

Set Theory

Definition 1.1 Set $A = \{1, 3, 2\}$: is a well-defined group of distinct items that are considered as an object in its own right. The arrangement/order of the objects does not matter but each member of the set must be unique.

Definition 1.2 Empty Set $\{\}/\emptyset$: is the unique set having no elements/cardinality^[def. 1.5] zero.

Definition 1.3 Multiset/Bag: Is a set-like object in which multiplicity^[def. 1.4] matters, that is we can have multiple elements of the same type.
I.e. $\{1, 1, 2, 3\} \neq \{1, 2, 3\}$

Definition 1.4 Multiplicity: The multiplicity n_a of a member a of a multiset^[def. 1.3] S is the number of times it appears in that set.

Definition 1.5 Cardinality $|S|$: Is the number of elements that are contained in a set.

Definition 1.6 The Power Set $\mathcal{P}(S)/2^S$: The power set of any set S is the set of all subsets of S , including the empty set and S itself. The cardinality of the power set is 2^S is equal to $2^{|S|}$.

1. Closure

Definition 1.7 Closure: A set is *closed* under an operation Ω if performance of that operations onto members of the set always produces a member of that set.

2. Open vs. Closed Sets

Definition 1.8 Open Sets:

- Euclidean Spaces**:
A subset $U \in \mathbb{R}$ is open, if for every $x \in U$ it exists $\epsilon(x) \in \mathbb{R}_+$ s.t. a point $y \in \mathbb{R}$ belongs to U if:
$$\|x - y\|_2 < \epsilon(x) \tag{1.1}$$
- Metric Spaces**^[def. 10.65]: a Subset U of a metric space (M, d) is open if:
$$\exists \epsilon > 0 : \quad \text{if} \quad d(x, y) < \epsilon \quad \forall y \in M, \forall x \in U \implies y \in U \tag{1.2}$$
- Topological Spaces**^[def. 12.2]: Let (X, τ) be a topological space. A set A is said to be open if it is contained in τ .



Definition 1.9 Closed Set: Is the complement of an open set^[def. 1.8].

Definition 1.10 Bounded Set: A set $S \subset \mathbb{R}^n$ is *bounded* if there exists a constant K s.t. the absolute value of every component of every element of S is less or equal to K .

3. Number Sets

3.1. The Real Numbers

3.1.1. Intervals

Definition 1.11 Closed Interval $[a, b]$: The closed interval of a and b is the set of all real numbers that are within a and b , including a and b :
$$[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\} \tag{1.3}$$

Definition 1.12 Open Interval (a, b) : The open interval of a and b is the set of all real numbers that are within a and b :
$$(a, b) = \{x \in \mathbb{R} \mid a < x \leq b\} \tag{1.4}$$

3.2. The Rational Numbers

Example 1.1 Power Set/Cardinality of $S = \{x, y, z\}$: The subsets of S are:
 $\{\emptyset\}, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}$
and hence the power set of S is $\mathcal{P}(S) = \{\{\emptyset\}, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}\}$ with a cardinality of $|S| = 2^3 = 8$.

4. Set Functions

4.1. Submodular Set Functions

Definition 1.13 Submodular Set Functions: A submodular function $f : 2^\Omega \rightarrow \mathbb{R}$ is a function that satisfies:
$$f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B) \quad \forall A \subseteq B \subset \Omega \quad \{x\} \in \Omega \setminus B \tag{1.5}$$

Explanation 1.1 (Definition 1.13). *Adding an element x to the the smaller subset A yields at least as much information/-value gain as adding it to the larger subset B .*

Definition 1.14 Montone Submodular Function: A *monotone* submodular function is a submodular function^[def. 1.13] that satisfies:
$$f(A) \leq f(B) \quad \forall A \subseteq B \subset \Omega \tag{1.6}$$

Explanation 1.2 (Definition 1.14). *Adding more elements to a set will always increase the information/value gain.*

4.2. Complex Numbers

Definition 1.15 Complex Conjugate \bar{z} : The complex conjugate of a complex number $z = x + iy$ is defined as:
$$\bar{z} = x - iy \tag{1.7}$$



Corollary 1.1 Complex Conjugate Of a Real Number: The complex conjugate of a real number $x \in \mathbb{R}$ is x :
$$\bar{x} = x \implies x \in \mathbb{R} \tag{1.8}$$

Formula 1.1 Euler's Formula:
$$e^{\pm ix} = \cos x \pm i \sin x \tag{1.9}$$

Formula 1.2 Euler's Identity:
$$e^{\pm i} = -1 \tag{1.10}$$

Note
$$e^n = 1 \Leftrightarrow n = i 2\pi k, \quad k \in \mathbb{N} \tag{1.11}$$

Sequences&Series

Definition 2.1 Index Set: Is a set^[def. 1.1] A , whose members are labels to another set S . In other words its members index member of another set. An index set is build by enumerating the members of S using a function f s.t.
$$f : A \mapsto S \quad A \in \mathbb{N} \tag{2.1}$$

Definition 2.2 Sequence $(a_n)_{n \in A}$: A sequence is an by an *index set* A enumerated multiset^[def. 1.3] (repetitions are allowed) of objects in which *order does matter*.

Definition 2.3 Series: is an infinite ordered set of terms combined together by addition.

1. Types of Sequences

1.1. Arithmetic Sequence

Definition 2.4 Arithmetic Sequence: Is a sequence where the *difference* between two consecutive terms constant i.e. $(2, 4, 8, 16, 32, \dots)$.
$$t_n = t_0 + nd \quad d : \text{difference between two terms} \tag{2.2}$$

1.2. Geometric Sequence

Definition 2.5 Geometric Sequence: Is a sequence where the *ratio* between two consecutive terms constant i.e. $(2, 4, 8, 16, 32, \dots)$.
$$t_n = t_0 \cdot r^n \quad r : \text{ratio between two terms} \tag{2.3}$$

Property 2.1 Sum of Geometric Sequence:
$$\sum_{k=1}^n ar^{k-1} = \frac{a(1 - r^n)}{1 - r} \tag{2.4}$$

2. Converging Sequences

2.1. Pointwise Convergence

Definition 2.6 Pointwise Convergence^[?]: $\lim_{n \rightarrow \infty} f_n = f$ **pointwise**
Let (f_n) be a sequence of functions with the same domain^[def. 5.8] and codomain^[def. 5.9]. The sequence is said to convergence pointwise to its *pointwise limit function* f if it satisfies:
$$\lim_{n \rightarrow \infty} |f_n(x) - f(x)| = 0 \quad \forall x \in \text{dom}(f_i) \tag{2.5}$$

2.2. Uniform Convergence

Definition 2.7 Uniform Convergence^[?]: $\lim_{n \rightarrow \infty} f_n = f$ **uniform**/ $f_n \xrightarrow{\infty} f$
Let (g_n) be a sequence of functions with the same domain^[def. 5.8] and codomain^[def. 5.9]. The sequence is said to convergence uniformly to its *pointwise limit function* f if it satisfies:
$$\exists \epsilon > 0 : \exists n \geq 1 \quad \sup_{x \in \text{dom}(f_i)} |g_n(x) - f(x)| < \epsilon \quad \forall x \in \text{dom}(f_i) \tag{2.6}$$

Note

Uniform convergence is characterized by the uniform norm??, and is stronger than pointwise convergence.

Topology

Definition 3.1 Topological Space^[?] (X, τ) : Is an ordered pair (X, τ) , where X is a set and τ is a topology^[def. 12.1] on X .

Definition 3.2 Topological Space^[?] (X, τ) : Is an ordered pair (X, τ) , where X is a set and τ is a topology^[def. 12.1] on X .

1. Weak Topologies

Definition 3.3 Weak Topology $\mathcal{C}(\mathcal{K}; \mathbb{R})$: Is the corests topology s.t all cont. linear functionals w.r.t. to the strong topology are continuous.
Neighbourhood Basis:
 $\{f \mid |l_1| < \epsilon_1, \dots, |l_n| < \epsilon_n, \forall \epsilon_i, \forall n, \forall \text{lin. functions } f\} \tag{3.1}$

Note

The weak closure:

- is usually larger as the uniform closure, as for the weak closure there are many more convergence sequences
- is easier to calculate than the uniform closure

2. Compact Space

Corollary 3.1 Euclidean Space: In the euclidean case, a set $X \in \mathbb{R}$ is compact iff:

- it is closed^[def. 1.9]
- bounded

3. Closure

Definition 3.4 Closure of a Set^[?] $\text{cl}_{X, \tau}(S) / \bar{S}$: The closure of a subset S of a topological space^[def. 12.2] (X, τ) is defined equivalantly by:

- Is the union of S and its boundary ∂S .
- is the set S together with its limit points.

Note

If the topological space X, τ is clear from context, then the closure of a set S is often written simply as \bar{S} .

Corollary 3.2 Uniform Closure $\|\cdot\|_\infty$: The uniform closure of a set of functions A is the *space of all functions that can be approximated* by a sequence (f_n) of uniformly-converging functions from A .^[def. 2.7] functions

Corollary 3.3 Weak Closure:

Logic

1. Boolean Algebra

1.1. Basic Operations

Definition 4.1 **Conjunction**/AND \wedge :

Definition 4.2 **Disjunction**/OR \vee :

Definition 4.3 **Negation**/NOT \neg :

1.1.1. Expression as Integer

If the truth values $\{0, 1\}$ are interpreted as integers then the basic operations can be represent with basic arithmetic operations.

$$\begin{aligned}x \wedge y &= xy = \min(x, y) \\x \vee y &= x + y = \max(x, y) \\ \neg x &= 1 - x \\x \oplus y &= (x + y) \cdot (\neg x + \neg y) = x \cdot \neg y + \neg x \cdot y\end{aligned}$$

Note: non-linearity of XOR

$$(x + y) \cdot (\neg x + \neg y) = -x^2 - y^2 - 2xy + 2x + 2y$$

1.2. Boolean Identities

Property 4.1 Idempotence:
 $x \wedge x \equiv x$ and $x \vee x \equiv x$ (4.1)

Property 4.2 Identity Laws:
 $x \wedge \text{true} \equiv x$ and $x \vee \text{false} \equiv x$ (4.2)

Property 4.3 Zero Law's:
 $x \wedge \text{false} \equiv \text{false}$ and $x \vee \text{true} \equiv \text{true}$ (4.3)

Property 4.4 Double Negation:
 $\neg \neg x \equiv x$ (4.4)

Property 4.5 Complementation:
 $x \wedge \neg x \equiv \text{false}$ and $x \vee \neg x \equiv \text{true}$ (4.5)

Property 4.6 Commutativity:
 $x \vee y \equiv y \vee x$ and $x \wedge y \equiv y \wedge x$ (4.6)

Property 4.7 Associativity:
 $(x \vee y) \vee z \equiv x \vee (y \vee z)$ (4.7)
 $(x \wedge y) \wedge z \equiv x \wedge (y \wedge z)$ (4.8)

Property 4.8 Distributivity:
 $x \vee (y \wedge z) \equiv (x \vee y) \wedge (x \vee z)$ (4.9)
 $x \wedge (y \vee z) \equiv (x \wedge y) \vee (x \wedge z)$ (4.10)

Property 4.9 De Morgan's Laws:
 $\neg(x \vee z) \equiv (\neg x \wedge \neg y)$ (4.11)
 $\neg(x \wedge z) \equiv (\neg x \vee \neg y)$ (4.12)

Note

The algebra axioms come in pairs that can be obtained by interchanging \wedge and \vee .

1.3. Normal Forms

Definition 4.4 **Literal** [example 4.1]:
Literals are atomic formulas or their negations

Definition 4.5 **Negation Normal Form (NNF)**: A formula F is in negation normal form if the negation operator \neg is only applied to literals^[def. 4.4] and the only other operators are \wedge and \vee .

Definition 4.6 **Conjunctive Normal Form (CNF)**: An boolean algebraic expression F is in CNF if it is a *conjunction* of *clauses*, where each clause is a disjunction of *literals*^[def. 4.4] $L_{i,j}$:

$$F_{\text{CNF}} = \bigwedge_{i=1}^n \left(\bigvee_{j=1}^{m_i} L_{i,j} \right) \quad (4.13)$$

Definition 4.7 **Disjunctive Normal Form (DNF)**: An boolean algebraic expression F is in DNF if it is a *disjunction* of *clauses*, where each clause is a conjunction of *literals*^[def. 4.4] $L_{i,j}$:

$$F_{\text{DNF}} = \bigvee_{i=1}^n \left(\bigwedge_{j=1}^{m_i} L_{i,j} \right) \quad (4.14)$$

Note

- true is a CNF with no clause and a single literal.
- false is a CNF with a single clause and no literals

1.3.1. Transformation to CNF and DNF

DNF

Algorithm 4.1:

- ① Using *De Morgan's laws*Property 4.9 and double negationProperty 4.4 transform F into *Negation Normal Form*^[def. 4.5]:

$$\begin{array}{lll} \neg \neg x & \text{by} & x \\ \neg(x \wedge y) & \text{by} & (\neg x \vee \neg y) \\ \neg(x \vee y) & \text{by} & (\neg x \wedge \neg y) \\ \neg \text{true} & \text{by} & \text{false} \\ \neg \text{false} & \text{by} & \text{true} \end{array}$$

- ② Using distributive lawsProperty 4.8 substitute all:

$$\begin{array}{lll} x \wedge (y \vee z) & \text{by} & (x \wedge y) \vee (x \wedge z) \\ (y \vee z) \wedge x & \text{by} & (y \wedge x) \vee (z \wedge x) \\ x \wedge \text{true} & \text{by} & \text{true} \\ \text{true} \wedge x & \text{by} & \text{true} \end{array}$$

- ③ Using the identityProperty 4.2 and zero laws Property 4.3 remove true from any cause and delete all clauses containing false.

Note

For the CNF form simply use duality for step 2 and 3 i.e. swap \wedge and \vee and true and false.

Using Truth Tables [example 4.2]

To obtain a DNF formula from a truth table we need to have a *conjunctive*^[def. 4.3] for each row where F is true.

2. Examples

Example 4.1 **Literals:**

Boolean literals: $x, \neg y, s$

Not boolean literals: $\neg \neg x, (x \wedge y)$

Example 4.2 **DNF from truth tables:**

	x	y	z	F
	0	0	0	1
Need a conjunction of:	0	0	1	0
• $(\neg x \wedge \neg y \wedge \neg z)$	0	1	0	0
• $(\neg x \wedge y \wedge z)$	0	1	1	1
• $(x \wedge \neg y \wedge \neg z)$	1	0	0	1
• $(x \wedge y \wedge z)$	1	0	1	0
	1	1	0	0
	1	1	1	1

$$(\neg x \wedge \neg y \wedge \neg z) \wedge (\neg x \wedge y \wedge z) \wedge (x \wedge \neg y \wedge \neg z) \wedge (x \wedge y \wedge z)$$

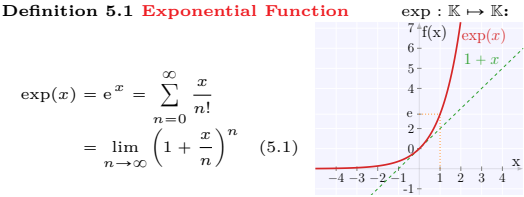
Calculus and Analysis

1. Functional Analysis

1.1. Elementary Functions

1.1.1. Exponential Numbers

Definition 5.1 Exponential Function



Definition 5.2 Exponential/Euler Number

$e :=$

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = 2.7182 \quad (5.2)$$

Properties Defining the Exponential Function

Property 5.1:

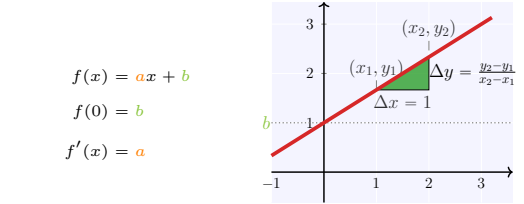
$$\exp(x + y) = \exp(x) + \exp(y) \quad (5.3)$$

Property 5.2:

$$\exp(x) \leq 1 + x \quad (5.4)$$

1.1.2. Affine Linear Functions

Definition 5.3 Affine Linear Function $f(x) = ax + b$:
An affine linear function are functions that can be defined by a scaling $s_a(x) = ax$ plus a translation $t_b(x) = x + b$:
 $M = \{f : \mathbb{R} \mapsto \mathbb{R} | f(x) = (s_a \circ t_b)(x) = ax + b, \quad a, b \in \mathbb{R}\}$ (5.5)



Formula 5.1 [proof 5.1]
Linear Function from Point and slope $f(x_0) = y_1$:
Given a point (x_1, y_1) and a slope a we can derive:
 $f(x) = a \cdot (x - x_0) + y_0 = ax + (y_1 - ax_0)$ (5.6)

Formula 5.2 Linear Function from two Points:

$$f(x) = a \cdot (x - x_p) + y_p = ax + (y_p - ax_p) \quad (5.7)$$
$$a = \frac{y_1 - y_0}{x_1 - x_0} \quad p = \{0 \text{ or } 1\}$$

1.1.3. Polynomials

Definition 5.4 Polynomial: A function $\mathcal{P}_n : \mathbb{R} \mapsto \mathbb{R}$ is called *Polynomial*, if it can be represented in the form:
 $\mathcal{P}_n(x) = a_0 + a_1x + a_2x^2 + \dots + a_{n-1}x^{n-1} + a_nx^n$ (5.8)

Corollary 5.1 Degree n-of a Polynomial $\deg(\mathcal{P}_n)$: the *degree* of the polynomial is the highest exponent of the variable x , among all non-zero coefficients $a_i \neq 0$.

Definition 5.5 Monomial: Is a polynomial with only one term.

Cubic Polynomials

Definition 5.6 Cubic Polynomials: Are polynomials of degree^[cor. 5.1] 3 and have four coefficients:
 $f(x) = a_3x^3 + a_2x^2 + a_1x + a_0$ (5.9)

1.2. Functional Compositions

Definition 5.7 Functional Compositions $f \circ g$:
Let $f : A \mapsto B$ and $g : D \mapsto C$ be two mappings s.t. $\text{codom}(f) \subseteq D$ then we can define a composition function $(f \circ g)A \mapsto D$ as:
 $h(x) = (g \circ f)(x) = g(f(x))$ with $x \in A$ (5.10)

Corollary 5.2 Nested Functional Composition:

$$F_{k;1}(x) = (F_k \circ \dots \circ F_1)(x) = F_k(F_{k-1} \circ \dots \circ (F_1(x))) \quad (5.11)$$

2. Proofs

Proof 5.1 formula 5.1:

$$f(x_0) = y_0 = ax_0 + b \quad \Rightarrow \quad b = y_0 - ax_0$$

Theorem 5.1

First Fundamental Theorem of Calculus:

Let f be a continuous real-valued function defined on a closed interval $[a, b]$.
Let F be the function defined $\forall x \in [a, b]$ by:

$$F(x) = \int_a^x f(t) dt \quad (5.12)$$

Then it follows:

$$F'(x) = f(x) \quad \forall x \in (a, b) \quad (5.13)$$

Theorem 5.2

Second Fundamental Theorem of Calculus:

Let f be a real-valued function on a closed interval $[a, b]$ and F an antiderivative of f in $[a, b]$: $F'(x) = f(x)$, then it follows if f is Riemann integrable on $[a, b]$:

$$\int_a^b f(t) dt = F(b) - F(a) \quad \Longleftrightarrow \quad \int_a^x \frac{\partial}{\partial x} F(t) dt = F(x) \quad (5.14)$$

Definition 5.8 Domain of a function

$\text{dom}(\cdot)$:
Given a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, the set of all possible input values \mathcal{X} is called the domain of $f - \text{dom}(f)$.

Definition 5.9

Codomain/target set of a function

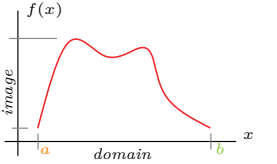
$\text{codom}(\cdot)$:
Given a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, the codomain of that function is the set \mathcal{Y} into which all of the output of the function is constrained to fall.

Definition 5.10 Image (Range) of a function: $f[\cdot]$

Given a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, the image of that function is the set to which the function can actually map:

$$\{y \in \mathcal{Y} | y = f(x), \quad \forall x \in \mathcal{X}\} := f[\mathcal{X}] \quad (5.15)$$

Evaluating the function f at each element of a given subset A of its domain $\text{dom}(f)$ produces a set called the *image* of A under (or through) f .
The image is thus a subset of a function's codomain.



Misnomer Range: The term Range is ambiguous s.t. certain books refer to it as codomain and other as image.

Definition 5.11 Inverse Image/Preimage $f^{-1}(\cdot)$:

Let $f : X \mapsto Y$ be a function, and A a subset set of its codomain Y .
Then the preimage of A under f is the set of all elements of the domain X , that map to elements in A under f :

$$f^{-1}(A) = \{x \subseteq X : f(x) \subseteq A\} \quad (5.16)$$

Example 5.1 :

Given $f : \mathbb{R} \rightarrow \mathbb{R}$
defined by $f : x \mapsto x^2 \Longleftrightarrow f(x) = x^2$
 $\text{dom}(f) = \mathbb{R}$, $\text{codom}(f) = \mathbb{R}$ but its image is $f[\mathbb{R}] = \mathbb{R}_+$.

Image (Range) of a subset

The image of a subset $A \subseteq \mathcal{X}$ under f is the subset $f[A] \subseteq \mathcal{Y}$ defined by:

$$f[A] = \{y \in \mathcal{Y} | y = f(x), \quad \forall x \in A\} \quad (5.17)$$

Note: Range

The term range is ambiguous as it may refer to the image or the codomain, depending on the definition.
However, modern usage almost always uses range to mean image.

Definition 5.12 (strictly) Increasing Functions:

A function f is called monotonically increasing/increasing/non-decreasing if:
 $x \leq y \Longleftrightarrow f(x) \leq f(y) \quad \forall x, y \in \text{dom}(f)$ (5.18)

And **strictly increasing** if:

$$x < y \Longleftrightarrow f(x) < f(y) \quad \forall x, y \in \text{dom}(f) \quad (5.19)$$

Definition 5.13 (strictly) Decreasing Functions:

A function f is called monotonically decreasing/decreasing or non-increasing if:

$$x \geq y \Longleftrightarrow f(x) \geq f(y) \quad \forall x, y \in \text{dom}(f) \quad (5.20)$$

And **strictly** decreasing if:

$$x > y \Longleftrightarrow f(x) > f(y) \quad \forall x, y \in \text{dom}(f) \quad (5.21)$$

Definition 5.14 Monotonic Function:

A function f is called monotonic iff either f is **increasing** or **decreasing**.

Definition 5.15 Linear Function:

A function $L : \mathbb{R}^n \mapsto \mathbb{R}^m$ is linear if and only if:

$$L(x + y) = L(x) + L(y)$$
$$L(\alpha x) = \alpha L(x) \quad \forall x, y \in \mathbb{R}^n, \quad \alpha \in \mathbb{R}$$

Corollary 5.3 Linearity of Differentiation: The derivative of **any** linear combination of functions equals the same linear combination of the derivatives of the functions:

$$\frac{d}{dx} (af(x) + bg(x)) = a \frac{d}{dx} f(x) + b \frac{d}{dx} g(x) \quad a, b \in \mathbb{R} \quad (5.22)$$

Definition 5.16 Quadratic Function:

A function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is quadratic if it can be written in the form:

$$f(x) = \frac{1}{2} x^T A x + b^T x + c \quad (5.23)$$

3. Norms

3.1. Infinity/Supremum Norm

Definition 5.17 Infinity/Supremum Norm:

$$\|f\|_{\infty} := \sup_{x \in \text{dom}(f)} |f(x)| \quad (5.24)$$

Note

In order to make this a proper norm one usually considers *bounded functions* s.t.:

$$\|f\|_{\infty} \leq M < \infty$$

Corollary 5.4 Infinity Norm induced Metric: The infinity norm naturally induces a metric^[def. 10.64]:

$$d := (f, g) := \|f - g\|_{\infty} \quad (5.25)$$

4. Smoothness

Definition 5.18 Smoothness of a Function C^k :

Given a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, the function is said to be of class k if it is differentiable up to order k **and** continuous, on its entire domain:
 $f \in C^k(\mathcal{X}) \Longleftrightarrow \exists f', f'', \dots, f^{(k)}$ continuous (5.26)

Note

- P.w. continuous \neq continuous.
- A function of that is k times differentiable must at least be of class C^{k-1} .
- $C^m(\mathcal{X}) \subset C^{m-1}, \dots, C^1 \subset C^0$
- Continuity is implied by the differentiability of all **derivatives** of up to order $k - 1$.

