

# Mathematics Refresher

Daniel Wehner, M.Sc.

SAP SE / DHBW Mannheim

Summer term 2025



Find all slides on [GitHub](https://github.com/DaWe1992/Applied_ML_Fundamentals) (DaWe1992/Applied\_ML\_Fundamentals)

# Agenda for this Unit

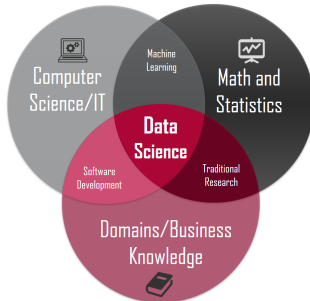
- 1 Introduction
- 2 Linear Algebra
- 3 Probability Theory and Statistics
- 4 Wrap-Up

## Section: Introduction

Introduction  
Math is important!

# Introduction

**Mathematics play a major role in data science and machine learning!**

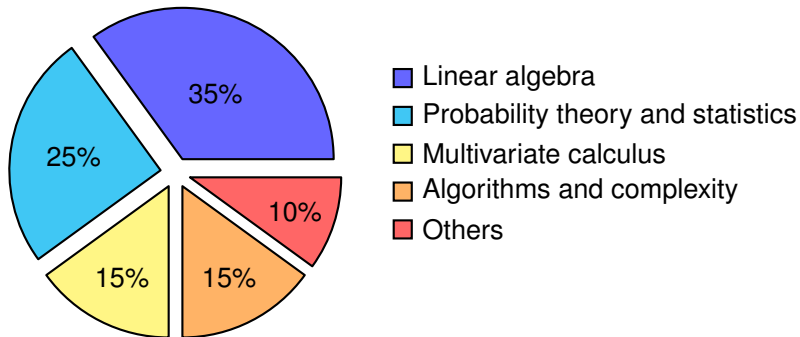


You will need it to understand:

- **Statistical** machine learning
- How **optimization** is used in learning and empirical risk minimization
- How linear algebra, calculus, and statistics are used to make learning and inference more efficient

# Math is important!

Rough importance of mathematical disciplines in data science and machine learning:



## Section:

# Linear Algebra

Vectors and Vector Operations  
Matrix Operations  
Determinants and Inverses  
Eigenvectors and Eigenvalues  
Symmetric Matrices and Definiteness

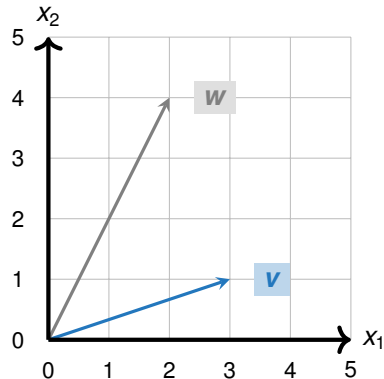
# What is a Vector?

General:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} \in \mathbb{R}^D$$

Example: ( $D = 2$ )

$$\mathbf{v} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \quad \mathbf{w} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$



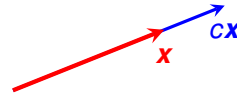
# Multiplication of Vectors by Scalars

General: Let  $c \in \mathbb{R}$  and  $\mathbf{x} \in \mathbb{R}^D$ :

$$c\mathbf{x} = c \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} = \begin{pmatrix} cx_1 \\ \vdots \\ cx_D \end{pmatrix}$$

Example: ( $D = 2$ )

$$2\mathbf{v} = 2 \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 2 \end{pmatrix}$$





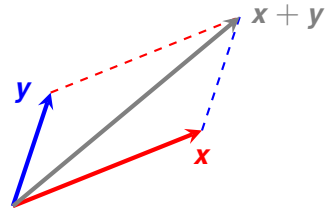
# Addition of Vectors

General: Let  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$

$$\mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_D \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_D + y_D \end{pmatrix}$$

Example: ( $D = 2$ )

$$\mathbf{v} + \mathbf{w} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} + \begin{pmatrix} 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$$



# Linear Combination of Vectors and Span

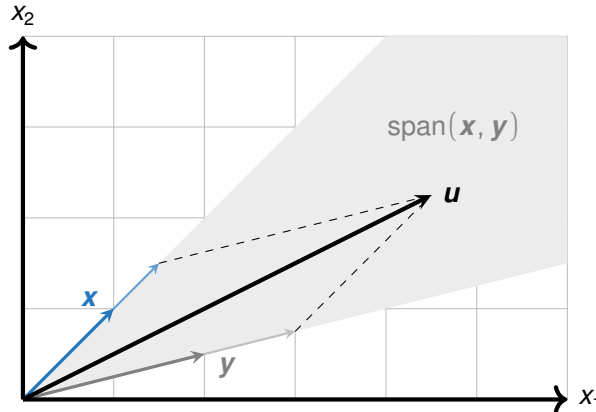
- Let  $c_1, \dots, c_N \in \mathbb{R}$  and  $\mathbf{x}^1, \dots, \mathbf{x}^N \in \mathbb{R}^D$
- A **linear combination** of these vectors is given by  $\mathbf{u} \in \mathbb{R}^D$ :

$$\mathbf{u} := \sum_{n=1}^N c_n \mathbf{x}^n \quad (1)$$

- The **span** (German: *lineare Hülle*) of  $\mathbf{x}^1, \dots, \mathbf{x}^N \in \mathbb{R}^D$  is defined by:

$$\text{span}(\mathbf{x}^1, \dots, \mathbf{x}^N) := \left\{ \mathbf{u} \in \mathbb{R}^D : \exists c_1, \dots, c_N : \mathbf{u} = \sum_{n=1}^N c_n \mathbf{x}^n \right\} \quad (2)$$

