

# \*\*\* Applied Machine Learning Fundamentals \*\*\*

## Mathematical Foundations

Daniel Wehner, M.Sc.

SAP SE / DHBW Mannheim

Winter term 2023/2024



Find all slides on [GitHub](#) (DaWe1992/Applied\_ML\_Fundamentals)

# Lecture Overview

Unit I	Machine Learning Introduction
<b>Unit II</b>	<b>Mathematical Foundations</b>
Unit III	Bayesian Decision Theory
Unit IV	Regression
Unit V	Classification I
Unit VI	Evaluation
Unit VII	Classification II
Unit VIII	Clustering
Unit IX	Dimensionality Reduction

# Agenda for this Unit

- 1 Introduction
- 2 Linear Algebra

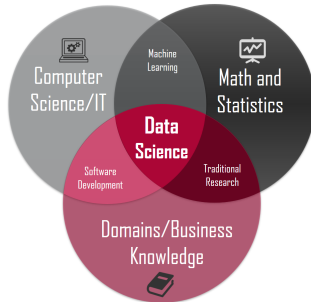
- 3 Probability Theory and Statistics
- 4 Optimization Techniques
- 5 Wrap-Up

## Section: Introduction

Introduction  
Math is important!

# Introduction

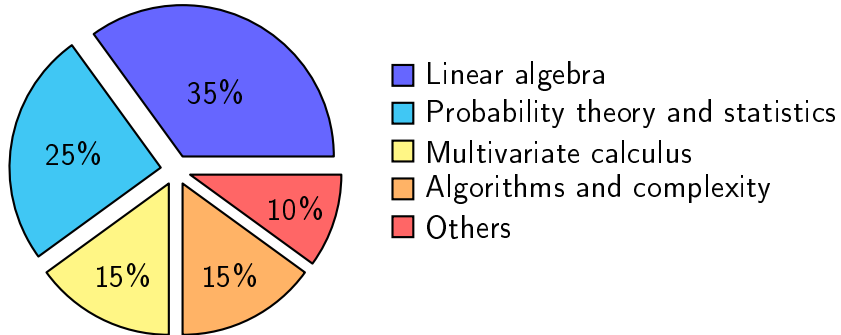
Maths play a major role in data science and machine learning!



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!



## Section: Linear Algebra

Vector Operations  
Matrix Operations  
Determinants and Inverses  
Eigenvectors and Eigenvalues  
Miscellaneous

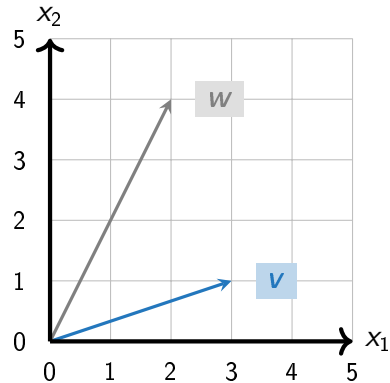
# What is a Vector?

General:

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

Example: ( $m = 2$ )

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





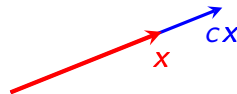
# Multiplication by a Scalar

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

$$c\mathbf{x} = c \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} cx_1 \\ \vdots \\ cx_m \end{pmatrix}$$

Example: ( $m = 2$ )

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



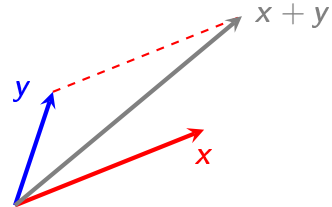
# Addition of Vectors

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

$$\mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_m + y_m \end{pmatrix}$$

Example: ( $m = 2$ )

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

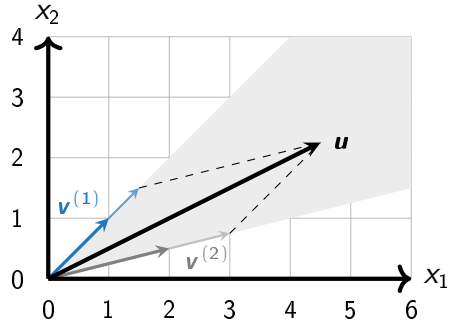


# Linear Combination of Vectors

Let  $c_1, \dots, c_n \in \mathbb{R}$  be scalars and  $\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(n)} \in \mathbb{R}^m$  be vectors.

A **linear combination** of these vectors is given by  $\mathbf{u} \in \mathbb{R}^m$ :

$$\mathbf{u} := c_1 \mathbf{v}^{(1)} + c_2 \mathbf{v}^{(2)} + \dots + c_n \mathbf{v}^{(n)} \quad (1)$$





# Vector Transpose, inner and outer Product

- **Vector transpose:**

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

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

$$\mathbf{x}^T \mathbf{y} \equiv \langle \mathbf{x}, \mathbf{y} \rangle \equiv \mathbf{x} \cdot \mathbf{y} := \sum_{j=1}^m x_j y_j \quad (2)$$

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



# Vector Transpose and inner and outer Product (Ctd.)

- Outer product: Let  $x, y \in \mathbb{R}^m$

$$xy^T := \begin{pmatrix} x_1y_1 & x_1y_2 & \dots & x_1y_m \\ x_2y_1 & x_2y_2 & \dots & x_2y_m \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \dots & x_my_m \end{pmatrix} \quad (3)$$
$$vw^T = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \begin{pmatrix} 2 & 4 \end{pmatrix} = \begin{pmatrix} 6 & 12 \\ 2 & 4 \end{pmatrix}$$

**The inner product yields a scalar value, the results of an outer product is a matrix!**

# Length of a Vector

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

$$\|\mathbf{x}\| := \sqrt{\mathbf{x}^\top \mathbf{x}} \quad (4)$$

$$\|c\mathbf{x}\| = |c| \cdot \|\mathbf{x}\| \quad (5)$$

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad (6)$$

- Example:

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

# Angle between Vectors

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

$$\cos \angle(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} = \frac{\sum_{j=1}^m x_j y_j}{\sqrt{\sum_{j=1}^m x_j^2} \cdot \sqrt{\sum_{j=1}^m y_j^2}} \quad (7)$$

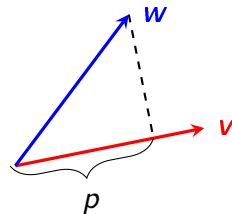
$$\cos \angle(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v}^\top \mathbf{w}}{\|\mathbf{v}\| \cdot \|\mathbf{w}\|} = \frac{10}{\sqrt{10} \cdot \sqrt{20}} \approx 0.71$$

