*** 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
Unit II	**Mathematical Foundations**
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

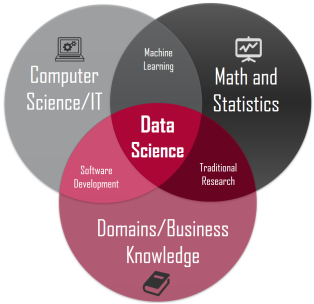
Section:

Introduction

Introduction
Math is important!

Introduction

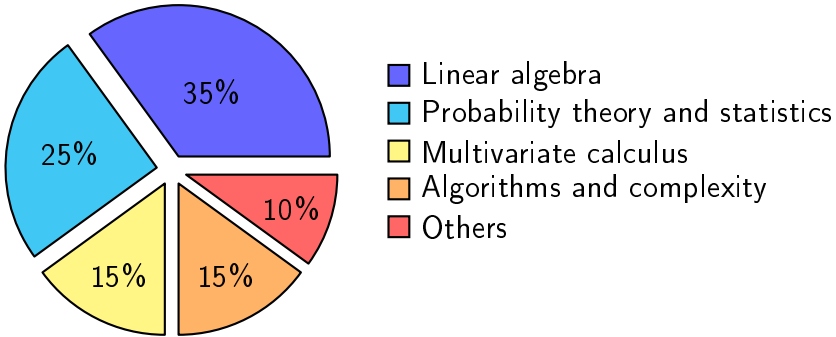
Maths play a major role in data science and machine learning!



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

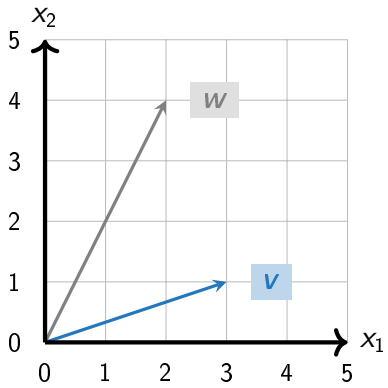
What is a Vector?

General:

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

Example: $(m = 2)$

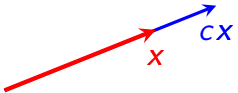
$$\boldsymbol{v} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \qquad \boldsymbol{w} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$



Multiplication by a Scalar

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

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



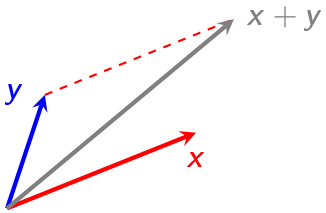
Example: $(m = 2)$

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

Addition of Vectors

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

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



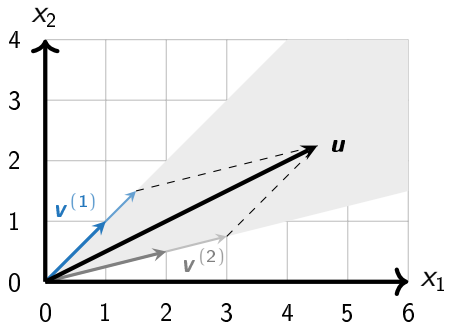
Example: $(m = 2)$

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

Linear Combination of Vectors

Let $c_1, \ldots, c_n \in \mathbb{R}$ be scalars and $\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n)} \in \mathbb{R}^m$ be vectors.
A **linear combination** of these vectors is given by $\boldsymbol{u} \in \mathbb{R}^m$:

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



Vector Transpose, inner and outer Product

- **Vector transpose:**

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

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

$$\boldsymbol{x}^\intercal \boldsymbol{y} \equiv \langle \boldsymbol{x}, \boldsymbol{y} \rangle \equiv \boldsymbol{x} \cdot \boldsymbol{y} := \sum_{j=1}^{m} x_j y_j \tag{2}$$

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

Vector Transpose and inner and outer Product (Ctd.)