# Linear Combination of Vectors and Span (Ctd.)



# Vector Transpose, Inner and Outer Product

- Let  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$  be given
- Transposition:**

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} \quad \mathbf{x}^\top = \begin{pmatrix} x_1 & \dots & x_D \end{pmatrix} \quad (3)$$

- Inner product** (also referred to as **dot product** or **scalar product**):

$$\mathbf{x}^\top \mathbf{y} \equiv \langle \mathbf{x}, \mathbf{y} \rangle := \begin{pmatrix} x_1 & \dots & x_D \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_D \end{pmatrix} = \sum_{d=1}^D x_d y_d \quad (4)$$



# Vector Transpose, Inner and Outer Product (Ctd.)

- Outer product:

$$\mathbf{xy}^T = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} \begin{pmatrix} y_1 & \dots & y_D \end{pmatrix} = \begin{pmatrix} x_1 y_1 & x_1 y_2 & \dots & x_1 y_D \\ x_2 y_1 & x_2 y_2 & \dots & x_2 y_D \\ \vdots & \vdots & \ddots & \vdots \\ x_D y_1 & x_D y_2 & \dots & x_D y_D \end{pmatrix} \quad (5)$$

**Remember:** The inner product yields a scalar value; The result of an outer product is a matrix!

## Example: Vector Transpose, Inner and Outer Product

- Let  $\mathbf{v} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \in \mathbb{R}^2$  and  $\mathbf{w} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} \in \mathbb{R}^2$

- Transposition:**

$$\mathbf{v}^T = \begin{pmatrix} 3 & 1 \end{pmatrix}$$

- Inner product:**

$$\mathbf{v}^T \mathbf{w} = \begin{pmatrix} 3 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \end{pmatrix} = 3 \cdot 2 + 1 \cdot 4 = 10$$

## Example: Vector Transpose, Inner and Outer Product (Ctd.)

- **Outer product:**

$$\mathbf{vw}^T = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \begin{pmatrix} 2 & 4 \end{pmatrix} = \begin{pmatrix} 6 & 12 \\ 2 & 4 \end{pmatrix}$$

# Length of a Vector and Norms

- Length of a vector (**EUCLIDEAN norm**): Let  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$  and  $c \in \mathbb{R}$

$$\|\mathbf{x}\| := \sqrt{\mathbf{x}^\top \mathbf{x}} \quad (6)$$

$$\|c\mathbf{x}\| = |c| \cdot \|\mathbf{x}\| \quad (7)$$

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad (8)$$

- Example: Let  $\mathbf{v} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \in \mathbb{R}^2$

$$\|\mathbf{v}\| = \sqrt{3^2 + 1^2} = \sqrt{10}$$



# Angle between Vectors

- The **angle** between two vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$  is given by:

$$\cos \angle(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} = \frac{\sum_{d=1}^D x_d y_d}{\sqrt{\sum_{d=1}^D x_d^2} \cdot \sqrt{\sum_{d=1}^D y_d^2}} \quad (9)$$

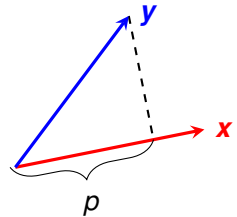
$$\cos \angle(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v}^\top \mathbf{w}}{\|\mathbf{v}\| \cdot \|\mathbf{w}\|} = \frac{10}{\sqrt{10} \cdot \sqrt{20}} \approx 0.71$$

- Inner product:  $\mathbf{x}^\top \mathbf{y} = \|\mathbf{x}\| \cdot \|\mathbf{y}\| \cdot \cos \angle(\mathbf{x}, \mathbf{y})$

## (Orthogonal) Projection of Vectors

- How is the projection of  $\mathbf{y}$  onto  $\mathbf{x}$  defined?
- Formally, we have:

$$\begin{aligned} p &= \|\mathbf{y}\| \cos \angle(\mathbf{x}, \mathbf{y}) \\ &= \|\mathbf{y}\| \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} \\ &= \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|} \end{aligned} \tag{10}$$



- Note that  $p$  is **not a vector**!

# What is a Matrix?

- **General case:**

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{pmatrix} \in \mathbb{R}^{N \times M}$$

- $x_{nm}$  is the entry in row  $n$  and column  $m$

**Remember: Zeilen zuerst, Spalten später (German)**

## Example: Matrices

$$\mathbf{A} = \begin{pmatrix} 3 & 4 & 5 \\ 1 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{2 \times 3}$$

$\mathbf{A}$  has three rows, but only two columns.

$$\mathbf{B} = \begin{pmatrix} 10 & 1 \\ 11 & 2 \end{pmatrix} \in \mathbb{R}^{2 \times 2}$$

$\mathbf{B}$  is a **square matrix** as it has the same number of rows and columns.

