Mathematics of Data Science - Summary

Jannis Zeller

November 29, 2021

Contents

1 Matrix Decomposition and Optimization

1.1 Matrices Introduction

Basic Notations

- Matrices with m rows and n columns will be denoted as $A \in \mathbb{K}^{m\times n}$, where typically $\mathbb{K} \in \{\mathbb{C}, \mathbb{R}\}$.
- A single entry of the matrix A is therefore denoted as A_{ij}.
- Complete rows and columns (as vectors) will be denoted as $A^{(i)}$ for rows and $A_{(j)}$ for columns.
- The $n \times n$-identity matrix is denoted as I_n.
- The matrixproduct between $A \in \mathbb{K}^{l\times m}$ and $B \in \mathbb{K}^{m\times n}$ is a matrix $C \in \mathbb{K}^{l\times n}$ with the entries

$$C_{ik} = \sum_{j=1}^{m} A_{ij} B_{jk} \,.$$

- A matrix-vector-product is therefore denoted as:

$$(Av)_k = \sum_{i=1}^{m} A_{km} v_m$$

- An $m \times n$ matrix can be constructed by the "matrix"-product of an $\mathbb{R}^m$-vector a with a transposed $\mathbb{R}^n$-vector b:

$$A = ab^T = \begin{pmatrix} a_1b_1 & a_1b_2 & \dots & a_1b_n \\ a_2b_1 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_mb_1 & \dots & \dots & a_mb_n \end{pmatrix} \qquad \Rightarrow \qquad A_{(j)} = b_j a \quad \& \quad A^{(i)} = a_i b^T$$

Trasopsed and Adjoint Matrices

- The transopsed A^T of a matrix A has the entries $(A^T)_{ij} = A_{ji}$.
- The adjoint or Hermetian transposed A^* of a matrix A has the entries $(A^*)_{ij} = \bar{A}_{ji}$, where $\bar{a}$ denotes the complex conjugated of a comlpex number $a \in \mathbb{C}$.
- Every quadratic Matrix $A \in \mathbb{C}^{n\times n}$ can be decomposed into a Hermitian and an anti-Hermitian part via:

$$A = \frac{1}{2} \underbrace{(A + A^*)}_{\text{Hermitian}} + \frac{1}{2} \underbrace{(A - A^*)}_{\text{anti-Hermitian}}$$

- In the real case, Hermetian matrices are called symmetric matrices and all the properties still hold replacing $\cdot^*$ with $\cdot^T$.

Scalar Product and basis

- The (complex) scalar product for $x, y \in \mathbb{C}^m$ is defined as

$$\langle x, y\rangle = \sum_{i=1}^{m} x_i \bar{y}_i \,,$$

  which is called the Euclidiean scalar product in the case $x, y \in \mathbb{R}^m$.

- x, y are called orthogonal if $\langle x, y\rangle = 0$.

- A basis $x^1, ..., x^m$ of $\mathbb{K}^m$ is called orthonormal if

$$\langle x^i, y^j\rangle = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} .$$

Semi Symmetric inner Product

- Given a $\mathbb{C}$-vector space V a mapping $\langle .\,, .\rangle : V \times V \to \mathbb{C}$ is called a semi symmetric inner product if the following properties hold:

  (i) Positive Definiteness: For all $v \in V$, $\langle v,\, v\rangle \geq 0$.

  (ii) Definiteness: If $\langle v,\, v\rangle = 0$, then $v = 0$.

  (iii) Linearity (in the first argument): For all $\lambda \in \mathbb{C}$ and $u, v, w \in V$

$$\langle v + \lambda u,\, w\rangle = \langle v,\, w\rangle + \lambda\langle u,\, w\rangle \,.$$

  (iv) (Semi) Symmetry: For all $u, v \in V$: $\langle u,\, v\rangle = \overline{\langle v,\, u\rangle}$.

- (iii) and (iv) together yield: $\langle v,\, \lambda u + w\rangle = \bar{\lambda}\langle v,\, u\rangle + \langle v,\, w\rangle$.

- If we constrain V to $V \in \mathbb{R}$, then this inner product is truly symmetric and thus linear in both arguments.

Orthogonal und unitary matrices

- $A \in \mathbb{R}^{m\times m}$ is called orthogonal if $A^T A = AA^T = I_m$.

- $A \in \mathbb{C}^{m\times m}$ is called unitary if $A^* A = AA^* = I_m$.

Range and rank of a matrix

- The range of a matrix $A \in \mathbb{K}^{m\times n}$ is defined as

$$\mathrm{ran}(A) = \{Ax : x \in \mathbb{K}^m\} = \mathrm{span}\{A_{(j)} : j = 1, ..., n\} \subset \mathbb{K}^m$$

- The rank of a matrix $A \in \mathbb{K}^{m\times n}$ is defined as: $\mathrm{rank}(A) = \dim(\mathrm{ran}A) = \dim(\mathrm{ran}A^T)$.

- A matrix $A \in \mathbb{K}^{m\times n}$ has full rank if $\mathrm{rank}(A) = \min\{m, n\}$.

- The rank of a matrix can be described as the number of linear independent columns or rows in the matrix.

Eigenvalues and eigenvectors

- For $A \in \mathbb{K}^{m\times n}$, $\lambda \in \mathbb{K}$ is called an eigenvalue of A with corresponding eigenvector $v \in \mathbb{K}^m \setminus \{0\}$ if $Av = \lambda v$.

- Eigenvalues are the roots of the characteristic polynomial $\chi_A(\lambda) = \det(A - \lambda I)$.

- If $A \in \mathbb{K}^{m\times m}$ is Hermitian ($A = A^*$), then all eigenvalues are real and there exists an orthonormal basis $v_1, ..., v_n \in \mathbb{K}^m$ of eigenvectors. With $V = (v_1|...|v_n) \in \mathbb{K}^{m\times m}$ it holds

$$A = VDV^* = \sum_{j=1}^{m} \lambda_j v_j v_j^* \qquad \text{with} \qquad D = \mathrm{diag}(\lambda_1, ..., \lambda_m)\,.$$

Trace

- The trace $\mathrm{tr} : \mathbb{C}^{m\times m} \to \mathbb{C}$ is defined as the sum of the entries of the diagonal:

$$\mathrm{tr}(A) = \sum_{i=1}^{m} A_{ii}\,.$$

- The trace has the properties:

  (i) Cyclicity: $\mathrm{tr}(AB) = \mathrm{tr}(BA)$

  (ii) Invariance under unitary conjugation:

$$\mathrm{tr}(UAU^*) = \mathrm{tr}(A)$$

  (iii) Trace and eigenvalues: If $\lambda_1, ..., \lambda_m$ are the eigenvalues of $A \in \mathbb{C}^{m\times m}$, the trace is calculated by

$$\mathrm{tr}(A) = \sum_{i=1}^{m} \lambda_i$$

- The **Frobenius scalar product** of two matrices $A, B \in \mathbb{C}^{m \times n}$ is defined as

$$\langle A, B \rangle_F = \text{tr}(AB^*) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} \bar{B}_{ij}$$

- The **Frobenius norm / Hilbert-Schmidt-norm** is therefore defined as

$$||A||_F = \sqrt{\langle A, A \rangle_F} = ... = \sqrt{\sum_{i=1}^{m} \lambda_i(A^*A)},$$

  where $\lambda_i(A^*A)$ are the eigenvalues of A^*A. (See below for confirmation that it is well defined and indeed a norm).

Norms

- For a $\mathbb{K}$-vectorspace V a norm $||.|| : V \to \mathbb{R}_+ = \{x \in \mathbb{R}, x \geq 0\}$ is a function satisfying

  (i) $||v|| = 0$ if and only if $v = 0$

  (ii) $||\lambda v|| = |\lambda| \, ||v||$ for all $v \in V$ and $\lambda \in \mathbb{K}$.

  (iii) Tranglie inequality: $||v + w|| \leq ||x|| + ||w||$ for all $v, w \in V$.

- If (i) is weakened to $||v|| = 0$ if $v = 0$ (thus $||v|| = 0$ does not imply $\Rightarrow v = 0$), $||.||$ is calles a semi-norm.

- The probably most important norm in linear algebra is the euclidian norm

$$||x||_2 = \sqrt{\langle x, x \rangle}\,.$$

- Indeed every inner product induces a corresponding norm.

Definite matrices

- A hermitian matrix $A = A^* \in \mathbb{C}^{m \times m}$ is called positive (semi) definite if

$$x^*Ax > 0 \;\; (\geq 0) \;\; \text{for all } x \in \mathbb{C}^m \setminus \{0\},$$

  and negative (semi) definite respectivley.

- A hermitian matrix $A = A^* \in \mathbb{C}^{m \times m}$ is positive (semi) definite if and only if $\lambda_i(A) > 0 \;\; (\geq 0)$ for all $i = 1, ..., m$.

- Note that in case of the Frobenius norm A^*A is obviously hermitian and indeed positive semidefinite[1], thus the Frobenius norm is well defined.

[1] Note: $x^*A^*Ax = \langle Ax, Ax \rangle = ||Ax||^2 \geq 0$.

Operator norm

- For $X = (\mathbb{R}^n, ||.||_X)$, $Y = (\mathbb{R}^m, ||.||_Y)$ the operator norm of $A \in \mathbb{R}^{m\times n}$ (where $A : X \to Y$) is defined as

$$||A||_{X\to Y} = \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{||Ax||_Y}{||x||_X} = \sup_{\substack{x\in\mathbb{R}^n \\ ||x||_X=1}} ||Ax||_Y$$

- An concrete example for an operator norm is the spectral norm

$$||A||_{l^2\to l^2} = \max_j \sqrt{\lambda(A^*A)}\,,$$

which can be shown to be unitary invariant (because of properties of the determinant and such the characteristic polynomal).

Cauchy-Schwarz inequality

Let V be a $\mathbb{C}$-vectorspace and $a, b \in V$. Additionally let $\langle .\,, .\rangle : V \times V \to \mathbb{C}$ be an semi symmetric inner product with corresponding norm $||.||$. Then the following equation holds:

$$\boxed{|\langle a,\, b\rangle| \leq ||a|| \cdot ||b||}$$

Hölder inequality

Let $u, v \in \mathbb{R}^n$ and $p, q \in \mathbb{R}$ such that $1/p + 1/q = 1$. Then

$$\boxed{\sum_{i=1}^n |u_i v_i| \leq \left(\sum_{i=1}^n |u_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^n |v_i|^q\right)^{\frac{1}{q}}}$$

Multidimensioanl Taylors theorem:

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Then for some $z \in \{\lambda x+(1-\lambda)y | \lambda \in [0, 1]\}$ one can evaluate $f(y)$ via:

$$f(y) = f(x) + \langle\nabla f(x), y - x\rangle + \frac{1}{2}(y - x)^T D^2 f(z)(y - x),$$

which is the taylor formula with remainder for $n = 2$. $D^2 f$ denotes the Hessian Matrix as declared below.