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

$$xy^\intercal := \begin{pmatrix} x_1y_1 & x_1y_2 & \dots & x_1y_m \\ x_2y_1 & x_2y_2 & \dots & x_2y_m \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \dots & x_my_m \end{pmatrix} \quad (3)$$

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

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 $x, y \in \mathbb{R}^m$ and $c \in \mathbb{R}$

$$\|x\| := \sqrt{x^\intercal x} \tag{4}$$

$$\|cx\| = |c| \cdot \|x\| \tag{5}$$

$$\|x + y\| \leqslant \|x\| + \|y\| \tag{6}$$

- Example:

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

Angle between Vectors

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

$$\cos \measuredangle(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^\intercal \boldsymbol{y}}{\|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|} = \frac{\sum_{j=1}^{m} x_j y_j}{\sqrt{\sum_{j=1}^{m} x_j^2} \cdot \sqrt{\sum_{j=1}^{m} y_j^2}} \tag{7}$$

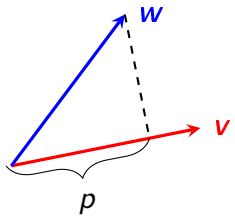
$$\cos \measuredangle(\boldsymbol{v}, \boldsymbol{w}) = \frac{\boldsymbol{v}^\intercal \boldsymbol{w}}{\|\boldsymbol{v}\| \cdot \|\boldsymbol{w}\|} = \frac{10}{\sqrt{10} \cdot \sqrt{20}} \approx 0.71$$

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

(Orthogonal) Projection of Vectors

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

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



- 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_{ij} is the entry in row i and column j
(*'**Z**eilen zuerst, **S**palten später'*)

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

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

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

Matrix Transpose

Transpose of a matrix:

$$\boldsymbol{M}^{\mathsf{T}} = \begin{pmatrix} 3 & 4 & 5 \\ 1 & 0 & 1 \end{pmatrix}^{\mathsf{T}} = \begin{pmatrix} 3 & 1 \\ 4 & 0 \\ 5 & 1 \end{pmatrix} \tag{9}$$

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

Matrix Addition

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

$$\boldsymbol{X} + \boldsymbol{Y} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix} + \begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1m} \\ Y_{21} & Y_{22} & \dots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \dots & Y_{nm} \end{pmatrix}$$

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

Multiplication of Matrices by Scalars

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

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

Multiplication of Matrices by Vectors

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

$$\boldsymbol{X}y = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix} \begin{pmatrix} y_1 \\ v_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{m} X_{1j} y_j \\ \sum_{j=1}^{m} X_{2j} y_j \\ \vdots \\ \sum_{j=1}^{m} X_{nj} y_j \end{pmatrix} \in \mathbb{R}^n \qquad (12)$$

Matrix Multiplication (Ctd.)

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

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

where:

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

Determinants

- The **determinant** of a 2×2 matrix is given by:

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

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

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

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

Matrix Inversion

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

$$\boldsymbol{X}^{-1}\boldsymbol{X} = \boldsymbol{X}\boldsymbol{X}^{-1} = \boldsymbol{I}_n \tag{17}$$

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

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

Matrix Inversion (Ctd.)

Let $\boldsymbol{X} \in \mathbb{R}^{n \times n}$. The following statements are equivalent:

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

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

$\iff \boldsymbol{X}$ has rank n (full rank)

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

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

Matrix Inversion (Ctd.)

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

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

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

Matrix Inversion Example

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

Please verify for yourself!

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

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

Eigenvectors and Eigenvalues

- Let $\boldsymbol{X} \in \mathbb{R}^{n \times n}$. Some vectors $\boldsymbol{v} \in \mathbb{R}^n$ only change their length (but not their direction) when multiplied by $\boldsymbol{X}$
- **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 $\boldsymbol{X}$, the scaling factors are known as **eigenvalues** of $\boldsymbol{X}$
- More general:

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

Eigenvectors form a Basis (Eigenbasis)

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

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

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

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

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

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

- For each eigenvalue λ_j we have to solve the homogeneous system of linear equations (see $\Rightarrow$ here)

$$(\boldsymbol{X} - \lambda_j \boldsymbol{I}_n)\boldsymbol{v} = \boldsymbol{0} \tag{21}$$

to obtain the respective eigenvectors

Example: Computation of Eigenvalues and Eigenvectors

- Let $\boldsymbol{A} := \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \in \mathbb{R}^{2\times 2}$.
- The eigenvalues are the roots of the **characteristic polynomial** given by

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

- 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 $\lambda_1 = 1$:

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

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

Diagonalizable Matrices

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

$$\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S} = \boldsymbol{B} \in \mathbb{R}^{n \times n} \tag{22}$$

- The columns of $\boldsymbol{S}$ are formed 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**