Square matrices play a special role in mathematics, e.g. matrix inversion and determinants are only defined for square matrices.

$$\mathbf{C} = \begin{pmatrix} 3 & 0 \\ 0 & 7 \end{pmatrix} \in \mathbb{R}^{2 \times 2}$$

$\mathbf{C}$  is a square matrix, but also a **diagonal matrix** because  $c_{ij} = 0$  for  $i \neq j$ . We often write  $\mathbf{C} := \text{diag}(3, 7)$ .

# Special Matrices

**Identity matrix:**

$$I_N := \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{N \times N}$$

**Zero matrix:**

$$\mathbf{0} := \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{N \times N}$$



# Matrix Transpose

- **Matrix transposition:**

$$\mathbf{X}^T = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{N1} \\ x_{12} & x_{22} & \dots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1M} & x_{2M} & \dots & x_{NM} \end{pmatrix} \in \mathbb{R}^{M \times N} \quad (11)$$

- **Please note:**  $\mathbf{X} \in \mathbb{R}^{N \times M}$ , but  $\mathbf{X}^T \in \mathbb{R}^{M \times N}$



# Matrix Addition

- **Addition of matrices:** Let  $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{N \times M}$

$$\begin{aligned}\mathbf{X} + \mathbf{Y} &= \begin{pmatrix} x_{11} & \dots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{NM} \end{pmatrix} + \begin{pmatrix} y_{11} & \dots & y_{1M} \\ \vdots & \ddots & \vdots \\ y_{N1} & \dots & y_{NM} \end{pmatrix} \\ &= \begin{pmatrix} x_{11} + y_{11} & \dots & x_{1M} + y_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} + y_{N1} & \dots & x_{NM} + y_{NM} \end{pmatrix} \in \mathbb{R}^{N \times M} \end{aligned} \tag{12}$$

- **Please note:**  $\mathbf{X}$  and  $\mathbf{Y}$  must be of the same size!



# Multiplication of Matrices by Scalars

- **Multiplication by scalars:** Let  $\mathbf{X} \in \mathbb{R}^{N \times M}$  and  $c \in \mathbb{R}$

$$c\mathbf{X} = c \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{pmatrix} = \begin{pmatrix} cx_{11} & cx_{12} & \dots & cx_{1M} \\ cx_{21} & cx_{22} & \dots & cx_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ cx_{N1} & cx_{N2} & \dots & cx_{NM} \end{pmatrix} \in \mathbb{R}^{N \times M} \quad (13)$$

- This is defined for all matrices





# Multiplication of Matrices by Vectors

- **Matrix-vector multiplication:** Let  $\mathbf{X} \in \mathbb{R}^{N \times M}$  and  $\mathbf{y} \in \mathbb{R}^M$

$$\mathbf{X}\mathbf{y} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{pmatrix} = \begin{pmatrix} \sum_{m=1}^M x_{1m}y_m \\ \sum_{m=1}^M x_{2m}y_m \\ \vdots \\ \sum_{m=1}^M x_{Nm}y_m \end{pmatrix} \in \mathbb{R}^N \quad (14)$$

- **Please note:** The number of columns of  $\mathbf{X}$  and the number of rows of  $\mathbf{y}$  must be equal in order for the matrix-vector product to exist!
- The order is important:  $\mathbf{X}\mathbf{y}$  is defined, but  $\mathbf{y}\mathbf{X}$  is not, if  $n > 1$



# Matrix Multiplication (Ctd.)

- **Matrix-matrix multiplication:** Let  $\mathbf{X} \in \mathbb{R}^{L \times M}$  and  $\mathbf{Y} \in \mathbb{R}^{M \times N}$

$$\mathbf{XY} = \begin{pmatrix} x_{11} & \dots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{L1} & \dots & x_{LM} \end{pmatrix} \begin{pmatrix} y_{11} & \dots & y_{1N} \\ \vdots & \ddots & \vdots \\ y_{M1} & \dots & y_{MN} \end{pmatrix} = \begin{pmatrix} z_{11} & \dots & z_{1N} \\ \vdots & \ddots & \vdots \\ z_{L1} & \dots & z_{LN} \end{pmatrix} \quad (15)$$

where:

$$z_{\ell n} = \sum_{m=1}^M x_{\ell m} y_{m n} \quad (16)$$

- **Please note:** The number of columns of  $\mathbf{X}$  and the number of rows of  $\mathbf{Y}$  must match!

# Determinants of Square Matrices

- **Determinants are defined for square matrices only!**
- The **determinant** of a  $(2 \times 2)$ -matrix is given by:

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} := ad - bc \quad (17)$$

- The determinant of a  $(3 \times 3)$ -matrix is given by **(rule of SARRUS)**:

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} := aei + bfg + cdh - gec - hfa - idb \quad (18)$$

# LAPLACE Expansion

Use the **LAPLACE expansion** for  $(N \times N)$ -matrices if  $N > 3$ :

**LAPLACE expansion:** Let  $\mathbf{X} \in \mathbb{R}^{N \times N}$  be given. Then

$$\det(\mathbf{X}) = \sum_{n=1}^N x_{nm} \cdot (-1)^{n+m} \cdot \det(\mathbf{X}_{nm}), \quad (19)$$

where  $\mathbf{X}_{nm}$  is the matrix obtained by removing row  $n$  and column  $m$  from  $\mathbf{X}$ .