1.2 Singular value decomposition

Definition of the SVD

- The SVD of a matrix $A \in \mathbb{R}^{m\times n}$ consists of the factorisation of A into three matrices:

$$\boxed{A = U\Sigma V^T\,,}$$

  where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal and $\Sigma \in \mathbb{R}^{m\times n}$ is diagonal with $\Sigma = \operatorname{diag}(\sigma_1, ..., \sigma_k),\ k = \min\{m, n\}$, with $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_k$.

- In contrast to the eigenvalue decomposition (EVD), Σ is not necessarily quadratic, such that with $\Sigma' \in \mathbb{R}^{k\times k}$ defined, Σ takes the form ($\mathbf{0}$ is matrix of zeros in needed size):

$$\Sigma = \begin{pmatrix}\Sigma' & \mathbf{0}\end{pmatrix}\,,\ m \leq n\,, \qquad \Sigma = \begin{pmatrix}\Sigma' \\ \mathbf{0}\end{pmatrix}\,,\ n < m \quad \text{and} \quad \Sigma = \begin{pmatrix}\Sigma' & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{pmatrix}\,. \tag{1.1}$$

Left singular Matrix and Linear Subspaces for data

- I. e. Approach: Given „data points“ in (the columns of) a matrix A, one would like to find the k-dimensional subspace, that best finds the data points $A_{(1)}, ..., A_{(n)} \in \mathbb{R}^m$, i. e. in the case of $k = 1$ or $k = 2$ the best line or plane.

- I. e. $m = 2$ and $k = 1$ via Pythagoras the task is derived, to find u such that:

$$\max_{||u||_2=1} \sum_{j=1}^{n} |\langle A_{(j)},\, u\rangle|^2 = \max_{||u||_2=1} ||A^T u||_2^2 = \max_{||u||_2=1} ||Au||_2^2$$

- There are other possible choices than to minimize the euclidean distance, which are computationally more complex but might have advantages in some cases.

- **Left singular vectors and values**: For $k < m \in \mathbb{N}$ the left singular values are given by:

$$\boxed{\begin{aligned} \sigma_1 &= \max_{||u||_2=1} ||Au||_2 = ||A||_{l^2\to l^2}\,, \\ \sigma_r &= \max_{\substack{u\perp u_1,...,u_{r-1} \\ ||u||_2=1}} ||Au||_2\,, \end{aligned}}$$

  and stop once $\sigma_{r+1} = 0$ or $r = m$, with the corresponding left singular vectors as:

$$\boxed{\begin{aligned} u_1 &= \operatorname*{argmax}_{||u||_2=1} ||Au||_2 = ||A||_{l^2\to l^2}\,, \\ u_r &= \operatorname*{argmax}_{\substack{u\perp u_1,...,u_{r-1} \\ ||u||_2=1}} ||Au||_2\,, \end{aligned}}$$

- For a matrix $A \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1, ..., \sigma_r$ it holds

$$||A||_F = \sqrt{\sum_{j=1}^{r} \sigma_r^2}\,.$$

- It can be shown, that this is in a sense actually the best possible k-dimensional subspace, by using orthogonal projections.

Orthogonal projection

- Let $W \subset \mathbb{R}^m$ be a subspace of dimension k with orthonormal basis (ONB) $w_1, ..., w_k \in \mathbb{R}^m$ and let $u \in \mathbb{R}^m$.

  a) The minimizer $\hat{w}$ of $\min_{w\in W} ||u - w||_2$ exists, is unique and given by

  $$\hat{w} = \sum_{j=1}^{k} \langle u,\, w_j\rangle w_j\,.$$

  Therefore the **orthogonal projection**

  $$\boxed{P_w : \mathbb{R}^m \to W,\ u \mapsto \underset{w\in W}{\operatorname{argmin}} ||u - w||_2}$$

  is linear and $P_w u$ is given by

  $$\boxed{P_w u = \sum_{j=1}^{k} \langle u,\, w_j\rangle w_j\,.}$$

  b) The difference vector $u - \hat{w}$ and $\hat{w}$ are orthogonal.

  c) It holds:

  $$||\hat{w}||_2^2 = ||P_w u||_2^2 = \sum_{j=1}^{k} \langle u,\, w_j\rangle^2 \quad \text{and} \quad ||u - \hat{w}||_2^2 = ||u||_2^2 - \sum_{j=1}^{k} \langle u,\, w_j\rangle^2$$

- Using this definition one can show the optimality of the decomposition above:

Left singular subspace minimizes distance: Let $A \in \mathbb{R}^{m\times n}$ with left singular vectors $u_1, ..., u_r$. For $k = 1, ..., r$ define $U_k = \operatorname{span}\{u_1, ..., u_k\}$. Then U_k minimizes

$$d(A, V)^2 = \sum_{j=1}^{n} d(A_{(j)}, V)^2 = \sum_{j=1}^{n} ||A_{(j)} - P_V A_{(j)}||_2^2$$

over all k-dimensional subspaces V, i. e. U_k is the best k-dimensional approximation to the columns of A.

Right singular vectors

- Given a matrix $A \in \mathbb{R}^{m\times n}$ with left singular values and vectors denoted as above one can define the **right singular vectors** as

$$\boxed{v_j = \frac{1}{\sigma_j} A^T u_j \; , \; j = 1, ..., r \; .}$$

  These are **orthonormal**.

- In this case it holds $\operatorname{ran}(A) = \operatorname{span}\{v_1, ..., v_r\}$ and thus $\operatorname{rank}(A) = r$.

Formulation and computation of the SVD

- **Reduced singular value decomposition:** Given a matrix $A \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1, ..., \sigma_r$, left singular vectors $u_1, ..., u_r$ and right singular vectors $v_1, ..., v_r$ one can write:

$$\boxed{A = \sum_{j=1}^{r} \sigma_j u_j v_j^T = U \Sigma V^T \; ,}$$

  with $U = (u_1|...|u_r) \in \mathbb{R}^{m\times r}$, $V = (v_1|...|v_r) \in \mathbb{R}^{n\times r}$ and $\Sigma = \operatorname{diag}(\sigma_1, ..., \sigma_r)$.

- In case of the reduced SVD one has to be careful with the following: Because r might be smaller than $\min\{m, n\}$, i. e. A is not of full rank, U **and** V **are not orthogonal in the following sense.** If i. e. $r < m$, then UU^T is an $m \times m$ matrix, but only of rank r such that UU^T is not the identity matrix. The only thing that always (also in case of the reduced SVD) holds is: $U^T U = I_r$, and for V respectively.

- **Singular value decomposition:** In the case $r < m$ complete $u_1, ..., u_r$ to an ONB of $\mathbb{R}^m$ and set $\tilde{U} = (u_1|...|u_m) \in \mathbb{R}^{m\times m}$. In the case $r < n$ complete $v_1, ..., v_r$ to an ONB of $\mathbb{R}^n$ and set $\tilde{V} = (v_1|...|v_m) \in \mathbb{R}^{n\times n}$. Then the singular value decomposition of A is given as

$$\boxed{A = \tilde{U} \tilde{\Sigma} \tilde{V}^T \; ,}$$

  with $\tilde{\Sigma}$ as in (1.1).

- **Obtaining singular values:** The singular values of A are the square roots of the nonzero eigenvalues of AA^T and A^TA:

$$\sigma_j = \sqrt{\lambda(A^TA)} = \sqrt{\lambda(AA^T)}\,.$$

- **Obtaining singular vectors:** The left singular vectors of A are the corresponding eigenvectors of AA^T. The right singular vectors of A are the corresponding eigenvectors of A^TA.

- **The Moore-Penrose-Pseudo-Inverse:** Given a Matrix $A \in \mathbb{R}^{m\times n}$ with reduced SVD as above, the Moore-Penrose-Pseudo-Inverse $A^+ \in \mathbb{R}^{n\times m}$ is given by

  $$A^\dagger = V\Sigma^{-1}U^T\,.$$

  If $A^TA \in \mathbb{R}^{n\times n}$ or $AA^T \in \mathbb{R}^{m\times m}$ is invertible then

  $$A^\dagger = (A^TA)^{-1}A^T \quad \text{or} \quad A^\dagger = A^T(AA^T)^{-1} \quad \text{respectively.}$$

Least Squares Problems

- Let $A \in \mathbb{R}^{m\times n}$ and $y \in \mathbb{R}^m$. Define $M := \underset{x\in\mathbb{R}^n}{\operatorname{argmin}}||Ax - y||_2$, i. e. the set of minimizers of $||Ax - y||_2$. The optimization problem

  $$\min_{x\in M} ||x||_2$$

  possesses the unique minimizer

  $$\boxed{\hat{x} = A^\dagger y.}$$

- Let $A \in \mathbb{R}^{m\times n}$, $m \geq n$, be of full rank ($\text{rank}A = n$) and $y \in \mathbb{R}^m$. Then the least squares problem $\min_{x\in\mathbb{R}^n} ||Ax - y||_2$, has the unique solution

  $$\boxed{\hat{x} = A^\dagger y.}$$

Principal Component Analysis (PCA)

- Start-point: $A \in \mathbb{R}^{m\times n}$ data matrix, with columns $A_{(j)}$ as data points.

- Idea: Fir an m-dimensional ellipsoid to data such that each axis of the ellipsoid represents a principle component.

- Process: PCA transforms the data via orthogonal basis transformation with the left singular vector matrix:

$$T_{(j)} = (A^T U)_{(j)} \quad \Leftrightarrow \quad T = (A^T U)^T = \Sigma V^T.$$

Thus the entries of of $(T_{(j)})_i, j = 1, ..., n$ have decreasing empirical variance with i.

- Replacing T with T_k (V with V_k) yields a low rank approximation. Doing this, one omits the dimensions with low variance. I. e. if $m = 3$ the data points are 3D-vectors. If these data points essentially align with a given plane, a PCA with a 2D-approximation would be legitimized. One would get two orthogonal vectors which span the observed plane.

Covariance matrices

- The covariance matrix of a mean zero random vector X is defined as $C = \mathrm{E}(XX^T)$. The corresponding empirical covariance matrix given data points as columns of a matrix A is therefore:

$$\hat{C} = \frac{1}{n} AA^T . \tag{1.2}$$

- Performing a PCA (not low rank!) on A yields:

$$TT^T = \Sigma\Sigma^T ,$$

which means, that the basis transformation with U diagonalizes the covariance matrix.

1.3 Optimization

Unconstrained optimization problems

- For a function $f : \mathbb{R}^n \to \mathbb{R}$ an **unconstrained optimization problem** is of the form $\min_{x\in\mathbb{R}^n} f(x)$ or $\max_{x\in\mathbb{R}^n} f(x)$.

- A point $x_0 \in \mathbb{R}^n$ with $f(x_0) \leq f(x)$ $(f(x_0) \geq f(x))$ for all $x \in \mathbb{R}^n$ is called a (global) **minimizer** (maximizer) and the value $f(x_0)$ is called global **minimum** (maximum).