4.0.1. Continuous Functions

Definition 5.19 Continuous Function C^0 : Functions that do not have any jumps or peaks.

4.0.2. Piece wise Continuous Functions

Definition 5.20 Piecewise Linear Functions C^0_{pw} :

4.0.3. Continously Differentiable Function

Corollary 5.5 Continuously Differentiable Function C^1 : Is the class of functions that consists of all differentiable functions whose derivative is continuous.

Hence a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ of the class must satisfy:

$$f \in C^1(\mathcal{X}) \Longleftrightarrow f' \text{ continuous} \quad (5.27)$$

4.0.4. Smooth Functions

Corollary 5.6 Smooth Function C^∞ : Is a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ that has derivatives infinitely many times differentiable.

$$f \in C^\infty(\mathcal{X}) \iff f', f'', \dots, f^{(\infty)} \quad (5.28)$$

4.1. Lipschitz Continuous Functions

Often functions are not differentiable but we still want to state something about the rate of change of a function \Rightarrow hence we need a weaker notion of differentiability.

Definition 5.21 Lipschitz Continuity:
A Lipschitz continuous function is a function f whose rate of change is bound by a Lipschitz Constant L :

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y}, \quad L > 0 \quad (5.29)$$

Note

This property is useful as it allows us to conclude that a small perturbation in the input (i.e. of an algorithm) will result in small changes of the output \Rightarrow tells us something about robustness.

4.1.1. Lipschitz Continuous Gradient

Definition 5.22 Lipschitz Continuous Gradient:
A continuously differentiable function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has L -Lipschitz continuous gradient if it satisfies:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f), \quad L > 0 \quad (5.30)$$

if $f \in C^2$, this is equivalent to:

$$\nabla^2 f(\mathbf{x}) \leq L \mathbf{I} \quad \forall \mathbf{x} \in \text{dom}(f), \quad L > 0 \quad (5.31)$$

Lemma 5.1 Descent Lemma [Poorfs 5.5,??]:
If a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has Lipschitz continuous gradient eq. (5.30) over its domain, then it holds that:

$$\|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})\| \leq \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \quad (5.32)$$

Note

If f is twice differentiable then the largest eigenvalue of the Hessian (Definition 6.8) of f is uniformly upper bounded by L

4.2. L-Smooth Functions

Definition 5.23 L-Smoothness:
A L -smooth function is a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ that satisfies:

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2$$

with $\forall \mathbf{x}, \mathbf{y} \in \text{dom}(f), \quad L > 0 \quad (5.33)$

If f is a twice differentiable this is equivalent to:

$$\nabla^2 f(\mathbf{x}) \leq L \mathbf{I} \quad L > 0 \quad (5.34)$$

Theorem 5.3 [proof 5.6]
L-Smoothness of convex functions:
A convex and L -Smooth function (def. 5.23) has a Lipschitz continuous gradient eq. (5.30) thus it holds that:

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \quad (5.35)$$

Note

L -smoothness is a weaker condition than L -Lipschitz continuous gradients

5. Convexity and Concavity

Read stuff about uniqueness and so on again in NPDE/or NUM CSE and add proofs

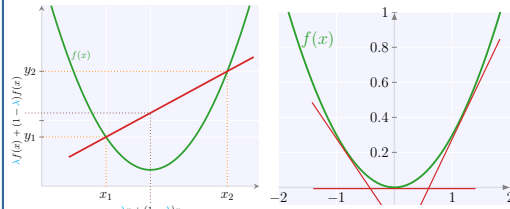
Definition 5.24 Convex Functions:
A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex if it satisfies:

$$f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) \quad \forall \lambda \in [0, 1] \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f) \quad (5.36)$$

If f is a differentiable function this is equivalent to:

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f) \quad (5.37)$$

If f is a twice differentiable function this is equivalent to:

$$\nabla^2 f(\mathbf{x}) \geq 0 \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f) \quad (5.38)$$


Definition 5.25 Concave Functions:
A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is concave if it satisfies:

$$f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) \geq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f) \quad \forall \lambda \in [0, 1] \quad (5.39)$$

Corollary 5.7 Convexity \rightarrow global minimima: Convexity implies that all local minima (if they exist) are global minima.

5.1. Properties

Property 5.3 Monotonicity of the Derivative:
If $f : \mathbb{R} \mapsto \mathbb{R}$ is

convex	$f'(a) < f'(b)$	$a < b, \quad a, b \in \mathbb{R}$
concave	$f'(a) > f'(b)$	

(5.40)

5.1.1. Properties that preserve convexity

Property 5.4 Non-negative weighted Sums: Let f be a convex function then $g(\mathbf{x})$ is convex as well:

$$g(\mathbf{x}) = \sum_{i=1}^n \alpha_i f_i(\mathbf{x}) \quad \forall \alpha_j > 0$$

Property 5.5 Composition of Affine Mappings: Let f be a convex function then $g(\mathbf{x})$ is convex as well:

$$g(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$$

Property 5.6 Pointwise Maxima: Let f be a convex function then $g(\mathbf{x})$ is convex as well:

$$g(\mathbf{x}) = \max_i \{f_i(\mathbf{x})\}$$

5.2. Strict Convexity/Concavity

Definition 5.26 Strictly Convex Functions:
A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is strictly convex if it satisfies:

$$f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) < \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f) \quad \forall \lambda \in [0, 1]$$

If f is a differentiable function this is equivalent to:

$$f(\mathbf{x}) > f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f) \quad (5.41)$$

If f is a twice differentiable function this is equivalent to:

$$\nabla^2 f(\mathbf{x}) > 0 \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f) \quad (5.42)$$

Intuition

- Convexity implies that a function f is bound by/below a linear interpolation from x to y and strong convexity that f is strictly bound/below.
- eq. (5.41) implies that $f(\mathbf{x})$ is above the tangent $f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in \text{dom}(f)$
- ?? implies that $f(\mathbf{x})$ is flat or curved upwards

Corollary 5.8 Strict Convexity \rightarrow Uniqueness:
Strict convexity implies a unique minimizer \iff at most one global minimum.

Corollary 5.9 : A twice differentiable function of one variable $f : \mathbb{R} \rightarrow \mathbb{R}$ is **convex** on an interval $\mathcal{X} = [a, b]$ if and only if its second derivative is non-negative on that interval \mathcal{X} :

$$f''(\mathbf{x}) \geq 0 \quad \forall \mathbf{x} \in \mathcal{X} \quad (5.43)$$

5.3. Strong Convexity/Concavity

Definition 5.27 μ -Strong Convexity:
Let \mathcal{X} be a Banach space over $\mathbb{K} = \mathbb{R}, \mathbb{C}$. A function $f : \mathcal{X} \rightarrow \mathbb{R}$ is called strongly convex iff the following equation holds:

$$f(t\mathbf{x} + (1 - t)\mathbf{y}) \leq tf(\mathbf{x}) + (1 - t)f(\mathbf{y}) - \frac{t(1 - t)}{2} \mu \|\mathbf{x} - \mathbf{y}\|^2 \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}, \quad t \in [0, 1], \quad \mu > 0$$

If $f \in C^1 \iff f$ is differentiable, this is equivalent to:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2 \quad (5.44)$$

If $f \in C^2 \iff f$ is twice differentiable, this is equivalent to:

$$\nabla^2 f(\mathbf{x}) \geq \mu \mathbf{I} \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad \mu > 0 \quad (5.45)$$

Corollary 5.10 Strong Convexity implies Strict Convexity:
<https://math.stackexchange.com/questions/2090991/proof-for-strongly-convex-function-is-strictly-convex>

Property 5.7:

$f(\mathbf{y}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{1}{2\mu} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \quad (5.46)$

Intuition

Strong convexity implies that a function f is lower bounded by its second order (quadratic) approximation, rather then only its first order (linear) approximation.

Size of μ

The parameter μ specifies how strongly the bounding quadratic function/approximation is.

Proof 5.2: eq. (5.45) analogously to **Proof** eq. (5.34)

Note

If f is twice differentiable then the smallest eigenvalue of the Hessian (def. 6.8) of f is uniformly lower bounded by μ
Hence strong convexity can be considered as the analogous to smoothness

Example 5.2 Quadratic Function: A quadratic function eq. (5.23) is convex if:

$$\nabla_{\mathbf{x}}^2 \text{eq. (5.23)} = \mathbf{A} \geq 0 \quad (5.47)$$

Corollary 5.11 :
Strong convexity \Rightarrow Strict convexity \Rightarrow Convexity

Functions

Even Functions: have rotational symmetry with respect to the origin.
 \Rightarrow **Geometrically:** its graph remains unchanged after reflection about the y-axis.

$$f(-x) = f(x) \quad (5.48)$$

Odd Functions: are symmetric w.r.t. to the y -axis.
 \Rightarrow **Geometrically:** its graph remains unchanged after rotation of 180 degrees about the origin.

$$f(-x) = -f(x) \quad (5.49)$$

Theorem 5.4 Rules:
Let f be even and f odd respectively.

$g =: f \cdot f$ is even	$g =: f \cdot f$ is even
$g =: f \cdot f$ is odd	the same holds for division

Examples

Even: $\cos x, |x|, c, x^2, x^4, \dots \exp(-x^2/2)$.
Odd: $\sin x, \tan x, x, x^3, x^5, \dots$

x-Shift: $f(x - c) \Rightarrow$ shift to the right
 $f(x + c) \Rightarrow$ shift to the left
 $f(x) \pm c \Rightarrow$ shift up/down

(5.50)
(5.51)

Proof 5.3: eq. (5.50) $f(x_n - c)$ we take the x -value at x_n but take the y -value at $x_0 := x_n - c$
 \Rightarrow we shift the function to x_n .

Euler's formula

$$e^{\pm ix} = \cos x \pm i \sin x \quad (5.52)$$

Euler's Identity

$$e^{\pm i} = -1 \quad (5.53)$$

Note

$$e^n = 1 \Leftrightarrow n = i2\pi k, \quad k \in \mathbb{N} \quad (5.54)$$

Corollary 5.12 Every norm is a convex function: By using definition [def. 5.24] and the triangular inequality it follows (with the exception of the L_0 -norm):

$$\|\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}\| \leq \lambda \|\mathbf{x}\| + (1 - \lambda) \|\mathbf{y}\|$$

5.4. Taylor Expansion

Definition 5.28 Taylor Expansion:

$$T_n(\mathbf{x}) = \sum_{i=0}^n \frac{1}{n!} f^{(i)}(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)^{(i)} \quad (5.55)$$

$$= f(\mathbf{x}_0) + f'(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2} f''(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)^2 + \mathcal{O}(x^3) \quad (5.56)$$

Definition 5.29 Incremental Taylor:
Goal: evaluate $T_n(\mathbf{x})$ (eq. (5.56)) at the point $\mathbf{x}_0 + \Delta \mathbf{x}$ in order to propagate the function $f(\mathbf{x})$ by $h = \Delta \mathbf{x}$:

$$T_n(\mathbf{x}_0 \pm h) = \sum_{i=0}^n \frac{h^i}{n!} f^{(i)}(\mathbf{x}_0) i^{-1} \quad (5.57)$$

$$= f(\mathbf{x}_0) \pm h f'(\mathbf{x}_0) + \frac{h^2}{2} f''(\mathbf{x}_0) \pm f'''(\mathbf{x}_0)(h)^3 + \mathcal{O}(h^4)$$

Note

If we chose $\Delta \mathbf{x}$ small enough it is sufficient to look only at the first two terms.

Definition 5.30 Multidimensional Taylor: Suppose $X \in \mathbb{R}^n$ is open, $\mathbf{x} \in X, f : X \mapsto \mathbb{R}$ and $f \in C^2$ then it holds that

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla_{\mathbf{x}} f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^\top H(\mathbf{x} - \mathbf{x}_0) \quad (5.58)$$

Definition 5.31 Argmax: The argmax of a function defined on a set D is given by:

$$\arg \max_{\mathbf{x} \in D} f(\mathbf{x}) = \{\mathbf{x} | f(\mathbf{x}) \geq f(\mathbf{y}), \forall \mathbf{y} \in D\} \quad (5.59)$$

Definition 5.32 Argmin: The argmin of a function defined on a set D is given by:

$$\arg \min_{\mathbf{x} \in D} f(\mathbf{x}) = \{\mathbf{x} | f(\mathbf{x}) \leq f(\mathbf{y}), \forall \mathbf{y} \in D\} \quad (5.60)$$

Corollary 5.13 Relationship $\arg \min \leftrightarrow \arg \max$:

$$\arg \min_{\mathbf{x} \in D} f(\mathbf{x}) = \arg \max_{\mathbf{x} \in D} -f(\mathbf{x}) \quad (5.61)$$

Property 5.8 Argmax Identities:

1. **Shifting:**
 $\forall \lambda \text{ const} \quad \arg \max_{\mathbf{x}} f(\mathbf{x}) = \arg \max_{\mathbf{x}} f(\mathbf{x}) + \lambda \quad (5.62)$

2. **Positive Scaling:**
 $\forall \lambda > 0 \text{ const} \quad \arg \max_{\mathbf{x}} f(\mathbf{x}) = \arg \max_{\mathbf{x}} \lambda f(\mathbf{x}) \quad (5.63)$

3. **Negative Scaling:**
 $\forall \lambda < 0 \text{ const} \quad \arg \max_{\mathbf{x}} f(\mathbf{x}) = \arg \min_{\mathbf{x}} \lambda f(\mathbf{x}) \quad (5.64)$

4. **Positive Functions:**
 $\forall \arg \max f(\mathbf{x}) > 0, \forall \mathbf{x} \in \text{dom}(f)$
 $\arg \max f(\mathbf{x}) = \arg \min \frac{1}{f(\mathbf{x})} \quad (5.65)$

5. **Strictly Monotonic Functions:** for all strictly monotonic increasing functions [def. 5.12] g it holds that:

$$\arg \max g(f(\mathbf{x})) = \arg \max f(\mathbf{x}) \quad (5.66)$$

Definition 5.33 Max: The maximum of a function f defined on the set D is given by:

$$\max_{x \in D} f(x) = f(x^*) \quad \text{with} \quad \forall x^* \in \arg \max_{x \in D} f(x) \quad (5.67)$$

Definition 5.34 Min: The minimum of a function f defined on the set D is given by:

$$\min_{x \in D} f(x) = f(x^*) \quad \text{with} \quad \forall x^* \in \arg \min_{x \in D} f(x) \quad (5.68)$$

Corollary 5.14 Relationship $\min \leftrightarrow \max$:

$$\min_{x \in D} f(x) = - \max_{x \in D} -f(x) \quad (5.69)$$

Property 5.9 Max Identities:

1. **Shifting:**

$$\forall \lambda \text{ const} \quad \max \{f(x) + \lambda\} = \lambda + \max f(x) \quad (5.70)$$

2. **Positive Scaling:**

$$\forall \lambda > 0 \text{ const} \quad \max \lambda f(x) = \lambda \max f(x) \quad (5.71)$$

3. **Negative Scaling:**

$$\forall \lambda < 0 \text{ const} \quad \max \lambda f(x) = \lambda \min f(x) \quad (5.72)$$

4. **Positive Functions:**

$$\forall \arg \max f(x) > 0, \forall x \in \text{dom}(f) \quad \max \frac{1}{f(x)} = \frac{1}{\min f(x)} \quad (5.73)$$

5. **Stricly Monotonic Functions:** for all strictly monotonic increasing functions^[def. 5.12] g it holds that:

$$\max g(f(x)) = g(\max f(x)) \quad (5.74)$$

Definition 5.35 Supremum: The supremum of a function defined on a set D is given by:

$$\sup_{x \in D} f(x) = \{y|y \geq f(x), \forall x \in D\} = \min_{y|y \geq f(x), \forall x \in D} y \quad (5.75)$$

and is the smallest value y that is equal or greater $f(x)$ for any $x \iff$ smallest upper bound.

Definition 5.36 Infimum: The infimum of a function defined on a set D is given by:

$$\inf_{x \in D} f(x) = \{y|y \leq f(x), \forall x \in D\} = \max_{y|y \leq f(x), \forall x \in D} y \quad (5.76)$$

and is the biggest value y that is equal or smaller $f(x)$ for any $x \iff$ largest lower bound.

Corollary 5.15 Relationship $\sup \leftrightarrow \inf$:

$$\sup_{x \in D} f(x) = - \inf_{x \in D} -f(x) \quad (5.77)$$

Note

The supremum/infimum is necessary to handle unbound function that seem to converge and for which the max/min does not exist as the argmax/argmin may be empty. E.g. consider $-e^x/e^x$ for which the max/min converges toward 0 but will never reached s.t. we can always choose a bigger $x \Rightarrow$ there exists no argmax/argmin \Rightarrow need to bound the functions from above/below \iff infimum/supremum.

Definition 5.37 Time-invariant system (TIS): A function f is called time-invariant, if shifting the input in time leads to the same output shifted in time by the same amount.

$$y(t) = f(x(t), t) \xrightarrow[\forall \tau]{\text{time-invariance}} y(t - \tau) = f(x(t - \tau), t) \quad (5.78)$$

Definition 5.38 Inverse Function $g = f^{-1}$:
A function g is the inverse function of the function $f : A \subset \mathbb{R} \rightarrow B \subset \mathbb{R}$ if

$$f(g(x)) = x \quad \forall x \in \text{dom}(g) \quad (5.79)$$

and

$$g(f(u)) = u \quad \forall u \in \text{dom}(f) \quad (5.80)$$

Property 5.10

Reflective Property of Inverse Functions: f contains (a, b) if and only if f^{-1} contains (b, a) .

The line $y = x$ is a symmetry line for f and f^{-1} .

Theorem 5.5 The Existence of an Inverse Function:
A function has an inverse function if and only if it is one-to-one.

Corollary 5.16 Inverse functions and strict monotonicity: If a function f is **strictly monotonic**^[def. 5.14] on its entire domain, then it is one-to-one and therefore has an inverse function.

6. Special Functions

6.1. The Gamma Function

Definition 5.39 The gamma function $\Gamma(\alpha)$: Is extension of the factorial function (??) to the real and complex numbers (with a positive real part):

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx \quad \Re(z) > 0 \quad (5.81)$$

$$\Gamma(n) \stackrel{n \in \mathbb{N}}{\iff} \Gamma(n) = (n-1)!$$

7. Proofs

Proof 5.4: lemma 5.1 for C^1 functions:

Let $g(t) \equiv f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))$ from the FToC (theorem 5.2) we know that:

$$\int_0^1 g'(t) dt = g(1) - g(0) = f(\mathbf{x}) - f(\mathbf{y})$$

It then follows from the reverse:

$$\begin{aligned} & |f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})| \\ & \stackrel{\text{Chain. R}}{\stackrel{\text{FToC}}{=}} \left| \int_0^1 \nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) dt - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \right| \\ & = \left| \int_0^1 (\nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) dt \right| \\ & = \left| \int_0^1 (\nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) dt \right| \\ & \stackrel{\text{C.S.}}{\leq} \left| \int_0^1 \|\nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(\mathbf{y})\| \cdot \|\mathbf{x} - \mathbf{y}\| dt \right| \\ & \stackrel{\text{eq. (5.30)}}{=} \left| \int_0^1 L \|\mathbf{y} + t(\mathbf{x} - \mathbf{y}) - \mathbf{y}\| \cdot \|\mathbf{x} - \mathbf{y}\| dt \right| \\ & = \left| L \|\mathbf{x} - \mathbf{y}\|^2 \int_0^1 t dt \right| = \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 \end{aligned}$$

Proof 5.5: ?? for C^2 functions:

$$f(\mathbf{y}) \stackrel{\text{Taylor}}{=} f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2} (\mathbf{y} - \mathbf{x})^\top \nabla^2 f(z) (\mathbf{y} - \mathbf{x})$$

Now we plug in $\nabla^2 f(\mathbf{x})$ and recover eq. (5.33):

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2} (\mathbf{y} - \mathbf{x})^\top L (\mathbf{y} - \mathbf{x})$$

Proof 5.6: theorem 5.3:

With the definition of convexity for a differentiable function (eq. (5.41)) it follows

$$\begin{aligned} & f(x) - f(y) + \nabla f(y)^\top (x - y) \geq 0 \\ & \Rightarrow |f(x) - f(y) + \nabla f(y)^\top (x - y)| \\ & \stackrel{\text{if eq. (5.41)}}{=} f(x) - f(y) + \nabla f(y)^\top (x - y) \end{aligned}$$

with lemma 5.1 and ^[def. 5.23] it follows theorem 5.3

Differential Calculus

1. Mean Value Theorem

Theorem 6.1 Mean Value Theorem: Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous function, differentiable on the open interval (a, b) , with $a < b$. Then there exist some $c \in (a, b)$ s.t.

$$f'(c) = \frac{f(b) - f(a)}{b - a} = \frac{1}{b - a} \int_a^b f(x) dx \quad (6.1)$$

2. The Product Rule

Rule 6.1 (Product /Leibniz Rule).

Let u, v be two differentiable functions $u, v \in \mathcal{C}^1$ then it holds that:

$$\frac{d(u(x)v(x))}{dx} = (uv)' = u'v + v'u \quad (6.2)$$

3. The Chain Rule

Formula 6.1 Generalized Chain Rule:

Let $\mathbf{F} : \mathbb{R}^n \rightarrow \mathbb{R}^k$ and $\mathbf{G} : \mathbb{R}^k \rightarrow \mathbb{R}^m$ be to general maps then it holds:

$$\frac{\partial (\mathbf{G} \circ \mathbf{F})}{\mathbb{R}^n \mapsto (\mathbb{R}^m \times k, \mathbb{R}^k \times n)} = \left(\frac{\partial \mathbf{G} \circ \mathbf{F}}{\partial \mathbf{F}} \right) \cdot \frac{\partial \mathbf{F}}{\partial \mathbf{x}} \quad \frac{\partial \mathbf{F}}{\mathbb{R}^n \mapsto \mathbb{R}^{k \times n}} \quad \frac{\partial \mathbf{G}}{\mathbb{R}^k \mapsto \mathbb{R}^{m \times k}} \quad (6.3)$$

4. Directional Derivative

5. Partial Differentiation

Definition 6.1 Partial Derivative:

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a real valued function, its partial derivative $\partial_i f : \mathbb{R}^n \mapsto \mathbb{R}$ is defined as the directional derivative?? along the coordinate axis of one of its variables:

$$\begin{aligned} \partial_i f(\mathbf{x}) &= \frac{\partial f}{\partial x_i} = D_{x_i} f = \lim_{h \rightarrow 0} \frac{f(\mathbf{x}, x_i \leftarrow x_i + h) - f(\mathbf{x})}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(x_1, \dots, x_i + h, \dots, x_n) - f(x_1, \dots, x_i, \dots, x_n)}{h} \end{aligned} \quad (6.4)$$

5.1. The Gradient

5.1.1. The Nabla Operator

Definition 6.2 Nabla Operator/Del ∇ : Given a cartesian coordinate system \mathbb{R}^n with coordinates x_1, \dots, x_n and associated unit vectors $\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_n$ its *del* operator is defined as:

$$\nabla = \sum_{i=1}^n \frac{\partial}{\partial x_i} \hat{\mathbf{e}}_i = \begin{bmatrix} \frac{\partial}{\partial x_1}(\mathbf{x}) \\ \frac{\partial}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_n}(\mathbf{x}) \end{bmatrix} \quad (6.5)$$

Definition 6.3 Gradient:

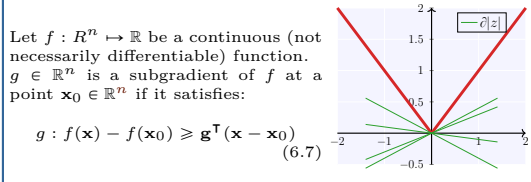
Given a *scalar valued* function $f : \mathbb{R}^n \mapsto \mathbb{R}$ its gradient $\nabla f : \mathbb{R}^n \mapsto \mathbb{R}^n$ is defined as vector \mathbb{R}^n of the partial derivatives^[def. 6.1] w.r.t. all coordinate axes:

$$\text{grad } f(\mathbf{x}) := \nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{bmatrix} = \left(\frac{\partial f}{\partial \mathbf{x}} \right)^T \quad (6.6)$$

5.1.2. The Subderivative

Definition 6.4

Subgradient



Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a continuous (not necessarily differentiable) function. $g \in \mathbb{R}^n$ is a subgradient of f at a point $\mathbf{x}_0 \in \mathbb{R}^n$ if it satisfies:

$$g : f(\mathbf{x}) - f(\mathbf{x}_0) \geq \mathbf{g}^T (\mathbf{x} - \mathbf{x}_0) \quad (6.7)$$

Definition 6.5

Subderivative

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a continuous (not necessarily differentiable) function. The subdifferential of f at a point $\mathbf{x}_0 \in \mathbb{R}^n$ is defined as the set of all possible subgradients^[def. 6.4] g :

$$\partial f(\mathbf{x}_0) \{ g : f(\mathbf{x}) - f(\mathbf{x}_0) \geq \mathbf{g}^T (\mathbf{x} - \mathbf{x}_0) \quad \forall \mathbf{x} \in \mathbb{R}^n \} \quad (6.8)$$

Heuristic

We can guess the sub derivative at a point by looking at all the slopes that are smaller then the graph.