# Matrix Inversion

- **Matrix inversion is defined for square matrices only!**
- $\mathbf{X} \in \mathbb{R}^{N \times N}$  multiplied by its inverse  $\mathbf{X}^{-1} \in \mathbb{R}^{N \times N}$  gives the identity matrix  $\mathbf{I}_N$ :

$$\mathbf{X}\mathbf{X}^{-1} = \mathbf{I}_N \quad (20)$$

- Also, the order is not important, i. e.:

$$\mathbf{X}^{-1}\mathbf{X} = \mathbf{I}_N \quad (21)$$

- We call  $\mathbf{X}$  **non-singular** or **invertible**, if  $\mathbf{X}^{-1}$  exists

## Matrix Inversion (Ctd.)

Let  $\mathbf{X} \in \mathbb{R}^{N \times N}$  be a square matrix. The following statements are equivalent:

$\mathbf{X}$  is invertible  $\iff \mathbf{X}$  is non-singular

$\iff \det(\mathbf{X}) \neq 0$

$\iff \mathbf{X}$  has rank  $N$  (full rank)

$\iff \mathbf{X}$  does not have eigenvalue 0

$\iff$  The **reduced row echelon form** of  $\mathbf{X}$  is the identity matrix  $\mathbf{I}_N$

## Matrix Inversion (Ctd.)

- The inverse of a matrix can be computed using the **GAUSS-JORDAN algorithm**
- Special case:** Do not use the GAUSS-JORDAN algorithm for  $(2 \times 2)$ -matrices!

You can be more efficient using the identity

$$\mathbf{X} \operatorname{adj}(\mathbf{X}) = \det(\mathbf{X}) \mathbf{I}_N \quad \stackrel{\det(\mathbf{X}) \neq 0}{\iff} \quad \mathbf{X} \overbrace{\frac{1}{\det(\mathbf{X})} \operatorname{adj}(\mathbf{X})}^{=: \mathbf{X}^{-1}} = \mathbf{I}_N, \quad (22)$$

where  $\operatorname{adj}(\mathbf{X})$  is the **adjugate matrix** of  $\mathbf{X}$ .

## Matrix Inversion (Ctd.)

- In general,  $\text{adj}(\mathbf{X})$  is hard to compute, but it is easy for  $(2 \times 2)$ -matrices
- Let

$$\mathbf{X} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbb{R}^{2 \times 2}$$

- The adjugate matrix  $\text{adj}(\mathbf{X})$  of  $\mathbf{X}$  is then given by:

$$\text{adj}(\mathbf{X}) := \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \in \mathbb{R}^{2 \times 2} \quad (23)$$





# Eigenvectors and Eigenvalues

- Let  $\mathbf{X} \in \mathbb{R}^{N \times N}$  be a square matrix
- Some vectors  $\mathbf{v} \in \mathbb{R}^N$  only change their length (but not their direction) when multiplied by  $\mathbf{X}$
- Such vectors are called **eigenvectors** of  $\mathbf{X}$  and the scaling factors are known as **eigenvalues** of  $\mathbf{X}$

**Eigenvectors and eigenvalues satisfy the equation:**

$$\mathbf{X}\mathbf{v} = \lambda\mathbf{v} \quad (24)$$

## Example: Eigenvectors and Eigenvalues

- Let  $\mathbf{X} := \begin{pmatrix} 4 & -1 \\ 2 & 1 \end{pmatrix} \in \mathbb{R}^{2 \times 2}$  be given

- We have (*please verify for yourself*):

$$\begin{pmatrix} 4 & -1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

- Thus,  $\mathbf{v} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$  is an eigenvector of  $\mathbf{X}$  and  $\lambda = 2$  is the corresponding eigenvalue

# Eigenvectors form a Basis (Eigenbasis)

- Let  $\mathbf{v}^1, \mathbf{v}^2, \dots, \mathbf{v}^N$  be  $N$  eigenvectors of  $\mathbf{X} \in \mathbb{R}^{N \times N}$  with corresponding eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_N$
- **Theorem (\*)**:
  - If  $\mathbf{v}^1, \mathbf{v}^2, \dots, \mathbf{v}^N$  correspond to **distinct eigenvalues**  $\lambda_1, \lambda_2, \dots, \lambda_N$ , then the system  $(\mathbf{v}^1, \mathbf{v}^2, \dots, \mathbf{v}^N)$  represents a basis of  $\mathbb{R}^N$  (**eigenbasis**)
  - Hence,  $\mathbf{v}^1, \dots, \mathbf{v}^N$  are **linearly independent** and any vector  $\mathbf{u} \in \mathbb{R}^N$  can be **uniquely** expressed as a **linear combination** of the eigenvectors of  $\mathbf{X}$ , i. e.  
 $\exists c_1, \dots, c_N \in \mathbb{R}$  such that

$$\mathbf{u} = \sum_{n=1}^N c_n \mathbf{v}^n \quad \forall \mathbf{u} \in \mathbb{R}^N$$