- A point $x_0 \in \mathbb{R}^n$ is called **local minimizer** (maximizer) if there exists $\delta > 0$ such that $f(x_0) \leq f(x)$ $(f(x_0) \geq f(x))$ for all $x \in \mathbb{R}^n$ with $||x - x_0||_2 < \delta$. The value $f(x_0)$ is then called **local minimum** (maximum).

- Sometimes the optimization problems are defined over subsets $D \subset \mathbb{R}^n$.

Optimization problems in data science

- **Supervised learning:** In supervised learning one is given input/output pairs (x_i, y_i), $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$, $i = 1, ..., m$ and wants to find a (hypothesis) function h such that $h(x_i) \approx y_i$.

Typically h is chosen from a pre defined function-space $\mathcal{H}$. Then one optimizes the parameters b of h_b by minimizing a function of the kind:

$$f(b) = \sum_{i=1}^{m} l(h_b(x_i), y_i) \tag{1.3}$$

where l is a **loss function** (i. e. $l(z, y) = (z - y)^2$, $l(z, y) = |z - y|$, $l(z, y) = \max\{0, 1 - zy\}$).

- **Norm approximation:** Given a data vector $y \in \mathbb{R}^m$ and fixed vectors $a_1, ..., a_n \in \mathbb{R}^m$, one wants to fit a linear combination Ax, were $A = (a_1|...|a_n)$ that minimizes a given norm $||.||$ on $\mathbb{R}^m$. This leads to

  $$\min_{x\in\mathbb{R}^m} ||Ax - y|| \, .$$

  Sometimes this is „regularized“ by adding a term $\lambda \cdot ||x||'$ with another norm $||.||'$, which results in a tendency to keep the values of x rather small.

- **Maximum likelihood estimation:** Consider a family $p_x : \mathbb{R}^m \to \mathbb{R}$ of distribution functions on $\mathbb{R}^m$, indexed b a parameter $x \in \mathbb{R}^n$. Given data $y_1, ..., y_k$ distributed according to p_x for unknown $x \in \mathbb{R}^n$, the task is to estimate the distribution p_x, i. e. x for the data. Given data $y \in \mathbb{R}^m$, a maximum likelihood estimator is a maximiser $\hat{x}_{ML}$ of

  $$\max_{x\in\mathbb{R}^m} p_x(y) \, .$$

- **Linear measurements with i. i. d. noise:** Consider a model $y_i = \langle a_i,\, x\rangle + \xi_i,\ a_i \in \mathbb{R}^n,\ i = 1, ..., m$, where ξ_i are independent random variables distributed according to p. Then (by independence)

  $$p_x(y) = \prod_{i=1}^{m} p(y_i - \langle a_i,\, x\rangle) \, .$$

  Maximising $p_x(y)$ is furthermore equivalent to maximizing $log(p_x(y))$. In case of Gaussian distribution of the xi_i this becomes a least square problem, i. e. a norm approximation with the 2-norm.

Optimality conditions

- **Gradient** For a differentiable function $f : D \to \mathbb{R}$, $D \subset \mathbb{R}^n$ open, the gradient is defined as

  $$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x),\, ... \frac{\partial f}{\partial x_n}(x) \right)^T \, .$$

- **Hessian Matrix:** For a twice differentiable function $f : D \to \mathbb{R}$, $D \subset \mathbb{R}^n$ open, the Hessian is defined as

  $$\big[D^2 f(x)\big]_{ij} = \frac{\partial^2 f}{\partial x_i \, \partial x_j} \,.$$

  Theorem of Schwarz: If f is twice continuously differentiable then $D^2 f^T = D^2 f$, because for such function the derivatives commute.

- **Necessary condition for local optimizers:** Let $f : D \to \mathbb{R}$, $D \subset \mathbb{R}^n$ open, be a differentiable function and $x \in D$ a local minimizer or maximizer of f. Then $\nabla f(x) = 0$. A point $x \in D$ with $\nabla f(x) = 0$ is also called **critical or stationary point** of f.

- Note: If $x \in D$ is such that f is not differentiable in x, x might still be a local optimum (recall i. e. $f(x) = |x|$).

- **Saddle point:** A point $(x_0, y_0) \in D \subset (\mathbb{R}^n \times \mathbb{R}^k)$ open is called a saddle point of $f : D \to \mathbb{R}$ if f is a critical point, but neither a local minimum nor a local maximum.

- **Sufficient conditions:** Let $f : D \to \mathbb{R}$, $D \subset \mathbb{R}^n$ open, be twice continuously differentiable. A point $x_0 \in D$ with $\nabla f(x_0) = 0$ is

  (i) a local minimizer if $D^2 f$ is positive definite.

  (ii) a local maximizer if $D^2 f$ is negative definite.

  (iii) a saddle point if $D^2 f$ is indefinite.

Convexity

- **Convex sets:** A subset $D \subset \mathbb{R}^n$ is called convex if for all $x, y \in D$, the convex combination $\lambda x + (1 - \lambda) y \in D$ for all $\lambda \in [0, 1]$.

- **Convex functions:** A function $f : D \to \mathbb{R}$, $D \subset \mathbb{R}^n$ is called convex, if D is convex and for all $x, y \in D$ and $\lambda \in [0, 1]$

  $$f\Big(\lambda x + (1 - \lambda) y\Big) \le \lambda f(x) + (1 - \lambda) f(y) \,.$$

  If the inequality is strict, then the function is called strictly convex. Contrary f is called (strictly) concave if $-f$ is (strictly) convex.

- **Optimizers of convex functions:** Local optimizers of convex functions are global optimizers.

- **Conditions for convexity:** Let $f : D \to \mathbb{R}$, $D \subset \mathbb{R}^n$ open and convex.

  (i) If f is differentiable, then f is convex if and only if for all $x, y \in D$

  $$f(y) \ge f(x) + \langle \nabla f(x), \, y - x \rangle \,.$$

(*ii*) If f is twice continuously differentiable, then f is convex if and only if $D^2 f(x)$ is positive semi-definite for all $x \in D$.

Descent methods

- **Goal:** The goal of descent methods is to find at least an approximation of minimizers x of $\min_{x\in D} f(x)$ by computing a sequence $x^{(0)}, x^{(1)}, x^{(2)}, ...$ in D that converges to x, which is then called a minimizing sequence.
  Usually, the iteration starts with some initial $x_0 \in D$ and take the form $x^{(k+1)} + t^{(k)}\, \Delta x^{(k)}$, where $\Delta x^{(k)}$ is the search direction and $t^{(k)}$ is the step size or step length.

- **Descent direction:** A descent direction always fulfils

$$\boxed{\langle \nabla f(x), \Delta x \rangle < 0\,.}$$

- **Descent methods:** In case of a descent method

$$\boxed{f\left(x^{(k+1)}\right) < f\left(x^{(k)}\right)}$$

  is required except when $x^{(k)}$ is optimal. If f is convex and differentiable then $\langle \nabla f(x^{(k)}),\, y - x^{(k)} \rangle \geq 0$ would imply $f(y) \geq f(x^{(k)})$.

- **Gradient descent:** In case of the gradient descent method one chooses $\Delta x^{(k)} = -\nabla f(x^{(k)})$.

- **Determine step sizes:** There are plenty methods for determine suitable step sizes for descent methods. One of it is the **backtracking line search** algorithm. Starting with $t = 1$ one repeats updating $t = \beta t$ with $\beta \in (0, 1)$ until the stopping condition

$$f(x + t\Delta x) \leq f(x) + \alpha\, t\, \langle \nabla f(x), \Delta x \rangle, \quad \alpha \in (0, 1/2)$$

  is fulfilled. This is called “line search” because there is a graphical interpretation of the condition as a line in t. A complete descent algorithm with backtracking line search would then first set up an initial x_0, then perform backtracking to determine the t-value for the first step with backtracking and update x with the found step size t. This is iteratively executed until the stopping condition (see below) is fulfilled.

- **Stopping condition:** The stopping condition for descent methods is typically of the kind $||\nabla f(x^{(k)})||_2 < \varepsilon$ for some desired accuracy parameter $\varepsilon > 0$.

- **Stochastic gradient descent:** Gradient descent is typically used for supervised learning techniques and corresponding cost functions (like in eq. 1.3). Because calculating the gradient for large datasets can be very computationally expensive, a common variation of the standards gradient descent (also called batch gradient descent) is, to use only a random subset of the data in each step.

Constrained Optimization

- **Objective and constraints:** A general constrained optimization problem is of the form

$$\min_{x\in D} f_0(x) \quad \text{subject to} \quad \begin{cases} f_i(x) \leq 0 & i=1,...,m \\ h_i(x)=0 & i=1,...,p \end{cases}$$

  where $f_0 : D \to \mathbb{R}$, $D \subset \mathbb{R}^n$ is called the objective functions, $f_i : D \to \mathbb{R}$ are called inequality constraint functions and $h_i : D \to \mathbb{R}$ are called equality constraint functions. A point $x \in D$ is called **feasible** if all constraints are satisfied. The set of feasible points is called feasible set or **constraint set.**

- **Optimal value:** The optimal value of a constrained optimization problem is defined as:

$$p^* = \inf \left\{ f_0(x) \middle| f_i(x) \leq 0,\ i=1,...,m\,,\ h_j(x)=0, j=1,...,p \right\}$$

  where $p^* = \pm \inf$ is actually allowed. If a problem is infeasible, then $p^* = \inf$.

- **Equivalent optimization problems:** A key procedure to solve constrained optimization problems is to convert the problem into an easier solvable, for example obtaining the feasible set beforehand.

- **Convex optimization problems:** A convex optimization problem is of the form

$$\min_{x\in D} f_0(x) \quad \text{subject to} \quad \begin{cases} f_i(x) \leq 0 & i=1,...,m \\ Ax=0 & A \in \mathbb{R}^{p\times n} \end{cases}$$

  where $f_0, f_1, ..., f_m$ are convex functions on $\mathbb{R}^n$.

- As suggested by the name a local minimizer of a convex optimization problem is a global optimizer.

- **Optimality condition:** Assume $f_0 : D \to \mathbb{R}$, $D \subset \mathbb{R}^n$ open and convex, is convex and differentiable. Let X be the feasible set. Then $x \in X$ is optimal if and only if

$$\boxed{\langle \nabla f_0(x),\, y-x \rangle \geq 0 \quad \text{for all} \quad y \in X\,.}$$

The Lagrangian

- **Definition of the Lagrangian:** Consider a constraint optimization function as above. Then the Lagrangian $L : D \times R^m \times R^p \to \mathbb{R}$ is defined as

  $$\boxed{L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x)\,.}$$

  The coefficients λ_i and ν_j are called Lagrange multipliers. This can also be written with $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_m)^T$ and $\tilde{\mathbf{f}}(x) = (f_1(x), ..., f_p(x))^T$ using $\boldsymbol{\lambda}^T \tilde{\mathbf{f}}(x)$ and for the ν's and h's respectively.