5.2. The Jacobian

Definition 6.6

Jacobian/Jacobi Matrix

Given a *vector valued* function

$$\mathbf{f} : \mathbb{R}^n \mapsto \mathbb{R}^m \quad \text{its derivative} \quad \mathbf{J}_{\mathbf{f}} : \mathbb{R}^n \mapsto \mathbb{R}^{m \times n}$$

with components $\partial_{ij} \mathbf{f} = \partial_i f_j : \mathbb{R}^n \mapsto \mathbb{R}$ is a vector valued function defined as:

$$\begin{aligned} \mathbf{J}(\mathbf{f}(\mathbf{x})) &= \mathbf{J}_{\mathbf{f}}(\mathbf{x}) = \mathbf{Df} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}) = \frac{\partial (f_1, \dots, f_m)}{\partial (x_1, \dots, x_n)}(\mathbf{x}) \quad (6.9) \\ &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^T f_1 \\ \vdots \\ \nabla^T f_m \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \frac{\partial f_m}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial f_m}{\partial x_n}(\mathbf{x}) \end{bmatrix} \end{aligned}$$

Explanation 6.1. Rows of the Jacobian are transposed gradients^[def. 6.3] of the component functions f_1, \dots, f_m .

Corollary 6.1 :

6. Second Order Derivatives

Definition 6.7 Second Order Derivative $\frac{\partial^2}{\partial x_i \partial x_j}$:

Theorem 6.2

Symmetry of second derivatives/Schwartz's Theorem: Given a continuous and twice differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ then its second order partial derivatives commute:

$$\frac{\partial}{\partial x_i} \frac{\partial f}{\partial x_j} = \frac{\partial}{\partial x_j} \frac{\partial f}{\partial x_i}$$

6.1. The Hessian

Definition 6.8 Hessian Matrix:

Given a function $f : \mathbb{R} \mapsto \mathbb{R}^n$ its Hessian $\mathbb{R}^{n \times n}$ is defined as:

$$\mathbf{H}(\mathbf{f})(\mathbf{x}) = \mathbf{H}_{\mathbf{f}}(\mathbf{x}) = \mathbf{J}(\nabla \mathbf{f}(\mathbf{x}))^T \quad (6.10)$$

$$= \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{x}) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x}) & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\mathbf{x}) & \dots & \frac{\partial^2 f}{\partial x_n^2}(\mathbf{x}) \end{bmatrix}$$

and it corresponds to the Jacobian of the Gradient.

Due to the differentiability and theorem 6.2 it follows that the Hessian is (if it exists):

- Symmetric
- Real

Corollary 6.2 Eigenvector basis of the Hessian: Due to the fact that the Hessian is real and symmetric we can decompose it into a set of real eigenvalues and an orthogonal basis of eigenvectors $\{(\lambda_1, \mathbf{v}_1), \dots, (\lambda_n, \mathbf{v}_n)\}$.

Not let \mathbf{d} be a directional unit vector then the second derivative in that direction is given by:

$$\mathbf{d}^T \mathbf{H} \mathbf{d} \iff \mathbf{d}^T \sum_{i=1}^n \lambda_i \mathbf{v}_i \iff \text{if } \mathbf{d} = \mathbf{v}_j \quad \mathbf{d}^T \lambda_j \mathbf{v}_j$$

- The eigenvectors that have smaller angle with \mathbf{d} have bigger weight/eigenvalues
- The minimum/maximum eigenvalue determines the minimum/maximum second derivative

7. Extrema

Definition 6.9 Critical/Stationary Point: Given a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, that is differentiable at a point \mathbf{x}_0 then it is called a **critical point** if the functions derivative vanishes at that point:

$$f'(\mathbf{x}_0) = 0 \iff \nabla_{\mathbf{x}} f(\mathbf{x}_0) = 0$$

Corollary 6.3 Second Derivative Test $f : \mathbb{R} \mapsto \mathbb{R}$:

Suppose $f : \mathbb{R} \mapsto \mathbb{R}$ is twice differentiable at a stationary point x ^[def. 6.9] then it follows that:

- $f''(x) > 0 \iff \begin{matrix} f'(x + \epsilon) > 0 & \text{slope points uphill} \\ f'(x - \epsilon) < 0 & \text{slope points downhill} \end{matrix}$
 $f(x)$ is a local minimum
- $f''(x) < 0 \iff \begin{matrix} f'(x + \epsilon) > 0 & \text{slope points downhill} \\ f'(x - \epsilon) < 0 & \text{slope points uphill} \end{matrix}$
 $f(x)$ is a local maximum

$\epsilon > 0$ sufficiently small enough

Corollary 6.4 Second Derivative Test $f : \mathbb{R}^n \mapsto \mathbb{R}$:

Suppose $f : \mathbb{R}^n \mapsto \mathbb{R}$ is twice differentiable at a stationary point \mathbf{x} ^[def. 6.9] then it follows that:

- If \mathbf{H} is **p.d** $\iff \forall \lambda_i > 0 \in \mathbf{H} \rightarrow f(\mathbf{x})$ is a local min.
- If \mathbf{H} is **n.d** $\iff \forall \lambda_i < 0 \in \mathbf{H} \rightarrow f(\mathbf{x})$ is a local max.
- If $\exists \lambda_i > 0 \in \mathbf{H}$ and $\exists \lambda_i < 0 \in \mathbf{H}$ then \mathbf{x} is a local maximum in one cross section of f but a local minimum in another
- If $\exists \lambda_i = 0 \in \mathbf{H}$ and all other eigenvalues have the same sign the test is inclusive as it is inconclusive in the cross section corresponding to the zero eigenvalue.

Note

If \mathbf{H} is positive definite for a minima \mathbf{x}^* of a *quadratic* function f then this point must be a global minimum of that function.

8. Proofs

Proof 6.1: Definition 6.4 $f(\mathbf{x}) \geq f(\mathbf{x}_0) + \mathbf{g}^T (\mathbf{x} - \mathbf{x}_0) \quad \forall \mathbf{x} \in \mathbb{R}^n$ corresponds to a line (see formula 5.1) at the point \mathbf{x}_0 with slope \mathbf{g}^T . Thus we search for all lines with smaller slope then function graph.

9. Examples

Example 6.1 Subderivatives Absolute Value Function

$|x|$: $f : \mathbb{R} \mapsto \mathbb{R}$ with $f(x) = |x|$ at the point $x = 0$ it holds:

$$f(x) - f(0) \geq gx \iff \text{the interval } [-1; 1]$$

For $x \neq 0$ the subgradient is equal to the gradient. Thus it follows for the subderivatives/differentials:

$$\partial |x| = \begin{cases} -1 & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Integral Calculus

Theorem 7.1 Important Integral Properties:

Addition $\int_a^b f(x) \, dx = \int_a^c f(x) \, dx + \int_c^b f(x) \, dx$ (7.1)

Reflection $\int_a^b f(x) \, dx = - \int_b^a f(x) \, dx$ (7.2)

Translation $\int_a^b f(x) \, dx \stackrel{u:=x\pm c}{=} \int_{a\pm c}^{b\pm c} f(x \mp c) \, dx$ (7.3)

f Odd $\int_{-a}^a f(x) \, dx = 0$ (7.4)

f Even $\int_{-a}^a f(x) \, dx = 2 \int_0^a f(x) \, dx$ (7.5)

Proof 7.1: eqs. (7.4) and (7.5)

$$\begin{aligned} I &:= \int_{-a}^a f(x) \, dx = \int_{-a}^0 f(x) \, dx + \int_0^a f(x) \, dx \\ &\stackrel{t=-x}{=} \int_a^0 f(-x) \, dx + \int_0^a f(x) \, dx \\ &= \int_0^a f(-x) + f(x) \, dx = \begin{cases} 0 & \text{if } f \text{ odd} \\ 2I & \text{if } f \text{ even} \end{cases} \end{aligned}$$

Definition 7.1 Integration by Parts:

$$\int_a^b u \, dv = uv \Big|_a^b - \int_a^b v \, du \tag{7.6}$$

1. Integral Theorems

1.1. Greens Identities

Theorem 7.2 Greens First Identity:

Let $\bar{\Omega} = \Omega \cup \partial\Omega$, for all vector fields $\mathbf{j} \in (C^1_{\text{pw}}(\bar{\Omega}))^d$ and scalar functions $v \in C^1_{\text{pw}}(\bar{\Omega})$ it holds:

$$\int_{\Omega} \mathbf{j}^\top \text{grad } v \, d\mathbf{x} = - \int_{\Omega} \text{div } \mathbf{j} v \, d\mathbf{x} + \int_{\partial\Omega} \mathbf{j}^\top \mathbf{n} v \, dS \tag{7.7}$$

add multidimensional product rule and gauss theorem from NPDE

Differential Equations

Definition 7.2

Differential Operator:

A differential operator \mathcal{L} is a mapping of a suitable function space onto another function space, involving only values of the function argument and its derivatives in the same point:
 $\mathcal{L} : C^n(\Omega) \mapsto C^k(\Omega), \quad k < n$

Note: \mathcal{L} is a differential operator of order $k - n$.

Definition 7.3 Linear Differential Operator:

Is a differential operator \mathcal{L} that satisfies:
 $\mathcal{L}(\alpha u + \beta v) = \alpha \mathcal{L}(u) + \beta \mathcal{L}(v) \quad \forall \alpha, \beta \in \mathbb{R} \quad (7.8)$

Ordinary Differential Quations

Partial Differential Equations (PDE)s

Definition 9.1 Partial Differential Equation:

Let $\mathbf{u} = \mathbf{u}(x_1, \dots, x_n) : \mathbb{R}^k \mapsto \mathbb{R}$ be an unknown function depending on $\mathbf{x} = (x_1, \dots, x_k)$ and let f be a known function.

The known function \mathcal{F} , depending on differentials of the non-known function \mathbf{u} is called a Partial Differential equation:

$$\mathcal{F}\left(\mathbf{u}, \frac{\partial \mathbf{u}}{\partial x_1}, \dots, \frac{\partial \mathbf{u}^n}{\partial x_i}, \dots, \frac{\partial \mathbf{u}^n}{\partial x_j}, \dots, f\right) = \mathcal{F}(\mathbf{u}, D\mathbf{u}, \dots, D^n \mathbf{u}, f) = 0$$

or
$$\mathcal{L}(\mathbf{u}) = f \quad \text{in } \Omega \quad (9.1)$$

Corollary 9.1 Dependent Variables:

$$\mathbf{u} : \mathbb{R}^k \mapsto \mathbb{R}^l \quad (9.2)$$

Corollary 9.2 Independent Variables:

$$\mathbf{x} = (x_1, \dots, x_k) \quad (9.3)$$

Definition 9.2 Order

Is the highest partial derivative that appears in a PDE.

1. Algebraic Types

1.1. Linearity

Definition 9.3

Linear PDEs:

A linear PDE naturally defines a linear operator [def. 7.3]. A linear PDE must be linear regarding the unknown function \mathbf{u} . In other words all dependent variables \mathbf{u} and their corresponding derivatives depend only on the independent variables x_1, x_2, \dots, x_m :

$$a(x, y)\mathbf{u}_x + b(x, y)\mathbf{u}_y + c(x, y)\mathbf{u} = d(x, y) \quad (9.4)$$

Definition 9.4

Semilinear PDEs:

Are PDEs whose coefficients of the highest order n -terms are functions depending only on the independent variables but not onto the dependent variables \mathbf{u} or their derivatives.

Thus the PDE is linear regarding to the highest order terms:

$$a(x, y)\mathbf{u}_x + b(x, y)\mathbf{u}_y = c(x, y, \mathbf{u}) \quad (9.5)$$

Definition 9.5

Quasilinear PDEs:

Are PDEs whose coefficients of the highest order (n) terms are functions only depending on the independent variables and on the dependent variables \mathbf{u} and their derivatives up to an order $m < n$, that is smaller than the highest order terms n :

$$a(x, y, \mathbf{u})\mathbf{u}_x + b(x, y, \mathbf{u})\mathbf{u}_y = c(x, y, \mathbf{u}) \quad (9.6)$$

Definition 9.6

Fully Non-linear PDEs:

Are PDEs where all terms of the highest order n are non-linear:

$$a(x, y, \mathbf{u}, \mathbf{u}')\mathbf{u}_x + b(x, y, \mathbf{u}, \mathbf{u}')\mathbf{u}_y = c(x, y, \mathbf{u}) \quad (9.7)$$

Note: $\neg(\text{Quasilinear} \Leftrightarrow \text{Fully Nonlinear})$

1.2. Homogeneity

Definition 9.7 Homogeneous

All terms depend on \mathbf{u} or on derivatives of \mathbf{u} .

Definition 9.8 Non-Homogeneous

There exists non-zero terms f that do not depend on \mathbf{u} or on derivatives of \mathbf{u} .

1.3. Constant Coefficients

Definition 9.9 PDEs with Constant Coefficients:

Is a PDE whose coefficients a, b, c, \dots are constants i.e. independent variables.

1.4. 2nd-Order Linear PDEs in two variables

Definition 9.10

2nd-Order Linear PDEs in two Variables:

$$\mathcal{L}(\mathbf{u}) = a\mathbf{u}_{xx} + 2b\mathbf{u}_{xy} + c\mathbf{u}_{yy} + d\mathbf{u}_x + e\mathbf{u}_y + f\mathbf{u} = g \quad (9.8)$$

where a, b, \dots, g are functions depending on x and y .

Definition 9.11 Principal Part:

Is the operator \mathcal{L}_0 , that consists of the second-(=highest) order parts of \mathcal{L} :
$$\mathcal{L}_2(\mathbf{u}) := a\mathbf{u}_{xx} + 2b\mathbf{u}_{xy} + c\mathbf{u}_{yy}$$

Definition 9.12 PDEs Discriminante:

Is defined by:

$$\delta(\mathcal{L}) := -\det \begin{pmatrix} a & b \\ b & c \end{pmatrix} = b^2 - ac \quad (9.9)$$

Explanation 9.1.

It turns out that many fundamental properties of the solution of eq. (9.8) are determined by its principal part, or rather by the sign of the discriminant $\delta(\mathcal{L})$.

Definition 9.13

Parabolic PDEs:

Let [def. 9.10] be a PDE defined on $\Omega \subset \mathbb{R}^2$, then the PDE is called hyperbolic if:

$$\delta(\mathcal{L}) = b^2 - ac = 0 \quad (9.10)$$

Definition 9.14

Hyperbolic PDEs:

Let [def. 9.10] be a PDE defined on $\Omega \subset \mathbb{R}^2$, then the PDE is called hyperbolic if:

$$\delta(\mathcal{L}) = b^2 - ac > 0 \quad (9.11)$$

Definition 9.15

Parabolic PDEs:

Let [def. 9.10] be a PDE defined on $\Omega \subset \mathbb{R}^2$, then the PDE is called elliptic if:

$$\delta(\mathcal{L}) = b^2 - ac < 0 \quad (9.12)$$

Explanation 9.2.

The reason for this categorization are normal quadratic equations in two variables:

$$Ax^2 + By^2 + Cxy + Dx + Ey + f = 0$$

If $B^2 - 4AC = 0 \Leftrightarrow$ the equation is a parabola.

If $B^2 - 4AC > 0 \Rightarrow$ the equation is a hyperbola.

If $B^2 - 4AC < 0 \Rightarrow$ the equation is an ellipse.

2. Method Of Characteristics

Is a method that makes use of geometrical aspects in order to solve 1st-order PDEs with two variables by constructing integral surfaces and can be used to solve PDEs of the type:

Linear: $a(x, y)\mathbf{u}_x + b(x, y)\mathbf{u}_y = c(x, y) \quad (9.13)$

Semilin.: $a(x, y)\mathbf{u}_x + b(x, y)\mathbf{u}_y = c(x, y, \mathbf{u}) \quad (9.14)$

Quasilin.: $a(x, y, \mathbf{u})\mathbf{u}_x + b(x, y, \mathbf{u})\mathbf{u}_y = c(x, y, \mathbf{u}) \quad (9.15)$

Formula 9.1 Method of Characteristics:

$$x := x(r; s) \quad y := y(r; s) \quad z := u(r; s)$$

Parameter.: $\lambda(r; s) := x(r; s)\mathbf{e}_x + y(r; s)\mathbf{e}_y + z(r; s)\mathbf{e}_z$

$$\frac{\partial \lambda}{\partial r}(r; s) = (a, b, c)$$

$$v := v(x(r; s), y(r; s), z(r; s))$$

E.g.
$$\frac{\partial x}{\partial r}(r; s) = \dot{x} = a(\lambda_s(r))$$

$$\frac{\partial y}{\partial r}(r; s) = \dot{y} = b(\lambda_s(r))$$

$$\frac{\partial z}{\partial r}(r; s) = \dot{z} = c(\lambda_s(r))$$

Compact:

$$\dot{x} = a(x, y, u) \quad \dot{y} = b(x, y, u) \quad \dot{u} = c(x, y, u)$$

I.C.: $x(0; s) = x_0(s) \quad y_0(0; s) = y_0(s) \quad u_0(0; s) = u_0(s)$

Definition 9.16 Integral Surface

An function $\phi : \mathbb{R}^3 \mapsto \mathbb{R}$ is an integral surface of a vector field $\mathbf{V} : \mathbb{R}^3 \mapsto \mathbb{R}^3$ if ϕ is a surface that has in every point a tangent plane containing a vector $\mathbf{v} = (a \ b \ c)$ of \mathbf{V} .

Corollary 9.3 PDEs and Integral Surfaces:

The solution of a PDE $\mathbf{u}(x, y)$ can be thought of as an integral surface:

$$z = u(x, y) \quad \text{or implicitly} \quad \phi(x, y, z) = u(x, y) - z \quad (9.16)$$

Explanation 9.3

(Integral Surface and PDEs).

The solution $\mathbf{u}(x, y)$ of eq. (9.13) can be sought of as an surface $z = \mathbf{u}(x, y)$ in \mathbb{R}^3 or in implicit form $\phi(x, y, z) := \mathbf{u}(x, y) - z$.

Let: $\mathbf{n}(x, y) := \text{grad } \phi = \begin{pmatrix} \phi_x \\ \phi_y \\ \phi_z \end{pmatrix} = \begin{pmatrix} \mathbf{u}_x \\ \mathbf{u}_y \\ -1 \end{pmatrix}$

and

Let $\mathbf{V} := \begin{pmatrix} a(x, y) \\ b(x, y) \\ c(x, y) \end{pmatrix}$

be a vector field $\mathbb{R}^3 \mapsto \mathbb{R}^3$ and

$$\mathbf{n}(x, y) := \text{grad } \phi = \begin{pmatrix} \phi_x \\ \phi_y \\ \phi_z \end{pmatrix} = \begin{pmatrix} \mathbf{u}_x \\ \mathbf{u}_y \\ -1 \end{pmatrix}$$

Idea: we can rewrite eq. (9.13) as:

$$\left\langle \begin{pmatrix} a & b & c \end{pmatrix}^\top, \nabla \phi(x, y, z) \right\rangle = \left\langle \begin{pmatrix} a(x, y) \\ b(x, y) \\ c(x, y) \end{pmatrix}, \begin{pmatrix} \mathbf{u}_x \\ \mathbf{u}_y \\ -1 \end{pmatrix} \right\rangle = 0$$

Geometric Interpretation:

\mathbf{v} is orthogonal to the normal \mathbf{n} for all points $(x, y, \mathbf{u}(x, y))$.

Hence every vector $\mathbf{v} = (a \ b \ c)^\top$ lies in the tangent plane containing ϕ .

Consequently in order to find a surface ϕ (and thus also a solution \mathbf{u}), we need to search for ϕ s.t. the vector \mathbf{v} lies in the tangent plane for every possible point of ϕ .

Idea

We first simplify the task and start by constructing/finding integral curves λ and then we construct the integral surface ϕ out of this curves.

3. Linear Equations

Definition 9.17

Characteristic/Integral Curve

$\lambda_s(r) = \lambda(r; s)$: Given a vector field \mathbf{V} an integral curve $\lambda(r)$ of that vector field, is a curve parameterized by parameter r :

$$\lambda(r) := x(r)\mathbf{e}_x + y(r)\mathbf{e}_y + z(r)\mathbf{e}_z = \begin{pmatrix} x(r) \\ y(r) \\ z(r) \end{pmatrix} \quad (9.17)$$

s.t. at each point r of the curve a vector \mathbf{v} of the vector field:

$$\mathbf{v} = \begin{pmatrix} a(x(r), y(r)) \\ b(x(r), y(r)) \\ c(x(r), y(r)) \end{pmatrix} \in \mathbf{V} \quad (9.18)$$

is tangent to the curve:

$$\frac{d\lambda(r)}{dr} = \mathbf{V}(\lambda(r)) = \begin{pmatrix} a(x(r), y(r)) \\ b(x(r), y(r)) \\ c(x(r), y(r)) \end{pmatrix} = \begin{pmatrix} a(\lambda(r)) \\ b(\lambda(r)) \\ c(\lambda(r)) \end{pmatrix} \quad (9.19)$$

Definition 9.18 Characteristic Equations:

The set of ordinary differential equations of a PDE arising from Equation (9.19) are called characteristic equations:

$$\frac{dx(r)}{dr} = \dot{x} = \underline{a(\lambda(r))} = a(r) \quad (9.20)$$

$$\frac{\partial y(r)}{\partial r} = \dot{y} = \underline{b(\lambda(r))} = b(r) \quad (9.21)$$

$$\frac{\partial z(r)}{\partial r} = \dot{z} = \underline{c(\lambda(r))} = c(r) \quad (9.22)$$

Problem: in order to get a unique solution we need to specify initial conditions.

Idea: If a characteristic has an arbitrary point in common with the integral surface ϕ then the whole characteristic λ will lie in the integral surface.

Proof 9.1: Let: $\phi(\lambda(r)) = u(x(r), y(r)) - z(r)$

$$\Rightarrow \frac{d\phi}{dr} = u_x \frac{dx}{dr} + u_y \frac{dy}{dr} - 1 \frac{dz}{dr} =$$

$$= \begin{pmatrix} u_x \\ u_y \\ -1 \end{pmatrix} \begin{pmatrix} a(x(r), y(r)) \\ b(x(r), y(r)) \\ c(x(r), y(r)) \end{pmatrix} = \begin{pmatrix} u_x \\ u_y \\ -1 \end{pmatrix} \dot{\lambda}(r) = 0$$

Thus: $\phi(\lambda(r_0)) = 0 \Leftrightarrow \phi(\lambda(r)) = 0, \quad \forall r$

Definition 9.19

Characteristic (Curve)

$\lambda_s(r) = \lambda(r; s)$: is an integral curve of the vector field \mathbf{V} that is uniquely determined by a parameter s .