# How to compute Eigenvalues and Eigenvectors

- Let  $\mathbf{X} \in \mathbb{R}^{N \times N}$  be a square matrix
- The eigenvalues are the roots (German: *Nullstellen*) of the **characteristic polynomial** defined by

$$\chi_{\mathbf{X}}(\lambda) := \det(\lambda \mathbf{I}_N - \mathbf{X}) \quad (25)$$

- For each eigenvalue  $\lambda_n$  we have to solve the **homogeneous system of linear equations**

$$(\mathbf{X} - \lambda_n \mathbf{I}_N) \mathbf{v} = \mathbf{0} \quad (26)$$

to obtain the respective eigenvectors (see  $\Rightarrow$  [here](#))

## Example: Computation of Eigenvalues and Eigenvectors

### Computation of the eigenvalues:

- Let  $\mathbf{A} := \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \in \mathbb{R}^{2 \times 2}$

- The eigenvalues are the roots of (25) which is given by

$$\chi_{\mathbf{A}} := \det \begin{pmatrix} \lambda - 2 & -1 \\ -1 & \lambda - 2 \end{pmatrix} = (\lambda - 2)^2 - 1 = (\lambda - 1)(\lambda - 3)$$

- We directly see:  $\lambda_1 = 1, \lambda_2 = 3$

## Example: Computation of Eigenvalues and Eigenvectors (Ctd.)

### Computation of the eigenspaces:

- We start with the eigenvalue  $\lambda_1 = 1$  and solve (26):

$$(\mathbf{A} - \mathbf{I}_2) = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \xrightarrow{I+II} \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix}$$

- The eigenspace for eigenvalue 1 is therefore given by  $\mathcal{E}(1) = \{t \cdot (-1, -1)^\top : t \in \mathbb{R}\}$
- Similarly, we can show that  $\mathcal{E}(3) = \{t \cdot (1, -1)^\top : t \in \mathbb{R}\}$

# Diagonalizable Matrices

- Let  $\mathbf{A} \in \mathbb{R}^{N \times N}$  be a square matrix
- If the conditions of theorem (\*) are met, then we can find a non-singular matrix  $\mathbf{S} \in \mathbb{R}^{N \times N}$  such that:

$$\mathbf{S}^{-1} \mathbf{A} \mathbf{S} = \mathbf{B} \in \mathbb{R}^{N \times N} \quad (27)$$

- The columns of  $\mathbf{S}$  are given by the eigenvectors of  $\mathbf{A}$ , and  $\mathbf{B} := \text{diag}(\lambda_1, \dots, \lambda_N)$  is a diagonal matrix containing the eigenvalues of  $\mathbf{A}$
- We say  $\mathbf{A}$  is a **diagonalizable matrix**
- $\mathbf{A}$  and  $\mathbf{B}$  are called **similar matrices**



# Symmetric Matrices

- A square  $(N \times N)$ -matrix  $\mathbf{X}$  is called **symmetric**, if and only if

$$\mathbf{X} = \mathbf{X}^\top \quad (28)$$

- **Some properties:**
  - The inverse  $\mathbf{X}^{-1}$  of  $\mathbf{X}$  is also a symmetric matrix
  - **Eigen-decomposition:** Let  $\mathbf{X}$  be a symmetric matrix. In this case the conditions of theorem (\*) are met and we can find an **orthogonal matrix**  $\mathbf{Q}$  (i. e.  $\mathbf{Q}^{-1} = \mathbf{Q}^\top$ ) such that  $\mathbf{Q}^\top \mathbf{X} \mathbf{Q} = \mathbf{D}$ . The columns of  $\mathbf{Q}$  are given by the normalized eigenvectors of  $\mathbf{X}$ , and  $\mathbf{D}$  is a diagonal matrix whose entries are the corresponding eigenvalues



## Example: Eigen-Decomposition

- Consider  $\mathbf{A} := \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \in \mathbb{R}^{2 \times 2}$
- We choose one eigenvector for each eigenvalue and divide them by their lengths:

$$\text{Eigenvalue 1: } \left( \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)^\top \quad \text{Eigenvalue 3: } \left( \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} \right)^\top$$

Thus, the eigen-decomposition of  $\mathbf{A}$  is given by:

$$\mathbf{D} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}^\top = \mathbf{Q}^\top \mathbf{A} \mathbf{Q}$$

# Positive (semi-)definite Matrices

- A symmetric matrix  $\mathbf{X} \in \mathbb{R}^{N \times N}$  is called **positive definite** (notation:  $\mathbf{X} \succ 0$ ), if

$$\mathbf{z}^T \mathbf{X} \mathbf{z} > 0 \quad \forall \mathbf{z} \in \mathbb{R}^N \setminus \{\mathbf{0}\} \quad (29)$$

- Or **positive semi-definite** (notation:  $\mathbf{X} \succeq 0$ ), if

$$\mathbf{z}^T \mathbf{X} \mathbf{z} \geq 0 \quad \forall \mathbf{z} \in \mathbb{R}^N \quad (30)$$

**Such matrices are important in machine learning.** For instance, the covariance matrices  $\Sigma$  are always positive semi-definite.