- 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  $\mathbf{w}$  onto  $\mathbf{v}$  defined?
- Formally, we have:

$$\begin{aligned}
 p &= \|\mathbf{w}\| \cos \angle(\mathbf{v}, \mathbf{w}) \\
 &= \|\mathbf{w}\| \frac{\mathbf{v}^T \mathbf{w}}{\|\mathbf{v}\| \cdot \|\mathbf{w}\|} \\
 &= \frac{\mathbf{v}^T \mathbf{w}}{\|\mathbf{v}\|}
 \end{aligned}
 \tag{8}$$



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



# What is a Matrix?

General case:

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix} \in \mathbb{R}^{n \times m}$$

$X_{ij}$  is the entry in row  $i$  and column  $j$   
 ('**Z**eilen **z**uerst, **S**palten **s**päter')

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

$$\mathbf{N} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{3 \times 3}$$

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



# Matrix Transpose

Transpose of a matrix:

$$\mathbf{M}^T = \begin{pmatrix} 3 & 4 & 5 \\ 1 & 0 & 1 \end{pmatrix}^T = \begin{pmatrix} 3 & 1 \\ 4 & 0 \\ 5 & 1 \end{pmatrix} \quad (9)$$

Please note:  $\mathbf{M} \in \mathbb{R}^{2 \times 3}$ , but  $\mathbf{M}^T \in \mathbb{R}^{3 \times 2}$

# Matrix Addition

Addition of matrices: Let  $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times m}$

$$\begin{aligned}\mathbf{X} + \mathbf{Y} &= \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix} + \begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1m} \\ Y_{21} & Y_{22} & \dots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \dots & Y_{nm} \end{pmatrix} \\ &= \begin{pmatrix} X_{11} + Y_{11} & X_{12} + Y_{12} & \dots & X_{1m} + Y_{1m} \\ X_{21} + Y_{21} & X_{22} + Y_{22} & \dots & X_{2m} + Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} + Y_{n1} & X_{n2} + Y_{n2} & \dots & X_{nm} + Y_{nm} \end{pmatrix} \in \mathbb{R}^{n \times m} \quad (10)\end{aligned}$$



# Multiplication of Matrices by Scalars

Multiplication by scalars: Let  $\mathbf{X} \in \mathbb{R}^{n \times m}$  and  $c \in \mathbb{R}$

$$c\mathbf{X} = c \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix} = \begin{pmatrix} cX_{11} & cX_{12} & \dots & cX_{1m} \\ cX_{21} & cX_{22} & \dots & cX_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ cX_{n1} & cX_{n2} & \dots & cX_{nm} \end{pmatrix} \in \mathbb{R}^{n \times m} \quad (11)$$



# Multiplication of Matrices by Vectors

Matrix-vector multiplication: Let  $\mathbf{X} \in \mathbb{R}^{n \times m}$  and  $\mathbf{y} \in \mathbb{R}^m$

$$\mathbf{X}\mathbf{y} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^m X_{1j}y_j \\ \sum_{j=1}^m X_{2j}y_j \\ \vdots \\ \sum_{j=1}^m X_{nj}y_j \end{pmatrix} \in \mathbb{R}^n \quad (12)$$



# Matrix Multiplication (Ctd.)

Matrix-matrix multiplication: Let  $\mathbf{X} \in \mathbb{R}^{\ell \times m}$  and  $\mathbf{Y} \in \mathbb{R}^{m \times n}$

$$\mathbf{XY} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{\ell 1} & X_{\ell 2} & \dots & X_{\ell m} \end{pmatrix} \begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1n} \\ Y_{21} & Y_{22} & \dots & Y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{m1} & Y_{m2} & \dots & Y_{mn} \end{pmatrix} = \begin{pmatrix} Z_{11} & Z_{12} & \dots & Z_{1n} \\ Z_{21} & Z_{22} & \dots & Z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{\ell 1} & Z_{\ell 2} & \dots & Z_{\ell n} \end{pmatrix} \quad (13)$$

where:

$$Z_{ik} = \sum_{j=1}^m X_{ij} Y_{jk} \quad (14)$$

# Determinants

- The **determinant** of a  $2 \times 2$  matrix is given by:

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} := ad - bc \quad (15)$$

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

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} := aei + bfg + cdh - gec - hfa - idb \quad (16)$$

- Use the **Laplace expansion** for larger matrices



# Matrix Inversion

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

$$\mathbf{X}^{-1}\mathbf{X} = \mathbf{X}\mathbf{X}^{-1} = \mathbf{I}_n \quad (17)$$

$$\mathbf{I}_n := \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

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



## Matrix Inversion (Ctd.)

Let  $\mathbf{X} \in \mathbb{R}^{n \times n}$ . 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.)

- $\mathbf{X} \in \mathbb{R}^{n \times n}$  is invertible if and only if  $\det(\mathbf{X}) \neq 0$
- The inverse of a matrix can be computed using the **Gauß-Jordan algorithm**
- **Special case:** Do not use Gauß-Jordan for  $2 \times 2$  matrices! You can be more efficient:

$$\mathbf{X} := \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad \mathbf{X}^{-1} := \frac{1}{\det(\mathbf{X})} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \qquad (18)$$

- $\det(\mathbf{X}) := ad - bc$  is the determinant of  $\mathbf{X}$

## Matrix Inversion Example

$$\mathbf{X} = \begin{pmatrix} 1 & 1/2 \\ -1 & 1 \end{pmatrix} \qquad \mathbf{X}^{-1} = \begin{pmatrix} 2/3 & -1/3 \\ 2/3 & 2/3 \end{pmatrix}$$

Please verify for yourself!

$$\mathbf{X}\mathbf{X}^{-1} = \mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{X}^{-1}\mathbf{X}$$

Check the result using Gauß-Jordan as well as equation (18)!