Consequence:

For every characteristic s we need to specify one initial point on the integral surface in order to have all the characteristics lie within the integralsurface.

Idea: we define another curve $\Gamma(s)$ on the integralsurface that transverses all the characteristic curves $\lambda_s(r)$ transversal (=angle between $\Gamma(s)$ and $\lambda_s(r)$ is never zero $\Leftrightarrow \Gamma(s) \nparallel \lambda_s(r)$).

Definition 9.20 Initial Condition:

$s \mapsto \Gamma(s), \quad \Gamma : \mathbb{R} \mapsto \mathbb{R}^3$

$$\lambda_s(r) = \begin{pmatrix} x_s(r) \\ y_s(r) \\ z_s(r) \end{pmatrix}, \quad \Gamma(s) = \begin{pmatrix} x_0(s) \\ y_0(s) \\ z_0(s) \end{pmatrix} \quad \lambda_s(0) \stackrel{!}{=} \Gamma(s)$$

$$\Rightarrow \underline{x_s(0)} = \underline{x_0(s)} \quad \underline{y_s(0)} = \underline{y_0(s)} \quad \underline{z_s(0)} = \underline{z_0(s)}$$

Definition 9.21

Projected Characteristic Curves

$\gamma(\tau)$: Are curves in the plane of the independent variables of our PDE, along which u is constant or satisfies certain conditions. If u is constant along $g(\tau)$ then the initial data is simply propagated along those characteristic curves:

$$\frac{d}{d\gamma} u(\gamma(\tau), \tau) = 0 \Leftrightarrow u(\gamma(\tau), \tau) = u_0(\gamma(\tau)) \quad (9.23)$$

Hint: If the PDE is linear, then the two first characteristics do not depend on u and can be solved directly, u will then be constant along those characteristics:

$$\begin{aligned} &a(x,y)\mathbf{u}_x + b(x,y)\mathbf{u}_y = c(x,y) \\ \frac{dx}{dr} = a \quad \frac{dy}{dr} = b \quad \frac{du}{dr} = c \quad &\text{implies} \quad \frac{dy}{dx} = \frac{b(x,y)}{a(x,y)} \end{aligned}$$

Hint: If we divide the PDE by *a* we have to solve a PDE less, beacause the first ODE will allways be:

$$\dot{x} = 1 \Rightarrow \quad x = r \Rightarrow \quad x_s(r) = x_0(s)$$

4. Quasilinear Equations

Solving Quasilinear Equations

$$\begin{aligned} &a(x,y,u)\mathbf{u}_x \quad + b(x,y,u)\mathbf{u}_y \quad = c(x,y,u) \\ u|_{\Gamma}(r,s) &= \phi(s) \\ \frac{dx}{dr} = a(x,y,u) \quad \frac{dy}{dr} = b(x,y,u) \quad \frac{du}{dr} &= c(x,y,u) \\ x_s(0) = x_0(s) \quad y_s(0) = y_0(s) \quad z_s(0) &= \phi(s) \end{aligned}$$

Results

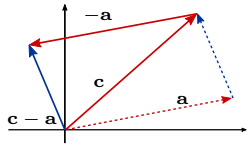
Now the projected characteristic curves may depend on u as well as on x,y. **Thus** the first two characteristics are no longer decoupled form the third one.

1. We may get projected characteristic curves crossing themselves.
2. u is no longer constant along the projected characteristic curves, rather the PDE reduces to an ODE satisfying certain conditions along this curves.

Linear Algebra

1. Vectors

Definition 10.1 Vector Subtraction:



$$\mathbf{b} = \mathbf{c} - \mathbf{a} \quad (10.1)$$

2. Linear Systems of Equations

2.1. Gaussian Elimination

2.1.1. Rank

Definition 10.2 Matrix Rank

The ranks of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is defined as the dimension^[def. 10.13] of the vector space spanned^[def. 10.9] by its row or column vectors:

$$\begin{aligned} \text{rank}(\mathbf{A}) &= \dim(\{\mathbf{a}_{:,1}, \dots, \mathbf{a}_{:,n}\}) \\ &= \dim(\{\mathbf{a}_{1,:}, \dots, \mathbf{a}_{m,:}\}) \\ \text{def. 10.50} \quad &= \dim(\mathfrak{R}(\mathbf{A})) \end{aligned} \quad (10.2)$$

Corollary 10.1 :

- The column-and row-ranks of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ are equal.
- The rank of a non-symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is limited by the smaller dimension:

$$\text{rank}(\mathbf{A}) \leq \min\{n, m\} \quad (10.3)$$

Property 10.1 Rank of Matrix Product: Let $\mathbf{A} \in \mathbb{R}^{m,n}$ and $\mathbf{B} \in \mathbb{R}^{n,p}$ then the rank of the matrix product is limited:

$$\text{rank}(\mathbf{AB}) \leq \min\{\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})\} \quad (10.4)$$

Rank-1 Matrix

Definition 10.3 Rank-1 Matrix:

Is a matrix of rank one. A tensor product of two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ results in a rank one matrix:

$$\mathbf{uv}^T = \mathbf{A} \in \mathbb{R}^{n,n} \quad (10.5)$$

Definition 10.4 Rank-1 Modification/Update:

Adding a rank-1 matrix to another matrix is called rank-1 modification:

$$\mathbf{X} = \mathbf{X} + \mathbf{uv}^T \quad (10.6)$$

3. Sparse Linear Systems

Definition 10.5 Sparse Matrix

$\mathbf{A} \in \mathbb{K}^{m,n}$, $m, n \in \mathbb{N}_{>0}$: A matrix \mathbf{A} is sparse if:

$$\begin{aligned} \text{nnz}(\mathbf{A}) &\ll mn & \mathbf{A} &\in \mathbb{K}^{m,n}, m, n \in \mathbb{N}_{>0} \\ \text{nnz} &:= \#\{(i, j) \in \{1, \dots, m\} \times \{1, \dots, n\} : a_{i,j} \neq 0\} \end{aligned} \quad (10.7)$$

4. Vector Spaces

4.1. Vector Space

Definition 10.6 Vector Space: TODO

4.2. Vector Subspace

Definition 10.7 Vector Subspaces:

A non-empty subset U of a \mathbb{K} -vector space \mathcal{V} is called a subspace of \mathcal{V} if it satisfies:

$$\begin{aligned} \mathbf{u}, \mathbf{v} \in U &\implies \mathbf{u} + \mathbf{v} \in U & (10.8) \\ \mathbf{u} \in U &\implies \lambda \mathbf{u} \in U & \forall \lambda \in \mathbb{K} \quad (10.9) \end{aligned}$$

Definition 10.8 Linear combination:

Let $X = \{\mathbf{v}_1, \dots, \mathbf{v}_n\} \subset \mathcal{V}$ be a non-empty and finite subset of vectors of an \mathbb{K} -vector space \mathcal{V} . A linear combination of X is a combination of the vectors defined as:

$$\mathbf{v} = \sum_{i=1}^n \lambda_i \mathbf{v}_i = \alpha_1 \mathbf{v}_1 + \dots + \alpha_n \mathbf{v}_n \quad \alpha_i \in \mathbb{K} \quad (10.10)$$

Definition 10.9

Span/Linear Hull

Is the set of all possible linear combinations^[def. 10.8] of finite set $X = \{\mathbf{v}_1, \dots, \mathbf{v}_n\} \subset \mathcal{V}$ of a \mathbb{K} vector space \mathcal{V} :

$$\langle X \rangle = \text{span}(X) = \left\{ \mathbf{v} \mid \sum_{i=1}^n \alpha_i \mathbf{v}_i, \forall \alpha_i \in \mathbb{K} \right\} \quad (10.11)$$

Definition 10.10 Generating Set: A generating set of vectors $X = \{\mathbf{v}_1, \dots, \mathbf{v}_m\} \in \mathcal{V}$ of a vector spaces \mathcal{V} is a set of vectors that span^[def. 10.9] \mathcal{V} :

$$\text{span}(\mathbf{v}_1, \dots, \mathbf{v}_m) = \mathcal{V} \quad (10.12)$$

Explanation 10.1 (Definition 10.10).

The generating set of vector space (or set of vectors) \mathcal{V} i.e. \mathbb{R}^n is a subset $X = \{\mathbf{v}_1, \dots, \mathbf{v}_m\} \subset \mathcal{V}$ s.t. every element of \mathcal{V} can be produced by span(X).

Definition 10.11 Linear Independence: A set of vector $\{\mathbf{v}_1, \dots, \mathbf{v}_n\} \in \mathcal{V}$ is called linear independent if the satisfy:

$$\mathbf{v} = \sum_{i=1}^n \lambda_i \mathbf{v}_i = \mathbf{0} \iff \alpha_1 = \dots = \alpha_n = 0 \quad (10.13)$$

Corollary 10.2 : A set of vector $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \in \mathcal{V}$ is called linear independent, if for every subset $X = \mathbf{x}_1, \dots, \mathbf{x}_m \subseteq \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ it holds that:

$$\langle X \rangle \subsetneq \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \quad (10.14)$$

4.3. Basis

Definition 10.12 Basis \mathfrak{B} :

A subset $\mathfrak{B} = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ of a \mathbb{K} -vector space \mathcal{V} is called a basis of \mathcal{V} if:

$$\langle \mathfrak{B} \rangle = \mathcal{V} \quad \text{and} \quad \mathfrak{B} \text{ is a linear independent generating set} \quad (10.15)$$

Corollary 10.3 : The unit vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ build a standard basis of the \mathbb{R}^n .

Corollary 10.4 Basis Representation:

Let \mathfrak{B} be a basis of a \mathbb{K} -vector space \mathcal{V} , then it holds that every vector $\mathbf{v} \in \mathcal{V}$ can be represented as a linear combination^[def. 10.8] of \mathfrak{B} by a unique set of coefficients α_i :

$$\mathbf{v} = \sum_{i=1}^n \alpha_i \mathbf{b}_i \quad \begin{matrix} \alpha_1, \dots, \alpha_n \in \mathbb{K} \\ \mathbf{b}_1, \dots, \mathbf{b}_n \in \mathfrak{B} \end{matrix} \quad (10.16)$$

4.3.1. Dimensionality

Definition 10.13 Dimension of a vector space $\dim(\mathcal{V})$: Let \mathcal{V} be a vector space. The dimension of \mathcal{V} is defined as the number of necessary basis vectors $\mathfrak{B} = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ in order to span \mathcal{V} :

$$\dim(\mathcal{V}) := |\mathfrak{B}| = n \in \mathbb{N}_0 \quad (10.17)$$

Corollary 10.5 : n -linearly independent vectors of a \mathbb{K} -vector space \mathcal{V} with finite dimension n constitute a basis.

Note

If \mathcal{V} is infinite $\dim(\mathcal{V}) = \infty$.

4.4. Affine Subspaces

Definition 10.14 Affine Subspaces: Given a \mathbb{K} -vector space \mathcal{V} of dimension $\dim(\mathcal{V}) \geq 2$ a sub vector space^[def. 10.7] U of \mathcal{V} defined as:

$$\mathcal{W} := \mathbf{v} + U = \{\mathbf{v} + \mathbf{x} \mid \mathbf{x} \in U\} \quad \mathbf{v} \in \mathcal{V} \quad (10.18)$$

Corollary 10.6 Direction: The sub vector spaces U are called directions of \mathcal{V} and it holds:

$$\dim(\mathcal{W}) := \dim(U) \quad (10.19)$$

4.4.1. Hyperplanes

Definition 10.15 Hyperplane

A hyperplane is a $d-1$ dimensional subspace of an d -dimensional ambient space that can be specified by the hess normal form^[def. 10.16]:

$$\mathcal{H} = \{\mathbf{x} \in \mathbb{R}^d \mid \hat{\mathbf{n}}^T \mathbf{x} - d = 0\} \quad (10.20)$$

Corollary 10.7 Half spaces: A hyperplane $\mathcal{H} \in \mathbb{R}^{d-1}$ separates its d -dimensional ambient space into two half spaces:

$$\mathcal{H}^+ = \{x \in \mathbb{R}^d \mid \hat{\mathbf{n}}^T \mathbf{x} + b > 0\} \quad (10.21)$$

$$\mathcal{H}^- = \{x \in \mathbb{R}^d \mid \hat{\mathbf{n}}^T \mathbf{x} + b < 0\} = \mathbb{R}^d - \mathcal{H}^+ \quad (10.22)$$

Notes

Hyperplanes in \mathbb{R}^2 are lines and hyperplanes in \mathbb{R}^3 are planes.

Hess Normal Form

Definition 10.16 Hess Normal Form: Is an equation to describe hyperplanes^[def. 10.15] in \mathbb{R}^d :

$$\mathbf{r}^T \hat{\mathbf{n}} - d = 0 \iff \hat{\mathbf{n}}^T (\mathbf{r} - \mathbf{r}_0) \quad \mathbf{r}_0 := \mathbf{r}^T d \geq 0 \quad (10.23)$$

where all points described by the vector $\mathbf{r} \in \mathbb{R}^d$, that satisfy this equations lie on the hyperplane.

Note

The direction of the unit normal vector is usually chosen s.t. $\mathbf{r}^T \hat{\mathbf{n}} \geq 0$.

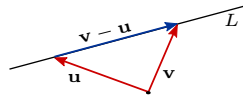
4.4.2. Lines

Definition 10.17 Lines: Lines are a set^[def. 1.1] of the form:

$$L = \mathbf{u} + \mathbb{K} \mathbf{v} = \{\mathbf{u} + \lambda \mathbf{v} \mid \lambda \in \mathbb{K}\} \quad \mathbf{u}, \mathbf{v} \in \mathcal{V}, \mathbf{v} \neq \mathbf{0} \quad (10.24)$$

Two Point Formula

Definition 10.18 Two Point Formula:



$$L = \mathbf{u} + \mathbb{K} \mathbf{v} \quad (10.25)$$

4.4.3. Planes

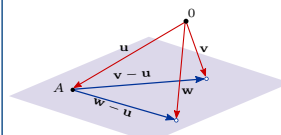
Definition 10.19 Planes: Planes are sets defined as:

$$E = \mathbf{u} + \mathbb{K} \mathbf{v} + \mathbb{K} \mathbf{w} = \{\mathbf{u} + \lambda \mathbf{v} + \mu \mathbf{w} \mid \lambda, \mu \in \mathbb{K}\} \quad (10.26)$$

$$\mathbf{u}, \mathbf{w} \in \mathcal{V} \quad \text{s.t. } \mathbf{v}, \mathbf{u} \neq \mathbf{0} \quad \text{and} \quad \mathbf{v}, \mathbf{w} \text{ lin. indep.}$$

Parameterform

Definition 10.20 Two Point Formula:

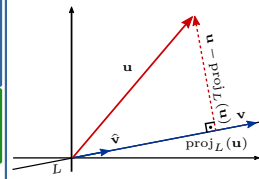


$$E = \mathbf{u} + \mathbb{K}(\mathbf{v} - \mathbf{u}) + \mathbb{K}(\mathbf{w} - \mathbf{u}) \quad (10.27)$$

4.4.4. Minimal Distance of Vector Subspaces

Projections in 2D

Definition 10.21 2D Vector Projection
[Proof 10.17,10.18]:



$$\begin{aligned} \mathbf{u}_v &= \text{proj}_L(\mathbf{u}) \\ &= u_v \hat{\mathbf{v}} = (\mathbf{u}^T \hat{\mathbf{v}}) \hat{\mathbf{v}} \\ &= \frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{v}\|^2} \mathbf{v} = \frac{\mathbf{u}^T \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \mathbf{v} \end{aligned} \quad (10.28)$$

Corollary 10.8

2D Projection Matrix \mathbf{P} : Is the matrix that satisfies:

$$\mathbf{P} \mathbf{u} = \text{proj}_L(\mathbf{u}) \quad \mathbf{P} = \frac{\mathbf{v} \mathbf{v}^T}{\mathbf{v}^T \mathbf{v}} = \frac{\mathbf{v} \mathbf{v}^T}{\|\mathbf{v}\|^2} \quad (10.29)$$

Proof 10.1: [Corollary 10.8]

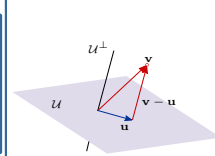
$$\frac{1}{\mathbf{v}^T \mathbf{v}} \mathbf{u}^T \mathbf{v} \mathbf{v} = \frac{1}{\mathbf{v}^T \mathbf{v}} \mathbf{v} (\mathbf{v}^T \mathbf{u}) = \frac{1}{\mathbf{v}^T \mathbf{v}} (\mathbf{v} \mathbf{v}^T) \mathbf{u}$$

General Projections

Definition 10.22

General Vector Projection:

Is the orthogonal projection \mathbf{u} of a vector \mathbf{v} onto a sub-vector space \mathcal{U}



$$\mathbf{u} = \sum_{i=1}^n \alpha_i \mathbf{b}_i \quad (10.30)$$

$$\mathbf{A} \mathbf{A}^T \alpha_i = \mathbf{A}^T \mathbf{v} \quad \mathbf{A} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{pmatrix}$$

where $\mathfrak{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ is a basis of the vector subspace \mathcal{U} .

Theorem 10.1 Projection Theorem: Let \mathcal{U} a sub vector space of a finite euclidean vector space \mathcal{V} . Then there exists for every vector $\mathbf{v} \in \mathcal{V}$ a vector $\mathbf{u} \in \mathcal{U}$ obtained by an orthogonal^[def. 10.67] projection

$$p: \begin{cases} \mathcal{V} \rightarrow \mathcal{U} \\ \mathbf{v} \mapsto \mathbf{u} \end{cases} \quad (10.31)$$

the vector $\mathbf{u}' := \mathbf{v} - \mathbf{u}$ representing the distance between \mathbf{u} and \mathbf{v} and is minimal:

$$\|\mathbf{u}'\| = \|\mathbf{v} - \mathbf{u}\| \leq \|\mathbf{v} - \mathbf{w}\| \quad \forall \mathbf{w} \in \mathcal{U} \quad \mathbf{u}' \in \mathcal{U}^\perp \quad (10.32)$$

4.5. Affine Subspaces

4.6. Planes

<https://math.stackexchange.com/questions/1485509/show-that-two-planes-are-parallel-and-find-the-distance-between-them>

5. Matrices

Special Kind of Matrices

5.1. Symmetric Matrices

Definition 10.23 Symmetric Matrices: A matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ is called *symmetric* if it satisfies:

$$\mathbf{A} = \mathbf{A}^T \quad (10.33)$$

Property 10.2 [proof ??]
Eigenvalues of real symmetric Matrices: The eigenvalues of a real symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ are real:

$$\text{spectrum}(\mathbf{A}) \in \{\mathbb{R}_{\geq 0}\}_{i=1}^n \quad (10.34)$$

Property 10.3 [proof ??]
Orthogonal Eigenvector basis: Eigenvectors of real symmetric matrices with distinct eigenvalues are orthogonal.

Corollary 10.9
Eigendecomposition Symmetric Matrices: If $\mathbf{A} \in \mathbb{R}^{n,n}$ is a real *symmetric*^[def. 10.23] matrix then its eigenvectors are *orthogonal* and its eigen-decomposition^[def. 10.86] is given by:

$$\mathbf{A} = \mathbf{X} \mathbf{\Lambda} \mathbf{X}^T \quad (10.35)$$

5.2. Orthogonal Matrices

Definition 10.24 Orthogonal Matrix: A real valued square matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is said to be orthogonal if its row vectors (and respectively its column vectors) build an orthonormal^[def. 10.68] basis:

$$\langle \mathbf{q}_{:,i}, \mathbf{q}_{:,j} \rangle = \delta_{ij} \quad \text{and} \quad \langle \mathbf{q}_i, \mathbf{q}_j \rangle = \delta_{ij} \quad (10.36)$$

This is exactly true if the inverse of \mathbf{Q} equals its transpose:

$$\mathbf{Q}^{-1} = \mathbf{Q}^T \iff \mathbf{Q} \mathbf{Q}^T = \mathbf{Q}^T \mathbf{Q} = \mathbf{I}_n \quad (10.37)$$

Attention: *Orthogonal* matrices are sometimes also called *orthonormal matrices*.

5.3. Hermitian Matrices

Definition 10.25 Conjugate Transpose $\mathbf{A}^H / \mathbf{A}^*$
Hermitian Conjugate/Adjoint Matrix:
The conjugate transpose of a matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$ is defined as:

$$\mathbf{A}^H := (\mathbf{A}^T)^* = \overline{\mathbf{A}^T} \iff \mathbf{a}_{i,j}^H = \overline{\mathbf{a}_{j,i}} \quad \begin{matrix} 1 \leq i \leq n \\ 1 \leq j \leq m \end{matrix} \quad (10.38)$$

Definition 10.26
Hermitian/Self-Adjoint Matrices $\mathbf{A} = \mathbf{A}^H$:
A hermitian matrix is complex square matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ who is equal to its own *conjugate transpose*^[def. 10.25]:

$$\mathbf{A} = \mathbf{A}^H = \overline{\mathbf{A}^T} \iff \mathbf{a}_{i,j} = \overline{\mathbf{a}_{j,i}} \quad i \in \{1, \dots, n\} \quad (10.39)$$

Corollary 10.10 : ^[def. 10.25] implies that \mathbf{A} must be a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$.

Corollary 10.11 Real Hermitian Matrices: From ^[cor. 1.1] it follows:

$$\mathbf{A} \in \mathbb{R}^{n \times n} \text{ hermitian} \implies \mathbf{A} \text{ real symmetric}^{\text{[def. 10.23]}} \quad (10.40)$$

Property 10.4 [proof 10.15]
Eigenvalues of Hermitan Matrices: The eigenvalues of a hermitian matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ are real:

$$\text{spectrum}(\mathbf{A}) \in \{\mathbb{R}_{\geq 0}\}_{i=1}^n \quad (10.41)$$

Property 10.5 [proof 10.16]
Orthogonal Eigenvector basis: Eigenvectors of hermitian matrices with distinct eigenvalues are orthogonal.

Corollary 10.12
Eigendecomposition Symmetric Matrices: If $\mathbf{A} \in \mathbb{C}^{n,n}$ is a hermitian matrix^[def. 10.26] then its eigendecomposition^[def. 10.86] is given by:

$$\mathbf{A} = \mathbf{X} \mathbf{\Lambda} \mathbf{X}^H \quad (10.42)$$

5.4. Unitary Matrices

Definition 10.27 Unitary Matrix $\mathbf{U} \mathbf{U}^H$:
is a complex square matrix $\mathbf{U} \in \mathbb{C}^{n \times n}$ whose inverse^[def. 10.41] is equal to its *conjugate transpose*^[def. 10.25]:

$$\mathbf{U}^H \mathbf{U} = \mathbf{U} \mathbf{U}^H = \mathbf{I} \quad (10.43)$$

Corollary 10.13 Real Unitary Matrix: A real matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ that is unitary is an *orthogonal matrix*^[def. 10.24].

