The University of Danang

University of Science and Technology

MATHEMATICS FOR COMPUTER SCIENCE

Chap 2. . Linear Algebra

Faculty of Information Technology

PHAM Cong Thang, PhD

References

1. M. P. Deisenroth, A. A. Faisal, and C. S. Ong, Mathematics for Machine Learning, Cambridge University Press; 1 edition (April 23, 2020), 398 pages.
2. E. Lehman, F. T. Leighton, A. R. Meyer, 2017, Mathematics for Computer Science, Eric Lehman Google Inc, 998 pages
3. A. Laaksonen, Competitive Programmer's Handbook, 2018, 286 pages.
4. W. H. Press, S. A. Teukolsky, W.T. Vetterling, B. P. Flannery Numerical Recipes: The Art of Scientific Computing, Third Edition, Cambridge University Press, 1262 pages.
5. Other online/offline learning resources

Linear Algebra

- **Introduction**
- Eigenvectors and finding Eigenvectors
- Matrix Decompositions

Introduction

- When formalizing intuitive concepts, a common approach is to construct a set of objects (symbols) and a set of rules to manipulate these objects. This is known as an *algebra*.
- Linear algebra is the study of vectors and certain algebra rules to manipulate vectors.
  - The vectors many of us know from school are called "**geometric vectors**", which are usually denoted by a small arrow above the letter, e.g., $\vec{a}$, $\vec{b}$.
  - A bold letter is used to represent them, e.g., $\boldsymbol{a}$ and $\boldsymbol{b}$.
  - In general, vectors are special objects that can be added together and multiplied by scalars to produce another object of the same kind.

Introduction

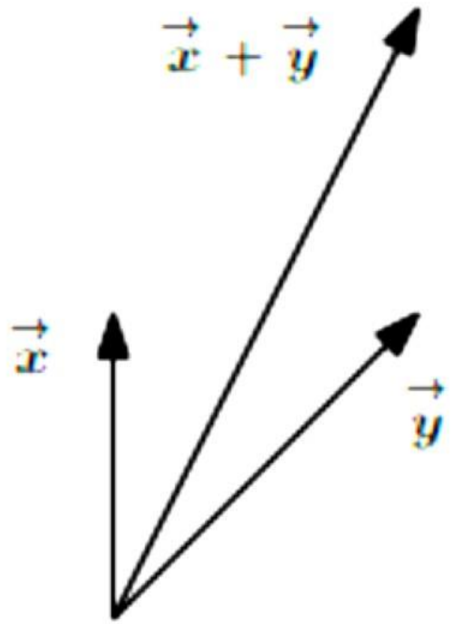
- From an abstract mathematical viewpoint, any object that satisfies these two properties can be considered a vector. Here are some examples of such vector objects:
  - **Geometric vectors**.
  - **Polynomials**
  - **Audio signals are vectors**
  - **Elements of $\mathbb{R}^n$ (tuples of n real numbers) are vectors**

Introduction

- Linear algebra focuses on the similarities between these vector concepts.
  - We can add them together and multiply them by scalars.
  - We will largely focus on vectors in $\mathbb{R}^n$ since most algorithms in linear algebra are formulated in $\mathbb{R}^n$.
  - We often consider data to be represented as vectors in $\mathbb{R}^n$.
  - We will focus on finite-dimensional vector spaces, in which case there is a 1:1 correspondence between any kind of vector and $\mathbb{R}^n$. When it is convenient, we will use intuitions about geometric vectors and consider array-based algorithms.
  - **Linear algebra plays an important role in machine learning and general Mathematics. The concept of a vector space and its properties underlie much of machine learning**

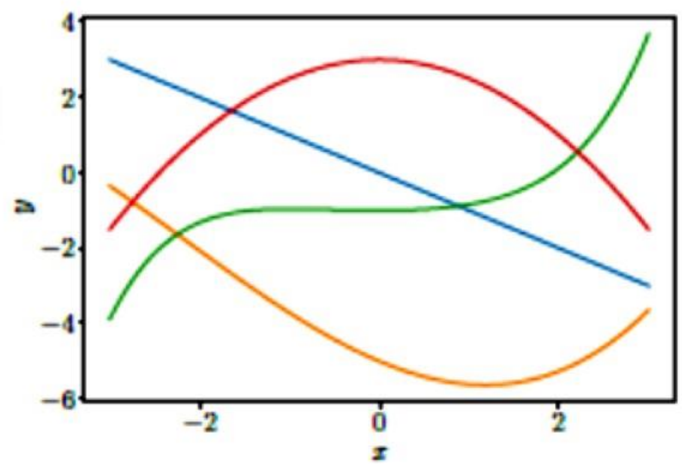
Introduction - vector objects

- **Geometric vectors**. This example of a vector may be familiar from high school mathematics and physics. Geometric vectors are directed segments, which can be drawn (at least in two dimensions).
  - Two geometric vectors $\vec{x}, \vec{y}$ can be added, such that $\vec{x} + \vec{y} = \vec{z}$ is another geometric vector.
  - Multiplication by a scalar $\lambda\vec{x}$ is also a geometric vector. In fact, it is the original vector scaled by λ
  - Interpreting vectors as geometric vectors enables us to use our intuitions about direction and magnitude to reason about mathematical operations.



Introduction - vector objects

- **Polynomials** are also vectors:
  - Two polynomials can be added together, which results in another polynomial
  - They can be multiplied by a scalar $\lambda \in \mathbb{R}$, and the result is a polynomial as well. Therefore, polynomials are (rather unusual) instances of vectors.
  - Note that polynomials are very different from geometric vectors. While geometric vectors are concrete "drawings", polynomials are abstract concepts.



Introduction - vector objects

- **Audio signals** are vectors.
  - Audio signals are represented as a series of numbers.
  - We can add audio signals together, and their sum is a new audio signal.
  - If we scale an audio signal, we also obtain an audio signal.