- **Lagrange dual function:** The Lagrange dual function $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is defined by

  $$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu)\,.$$

- **Lagrange dual problem:** The Lagrange dual problem is given by:

  $$\max g(\lambda, \nu) \quad \text{subject to} \quad \lambda \geq 0\,.$$

  The original problem is also called **primal problem**. A maximizer (λ^*, ν^*) of the dual problem is called **dual optimal** or **optimal Lagrange multiplier**.

- **Duality:** Let $d^* = \sup\{g(\lambda, \nu) | \lambda \geq 0, \nu \in \mathbb{R}^p\}$ be the optimal value of a dual problem for some given primal problem with optimal value p^*. Then **weak duality** always holds, which means:

  $$\boxed{d^* \leq p^*\,.}$$

  Solving the dual problem thus yields a lower bound for p^*. The value $p^* - d^*$ is called **duality gap**.
  Furthermore if

  $$\boxed{d^* = p^*}$$

  we say that **strong duality** holds.

- **Critical points using the Lagrangian:** In case of only equality constraint functions can find critical points of the optimization problem by solving

  $$\boxed{\nabla_{x,\nu} L(x, \nu) = 0\,.}$$

- **Slaters theorem:** Given primal and dual problem of the discussed forms and the primal problem is convex with $D = \mathbb{R}^n$ and that there exists $x \in \mathbb{R}^n$ with

$$f_i(x) < 0,\ i = 1, ..., m \quad \text{and} \quad Ax = b\,.$$

Then strong duality holds. Additionally if the first k functions f_i, $i = 1, ..., k$ are affine[2], then we only need that x is feasible and $f_i(x) < 0$, for $i = k + 1, ..., m$.

Example for a constrained optimization Problem

- Lets consider the function $f : \mathbb{R}^n \to \mathbb{R}, x \mapsto ||x||_2^2$, which is to be minimized, with constrain $Ax = y$ for $A \in \mathbb{R}^{m \times n},\ b \in \mathbb{R}^m$.

- The Lagragian is therefore given by

$$L(x, \nu) = x^T x + \sum_{j=1}^{m} \nu_j (a_j^T x - b_j) = ||x||_2^2 + \nu^T (Ax - b)\,.$$

- The gradient and Hessian of the Lagrangian with respect to x are given by

$$\nabla_x(x, \nu)L = 2x + A^T \nu \quad \text{and} \quad D_x^2 L(x, \nu) = 2 \cot I_n.$$

  Thus L is a convex function (in x) and $x = -A^T \nu / 2$ is a global minimum.

- Therefore the Lagrange dual function is given by

$$g(\nu) = \inf_x L(x, \nu) = -\frac{1}{4} ||A^T \nu||_2^2 - \nu^T b\,.$$

- Using the definition $p^* = \inf\{||x||_2^2 \,|\, Ax = b\}$ and using the fact, that weak duality always holds one directly obtains:

$$g(\nu) \le p^* \quad \forall \ \ \nu \in \mathbb{R}^m\,.$$

- By Slaters theorem additionally strong duality holds, if the problem is feasible, meaning if $b \in \text{range}(A)$.

Optimality conditions using Lagrangians

- **Complementary slackness:** Assume strong duality holds and let x^* and (λ^*, ν^*) be primal and dual optimal. Then it holds

$$\lambda^{*T} \tilde{\mathbf{f}}(x^*) = 0\,.$$

[2] Affine functions are linear functions added with a constant: Simplified $f(x) = Ax + b$, $A \in R^{m \times n}$.

- **Karush-Kuhn-Tucker (KKT) optimality conditions:** Assume $f_0, f_1, ..., f_m$ and $h_1, ..., h_p$ are differentiable. Since x^* minimizes $x \mapsto L(x, \lambda^*, \nu^*)$ and due to feasibility and complementary slackness the following 5 conditions hold:

  (i) $\nabla_x L(x^*, \lambda^*, \nu^*) = 0$

  (ii) $f_i(x^*) \leq 0, \quad i = 1, ..., m$

  (iii) $h_i(x^*) = 0, \quad i = 1, ..., p$

  (iv) $\lambda_i^* \geq 0, \quad i = 1, ..., m$

  (v) $\lambda_i^* f_i(x^*) = 0, \quad i = 1, ..., m$

  For any optimization problem with differentiable objective and constraint functions for which strong duality holds, any pair of primal and dual optimal points must satisfy the KKT conditions. Conversely, if the primal problem is convex and $x^*,\ (\lambda^*, \nu^*)$ satisfy the KKT conditions they are a primal/dual optimal pair.

2 Probability

Sets and Basic Set Operations

- Just a few important properties and formulas:

  (i) Commutative laws: $B \cup C = C \cup B$, and $B \cap C = C \cap B$

  (ii) Associative laws: $A \cup (B \cup C) = (A \cup B) \cup C$, and $A \cap (B \cap C) = (A \cap B) \cap C$

  (iii) **de Morgan laws:**

$$\left(\bigcup_{i \in I} A_i \right)^c = \bigcap_{i \in I} A_i^c \qquad \text{and} \qquad \left(\bigcap_{i \in I} A_i \right)^c = \bigcup_{i \in I} A_i^c$$

  (iv) Further properties:

$$\begin{aligned} A \setminus B &= A \cap B^c \\ (A \setminus B) \cap B &= \emptyset \\ (A \setminus B) \cup B &= A \cup B \\ A \setminus (A \setminus B) &= A \cap B \\ A \setminus (B \cup C) &= (A \setminus B) \cap (A \setminus C) \\ A \setminus (B \cap C) &= (A \setminus B) \cup (A \setminus C) \end{aligned}$$

- If $A_k,\ k \in I$, are **pairwise disjoint**, that is $A_l \cap A_k = \emptyset$ for all $l, k \in I$ with $l \neq k$, then the union of these sets will also be denoted by

$$\sum_{k\text{ß}I} A_k \, .$$

- **Cartesian product** For sets A and B, the set

$$A \times B = \{(a, b) | a \in A, b \in B\}$$

  is called Cartesian product of A and B. It is generalized to $k \geq 2$ sets via the symbol $\bigtimes_{i=1}^k A_i$. An important special case is $\mathbb{R}^k = \bigtimes_{i=1}^k \mathbb{R} = \{(x_1, ..., x_k) | x_i \in \mathbb{R},\ i \in \mathbb{N} \cap [1, k]\}$.

2.1 Foundations of Probability

Defining probability and stochastic models

- **Sigma-algebra:** Let $\Omega \neq \emptyset$. A family $\mathscr{A}$ of subsets of Ω is called σ-algebra (-filed) if

  (i) $\Omega \in \mathcal{A}$,

  (ii) if $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$ for all $A \in \mathcal{A}$, (*iii*) $\bigcup_{i=1}^{\infty} A_n \in \mathcal{A}$ for all sequences $A_1, A_2, ...$ of sets from $\mathcal{A}$.

$(\Omega, \mathcal{A})$ is called a **(measurable) space**.

- **Kolmogorov' axioms:** A map $P : \mathcal{A} \to [0,1]$ defined on a σ-algebra $\mathcal{A}$ on $\Omega \neq \emptyset$ is called probability measure of probability distribution if

  (*i*) $P(\Omega) = 1$,

  (*ii*) $P\left(\bigcup_{i=1}^{\infty} A_n\right) = \sum_{i=1}^{\infty} P(A_n)$ for all pairwise disjoint $(A_i)_{i\in\mathbb{N}} \subset \mathcal{A}$.

  $(\Omega, \mathcal{A}, P)$ is called **probability space**.

- A sigma algebra can in most discrete cases simply be defined as the power set $\mathcal{P}(\Omega)$ of Ω. In the continuous case one has to define another type of set with special features called the Borel-algebra.

- **Laplace space:** Let $n \in \mathbb{N}$ and $\omega_1, ..., \omega_n$ be different objects. Then, the probability space $(\Omega, \mathcal{A}, P)$ with $\Omega = \omega_1, ..., \omega_n$, $\mathcal{A} = \mathcal{P}(\Omega)$ and $P(A) = |A|/n$ for $A \subseteq \Omega$ is called Laplace space.

- **Properties of probability measures:** Let $(\Omega, \mathcal{A}, P)$ be a probability space and $A, B \in \mathcal{A}$. Then:

  (*i*) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

  (*ii*) $P(B \setminus A) = P(B) - P(A)$ if $A \subseteq B$

  (*iii*) $P(A^c) = 1 - P(A)$

  (*iv*) $A \subseteq B \Rightarrow P(A) \leq P(B)$

Conditional Probability

- **Definition of conditional probability:** Let $(\Omega, \mathcal{A}, P)$ be a probability space. For every $B \in \mathcal{A}$ with $P(B) > 0$

$$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad A \in \mathcal{A},}$$

  defines a probability measure $P(\cdot|B)$ on $\mathcal{A}$ which is called conditional probability measure given (the hypothesis) B. $P(A|B)$ is called probability of A given B.

- In multi-step random experiments each connection between two nodes in a **tree diagram** refers to a conditional probability. The total probability of a branch of the tree can be calculated using the following rule:

- **Law of total probability:** Let $(\Omega, \mathcal{A}, P)$ be a probability space and $A \in \mathcal{A}$, $(B_n)_n \in \mathcal{A}$, where B_n are pairwise disjoint sets with $A \subset \bigcup_{i=1}^{\infty} B_n$. Then

$$P(A) = \sum_{n=1}^{\infty} P(A \cap B_n) = \sum_{i=1}^{\infty} P(A|B_n) \cdot P(B_n)\,.$$

- **Baye's rule:** Let $(\Omega, \mathcal{A}, P)$ be a probability space and $A \in \mathcal{A}$, $(B_n)_n \in \mathcal{A}$, where B_n are pairwise disjoint sets with $A \subset \bigcup_{i=1}^{\infty} B_n$ and $P(A) > 0$. Then, for $k \in \mathbb{N}$,

$$\boxed{P(B_k|A) = \frac{P(B_k) \cdot P(A|B_k)}{\sum_{i=1}^{\infty} P(A|B_n) \cdot P(B_n)}\,.}$$

  A simple yet very common special case with using B and B^c.

- **Independence Laws:** Let $(\Omega, \mathcal{A}, P)$ be a probability space and $A, B, A_1, A_2, ... \in \mathcal{A}$. Then

  (i) A and B are called (stochastically) independent if

$$\boxed{P(A \cap B) = P(A) \cdot P(B)\,.}$$

  (ii) A_i, $i \in I$ are called (joint stochastically) independent if for each finite selection $\{i_1, ..., i_s\} \subseteq I$ we have

$$\boxed{P\left(\bigcap_{l=1}^{s} A_{i_l}\right) = \prod_{l=1}^{s} P(A_{i_l})\,.}$$

2.2 Distributions

Discrete distributions

- **Binomial distribution:** A binomial distribution $\text{bin}(n, p)$ with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$ has the PMF

$$\boxed{p_k = f(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in [0, n] \cap \mathbb{N}_0}$$

- **Bernoulli distribution:** The special case of a binomial distribution $\text{bin}(1, p)$ is called Bernoulli distribution.