Symmetric Matrices

- A squared $n \times n$ matrix $\boldsymbol{X}$ is called **symmetric**, if and only if

$$\boldsymbol{X} = \boldsymbol{X}^\intercal \tag{23}$$

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

Example: Eigen-Decomposition

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

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

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

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

Positive (semi-)definite Matrices

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

$$z^\intercal \boldsymbol{X} z > 0 \tag{24}$$

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

$$z^\intercal \boldsymbol{X} z \geqslant 0 \tag{25}$$

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**

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

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

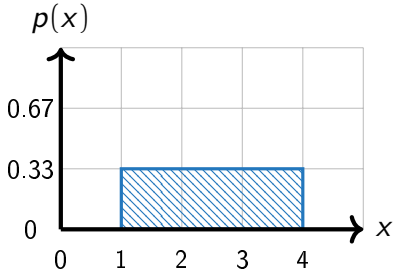
Random Variables

- 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**

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

  - 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):

$$\mathcal{X} \in \{0, 1\} \tag{26}$$
$$p(\mathcal{X} = 1|\mu) = \mu \tag{27}$$
$$p(\mathcal{X} = 0|\mu) = 1 - \mu \tag{28}$$
$$\mathbb{E}\{\mathcal{X}\} = \mu \tag{29}$$
$$\mathbb{V}\{\mathcal{X}\} = \mu(1 - \mu) \tag{30}$$

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

Binomial Distribution

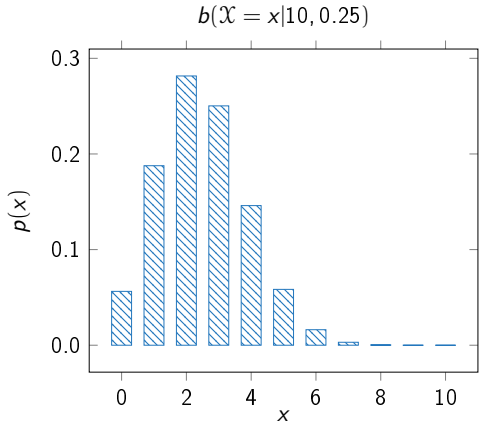
- Repeating a Bernoulli experiment n times leads to the **binomial distribution**
- **Example:** What is the probability of getting $k \in \mathbb{N}$ heads in n trials?

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

$$\mathbb{E}\{\mathcal{X}\} = n\mu \tag{32}$$

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

Binomial Distribution (Ctd.)



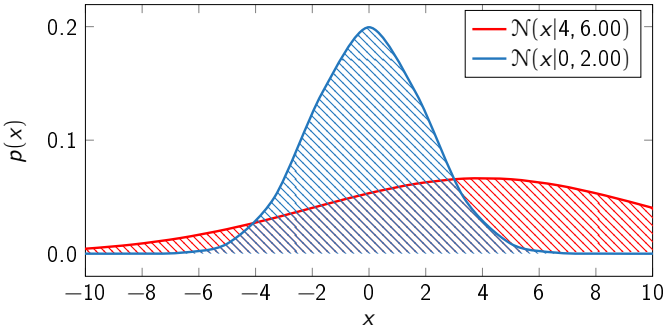
Continuous Distributions

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:

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

Gaussian Distribution



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

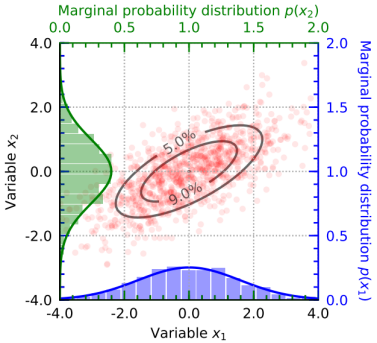
Central Limit Theorem

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

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



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

$$p(x \cap y) \tag{36}$$

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

$$p(y) = \int_x p(x \cap y)\, \mathrm{d}x \tag{37}$$

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

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

Basic Rules of Probability (Ctd.)

- **Probabilistic independence:**

$$p(x \cap y) = p(x)p(y) \tag{39}$$

- **Chain rule of probabilities:**

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

- **Bayes' rule:**

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

Expectation

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

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

- Expectations of functions:

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

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

Expectation (Ctd.)