Property 10.6 [proof 10.14]:
Preservation of Euclidean Norm
Orthogonal and unitary matrices $\mathbf{Q} \in \mathbb{K}^{n,n}$ do not affect the 2-norm:

$$\|\mathbf{Q}\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \quad \forall \mathbf{x} \in \mathbb{K}^n \quad (10.44)$$

5.5. Similar Matrices

Definition 10.28 Similar Matrices: Two square matrices $\mathbf{A} \in \mathbb{K}^{n \times n}$ and $\mathbf{B} \in \mathbb{K}^{n \times n}$ are called *similar* if there exists a invertible matrix $\mathbf{S} \in \mathbb{K}^{n \times n}$ s.t.:

$$\exists \mathbf{S} : \quad \mathbf{B} = \mathbf{S}^{-1} \mathbf{A} \mathbf{S} \quad (10.45)$$

Corollary 10.14
Similarity Transformation/Conjugation:
The mapping:

$$\mathbf{A} \mapsto \mathbf{S}^{-1} \mathbf{A} \mathbf{S} \quad (10.46)$$

is called *similarity transformation*

Corollary 10.15 [proof 10.13]:
Eigenvalues of Similar Matrices
If $\mathbf{A} \in \mathbb{K}^{n \times n}$ has the eigenvalue-eigenvector pairs $\{\{\lambda_i, \mathbf{v}_i\}\}_{i=1}^n$ then its *conjugate*eq. (10.46) \mathbf{B} has the same eigenvalues with transformed eigenvectors:

$$\{\{\lambda_i, \mathbf{u}_i\}\}_{i=1}^n \quad \mathbf{u}_i := \mathbf{S}^{-1} \mathbf{v}_i \quad (10.47)$$

5.6. Skew Symmetric Matrices

Definition 10.29
Key Symmetric/Antisymmetric Matrices:

$$\mathbf{A}^T = -\mathbf{A} \quad (10.48)$$

5.7. Triangular Matrix

Definition 10.30 Triangular Matrix: An upper (lower) triangular matrix, is a matrix whose element's below (above) the main diagonal are all zero:

Figure 1: Lower Tri. Mat. $\begin{pmatrix} l_{11} & & \\ l_{21} & l_{22} & \\ \vdots & \vdots & \ddots \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{pmatrix}$ Figure 2: Upper Tri. Mat. $\begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ & u_{22} & \dots & u_{2n} \\ & & \ddots & \vdots \\ & & & u_{nn} \end{pmatrix}$

Figure 1: Lower Tri. Mat. Figure 2: Upper Tri. Mat.

5.7.1. Unitriangular Matrix

Definition 10.31 Unitriangular Matrix: An upper (lower) unitriangular matrix, is a upper (lower) triangular matrix^[def. 10.30] whose diagonal elements are all ones.

5.7.2. Strictly Triangular Matrix

Definition 10.32 Strictly Triangular Matrix: An upper (lower) strictly triangular matrix, is a upper (lower) triangular matrix^[def. 10.30] whose diagonal elements are all zero.

5.8. Block Partitioned Matrices

Definition 10.33 Block Partitioned Matrix:
A matrix $\mathbf{M} \in \mathbb{R}^{k+l, k+l}$ can be partitioned into a *block partitioned matrix*:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \quad \mathbf{A} \in \mathbb{R}^{k,k}, \mathbf{B} \in \mathbb{R}^{k,l}, \mathbf{C} \in \mathbb{R}^{l,k}, \mathbf{D} \in \mathbb{R}^{l,l} \quad (10.49)$$

Definition 10.34 Block Partitioned Linear System:
A linear system $\mathbf{M}\mathbf{x} = \mathbf{b}$ with $\mathbf{M} \in \mathbb{R}^{k+l, k+l}$ and $\mathbf{x}, \mathbf{b} \in \mathbb{R}^{k+l}$ can be partitioned into a *block partitioned system*:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} \quad \mathbf{A} \in \mathbb{R}^{k,k}, \mathbf{B} \in \mathbb{R}^{k,l}, \mathbf{C} \in \mathbb{R}^{l,k}, \mathbf{D} \in \mathbb{R}^{l,l}, \mathbf{x}_1, \mathbf{b}_1 \in \mathbb{R}^k, \mathbf{x}_2, \mathbf{b}_2 \in \mathbb{R}^l \quad (10.50)$$

5.8.1. Schur Complement

Definition 10.35 Schur Complement: Given a block partitioned matrix^[def. 10.33] $\mathbf{M} \in \mathbb{R}^{k+l, k+l}$ its Schur complements are given by:

$$\mathbf{S}_A = \mathbf{D} - \mathbf{C} \mathbf{A}^{-1} \mathbf{B} \quad \mathbf{S}_D = \mathbf{A} - \mathbf{B} \mathbf{D}^{-1} \mathbf{C} \quad (10.51)$$

5.8.2. Inverse of Block Partitioned Matrix

Definition 10.36 proof 10.3
Inverse of a Block Partitioned Matrix:
Given a block partitioned matrix^[def. 10.33] $\mathbf{M} \in \mathbb{R}^{k+l, k+l}$ its inverse \mathbf{M}^{-1} can be partitioned as well:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \quad \mathbf{M}^{-1} = \begin{bmatrix} \tilde{\mathbf{A}} & \tilde{\mathbf{B}} \\ \tilde{\mathbf{C}} & \tilde{\mathbf{D}} \end{bmatrix} \quad (10.52)$$
$$\begin{aligned} \tilde{\mathbf{A}} &= \mathbf{A}^{-1} + \mathbf{A}^{-1} \mathbf{B} \mathbf{S}_A^{-1} \mathbf{C} \mathbf{A}^{-1} & \tilde{\mathbf{C}} &= -\mathbf{S}_A^{-1} \mathbf{C} \mathbf{A}^{-1} \\ \tilde{\mathbf{B}} &= -\mathbf{A}^{-1} \mathbf{B} \mathbf{S}_A^{-1} & \tilde{\mathbf{D}} &= \mathbf{S}_A^{-1} \end{aligned}$$

where $\mathbf{S}_A = \mathbf{D} - \mathbf{C} \mathbf{A}^{-1} \mathbf{B}$ is the Schur complement of \mathbf{A} .

5.9. Properties of Matrices

5.9.1. Square Root of p.s.d. Matrices

Definition 10.37 Square Root:

5.9.2. Trace

Definition 10.38 Trace: The trace of an $\mathbf{A} \in \mathbb{R}^{n \times n}$ matrix is defined as:

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii} = a_{11} + a_{22} + \dots + a_{nn} \quad (10.53)$$

Property 10.7 Trace of a Scalar:

$$\text{tr}(\mathbb{R}) = \mathbb{R} \quad (10.54)$$

Property 10.8 Trace of Transpose:

$$\text{tr}(\mathbf{A}^T) = \text{tr}(\mathbf{A}) \quad (10.55)$$

Property 10.9 Trace of multiple Matrices:

$$\text{tr}(\mathbf{A} \mathbf{B} \mathbf{C}) = \text{tr}(\mathbf{B} \mathbf{C} \mathbf{A}) = \text{tr}(\mathbf{C} \mathbf{B} \mathbf{A}) \quad (10.56)$$

6. Matrices and Determinants

6.1. Determinants

6.1.1. Laplace/Cofactor Expansion

Definition 10.39 Minor:

Definition 10.40 Cofactors:

Properties

Property 10.10 Determinant times Scalar $\det(\alpha \mathbf{A})$:
Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ it holds:

$$\det(\alpha \cdot \mathbf{A}) = \alpha^n \mathbf{A} \quad (10.57)$$

6.2. Inverse of Matrices

Definition 10.41 Inverse Matrix \mathbf{A}^{-1} :

6.2.1. Invertability

Definition 10.42
Singular/Non-Invertible Matrix $\det(\mathbf{A}) = 0$:
A square matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ is singular or non-invertible if it satisfies the following and equal conditions:

- $\det(\mathbf{A}) = 0$
- $\mathbf{A} \mathbf{x} = \mathbf{b}$ has either
 - no solution \mathbf{x}
 - infinitely many solutions \mathbf{x}
- $\dim(\mathbf{A}) < n$
- $\nexists \mathbf{B} : \mathbf{B} = \mathbf{A}^{-1}$

Transformations And Mapping

7. Linear & Affine Mappings/Transformations

7.1. Linear Mapping

Definition 10.43
Linear Mapping: A linear mapping, function or transformation is a map $l : V \mapsto W$ between two \mathbb{K} -vector spaces^[def. 10.6] V and W if it satisfies:

$l(\mathbf{x} + \mathbf{y}) = l(\mathbf{x}) + l(\mathbf{y})$ (Additivity) (10.58)

$l(\alpha \mathbf{x}) = \alpha l(\mathbf{x}) \quad \forall \alpha \in \mathbb{K}$ (Homogenitivity) (10.59)

$\forall \mathbf{x}, \mathbf{y} \in V$

Proposition 10.1^[proof 10.8]
Equivalent Formulations: Definition 10.43 is equivalent to:

$l(\alpha \mathbf{x} + \beta \mathbf{y}) = \alpha l(\mathbf{x}) + \beta l(\mathbf{y}) \quad \forall \alpha, \beta \in \mathbb{K}$
 $\forall \mathbf{x}, \mathbf{y} \in V$ (10.60)

Corollary 10.16 Superposition Principle:
Definition 10.43 is also known as the superposition principle: “the net response caused by two or more signals is the sum of the responses that would have been caused by each signal individually.”

Corollary 10.17^[proof 10.10]
A linear mapping $\iff \mathbf{A}\mathbf{x}$:
For every matrix $\mathbf{A} \in \mathbb{K}^{m \times n}$ the map:

$l_{\mathbf{A}} : \begin{cases} \mathbb{K}^n & \rightarrow & \mathbb{K}^m \\ \mathbf{x} & \mapsto & \mathbf{A}\mathbf{x} \end{cases}$ (10.61)

is a linear map and every linear map l can be represented by a matrix vector product:

l is linear $\iff \exists \mathbf{A} \in \mathbb{K}^{n \times m} : f(\mathbf{x}) = \mathbf{A}\mathbf{x} \quad \forall \mathbf{x} \in \mathbb{K}^m$ (10.62)

Principle 10.1^[proof 10.9]
Principle of linear continuation: A linear mapping $l : \mathcal{V} \mapsto \mathcal{W}$ is determined by the image of the basis \mathfrak{B} of \mathcal{V} :

$l(\mathbf{v}) = \sum_{i=1}^n \beta_i l(b_i) \quad \mathfrak{B}(\mathcal{V}) = \{b_1, \dots, b_n\}$ (10.63)

Property 10.11^[proof 10.11]
Compositions of linear mappings are linear $f \circ g$: Let g, f be linear functions mapping from \mathcal{V} to \mathcal{W} (i.e. matching) then it holds that $f \circ g$ is a linear^[def. 10.43].

Definition 10.44 Level Sets:

7.2. Affine Mapping

Definition 10.45 Affine Transformation/Map:
Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$ then:

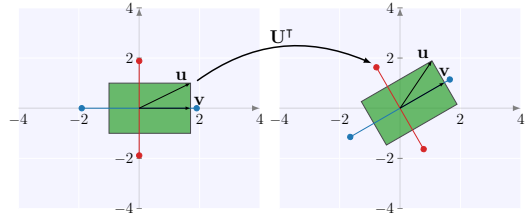
$\mathbf{Y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ (10.64)

is called an affine transformation of \mathbf{x} .

7.3. Orthogonal Transformations

Definition 10.46 Orthogonal Transformation:
A linear transformation $T : \mathcal{V} \mapsto \mathcal{V}$ of an inner product space^[def. 10.78] is an orthogonal transformation if preserves the inner product:

$T(\mathbf{u}) \cdot T(\mathbf{v}) = \mathbf{u} \cdot \mathbf{v} \quad \forall \mathbf{u}, \mathbf{v} \in \mathcal{V}$ (10.65)



Corollary 10.18 Orthogonal Matrix Transformation:
An orthogonal matrix^[def. 10.24] \mathbf{Q} provides an orthogonal transformation:

$(\mathbf{Q}\mathbf{u})^T (\mathbf{Q}\mathbf{v}) = \mathbf{u}\mathbf{v}$ (10.66)

Explanation 10.2 (Improper Rotations).
Orthogonal transformations in two or three dimensional euclidean space^[def. 10.46] represent improper rotations:

- Stiff Rotations
- Reflections
- Reflections+Rotations

Corollary 10.19 Preservation of Orthogonality: Orthogonal transformation preserves orthogonality.

Corollary 10.20^[proof 10.6]
Preservation of Norm:
An orthogonal transformation $\mathbf{Q} : \mathcal{V} \mapsto \mathcal{V}$ preserves the length/norm:

$\|\mathbf{u}\|_{\mathcal{V}} = \|\mathbf{Q}\mathbf{u}\|_{\mathcal{V}}$ (10.67)

Corollary 10.21 Preservation of Angle:
An orthogonal transformation T preserves the angle^[def. 10.66] of its vectors:

$\angle(\mathbf{u}, \mathbf{v}) = \angle(T(\mathbf{u}), T(\mathbf{v}))$ (10.68)

7.4. Kernel & Image

7.4.1. Kernel

Definition 10.47 Kernel/Null Space $\mathbb{N}/\varphi^{-1}(\{0\})$:
Let φ be a linear mapping^[def. 10.43] between two \mathbb{K} -vector spaces $\varphi : \mathcal{V} \mapsto \mathcal{W}$.
The kernel of φ is defined as:

$\mathbb{N}(\varphi) := \varphi^{-1}(\{0\}) = \{\mathbf{v} \in \mathcal{V} \mid \varphi(\mathbf{v}) = \mathbf{0}\} \subseteq \mathcal{V}$ (10.69)

Definition 10.48 Right Null Space $\mathbb{N}(\mathbf{A})$:
If $\varphi = \mathbf{A} \in \mathbb{K}^{m \times n}$ then the eq. (10.69) is equal to:

$\mathbb{N}(\mathbf{A}) = \varphi_{\mathbf{A}}^{-1}(\{0\}) = \{\mathbf{v} \in \mathbb{K}^n \mid \mathbf{A}\mathbf{v} = \mathbf{0}\} \in \mathbb{K}^m$ (10.70)

Definition 10.49 Left Null Space $\mathbb{N}(\mathbf{A}^T)$:
If $\varphi = \mathbf{A} \in \mathbb{K}^{m \times n}$ then the left null space is defined as:

$\mathbb{N}(\mathbf{A}^T) = \varphi_{\mathbf{A}^T}^{-1}(\{0\}) = \{\mathbf{v} \in \mathbb{K}^m \mid \mathbf{A}^T \mathbf{v} = \mathbf{0}\} \in \mathbb{K}^n$ (10.71)

Note
The term left null space stems from the fact that:

$(\mathbf{A}^T \mathbf{x})^T = \mathbf{0}$ is equal to $\mathbf{x}^T \mathbf{A} = \mathbf{0}$

7.4.2. Image

Definition 10.50 Image/Range \mathfrak{R}/φ :
Let φ be a linear mapping^[def. 10.43] between two \mathbb{K} -vector spaces $\varphi : \mathcal{V} \mapsto \mathcal{W}$.
The image of φ is defined as:

$\mathfrak{R}(\varphi) := \varphi(\mathcal{V}) = \{\varphi(\mathbf{v}) \mid \mathbf{v} \in \mathcal{V}\} \subseteq \mathcal{W}$ (10.72)

Definition 10.51 Column Space $\mathbf{A}\mathbf{x}$:
If $\varphi = \mathbf{A} = (\mathbf{c}_1 \dots \dots \mathbf{c}_n) \in \mathbb{K}^{m \times n}$ then eq. (10.72) is equal to:

$\mathfrak{R}(\mathbf{A}) = \varphi_{\mathbf{A}}(\mathbb{K}^n) = \{\mathbf{A}\mathbf{x} \mid \forall \mathbf{x} \in \mathbb{K}^n\} = \left\langle (\mathbf{c}_1 \dots \dots \mathbf{c}_n) \right\rangle$

$= \left\{ \mathbf{v} \mid \sum_{i=1}^n \alpha_i \mathbf{c}_i, \forall \alpha_i \in \mathbb{K} \right\}$ (10.73)

Definition 10.52 Row Space $\mathbf{A}^T \mathbf{x}$:
If $\varphi = \mathbf{A} = (\mathbf{r}_1^T \dots \dots \mathbf{r}_m^T) \in \mathbb{K}^{m \times n}$ then the column space is defined as:

$\mathfrak{R}(\mathbf{A}^T) = \varphi_{\mathbf{A}}(\mathbb{K}^m) = \{\mathbf{A}^T \mathbf{x} \mid \forall \mathbf{x} \in \mathbb{K}^m\} = \left\langle (\mathbf{r}_1 \dots \dots \mathbf{r}_m) \right\rangle$

$= \left\{ \mathbf{v} \mid \sum_{i=1}^m \alpha_i \mathbf{r}_i, \forall \alpha_i \in \mathbb{K} \right\}$ (10.74)

From orthogonality it follows $\mathbf{x} \in \mathfrak{R}(\mathbf{A}), \mathbf{y} \in \mathbb{N}(\mathbf{A}) \Rightarrow \mathbf{x}^T \mathbf{y} = 0$.

8. Eigenvalues and Vectors

Definition 10.53 Eigenvalues: Given a square matrix $\mathbf{A} \in \mathbb{K}^{n,n}$ the eigenvalues

Definition 10.54 Spectrum: The spectrum of a square matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ is the set of its eigenvalues^[def. 10.53]:

$\text{spectrum}(\mathbf{A}) = \lambda(\mathbf{A}) = \{\lambda_1, \dots, \lambda_n\}$ (10.78)

Formula 10.1 Eigenvalues of a 2x2 matrix: Given a 2x2-matrix \mathbf{A} its eigenvalues can be calculated by:

$\{\lambda_1, \lambda_2\} \in \frac{\text{tr}(\mathbf{A}) \pm \sqrt{\text{tr}(\mathbf{A})^2 - 4 \det(\mathbf{A})}}{2}$ (10.79)

with $\text{tr}(\mathbf{A}) = a + d$ $\det(\mathbf{A}) = ad - bc$

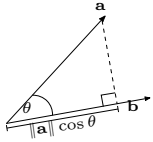
9. Vector Algebra

9.1. Dot/Standard Scalar Product

Definition 10.55 Scalar Projection

The scalar projection of a vector \mathbf{a} onto a vector \mathbf{b} is the *scalar* magnitude of the shadow/projection of the vector \mathbf{a} onto \mathbf{b} :

$$a_b = \|\mathbf{a}\| \cos \theta_{a,b} = \mathbf{a} \cdot \tilde{\mathbf{b}} \quad (10.80)$$



Definition 10.56 Standard Scalar/Dot Product:

Given two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ the standard scalar product is defined as:

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \mathbf{v} = \langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^n u_i v_i = u_1 v_1 + \dots + u_n v_n$$

$$= \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta = u_v \tilde{\mathbf{v}} = v_u \tilde{\mathbf{u}} \quad \theta \in [0, \pi] \quad (10.81)$$

Explanation 10.3 (Geometric Interpretation).

It is the magnitude of one vector times the magnitude of the shadow/scalar projection of the other vector.

Thus the dot product tells you:

- How much are two vectors pointing into the same direction
- With what magnitude

Property 10.12 Orthogonal Direction

For $\theta \in [-\pi, \pi/2]$ rad $\cos \theta = 0$ and it follows:

$$\mathbf{u} \cdot \mathbf{v} = 0 \iff \mathbf{u} \perp \mathbf{v} \quad (10.82)$$

Note: Perpendicular

Perpendicular corresponds to orthogonality of two lines.

Property 10.13 Maximizing Direction:

For $\theta = 0$ rad $\cos \theta = 1$ and it follows:

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| \quad (10.83)$$

Property 10.14 Minimizing Direction:

For $\theta = \pi$ rad $\cos \theta = -1$ and it follows:

$$\mathbf{u} \cdot \mathbf{v} = -\|\mathbf{u}\| \|\mathbf{v}\| \quad (10.84)$$

Definition 10.57 Vector Projection:

General Projection via normal equation into inner product stuff i.e. with projection theorem

9.2. Cross Product

9.3. Outer Product

Definition 10.58 Outer Product $\mathbf{uv}^T = \mathbf{u} \otimes \mathbf{v}$: Given two vectors $\mathbf{u} \in \mathbb{K}^m$, $\mathbf{v} \in \mathbb{K}^n$ their outer product is defined as:

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{uv}^H = \begin{bmatrix} u_1 & \dots & u_m \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \quad (10.85)$$

$$= \begin{bmatrix} u_1 \odot v_1 & \vdots & u_m \odot v_1 \\ \vdots & \ddots & \vdots \\ u_1 \odot v_n & \vdots & u_m \odot v_n \end{bmatrix} = \begin{bmatrix} u_1 \tilde{v}_1 & \dots & u_1 \tilde{v}_n \\ u_2 \tilde{v}_1 & \dots & u_2 \tilde{v}_n \\ \vdots & \ddots & \vdots \\ u_m \tilde{v}_1 & \dots & u_m \tilde{v}_n \end{bmatrix}$$

Proposition 10.2 [proof 10.5]
Rank of Outer Product: The outer product of two vectors is of rank one:

$$\text{rank}(\mathbf{u} \otimes \mathbf{v}) = 1 \quad (10.86)$$

9.4. Vector Norms

Definition 10.59 Norm $\|\cdot\|_{\mathcal{V}}$:

Let \mathcal{V} be a vector space over a field F , a norm on \mathcal{V} is a map: $\|\cdot\|_{\mathcal{V}} : \mathcal{V} \mapsto \mathbb{R}_+$ (10.87)

that satisfies:

$$\|\mathbf{x}\|_{\mathcal{V}} = 0 \iff \mathbf{x} = 0 \quad (\text{Definitness}) \quad (10.88)$$

$$\|\alpha \mathbf{x}\|_{\mathcal{V}} = |\alpha| \|\mathbf{x}\|_{\mathcal{V}} \quad (\text{Homogeneity}) \quad (10.89)$$

$$\|\mathbf{x} + \mathbf{y}\|_{\mathcal{V}} \leq \|\mathbf{x}\|_{\mathcal{V}} + \|\mathbf{y}\|_{\mathcal{V}} \quad (\text{Triangular Inequality}) \quad (10.90)$$

$$\alpha \in \mathbb{K} \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{V}$$

Explanation 10.4 (Definition 10.59).

A norm is a measures of the size of its argument.

Corollary 10.24 Normed vector space: Is a vector space \mathcal{V} over a field F , on which a norm $\|\cdot\|_{\mathcal{V}}$ can be defined.

9.4.1. Cauchy Schwartz

Definition 10.60 Cauchy Schwartz Inequality:

$$|\mathbf{u}^T \mathbf{v}| \leq \|\mathbf{u}\| \|\mathbf{v}\| \quad (10.91)$$

9.4.2. Triangular Inequality

Definition 10.61 [proof 10.22]
Triangular Inequality: States that the length of the sum of two vectors is lower or equal to the sum of their individual lengths:

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\| \quad (10.92)$$

Corollary 10.25 Reverse Triangular Inequality:

$$-\|\mathbf{x} - \mathbf{y}\|_{\mathcal{V}} \leq \|\mathbf{x}\|_{\mathcal{V}} - \|\mathbf{y}\|_{\mathcal{V}} \leq \|\mathbf{x} - \mathbf{y}\|_{\mathcal{V}}$$

resp. $\|\|\mathbf{x}\|_{\mathcal{V}} - \|\mathbf{y}\|_{\mathcal{V}}\| \leq \|\mathbf{x} - \mathbf{y}\|_{\mathcal{V}}$

9.5. Distances

Definition 10.62

Distance Function/Measure $d : S \times S \mapsto \mathbb{R}_+$: Let S be a set, a distance functions is a mapping d that satisfies:

$$d(x, x) = 0 \quad (\text{Zero Identity Distance}) \quad (10.93)$$

$$d(x, y) = d(y, x) \quad (\text{Symmetry}) \quad (10.94)$$

$$d(x, z) \leq d(x, y) + d(y, z) \quad (\text{Triangular Identity}) \quad (10.95)$$

$$\forall x, y, z \in S$$

Explanation 10.5 (Definition 10.62).

Is measuring the distance between two things.