# Eigenvectors and Eigenvalues

- Let  $\mathbf{X} \in \mathbb{R}^{n \times n}$ . Some vectors  $\mathbf{v} \in \mathbb{R}^n$  only change their length (but not their direction) when multiplied by  $\mathbf{X}$

- **Example:**

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

- Such vectors are called **eigenvectors** of  $\mathbf{X}$ , the scaling factors are known as **eigenvalues** of  $\mathbf{X}$
- More general:

$$\mathbf{X}\mathbf{v} = \lambda\mathbf{v} \tag{19}$$

# Eigenvectors form a Basis (Eigenbasis)

- Let us assume that there are  $n$  eigenvectors with corresponding eigenvalues:

$$\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \dots, \mathbf{v}^{(n)} \qquad \lambda_1, \lambda_2, \dots, \lambda_n$$

- Theorem (\*):**

- Let  $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \dots, \mathbf{v}^{(n)}$  be the eigenvectors of an  $n \times n$  matrix. If they correspond to **distinct** eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_n$ , then the system  $(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \dots, \mathbf{v}^{(n)})$  is linearly independent
- Hence, any vector  $\mathbf{v} \in \mathbb{R}^n$  can be **uniquely** expressed as a linear combination of the eigenvectors:

$$\mathbf{v} = c_1 \mathbf{v}^{(1)} + c_2 \mathbf{v}^{(2)} + \dots + c_n \mathbf{v}^{(n)}$$



# 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

$$\chi_{\mathbf{X}}(\lambda) := \det(\lambda \mathbf{I}_n - \mathbf{X}) \quad (20)$$

- For each eigenvalue  $\lambda_j$  we have to solve the homogeneous system of linear equations (see [⇒ here](#))

$$(\mathbf{X} - \lambda_j \mathbf{I}_n) \mathbf{v} = \mathbf{0} \quad (21)$$

to obtain the respective eigenvectors

## Example: Computation of Eigenvalues and Eigenvectors

- Let  $\mathbf{A} := \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \in \mathbb{R}^{2 \times 2}$ .
- The eigenvalues are the roots of the **characteristic polynomial** 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 set the characteristic polynomial to zero and directly see:

$$\lambda_1 = 1, \lambda_2 = 3$$

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

## Computation of the eigenspaces

- We start with  $\lambda_1 = 1$ :

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

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



# Diagonalizable Matrices

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

$$\mathbf{S}^{-1} \mathbf{A} \mathbf{S} = \mathbf{B} \in \mathbb{R}^{n \times n} \quad (22)$$

- The columns of  $\mathbf{S}$  are formed 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 squared  $n \times n$  matrix  $\mathbf{X}$  is called **symmetric**, if and only if

$$\mathbf{X} = \mathbf{X}^T \quad (23)$$

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

## Example: Eigen-Decomposition

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

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

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

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

# Positive (semi-)definite Matrices

- A square matrix  $\mathbf{X} \in \mathbb{R}^{n \times n}$  is called **positive definite** ( $\mathbf{X} \succ 0$ ), if for any non-zero real-valued vector  $\mathbf{z} \in \mathbb{R}^n$ :

$$\mathbf{z}^T \mathbf{X} \mathbf{z} > 0 \quad (24)$$

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

$$\mathbf{z}^T \mathbf{X} \mathbf{z} \geq 0 \quad (25)$$

Such matrices are important in machine learning. For instance, the covariance matrix is 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**

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

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

# Random Variables

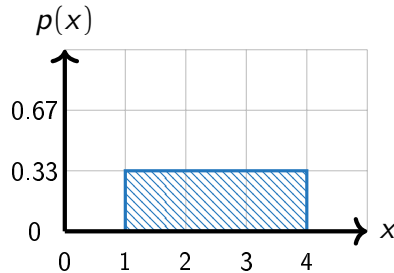
- What is a **random variable**?
  - It's a random number determined by chance (according to an underlying distribution). To be precise: A random variable  $\mathcal{X}$  is a **measurable function**

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

- Random variables in machine learning: Input data, output data, noise
- What is a **probability distribution**?
  - 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  $\mathbb{R} := [a, b]$

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

# 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:

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

- 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):

$$\mathcal{X} \in \{0, 1\} \quad (26)$$

$$p(\mathcal{X} = 1|\mu) = \mu \quad (27)$$

$$p(\mathcal{X} = 0|\mu) = 1 - \mu \quad (28)$$

$$\mathbb{E}\{\mathcal{X}\} = \mu \quad (29)$$

$$\mathbb{V}\{\mathcal{X}\} = \mu(1 - \mu) \quad (30)$$

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

# Binomial Distribution

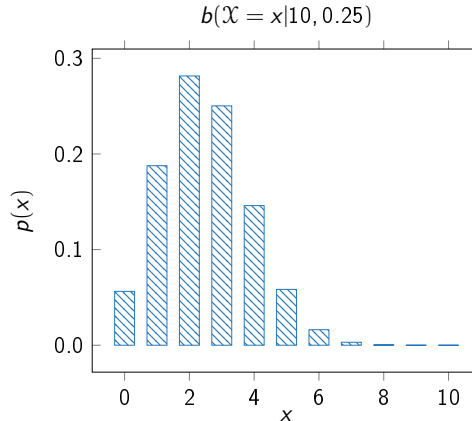
- Repeating a Bernoulli experiment  $n$  times leads to the **binomial distribution**
- **Example:** What is the probability of getting  $k \in \mathbb{N}$  heads in  $n$  trials?

$$b(\mathcal{X} = k | n, \mu) := \binom{n}{k} \mu^k (1 - \mu)^{n-k} \quad (31)$$

$$\mathbb{E}\{\mathcal{X}\} = n\mu \quad (32)$$

$$\mathbb{V}\{\mathcal{X}\} = n\mu(1 - \mu) \quad (33)$$

# Binomial Distribution (Ctd.)



# Continuous Distributions

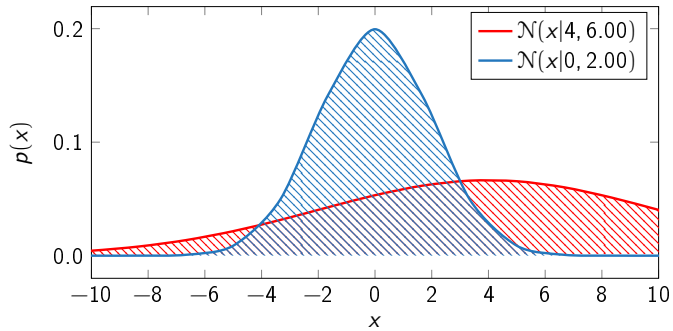
The random variables take on **continuous values**

- 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:

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



# Gaussian Distribution



$$p(\mathcal{X} = x) := \mathcal{N}(\mathcal{X} = x | \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad (34)$$