Introduction - vector objects

- **Elements of $\mathbb{R}^n$ (tuples of n real numbers)** are vectors
  - $\mathbb{R}^n$ is more abstract than polynomials, and it is the concept we focus on in this subject.
  - For instance, $a = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} \in \mathbb{R}^3$ is an example of a triplet of numbers. Adding two vectors $a,\ b \in \mathbb{R}^n$ component-wise results in another vector: $a + b = c \in \mathbb{R}^n$
  - Multiplying $a \in \mathbb{R}^n$ by $\lambda \in \mathbb{R}$ results in a scaled vector $\lambda a \in \mathbb{R}^n$
  - Considering vectors as elements of $\mathbb{R}^n$ has an additional benefit that it loosely corresponds to arrays of real numbers on a computer.
  - **Many programming languages support array operations, which allow for convenient implementation of algorithms that involve vector operations.**

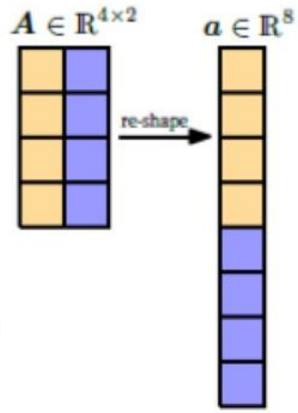
Introduction – Matrix

- Matrices
  - Matrices play a central role in linear algebra.
  - They can be used to compactly represent systems of linear equations, but they also represent linear functions (linear mappings)

Introduction – Matrix

- With $m; n \in N$ a real-valued $(m; n)$ *matrix* A is an $m \cdot n$-tuple of *elements* a_{ij} , $i = 1, \ldots m, j = 1 \ldots n$, which is ordered according to a rectangular scheme consisting of m rows and n columns:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad a_{ij} \in \mathbb{R}$$



  - By convention $(1; n)$ −matrices are called *rows* and $(m; 1)$ −matrices are called column *columns*. These special matrices are also called *row/column vectors*.
  - Rmn is the set of all real-valued $(m; n)$ −matrices. $A \in R^{m\times n}$ can be equivalently represented as a $\in R^{m\times n}$ by stacking all n columns of the matrix into a long vector

Introduction – Matrix operations

- **Matrix Addition and Multiplication:**
  - The sum of two matrices $A \in R^{m \times n}, B \in R^{mn}$ is defined as the elementwise sum, i.e.,

$$\boldsymbol{A} + \boldsymbol{B} := \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

  - For matrices$A \in R^{m \times n}, B \in \mathbb{R}^{n \times k}$, the elements of the product:

$C = AB \in \mathbb{R}^{m \times k}$ are computed as

$$c_{ij} = \sum_{l=1}^{n} a_{il} b_{lj}, \qquad i = 1, \ldots, m, \quad j = 1, \ldots, k$$

Introduction – Matrix operations

- **Inverse and Transpose:**
  - **Inverse:** Consider a square matrix $A \in \mathbb{R}^{n \times n}$. Let matrix $B \in \mathbb{R}^{n \times n}$ have the property that $AB = I_n = BA$. B is called the inverse of A and denoted by A^{-1}.
  - Unfortunately, not every matrix A possesses an inverse A^{-1}.
    - If this regular inverse does exist, A is called *regular/invertible/nonsingular*,
    - otherwise *singular/noninvertible*.

Introduction – Matrix operations

- **Inverse and Transpose:**
  - **Transpose:** For $A \in \mathbb{R}^{m \times n}$ the matrix $B \in \mathbb{R}^{m \times n}$ with $b_{ij} = a_{ji}$ is called the *transpose* of A. We write $B = A^T$
    - In general,A can be obtained by writing the columns of A as the rows of A
    - *Matrix $A \in \mathbb{R}^{n \times n}$ is symmetric if $A = A^T$*
    - *The **sum of symmetric matrices is always symmetric**. However, **their product is always defined, it is generally not symmetric***
- **Multiplication by a Scalar:** Matrices are multiplied by a scalar $\lambda \in \mathbb{R}$.
  - Let $A \in \mathbb{R}^{m \times n}$ and $\lambda \in \mathbb{R}$ then $\lambda A = K$, $K_{ij} = \lambda\, a_{ij}$. Practically, λ scales each element of A

Introduction – Matrix operations

- The determinant $\det(A)$ or $|A|$ of a square matrix A:
  - A number encoding certain properties of the matrix.
  - A matrix is invertible if and only if its determinant is nonzero.
  - The determinant of a product of square matrices equals the product of their determinants: $\det(AB) = \det(A) \cdot \det(B$
  - The determinant of a matrix tell us whether or not the matrix is invertible, can be used to find a formula for the inverse of a matrix, and give us a method for solving linear equations
  - If A has one row of zeros then $detA = 0$
  - If A is a triangular matrix then $detA$ is the product of the entries on the main diagonal

Introduction – Matrix operations

- The determinant $\det(A)$ or $|A|$ of a square matrix A:
  - 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 $A \in \mathbb{R}^{n \times n}$, i.e., matrices with the same number of rows and columns.

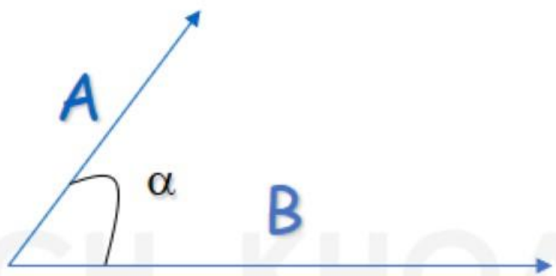
$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}$$

Introduction – Matrix operations

- The determinant $\det(A)$ or $|A|$ of a square matrix A:
  - Adding a multiple of any row to another row, or a multiple of any column to another column, does not change the determinant.
  - Interchanging two rows or two columns affects the determinant by multiplying it by −1.
  - Multiplying one row (or column) by a nonzero number k affects the determinant by multiplying it by k.