Rules of expectations:

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

$$\mathbb{E}\{a\mathcal{X} + b\mathcal{Y}\} = a\mathbb{E}\{\mathcal{X}\} + b\mathbb{E}\{\mathcal{Y}\} \tag{44}$$

- 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**:

$$\mathcal{X} \leqslant \mathcal{Y} \Longrightarrow \mathbb{E}\{\mathcal{X}\} \leqslant \mathbb{E}\{\mathcal{Y}\} \tag{45}$$

Variance

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

$$\mathbb{V}\{\mathcal{X}\} := \mathbb{E}\{\mathcal{X} - \mathbb{E}^2\{\mathcal{X}\}\} = \mathbb{E}\{\mathcal{X}^2\} - \mathbb{E}^2\{\mathcal{X}\} \tag{46}$$

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

$$\mathbb{V}\{a + b\mathcal{X}\} = b^2\mathbb{V}\{\mathcal{X}\} \tag{47}$$

$$\mathbb{V}\{\mathcal{X} + \mathcal{Y}\} = \mathbb{V}\{\mathcal{X}\} + \mathbb{V}\{\mathcal{Y}\} + \mathrm{cov}\{\mathcal{X}, \mathcal{Y}\} \tag{48}$$

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

$$\mathbb{V}\{\mathcal{X} + \mathcal{Y}\} = \mathbb{V}\{\mathcal{X}\} + \mathbb{V}\{\mathcal{Y}\} \tag{49}$$

Covariance

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

$$\begin{aligned}\mathrm{cov}\{\mathcal{X}, \mathcal{Y}\} &:= \mathbb{E}\Big\{\big(\mathcal{X} - \mathbb{E}\{\mathcal{X}\}\big)\big(\mathcal{Y} - \mathbb{E}\{\mathcal{Y}\}\big)\Big\} \\ &= \mathbb{E}\{\mathcal{X}\mathcal{Y}\} - \mathbb{E}\{\mathcal{X}\}\mathbb{E}\{\mathcal{Y}\} \end{aligned} \tag{50}$$

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

$$\mathbb{V}\{\mathcal{X}\} = \mathrm{cov}\{\mathcal{X}, \mathcal{X}\}$$

Kullback-Leibler Divergence

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

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

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

Section:
Optimization Techniques

Partial Derivatives, Gradients, and Hessian Matrix
Cost Functions and Convexity
Unconstrained Optimization (analytical)
Constrained Optimization and Lagrange Multipliers (analytical)
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 $f : D \to \mathbb{R}$. We say that f is in $x_0 \in D$ **partially differentiable** with respect to the i-th input variable x_i, if and only if the limit

$$\lim_{h \to 0} \left(\frac{1}{h} \big(f(x_0 + h e_i) - f(x_0) \big) \right) \tag{52}$$

exists (e_i is the i-th standard unit vector)

- The limit (52) is denoted by

$$\frac{\partial}{\partial x_i} f(x_0) \tag{53}$$

Gradients

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

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

- The symbol ∇ is pronounced *'nabla'* (it is **not** a capital delta Δ!)

Gradients (Ctd.)

The gradient ∇f points into the direction of steepest ascent.

Proof: Let $v \in \mathbb{R}^m$ be a unit vector, i. e. $\|v\| = 1$. The directional derivative of f in the direction of $\boldsymbol{v}$ is given by the scalar product $\boldsymbol{v}^\intercal \nabla f$. We aim to find the vector $\boldsymbol{v}$ 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 $\boldsymbol{v}^\intercal \nabla f = \|\boldsymbol{v}\| \cdot \|\nabla f\| \cdot \cos(\alpha)$. Since $\boldsymbol{v}$ has length 1, the equation simplifies to $\boldsymbol{v}^\intercal \nabla f = \|\nabla f\| \cdot \cos(\alpha)$. Only the angle α can be altered to maximize this quantity (since the gradient is given). The cosine reaches a maximum value for $\alpha = 0$. This means ∇f is parallel to $\boldsymbol{v}$, i. e. ∇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 one variable 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:

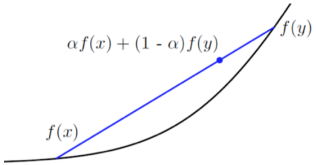
$$\nabla^2 f(\boldsymbol{x}_0) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f(\boldsymbol{x}_0) & \frac{\partial^2}{\partial x_1 \partial x_2} f(\boldsymbol{x}_0) & \dots & \frac{\partial^2}{\partial x_1 \partial x_m} f(\boldsymbol{x}_0) \\ \frac{\partial^2}{\partial x_2 \partial x_1}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

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

Convexity – Convex Functions

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

$$f(\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{y}) \leqslant \alpha f(\boldsymbol{x}) + (1 - \alpha) f(\boldsymbol{y}) \tag{56}$$