9.5.1. Contraction

Definition 10.63 Contraction: Given a metric space (M, d) is a mapping $f : M \mapsto M$ that satisfies:

$$d(f(x), f(y)) \leq \lambda d(x, y) \quad \lambda \in [0, 1] \quad (10.96)$$

Add metric spaces

9.6. Metrics

Definition 10.64 Metric

Is a distance measure [def. 10.62] that additionally satisfies the **identity of indiscernibles**:

$$d(x, y) = 0 \iff x = y \quad \forall x, y \in S$$

Corollary 10.26 Metric \rightarrow Norm: Every norm $\|\cdot\|_{\mathcal{V}}$ on a vector space \mathcal{V} over a field F induces a metric by:

$$d(x, y) = \|x - y\|_{\mathcal{V}} \quad \forall x, y \in \mathcal{V}$$

metric induced by norms additionally satisfy: $\forall x, y \in \mathcal{V}$, $\alpha \in F \subseteq \mathbb{K}$ $K = \mathbb{R}$ or \mathbb{C}

- Homogeneity/Scaling:** $d(\alpha x, \alpha y)_{\mathcal{V}} = |\alpha| d(x, y)_{\mathcal{V}}$
- Translational Invariance:** $d(x + \alpha, y + \alpha) = d(x, y)$

Conversely not every metric induces a norm **but** if a metric d on a vector space \mathcal{V} satisfies the properties then it induces a norm of the form:

$$\|\mathbf{x}\|_{\mathcal{V}} := d(\mathbf{x}, 0)_{\mathcal{V}}$$

Note

Similarity measure is a much weaker notion than a metric as triangular inequality does not have to hold.

Hence: If \mathbf{a} is similar to \mathbf{b} and \mathbf{b} is similar to \mathbf{c} it does not imply that \mathbf{a} is similar to \mathbf{c} .

Note

(bilinear form $\xrightarrow{\text{induces}}$)

inner product $\xrightarrow{\text{induces}}$ norm $\xrightarrow{\text{induces}}$ metric.

9.6.1. Metric Space

Definition 10.65 Metric Space

A metric space is a pair (M, d) of a set M and a metric d defined on M :

$$d : M \times M \mapsto \mathbb{R}_+ \quad (10.97)$$

10. Angles

Definition 10.66 Angle between Vectors $\angle(\mathbf{u}, \mathbf{v})$: Let $\mathbf{u}, \mathbf{v} \in \mathbb{K}^n$ be two vectors of an inner product space [def. 10.78] \mathcal{V} . The angle $\alpha \in [0, \pi]$ between \mathbf{u}, \mathbf{v} is defined by:

$$\angle(\mathbf{u}, \mathbf{v}) := \alpha \quad \cos \alpha = \frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} \quad \mathbf{u}, \mathbf{v} \in \mathcal{V} \quad \alpha \in [0, \pi] \quad (10.98)$$

11. Orthogonality

Definition 10.67 Orthogonal Vectors: Let \mathcal{V} be an inner-product space [def. 10.78]. A set of vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_n\} \in \mathcal{V}$ is called *orthogonal* iff:

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0 \quad \forall i \neq j \quad (10.99)$$

11.1. Orthonormality

Definition 10.68 Orthonormal Vectors: Let \mathcal{V} be an inner-product space [def. 10.78]. A set of vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_n, \dots\} \in \mathcal{V}$ is called *orthonormal* iff:

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad \forall i, j \quad (10.100)$$

12. Special Kind of Vectors

12.1. Binary/Boolean Vectors

Definition 10.69

Binary/Boolean Vectors/Bit Maps \mathbb{B}^n : Are vectors that contain only zero or one values:

$$\mathbb{B}^n = \{0, 1\}^n \quad (10.101)$$

Definition 10.70

R-Sparse Boolean Vectors \mathbb{B}_r^n : Are boolean vectors that contain exact r one values:

$$\mathbb{B}_r^n = \left\{ \mathbf{x} \in \{0, 1\}^n : \mathbf{x}^T \mathbf{x} = \sum_{i=1}^n x_i = r \right\} \quad (10.102)$$

12.2. Probabilistic Vectors

Definition 10.71 Probabilistic Vectors: Are vectors that represent probabilities and satisfy:

$$\left\{ \mathbf{x} \in [0, 1]^n : \sum_{i=1}^n x_i = 1 \right\} \quad (10.103)$$

13. Vector Spaces and Measures

13.1. Bilinear Forms

13.2. Quadratic Forms

13.2.1. Min/Max Value

Corollary 10.27 [proof 10.20]

Extreme Value: The minimum/maximum of a quadratic form?? with a quadratic matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ is given by the eigenvector corresponding to the smallest/largest eigenvector of \mathbf{A} :

$$\mathbf{v}_1 \in \arg \min \mathbf{x}^T \mathbf{A} \mathbf{x} \quad \mathbf{v}_1 \in \arg \max \mathbf{x}^T \mathbf{A} \mathbf{x} \quad (10.104)$$

$$\mathbf{x}^T \mathbf{x} = 1 \quad \mathbf{x}^T \mathbf{x} = 1$$

Note

$$(\mathbf{Q}^T \tilde{\mathbf{n}})^T \mathbf{Q}^T \tilde{\mathbf{n}} = \tilde{\mathbf{n}}^T \mathbf{Q} \mathbf{Q}^T \tilde{\mathbf{n}} = \tilde{\mathbf{n}}^T \tilde{\mathbf{n}} = 1$$

13.2.2. Skew Symmetric Matirx

Corollary 10.28

Quadratic Form of Skew Symmetric matrix: The quadratic form of a skew symmetric matrix [def. 10.29] vanishes:

$$\alpha = \mathbf{x}^T \mathbf{A}_{\text{skew}} \mathbf{x} = (\mathbf{x}^T \mathbf{A}_{\text{skew}}^T \mathbf{x})^T = (\mathbf{x}^T \mathbf{A}_{\text{skew}} \mathbf{x})^T = -\alpha \quad (10.105)$$

Which can only hold iff $\alpha = 0$.

13.3. Inner Product – Generalization of the dot product

Definition 10.72 Bilinear Form/Functional:

Is a mapping $a : \mathcal{V} \times \mathcal{V} \mapsto F$ on a field of scalars $F \subseteq \mathbb{K}$, $K = \mathbb{R}$ or \mathbb{C} that satisfies:

$$a(\alpha u + \beta v, w) = \alpha a(u, w) + \beta a(v, w)$$

$$a(u, \alpha v + \beta w) = \alpha a(u, v) + \beta a(u, w)$$

$$\forall u, v, w \in \mathcal{V}, \quad \forall \alpha, \beta \in \mathbb{K}$$

Thus: a is linear w.r.t. each argument.

Definition 10.73 Symmetric bilinear form: A bilinear form a on \mathcal{V} is symmetric if and only if:

$$a(u, v) = a(v, u) \quad \forall u, v \in \mathcal{V}$$

Definition 10.74 Positive (semi) definite bilinear form:

A symmetric bilinear form a on a vector space \mathcal{V} over a field F is **positive definite** if and only if:

$$a(u, u) > 0 \quad \forall u \in \mathcal{V} \setminus \{0\} \quad (10.106)$$

$$\text{And positive semidefinite} \iff \geq \quad (10.107)$$

Corollary 10.29 Matrix induced Bilinear Form:

For finite dimensional inner product spaces $\mathcal{X} \in \mathbb{K}^n$ any *symmetric* matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ induces a **bilinear form**:

$$a(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{A} \mathbf{x}' = (\mathbf{A} \mathbf{x}')^T \mathbf{x}$$

Definition 10.75 Positive (semi) definite Matrix $>$:

A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is **positive definite** if and only if:

$$\mathbf{x}^T \mathbf{A} \mathbf{x} > 0 \iff \mathbf{A} > 0 \quad \forall \mathbf{x} \in \mathbb{R}^n \setminus \{0\} \quad (10.108)$$

$$\text{And positive semidefinite} \iff \geq \quad (10.109)$$

Corollary 10.30

Eigenvalues of positive (semi) definite matrix: A positive definite matrix is a matrix where every eigenvalue is *strictly* positive and positive semi definite if every eigenvalue is *positive*.

$$\forall \lambda_i \in \text{eigen}(\mathbf{A}) > 0 \quad (10.110)$$

$$\text{And positive semidefinite} \iff \geq \quad (10.111)$$

Note

Positive definite matrices are often assumed to be symmetric but that is not necessarily true.

Proof 10.2: ?? 10.2 (for real matrices):

Let \mathbf{v} be an eigenvector of \mathbf{A} then it follows:

$$\stackrel{?? 10.2}{0} < \mathbf{v}^T \mathbf{A} \mathbf{v} = \mathbf{v}^T \lambda \mathbf{v} = \|\mathbf{v}\| \lambda$$

Corollary 10.31 Positive Definiteness and Determinant: The determinant of a positive definite matrix is always positive. Thus a positive definite matrix is always *nonsingular*

Definition 10.76 Negative (semi) definite Matrix <:
A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is **negative definite** if and only if:
 $\mathbf{x}^\top \mathbf{A} \mathbf{x} < 0 \iff \mathbf{A} < 0 \quad \forall \mathbf{x} \in \mathbb{R}^n \setminus \{0\}$ (10.112)
And **negative semidefinite** $\iff \leq$ (10.113)

Theorem 10.3 Sylvester's criterion: Let \mathbf{A} be *symmetric/Hermitian* matrix and denote by $\mathbf{A}^{(k)}$ the $k \times k$ upper left sub-matrix of \mathbf{A} .
Then it holds that:

- $\mathbf{A} > 0 \iff \det(\mathbf{A}^{(k)}) > 0 \quad k = 1, \dots, n$ (10.114)
- $\mathbf{A} < 0 \iff (-1)^k \det(\mathbf{A}^{(k)}) > 0 \quad k = 1, \dots, n$ (10.115)
- \mathbf{A} is indefinite if the first $\det(\mathbf{A}^{(k)})$ that breaks both of the previous patterns is on the wrong side.
- Sylvester's criterion is inconclusive (\mathbf{A} can be anything of the previous three) if the first $\det(\mathbf{A}^{(k)})$ that breaks both patterns is 0.

14. Inner Products

Definition 10.77 Inner Product: Let \mathcal{V} be a vector space over a field $F \in \mathbb{K}$ of scalars. An inner product on \mathcal{V} is a map:
 $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \mapsto F \subseteq \mathbb{K} \quad K = \mathbb{R} \text{ or } \mathbb{C}$ (10.116)
that satisfies:

- (Conjugate) Symmetry:** $\langle x, y \rangle = \overline{\langle y, x \rangle}$.
- Linearity** in the first argument:
 $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$
- Positive-definiteness:**
 $\langle x, x \rangle \geq 0 : x = 0 \iff \langle x, x \rangle = 0$

Definition 10.78 Inner Product Space $(\mathcal{V}, \langle \cdot, \cdot \rangle_{\mathcal{V}})$: Let $F \in \mathbb{K}$ be a field of scalars.
An inner product space \mathcal{V} is a vector space over a field F together with an **inner product** $\langle \cdot, \cdot \rangle_{\mathcal{V}}$.

Corollary 10.32 Inner product \rightarrow S.p.d. Bilinear Form:
Let \mathcal{V} be a vector space over a field $F \in \mathbb{K}$ of scalar.
An **inner product** on \mathcal{V} is a positive definite symmetric bilinear form on \mathcal{V} .

Example: scalar prodct

Let $a(u, v) = u^\top \mathbf{I} v$ then the standard scalar product can be defined in terms of a bilinear form vice versa the standard scalar product induces a bilinear form.

Note

Inner products must be positive definite by definition $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, whereas bilinear forms must not.

Corollary 10.33 Inner product induced norm
 $\langle \cdot, \cdot \rangle_{\mathcal{V}} \rightarrow \|\cdot\|_{\mathcal{V}}$: Every inner product $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ induces a norm of the form:

$$\|\mathbf{x}\|_{\mathcal{V}} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \quad \mathbf{x} \in \mathcal{V}$$

Thus We can define function spaces by their associated norm $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ and inner product spaces lead to normed vector spaces and vice versa.

Corollary 10.34 Energy Norm: A *s.p.d.* bilinear form $a : \mathcal{V} \times \mathcal{V} \mapsto F$ induces an energy norm:
 $\|\mathbf{x}\|_a := (a(\mathbf{x}, \mathbf{x}))^{\frac{1}{2}} = \sqrt{a(\mathbf{x}, \mathbf{x})} \quad \mathbf{x} \in \mathcal{V}$

15. Matrix Algebra

16. Matrix Norms

16.1. Operator Norm

Definition 10.79 Operator/Induced Norm:
Let $\|\cdot\|_{\mu} : \mathbb{K}^m \mapsto \mathbb{R}$ and $\|\cdot\|_{\nu} : \mathbb{K}^n \mapsto \mathbb{R}$ be vector norms.
The operator norm is defined as:
 $\|\mathbf{A}\|_{\mu, \nu} := \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \mathbf{x} \neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_{\mu}}{\|\mathbf{x}\|_{\nu}} = \sup_{\|\mathbf{x}\|_{\nu}=1} \|\mathbf{A}\mathbf{x}\|_{\mu} \quad \|\cdot\|_{\mu} : \mathbb{K}^m \mapsto \mathbb{R}$ (10.117)

Explanation 10.6 (Definition 10.79). *Is a measure for the largest factor by which a matrix \mathbf{A} can stretch a vector $\mathbf{x} \in \mathbb{R}^n$.*

16.2. Induced Norms

Corollary 10.35 Induced Norms: Let $\|\cdot\|_p : \mathbb{K}^{m \times n} \mapsto \mathbb{R}$ defined as:

$$\|\mathbf{A}\|_p := \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \mathbf{x} \neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \sup_{\|\mathbf{y}\|_p=1} \|\mathbf{A}\mathbf{y}\|_p \quad (10.118)$$

Explanation 10.7 ([Corollary 10.35]).
Induced norms are matrix norms induced by vector norms as we:

- Only work with vectors $\mathbf{A}\mathbf{x}$*
- And use the normal p -vector norms $\|\cdot\|_p$*

Note supremum

The set of vectors $\{\mathbf{y} | \|\mathbf{y}\| = 1\}$ is compact, thus if we consider finite matrices the supremum is attained and we may replace it by the max.

16.3. Induced Norms

16.3.1. 1-Norm

Definition 10.80 Column Sum Norm $\|\mathbf{A}\|_1$:

$$\|\mathbf{A}\|_1 = \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \mathbf{x} \neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}| \quad (10.119)$$

16.3.2. ∞ -Norm

Definition 10.81 Row Sum Norm $\|\mathbf{A}\|_{\infty}$:

$$\|\mathbf{A}\|_{\infty} = \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \mathbf{x} \neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_{\infty}}{\|\mathbf{x}\|_{\infty}} = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}| \quad (10.120)$$

16.3.3. Spectral Norm L2-Norm Spectral Radius & Singular Value

Definition 10.82 Spectral Radius $\rho(\mathbf{A})$:
The spectral radius is defined as the largest eigenvalue of a matrix:
 $\rho(\mathbf{A}) = \max \{|\lambda| \in \text{eigenval}(\mathbf{A})\}$ (10.121)

Definition 10.83 Singular Value σ_i :
Given a matrix $\mathbf{A} \in \mathbb{K}^{m \times n}$ its n real and positive singular values are defined as:
 $\sigma(\mathbf{A}) := \left\{ \left\{ \sqrt{\lambda_i} \right\}_{i=1}^n \mid \lambda_i \in \text{eigenval}(\mathbf{A}^\top \mathbf{A}) \right\}$ (10.122)

Spectral Norm

Definition 10.84 L2/Spectral Norm $\|\mathbf{A}\|_2$:

$$\|\mathbf{A}\|_2 = \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \|\mathbf{x}\|_2=1}} \|\mathbf{A}\mathbf{x}\|_2 = \max_{\|\mathbf{x}\|_2=1} \sqrt{\mathbf{x}^\top \mathbf{A}^\top \mathbf{A} \mathbf{x}} \quad (10.123)$$

$$= \max_{\|\mathbf{x}\|_2=1} \sqrt{\rho(\mathbf{A}^\top \mathbf{A})} =: \sigma_{\max}(\mathbf{A}) \quad (10.124)$$

16.4. Energy Norm

16.5. Forbenius Norm

Definition 10.85 Forbenius Norm $\|\mathbf{A}\|_F$:
The *Forbenius norm* $\|\cdot\|_F : \mathbb{K}^{m \times n} \mapsto \mathbb{R}$ is defined as:

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}^2|} = \sqrt{\text{tr}(\mathbf{A}\mathbf{A}^\text{H})} \quad (10.125)$$

16.6. Distance

17. Decompositions

17.1. Eigen/Spectral decomposition

Definition 10.86 $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$, [proof 10.25]
Eigendecomposition/ Spectral Decomposition :
Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a *diagonalizable* square matrix and define by $\mathbf{X} = [\mathbf{x}_1 \dots \dots \mathbf{x}_n] \in \mathbb{R}^{n \times n}$ a non-singular matrix whose column vectors are the eigenvectors of \mathbf{A} with associated eigenvalue matrix $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$. Then \mathbf{A} can be represented as:
 $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$ (10.126)

Proposition 10.3 Diagonalization: If non of \mathbf{A} eigenvalues are zero it can be diagonalized:
 $\mathbf{S}^{-1} \mathbf{A} \mathbf{S} = \mathbf{\Lambda}$ (10.127)

Proposition 10.4 Existence:
 $\exists \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1} \iff \mathbf{A}$ diagonalizable (10.128)

17.2. QR-Decompositions

17.3. Singular Value Decomposition

Definition 10.87
Singular Value Decomposition (SVD) $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\text{H}$:
For any matrix $\mathbf{A} \in \mathbb{K}^{m,n}$ there exist unitary matrices^[def. 10.27]
 $\mathbf{U} \in \mathbb{K}^{m,m} \quad \mathbf{V} \in \mathbb{K}^{n,n}$
and a (generalized) digonal matrix:
 $\mathbf{\Sigma} \in \mathbb{R}^{m,n} \quad p := \min\{m, n\}$
 $\mathbf{\Sigma} = \text{gendia}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{m,n}$

such that:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\text{H} \quad (10.129)$$

$$= \left(\begin{array}{c|c|c|c} \text{u}_1 & \text{u}_r & \text{u}_{r+1} & \text{u}_m \\ \hline \vdots & \vdots & \vdots & \vdots \\ \hline \text{u}_1 & \text{u}_r & \text{u}_{r+1} & \text{u}_m \\ \hline \end{array} \right) \left(\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r & \\ & & & 0 & \\ & & & & \ddots & \\ & & & & & & 0 \end{array} \right) \left(\begin{array}{c} \text{v}_1^\text{H} \\ \text{v}_2^\text{H} \\ \text{v}_3^\text{H} \\ \vdots \\ \text{v}_n^\text{H} \end{array} \right)$$

(Image full, economical, alternative representation, range and kernel
https://math.ubt.uni-bonn.de/~math/lehre/ss19/lineal/190920_SVD_Kernel.pdf

17.3.1. Eigenvalues

Proposition 10.5 [proof 10.23]:
The eigenvalues of a matrix $\mathbf{A}^\top \mathbf{A}$ are positive.

Proposition 10.6 [proof 10.24]
Similarity Transformation: The unitary matrix \mathbf{V} provides a *similarity transformation*^[cor. 10.14] of $\mathbf{A}^\top \mathbf{A}$ into a diagonal matrix $\mathbf{\Sigma}^\text{T} \mathbf{\Sigma}$:

$$\mathbf{\Sigma}^\text{T} \mathbf{\Sigma} \mapsto \mathbf{V}^\text{H} \mathbf{A}^\top \mathbf{A} \mathbf{V} \quad (10.130)$$

Corollary 10.36 eigenval($\mathbf{A}^\top \mathbf{A}$) = eigenval($\mathbf{\Sigma}^\text{T} \mathbf{\Sigma}$):
From proposition 10.6 and ^[cor. 10.15] it follows that:
 $\text{eigenval}(\mathbf{A}^\top \mathbf{A}) = \text{eigenval}(\mathbf{\Sigma}^\text{T} \mathbf{\Sigma})$ (10.131)
 $\implies \|\mathbf{A}\|_2 = \sqrt{\rho(\mathbf{A}^\top \mathbf{A})} = \sqrt{\lambda_{\max}} = \sigma_{\max}$

Note

λ and *singularvalue* corresponds to the eigenvalues/singular-values of $\mathbf{A}^\top \mathbf{A}$ and not \mathbf{A}

17.3.2. Best Lower Rank Approximation

Theorem 10.4 Eckart Young Theorem: Given a matrix $\mathbf{X} \in \mathbb{K}^{m,n}$ the *reduced SVD* \mathbf{X} defined as:

$$\mathbf{X}_k := \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\text{H} \quad \mathbf{U}_k := [\mathbf{u}_{:,1} \dots \dots \mathbf{u}_{:,k}] \in \mathbb{K}^{m,k}$$

$$\mathbf{\Sigma}_k = \text{diag}(\sigma_1, \dots, \sigma_k) \in \mathbb{R}^{k,k}$$

$$\mathbf{V}_k = [\mathbf{v}_{:,1} \dots \dots \mathbf{v}_{:,k}] \in \mathbb{K}^{n,k}$$

$k \leq \min\{m, n\}$

provides the best lower k rank approximation of \mathbf{X} :

$$\min_{\mathbf{Y} \in \mathbb{K}^{n,m} : \text{rank}(\mathbf{Y}) \leq k} \|\mathbf{X} - \mathbf{Y}\|_F = \|\mathbf{X} - \mathbf{X}_k\|_F \quad (10.132)$$

18. Matric Calculus

18.1. Derivatives

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{b}^\top \mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^\top \mathbf{b}) = \mathbf{b}$$

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^\top \mathbf{x}) = 2\mathbf{x}$$

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{A} \mathbf{x} = \mathbf{A} \quad (10.133)$$

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top \mathbf{A} \mathbf{x} = (\mathbf{A} + \mathbf{A}^\top) \mathbf{x} \quad (10.134)$$

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{b}^\top \mathbf{A} \mathbf{x}) = \mathbf{A}^\top \mathbf{b} \quad \frac{\partial}{\partial \mathbf{X}} (\mathbf{c}^\top \mathbf{X} \mathbf{b}) = \mathbf{c} \mathbf{b}^\top \quad \frac{\partial}{\partial \mathbf{x}} (\|\mathbf{x} - \mathbf{b}\|_2) = \frac{\mathbf{x} - \mathbf{b}}{\|\mathbf{x} - \mathbf{b}\|_2}$$

$$\frac{\partial}{\partial \mathbf{x}} (\|\mathbf{x}\|_2^2) = \frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^\top \mathbf{x}) = 2\mathbf{x} \quad \frac{\partial}{\partial \mathbf{X}} (\|\mathbf{X}\|_F^2) = 2\mathbf{X}$$

$$\frac{\partial}{\partial \mathbf{x}} \|\mathbf{x}\|_1 = \frac{\mathbf{x}}{|\mathbf{x}|}$$

$$\frac{\partial}{\partial \mathbf{x}} (\|\mathbf{A} \mathbf{x} - \mathbf{b}\|_2^2) = 2(\mathbf{A}^\top \mathbf{A} \mathbf{x} - \mathbf{A}^\top \mathbf{b}) \quad \frac{\partial}{\partial \mathbf{X}} (|\mathbf{X}|) = |\mathbf{X}| \cdot \mathbf{X}^{-1}$$

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{Y}^{-1}) = -\mathbf{Y}^{-1} \frac{\partial \mathbf{Y}}{\partial \mathbf{x}} \mathbf{Y}^{-1}$$

19. Proofs

Proof 10.3: ^[def. 10.36]

$$\mathbf{M}\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{I}_{k,k} & \mathbf{0}_{l,k} \\ \mathbf{0}_{l,k} & \mathbf{I}_{l,l} \end{bmatrix} \quad (10.135)$$

19.1. Vector Algebra

Proof 10.4 Definition 10.56:
eq. (11.19)

- $\|a - b\| = \|a\|^2 + \|b\|^2 - 2\|a\| \|b\| \cos \theta$
- $\|a - b\| = (a - b)(a - b) = \|a\|^2 + \|b\|^2 - 2(ab)$

$$\|a - b\| = \|a - b\| \implies ab = \|a\| \|b\| \cos \theta$$

Proof 10.5 Proposition 10.2: The outer product of \mathbf{u} with \mathbf{v} corresponds to a scalar multiplication of \mathbf{v} , which is a vector and hence of rank 1

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u} \mathbf{v}^\text{H} = \begin{bmatrix} \mathbf{u}_1 \odot \bar{\mathbf{v}}_1 \\ \vdots \\ \mathbf{u}_m \odot \bar{\mathbf{v}}_n \end{bmatrix}$$

19.2. Mappings

Proof 10.6: Corollary 10.20
 $\|\mathbf{Q}\mathbf{x}\|^2 = (\mathbf{Q}\mathbf{x})^\top \mathbf{Q}\mathbf{x} = \mathbf{x}^\top \mathbf{Q}^\text{T} \mathbf{Q}\mathbf{x} = \mathbf{x}^\top \mathbf{x} = \|\mathbf{x}\|^2$

Proof 10.7: Corollary 10.21 Follows immediately from definition 10.66 in combination with eqs. (10.65) and (10.67).