# Central Limit Theorem

## Central Limit Theorem:

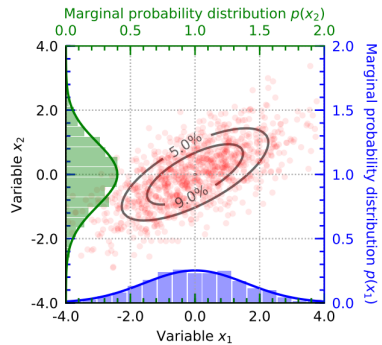
The distribution of the sum of  $n$  i.i.d. (independent and identically distributed) random variables becomes increasingly Gaussian as  $n$  increases

- The Gaussian distribution is one of the most important distributions
- Gaussians 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

$$p_D(\mathbf{x}) := \mathcal{N}_D(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{1}{\sqrt{(2\pi)^D \det(\boldsymbol{\Sigma})}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \quad (35)$$



**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:** (Probability of  $x$  **and**  $y$ )

$$p(x \cap y) \quad (36)$$

- **Marginal distribution:** (summing out)

$$p(y) = \int_x p(x \cap y) dx \quad (37)$$

- **Conditional distribution:** (Probability of  $y$  **given**  $x$ )

$$p(y|x) = \frac{p(x \cap y)}{p(x)} \quad (38)$$

# Basic Rules of Probability (Ctd.)

- Probabilistic independence:

$$p(x \cap y) = p(x)p(y) \quad (39)$$

- Chain rule of probabilities:

$$\begin{aligned} p(x_1 \cap \dots \cap x_n) &= p(x_1 | x_2 \cap \dots \cap x_n) p(x_2 \cap \dots \cap x_n) \\ &= p(x_1 | x_2 \cap \dots \cap x_n) p(x_2 | x_3 \cap \dots \cap x_n) \dots p(x_{n-1} | x_n) p(x_n) \end{aligned} \quad (40)$$

- Bayes' rule:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \quad (41)$$



# Expectation

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

$$\mathbb{E}\{\mathcal{X}\} := \sum_{k \in \Omega(\mathcal{X})} k \cdot p(\mathcal{X} = k) \quad (42)$$

- Expectations of functions:

$$\mathbb{E}_x\{f\} := \sum_x p(x) f(x) \quad (43)$$

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

# Expectation (Ctd.)

## Rules of expectations:

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

$$\mathbb{E}\{a\mathcal{X} + b\mathcal{Y}\} = a\mathbb{E}\{\mathcal{X}\} + b\mathbb{E}\{\mathcal{Y}\} \quad (44)$$

- More general:  $\mathbb{E}\{\sum_{i=1}^n a_i \mathcal{X}_i\} = \sum_{i=1}^n a_i \mathbb{E}\{\mathcal{X}_i\}$
- 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**:

$$\mathcal{X} \leq \mathcal{Y} \implies \mathbb{E}\{\mathcal{X}\} \leq \mathbb{E}\{\mathcal{Y}\} \quad (45)$$



# Variance

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

$$\mathbb{V}\{\mathcal{X}\} := \mathbb{E}\{\mathcal{X} - \mathbb{E}^2\{\mathcal{X}\}\} = \mathbb{E}\{\mathcal{X}^2\} - \mathbb{E}^2\{\mathcal{X}\} \quad (46)$$

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

$$\mathbb{V}\{a + b\mathcal{X}\} = b^2\mathbb{V}\{\mathcal{X}\} \quad (47)$$

$$\mathbb{V}\{\mathcal{X} + \mathcal{Y}\} = \mathbb{V}\{\mathcal{X}\} + \mathbb{V}\{\mathcal{Y}\} + \text{cov}\{\mathcal{X}, \mathcal{Y}\} \quad (48)$$

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

$$\mathbb{V}\{\mathcal{X} + \mathcal{Y}\} = \mathbb{V}\{\mathcal{X}\} + \mathbb{V}\{\mathcal{Y}\} \quad (49)$$

# Covariance

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

$$\begin{aligned}\text{cov}\{x, y\} &:= \mathbb{E}\left\{ (x - \mathbb{E}\{x\})(y - \mathbb{E}\{y\}) \right\} \\ &= \mathbb{E}\{xy\} - \mathbb{E}\{x\}\mathbb{E}\{y\}\end{aligned}\tag{50}$$

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

$$\mathbb{V}\{x\} = \text{cov}\{x, x\}$$





# Kullback-Leibler Divergence

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

$$\text{KL}(p\|q) := \sum_x p(x) \cdot \log \frac{p(x)}{q(x)} \quad (51)$$

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

## Section: Optimization Techniques

Partial Derivatives, Gradients, and Hessian Matrix  
Cost Functions and Convexity  
Unconstrained Optimization  
Constrained Optimization and Lagrange Multipliers  
Numerical Optimization Techniques

# Motivation

- In every machine learning problem, you will have:
  - ① An **objective function** you want to optimize
  - ② **Data** you want to learn from
  - ③ **Parameters** which need to be learned
  - ④ Assumptions about the problem and the data
- We would like to have general solutions to the problem of learning
- Different algorithms embody different objective functions and assumptions

**Every machine learning problem is an optimization problem!**

# Partial Derivatives

- **Definition:** Let  $D \subset \mathbb{R}^m$  be an open set and  $f : D \rightarrow \mathbb{R}$ . We say that  $f$  is in  $\mathbf{x_0} \in D$  **partially differentiable** with respect to the  $i$ -th input variable  $x_i$ , if and only if the limit

$$\lim_{h \rightarrow 0} \left( \frac{1}{h} (f(\mathbf{x}_0 + h\mathbf{e}_i) - f(\mathbf{x}_0)) \right) \quad (52)$$

exists ( $\mathbf{e}_i$  is the  $i$ -th standard unit vector)

- The limit (52) is denoted by

$$\frac{\partial}{\partial x_i} f(\mathbf{x}_0) \quad (53)$$

# Gradients

- The **gradient** of  $f$  is a vector comprising all partial derivatives of  $f$ :

$$\nabla f(\mathbf{x}_0) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}_0) \\ \frac{\partial}{\partial x_2} f(\mathbf{x}_0) \\ \vdots \\ \frac{\partial}{\partial x_m} f(\mathbf{x}_0) \end{pmatrix} \quad (54)$$

- The symbol  $\nabla$  is pronounced '*nabla*' (it is **not** a capital delta  $\Delta$ !)



## Gradients (Ctd.)

The gradient  $\nabla f$  points into the direction of steepest ascent.