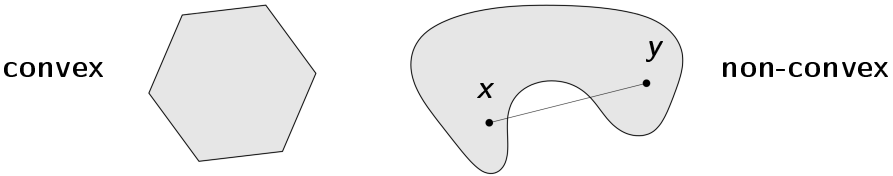
- Examples are (affine-)linear functions $f(\boldsymbol{x}) = \boldsymbol{a}^\intercal \boldsymbol{x} + b$ and quadratic functions $f(\boldsymbol{x}) = \boldsymbol{x}^\intercal \boldsymbol{A} \boldsymbol{x} + \boldsymbol{b}^\intercal \boldsymbol{x} + c$

Convexity – Convex Sets

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

$$\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{y} \in C \tag{57}$$

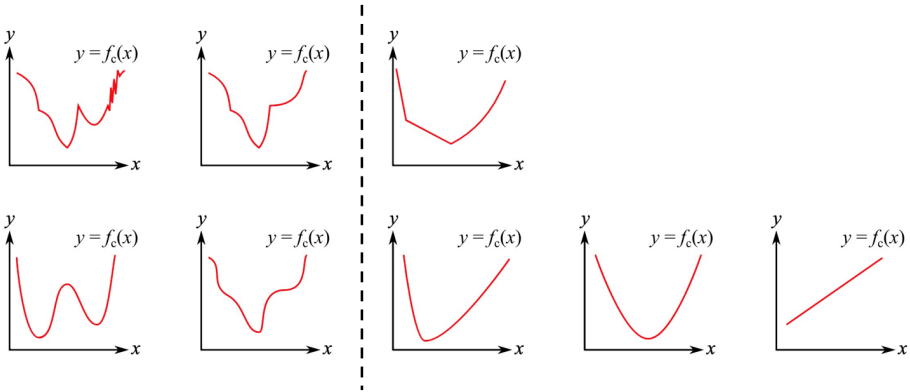
- This is the equation of the line segment between $\boldsymbol{x}$ and $\boldsymbol{y}$



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 $\begin{cases} f : \mathbb{R}^m \to \mathbb{R} \\ x \mapsto f(x) \end{cases}$ be given
- How do we find minimums and maximums of f?
- **Necessary condition:** A point $x_0 \in \mathbb{R}^m$ is a potential candidate if

$$\nabla f(x_0) = \mathbf{0} \tag{58}$$

- Then check if the Hessian matrix $\nabla^2 f(x_0)$ is positive definite ($\Longrightarrow$ minimum) or negative definite ($\Longrightarrow$ maximum)

Constrained Optimization

Formalization:

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

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

$$g(\boldsymbol{\theta}) \geqslant 0 \quad \longleftarrow \text{inequality constraints}$$

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

Constrained Optimization (Ctd.)

Answer:

$$\min_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta}) \quad \longleftarrow \textbf{convex} \text{ function}$$

$$\text{subject to } f(\boldsymbol{\theta}) = 0 \quad \longleftarrow \textbf{linear} \text{ function}$$

$$g(\boldsymbol{\theta}) \geqslant 0 \quad \longleftarrow \textbf{convex} \text{ set}$$

Constrained Optimization

How to solve this optimization problem?

$$\mathcal{J}(x, y) := 2y + x$$

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

Solution:

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

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 λ

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 ❷: Set derivatives to zero

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

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

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

Step ❸: Substitute

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

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

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

Exercise

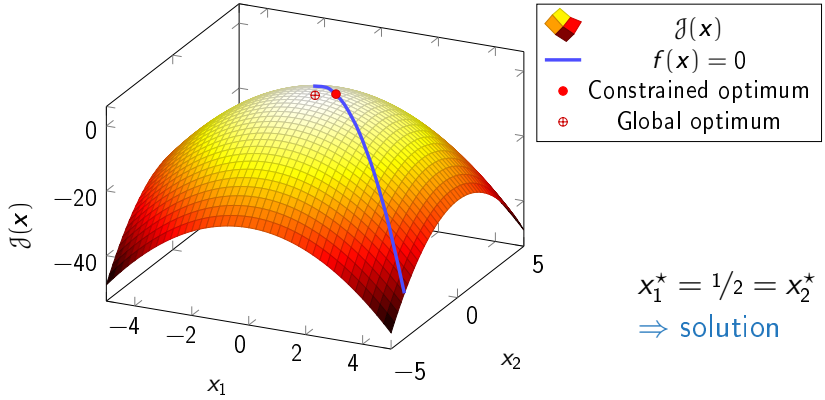
Optimize the following function:

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

subject to:

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

Exercise Solution



$x_1^\star = 1/2 = x_2^\star$

$\Rightarrow$ 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:

$$\boldsymbol{\theta}_{\text{new}} \longleftarrow \boldsymbol{\theta}_{\text{old}} + \Delta\boldsymbol{\theta} \tag{59}$$

- 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

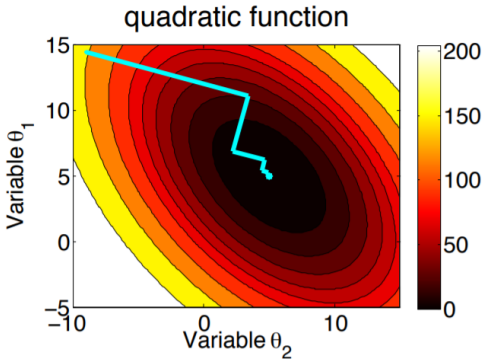
- Gradient descent is a basic, but most commonly used algorithm im ML
- **General idea:** Go into the direction of the **steepest descent**
- The gradient points into the direction of steepest ascent, therefore subtract the gradient from the current estimate:

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

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

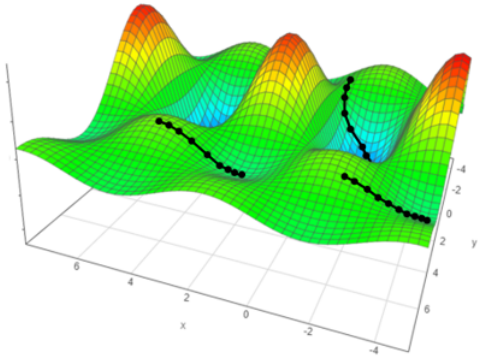
Gradient Descent (Ctd.)

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



Gradient Descent (Ctd.)

Initialization also matters...



Taylor Expansion (one-dimensional)

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

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

- Alternatively we can write:

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

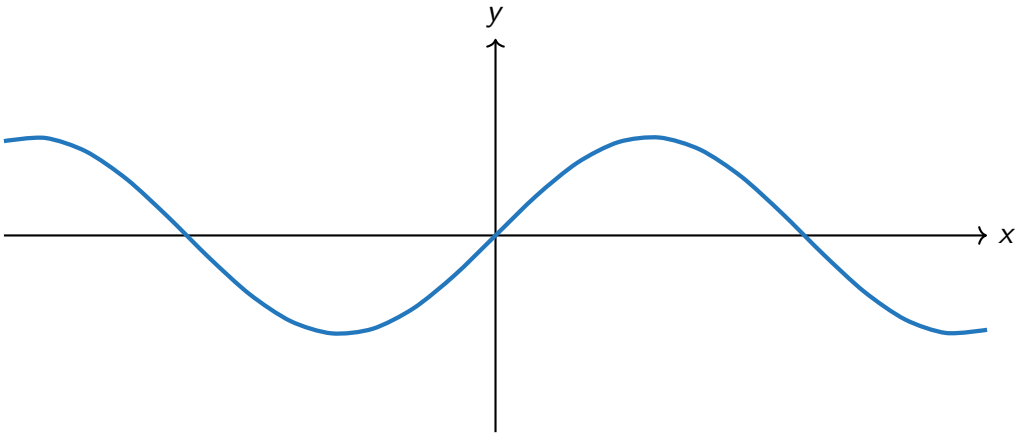
Example: Taylor Polynomial for $\sin(x)$

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

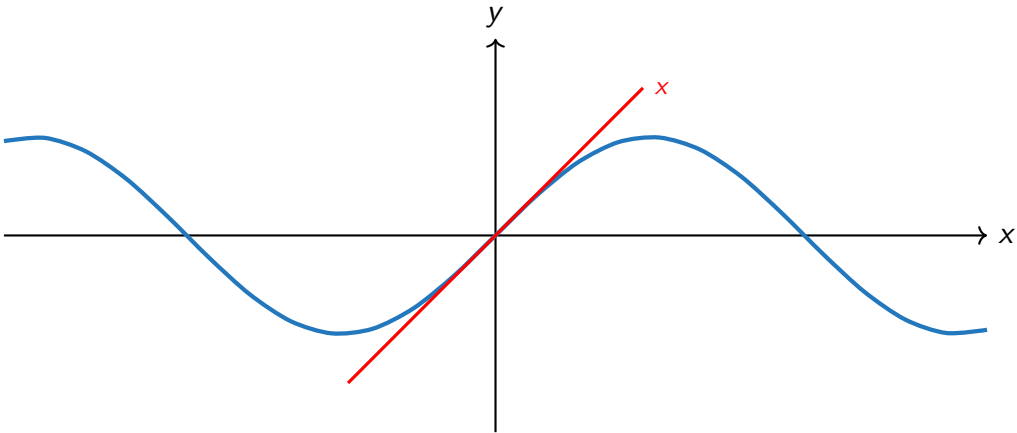
$$\sin(x) \approx \sin(0) + x\cos(0) - \frac{x^2}{2}\sin(0) - \frac{x^3}{6}\cos(0) + \frac{x^4}{24}\sin(0) + \frac{x^5}{120}\cos(0)$$
$$= x - \frac{x^3}{6} + \frac{x^5}{120}$$

