\|_F$$

- Eckart–Young:

  $$\mathbf{W}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\top$$

- Meaning:
  - Networks learn compressed linear operators.
  - Redundant directions collapse spectrally.
- In CNNs:

  $$\text{Filters} \Rightarrow \text{learned basis functions}$$

Theory Link (3): Spectral Geometry of Representations

- Deep features:

$$\mathbf{H} = f(\mathbf{X})$$

- Feature covariance:

$$\mathbf{C}_H = \frac{1}{n}\mathbf{H}^\top \mathbf{H}$$

- Spectrum:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

- If:

$$\lambda_1 \gg \lambda_2 \gg \cdots$$

  Then:

$$\mathrm{rank}(\mathbf{H}) \approx k$$

- Trace identity:

$$\mathrm{tr}(\mathbf{C}_H) = \sum_i \lambda_i$$

- Interpretation:
  - Total variance in learned representation.

Theory Link (4): Spectral View of Optimization Dynamics

- Local training behavior governed by Hessian:

  $$\mathbf{H} = \nabla^2 L(\theta)$$

- Quadratic approximation:

  $$L(\theta + \delta) \approx L(\theta) + \frac{1}{2}\delta^\top \mathbf{H}\delta$$

- Eigen-decomposition:

  $$\mathbf{H} = \mathbf{Q}\Lambda\mathbf{Q}^\top$$

- Dynamics:
  - Large $\lambda_i \Rightarrow$ sharp curvature
  - Small $\lambda_i \Rightarrow$ flat directions
- Generalization hypothesis:

  Flat minima $\Leftrightarrow$ smaller dominant eigenvalues

- Training becomes:

  Spectral shaping of the loss landscape

Theory Link (5): Operator View of Deep Networks

- Linear layer:

  $$\mathbf{x} \mapsto \mathbf{W}\mathbf{x}$$

- SVD:

  $$\mathbf{W} = \mathbf{U}\Sigma\mathbf{V}^\top$$

- Interpretation:
  - $\mathbf{V}^\top$: change of basis (input space)
  - Σ: anisotropic scaling
  - $\mathbf{U}$: rotation (output space)
- Thus each layer:

  $$\text{Rotate} \rightarrow \text{Scale} \rightarrow \text{Rotate}$$

- Deep network:

  $$\mathbf{W}_L \cdots \mathbf{W}_2\mathbf{W}_1$$

- Theoretical insight:
  - Learning = hierarchical spectral factorization.
  - Vision models learn multi-scale spectral operators.

PhD Level Content Only

This section connects matrix decompositions to modern theoretical foundations of deep learning:

- Neural Tangent Kernel (NTK) spectrum
- Fisher Information geometry
- Spectral bias theory
- Implicit regularization
- Operator compression view

Mathematical maturity assumed

PhD Theory (1): Neural Tangent Kernel Spectrum

- Neural network:

  $$f_\theta(x)$$

- Linearization near initialization:

  $$f_\theta(x) \approx f_{\theta_0}(x) + \nabla_\theta f_{\theta_0}(x)^\top (\theta - \theta_0)$$

- NTK matrix:

  $$K(x_i, x_j) = \nabla_\theta f(x_i)^\top \nabla_\theta f(x_j)$$

- Spectral decomposition:

  $$K = \mathbf{Q}\Lambda\mathbf{Q}^\top$$

- Training dynamics:

  $$f_t = (I - e^{-\eta\Lambda t})f^*$$

- Insight:
  - Large eigenvalues learn faster
  - Small eigenvalues learn slower
- Deep learning is spectrally biased

PhD Theory (2): Fisher Information Geometry

- Fisher matrix:

$$\mathbf{F} = \mathbb{E}\left[\nabla_\theta \log p(x|\theta) \nabla_\theta \log p(x|\theta)^\top\right]$$

- Spectral form:

$$\mathbf{F} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$$

- Interpretation:
  - Eigenvectors = sensitive parameter directions
  - Eigenvalues = curvature strength
- Natural gradient:

$$\theta_{t+1} = \theta_t - \eta \mathbf{F}^{-1} \nabla L$$

- Theoretical meaning:

  Learning occurs along dominant Fisher modes

- Connection:
  - Compression
  - Pruning
  - Information bottleneck

PhD Theory (3): Spectral Bias of Deep Networks

- Target function decomposition:

$$f(x) = \sum_i \alpha_i \phi_i(x)$$

- Where ϕ_i are eigenfunctions of the NTK.
- Gradient descent solution:

$$\alpha_i(t) = \alpha_i(0) + (1 - e^{-\eta \lambda_i t})\alpha_i^*$$

- Consequence:
  - Low-frequency modes learned first
  - High-frequency modes learned later
- Known as:

  Spectral Bias (Frequency Principle)

- Vision implication:
  - Networks capture global structure first
  - Fine details later

PhD Theory (4): Implicit Low-Rank Regularization

- Consider linear network:

$$\mathbf{W} = \mathbf{W}_L \cdots \mathbf{W}_1$$

- Gradient descent solution minimizes:

$$\|\mathbf{W}\|_* \quad \text{(nuclear norm)}$$

- Nuclear norm:

$$\|\mathbf{W}\|_* = \sum_i \sigma_i$$

- Equivalent to:

  Promoting low-rank structure

- Hence:
  - Deep networks favor compressed operators
  - Rank collapse emerges naturally
- Explains:
  - Generalization
  - Compression without explicit constraints

PhD Theory (5): Deep Networks as Hierarchical Operators

- Each layer:

$$\mathbf{W} = \mathbf{U}\Sigma\mathbf{V}^\top$$

- Full network:

$$\mathbf{W}_L \cdots \mathbf{W}_1 = (\mathbf{U}_L\Sigma_L\mathbf{V}_L^\top)\cdots(\mathbf{U}_1\Sigma_1\mathbf{V}_1^\top)$$

- Interpretation:
  - Successive rotations
  - Anisotropic scalings
  - Subspace projections
- Depth creates:

  Hierarchical spectral filtering

- Theoretical view:
  - CNNs $\approx$ multi-scale operator decompositions
  - Vision models $\approx$ spectral manifold learners