**Proof:** Let  $\mathbf{v} \in \mathbb{R}^m$  be a unit vector, i. e.  $\|\mathbf{v}\| = 1$ . The directional derivative of  $f$  in the direction of  $\mathbf{v}$  is given by the scalar product  $\mathbf{v}^T \nabla f$ . We aim to find the vector  $\mathbf{v}$  which maximizes the directional derivative. This vector will then point into the direction of steepest ascent. We can rewrite the scalar product using its geometrical definition and receive  $\mathbf{v}^T \nabla f = \|\mathbf{v}\| \cdot \|\nabla f\| \cdot \cos(\alpha)$ . Since  $\mathbf{v}$  has length 1, the equation simplifies to  $\mathbf{v}^T \nabla f = \|\nabla f\| \cdot \cos(\alpha)$ . Only  $\alpha$  can be changed to maximize this quantity (since the gradient is given). The cosine reaches a maximum value for  $\alpha = 0$ . This means  $\nabla f$  is parallel to  $\mathbf{v}$ , i. e.  $\nabla f$  points into the direction of steepest ascent. □

Similarly we can show that  $-\nabla f$  points into the direction of steepest descent!

# Computation of the Gradient

## How can we compute the gradient?

- Luckily, we can apply the same rules of differentiation we know from one-dimensional calculus: **Chain rule**, **product rule**, ...
- Differentiate with respect to the variable you are interested in and hold the others fixed
- **Example:**  $f(x_1, x_2) = \log(x_1 x_2) + \sin(x_2)$

$$\nabla f(x_1, x_2) = \begin{pmatrix} \frac{1}{x_1} \\ \frac{1}{x_2} + \cos(x_2) \end{pmatrix}$$

# Hessian Matrix

- The **Hessian matrix** contains all second-order partial derivatives of  $f$ :

$$\nabla^2 f(\mathbf{x}_0) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f(\mathbf{x}_0) & \frac{\partial^2}{\partial x_1 \partial x_2} f(\mathbf{x}_0) & \dots & \frac{\partial^2}{\partial x_1 \partial x_m} f(\mathbf{x}_0) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(\mathbf{x}_0) & \frac{\partial^2}{\partial x_2 \partial x_2} f(\mathbf{x}_0) & \dots & \frac{\partial^2}{\partial x_2 \partial x_m} f(\mathbf{x}_0) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_m \partial x_1} f(\mathbf{x}_0) & \frac{\partial^2}{\partial x_m \partial x_2} f(\mathbf{x}_0) & \dots & \frac{\partial^2}{\partial x_m \partial x_m} f(\mathbf{x}_0) \end{pmatrix} \in \mathbb{R}^{m \times m} \quad (55)$$

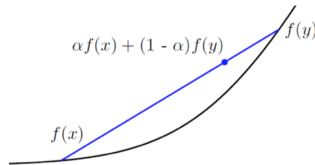
- $f$  has continuous second-order derivatives  $\implies \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x}_0) = \frac{\partial^2}{\partial x_j \partial x_i} f(\mathbf{x}_0)$   
**(Schwarz's theorem)**



# Convexity – Convex Functions

- A function  $f : \mathbb{R}^m \rightarrow \mathbb{R}$  is convex, if  $\forall \mathbf{x}, \mathbf{y} \in \text{dom}(f)$  and  $\forall \alpha \in [0, 1]$ :

$$f(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}) \leq \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}) \quad (56)$$



- Examples are (affine-)linear functions  $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b$  and quadratic functions  $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$

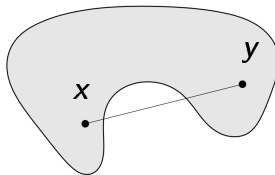
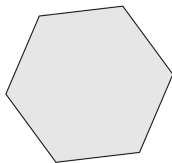
# Convexity – Convex Sets

- A set  $C \subset \mathbb{R}^m$  is convex, if  $\forall x, y \in C$  and  $\forall \alpha \in [0, 1]$ :

$$\alpha x + (1 - \alpha)y \in C \quad (57)$$

- This is the equation of the line segment between  $x$  and  $y$

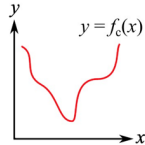
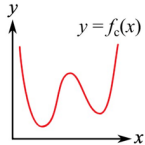
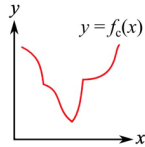
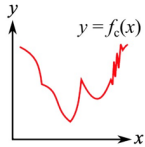
convex



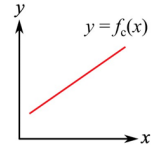
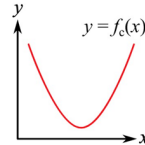
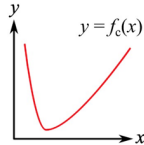
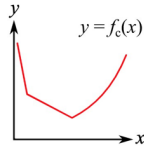
non-convex

# Convexity – Convex and Non-convex Functions

non-convex



convex



# Advantages of Convexity

## Why are convex cost functions so appealing?

- Local solutions are global optima
- Numeric optimizers do not get stuck in local optima
- Efficient implementations of optimizers are available

# Unconstrained Optimization

- Let a function  $\begin{cases} f : \mathbb{R}^m \rightarrow \mathbb{R} \\ \mathbf{x} \mapsto f(\mathbf{x}) \end{cases}$  be given
- How do we find minimums and maximums of  $f$ ?
- **Necessary condition:** A point  $\mathbf{x}_0 \in \mathbb{R}^m$  is a potential candidate if

$$\nabla f(\mathbf{x}_0) = \mathbf{0} \tag{58}$$

- Then check if the Hessian matrix  $\nabla^2 f(\mathbf{x}_0)$  is positive definite (minimum) or negative definite (maximum)

# Constrained Optimization

## Formalization:

$\min_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta})$   $\longleftarrow$  cost function / objective function

subject to  $f(\boldsymbol{\theta}) = 0$   $\longleftarrow$  equality constraints

$g(\boldsymbol{\theta}) \geq 0$   $\longleftarrow$  inequality constraints

**What should an ideal optimization problem, i. e. cost function and constraints, look like?**

# Constrained Optimization (Ctd.)

Answer:

$$\min_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta})$$

← **convex** function

$$\text{subject to } f(\boldsymbol{\theta}) = 0$$

← **linear** function

$$g(\boldsymbol{\theta}) \geq 0$$

← **convex** set



# Constrained Optimization

How to solve this optimization problem?

$$\begin{aligned} & \mathcal{J}(x, y) := 2y + x \\ \min_{x, y} \mathcal{J}(x, y) \quad & \text{subject to: } f(x, y) := y^2 + xy - 1 = 0 \end{aligned}$$