Introduction – Vector operations

- Inner (dot) Product: $v \cdot w$
  - The inner product is a **SCALAR**
  - $v \cdot w = (x_1, x_2) \cdot (y_1, y_2) = x_1 y_1 + x_2 y_2$
  - $v \cdot w = \|v\| \, \|v\| \cos(\alpha)$
  - $v \cdot w = 0 \iff v \perp w$
  - vectors $v \cdot w$ are "**columns**", then dot product is $w^T v$
  - $v \cdot w = (x_1, x_2) + (y_1, y_2) = (x_1 + y_1, x_2 + y_2)$
  - $\lambda \cdot v = \lambda(x_1, x_2) = (\lambda x_1, \lambda x_2)$



Introduction – Vector operations

- Dot Product
  - The dot product as a matrix multiplication

$$A \cdot B = A^T B = [a \quad b \quad c] \begin{bmatrix} d \\ e \\ f \end{bmatrix} = ad + be + cf$$

  - The dot product of a vector with itself $\|A\|^2 = A^T A = aa + bb + cc$
  - The dot product is also related to the angle between the two vectors

$$A \cdot B = \|A\| \, \|B\| \cos(\alpha)$$

Introduction – Vector operations

- Norms: the norm of a vector $\|x\|$
  - Absolutely homogeneous: $\lambda\|x\| = \|\lambda\|\|x\|$
  - Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$
  - Positive definite: $\|x\| \geq 0$ and $\|x\| = 0 \Leftrightarrow x = 0$
  - The Manhattan norm on $\mathbb{R}^n$ is defined for $x \in \mathbb{R}^n$ (is also called l_1 norm)

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|,$$

where $\|\cdot\|$ is the absolute value, The Manhattan norm is also called l_1 norm

  - Euclidean Norm on $\mathbb{R}^n$ is defined for $x \in \mathbb{R}^n$ (is also called l_2 norm)

$$\|x\|_1 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x^T x}$$

Introduction – Vector operations

- Norms:
  - l_p norm of x is defined is $(p \geq 1)$:

$$\|x\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$$

  - l_∞ norm (maximum norm or uniform norm) of x is defined

$$\|x\|_p = \max\{|x_1|, |x_2|, \ldots, |x_n|\}$$

Introduction – Vector operations

- Vector Addition:
  - $A + B = (x_1, x_2) + (y_1 + y_2) = (x_1 + x_2, y_1 + y_2)$
- Scalar Product:
  - $\lambda A = \lambda(x_1, x_2) = (\lambda x_1, \lambda x_2)$

Introduction - Matrix Norms

- Common matrix norms for a matrix A

column-sum norm $\quad \|A\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{n} |a_{ij}|$

Frobenius norm

row-sum norm $\quad \|A\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|$

spectral norm (2 norm) $\quad \|A\|_2 = (\mu_{\max})^{1/2}$

μ_{max} is the largest eigenvalue of $A^T A$.

Introduction – Vector operations

- Inner Products
  - Inner products allow for the introduction of intuitive geometrical concepts, such as the length of a vector and the angle or distance between two vectors.
  - A major purpose of inner products is to determine whether vectors are orthogonal to each other.
- *Cross product or vector product is an operation on two vectors in three-dimensional space*

Linear Algebra

- Introduction
- **Eigenvectors and finding Eigenvectors**
- Matrix Decompositions

Eigenvalues and Eigenvectors

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

$$Ax = \lambda x$$

This is the eigenvalue equation, x is called an **eigenvector** of A corresponding to λ

$$Ax = \underbrace{\lambda}_{\text{eigenvalue}} \underbrace{x}_{\text{eigenvector}}$$

Eigenvalues and Eigenvectors

- Computing Eigenvalues, Eigenvectors
  - **Step 1**: **Characteristic Polynomial.** definition of the eigenvector $x \neq 0$ and eigenvalue λ of A, there will be a vector such that $Ax = \lambda x$, i.e.,

$$(A - \lambda I)x = 0$$

  - **Step 2: Eigenvalues.** We need to compute the roots of the characteristic polynomial to find the eigenvalues.
  - **Step 3**: **Eigenvectors**. We find the eigenvectors that correspond to these eigenvalues by looking at vectors x

Eigenvalues and Eigenvectors

- For $A \in \mathbb{R}^{n \times n}$, the set of all eigenvectors of A associated with an eigenvalue spans a subspace eigenspace of $\mathbb{R}^n$, which is called the *eigenspace* of A with respect to λ and is denoted by E_λ . The set of all eigenvalues of A is called the *eigenspectrum*, or just spectrum *spectrum*, of A.
  - If λ is an eigenvalue of $A \in \mathbb{R}^{n \times n}$, then the corresponding eigenspace E_λ is the solution space of the homogeneous system of linear equations

$$(A - \lambda I)x = 0$$

Eigenspaces

- Suppose A is an $n \times n$ matrix and λ is an eigenvalue of A:
  - $E_\lambda(A) = \{x\colon Ax = \lambda x\}$ is an subspace of $\mathbb{R}^n$

$$A = \begin{bmatrix} -3 & -1 \\ 0 & 2 \end{bmatrix} \Rightarrow c_A(x) = \det(xI - A) = \begin{vmatrix} x+3 & 1 \\ 0 & x-2 \end{vmatrix} = (x+3)(x-2)$$

$$c_A(x) = 0 \Leftrightarrow x = -3 \vee x = 2$$

$$x = -3: \left[\begin{array}{cc|c} 0 & 1 & 0 \\ 0 & -5 & 0 \end{array}\right] \to \left[\begin{array}{cc|c} 0 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right] \Rightarrow X = \begin{bmatrix} t \\ 0 \end{bmatrix} \text{ (or X=}(t,0))$$

$$x = 2: \left[\begin{array}{cc|c} 5 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right] \Rightarrow X = \begin{bmatrix} t \\ -5t \end{bmatrix}$$

$$E_{-3} = \{X : AX = -3X\} = \{(t,0) : t \in \mathbb{R}\}$$

$$E_2 = \{X : AX = 2X\} = \{(t,-5t) : t \in \mathbb{R}\}$$

Subspace of $\mathbb{R}^n$

- Definition of subspace of $\mathbb{R}^n$.
  - Let $U \neq \emptyset$ be a subset of $\mathbb{R}^n$
  - U is called a subspace of $\mathbb{R}^n$ if:
    - The zero vector 0 is in U
    - If X, Y are in U then $X + Y$ is in U
    - If X is in U then aX is in U for all real number a.