## Section:

# Probability Theory and Statistics

Random Variables and Common Distributions  
Basic Rules of Probability  
Expectation and Variance  
Kullback-Leibler Divergence

# Random Variables

- What is a **random variable**?

# Random Variables

- What is a **random variable**?
  - It's a random number determined by chance (according to an underlying distribution).  
To be precise: A random variable  $\mathcal{X}$  is a **measurable function**

$$\mathcal{X} : \Omega \rightarrow \mathbb{R} \quad \text{where } \Omega \text{ is the } \mathbf{sample\ space}$$

- Examples of random variables in machine learning: Input data, output data, noise
- What is a **probability distribution**?

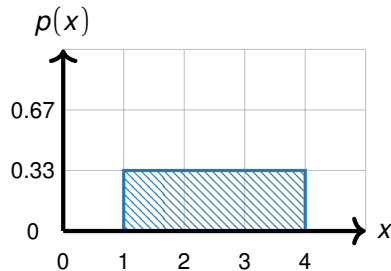
# Random Variables

- What is a **random variable**?
  - It's a random number determined by chance (according to an underlying distribution).  
To be precise: A random variable  $\mathcal{X}$  is a **measurable function**

$$\mathcal{X} : \Omega \rightarrow \mathbb{R} \quad \text{where } \Omega \text{ is the } \mathbf{sample\ space}$$

- Examples of random variables in machine learning: Input data, output data, noise
- What is a **probability distribution**?
  - It describes the probability that a random variable is equal to a certain value
  - It can be given by the physics of an experiment (e. g. throwing dice)
  - **Discrete** vs. **continuous** distributions

# Uniform Distribution



Every outcome is equally probable within a bounded region  $\mathcal{R} := [a, b]$

$$p(\mathcal{X} = x) := \frac{1}{b - a}$$

# Discrete Distributions

A **discrete random variable** takes on discrete values.

**Please note:** Discrete does not mean finite!

## Examples:

- When throwing a die, the possible values are given by the finite set:

$$\mathcal{X} \in \{1, 2, 3, 4, 5, 6\}$$

- The number of sand grains at the beach (countably infinite set):

$$\mathcal{X} \in \mathbb{N}$$



## Discrete Distributions (Ctd.)

- All probabilities sum up to 1, i. e.:

$$\sum_{x \in \mathcal{X}(\Omega)} p(\mathcal{X} = x) = 1$$

- Discrete distributions are particularly important in classification
- A discrete distribution is described by a **probability mass function** (also called **frequency function**)

# BERNOULLI Distribution

- A **BERNOULLI random variable** only takes on two values (e. g. 0 and 1):

$$\mathcal{X} \in \{0, 1\} \quad (31)$$

$$p(\mathcal{X} = 1; \mu) = \mu \quad (32)$$

$$p(\mathcal{X} = 0; \mu) = 1 - \mu \quad (33)$$

$$\mathbb{E}\{\mathcal{X}\} = \mu \quad (34)$$

$$\mathbb{V}\{\mathcal{X}\} = \mu(1 - \mu) \quad (35)$$

- The BERNOLLI distribution is governed only by the parameter  $\mu$ , the **probability of success**



# Binomial Distribution

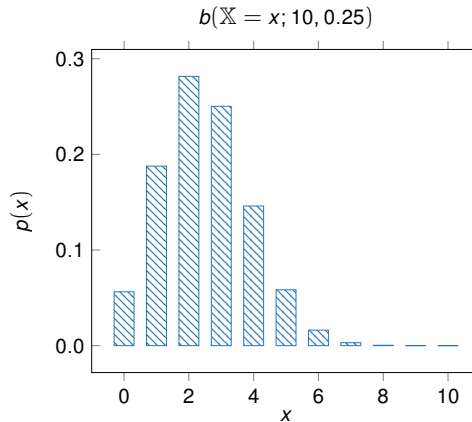
- Repeating a BERNOULLI experiment  $N$  times leads to the **binomial distribution**
- **Example:** What is the probability of getting  $n \in \mathbb{N}$  heads in  $N$  trials?

$$b(\mathcal{X} = n; N, \mu) := \binom{N}{n} \mu^n (1 - \mu)^{N-n} \quad (36)$$

$$\mathbb{E}\{\mathcal{X}\} = N\mu \quad (37)$$

$$\mathbb{V}\{\mathcal{X}\} = N\mu(1 - \mu) \quad (38)$$

## Binomial Distribution (Ctd.)



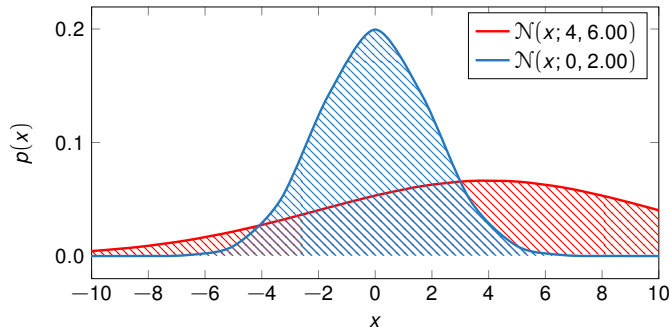
# Continuous Distributions

**Continuous random variables** take on continuous values

- Continuous distributions are discrete distributions where the **number of discrete values goes to infinity**, while the **probability of each value goes to zero**
- A continuous random variable  $\mathcal{X}$  is described by a **probability density function** which integrates to 1, i. e.:

$$\int_{-\infty}^{\infty} p(\mathcal{X} = x) dx = 1$$

# GAUSSIAN Distribution



$$\mathcal{N}(\mathcal{X} = x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad (39)$$



