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

Mathematics Refresher

Daniel Wehner, M.Sc.

SAP SE / DHBW Mannheim

Summer term 2025

Find all slides on GitHub (DaWe1992/Applied_ML_Fundamentals)

Agenda for this Unit

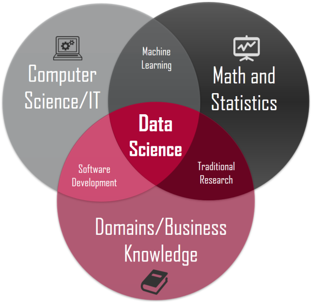
Section:
Introduction

Introduction

Math is important!

Introduction

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

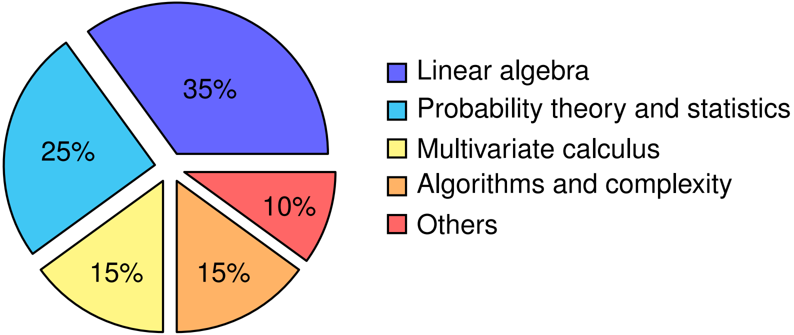


You will need it to understand:

- **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

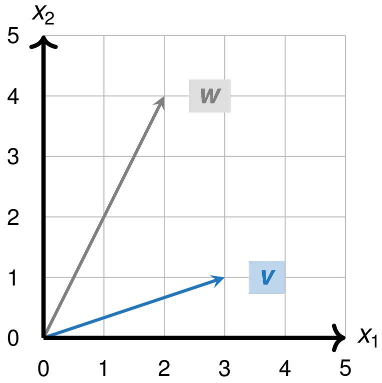
What is a Vector?

General:

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

Example: ($D = 2$)

$$\boldsymbol{v} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \qquad \boldsymbol{w} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$



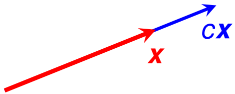
Multiplication of Vectors by Scalars

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

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

Example: $(D = 2)$

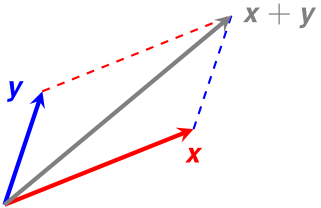
$$2\boldsymbol{v} = 2 \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 2 \end{pmatrix}$$



Addition of Vectors

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

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



Example: $(D = 2)$

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

Linear Combination of Vectors and Span

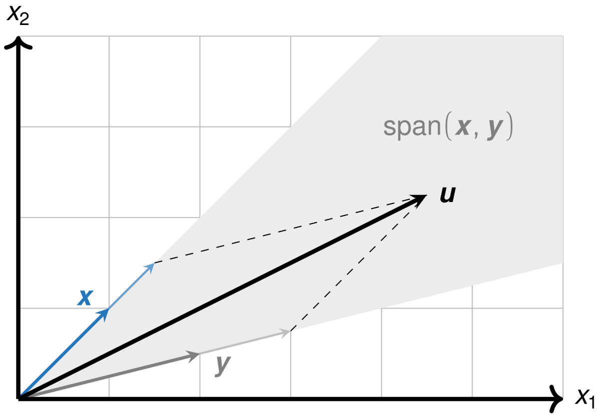
- Let $c_1, \dots, c_N \in \mathbb{R}$ and $\boldsymbol{x}^1, \dots, \boldsymbol{x}^N \in \mathbb{R}^D$
- A **linear combination** of these vectors is given by $\boldsymbol{u} \in \mathbb{R}^D$:

$$\boldsymbol{u} := \sum_{n=1}^{N} c_n \boldsymbol{x}^n \tag{1}$$

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

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

Linear Combination of Vectors and Span (Ctd.)



Vector Transpose, Inner and Outer Product

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

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} \qquad \boldsymbol{x}^\top = \begin{pmatrix} x_1 & \dots & x_D \end{pmatrix} \tag{3}$$

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

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

Vector Transpose, Inner and Outer Product (Ctd.)