Subspace of $\mathbb{R}^n$

- Definition of subspace of $\mathbb{R}^n$.
  - Ex1. U={(a,a,0)|$a \in \mathbb{R}$} is a subspace of $\mathbb{R}^3$:
    - the zero vector of R3, (0,0,0) $\in$ U
    - (a,a,0), (b,b,0) $\in$ U$\Rightarrow$ (a,a,0)+(b,b,0)=(a+b,a+b,0) $\in$ U
    - If (a,a,0) $\in$ U and k $\in$ $\mathbb{R}$, then k(a,a,0)=(ka,ka,0) $\in$ U
  - Ex2. U={(a,b,1): a,b $\in$ R} is not a subspace of $\mathbb{R}^3$: (0,0,0) $\notin$ U $\Rightarrow$ U is not a subspace
  - Ex3. U={(a,|a|,0)|a$\in$R} is not a subspace of $\mathbb{R}^3$: (-1,|-1|,0), (1,|1|,0)$\in$U but (0,2,0) $\notin$ U $\Rightarrow$ U is not a subspace

Subspace of $\mathbb{R}^n$

- A subspace either has only one or infinite many vectors
  - Example, {0} has only vector
  - If a subspace U has nonzero vector X then aX is also in U. Then U has infinite many vector

Spanning sets

- $\boldsymbol{y} = \boldsymbol{k_1 x_1} + \boldsymbol{k_2 x_2} + \cdots + \boldsymbol{k_n x_n}$ is called a linear combination of the vectors $x_1, x_2, \ldots, x_n$
- The set of all linear combinations of the the vectors $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ is called the span of these vectors, denoted by ***span{***$\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$***}***

$$\boldsymbol{span}\{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\} = \{\boldsymbol{k_1 x_1} + \boldsymbol{k_2 x_2} + \cdots + \boldsymbol{k_n x_n} : \boldsymbol{k_i} \in \mathbb{R} \text{ is arbitrary}\}$$

  - $U = Span\{X1, X2, \ldots, Xn\}$ is a subspace of $\mathbb{R}^n$. If W is a subspace of $\mathbb{R}^n$ such that x_i are in W then $U \subseteq W$
  - For example, span{(1,0,1),(0,1,1)}={a(1,0,1)+b(0,1,1) :a,b $\in$ $\mathbb{R}$}.
  - And we have (1,2,3) $\in$ span{(1,0,1),(0,1,1)} because (1,2,3)= 1(1,0,1)+ 2(0,1,1).
  - (2,3,2) $\notin$ span{(1,0,1),(0,1,1)} because (2,3,2)≠a(1,0,1)+b(0,1,1) for all a,b

Eigenvalues and Eigenvectors

- **Computing Eigenvalues, Eigenvectors, and Eigenspaces**

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

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

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

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

We factorize the characteristic polynomial and obtain

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

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

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

For $\lambda = 5$ we obtain

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

We solve this homogeneous system and obtain a solution space

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

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

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

Linear Algebra

- Introduction
- Eigenvectors and finding Eigenvectors
- **Matrix Decompositions**

Matrix Decompositions

- 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., $9 = 3 \times 3$.
  - For matrices, we need to be careful that we compute a square-root-like operation on positive quantities.
    - For symmetric, positive definite matrices we can Cholesky choose from a number of square-root equivalent operations. The *Cholesky* decomposition *decomposition/Cholesky factorization* provides a square-root equivalent operation on symmetric, positive definite matrices that is useful in practice.

Matrix Decompositions-Cholesky Decomposition

- A symmetric, positive definite matrix A can be factorized into a product $A = LL^T$, where L is a lower-triangular matrix with positive diagonal elements:

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

L is called the Cholesky factor of A, and L is unique.

Matrix Decompositions-Cholesky Decomposition

- Consider a symmetric, positive definite matrix $A \in \mathbb{R}^{3\times 3}$. We are interested in finding its Cholesky factorization $A = LL^T$, i.e.,

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \boldsymbol{L}\boldsymbol{L}^\top = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

  - Multiplying out the right-hand side yields

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

Matrix Decompositions-Cholesky Decomposition

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

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} l_{11}^2 & l_{21}l_{11} & l_{31}l_{11} \\ l_{21}l_{11} & l_{21}^2 + l_{22}^2 & l_{31}l_{21} + l_{32}l_{22} \\ l_{31}l_{11} & l_{31}l_{21} + l_{32}l_{22} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{bmatrix}$$

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

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

Matrix Decompositions-Cholesky Decomposition

- The Cholesky decomposition is an important tool for the numerical computations underlying machine learning.
- Symmetric positive definite matrices require frequent manipulation, e.g., the covariance matrix of a multivariate Gaussian variable is symmetric, positive definite.

Matrix Decompositions-Cholesky Decomposition

- 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
- The Cholesky decomposition also allows us to compute determinants very efficiently.
  - Given the Cholesky decomposition $A = LL^T$, we know that $\det(A) = \det(L)\det(L^T) = \det(L)^2$.
  - Since L is a triangular matrix, the determinant is simply the product of its diagonal entries so that $\det(A) = \prod_i l_{ii}^2$. Thus, many numerical software packages use the Cholesky decomposition to make computations more efficient.

Matrix Decompositions-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

$$\boldsymbol{D} = \begin{bmatrix} c_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c_n \end{bmatrix}$$

- They allow fast computation of determinants, powers, and inverses. The determinant is the product of its diagonal entries, a matrix power D^k is given by each diagonal element raised to the power k, and the inverse D^{-1} is the reciprocal of its diagonal elements if all of them are nonzero.
  - we will discuss how to transform matrices into diagonal form.