**Solution:**

- Convert the problem to an unconstrained one
- This is done using the concept of **Lagrange multipliers**  $\lambda$





# The Concept of Lagrange Multipliers

**Lagrange function:**

$$\mathcal{L}(x, y, \lambda) := \mathcal{J}(x, y) + \lambda f(x, y)$$

**Step ❶:** Differentiate with respect to  $x$ ,  $y$  and  $\lambda$

I. 
$$\frac{\partial}{\partial x} \mathcal{L}(x, y, \lambda) = 1 + \lambda y$$

II. 
$$\frac{\partial}{\partial y} \mathcal{L}(x, y, \lambda) = 2 + 2\lambda y + \lambda x$$

III. 
$$\frac{\partial}{\partial \lambda} \mathcal{L}(x, y, \lambda) = y^2 + xy - 1$$



# The Concept of Lagrange Multipliers (Ctd.)

**Step ②:** Set derivatives to zero

$$\text{I. } 1 + \lambda y \stackrel{!}{=} 0$$

$$\text{II. } 2 + 2\lambda y + \lambda x \stackrel{!}{=} 0$$

$$\text{III. } y^2 + xy - 1 \stackrel{!}{=} 0$$

**Step ③:** Substitute

$$\text{I. } \lambda = -\frac{1}{y}$$

$$\text{I.} \rightarrow \text{II. } x = 0$$

$$\text{II.} \rightarrow \text{III. } y = \pm 1$$

# Exercise

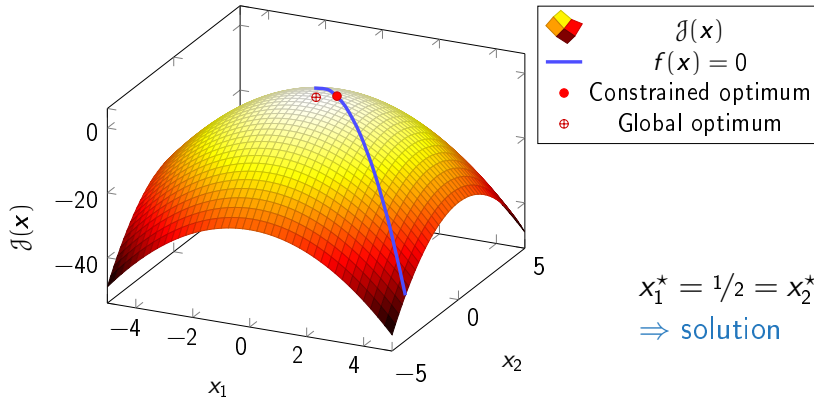
Optimize the following function:

$$\max_{x_1, x_2} \mathcal{J}(x_1, x_2) = 1 - x_1^2 - x_2^2$$

subject to:

$$f(x_1, x_2) = x_1 + x_2 - 1 = 0$$

# Exercise Solution





# Numerical Optimization

- There is a multitude of **numerical optimization algorithms** for optimizing functions on a computer if we **cannot solve it analytically**
- Incrementally update an estimate of the parameters:

$$\boldsymbol{\theta}_{\text{new}} \longleftarrow \boldsymbol{\theta}_{\text{old}} + \Delta\boldsymbol{\theta} \quad (59)$$

- Choose  $\Delta\boldsymbol{\theta}$  such that  $\mathcal{J}(\boldsymbol{\theta}_{\text{new}}) < \mathcal{J}(\boldsymbol{\theta}_{\text{old}})$
- The algorithms differ in the number of iterations required, the computational cost, the convergence guarantees, the robustness with noisy cost functions, and their memory usage

# Numerical Optimization Algorithms

Gradient based methods:

- **Gradient descent**
- **(L-)BFGS** (Broyden-Fletcher-Goldfarb-Shanno)
- **Conjugate gradient descent**

There are also non-gradient based methods like genetic algorithms, particle swarm, non-linear simplex, Nelder-Mead

**Numerical optimization techniques may not find the global optimum!**



# Gradient Descent

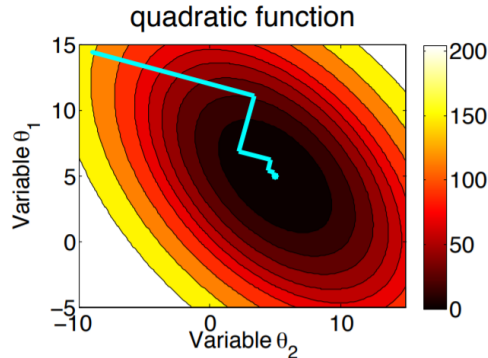
- Gradient descent is a basic, but most commonly used algorithm in ML
- **General idea:** Go into the direction of the **steepest descent**
- The gradient points into the direction of steepest ascent, therefore subtract the gradient from the current estimate:

$$\Delta\theta \longleftarrow -\alpha \nabla_{\theta} \mathcal{J}(\theta_{\text{old}}) \qquad \theta_{\text{new}} \longleftarrow \theta_{\text{old}} + \Delta\theta \qquad (60)$$

- $\alpha \in (0, 1)$  is called the **learning rate**; It stabilizes the learning process
- Gradient descent is a **first-order method**

## Gradient Descent (Ctd.)

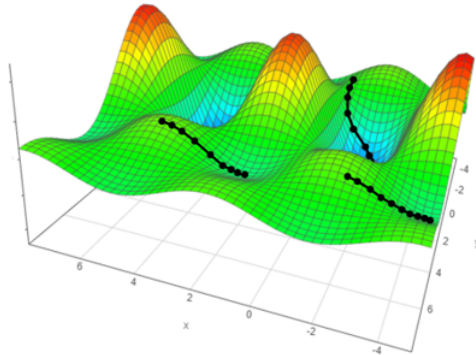
The parameter updates tend to 'zig-zag' down the valley:





# Gradient Descent (Ctd.)

Initialization also matters...