- **Outer product:**

$$\boldsymbol{x}\boldsymbol{y}^\top = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} \begin{pmatrix} y_1 & \dots & y_D \end{pmatrix} = \begin{pmatrix} x_1 y_1 & x_1 y_2 & \dots & x_1 y_D \\ x_2 y_1 & x_2 y_2 & \dots & x_2 y_D \\ \vdots & \vdots & \ddots & \vdots \\ x_D y_1 & x_D y_2 & \dots & x_D y_D \end{pmatrix} \tag{5}$$

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 $\boldsymbol{v} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \in \mathbb{R}^2$ and $\boldsymbol{w} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} \in \mathbb{R}^2$
- **Transposition:**

$$\boldsymbol{v}^\top = \begin{pmatrix} 3 & 1 \end{pmatrix}$$

- **Inner product:**

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

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

- **Outer product:**

$$\boldsymbol{v}\boldsymbol{w}^\top = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \begin{pmatrix} 2 & 4 \end{pmatrix} = \begin{pmatrix} 6 & 12 \\ 2 & 4 \end{pmatrix}$$

Important

Length of a Vector and Norms

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

$$\|\boldsymbol{x}\| := \sqrt{\boldsymbol{x}^\top \boldsymbol{x}} \tag{6}$$

$$\|c\boldsymbol{x}\| = |c| \cdot \|\boldsymbol{x}\| \tag{7}$$

$$\|\boldsymbol{x} + \boldsymbol{y}\| \leqslant \|\boldsymbol{x}\| + \|\boldsymbol{y}\| \tag{8}$$

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

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

Angle between Vectors

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

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

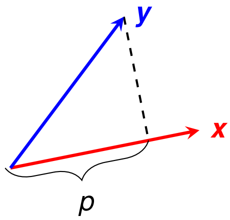
$$\cos \measuredangle(\boldsymbol{v}, \boldsymbol{w}) = \frac{\boldsymbol{v}^\top \boldsymbol{w}}{\|\boldsymbol{v}\| \cdot \|\boldsymbol{w}\|} = \frac{10}{\sqrt{10} \cdot \sqrt{20}} \approx 0.71$$

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

(Orthogonal) Projection of Vectors

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

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



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

What is a Matrix?

- **General case:**

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

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

Remember: Zeilen **z**uerst, **S**palten **s**päter (German)

Example: Matrices

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

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

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

$\boldsymbol{B}$ 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.

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

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

Special Matrices

Identity matrix:

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

Zero matrix:

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

Matrix Transpose

- **Matrix transposition:**

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

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

Important

Matrix Addition

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

$$\boldsymbol{X} + \boldsymbol{Y} = \begin{pmatrix} x_{11} & \dots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{NM} \end{pmatrix} + \begin{pmatrix} y_{11} & \dots & y_{1M} \\ \vdots & \ddots & \vdots \\ y_{N1} & \dots & y_{NM} \end{pmatrix}$$

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

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

Important

Multiplication of Matrices by Scalars

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

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

- This is defined for all matrices

Multiplication of Matrices by Vectors

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

$$\boldsymbol{X}\boldsymbol{y} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{pmatrix} \begin{pmatrix} y_1 \\ v_2 \\ \vdots \\ y_M \end{pmatrix} = \begin{pmatrix} \sum_{m=1}^{M} x_{1m} y_m \\ \sum_{m=1}^{M} x_{2m} y_m \\ \vdots \\ \sum_{m=1}^{M} x_{Nm} y_m \end{pmatrix} \in \mathbb{R}^{N} \tag{14}$$

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

Matrix Multiplication (Ctd.)

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

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

  where:

$$z_{\ell n} = \sum_{m=1}^{M} x_{\ell m} y_{nk} \tag{16}$$

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

Determinants of Square Matrices

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

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

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

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

LAPLACE Expansion

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

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

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

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

Important

Matrix Inversion

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

$$\boldsymbol{X}\boldsymbol{X}^{-1} = \boldsymbol{I}_N \tag{20}$$

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

$$\boldsymbol{X}^{-1}\boldsymbol{X} = \boldsymbol{I}_N \tag{21}$$

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

Matrix Inversion (Ctd.)

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

$$
\begin{aligned}
\boldsymbol{X} \text{ is invertible} &\iff \boldsymbol{X} \text{ is non-singular} \\
&\iff \det(\boldsymbol{X}) \neq 0 \\
&\iff \boldsymbol{X} \text{ has rank } N \text{ (full rank)} \\
&\iff \boldsymbol{X} \text{ does not have eigenvalue } 0 \\
&\iff \text{The } \textbf{reduced row echelon form} \text{ of } \boldsymbol{X} \text{ is the identity matrix } \boldsymbol{I}_N
\end{aligned}
$$

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 (2×2)-matrices!