Matrix Decompositions-Eigendecomposition and Diagonalization

- Two matrices A, D are similar if there exists an invertible matrix P, such that

$D = P^{-1}AP.$

  - More specifically, we will look at matrices A that are similar to diagonal matrices D that contain the eigenvalues of A on the diagonal

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

Matrix Decompositions-Eigendecomposition and Diagonalization

- **Diagonalizable.**
  - Let $A \in \mathbb{R}^{n\times n}$, let $\lambda_1, \dots, \lambda_n$ be a set of scalars, and let $p_1, \dots, p_n$ be a set of vectors in $\mathbb{R}^n$. We define $P = [p_1, \dots p_n]$ and let $D \in \mathbb{R}^{n\times n}$ be a diagonal matrix with diagonal entries $\lambda_1, \dots, \lambda_n$. Then we can show that

$$AP = PD$$

if and only if $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A and $p_1, \dots p_n$ are corresponding eigenvectors of A

Matrix Decompositions-Eigendecomposition and Diagonalization

- **Diagonalizable.**
  - We can see that this statement holds because

$$AP = A[p_1, \dots, p_n] = [Ap_1, \dots, Ap_n]$$

$$PD = [p_1, \dots, p_n] \begin{bmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{bmatrix} = [\lambda_1 p_1, \dots, \lambda_n p_n]$$

- Thus $Ap_1 = \lambda p_1, Ap_2 = \lambda p_2, \dots, Ap_n = \lambda p_n$

$\Leftrightarrow$ **the columns of P must be eigenvectors of A.**

Matrix Decompositions-Eigendecomposition and Diagonalization

- **Eigendecomposition.** A square matrix $A \in \mathbb{R}^{n\times n}$ can be factored into

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

where $P \in \mathbb{R}^{n\times n}$ and D is a diagonal matrix whose diagonal entries are the eigenvalues of A, if and only if the eigenvectors of A form a basis of $\mathbb{R}^n$

  - Only non-defective matrices can be diagonalized and that the columns of P are the n eigenvectors of A.
  - For symmetric matrices we can obtain even stronger outcomes for the eigenvalue decomposition.
  - A symmetric matrix $S \in \mathbb{R}^{n\times n}$ can always be diagonalized

Matrix Decompositions-Eigendecomposition and Diagonalization

Diagonalize the matrix $A = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & -1 \\ 1 & 3 & -2 \end{bmatrix}$

Solution. By Example 3.3.4, the eigenvalues of A are $\lambda_1 = 2$, $\lambda_2 = 1$, and $\lambda_3 = -1$, with corresponding basic eigenvectors $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, $\mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$, and $\mathbf{x}_3 = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ respectively. Since the matrix $P = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 3 \end{bmatrix}$ is invertible, Theorem 3.3.4 guarantees that

$$P^{-1}AP = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} = D$$

The reader can verify this directly—easier to check $AP = PD$.

Matrix Decompositions-Eigendecomposition and Diagonalization

Diagonalize the matrix $A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$

Solution. To compute the characteristic polynomial of A first add rows 2 and 3 of $xI - A$ to row 1:

$$c_A(x) = \det \begin{bmatrix} x & -1 & -1 \\ -1 & x & -1 \\ -1 & -1 & x \end{bmatrix} = \det \begin{bmatrix} x-2 & x-2 & x-2 \\ -1 & x & -1 \\ -1 & -1 & x \end{bmatrix}$$
$$= \det \begin{bmatrix} x-2 & 0 & 0 \\ -1 & x+1 & 0 \\ -1 & 0 & x+1 \end{bmatrix} = (x-2)(x+1)^2$$

Hence the eigenvalues are $\lambda_1 = 2$ and $\lambda_2 = -1$, with λ_2 repeated twice (we say that λ_2 has *multiplicity* two). However, A *is* diagonalizable. For $\lambda_1 = 2$, the system of equations $(\lambda_1 I - A)\mathbf{x} = \mathbf{0}$ has general solution $\mathbf{x} = t\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ as the reader can verify, so a basic λ_1-eigenvector is $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

Turning to the repeated eigenvalue $\lambda_2 = -1$, we must solve $(\lambda_2 I - A)\mathbf{x} = \mathbf{0}$. By gaussian elimination, the general solution is $\mathbf{x} = s\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + t\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$ where s and t are arbitrary. Hence the gaussian algorithm produces *two* basic λ_2-eigenvectors $\mathbf{x}_2 = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$ and $\mathbf{y}_2 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$ If we take $P = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$ we find that P is invertible. Hence $P^{-1}AP = \mathrm{diag}(2, -1, -1)$

Matrix Decompositions-Eigendecomposition and Diagonalization

- If A is an $n \times n$ matrix and P is an invertible $n \times n$ matrix then

$$\begin{aligned} A &= PDP^{-1} \\ A^2 &= PD^2P^{-1} \\ A^k &= PD^kP^{-1} \end{aligned}$$

  - Assume that the eigendecomposition $A = PDP^{-1}$ exists. Then

$$\det(A) = \det(PDP^{-1}) = \det(P)\det(D)\det(P^{-1}) = \det(D) = \prod_i d_{ii}$$

allows for an efficient computation of the determinant of A.

Matrix Decompositions-Eigendecomposition and Diagonalization

Let us compute the eigendecomposition of $\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$.

Step 1: Compute eigenvalues and eigenvectors. The characteristic polynomial of $\boldsymbol{A}$ is

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = \det\left(\begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix}\right) = (2-\lambda)^2 - 1 = \lambda^2 - 4\lambda + 3 = (\lambda - 3)(\lambda - 1).$$

Therefore, the eigenvalues of $\boldsymbol{A}$ are $\lambda_1 = 1$ and $\lambda_2 = 3$ (the roots of the characteristic polynomial), and the associated (normalized) eigenvectors are obtained via

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \boldsymbol{p}_1 = 1 \boldsymbol{p}_1, \quad \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \boldsymbol{p}_2 = 3 \boldsymbol{p}_2.$$

This yields

$$\boldsymbol{p}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \boldsymbol{p}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Step 2: Check for existence. The eigenvectors $\boldsymbol{p}_1, \boldsymbol{p}_2$ form a basis of $\mathbb{R}^2$. Therefore, $\boldsymbol{A}$ can be diagonalized.