# Taylor Expansion

- Let  $f$  be  $n$ -times differentiable at  $x_0 \in \mathbb{R}$ . We can approximate  $f$  using a **Taylor polynomial** of degree  $n$  given by

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \dots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n \quad (61)$$

- Alternatively we can write:

$$f(x + \Delta x) \approx f(x_0) + f'(x_0)\Delta x + \frac{f''(x_0)}{2!}(\Delta x)^2 + \dots + \frac{f^{(n)}(x_0)}{n!}(\Delta x)^n \quad (62)$$

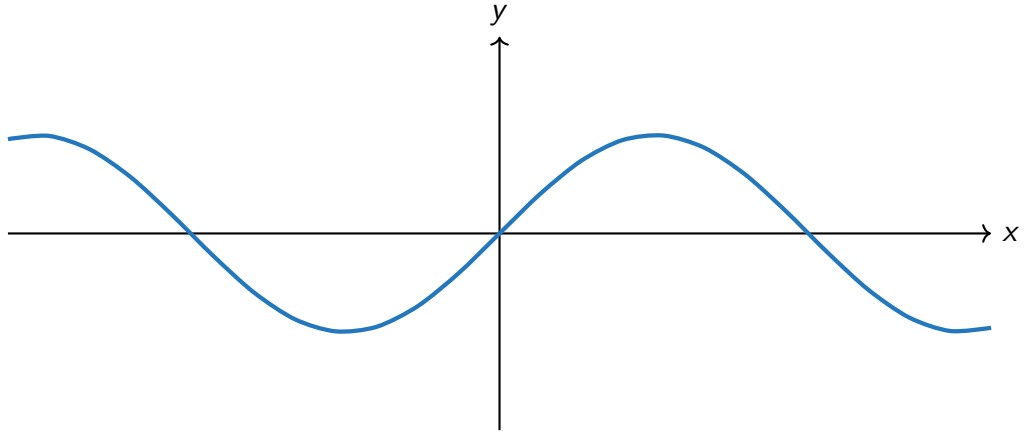
## Example: Taylor Polynomial for $\sin(x)$

- Let us compute a degree 5 Taylor expansion of  $\sin(x)$  at  $x_0 = 0$ :

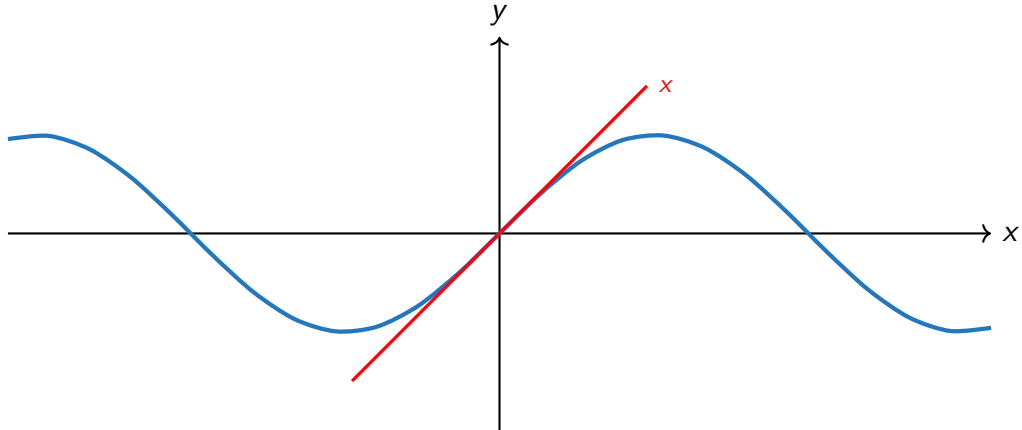
$$\begin{aligned}\sin(x) &\approx \sin(0) + \frac{x}{1} \cos(0) - \frac{x^2}{2} \sin(0) - \frac{x^3}{6} \cos(0) + \frac{x^4}{24} \sin(0) + \frac{x^5}{120} \cos(0) \\ &= x - \frac{x^3}{6} + \frac{x^5}{120}\end{aligned}$$

- We could continue the expansion forever because  $\sin(x)$  is infinitely differentiable:  $\sin(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}$

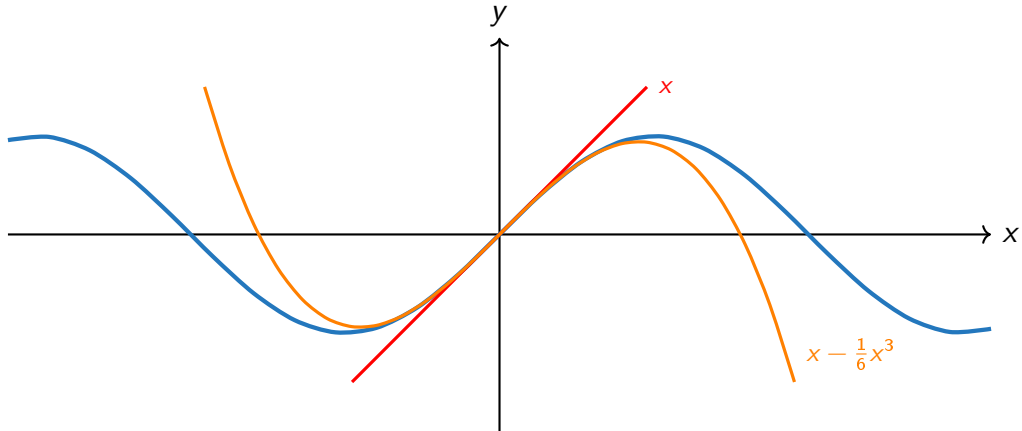
## Example: Taylor Polynomial for $\sin(x)$ (Ctd.)



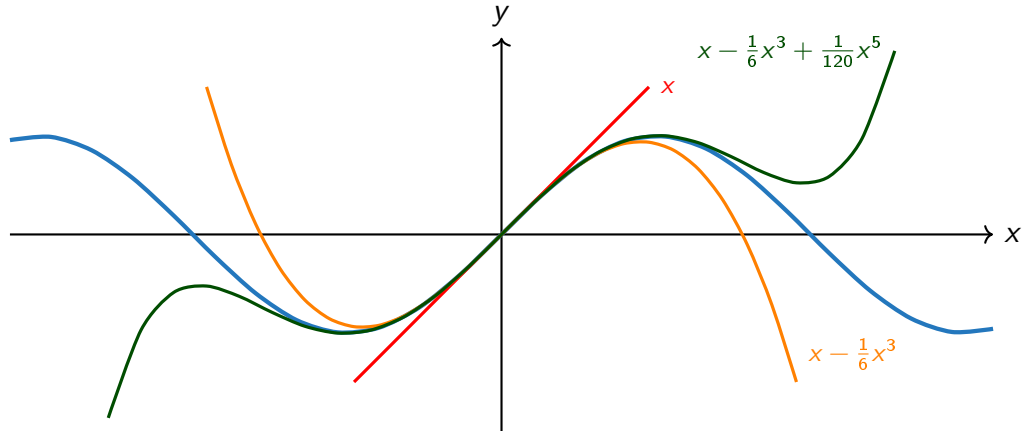
## Example: Taylor Polynomial for $\sin(x)$ (Ctd.)