You can be more efficient using the identity

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

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

Matrix Inversion (Ctd.)

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

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

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

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

Eigenvectors and Eigenvalues

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

Eigenvectors and eigenvalues satisfy the equation:

$$\boldsymbol{X}\boldsymbol{v} = \lambda \boldsymbol{v} \tag{24}$$

Example: Eigenvectors and Eigenvalues

- Let $\boldsymbol{X} := \begin{pmatrix} 4 & -1 \\ 2 & 1 \end{pmatrix} \in \mathbb{R}^{2 \times 2}$ 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, $\boldsymbol{v} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ is an eigenvector of $\boldsymbol{X}$ and $\lambda = 2$ is the corresponding eigenvalue

Eigenvectors form a Basis (Eigenbasis)

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

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

How to compute Eigenvalues and Eigenvectors

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

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

- For each eigenvalue λ_n we have to solve the **homogeneous system of linear equations**

$$(\boldsymbol{X} - \lambda_n \boldsymbol{I}_N)\boldsymbol{v} = \boldsymbol{0} \tag{26}$$

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

Example: Computation of Eigenvalues and Eigenvectors

Computation of the eigenvalues:

- Let $\boldsymbol{A} := \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \in \mathbb{R}^{2\times 2}$
- The eigenvalues are the roots of (25) which is given by

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

- 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):

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

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

Diagonalizable Matrices

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

$$\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S} = \boldsymbol{B} \in \mathbb{R}^{N \times N} \tag{27}$$

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

Important

Symmetric Matrices

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

$$\boldsymbol{X} = \boldsymbol{X}^{\top} \tag{28}$$

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

Example: Eigen-Decomposition

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

$$\text{Eigenvalue 1: } \left({}^1\!/\!\sqrt{2}, {}^1\!/\!\sqrt{2}\right)^\top \qquad \text{Eigenvalue 3: } \left({}^1\!/\!\sqrt{2}, -{}^1\!/\!\sqrt{2}\right)^\top$$

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

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

Positive (semi-)definite Matrices

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

$$\boldsymbol{z}^\top \boldsymbol{X}\boldsymbol{z} > 0 \quad \forall \boldsymbol{z} \in \mathbb{R}^N \backslash \{\boldsymbol{0}\} \tag{29}$$

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

$$\boldsymbol{z}^\top \boldsymbol{X}\boldsymbol{z} \geqslant 0 \quad \forall \boldsymbol{z} \in \mathbb{R}^N \tag{30}$$

Such matrices are important in machine learning. For instance, the covariance matrices $\boldsymbol{\Sigma}$ 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**

$$\mathcal{X} : \Omega \to \mathbb{R} \qquad \text{where } \Omega \text{ is the } \textbf{sample space}$$

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

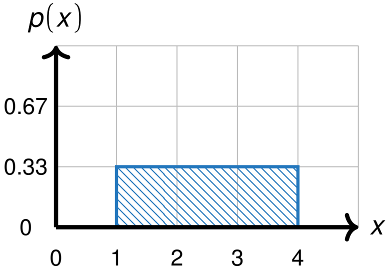
Random Variables

- 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**

$$\mathcal{X} : \Omega \to \mathbb{R} \qquad \text{where } \Omega \text{ is the } \textbf{sample space}$$

  - 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:

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

- 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):

$$\mathcal{X} \in \{0, 1\} \tag{31}$$

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

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

$$\mathbb{E}\{\mathcal{X}\} = \mu \tag{34}$$

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

- The BERNOULLI distribution is governed only by the parameter μ, the **probability of success**

Binomial Distribution

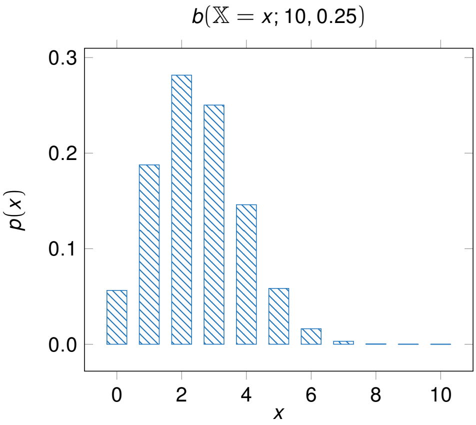
- Repeating a BERNOULLI experiment N times leads to the **binomial distribution**
- **Example:** What is the probability of getting $n \in \mathbb{N}$ heads in N trials?