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

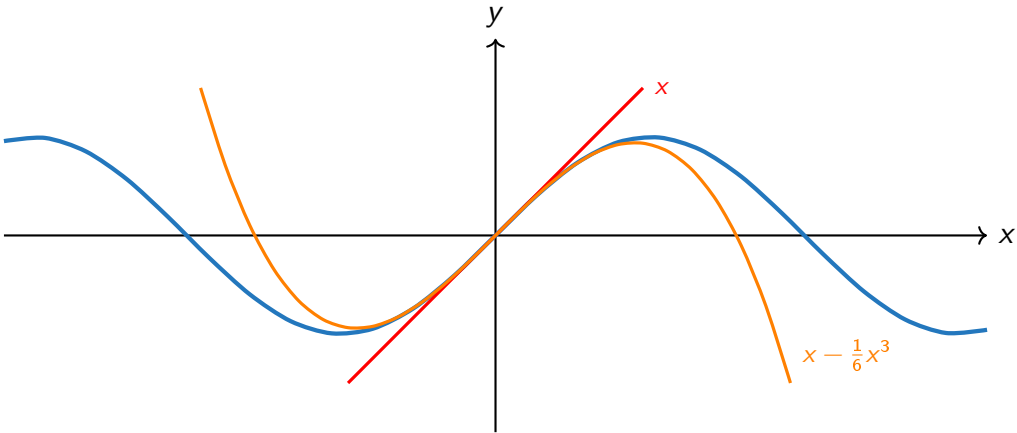
Example: Taylor Polynomial for $\sin(x)$ (Ctd.)



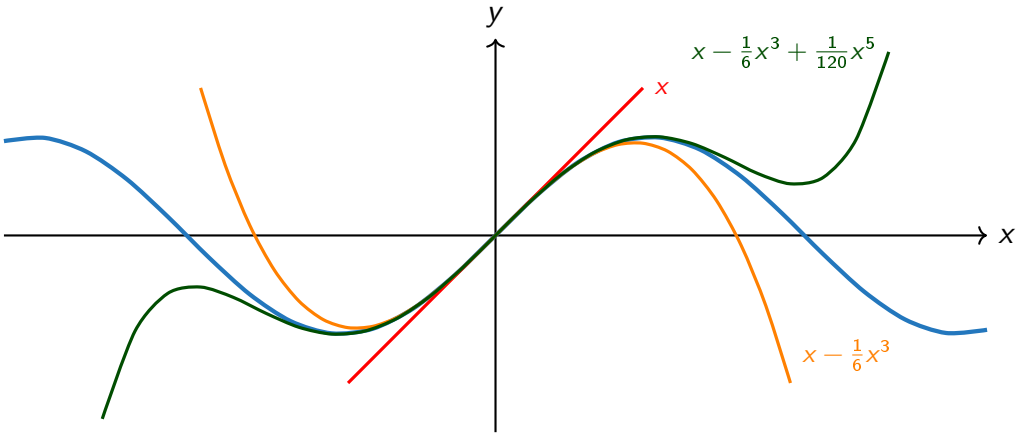
Example: Taylor Polynomial for $\sin(x)$ (Ctd.)



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



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



Taylor Expansion in higher Dimensions

- We can generalize this concept for multivariate functions
- Let $f : \mathbb{R}^m \to \mathbb{R}$ be 2-times differentiable at $x_0 \in \mathbb{R}$. We can approximate f using a **Taylor polynomial** of degree 2 given by

$$f(x + \Delta x) := f(x) + g(x)^\intercal \Delta x + \frac{1}{2} \Delta x^\intercal H(x) \Delta x \tag{63}$$

g is the gradient of f and H the Hessian matrix of f

(A Taylor polynomial of degree 2 is sufficient for our use cases. Of course we could consider polynomials of degree $n > 2$ as well.)