## Example: Taylor Polynomial for $\sin(x)$ (Ctd.)



## Example: Taylor Polynomial for $\sin(x)$ (Ctd.)



# Newton's Method

- **Newton's method** is a **second-order method**, also called **Newton-Raphson algorithm**
- We consider the second-order **Taylor expansion** of  $\mathcal{J}(\boldsymbol{\theta})$ :

$$\mathcal{J}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) := \mathcal{J}(\boldsymbol{\theta}) + \mathbf{g}(\boldsymbol{\theta})^\top \Delta\boldsymbol{\theta} + \frac{1}{2} \Delta\boldsymbol{\theta}^\top \mathbf{H}(\boldsymbol{\theta}) \Delta\boldsymbol{\theta} \quad (63)$$

where  $\mathbf{g}(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta})$  and  $\mathbf{H}(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}}^2 \mathcal{J}(\boldsymbol{\theta})$

- We differentiate (63) with respect to  $\Delta\boldsymbol{\theta}$  and set the derivative to zero:

$$\nabla_{\Delta\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) = \mathbf{g}(\boldsymbol{\theta}) + \mathbf{H}(\boldsymbol{\theta}) \Delta\boldsymbol{\theta} \stackrel{!}{=} \mathbf{0} \quad (64)$$



# Newton's Method (Ctd.)

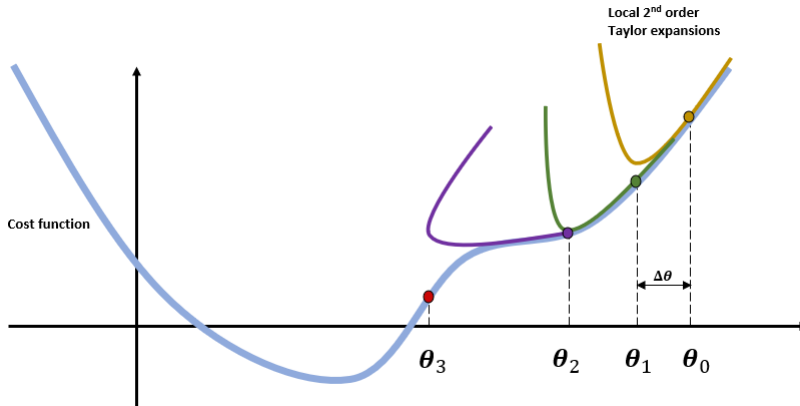
- We obtain:  $\Delta\theta = -H^{-1}(\theta)g(\theta)$
- The update rule then takes the form

$$\Delta\theta \longleftarrow -H^{-1}(\theta_{\text{old}})g(\theta_{\text{old}}) \quad (65)$$

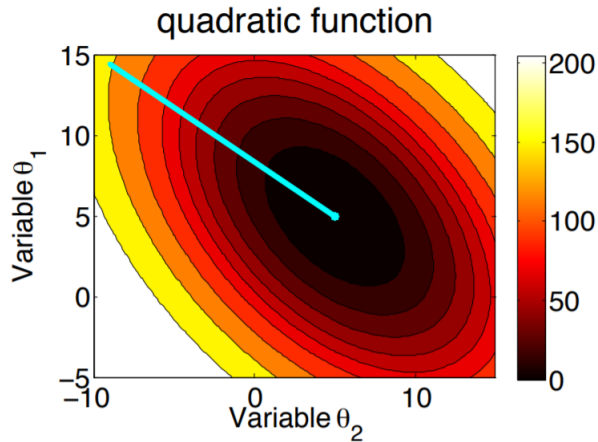
$$\theta_{\text{new}} \longleftarrow \theta_{\text{old}} + \Delta\theta \quad (66)$$

- The learning rate  $\alpha$  has been replaced by the inverse of the Hessian matrix
- Newton's method **converges faster** than gradient descent, but the inverse of the Hessian is **expensive to compute**

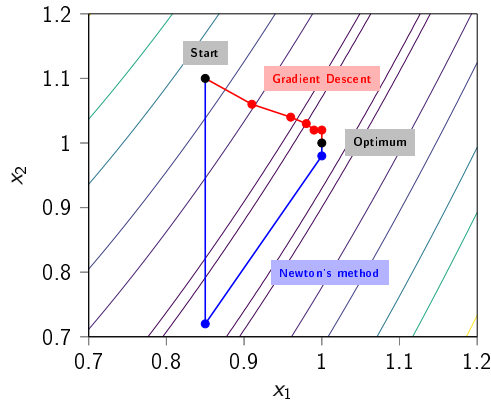
# Newton's Method (Ctd.)



# Newton's Method (Ctd.)



# Newton's Method in Comparison to Gradient Descent



## Section: Wrap-Up

Summary  
Self-Test Questions  
Lecture Outlook

# Summary I

- Machine learning is math!
- 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
  - The **eigendecomposition** plays an import role in many machine learning applications

# Summary II

- **Statistics:**
  - Random variables are numbers **determined by chance**
  - Probability distributions describe a **probability mass** or **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** is

# Summary III

- Optimization:
  - Every machine learning problem is an optimization problem!
  - Good cost functions are **convex**
  - You should be familiar with unconstrained and constrained optimization (Lagrange multipliers)
  - Closed-form solutions are not always feasible. In such cases we use **numerical optimization techniques**
  - The most prominent numerical technique is called **gradient descent**





# 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? Where do they play a role?
- ⑤ What are random variables and probability distributions?
- ⑥ Why is the Gaussian distribution so important?
- ⑦ What is Bayes' rule? Explain its components!
- ⑧ What is convexity? Why should cost functions be convex?
- ⑨ What is gradient descent?

# What's next...?

Unit I	Machine Learning Introduction
Unit II	Mathematical Foundations
<b>Unit III</b>	<b>Bayesian Decision Theory</b>
Unit IV	Regression
Unit V	Classification I
Unit VI	Evaluation
Unit VII	Classification II
Unit VIII	Clustering
Unit IX	Dimensionality Reduction

Thank you very much for the attention!

**Topic:** \*\*\* Applied Machine Learning Fundamentals \*\*\* Mathematical Foundations

**Term:** Winter term 2023/2024

**Contact:**

Daniel Wehner, M.Sc.

SAP SE / DHBW Mannheim

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

Do you have any questions?