# Central Limit Theorem

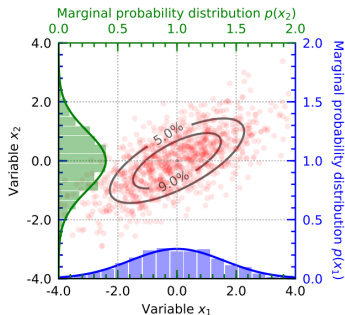
## Central Limit Theorem:

The distribution of the sum of  $N$  i.i.d. (independent and identically distributed) random variables **becomes increasingly GAUSSIAN as  $N$  increases**

- The GAUSSIAN distribution is one of the most important distributions
- GAUSSIAN distributions often are a good model (due to the central limit theorem)
- Working with GAUSSIANS leads to **analytical solutions for complex operations**

# Multivariate GAUSSian Distribution

$$\mathcal{N}_D(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{1}{\sqrt{(2\pi)^D \det(\boldsymbol{\Sigma})}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \quad (40)$$



**Please note:**  $\mathbf{x}$  and  $\boldsymbol{\mu}$  are vectors, while  $\boldsymbol{\Sigma}$  is a matrix.

The probability given by  $\mathcal{N}_D(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$  is still a scalar value!



# Basic Rules of Probability

- **Joint distribution:**

$$p(\mathcal{X} \cap \mathcal{Y}) \quad (41)$$

- **Marginal distribution:**

$$p(\mathcal{Y}) = \int_{\mathcal{X}} p(\mathcal{X} \cap \mathcal{Y}) \, d\mathcal{X} \quad (42)$$

- **Conditional distribution:**

$$p(\mathcal{Y}|\mathcal{X}) = \frac{p(\mathcal{X} \cap \mathcal{Y})}{p(\mathcal{X})} \quad (43)$$

# Basic Rules of Probability (Ctd.)

- **Probabilistic independence:**

$$p(\mathcal{X} \cap \mathcal{Y}) = p(\mathcal{X})p(\mathcal{Y}) \quad (44)$$

- **Chain rule of probabilities:**

$$\begin{aligned} p(\mathcal{X}_1 \cap \dots \cap \mathcal{X}_N) &= p(\mathcal{X}_1 | \mathcal{X}_2 \cap \dots \cap \mathcal{X}_N) p(\mathcal{X}_2 \cap \dots \cap \mathcal{X}_N) \\ &= p(\mathcal{X}_1 | \mathcal{X}_2 \cap \dots \cap \mathcal{X}_N) \dots p(\mathcal{X}_{N-1} | \mathcal{X}_N) p(\mathcal{X}_N) \end{aligned} \quad (45)$$

- **BAYES' rule:**

$$p(\mathcal{Y} | \mathcal{X}) = \frac{p(\mathcal{X} | \mathcal{Y}) p(\mathcal{Y})}{p(\mathcal{X})} \quad (46)$$



# Expectation

- The **expectation**  $\mathbb{E}$  of a random variable  $\mathcal{X}$  is defined by

$$\mathbb{E}\{\mathcal{X}\} := \sum_{k \in \Omega(\mathcal{X})} k \cdot p(\mathcal{X} = k) \quad (47)$$

- Expectations of functions:

$$\mathbb{E}_x\{f\} := \sum_x p(x) f(x) \quad (48)$$

- **Remark:** In the continuous case we have to replace  $\sum$  by  $\int$

## Expectation (Ctd.)

### Rules of expectations:

- The expectation is a **linear** operation:

$$\mathbb{E}\{a\mathcal{X} + b\mathcal{Y}\} = a\mathbb{E}\{\mathcal{X}\} + b\mathbb{E}\{\mathcal{Y}\} \quad (49)$$

- More general:  $\mathbb{E}\left\{\sum_{n=1}^N a_n \mathcal{X}_n\right\} = \sum_{n=1}^N a_n \mathbb{E}\{\mathcal{X}_n\}$
- If  $\mathcal{X}$  and  $\mathcal{Y}$  are independent:  $\mathbb{E}\{\mathcal{X}\mathcal{Y}\} = \mathbb{E}\{\mathcal{X}\}\mathbb{E}\{\mathcal{Y}\}$
- The expectation is **monotonous**:

$$\mathcal{X} \leq \mathcal{Y} \implies \mathbb{E}\{\mathcal{X}\} \leq \mathbb{E}\{\mathcal{Y}\} \quad (50)$$