Proof 10.8: Proposition 10.1:
 $\implies l(\alpha \mathbf{x} + \beta \mathbf{y}) \stackrel{10.58}{=} l(\alpha \mathbf{x}) + l(\beta \mathbf{y}) \stackrel{10.59}{=} \alpha l(\mathbf{x}) + \beta l(\mathbf{y})$
 $\leftarrow l(\alpha \mathbf{x} + \mathbf{0}) = \alpha l(\mathbf{x})$
 $l(1\mathbf{x} + 1\mathbf{y}) = l(\mathbf{x}) + l(\mathbf{y})$

Proof 10.9 principle 10.1:
Every vector $\mathbf{v} \in \mathcal{V}$ can be represented by a basis eq. (10.16) of \mathcal{V} . With *homogeneity*eq. (10.59) and *additivity*eq. (10.58) it follows for the image of all $\mathbf{v} \in \mathcal{V}$:

$$l(\mathbf{v}) = l(\alpha_1 b_1 + \dots + \alpha_n b_n) = l\alpha_1 (b_1) + \dots + l(\alpha_n b_n) \quad (10.136)$$

\implies the image of the basis of \mathcal{V} determines the linear mapping.

Proof 10.10 Proof [Corollary 10.17]:
 $\implies l_{\mathbf{A}}(\alpha \mathbf{x} + \mathbf{y}) = \mathbf{A}(\alpha \mathbf{x} + \beta \mathbf{y}) = \alpha \mathbf{A} \mathbf{x} + \beta \mathbf{A} \mathbf{y} = \alpha l(\mathbf{x}) + \beta l(\mathbf{y})$
 \longleftarrow Let \mathfrak{B} be a standard normal basis of \mathcal{V} with eq. (10.136):

$$l(\mathbf{x}) = \sum_{i=1}^n x_i l(\mathbf{e}_i) = \sum_{i=1}^n x_i \mathbf{A}_{:,i} = \mathbf{A} \mathbf{x} \quad \mathbf{A}_{:,i} := l(\mathbf{e}_i) \in \mathbb{R}^n$$

Proof 10.11 Proof Property 10.11:
 $(g \circ f)(\alpha \mathbf{x}) = g(f(\alpha \mathbf{x})) = g(\alpha f(\mathbf{x})) = \alpha (g \circ f)(\mathbf{x})$
 $(g \circ f)(\mathbf{x} + \mathbf{y}) = g(f(\mathbf{x} + \mathbf{y})) = g(f(\mathbf{x}) + f(\mathbf{y}))$
 $= (g \circ f)(\mathbf{x}) + (g \circ f)(\mathbf{y})$
 or even simpler as every linear form can be represented by a matrix product:
 $f(y) = \mathbf{A} \mathbf{y} \quad g(z) = \mathbf{B} \mathbf{z} \quad \Rightarrow \quad (f \circ g)(\mathbf{x}) = \mathbf{A} \mathbf{B} \mathbf{x} := \mathbf{C} \mathbf{x}$

Proof 10.12: [Corollary 10.22] Let $\mathbf{y} \in \mathcal{N}(\mathbf{A})$ ($\mathbf{z} \in \mathcal{N}(\mathbf{A}^\top)$) then it follows:
 $\mathcal{N}(\mathbf{A}) \perp \mathfrak{R}(\mathbf{A}^\top) \quad (\mathbf{A}^\top \mathbf{x})^\top \mathbf{y} = \mathbf{x}^\top \mathbf{A} \mathbf{y} = \mathbf{x}^\top \mathbf{0} = 0$
 $\mathcal{N}(\mathbf{A}^\top) \perp \mathfrak{R}(\mathbf{A}) \quad (\mathbf{A} \mathbf{x})^\top \mathbf{z} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{z} = \mathbf{x}^\top \mathbf{0} = 0$

19.3. Special Matrices

Proof 10.13 [Corollary 10.15]: Let $\mathbf{u} = \mathbf{S}^{-1} \mathbf{v}$ then it follows:
 $\mathbf{S}^{-1} \mathbf{A} \mathbf{S} \mathbf{u} = \mathbf{S}^{-1} \mathbf{A} \mathbf{S} \mathbf{v} = \lambda \mathbf{S}^{-1} \mathbf{v} = \lambda \mathbf{u}$

Proof 10.14 Property 10.6:
 $\|\mathbf{Q} \mathbf{x}\|_2^2 = (\mathbf{Q} \mathbf{x})^\top \mathbf{Q} \mathbf{x} = \mathbf{x}^\top \mathbf{Q}^\top \mathbf{Q} \mathbf{x} = \|\mathbf{x}\|_2^2$

Proof 10.15: Property 10.4
 Let $\mathbf{A} \in \mathbb{K}^{n \times n}$ be a hermitian matrix^[def. 10.26] and let $\lambda \in \mathbb{K}$ be an eigenvalue of \mathbf{A} with corresponding eigenvector $\mathbf{v} \in \mathbb{K}^n$:
 $\lambda(\bar{\mathbf{v}}^\top \mathbf{v}) = \bar{\mathbf{v}}^\top \lambda \mathbf{v} = \bar{\mathbf{v}}^\top \mathbf{A} \mathbf{v} = \overline{(\bar{\mathbf{v}}^\top \mathbf{A} \mathbf{v})} = \bar{\mathbf{A}} \mathbf{v}^\top \mathbf{v} = \bar{\lambda}(\bar{\mathbf{v}}^\top \mathbf{v})$
 $\lambda(\bar{\mathbf{v}}^\top \mathbf{v}) = \bar{\lambda}(\bar{\mathbf{v}}^\top \mathbf{v})$
 1. $\bar{\mathbf{v}} \mathbf{v} = \sum_{i=1}^n |v_i|^2 > 0$ as $\mathbf{v} \neq \mathbf{0}$
 2. $\lambda = \bar{\lambda}$ which can only hold for $\lambda \in \mathbb{R}$ (Equation (1.8))

Proof 10.16: ??

19.4. Vector Spaces

Proof 10.17 Definition 10.21: We know that $\text{proj}_L(\mathbf{u})$ must be a vector times a certain magnitude:
 $\text{proj}_L(\mathbf{u}) = \alpha \tilde{\mathbf{v}} \quad \alpha \in \mathbb{K} \quad (10.137)$
 the magnitude follows from the scalar projection^[def. 10.55] in the direction of \mathbf{v} which concludes the derivation.

Proof 10.18 Definition 10.21 (via orthogonality): We know that $\mathbf{u} - \text{proj}_L(\mathbf{u})$ must be orthogonal^[def. 10.67] to \mathbf{v}
 $(\mathbf{u} - \text{proj}_L(\mathbf{u}))^\top \mathbf{v} = (\mathbf{u} - \alpha \mathbf{v})^\top \mathbf{v} = 0 \Rightarrow \quad \alpha = \frac{\mathbf{u}^\top \mathbf{v}}{\mathbf{v}^\top \mathbf{v}}$

Proof 10.19: Definition 10.22 Let $\mathfrak{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ a basis of \mathcal{U} s.t. by ^[cor. 10.4]:

$$\mathbf{u} = \sum_{i=1}^n \alpha_i \mathbf{b}_i$$

the coefficients $\{\alpha_i\}_{i=1}^n$ need to be determined. We know that:
 $\mathbf{v} - \mathbf{u} \perp \mathbf{b}_1, \dots, \mathbf{v} - \mathbf{u} \perp \mathbf{b}_n$
 $\implies \left(\mathbf{v} - \sum_{i=1}^n \alpha_i \mathbf{b}_i \right) \cdot \mathbf{b}_j = 0 \quad j = 1, \dots, n$

this linear system of equations can be rewritten as:

$$(\mathbf{b}_1 \cdot \dots \cdot \mathbf{b}_n) \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{pmatrix} \mathbf{v}$$

Proof 10.20: Corollary 10.27
 Let $\mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top$ be the eigendecomposition^[cor. 10.12] of \mathbf{A} then it follows:

$$\begin{aligned} \min_{\tilde{\mathbf{n}}^\top \tilde{\mathbf{n}}=1} \tilde{\mathbf{n}}^\top \mathbf{A} \tilde{\mathbf{n}} &= \min_{\|\tilde{\mathbf{n}}\|=1} \tilde{\mathbf{n}}^\top (\mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top) \tilde{\mathbf{n}} \\ &= \min_{\|\tilde{\mathbf{n}}\|=1} (\mathbf{Q}^\top \tilde{\mathbf{n}})^\top \mathbf{\Lambda} (\mathbf{Q}^\top \tilde{\mathbf{n}}) \\ &= \min_{\mathbf{x}=1} \mathbf{x}^\top \mathbf{\Lambda} \mathbf{x} \quad \mathbf{x} := \mathbf{Q}^\top \tilde{\mathbf{n}} \\ &= \min_{\mathbf{x}=1} \sum_{i=1}^n \mathbf{x}_i^2 \mathbf{\Lambda}_{ii} = \min_{\mathbf{x}=1} \sum_{i=1}^n \mathbf{x}_i^2 \lambda_i \end{aligned}$$

Thus in order to obtain the minimum value we need to choose the eigenvector that leads to the smallest eigenvalue.

19.5. Norms

Proof 10.21: ?? 10.21
 $|\mathbf{u} \cdot \mathbf{v}| \stackrel{\text{eq. (10.81)}}{=} \|\mathbf{u}\| \|\mathbf{v}\| |\cos \theta| \leq \|\mathbf{u}\| \|\mathbf{v}\|$

Proof 10.22: Definition 10.61
 $\|\mathbf{u} + \mathbf{v}\|^2 = (\mathbf{u} + \mathbf{v})(\mathbf{u} + \mathbf{v}) = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2(\mathbf{u} \cdot \mathbf{v})$
 from cauchy schwartz we know:
 $\mathbf{u} \cdot \mathbf{v} \leq |\mathbf{u} \cdot \mathbf{v}| \stackrel{\text{eq. (10.91)}}{\leq} \|\mathbf{u}\| \|\mathbf{v}\|$
 $\|\mathbf{u} + \mathbf{v}\|^2 \leq \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2(\|\mathbf{u}\| \|\mathbf{v}\|) = (\|\mathbf{u}\| + \|\mathbf{v}\|)^2$

19.6. Decompositions

19.6.1. Symmetric - Antisemitic

Definition 10.88 Symmetric - Antisymmetric Decomposition: Any matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ can be decomposed into the sum of a *symmetric matrix*^[def. 10.23] \mathbf{A}^{sym} and a *skew-symmetric matrix*?? \mathbf{A}^{skes} :

$$\begin{aligned} \mathbf{A} &= \mathbf{A}^{\text{sym}} + \mathbf{A}^{\text{skew}} \\ \mathbf{A}^{\text{sym}} &= \frac{1}{2} \left(\mathbf{A} + \mathbf{A}^{\text{H}} \right) \\ \mathbf{A}^{\text{skew}} &= \frac{1}{2} \left(\mathbf{A} - \mathbf{A}^{\text{H}} \right) \end{aligned} \quad (10.138)$$

19.6.2. SVD

Proof 10.23 [Corollary 10.5]: $\mathbf{B} := \mathbf{A}^\top \mathbf{A}$ corresponds to a *symmetric positive definite* form^[def. 10.75]:
 $\mathbf{x}^\top \mathbf{B} \mathbf{x} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{A} \mathbf{x} = \|\mathbf{A} \mathbf{x}\|_2^2 > 0$
 thus Proposition 10.6 follows immediately form [Corollary 10.2].

Proof 10.24 Proposition 10.6:
 $\mathbf{A}^\top \mathbf{A} \stackrel{\text{SVD}}{=} \left(\mathbf{U} \mathbf{\Sigma} \mathbf{V}^{\text{H}} \right)^{\text{H}} \mathbf{U} \mathbf{\Sigma} \mathbf{V}^{\text{H}} = \mathbf{V} \mathbf{\Sigma}^{\text{H}} \underbrace{\mathbf{U}^{\text{H}} \mathbf{U}}_{\mathbf{I}_m} \mathbf{\Sigma} \mathbf{V}^{\text{H}} = \mathbf{V} \mathbf{\Sigma}^{\text{H}} \mathbf{\Sigma} \mathbf{V}^{\text{H}}$
 $\implies \mathbf{V}^{\text{H}} \mathbf{A}^\top \mathbf{A} \mathbf{V} = \mathbf{\Sigma}^\top \mathbf{\Sigma}$

19.6.3. Eigendecomposition

Proof 10.25 Definition 10.86:
 $\mathbf{A} \mathbf{X} = [\lambda_1 \mathbf{x}_1 \cdot \dots \cdot \lambda_n \mathbf{x}_n] = \mathbf{X} \mathbf{\Lambda}$

Geometry

Corollary 11.1 Affine Transformation in 1D: Given: numbers $x \in \hat{\Omega}$ with $\hat{\Omega} = [a, b]$
The **affine transformation** of $\phi : \hat{\Omega} \rightarrow \Omega$ with $y \in \Omega = [c, d]$ is defined by:

$$y = \phi(x) = \frac{d - c}{b - a} (x - a) + c \tag{11.1}$$

Proof 11.1: [cor. 11.1] By [def. 10.45] we want a function $f : [a, b] \rightarrow [c, d]$ that satisfies:

$$f(a) = c \qquad \text{and} \qquad f(b) = d$$

additionally $f(x)$ has to be a linear function ([def. 5.15]), that is the output scales the same way as the input scales.

Thus it follows:

$$\frac{d - c}{b - a} = \frac{f(x) - f(a)}{x - a} \qquad \Longleftrightarrow \qquad f(x) = \frac{d - c}{b - a} (x - a) + c$$

Trigonometry

0.1. Trigonometric Functions

0.1.1. Sine

Definition 11.1 Sine:

$$\sin \alpha = \frac{\text{opposite}}{\text{hypotenuse}} = \frac{a}{c} \tag{11.2}$$

0.1.2. Cosine

Definition 11.2 Cosine:

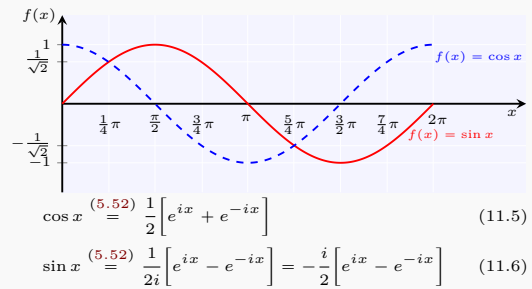
$$\cos \alpha = \frac{\text{adjacent}}{\text{hypotenuse}} = \frac{b}{c} \tag{11.3}$$

0.1.3. Tangens

Definition 11.3 Tangens:

$$\cos \alpha = \frac{\text{opposite}}{\text{adjacent}} = \frac{a}{b} = \frac{a/c}{b/c} = \frac{\sin \alpha}{\cos \alpha} \tag{11.4}$$

0.1.4. Trigonometric Functions and the Unit Circle
Sine and Cosine



Note

Using theorem 11.1 if follows:

$$\cos(\alpha \pm \pi) = -\cos \alpha \quad \text{and} \quad \sin(\alpha \pm \pi) = -\sin \alpha \tag{11.7}$$

0.1.5. Sinh

Definition 11.4 Sinh:

$$\sinh x \stackrel{(eq. (5.52))}{=} \frac{1}{2} \left[e^x - e^{-x} \right] = -i \sin(ix) \tag{11.8}$$

Property 11.1: $\sinh x = 0$ has a unique root at $x = 0$.

0.1.6. Cosh

Definition 11.5 Cosh:

$$\cosh x \stackrel{(5.52)}{=} \frac{1}{2} \left[e^x + e^{-x} \right] = \cos(ix) \tag{11.9}$$
$$\tag{11.10}$$

Property 11.2: $\cosh x$ is strictly positive.

Proof 11.2:

$$e^x = \cosh x + \sinh x \qquad e^{-x} = \cosh x - \sinh x \tag{11.11}$$

0.2. Addition Theorems

Theorem 11.1 Addition Theorems:

$$\sin(\alpha \pm \beta) = \sin \alpha \cos \beta \pm \cos \alpha \sin \beta \tag{11.12}$$

$$\cos(\alpha \pm \beta) = \cos \alpha \cos \beta \mp \sin \alpha \sin \beta \tag{11.13}$$

0.3. Werner Formulas

Werner Formulas

$$\sin \alpha \cos \beta = \frac{1}{2} \left[\sin(\alpha + \beta) + \sin(\alpha - \beta) \right] \tag{11.14}$$

$$\sin \alpha \sin \beta = \frac{1}{2} \left[\cos(\alpha - \beta) - \cos(\alpha + \beta) \right] \tag{11.15}$$

$$\cos \alpha \cos \beta = \frac{1}{2} \left[\cos(\alpha + \beta) + \cos(\alpha - \beta) \right] \tag{11.16}$$

Note

Using theorem 11.1 if follows:

$$\cos(\alpha \pm \pi) = -\cos \alpha \quad \text{and} \quad \sin(\alpha \pm \pi) = -\sin \alpha \tag{11.17}$$

0.4. Law of Cosines

Law 11.1 Law of Cosines

[proof 11.3]:

relates the three side of a *general* triangle to each other.

$$a^2 = b^2 + c^2 - 2bc \cos \theta_{b,c} \tag{11.18}$$

Law 11.2 Law of Cosines for Vectors

[proof 11.4]:

relates the length of vectors to each other.

$$\|\mathbf{a}\|^2 = \|\mathbf{c} - \mathbf{b}\|^2 = \|\mathbf{b}\|^2 + \|\mathbf{c}\|^2 - 2\|\mathbf{b}\|\|\mathbf{c}\| \cos \theta_{\mathbf{b},\mathbf{c}} \tag{11.19}$$

Law 11.3 Pythagorean theorem: special case of ?? for right triangle:

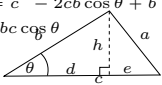
$$a^2 = b^2 + c^2 \tag{11.20}$$

1. Proofs

Proof 11.3: Law 11.1 From the definition of the sine and cosine we know that:

$$\sin \theta = \frac{h}{b} \Rightarrow \underline{h} \qquad \text{and} \qquad \cos \theta = \frac{d}{b} \Rightarrow \underline{d}$$

$$\begin{aligned} \underline{e} &= c - \underline{d} = c - b \cos \theta \\ a^2 &= \underline{e}^2 + \underline{h}^2 = c^2 - 2cb \cos \theta + b^2 \cos^2 \theta + b^2 \sin^2 \theta \\ &= c^2 + b^2 - 2bc \cos \theta \end{aligned}$$



Proof 11.4: Law 11.2 Notice that $\mathbf{c} = \mathbf{a} + \mathbf{b} \Rightarrow \mathbf{a} = \mathbf{c} - \mathbf{b}$ and we can either use ?? 11.3 or notice that:

$$\begin{aligned} \|\mathbf{c} - \mathbf{b}\|^2 &= (\mathbf{c} - \mathbf{b}) \cdot (\mathbf{c} - \mathbf{b}) \\ &= \mathbf{c} \cdot \mathbf{c} - 2\mathbf{c} \cdot \mathbf{b} + \mathbf{b} \cdot \mathbf{b} \\ &= \|\mathbf{c}\|^2 + \|\mathbf{b}\|^2 - 2(\|\mathbf{c}\|\|\mathbf{b}\| \cos \theta) \end{aligned}$$

Topology

Definition 12.1 Topology of set

τ :

Let X be a set. A collection τ of open?? subsets of X is called *topology* of X if it satisfies:

- $\emptyset \in \tau$ and $X \in \tau$
- Any finite or infinite union of subsets of τ is contained in τ :

$$\{U_i : i \in \mathbf{I}\} \subseteq \tau \qquad \Longrightarrow \qquad \cup_{i \in \mathbf{I}} U_i \in \tau \tag{12.1}$$

- The intersection of a finite number of elements of τ also belongs to τ :

$$\{U_i\}_{i=1}^n \in \tau \qquad \Longrightarrow \qquad U_1 \cap \dots \cap U_n \in \tau \tag{12.2}$$

Definition 12.2 Topological Space[?]

(X, τ) :

Is an ordered pair (X, τ) , where X is a set and τ is a topology^[def. 12.1] on X .

Numerical Methods

Machine Arithmetic's

Machine/Floating Point Numbers

Definition 13.1 (IEEE) **Institute of Electrical and Electronics Engineers:**
Is a engineering associations that defines a standard on how computers should treat machine numbers in order to have certain guarantees.

Definition 13.2 Machine/Floating Point Numbers **M:**
Computers are only capable to represent a *finite, discrete* set of the real numbers $\mathbb{M} \subset \mathbb{R}$

1.1.1. Floating Point Arithmetic's $x\tilde{\Omega}y = \mathfrak{fl}(x\Omega y)$

Corollary 13.1 Closure:
Machine numbers \mathbb{F} are not *closed*^[def. 1.7] under basic arithmetic operations:
$$\mathbb{F} \Omega \mathbb{F} \mapsto \not\mathbb{F} \quad \Omega = \{+, -, *, /\}$$
 (13.1)

Note
Corollary 13.1 provides a problem as the computer can only represent floating point number \mathbb{F} .

Definition 13.3 Overflow: Result is bigger then the biggest representable floating point number.

Definition 13.4 Underflow: Result is smaller then the smaller representable floating point number i.e. to close to zero.

1.1.2. The Rounding Unit

Definition 13.5 **Rounding Function/Unit** rd/\sim :
Let $x \in \mathbb{K}$ be a number real or complex number. The rounding function approximates x by the nearest machine number $\tilde{x} \in \mathbb{F}$:

$$\text{rd} : \begin{cases} \mathbb{R} \mapsto \mathbb{F} \\ x \mapsto \max \arg \min_{\tilde{x} \in \mathbb{F}} |x - \tilde{x}| \end{cases} \quad (13.2)$$

Notes

- If this is ambiguous (there are two possibilities), then it takes the larger one:
- Basic arithmetic rules such as associativity do no longer hold for operations such as addition and subtraction.