- **Poisson distribution:** The PMF of a Poisson distribution $\text{po}(\lambda)$ with parameter $\lambda > 0$

is given by

$$\boxed{p_k = f(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k \in \mathbb{N}_0 \,.}$$

- **Discrete uniform distribution:** The probability measure of a Laplace sapce is a discrete uniform distribution.

Continuous distributions

- For continuous values one has to redefine some properties of probability distributions, to prevent the occurrence of infinities (imagine a Laplace space with uncountably infinite many segments).
- **Riemann probability density functions:** A Riemann-integrable function $f : \mathbb{R} \to \mathbb{R}$ with

$$\boxed{f(x) \geq 0, \quad x \in \quad \text{and} \quad \int_{-\infty}^{\infty} f(x) \, \mathrm{d}x = 1}$$

  is called Riemann probability density function (PDF).

- **Probabilities in the continuous case:** Probabilities of $(-\infty, x] \subseteq \mathbb{R}$ are defined by the integral

$$\boxed{F(x) = P((-\infty, x]) = \int_{-\infty}^{x} f(y) \, \mathrm{d}y, \quad x \in \mathbb{R} \,.}$$

  The function $F : \mathbb{R} \to [0,1]$ is called **cumulative distribution function** (CDF). The probability of an interval $(a, b] \subseteq \mathbb{R}$ is defined by

$$P((a,b]) = \int_a^b f(x) \mathrm{d}x = F(b) - F(a) \,.$$

- Note that in the continuous case for a single $x \in \mathbb{R}$ always $P(\{x\}) = 0$, resulting in $P((a,b)) = P([a,b]) = P([a,b)) = P((a,b])$.
- **Uniform distribution:** The uniform distribution $U(a,b)$ with parameters $a, b \in \mathbb{R}$, $a < b$ is defined by the PDF

$$f(x) = \frac{1}{b-a} \mathbf{1}_{[a,b]}(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases} .$$

- **Exponential distribution**: The exponential distribution $\mathrm{Exp}(\lambda)$ with parameter $\lambda > 0$ is defined by the PDF

$$f(x) = \lambda e^{-\lambda x} \mathbf{1}_{(0,\infty)}(x) \,.$$

- **Weibull distribution:** The Weibull distribution $\mathrm{Wei}(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ is defined by the PDF

  $$f(x) = \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}} \cdot \mathbf{1}_{(0,\infty)}(x)\,,$$

  which results in the CDF

  $$F(x) = \left(1 - e^{-\alpha x^{\beta}}\right) \cdot \mathbf{1}_{(0,\infty)}(x)\,.$$

- **Power distribution:** Power distributions for a parameter $\alpha > 0$ are defined by the CDF

  $$F(x) = x^{\alpha} \cdot \mathbf{1}_{[0,1)}\,.$$

- **Normal distribution / Gaussian distribution:** The normal distribution / Gaussian distribution $N(\mu, \sigma^2)$ with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ is defined by the PDF

  $$\boxed{\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.}$$

  $N(0,1)$ is also called the **standard normal distribution**, for which the PDF is denoted by φ and the CDF is denoted by Φ.

- **Properties of the normal distribution's CDF:** The CDF of the normal distribution obeys the following properties:

  - The CDF Φ_{μ,σ^2} of $N(\mu, \sigma^2)$ is related to the standard normal distribution's CDF Φ via

    $$\Phi_{\mu,\sigma^2}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R}\,.$$

  - $\Phi(x) = 1 - \Phi(-x),\ x \in \mathbb{R}.$
  - Φ has only an integral representation.

- **Other commonly used continuous distributions** are i. e. the gamma, Erlang, χ^2 and beta distribution. Related important distributions are the t-distribution and the F-distribution.

- **Support:** Let f be a PDF with corresponding CDF F. The set $\{x \in \mathbb{R} | f(x) > 0\}$ is called support of the probability distribution P (or of f or of the CDF F). It is therefore denoted as $\mathrm{supp}(P),\ \mathrm{supp}(f)$ of $\mathrm{supp}(F)$.

- **Location-scale transformation:** Let F be a CDF. The location-scale family of a distribution with standard CDF F is defined via

$$F_{a,b}(x) = F\left(\frac{x-a}{b}\right), \quad x \in \mathbb{R},$$

with parameters $b > 0$ and $a \in \mathbb{R}$. a is called location parameter, b is called scale parameter. $\mathscr{F} = \{F_{a,b} | F_{a,b}(x) = F(\frac{x-a}{b}), x \in \mathbb{R}; a \in \mathbb{R}, b > 0\}$ is called location-scale family of distributions (with standard member F).

- **Location scale and support:** Let (α, ω) be the support of a CDF F. The support of a location scaled CDF $F_{a,b}$ is given by:

$$\operatorname{supp}(F_{a,b}) = (a + b\alpha, a + b\omega).$$

2.3 Random variables and their distribution

Construction of random variables

- **Measurable maps:** Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(\Omega', \mathcal{A}')$ be a (measurable) space. A map $X : \Omega \to \Omega'$ is called measurable if

$$X^{-1}(A') = \{\omega \in \Omega | X(\omega) \in A'\} \in \mathcal{A} \quad \text{for each } A' \in \mathcal{A}'.$$

The definition of random variable can be extended to more general images.

- **Borel's σ-algebra:** Let $n \in \mathbb{N}$, $(\Omega, \mathcal{A}, P)$ be a probability space and $(\mathbb{R}, \mathscr{B})$ be a measurable space where $\mathscr{B}$ denotes Borel's σ-algebra.

  (i) A measurable map $X : \Omega \to \mathbb{R}$ is called random variable.

  (ii) If $X_1, ..., X_n$ are random variables $X = (X_1, ..., X_n)$ is called random vector.

- **Distribution of a random veraible / vector:** A random variable / vector X defines a probability distribution P^X via

$$P^X(A) = P\Big(\{\omega \in \Omega | X(\omega) \in A\}\Big) = P(X \in A), \quad A \in \mathscr{B} \text{ (or } A \in \mathscr{B}^n).$$

P^X is called (probability) distribution of X (notation $X \sim P^X$).

- **CDF of random variables:** Let $X, X_1, ..., X_n$ be random variables defined on a probability space $(\Omega, \mathcal{A}, P)$. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random vector. Then,

  (i) the CDF of X is defined by

$$F^X(t) = P(X \leq t) = P\Big(\{\omega \in \Omega | X(\omega) \leq t\}\Big), \quad t \in \mathbb{R}.$$

(ii) the joint CDF of $X_1, ..., X_n$ is defined by

$$\begin{aligned} F^{\mathbf{X}}(\mathbf{t}) &= P(X_1 \leq t_1, ..., X_n \leq t_n) \\ &= P\Big(\{\omega \in \Omega | X_1(\omega) \leq t_1,\ \dots,\ X_n(\omega) \leq t_n\}\Big), \quad \mathbf{t} \in \mathbb{R}^n\,. \end{aligned}$$

(iii) X_1 and X_2 are called identically distributed when $F^{X_1}(t) = F^{X_2}(t)$ for every $t \in \mathbb{R}$. For short we write $X_1 \overset{d}{=} X_2$.

- Note that $X_1 \overset{d}{=} X_2 \quad \Rightarrow \quad P(X_1 \in A) = P(X_2 \in A)$ for every Borel set $A \in \mathscr{B}$.

- **Independence of random variables:** Let I be a set and $X_i, i \in I$, be random variables. Then $X_i, i \in I$ are called independent if

$$P(X_i \in A_i,\ i \in J) = \prod_{j \in J} P(X_i \in A_i) \quad \forall\ J \subseteq I,\ |J| < \infty\ \ \forall\ A_i \in \mathscr{B},\ i \in J\,.$$

Roughly speaking: Arbitrary probablilties of combinations of X_i's are decomosable into the products of probabilities.

- **Properties of independent random variables:**

(i) If $X_i,\ i \in I$ are independent and $g_i,\ i \in I$, are given functions, then $g_i(X_i),\ i \in I$, are independent as well.

(ii) $X_1, ..., X_n$ are independent if

$$\boxed{F^{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} F^{X_i}(x_i) \quad \forall\ \mathbf{x} \in \mathbb{R}^n.}$$

(iii) If $X_1, ..., X_n$ are **independent and identically distributet** (iid) one writes

$$X_1, ..., X_n \overset{\text{iid}}{\sim} P\,.$$

- **Quantile Function:** Let X be a random variable with CDF $F : \mathbb{R} \to [0, 1]$. Then $F^{-1} : (0, 1) \to \mathbb{R}$ defined by

$$F^{-1}(y) = \inf\{x \in \mathbb{R} | F(x) \geq y\}, \quad y \in (0, 1)\,,$$

is called quantile function (QF) of F (or X or P^X). $Q_p = F^{-1}(p),\ p \in (0, 1)$ is called the **p-th** quantile of F.

- **Properties of CDF and QF:** Let F be a CDF with QF F^{-1}. Then:

  - F and F^{-1} have the properties:

(i) $F \nearrow$, right continuous, $\lim_{x\to\infty} F(x) = 1$, $\lim_{x\to-\infty} F(x) = 0$.

(ii) $F^{-1} \nearrow$, left continuous.

(iii) $F(x) \geq y \quad \Leftrightarrow \quad x \geq F^{-1}(y)$.

(iv) F continuous, $X \sim F \quad \Rightarrow \quad F(X) \sim U(0,1)$.

(v) $Y \sim U(0,1) \quad \Rightarrow \quad F^{-1}(Y) \sim F$.

- If F is strictly increasing and continuous then Q_p is the unique solution of the equation $F(x) = p$.

- The α-Quantile of $N(0,1)$ is often denoted as u_α, for which $u_{1-\alpha} = -u_\alpha$ holds, because of the relation $\Phi(x) = 1 - \Phi(-x)$.

2.4 Discrete Random Vectors and Multivariate Probability Distributions

Multivariate discrete distributions

- **Multivatiate discrete distribution:** Let $n \in \mathbb{N}$, $T \subset \mathbb{R}^n$ be a finite or countable set and $f : \mathbb{R}^n \to [0,1]$ with

$$\sum_{(t_1,...,t_n)^T \in T} f(t_1,...,t_n) = 1, \quad f(t_1,...,t_n) = 0 \text{ if } (t_1,...,t_n)^T \notin T.$$

Then f is called a (joint) probability mass fumction (PMF) on T or $\mathbb{R}^n$. The set $\text{supp}(f) = \{(t_1,...,t_n)^T \in \mathbb{R}^n | f(t_1,...,t_n) > 0\}$ is called support of f. One can denote $\mathbf{t} = (t_1,...,t_n)^T$ and the equations become shorter.