# Variance

- The **variance**  $\mathbb{V}$  of a random variable  $\mathcal{X}$  is defined by

$$\mathbb{V}\{\mathcal{X}\} := \mathbb{E}\{\mathcal{X} - \mathbb{E}^2\{\mathcal{X}\}\} = \mathbb{E}\{\mathcal{X}^2\} - \mathbb{E}^2\{\mathcal{X}\} \quad (51)$$

- $\mathbb{V}$  is **not** linear:

$$\mathbb{V}\{a + b\mathcal{X}\} = b^2\mathbb{V}\{\mathcal{X}\} \quad (52)$$

$$\mathbb{V}\{\mathcal{X} + \mathcal{Y}\} = \mathbb{V}\{\mathcal{X}\} + \mathbb{V}\{\mathcal{Y}\} + \text{cov}\{\mathcal{X}, \mathcal{Y}\} \quad (53)$$

- BIENAYMÉ's identity:** If  $\mathcal{X}$  and  $\mathcal{Y}$  are **uncorrelated**, we get:

$$\mathbb{V}\{\mathcal{X} + \mathcal{Y}\} = \mathbb{V}\{\mathcal{X}\} + \mathbb{V}\{\mathcal{Y}\} \quad (54)$$

# Covariance

- **Covariances** give a measure of correlation, i. e. how much variables change together

$$\begin{aligned}\text{cov}\{X, Y\} &:= \mathbb{E}\left\{ (X - \mathbb{E}\{X\})(Y - \mathbb{E}\{Y\}) \right\} \\ &= \mathbb{E}\{XY\} - \mathbb{E}\{X\}\mathbb{E}\{Y\}\end{aligned}\tag{55}$$

- The variance  $\mathbb{V}$  of a random variable  $X$  is a special case:

$$\mathbb{V}\{X\} = \text{cov}\{X, X\}\tag{56}$$



# Kullback-Leibler Divergence

- The **Kullback-Leibler (KL) divergence** is a similarity measure between two distributions  $p$  and  $q$ :

$$\mathbb{KL}(p\|q) := \sum_x p(x) \cdot \log \frac{p(x)}{q(x)} \quad (57)$$

- Some properties:
  - It is not symmetric:  $\mathbb{KL}(p\|q) \neq \mathbb{KL}(q\|p)$
  - It is non-negative:  $\mathbb{KL}(p\|q) \geq 0$
  - If  $\forall x : p(x) = q(x) \implies \mathbb{KL}(p\|q) = 0$

## Section: Wrap-Up

Summary  
Recommended Literature  
Self-Test Questions  
Lecture Outlook



# Summary

- **Mathematics play a major role in machine learning!**
- **Linear algebra:**
  - You should know what vectors are and what you can do with them (addition, multiplication, transpose, ...)
  - The same applies to matrices
  - You should know the concept of **determinants** and how to **invert matrices**
  - **Eigenvectors** and **eigenvalues** are important tools in machine learning
  - The **eigen-decomposition** plays an import role in many machine learning applications

## Summary (Ctd.)

- **Probability theory and statistics:**
  - Random variables are numbers **determined by chance**
  - Probability distributions describe a **probability mass** or **probability density**
  - **Discrete distributions:** BERNOULLI, Binomial, Multinomial
  - **Continuous distribution:** GAUSSIAN distribution
  - GAUSSIANS are important in machine learning and have appealing properties
  - Terms you should know: Joint-, marginal- and conditional distribution, chain rule, probabilistic independence, Bayes' rule
  - You should know what **expectation** and **variance** of distributions are

# Recommended Literature

## 1 Linear algebra:

- [DEISENROTH.2019], chapter 2
- [DEISENROTH.2019], chapter 3
- [DEISENROTH.2019], chapter 4

## 2 Probability theory and statistics:

- [DEISENROTH.2019], chapter 6

(For free PDF versions, see list in GitHub readme!)



# Self-Test Questions

- 1 What is a vector and what is a matrix?
- 2 What is the result of an inner product / outer product?
- 3 How can you invert matrices? Is this always possible?
- 4 What is an eigenvalue problem? How can you compute eigenvectors and eigenvalues?
- 5 What are random variables and probability distributions?
- 6 Why is the GAUSSIAN distribution so important?
- 7 What is BAYES' rule? Explain its components!

# What's next...?

- **I** Machine Learning Introduction
- II** Optimization Techniques
- III** Bayesian Decision Theory
- IV** Non-parametric Density Estimation
- V** Probabilistic Graphical Models
- VI** Linear Regression
- VII** Logistic Regression
- VIII** Deep Learning
- IX** Evaluation
- X** Decision Trees
- XI** Support Vector Machines
- XII** Clustering
- XIII** Principal Component Analysis
- XIV** Reinforcement Learning
- XV** Advanced Regression

# Thank you very much for the attention!

**\*\*\* Artificial Intelligence and Machine Learning \*\*\***

**Topic:** Mathematics Refresher

**Term:** Summer term 2025

**Contact:**

Daniel Wehner, M.Sc.

SAP SE / DHBW Mannheim

[daniel.wehner@sap.com](mailto:daniel.wehner@sap.com)

## Do you have any questions?