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

$$\mathbb{E}\{\mathcal{X}\} = N\mu \tag{37}$$

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

Binomial Distribution (Ctd.)



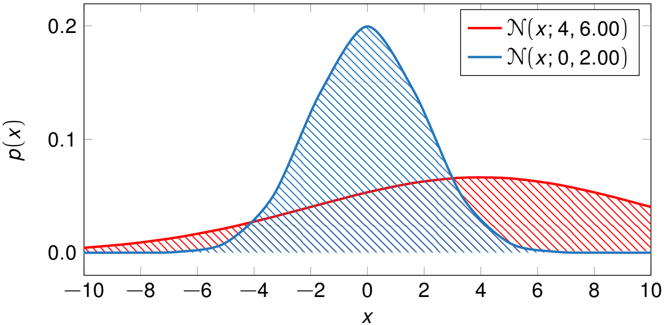
Continuous Distributions

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.:

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

GAUSSian Distribution



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

Central Limit Theorem

Central Limit Theorem:

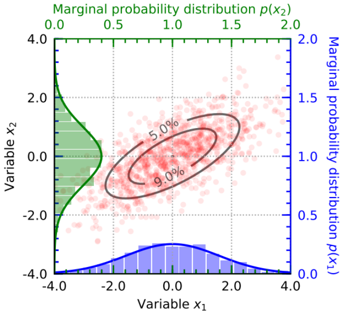
The distribution of the sum of N i. i. d. (independent and identically distributed) random variables **becomes increasingly GAUSSian as N increases**

- 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**

Important

Multivariate Gaussian Distribution

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



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

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

Basic Rules of Probability

- **Joint distribution:**

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

- **Marginal distribution:**

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

- **Conditional distribution:**

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

Basic Rules of Probability (Ctd.)

- **Probabilistic independence:**

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

- **Chain rule of probabilities:**

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

- **BAYES' rule:**

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

Expectation

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

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

- Expectations of functions:

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

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

Expectation (Ctd.)

Rules of expectations:

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

$$\mathbb{E}\{a\mathcal{X} + b\mathcal{Y}\} = a\mathbb{E}\{\mathcal{X}\} + b\mathbb{E}\{\mathcal{Y}\} \tag{49}$$

- More general: $\mathbb{E}\{\sum_{n=1}^{N} a_n\mathcal{X}_n\} = \sum_{n=1}^{N} a_n\mathbb{E}\{\mathcal{X}_n\}$
- 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**:

$$\mathcal{X} \leqslant \mathcal{Y} \Longrightarrow \mathbb{E}\{\mathcal{X}\} \leqslant \mathbb{E}\{\mathcal{Y}\} \tag{50}$$

Variance

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

$$\mathbb{V}\{\mathcal{X}\} := \mathbb{E}\{\mathcal{X} - \mathbb{E}^2\{\mathcal{X}\}\} = \mathbb{E}\{\mathcal{X}^2\} - \mathbb{E}^2\{\mathcal{X}\} \tag{51}$$

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

$$\mathbb{V}\{a + b\mathcal{X}\} = b^2\mathbb{V}\{\mathcal{X}\} \tag{52}$$

$$\mathbb{V}\{\mathcal{X} + \mathcal{Y}\} = \mathbb{V}\{\mathcal{X}\} + \mathbb{V}\{\mathcal{Y}\} + \mathrm{cov}\{\mathcal{X}, \mathcal{Y}\} \tag{53}$$

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

$$\mathbb{V}\{\mathcal{X} + \mathcal{Y}\} = \mathbb{V}\{\mathcal{X}\} + \mathbb{V}\{\mathcal{Y}\} \tag{54}$$

Covariance

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

$$\begin{aligned}\operatorname{cov}\{\mathcal{X}, \mathcal{Y}\} &:= \mathbb{E}\Big\{(\mathcal{X} - \mathbb{E}\{\mathcal{X}\})(\mathcal{Y} - \mathbb{E}\{\mathcal{Y}\})\Big\} \\ &= \mathbb{E}\{\mathcal{X}\mathcal{Y}\} - \mathbb{E}\{\mathcal{X}\}\mathbb{E}\{\mathcal{Y}\} \end{aligned} \tag{55}$$

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

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

Kullback-Leibler Divergence

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

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

- Some properties:
  - It is not symmetric: $\mathbb{KL}(p||q) \neq \mathbb{KL}(q||p)$
  - It is non-negative: $\mathbb{KL}(p||q) \geqslant 0$
  - 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!)

Important

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

Do you have any questions?