- Let $\mathbf{X} = (X_1,...,X_n)^T$ be a random vector and f be a PMF on $\mathbb{R}^n$ with support T. Then, a multivariate probability distribution $P^{\mathbf{X}}$ on $\mathbb{R}^n$ is uniquely defined by the PMF

$$P(\mathbf{X} = \mathbf{t}) = \begin{cases} f(\mathbf{t}), & \mathbf{t} \in T \\ 0, & \mathbf{t} \notin T \end{cases}.$$

The CDF of $\mathbf{X}$ is given by

$$F^{\mathbf{X}}(\mathbf{x}) = \sum_{\substack{\mathbf{t} \in T, \\ t_i \leq x_i,\ 1 \leq i \leq n}} f^{\mathbf{X}}(\mathbf{t}), \quad \mathbf{x} \in \mathbb{R}^n.$$

Generating discrete multivariate distributions via independence

- **Distribution of vector of independent random variables:** Let $X_1,...,X_n$ be independent random variables with PMFs f^{X_i} and supports T_i, $i = 1,...,n$ respectively. Then, the probability distribution $P^{\mathbf{X}} = P^{(X_1,...,X_n)}$ of the random vector $\mathbf{X} = (X_1,...,X_n)^T$ has the PMF

$$f^{\mathbf{X}}(\mathbf{k}) = \prod_{i=1}^{n} f^{X_i}(k_i), \quad \mathbf{k} \in \bigtimes_{i=1}^{n} T_i \, .$$

Furthermore, $\text{supp}(f^{\mathbf{X}}) = \bigtimes_{i=1}^{n} T_i$.

- **Bermoulli model:** Let $X_1, ..., X_n \overset{\text{iid}}{\sim} \text{bin}(1, p)$ with $p \in [0, 1]$. Then, the discrete probability distribution $P^{\mathbf{X}}$ has the PMF

  $$f^{\mathbf{X}}(\mathbf{k}) = p^{\sum_{i=1}^{n} k_i}(1-p)^{n-\sum_{i=1}^{n} k_i}, \quad \mathbf{k} \in \{0, 1\}^n \, .$$

  Furthermore $\text{supp}(f^{\mathbf{X}}) = \{0, 1\}^n$.

Generating discrete multivariate disributions via treansformations

- **Sums of discrete random variables:** Let X and Y be independent random variables on $\mathbb{Z}$ with PMF f and g. Then:

  (i) The random vector $(X, Y)'$ has the PMF $f^{(X,Y)}(x, y) = f(x)g(y), \; x, y \in \mathbb{Z}$.

  (ii) The sum $X + Y$ has the PMF h given by

  $$\begin{aligned} h(k) = P(X+Y=k) &= \sum_{j\in\mathbb{Z}} P(X=j, Y=k-j) = \sum_{j\in\mathbb{Z}} P(X=k-j, Y=j) \\ &= \sum_{j\in\mathbb{Z}} f(j)\cdot g(k-j) = \sum_{j\in\mathbb{Z}} f(k-j)\cdot g(j), \quad k \in \mathbb{Z}\, , \end{aligned}$$

  Where the first line also holds, if X and Y are not independent. h is called **convolution of the PMF f and g**. It es denoted by $h = f \star g$.

- **Special discrete convolutions:**

  - Let $X_1, ..., X_n \overset{\text{iid}}{\sim} \text{bin}(1, p)$ with $p \in [0, 1]$. Then

    $$\sum_{i=1}^{n} X_i \sim \text{bin}(n, p)$$

  - Let $X_1, ..., X_n$ be independent random variables with $X_i \sim \text{po}(\lambda_i), \; \lambda_i > 0, \; i = 1, ..., n$. Then,

    $$\sum_{i=1}^{n} X_i \sim \text{po}\left(\sum_{i=1}^{n} \lambda_i\right) .$$

  - Let $X_1, ...X_n \overset{\text{iid}}{\sim} \text{M}(1, p_1, ..., p_m)$ (see below). Then

    $$\sum_{k=1}^{n} \sim \text{M}(n, p_1, ..., p_m) \, .$$

Generation discrete multivariate distribution via multivariate PMF

- **Multinomial distribution:** The multinomial distribution (or polynomial distribution) $\mathrm{M}(n, p_1, ...p_m)$, with $n \in \mathbb{N}$ and parameters $p_1, ..., p_m \in [0,1]$ such that $\sum_{j=1}^m p_j = 1$ is defined by the PMF

$$f(k_1, ..., k_m) = \binom{n}{k_1, ..., k_m} \prod_{j=1}^m p_j^{k_j}\,, \qquad (k_1, ..., k_m)^T \in \left\{ (i_1, ..., i_k)^T \in \mathbb{N}_0^m \Big| \sum_{i=1}^m i_j = n \right\}.$$

$\binom{n}{k_1, ..., k_m} = \frac{n!}{k_1! \cdot ... \cdot k_m!}$ is called multinomial coefficient.

- Other multivariate PMF exist as well.

Marginalization

- As a concise notation, we write for a marginal vector of $X = (X_1, ..., X_n)^T$ and non-empty subset $K = \{i_1, ..., i_k\} \subseteq \{1, ..., n\}$ with $i_1 < ... < i_k$: $\mathbf{X}_k = \{X_{i_1}, ..., X_{i_k})^T$.

- **Marginal distribution:** Let $\mathbf{X} = (X_1, ..., X_n)^T$ be a random vector. Then:

  (i) The distribution of

$$\mathbf{X}_K = (X_{i_1}, ..., X_{i_k})^T$$

  with $k(< n)$ and $1 \leq i_1 < ... < i_k \leq n$ is called k-dimensional **marginal distribution** of $P^{\mathbf{X}}$.

  (ii) The distribution of (a single) X_i is called ith marginal (distribution).

  (iii) Let $\mathscr{X}_i$ be the (countable) set of all possible values of X_i. Then:

$$P(\mathbf{X}_K = x_K) = \sum_{x_j \in \mathscr{X}_j,\, j \notin \{i_1, ..., i_k\}} P(X_1 = x_1, ..., X_n = x_n)\,.$$

- **Marginal distributions and independence:** Let $\mathbf{X} = (X_1, ..., X_n)^T$ be a random vector with independent components and marginal PMFs f^{X_j}, $1 \leq j \leq n$ as well as $1 \leq i_1 < ... < i_k \leq n$, $1 \leq k \leq n$. Then, the marginal PMF of $(X_{i_1}, ..., X_{i_k})^T$ is given by the product

$$f^{(X_{i_1}, ..., X_{i_k})}(x_{i_1}, ..., x_{i_k}) = \prod_{j=1}^k f^{X_{i_j}}(x_{i_j}) \quad \text{for all } (x_{i_1}, ..., x_{i_k}).$$

  In case of independence, the joint PMF of $\mathbf{X}$ is uniquely specified by the one-dimensional marginal distributions.

- There are some equations for simple yet important marginal distribution of the multinomial distribution.

Conditional distributions

- **Conditional PMF:** Let $\mathbf{X} = (X_1, ..., X_n)^T$ be a random vector with values in $\mathscr{H}$ and $\emptyset \neq K, L \subseteq \{1, ..., n\}$ with $K \cap L = \emptyset$. Then:

  (i) The distribution of $\mathbf{X}_K$ given $\mathbf{X}_L = \mathbf{x}_L \in \mathscr{H}_L$ is defined by the conditional PMF

  $$P(\mathbf{X}_K = \mathbf{x}_K | \mathbf{X}_L = \mathbf{x}_L) = \begin{cases} \frac{P(\mathbf{X}_{L\cup K} = \mathbf{x}_{L\cup K})}{P(\mathbf{X}_L = \mathbf{x}_L)}, & P(\mathbf{X}_L = \mathbf{x}_L) > 0 \\ P(\mathbf{X}_K = \mathbf{x}_K), & P(\mathbf{X}_L = \mathbf{x}_L) = 0 \end{cases}, \quad \mathbf{x}_K \in \mathscr{H}_K \,.$$

  (ii) The corresponding conditional distribution is denoted by $P^{\mathbf{X}_K | \mathbf{X}_L = \mathbf{x}_L}$.

- **Conditional distributions in the iid-case:** Let $\mathbf{X}$ be a random vector with iid component $X_1, ..., X_n$ and $\emptyset \neq K, L \subseteq \{1, ..., n\}$ with $K \cap L = \emptyset$. Then:

  $$P^{\mathbf{X}_k | \mathbf{X}_L = \mathbf{x}_L} = P^{\mathbf{X}_k} \quad \text{for all} \;\; \mathbf{x}_L \in \mathbb{R}^{|L|} \,.$$

- The Multinomial distributions has some nice properties corresponding with conditional distrbutions.

2.5 Continuous Random Vectors and Multivariate Probability Distributions

- **Multivariate PDFs:** Let $n \in \mathbb{N}$ and $f : \mathbb{R}^n \to \mathbb{R}$ be a Riemann-integrable function with

  (i) $f(\mathbf{x}) \geq 0, \;\mathbf{x} \in \mathbb{R}^n$.

  (ii) Norm:

  $$\int_{\mathbb{R}^n} f(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(x_1, ..., x_n) \, \mathrm{d}x_1, ..., \mathrm{d}x_n = 1$$

  Then f is called a probability density function (PDF) on $\mathbb{R}^n$. The set $\mathrm{supp}(f) = \{\mathbf{t} | f(t) > 0, \mathbf{t} \in \mathbb{R}^n\}$ is called support of f.

- **Multivariate CDFs:** Let $\mathbf{X} = (X_1, ..., X_n)^T$ be a random vector and f be a PDF on $\mathbb{R}^n$ with support T. Then, a multivariat distribution $P^{\mathbf{X}}$ on $\mathbb{R}^n$ is uniquely specified by the PDF f through the CDF

  $$F^{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f(x_1, ..., x_n) \, \mathrm{d}x_1, ..., \mathrm{d}x_n \,, \quad \mathbf{x} \in \mathbb{R}^n \,.$$

  Futhermore

  (i) The CDF F is continuous.

(*ii*) Probabilities are obtained by

$$P(a_i \le X_i \le b_i,\ 1 \le i \le n) = \int_{\mathbf{a}}^{\mathbf{b}} f(\mathbf{y})\,\mathrm{d}\mathbf{y} = \int_{a_1}^{b_1} ... \int_{a_n}^{b_n} f(y_1, ..., y_n)\,\mathrm{d}x_1, ..., \mathrm{d}x_n$$

for $\mathbf{a} = (a_1, ..., a_n)^T \in \mathbb{R}^n,\ \mathbf{b} = (b_1, ..., b_n)^T \in \mathbb{R}^n$ with, $a_i \le b_i,\ 1 \le i \le n.$

(*iii*) The one-dimensional intervals $[a_i, b_i],\ 1 \le i \le n$ can be replaced by half-open or open versions.