Step 3: Construct the matrix $\boldsymbol{P}$ to diagonalize $\boldsymbol{A}$. We collect the eigenvectors of $\boldsymbol{A}$ in $\boldsymbol{P}$ so that

$$\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

We then obtain

$$\boldsymbol{P}^{-1} \boldsymbol{A} \boldsymbol{P} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} = \boldsymbol{D}.$$

Equivalently, we get (exploiting that $\boldsymbol{P}^{-1} = \boldsymbol{P}^\top$ since the eigenvectors $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ in this example form an ONB)

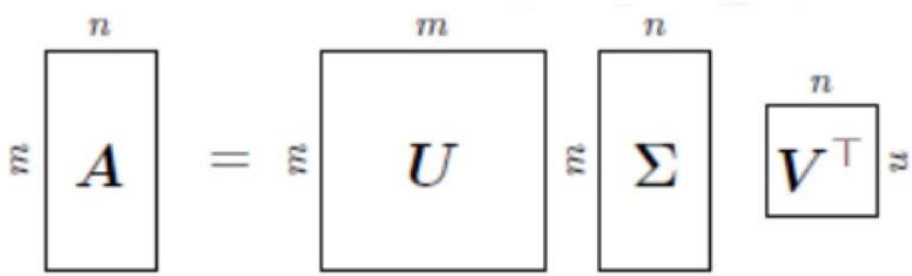
$$\underbrace{\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}}_{\boldsymbol{A}} = \underbrace{\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}}_{\boldsymbol{P}} \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}}_{\boldsymbol{D}} \underbrace{\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}}_{\boldsymbol{P}^\top}.$$

Matrix Decompositions - Singular Value Decomposition

- The singular value decomposition (SVD) of a matrix is a central matrix decomposition method in linear algebra.
  - It has been referred to as the “fundamental theorem of linear algebra” (Strang, 1993) because it can be applied to all matrices, not only to square matrices, and it always exists.

Matrix Decompositions - Singular Value Decomposition

- **SVD Theorem**
  - Let $A^{m \times n}$ be a rectangular matrix of rank $r \in\ [0; \min(m;\ n)]$
  - The SVD of A is a decomposition of the form



with an orthogonal matrix $U \in \mathbb{R}^{m \times m}$ with column vectors $u_i, i\ = 1, \dots, m$, and an orthogonal matrix $V \in \mathbb{R}^{n \times n}$with column vectors $v_j\ , j\ = 1, \dots, n$. Moreover, is an $m \times\ n$ matrix with $\Sigma_{ii}\ = \sigma_i >\ 0$ and $\Sigma_{ii} = 0, i \neq j$

Matrix Decompositions - Singular Value Decomposition

- The diagonal entries $\sigma_i, i = 1, \ldots, r$ of Σ are called the *singular values*, u_i are called the *left-singular vectors*, and v_j are called the *right-singular vectors*. By convention, the singular values are ordered, i.e., $\sigma_1 \geq \sigma_2 \geq \sigma_r \geq 0$
- The *singular value matrix* is unique, but it requires some attention. Observe that the $\Sigma \in \mathbb{R}^{m \times m}$ is rectangular. In particular, Σ is of the same size as A.
  - This means that has a diagonal submatrix that contains the singular values and needs additional zero padding.

Matrix Decompositions - Singular Value Decomposition

- The singular value matrix
  - If $m > n$, then the matrix Σ has diagonal structure up to row n and then consists of 0^T row vectors from $n + 1$ to m below so that:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix}$$

  - If $m < n$, the matrix has a diagonal structure up to column m and columns that consist of 0 from $m + 1$ to n:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & \ddots & 0 & 0 & & 0 \\ 0 & 0 & \sigma_m & 0 & \dots & 0 \end{bmatrix} .$$

  - The SVD exists for any matrix $A \in \mathbb{R}^{m \times n}$.

Matrix Decompositions - Singular Value Decomposition

- Construction of the SVD
  - The SVD of a general matrix shares some similarities with the eigendecomposition of a square matrix
  - Compare the eigendecomposition of an SVD matrix $S = S^T = PDP^T$

  with the corresponding SVD: $S = U\Sigma V^T$
  - If we set $U = P = V, D = \Sigma$, we see that the SVD of SPD matrices is their eigendecomposition
  - Computing the SVD of $A \in \mathbb{R}^{m\times m}$ is equivalent to finding two sets of orthonormal bases $U = (u_1, \dots, u_m)$ and $V = (v_1, \dots, v_n)$ of the codomain $\mathbb{R}^m$ and the domain $\mathbb{R}^n$, respectively. From these ordered bases, we will construct the matrices U and V .

Matrix Decompositions - Singular Value Decomposition

- Construction of the SVD
  - We start with constructing the orthonormal set of right-singular vectors $v_1, \dots, v_n \in \mathbb{R}^n$
  - Then, we then construct the orthonormal set of left-singular vectors $u_1, \dots, u_n \in \mathbb{R}^m$
  - Thereafter, we will link the two and require that the orthogonality of the v_i is preserved under the transformation of A
    - This is important because we know that the images Av_i form a set of orthogonal vectors. We will then normalize these images by scalar factors, which will turn out to be the singular values

Matrix Decompositions - Singular Value Decomposition

- Construction of the SVD
  - a symmetric matrix possesses an orthonormal basis of eigenvectors, which also means it can be diagonalized. Moreover,
  - we can always construct a symmetric, positive semidefinite matrix A$\in \mathbb{R}^{m\times n}$ from any rectangular matrix $S \in \mathbb{R}^{n\times n}$: $S = A^T A$
    - If $rank(A) = n, S = A^T A$ is symmetric, positive definite.
    - we can always diagonalize $A^T A$ and obtain

$$\boldsymbol{A}^{\top}\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{\top} = \boldsymbol{P}\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}\boldsymbol{P}^{\top}$$

where P is an orthogonal matrix, which is composed of the orthonormal eigenbasis. The $\lambda_i \geq 0$ are the eigenvalues of $A^T A$

Matrix Decompositions - Singular Value Decomposition

- Construction of the SVD
  - Let us assume the SVD of A exists

$$A^\top A = (U\Sigma V^\top)^\top(U\Sigma V^\top) = V\Sigma^\top U^\top U\Sigma V^\top$$

where U, V are orthogonal matrices

  - With $U^T U = I$ we obtain

$$A^\top A = V\Sigma^\top\Sigma V^\top = V\begin{bmatrix}\sigma_1^2 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \sigma_n^2\end{bmatrix}V^\top$$

  - We identify

$$V^\top = P^\top,$$
$$\sigma_i^2 = \lambda_i.$$

with

$$A^\top A = PDP^\top = P\begin{bmatrix}\lambda_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \lambda_n\end{bmatrix}P^\top$$