Newton's Method

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

$$\mathcal{J}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) := \mathcal{J}(\boldsymbol{\theta}) + \boldsymbol{g}(\boldsymbol{\theta})^\intercal \Delta\boldsymbol{\theta} + \frac{1}{2}\Delta\boldsymbol{\theta}^\intercal \boldsymbol{H}(\boldsymbol{\theta})\Delta\boldsymbol{\theta} \tag{64}$$

  where $\boldsymbol{g}(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}}\mathcal{J}(\boldsymbol{\theta})$ and $\boldsymbol{H}(\boldsymbol{\theta}) := \nabla^2_{\boldsymbol{\theta}}\mathcal{J}(\boldsymbol{\theta})$
- We differentiate (64) with respect to $\Delta\boldsymbol{\theta}$ and set the derivative to zero:

$$\nabla_{\Delta\boldsymbol{\theta}}\mathcal{J}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{\theta}) + \boldsymbol{H}(\boldsymbol{\theta})\Delta\boldsymbol{\theta} \stackrel{!}{=} \boldsymbol{0} \tag{65}$$

Newton's Method (Ctd.)

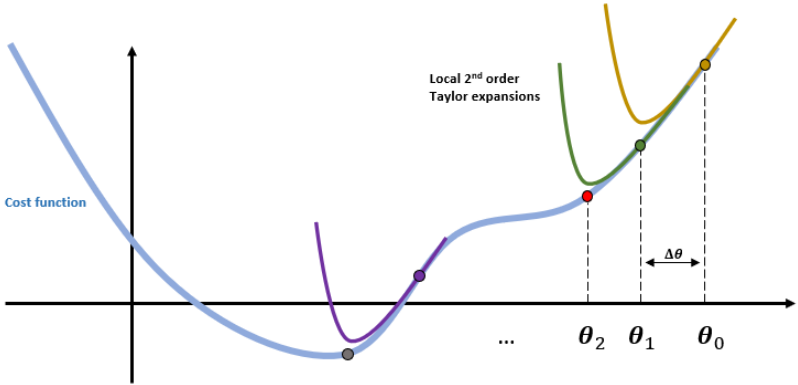
- We obtain: $\Delta\boldsymbol{\theta} = -\boldsymbol{H}^{-1}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})$
- The update rule then takes the form

$$\Delta\boldsymbol{\theta} \longleftarrow -\boldsymbol{H}^{-1}(\boldsymbol{\theta}_{\text{old}})\boldsymbol{g}(\boldsymbol{\theta}_{\text{old}}) \tag{66}$$

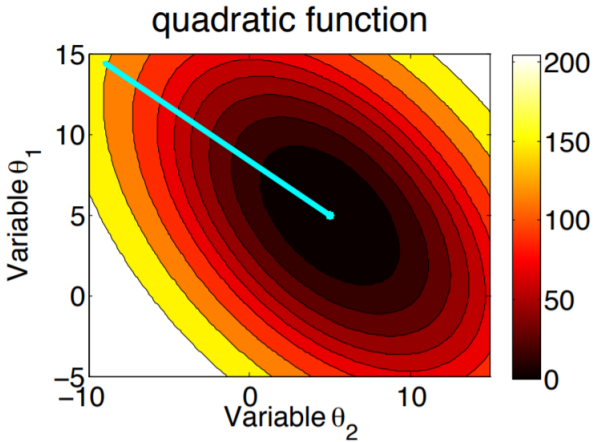
$$\boldsymbol{\theta}_{\text{new}} \longleftarrow \boldsymbol{\theta}_{\text{old}} + \Delta\boldsymbol{\theta} \tag{67}$$

- The learning rate α 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** $\sim \mathcal{O}(n^3)$

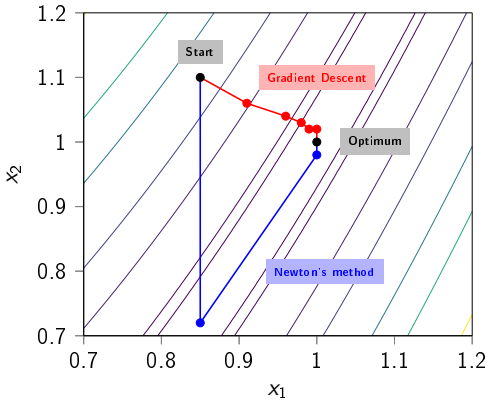
Newton's Method (Ctd.)



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
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

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

Do you have any questions?