- **Multivariate Marginals:** Let $\mathbf{X} = (X_1, ..., X_n)^T$ be a random vector with PDF $f^{\mathbf{X}}$ and $\emptyset \neq K \subseteq \{1, ..., n\}$. Then the marginal PDF of $\mathbf{X}_k = (X_{i_1}, ..., X_{i_k})^T$ is obtained by integrating over all remaining components:

$$f^{\mathbf{X}_K}(\mathbf{x}_K) = \int_{\mathbb{R}^{|K^c|}} f^{\mathbf{X}}(\mathbf{x})\,\mathrm{d}\mathbf{x}_{K^c}, \quad \mathbf{x}_K \in \mathbb{R}^{|K|}\,.$$

- **Multivariate Conditional PDFs:** Let $\mathbf{X} = (X_1, ..., X_n)^T$ be a random vector with PDF $f^{\mathbf{X}}$ and $K, L \subseteq \{1, ..., n\}$ with $K \cap L = \emptyset$. Then:

(*i*) The conditional PDF of $\mathbf{X}_K$ given $\mathbf{X}_L = \mathbf{x}_L$ is defined by

$$f^{\mathbf{X}_K|\mathbf{X}_L=\mathbf{x}_L}(\mathbf{x}_K) = \begin{cases} \frac{f^{\mathbf{X}_{L\cup K}}(\mathbf{x}_{L\cup K})}{f^{\mathbf{X}_L}(\mathbf{x}_L))}, & f^{\mathbf{X}_L}(\mathbf{x}_L) > 0 \\ f^{\mathbf{X}_K}(\mathbf{x}_K), & f^{\mathbf{X}_L}(\mathbf{x}_L) = 0 \end{cases}, \quad \mathbf{x}_K \in \mathscr{H}_K\,.$$

(*ii*) The conditional CDF of $\mathbf{X}_K$ given $\mathbf{X}_L = \mathbf{x}_L$ is defined by

$$F^{\mathbf{X}_K|\mathbf{X}_L=\mathbf{x}_L}(\mathbf{x}_K) = \int_{\mathbf{t}_K \le \mathbf{x}_K} f^{\mathbf{X}_K|\mathbf{X}_L=\mathbf{x}_L}(\mathbf{t}_K)\mathrm{d}\mathbf{t}_K\,.$$

- **Multivariate independent continuous random vectors:** Let $\mathbf{X} = (X_1, ..., X_n)^T$ be a random vector with independent components and marginal Riemann PDFs $f^{X_j},\ 1 \le j \le n$ as well as $1 \le i_1 < ... < i_m \le n,\ m \le n.$ Then, the marginal PDF of $(X_{i_1}, ..., X_{i_m})^T$ is given by the product

$$f^{X_{i_1},...,X_{i_m}}(x_{i_1}, ..., x_{i_m}) = \prod_{j=1}^{m} f^{X_{i_j}}(x_{i_j}) \quad \text{for all} \quad (x_{i_1}, ..., x_{i_m}).$$

In case of independence the joint PDF of $\mathbf{X}$ is uniquely specified by the one-dimensional marginal distributions.

- **Multivariat continuous conditional probabilties:** Let $\mathbf{X}$ be a random vector with iid components $X_1, ..., X_n$ and $\emptyset \neq K, L \subseteq \{1, ..., n\}$ with $K \cap L = \emptyset$. Then

$$P^{\mathbf{X}_K|\mathbf{X}_L=\mathbf{x}_L} = P^{\mathbf{X}_K} \quad \text{for all} \quad \mathbf{x} \in \mathbb{R}^{|L|}\,.$$

In particular, for $\mathbf{x}_L \in \mathbb{R}^{|L|}$,

$$P(\mathbf{X}_K \leq \mathbf{x}_K | \mathbf{X}_L = \mathbf{x}_L) = P(\mathbf{X}_K \leq \mathbf{x}_K), \quad \mathbf{x}_K \in \mathbb{R}^{|K|}\,.$$

2.6 Transformations of random vectors

- **Transformation theorem in** R**:** Let X be a random variable on $(\Omega, \mathcal{A}, P)$ with PDF f^X. It exists an open set $M \subseteq \mathbb{R}$ with

$$f^X(x) = 0 \quad \forall \;\; x \in M^c.$$

Further, $T : (\mathbb{R}, \mathscr{B}) \to (\mathbb{R}, \mathscr{B})$ is continuously differentiable mapping with

a) $\tilde{T} = T|_M$ is a bijective function.

b) the derivative of $\tilde{T}$ is continuous on M and satisfies

$$\Delta(x) = \frac{\partial}{\partial x}\tilde{T}(x) \neq 0 \quad \forall \;\; x \in M\,.$$

Then, the random variable $Y = \tilde{T}(X)$ has the PDF

$$\boxed{f^Y(y) = \frac{f^X\left(\tilde{T}^{-1}(y)\right)}{|\Delta(\tilde{T}^{-1}(y))|}\mathbf{1}_{\tilde{T}(M)}(y), \quad y \in \mathbb{R}\,.}$$

- **Transformation theorem in** R^p**:** Let $\mathbf{X} = (X_1, ..., X_p)^T$ be a random vector on $(\Omega, \mathcal{A}, P)$ with PDF $f^{\mathbf{X}}$. It exists an open set $M \subseteq \mathbb{R}^p$ with

$$f^{\mathbf{X}}(\mathbf{x}) = 0 \quad \forall \;\; \mathbf{x} \in M^c.$$

Further, $T : (\mathbb{R}^p, \mathscr{B}^p) \to (\mathbb{R}^p, \mathscr{B}^p)$ is continuously differentiable mapping with

a) $\tilde{T} = T|_M$ is a bijective function.

b) all partial derivatives of $\tilde{T}$ is continuous on M

c) the determinant of the Jacobian matrix $\mathcal{J}(\mathbf{x})$ satisfies

$$\Delta(\mathbf{x}) = \det(\mathcal{J}(\mathbf{x})) = \det\left(\frac{\partial \tilde{T}_i(\mathbf{x})}{\partial x_j}\right)_{1 \leq i,j \leq p} \neq 0 \quad \forall \;\; \mathbf{x} \in M\,.$$

Then, the random variable $\mathbf{Y} = \tilde{T}(\mathbf{X})$ has the PDF

$$\boxed{f^{\mathbf{Y}}(\mathbf{y}) = \frac{f^{\mathbf{X}}\left(\tilde{T}^{-1}(\mathbf{y})\right)}{|\Delta(\tilde{T}^{-1}(\mathbf{y}))|}\mathbf{1}_{\tilde{T}(M)}(\mathbf{y}), \quad \mathbf{y} \in \mathbb{R}^p\,.}$$

Note that $\Delta(\tilde{T}^{-1}(\mathbf{y}))$ is the Jacobian matrix of the mapping T evaluated at $\tilde{T}^{-1}(\mathbf{y})$.

- **Convolution in the continuous case:** Let X and Y be independent random variables with Riemann PDF f and g, respectively. Then $X+Y$ has the PDF h given by:

$$h(z) = \int_{-\infty}^{\infty} f(z-y)g(y)\,\mathrm{d}y = \int_{-\infty}^{\infty} f(x)g(z-x)\,\mathrm{d}x, \quad z \in \mathbb{R},$$

i. e., $P(X+Y \in (a,b)) = \int_a^b h(z)\,\mathrm{d}z,\ a < b$.

2.7 Moments, means and variance-covariance matrix

Here and in the following a general assumption is, that all random variables and vectors are defined on a suitable probability space $(\Omega, \mathcal{A}, P)$.

- **Mean / Expectation (value):** Let X be a random variable and $g : \mathbb{R} \to \mathbb{R}$. Then, the expectation of $g(X)$ is given by (provided, that the sum / integral exists)

$$\mathrm{E}g(X) = \begin{cases} \sum x \in \operatorname{supp}(P^X) g(x)P(X=x), & P^X \text{ discrete} \\ \int_{\mathbb{R}} g(x)f^X(x)\,\mathrm{d}x, & P^X \text{ continuous with PDF } f^X \end{cases}.$$

For $g(X) = X$ the value $\mathrm{E}X$ is called expectation value or mean.

- Alternatively one can take a look at $P^{g(X)}$ and use

$$\mathrm{E}g(X) = \sum_{t \in \operatorname{supp}(P^{g(X)})} \mathbf{t}P(g(X)=t)\,.$$

- **Moment, variance, moment generating function:** Let X be a random variable and $k \in \mathbb{N}$. Suppose all expectations exist. Then,

  (i) $\mathrm{E}(X^k)$ is called the kth moment of X.

  (ii) $\mathrm{E}(X-\mathrm{E}X)^2$ is called the variance of X and denoted by $\mathrm{Var}X$.

  (iii) The function

$$\psi_X : D \to \mathbb{R},\ \psi(t) = \mathrm{E}(e^{tX}), \quad t \in D,$$

  is called the moment generating function (MGF) where $D \subseteq \mathbb{R}$ denotes the set of reals where the expectation $\mathrm{E}(e^{tX})$ exists (is finite).

- **Covariance and correlation:** Let X, Y be random variables with existing second moments. Then:

  (i) $\mathrm{Cov}(X,Y) = \mathrm{E}\left[(X-\mathrm{E}X)(Y-\mathrm{E}Y)\right]$ is called covariance of X and Y.

(ii) If $\mathrm{Var}(X), \mathrm{Var}(Y) > 0$ then

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

is called correlation of X and Y.

- **Properties of means:** Let $X, Y, X_1, ..., X_n$ be random variables with existing means and $a \in \mathbb{R}$. Then:

  (i) $\mathrm{E}(aX) = a\,\mathrm{E}X$, $\mathrm{E}(a) = a$

  (ii) $\mathrm{E}(X+Y) = \mathrm{E}X + \mathrm{E}Y$

  (iii) $X \le Y \quad \Rightarrow \quad \mathrm{E}X \le \mathrm{E}Y$

  (iv) If $X_1, ..., X_n$ are independent, then

$$\mathrm{E}\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n \mathrm{E}X_i\,.$$

- **Properties of variance:** Let X, Y be random variables with existing mean and variance, respectivley. Then:

  (i) $\mathrm{Var}X = \mathrm{E}(X - \mathrm{E}X)^2 = \mathrm{E}X^2 - \mathrm{E}^2 X$

  (ii) $\mathrm{Var}(a+bX)0b^2\mathrm{Var}X, \quad a, b \in \mathbb{R}$

  (iii) $X \le Y \quad \Rightarrow \quad \mathrm{E}X \le \mathrm{E}Y$

  (iv) For a random variable X with $\mathrm{E}X = \mu,\ \mathrm{Var}X = \sigma^2 > 0$, the standardisation

$$Y = \frac{X - \mathrm{E}X}{\sqrt{\mathrm{Var}X}}$$

  satisfies $\mathrm{E}Y = 0,\ \mathrm{Var}Y = 1$.