Matrix Decompositions - Singular Value Decomposition

- Construction of the SVD
  - Therefore, the eigenvectors of $A^T A$ that compose P are the right-singular vectors V of A
  - The eigenvalues of $A^T A$ are the squared singular values of Σ

$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}\begin{bmatrix}\sigma_1^2 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \sigma_n^2\end{bmatrix}\boldsymbol{V}^\top$$

$$\boldsymbol{V}^\top = \boldsymbol{P}^\top,$$

$$\sigma_i^2 = \lambda_i .$$

$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top = \boldsymbol{P}\begin{bmatrix}\lambda_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \lambda_n\end{bmatrix}\boldsymbol{P}^\top$$

Matrix Decompositions - Singular Value Decomposition

- Construction of the SVD
  - To obtain the left-singular vectors U, we follow a similar procedure
  - We start by computing the SVD of the symmetric matrix $AA^T \in \mathbb{R}^{m\times m}$ (instead of the previous $AA^T \in \mathbb{R}^{n\times n}$). The SVD of A yields:

$$\boldsymbol{A}\boldsymbol{A}^\top = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top$$

$$= \boldsymbol{U}\begin{bmatrix}\sigma_1^2 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \sigma_m^2\end{bmatrix}\boldsymbol{U}^\top$$

  - AA^T can be diagonalized and we can find an orthogonal basis of eigenvectors of AA^T, which are collected in S. The orthonormal eigenvectors of AA^T are the left-singular vectors U and form an orthonormal basis set in the codomain of the SVD.
  - Since AA^T and AA^T have the same nonzero eigenvalues the nonzero entries of the Σ matrices in the SVD for both cases have to be the same.

Matrix Decompositions - Singular Value Decomposition

- Construction of the SVD
  - The last step is to link up all the parts we touched upon so far. We have an orthonormal set of right-singular vectors in V .
  - To finish the construction of the SVD, we connect them with the orthonormal vectors U.
  - To reach this goal, we use the fact the images of the v_i under A have to be orthogonal, too.
  - We require that the inner product between Av_i and Av_j must be 0 for $i \neq j$.
    - For any two orthogonal eigenvectors v_i, v_j , $i \neq j$, it holds that

$$(\boldsymbol{A}\boldsymbol{v}_i)^\top(\boldsymbol{A}\boldsymbol{v}_j) = \boldsymbol{v}_i^\top(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}_j = \boldsymbol{v}_i^\top(\lambda_j\boldsymbol{v}_j) = \lambda_j\boldsymbol{v}_i^\top\boldsymbol{v}_j = 0$$

    - For the case $m \geq r$, it holds that $\{Av_1, \ldots, Av_r\}$is a basis of an $r-$ dimensional subspace of $\mathbb{R}^m$

Matrix Decompositions - Singular Value Decomposition

- Construction of the SVD
  - To complete the SVD construction, we need left-singular vectors that are orthonormal: We normalize the images of the right-singular vectors Av_i and obtain

$$\boldsymbol{u}_i := \frac{\boldsymbol{A}\boldsymbol{v}_i}{\|\boldsymbol{A}\boldsymbol{v}_i\|} = \frac{1}{\sqrt{\lambda_i}}\boldsymbol{A}\boldsymbol{v}_i = \frac{1}{\sigma_i}\boldsymbol{A}\boldsymbol{v}_i$$

  - the eigenvalues of AA^T are such that $\sigma_i^2 = \lambda_i$
  - Therefore, the eigenvectors of AA^T , which we know are the right-singular vectors v_i, and their normalized images under A, the left-singular vectors u_i , form two self-consistent Orthonormal bases that are connected through the singular value matrix Σ

Matrix Decompositions - Singular Value Decomposition

- Construction of the SVD
  - We obtain the *singular value equation*

$$\boldsymbol{u}_i := \frac{\boldsymbol{A}\boldsymbol{v}_i}{\|\boldsymbol{A}\boldsymbol{v}_i\|} = \frac{1}{\sqrt{\lambda_i}}\boldsymbol{A}\boldsymbol{v}_i = \frac{1}{\sigma_i}\boldsymbol{A}\boldsymbol{v}_i \qquad \boxed{\boldsymbol{A}\boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i\,, \quad i = 1, \ldots, r}$$

  - This equation closely resembles the eigenvalue equation, but the vectors on the left- and the right-hand sides are not the same.
  - For $n > m$, the equation holds only for $i \leq m$ and says nothing about the u_i for $i > m$. However, we know by construction that they are orthonormal.
  - Conversely, for $m > n$, the equation holds only for $i \leq n$. For $i \geq n$, we have $Av_i = 0$ and we still know that the vi form an orthonormal set.
  - This means that the SVD also supplies an orthonormal basis of the kernel (null space) of A, the set of vectors x with $Ax = 0$

Matrix Decompositions - Singular Value Decomposition

- Construction of the SVD
  - Moreover, concatenating the v_i as the columns of V and the u_i as the columns of U yields

$$AV = U\Sigma$$

where Σ has the same dimensions as A and a diagonal structure for rows $1, \ldots, r$. Hence, right-multiplying with V^T yields $A = U\Sigma V^T >$, which is the SVD of A.

Matrix Decompositions - Singular Value Decomposition

Let us find the singular value decomposition of

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

The SVD requires us to compute the right-singular vectors v_j, the singular values σ_k, and the left-singular vectors u_i.

Step 1: Right-singular vectors as the eigenbasis of $A^\top A$.

We start by computing

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

We compute the singular values and right-singular vectors v_j through the eigenvalue decomposition of $A^\top A$, which is given as

$$A^\top A = \begin{bmatrix} \frac{5}{\sqrt{30}} & 0 & \frac{-1}{\sqrt{6}} \\ \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} 6 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{5}{\sqrt{30}} & \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{30}} \\ 0 & \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix} = PDP^\top,$$

and we obtain the right-singular vectors as the columns of P so that

$$V = P = \begin{bmatrix} \frac{5}{\sqrt{30}} & 0 & \frac{-1}{\sqrt{6}} \\ \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} \end{bmatrix}.$$

Step 2: Singular-value matrix.