- **Properties of covaraince and corraltion:** Let $X, Y, X_1, ..., X_n$ be random variables with existing mean and variance, respectivley. Then:

  (i) $\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}X)(Y - \mathrm{E}Y)] = \mathrm{E}(XY) - \mathrm{E}X\mathrm{E}Y$

  (ii) $\mathrm{Cov}(X,X) = \mathrm{Var}X, \quad \mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$

  (iii) $\mathrm{Cov}(a+bX, c+dY) = bd\,\mathrm{Cov}(X,Y), \quad a,b,c,d \in \mathbb{R}$

  (iv) X, Y independent implies $\mathrm{Cov}(X,Y) = 0$ (not vice versa).

  (v) The correlation is bounded by -1 and 1:

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}X \cdot \mathrm{Var}Y}} \in [-1,1] \ \text{ (if } \mathrm{Var}X, \mathrm{Var}Y > 0)\,.$$

(*vi*) If $X_1, ..., X_n$ are uncorrelated then

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}X_i\,,$$

which for $X_1, ..., X_n$ iid can be further simplified to $... = \mathrm{Var}X_1/n$.

- **Expectation of random vectors and matrices:**

  (*i*) The expectation of a random vector $\mathbf{X} = (X_1, ..., X_n)^T$ is defined by the vector of means $\mathrm{E}\mathbf{X} = (\mathrm{E}X_1, ..., \mathrm{E}X_n)^T$.

  (*ii*) The expectation of a random matrix $\mathfrak{X} = (X_{ij})_{1\le i\le p,\,1\le j\le q}$ is defined by the matrix of means $\mathrm{E}\mathfrak{X} = (\mathrm{E}X_{ij})_{1\le i\le p,\,1\le j\le q}$.

- **Transformations of the mean:**

  (*i*) Let $\mathbf{X} = (X_1, ..., X_p)^T$ be a p-dimensional random vector and $A \in \mathbb{R}^{k\times p}$, $\mathbf{b} \in \mathbb{R}^k$. Then:

$$\mathrm{E}(A\mathbf{X} + \mathbf{b}) = A\,\mathrm{E}\mathbf{X} + \mathbf{b}.$$

  (*ii*) Let $\mathbf{Z}_1, ..., \mathbf{Z}_n$ be a p-dimensional random vectors and $A_1, ..., A_n \in \mathbb{R}^{k\times p}$. Then:

$$\mathrm{E}\left(\sum_{j=1}^{n} A_j\mathbf{Z}_j\right) = \sum_{j=1}^{n} A_j\,\mathrm{E}(\mathbf{Z}_j).$$

- **Variance-covariance matrix:** Let $\mathbf{X} = (X_1, ..., X_p)^T$, $\mathbf{Y} = (Y_1, ..., Y_p)^T$ be random vectors. Then the covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$ is defined by

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \begin{pmatrix} \mathrm{Cov}(X_1, Y_1) & ... & \mathrm{Cov}(X_1, Y_q) \\ \vdots & ... & \vdots \\ \mathrm{Cov}(X_p, Y_1) & ... & \mathrm{Cov}(X_p, Y_q) \end{pmatrix}.$$

  The variance-covariance matrix of $\mathbf{X}$ is defined as $\Sigma = \mathrm{Cov}(\mathbf{X}, \mathbf{X}) := \mathrm{Cov}(\mathbf{X})$. This can be rewritten using the random matrix $\mathscr{C}_{\mathbf{X},\mathbf{Y}} = (\mathbf{X} - \mathrm{E}\mathbf{X})(\mathbf{Y} - \mathrm{E}\mathbf{Y})^T$.

- **Properties of the Covariance matrix:** Using the notation from above, one gets for $A \in \mathbb{R}^{k\times p}$, $B \in \mathbb{R}^{r\times q}$, $\mathbf{b} \in \mathbb{R}^k$, $\mathbf{c} \in \mathbb{R}^r$:

  (*i*) $\mathrm{Cov}(A\mathbf{X} + \mathbf{b},\, B\mathbf{Y} + \mathbf{c}) = A\mathrm{Cov}(\mathbf{X}, \mathbf{Y})B^T$.

  (*ii*) $\mathrm{Cov}(A\mathbf{X} + \mathbf{b}) = A\mathrm{Cov}(\mathbf{X})A^T$.

(iii) Let $(\mathbf{X}, \mathbf{Y})^T$ denote the combined vetor of $\mathbf{X}$ and $\mathbf{Y}$. Then:

$$\operatorname{Cov}\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix} = \begin{bmatrix}\operatorname{Cov}(\mathbf{X}) & \operatorname{Cov}(\mathbf{X},\mathbf{Y})\\ \operatorname{Cov}(\mathbf{Y},\mathbf{X}) & \operatorname{Cov}(\mathbf{Y})\end{bmatrix}.$$

(iv) $\operatorname{Cov}(\mathbf{X}, \mathbf{Y}) = \operatorname{Cov}(\mathbf{Y}, \mathbf{X})^T$.

- One also denotes shortly $\Sigma_{\mathbf{XY}} = \operatorname{Cov}(\mathbf{X}, \mathbf{Y})$.
- **Definitenes of the covariance matrix:** Let $\mathbf{X}$ be a p-dimensional random vector. Then $\Sigma = \operatorname{Cov}(\mathbf{X})$ is a positive semidefinite matrix.
- **MGF for random vectors:** Let $\mathbf{X} = (X_1, ..., X_k)^T$ be a k-dimensional random vector. Suppose all expectations exist. Then the function

$$\psi_{\mathbf{X}} : D \to \mathbb{R},\ \psi_{\mathbf{X}}(t) = \operatorname{E}\left(e^{\mathbf{t}^T\mathbf{X}}\right), \quad \mathbf{t} = (t_1, ..., t_k)^T \in D,$$

is called moment generating function where $D \subseteq \mathbb{R}^k$ denotes the set of real vectors where the expectation $\operatorname{E}(e^{\mathbf{t}^T\mathbf{X}})$ exists.

- **Propertires of the MGF:**

(i) For a random vector $\mathbf{X} = (X_1, ..., X_k)^T$ the MGF can be written as

$$\psi_{\mathbf{X}}(\mathbf{t}) = \operatorname{E}\left(e^{\mathbf{t}^T\mathbf{X}}\right) = \operatorname{E}\left(\exp\left(\sum_{j=1}^{k} t_j X_j\right)\right), \quad \mathbf{t} \in D\,.$$

(ii) $\psi_{\mathbf{X}}(\mathbf{0}) = 1$.

(iii) If $\mathbf{X}$ is a discrete random vector with finite support, then $\psi_{\mathbf{X}}$ always exists.

- **Moments and MGF:** Let $\mathbf{X} = (X_1, ..., X_k)^T$ be a random Vector with MGF $\psi_{\mathbf{X}}$ such that $\psi_{\mathbf{X}}$ exists for some open set $M \subseteq \mathbb{R}^k$ containing the zero vector. Then:

(i) Let $\psi_{\mathbf{Y}}$ be the MFG of a random vector $\mathbf{Y}$.

I f$\psi_{\mathbf{X}}(\mathbf{t}) = \psi_{\mathbf{Y}}(\mathbf{t})$ for $\mathbf{t} \in D \subseteq \mathbb{R}^k$, where D is an open set containing the zero vector, then $\mathbf{X} \overset{d}{=}$, that is the MGF determines the distribution uniquely.

(ii) $\psi_{\mathbf{X}}$ is infinetly often differentiable on M and, for $l_1, ..., l_k \in \mathbb{N}_0$, we have

$$\operatorname{E}\left(\prod_{j=1}^{k} X_j^{l_j}\right) = \frac{\partial^{l_1+...+l_k}}{\partial^{l_1}t_1...\partial^{l_k}t_k}\psi_{\mathbf{X}}(\mathbf{t})\Big|_{\mathbf{t}=0}.$$

- **Properties of the MGF:**

(i) The MGF of a marginal distribution (say of $\mathbf{X}_K$) is btained by choosing $\mathbf{t} = (t_1, ..., t_k)^T$ with $t_j = 0,\ j \notin K$.

(ii) Let $\mathbf{X} = (X_1, ..., X_k)^T$ be a random vector, $A \in \mathbb{R}^{m \times k},\ \mathbf{b} \in \mathbb{R}^m$, and $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$. Then:

$$\psi_{\mathbf{Y}}(\mathbf{t}) = e^{\mathbf{t}^T\mathbf{b}} \cdot \psi_{\mathbf{X}}(A^T\mathbf{t}) \quad \forall\ \ \mathbf{t} \text{ (such that the MGF exists).}$$

(iii) If $\mathbf{X}_1, ..., \mathbf{X}_k$ are independent random vectors and $\mathbf{X} = (\mathbf{X}_1^T, ..., \mathbf{X}_k^T)^T$ and $\mathbf{S} = \sum_{j=1}^k \mathbf{X}_j$, then

$$\psi_{\mathbf{X}}(\mathbf{t}_1, ..., \mathbf{t}_k) = \prod_{j=1}^k \psi_{\mathbf{X}_j}(\mathbf{t}_j) \quad \forall\ \ \mathbf{t}_1, ..., \mathbf{t}_k \text{ (such that the MGF exist).}$$
$$\psi_{\mathbf{S}}(\mathbf{t}) = \prod_{j=1}^k \psi_{\mathbf{X}_j}(\mathbf{t}) \quad \forall\ \ \mathbf{t} \text{ (such that the MGF exist).}$$

- **Conditional expectations:** Let $\mathbf{X}, \mathbf{Y}$ be random vectors with joint PMF or Riemann PDF $f^{\mathbf{X},\mathbf{Y}}$. Then:

(i) Conditional expectatioons are defined via the conditional PMF or PDF:

$$\mathrm{E}(\mathbf{X}|\mathbf{Y} = \mathbf{y}) = \begin{cases} \sum\limits_{\mathbf{x} \in \mathrm{supp}(P^{\mathbf{X}|\mathbf{Y}=y})} \mathbf{x} f^{\mathbf{X}|\mathbf{Y}=\mathbf{y}} & \text{discrete case} \\ \int \mathbf{x} f^{\mathbf{X}|\mathbf{Y}=\mathbf{y}}(\mathbf{x})\,\mathrm{d}\mathbf{x} & \text{continuous case} \end{cases}.$$

(ii) The same approach can be used to define conditional expectations of functions of random vectros (i. e. conditional (co-)variances). For some function h we get:

$$\mathrm{E}\Big(h(\mathbf{X}, \mathbf{Y}|\mathbf{Y} = \mathbf{y}\Big) = \mathrm{E}\Big(h(\mathbf{X}, \mathbf{y}|\mathbf{Y} = \mathbf{y}\Big)$$

(iii) If $\mathbf{X}, \mathbf{Y}$ are independent random vectors, then

$$\mathrm{E}\Big(h(\mathbf{X}, \mathbf{Y}|\mathbf{Y} = \mathbf{y}\Big) = \mathrm{E}\Big(h(\mathbf{X}, \mathbf{y})\Big).$$