As the singular values σ_i are the square roots of the eigenvalues of $A^\top A$ we obtain them straight from D. Since $\text{rk}(A) = 2$, there are only two nonzero singular values: $\sigma_1 = \sqrt{6}$ and $\sigma_2 = 1$. The singular value matrix must be the same size as A, and we obtain

$$\Sigma = \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Step 3: Left-singular vectors as the normalized image of the right-singular vectors.

We find the left-singular vectors by computing the image of the right-singular vectors under A and normalizing them by dividing them by their corresponding singular value. We obtain

$$u_1 = \frac{1}{\sigma_1} A v_1 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{5}{\sqrt{30}} \\ \frac{-2}{\sqrt{30}} \\ \frac{1}{\sqrt{30}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} \end{bmatrix},$$

$$u_2 = \frac{1}{\sigma_2} A v_2 = \frac{1}{1} \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix},$$

$$U = [u_1, u_2] = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}.$$

Note that on a computer the approach illustrated here has poor numerical behavior, and the SVD of A is normally computed without resorting to the eigenvalue decomposition of $A^\top A$.

Matrix Decompositions - Singular Value Decomposition

- Eigenvalue Decomposition vs. Singular Value Decomposition
  - Let us consider the eigendecomposition $A = PDP^{-1}$ and the SVD $A = U\Sigma V^T$ and review the core elements of the past sections.
    - The SVD always exists for any matrix $\mathbb{R}^{m\times n}$
    - The eigendecomposition is only defined for square matrices $\mathbb{R}^{n\times n}$ and only exists if we can find a basis of eigenvectors of $\mathbb{R}^n$

Matrix Decompositions - Singular Value Decomposition

- Eigenvalue Decomposition vs. Singular Value Decomposition
  - The vectors in the eigendecomposition matrix P are not necessarily orthogonal, i.e., the change of basis is not a simple rotation and scaling.
  - On the other hand, the vectors in the matrices U and V in the SVD are orthonormal, so they do represent rotations.
  - Both the eigendecomposition and the SVD are compositions of three linear mappings:
    - Change of basis in the domain
    - Independent scaling of each new basis vector and mapping from domain to codomain
    - Change of basis in the codomain

Matrix Decompositions - Singular Value Decomposition

- Eigenvalue Decomposition vs. Singular Value Decomposition
  - Example: Movie ratings of three people for four movies and its SVD

$$\begin{array}{r|ccc} & \text{Ali} & \text{Beatrix} & \text{Chandra} \\ \hline \text{Star Wars} & 5 & 4 & 1 \\ \text{Blade Runner} & 5 & 5 & 0 \\ \text{Amelie} & 0 & 0 & 5 \\ \text{Delicatessen} & 1 & 0 & 4 \end{array}$$

$$\begin{bmatrix} 5 & 4 & 1 \\ 5 & 5 & 0 \\ 0 & 0 & 5 \\ 1 & 0 & 4 \end{bmatrix} = \begin{bmatrix} -0.6710 & 0.0236 & 0.4647 & -0.5774 \\ -0.7197 & 0.2054 & -0.4759 & 0.4619 \\ -0.0939 & -0.7705 & -0.5268 & -0.3464 \\ -0.1515 & -0.6030 & 0.5293 & -0.5774 \end{bmatrix} \begin{bmatrix} 9.6438 & 0 & 0 \\ 0 & 6.3639 & 0 \\ 0 & 0 & 0.7056 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -0.7367 & -0.6515 & -0.1811 \\ 0.0852 & 0.1762 & -0.9807 \\ 0.6708 & -0.7379 & -0.0743 \end{bmatrix}$$

Matrix Decompositions - Singular Value Decomposition

- Eigenvalue Decomposition vs. Singular Value Decomposition
  - A key difference between the eigendecomposition and the SVD :
    - In the SVD, domain and codomain can be vector spaces of different dimensions.
    - In the SVD, the left- and right-singular vector matrices U and V are generally not inverse of each other (they perform basis changes in different vector spaces).
    - In the eigendecomposition, the basis change matrices P and P^{-1} are inverses of each other.
    - In the SVD, the entries in the diagonal matrix Σ are all real and nonnegative, which is not generally true for the diagonal matrix in the eigendecomposition.

Matrix Decompositions - Singular Value Decomposition

- Eigenvalue Decomposition vs. Singular Value Decomposition
  - A key difference between the eigendecomposition and the SVD :
    - The SVD and the eigendecomposition are closely related through their projections
      - The left-singular vectors of A are eigenvectors of AA^T
      - The right-singular vectors of A are eigenvectors of AA^T.
      - The nonzero singular values of A are the square roots of the nonzero eigenvalues of AA^T and are equal to the nonzero eigenvalues of AA^T
    - For symmetric matrices $A \in \mathbb{R}^{n\times n}$, the eigenvalue decomposition and the SVD are one and the same

Matrix Decompositions - Singular Value Decomposition

- It is possible to define the SVD of a $rank - r$ matrix A so that U is an $m \times r$ matrix, Σ a diagonal matrix $r \times r$, and V an $r \times n$ matrix.
  - This construction is very similar to our definition, and ensures that the diagonal matrix Σ has only nonzero entries along the diagonal. The main convenience of this alternative notation is that Σ is diagonal, as in the eigenvalue decomposition.
- A restriction that the SVD for A only applies to $m \times n$ matrices with $m > n$ is practically unnecessary.
  - When $m < n$, the SVD decomposition will yield Σ with more zero columns than rows and, consequently, the singular values $\sigma_{m+1}, \ldots, \sigma_n$ are 0.

Matrix Decompositions - Singular Value Decomposition

- The SVD is used in a variety of applications in machine learning from least-squares problems in curve fitting to solving systems of linear equations
  - These applications harness various important properties of the SVD, its relation to the rank of a matrix, and its ability to approximate matrices of a given rank with lower-rank matrices.
  - Substituting a matrix with its SVD has often the advantage of making calculation more robust to numerical rounding errors.
  - The SVD's ability to approximate matrices with "simpler" matrices in a principled manner opens up machine learning applications ranging from dimensionality reduction and topic modeling to data compression and clustering.