Definition 13.6 Floating Point Operation $\tilde{\Omega}$:
Is a basic arithmetic operation between two floating point numbers $x \in \mathbb{F}$ rounded back to the nearest floating point number:
$$\mathbb{F} \tilde{\Omega} \mathbb{F} \mapsto \mathbb{F} \quad \tilde{\Omega} := \text{rd} \circ \Omega \quad \Omega = \{+, -, *, /\}$$
 (13.3)

Definition 13.7 Absolute Error: Let $\tilde{x} \in \mathbb{K}$ be an approximation of $x \in \mathbb{K}$ then the absolute error is defined by:
$$\epsilon_{\text{abs}} := |x - \tilde{x}|$$
 (13.4)

Definition 13.8 Relative Error: Let $\tilde{x} \in \mathbb{K}$ be an approximation of $x \in \mathbb{K}$ then the relative error is defined by:
$$\epsilon_{\text{abs}} := \frac{|x - \tilde{x}|}{|x|} \quad (13.5)$$

Note
We are interested in the relative error as it controls the number of *correct/significant* digits l of the approximation \tilde{x} of $x \in \mathbb{K}$:
$$\epsilon_{\text{abs}} := \frac{|x - \tilde{x}|}{|x|} \leq 10^{-l} \quad l \in \mathbb{N}_{>0} \quad (13.6)$$

1.1.3. The Machine Epsilon

Definition 13.9 **The Machine Epsilon:** **EPS**
The machine epsilon EPS is the largest possible *relative* rounding error^[def. 13.8]:
$$\text{EPS} := \max_{x \in I \setminus 0} \frac{|\text{rd}(x) - x|}{|x|} \quad I := [\min|\mathbb{M}|, \max|\mathbb{M}|] \in \mathbb{K} \quad (13.7)$$

Corollary 13.2 Relative Error of Flop:
The *relative* error^[def. 13.8] of any floating point operation^[def. 13.6] is bounded by the machine epsilon^[def. 13.9]:
$$\text{EPS}_{\text{rel}} \left(\tilde{\Omega}(x, y) \right) := \frac{|\tilde{\Omega}(x, y) - \Omega(x, y)|}{|\Omega(x, y)|} = \frac{|(\text{rd} - \text{I}) \Omega(x, y)|}{|\Omega(x, y)|} \leq \text{EPS} \quad (13.8)$$

Corollary 13.3 EPS for Machine Number: For machine numbers EPS can be computed by:
$$\text{EPS} = \frac{1}{2} B^{1-m} \quad (13.9)$$

Type	EPS
double	$2.2 \cdot 10^{-16}$
float	$1.1 \cdot 10^{-23}$
FP16	$9.76 \cdot 10^{-4}$

Axiom of Round off Analysis

Axiom 13.1 Axiom of Round off Analysis:
Let $x, y \in \mathbb{F}$ be (normalized) floats and assume that $x\tilde{\Omega}y \in \mathbb{F}$ (i.e. no over/underflow). Then it holds that:
$$x\tilde{\Omega}y = (x\Omega y) (1 + \delta) \quad \Omega = \{+, -, *, /\} \quad (13.10)$$

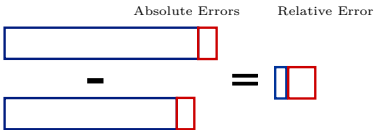
$$\tilde{f}(x) = f(x)(1 + \delta) \quad f \in \{\exp, \sin, \cos, \log, \dots\}$$

with $|\delta| < \text{EPS}$

Explanation 13.1 (axiom 13.1). *gives us a guarantee that for any two floating point numbers $x, y \in \mathbb{F}$, any operation involving them will give a floating point result which is within a factor of $1 + \delta$ of the true result $x\Omega y$.*

1.1.4. Cancellation

Definition 13.10 Cancellation:
Is the extreme amplification of *relative* errors^[def. 13.8] when subtracting numbers of almost equal size.



Roundoff Errors

2.0.1. Tricks

Log-Sum-Exp Trick

The sum exponential trick is at trick that helps to calculate the log-sum-exponential in a robust way by avoiding over/underflow. The log-sum-exponential^[def. 13.11] is an expression that arises frequently in machine learning i.e. for the cross entropy loss or for calculating the evidence of a posterior prediction.
The root of the problem is that we need to calculate the exponential $\exp(x)$, this comes with two different problems:

- If x is large (i.e. 89 for single precision floats) then $\exp(x)$ will lead to overflow
- If x is very negative $\exp(x)$ will lead to underflow/0. This is not necessarily a problem but if $\exp(x)$ occurs in the denominator or the logarithm for example this is catastrophic.

Definition 13.11 Log sum Exponential:
$$\text{LogSumExp}(x_1, \dots, x_n) := \log \left(\sum_{i=1}^n e^{x_i} \right) \quad (13.11)$$

Formula 13.1 [proof 13.3]
Log-Sum-Exp Trick:
$$\log \left(\sum_{i=1}^n e^{x_i} \right) = a + \log \sum_{i=1}^n e^{x_i - a} \quad a := \max_{i \in \{1, \dots, n\}} x_i \quad (13.12)$$

Explanation 13.2 (formula 13.1). *The value a can be any real value but for robustness one usually chooses the max s.t.*

- The leading digits are preserved by pulling out the maximum a .
- Inside the log only zero or negative numbers are exponentiated, so there can be no overflow.
- If there is underflow inside the log we know that at least the leading digits have been returned by the max.

Definition 13.12 Partition Π :
Given an interval $[0, T]$ a sequence of values $0 < t_0 < \dots < t_n < T$ is called a partition $\Pi(t_0, \dots, t_n)$ of this interval.

Asymptotic Complexity

3.1. O-Notation

3.1.1. Small $o(\cdot)$ Notation

Definition 13.13 Little o Notation:
$$f(n) = o(g(n)) \iff \lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0 \quad (13.13)$$

3.1.2. Big $\mathcal{O}(\cdot)$ Notation

3.2. Basic Operations

4. Rate Of Convergence

Definition 13.14 Rate of Convergence: Is a way to measure the rate of convergence of a sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ to a value to \mathbf{x}^* . Let $\rho \in [0, 1]$ be the *rate of convergence* and define:

$$\lim_{k \mapsto \infty} \frac{\left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\|}{\left\| \mathbf{x}^k - \mathbf{x}^* \right\|} = \rho \tag{13.14}$$
$$\iff \lim_{k \mapsto \infty} \left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\| \leq \rho \left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\| \quad \forall k \in \mathbb{N}_0$$

Definition 13.15 Linear/Exponential Convergence:
A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *linearly* to \mathbf{x}^* if in the asymptotic limit $k \rightarrow \infty$ if it satisfies:

$$\rho \in (0, 1) \qquad \qquad \qquad \forall k \in \mathbb{N}_0 \tag{13.15}$$

Definition 13.16 Superlinear Convergence:
A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *superlinear* to \mathbf{x}^* if in the asymptotic limit $k \rightarrow \infty$ if it satisfies:

$$\rho = 1 \tag{13.16}$$

Definition 13.17 Sublinear Convergence:
A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *sublinear* to \mathbf{x}^* if in the asymptotic limit $k \rightarrow \infty$ if it satisfies:

$$\rho = 0 \quad \iff \quad \left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\| = o \left(\left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\| \right) \tag{13.17}$$

Definition 13.18 Logarithmic Convergence:
A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *logarithmically* to \mathbf{x}^* if it converges *sublinear*^[def. 13.17] and additoinally satisfies

$$\rho = 0 \quad \iff \quad \left\| \mathbf{x}^{k+2} - \mathbf{x}^{k+1} \right\| = o \left(\left\| \mathbf{x}^{k+1} - \mathbf{x}^k \right\| \right) \tag{13.18}$$

add explanation why

Exponetial Convergence

Linear convergence is sometimes called exponential convergence. This is due to the fact that:

1. We often have expressions of the form:
- $$\left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\| \leq \underbrace{(1 - \alpha)}_{:= \rho} \left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\|$$
2. and that $(1 - \alpha) = \exp(-\alpha)$ from which follows that:
- eq. (13.19)
$$\iff \left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\| \leq e^{-\alpha} \left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\|$$

Definition 13.19 Convergence of order p : In order to distinguish *superlinear convergence* we define the order of convergence.
A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges superlinear with order $p \in \{2, \dots\}$ to \mathbf{x}^* if it satisfies:

$$\lim_{k \mapsto \infty} \frac{\left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\|}{\left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\|^p} = C \qquad C < 1 \tag{13.19}$$

Does this even exist/check if this is true

Definition 13.20 Exponential Convergence: A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges exponentially with rate ρ to \mathbf{x}^* if in the asymptotic limit $k \rightarrow \infty$ it satisfies:

$$\left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\| \leq \rho^k \left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\| \qquad \rho < 1 \tag{13.20}$$

$$\left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\| \in o \left(\left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\| \right) \tag{13.21}$$

5. Basic Operations

Operation	#mul/div	#add/sub	asympt. comp
Dot Prod.	n	$n - 1$	$\mathcal{O}(n)$
Tensor Prod.	nm	0	$\mathcal{O}(nm)$
Matrix Prod.	mnk	$mk(n - 1)$	$\mathcal{O}(nmk)$

Linear Systems of Equations

6.1. Direct Methods

6.1.1. Gaussian Elimination

Definition 13.21 Pivot Elements $a_{11}, a_{22}, \dots, a_{nn}$:
Are the diagonal elements of $\mathbf{A} \in \mathbb{R}^{n,n}$ that we use to zero out the column below.

Definition 13.22 Row Echelon Matrix: Is a rectangular matrix where:

- All non-zero rows are above any zero rows.
- Each pivot of a row has a larger column index then the pivot of the row above.
- All entries below a pivot are zero.

Corollary 13.4 Reduced Form Row Echelon Matirx: Is an echelon matrix^[def. 13.22] where:

- The leading entry in each non-zero row equals 1.
- Each leading one is the only entry in its colmun.

Note

In case of square matrix this is a unit diagonal matrix.

Definition 13.23 **Gaussian Elimination** $\mathbf{A} \in \mathbb{R}^{n,n}, \mathcal{O}(n^3)$:
Is an algorithm to solve linear systems of equations:

$\mathbf{Ax} = \mathbf{b}$

\iff

$a_{11}x_1 + a_{21}x_1 + \dots + a_{n1}x_1$

$+ a_{12}x_2 + a_{22}x_2 + \dots + a_{n2}x_2$

$+ \dots +$

$+ a_{1n}x_n + a_{2n}x_n + \dots + a_{nn}x_n$

$= b_1$

$= b_2$

$= b_n$

and consists of two steps:

① Forward Elimination $\mathcal{O}(n^3)$ – transforming \mathbf{A} into an upper diagonal form $[\mathbf{U}|\mathbf{b}^*]$:

$a_{11}x_1 + a_{22}^{(1)}x_2 + a_{33}^{(2)}x_3 + \dots + a_{nn}^{(n-1)}x_n$

$= b_1$

$= b_2$

$= b_3$

$= b_n$

② Back Substitution Elimination $\mathcal{O}(n^2)$ – calculating the unknown's \mathbf{x} from \mathbf{U} :

Gauss Jordan Elimination

Is in principle the same as Gauss elimination but reduce the matrix into row-reduced echelon form^[def. 13.22].

Forward Elimination

Algorithm 13.1 Forward Elimination:
Transforms $\mathbf{Ax} = \mathbf{b}$ into row-echelon form^[def. 13.22]:

Given:

```
1: for  $k = 1, \dots, n - 1$  do
2:   pivot  $\leftarrow \mathbf{A}(k, k)$ 
3:   for  $i = k + 1, \dots, n$  do
4:      $l_{ik} \leftarrow \frac{\mathbf{A}(i, k)}{\text{pivot}}$ 
5:     for  $j = k + 1, \dots, n$  do
6:        $a_{ij}^{(k)} = \mathbf{A}(i, j) - l_{ik} \mathbf{A}(k, j)$ 
7:     end for
8:   end for
```

Corollary 13.5 Complexity:

$$\sum_{i=1}^{n-1} (n-1)(2(n-i) + 3) = n(n-1) \left(\frac{2}{3}n + \frac{7}{6} \right) = \mathcal{O} \left(\frac{2}{3}n^3 \right)$$

(13.22)

Backward Substitution

Algorithm 13.2 Backward Substitution:

Given \mathbf{U} :

```
1:  $x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}$ 
2: for  $i = n - 1, n - 2, \dots, 1$  do
3:   
$$x_i = \frac{b_i^{(i-1)} - \sum_{j=i+1}^n a_{ij}^{(i-1)} x_j}{a_{ii}^{(i-1)}}$$

4: end for
```

Corollary 13.6 Complexity:

$$\sum_{i=1}^{n-1} 2(n-i) + 1 = \mathcal{O} \left(n^2 \right)$$

(13.23)

By Rank-1 Modifications

6.1.2. LU-Decomposition

Definition 13.24 LU Decomposition $\mathcal{O}(n^3)$:
Decomposes a matrix \mathbf{A} in an upper and lower triangular part in order to solve a system of linear equations.
Given: $\mathbf{PA} = \mathbf{LU}$ we can compute:

① $\mathbf{Ly} = \mathbf{Pb}$

② $\mathbf{Ux} = \mathbf{y}$

Corollary 13.7 [proof ??]
LU decomposition Complexity:

$$\frac{2}{3}n^3 + \frac{1}{3}n^2$$

Solving Multiple Systems of Equations

6.1.3. Symmetric Matrices

LDL-Decomposition

6.1.4. Symmetric Positive Definite Matrices

For linear systems with s.p.d.^[def. 10.75] matrices \mathbf{A} the LU-decomposition^[def. 13.24] simplifies to the Cholesky Decomposition^[def. 13.25].

Cholesky Decomposition

Definition 13.25 Cholesky Decomposition $\frac{1}{3}\mathcal{O}(n^3)$:
Let \mathbf{A} be a s.p.d.^[def. 10.75] then it can be factorized into:

$$\mathbf{A} = \mathbf{GG}^T \quad \text{with} \quad \mathbf{G} := \mathbf{LD}^{1/2}$$

(13.24)

Corollary 13.8 [proof 13.5]
Cholesky decomposition Complexity:

$$\frac{1}{3}n^3 + \frac{1}{3}n^2$$

6.2. Iterative Methods

7. Non-linear Systems of Equations

7.1. Iterative Methods

Definition 13.26

General Non-linear System of Equations (NLSE) F :
Is a system of non-linear equations F (that do **not** satisfy linearity??):
$$F:\subseteq \mathbb{R}^n \mapsto \mathbb{R}^n \quad \text{seek to find} \quad \mathbf{x} \in \mathbb{R}^n : F(\mathbf{x}) = \mathbf{0} \quad (13.25)$$

Definition 13.27 Stationary m -point Iteration ϕ_F :
Let $n, m \in \mathbb{R}$ and let $U \subseteq (R^n)^m = \mathbb{R}^n \times \dots \times \mathbb{R}^n$ be a set.
A function $\phi : U \mapsto \mathbb{R}^n$, is called (m -point) iteration function
if it produces an iterative sequence $\left(\mathbf{x}^{(k)}\right)_k$ of approximate
solutions to eq. (13.25), using the m most recent iterates:
$$\mathbf{x}^{(k)} = \phi_F\left(\mathbf{x}^{(k-1)}, \dots, \mathbf{x}^{(k-m)}\right) \quad (13.26)$$

Inital Guess $\mathbf{x}^{(0)}, \dots, \mathbf{x}^{(m-1)}$

Note

Stationary as ϕ does no explicitly depend on k .

Definition 13.28 Fixed Point \mathbf{x}^* :
Is a point \mathbf{x}^* for which the sequence does not change any-
more:
$$\mathbf{x}^{(k-1)} = \mathbf{x}^* \\ \vdots \\ \mathbf{x}^{(k-m)} = \mathbf{x}^* \\ (13.27)$$

$$\mathbf{x}^* = \phi_F\left(\mathbf{x}^{(k-1)}, \dots, \mathbf{x}^{(k-m)}\right) \quad \text{with}$$

7.1.1. Convergence

Question

Does the sequence $\left(\mathbf{x}^{(k)}\right)_k$ converge to a limit:
$$\lim_{k \rightarrow \infty} \mathbf{x}^{(k)} = \mathbf{x}^* \quad (13.28)$$

7.1.2. Consistency

Definition 13.29 Consistent m -point Iterative Method:
A stationary m -point method^[def. 13.27] is consistent with a non-
linear system of equations^[def. 13.26] F iff:
$$F\left(\mathbf{x}^*\right) \iff \phi_F\left(\mathbf{x}^*, \dots, \mathbf{x}^*\right) = \mathbf{x}^* \quad (13.29)$$

7.1.3. Speed of Convergence

add cvg, consistency, speed of cvng...

7.2. Fixed Point Iterations $m = 1$

Definition 13.30 Fixed Point Iteration: Is a 1-point
method $\phi_F : U \subset \mathbb{R}^n \mapsto \mathbb{R}^n$ that seeks a fixed point \mathbf{x}^*
to solve $F(\mathbf{x}) = 0$:
$$\mathbf{x}^{(k+1)} = \phi_F\left(\mathbf{x}^{(k)}\right) \quad \text{Inital Guess: } \mathbf{x}^{(0)} \quad (13.30)$$

Corollary 13.9 Consistency: If ϕ_F is continuous and $\mathbf{x}^* = \lim_{k \rightarrow \infty} x^{(k)}$ then \mathbf{x}^* is a fixed point^[def. 13.28] of ϕ .

Algorithm 13.3 Fixed Point Iteration:

Input: Inital Guess: $\mathbf{x}^{(0)}$
1: Rewrite $F(\mathbf{x}) = 0$ into a form of $\mathbf{x} = \phi_F(\mathbf{x})$
 \triangleright There exist many ways
2: **for** $k = 1, \dots, T$ **do**
3: Use the fixed point method:

$$\mathbf{x}^{(k+1)} = \phi_F\left(\mathbf{x}^{(k)}\right) \quad (13.31)$$

4: **end for**

add examples and rest

8. Numerical Quadrature

Definition 13.31 Order of a Quadrature Rule:
The **order** of a quadrature rule $\mathcal{Q}_n : C^0([a, b]) \rightarrow \mathbb{R}$ is defined as:
$$\text{order}(\mathcal{Q}_n) := \max \left\{ n \in \mathbb{N}_0 : \mathcal{Q}_n(p) = \int_a^b p(t) dt \quad \forall p \in \mathcal{P}_n \right\} + 1 \tag{13.32}$$

Thus it is the maximal degree+1 of polynomials (of degree maximal degree) $\mathcal{P}_{\text{maximal degree}}$ for which the quadrature rule yields exact results.

Note
Is a quality measure for quadrature rules.

8.1. Composite Quadrature

Definition 13.32 Composite Quadrature:
Given a mesh $\mathcal{M} = \{a = x_0 < x_1 < \dots < x_m = b\}$ apply a Q.R. \mathcal{Q}_n to each of the mesh cells $I_j := [x_{j-1}, x_j] \quad \forall j = 1, \dots, m \triangleq \text{p.w.}$ Quadrature:
$$\int_a^b f(t) dt = \sum_{j=1}^m \int_{x_{j-1}}^{x_j} f(t) dt = \sum_{j=1}^m \mathcal{Q}_n(f|_{I_j}) \tag{13.33}$$

Lemma 13.1 Error of Composite quadrature Rules:
Given a function $f \in C^k([a, b])$ with integration domain:
$$\sum_{i=1}^m h_i = |b - a| \quad \text{for } \mathcal{M} = \{x_j\}_{j=1}^m$$

Let: $h_{\mathcal{M}} = \max_j |x_j, x_{j-1}|$ be the **mesh-width**
Assume an equal number of quadrature nodes for each interval $I_j = [x_{j-1}, x_j]$ of the mesh \mathcal{M} i.e. $n_j = n$.
Then the error of a quadrature rule $\mathcal{Q}_n(f)$ of order q is given by:
$$\begin{aligned} \epsilon_n(f) &= \mathcal{O}\left(n^{-\min\{k, q\}}\right) = \mathcal{O}\left(h_{\mathcal{M}}^{\min\{k, q\}}\right) \quad \text{for } n \rightarrow \infty \\ &\stackrel{[\text{cor. 5.6}]}{=} \mathcal{O}\left(n^{-q}\right) = \mathcal{O}\left(h_{\mathcal{M}}^q\right) \quad \text{with } h_{\mathcal{M}} = \frac{1}{n} \end{aligned} \tag{13.34}$$

Definition 13.33 Complexity W : Is the number of function evaluations \triangleq number of quadrature points.
$$W(\mathcal{Q}(f)_n) = \# \text{f-eval} \triangleq n \tag{13.35}$$

Lemma 13.2 Error-Complexity $W(\epsilon_n(f))$: Relates the complexity to the quadrature error.
Assuming and quadrature error of the form :
$$\epsilon_n(f) = \mathcal{O}(n^{-q}) \iff \epsilon_n(f) = cn^{-q} \quad c \in \mathbb{R}_+$$

the error complexity is **algebraic** (??) and is given by:
$$W(\epsilon_n(f)) = \mathcal{O}(\epsilon_n^{1/q}) = \mathcal{O}\left(\sqrt[q]{\epsilon_n}\right) \tag{13.36}$$

Proof 13.1: lemma 13.2: **Assume:** we want to reduce the error by a factor of ϵ_n by increasing the number of quadrature points $n_{\text{new}} = a \cdot n_{\text{old}}$.
Question: what is the additional effort ($\#$ f-eval) needed in order to achieve this reduction in error?
$$\frac{c \cdot n_n^q}{c \cdot n_o^q} = \frac{1}{\epsilon_n} \implies n_n = n_o \cdot \sqrt[q]{\epsilon_n} = \mathcal{O}(\sqrt[q]{\epsilon_n}) \tag{13.37}$$

8.1.1. Simpson Integration

Definition 13.34 Simpson Integration:

Filtering Algorithms

10. Signals

Definition 13.35 Time Discrete Signal: Is a bounded sequence^[def. 2.2] $(x_j)_{j \in \mathbb{Z}} \in l^\infty(\mathbb{Z})$.

Definition 13.36 Sampling:

Corollary 13.10 Finite Time Discrete Signal:

11. Channels/Filters

Definition 13.37 Channel/Filter: F
Is a mapping of signals to signals $F ::$
$$F : l^\infty(\mathbb{Z}) \mapsto l^\infty(\mathbb{Z}) \tag{13.38}$$

Property 13.1 Finite Channel/Filter: A filter $F : l^\infty(\mathbb{Z}) \mapsto l^\infty(\mathbb{Z})$

Property 13.2 Causal Channel/Filter:

Explanation 13.3. *The response cannot start before the signal has been feed into the filter.*

Definition 13.38 Time Shift Operator: S_m

Property 13.3 Time-invariant Channel/Filter:

Explanation 13.4. *The response of the filter should not depend at which time we pass the signal to the filter.*

Property 13.4 Linear Channel/Filter:

Definition 13.39 Linear Time-invariant Finite Input Response Filter LT-FIR:

11.1. Impulse Responses

Definition 13.40 Impulse:

Definition 13.41 Impulse Response h:

Corollary 13.11 [proof 13.2]
Signal in terms of Impulse Responses: We can write any arbitrary discrete signal as weighted sum of time shifted impulses:
$$F(x_j) = \tag{13.39}$$
$$(F(x_j))_j = \tag{13.40}$$

Proof 13.2 ^[cor. 13.11]:

11.2. Discrete Convolution

Definition 13.42 LT-FIR formula:

Proofs

Proof 13.3 Log Sum Trickformula 13.1:
$$\begin{aligned} \text{LSE} &= \log \left(\sum_{i=1}^n e^{x_i} \right) = \log \left(\sum_{i=1}^n e^{x_i - a} e^a \right) \\ &= \log \left(e^a \sum_{i=1}^n e^{x_i - a} \right) = \log \left(\sum_{i=1}^n e^{x_i - a} \right) + \log(e^a) \\ &= \log \left(\sum_{i=1}^n e^{x_i - a} \right) + a \end{aligned}$$

Proof 13.4 LU-Complexity^[cor. 13.7]:
For eliminating the first column we need to eliminate $n - 1$ rows by n additions and n multiplications which equals $(n - 1)2n$. For the second column we need for $n - 2$ rows $n - 1$ additions and $n - 1$ multiplications which equals $(n - 2)2(n - 1)$ thus to eliminate all n columns we have:
$$\sum_{i=1}^n (n - i + 1) \cdot 2(n - i)$$
using the index $l = n - i + 1$ we can write this as:
$$\begin{aligned} \sum_{i=1}^n (n - i + 1) \cdot 2(n - i) &= 2 \sum_{l=0}^n (j + 1) \cdot (j) = 2 \sum_{l=0}^n j^2 + 1 \\ &= 2 \left(\frac{1}{3} n^3 - \frac{1}{3} n \right) \end{aligned}$$

add rules for sums of n and n^2

Proof 13.5 Cholesky Complexity^[cor. 13.8]: **U** and **L** “are the same” as we have a s.p.d. matrix s.t. we can simply half the forward elimination complexity of the LU-decomposition^[cor. 13.7]:
$$\frac{1}{2} \frac{2}{3} n^3 + \frac{1}{3} n^2 \tag{13.41}$$

Optimization

Definition 14.1 First Order Method: A first-order method is an algorithm that chooses the k -th iterate in $\mathbf{x}_0 + \text{span}\{\nabla f(\mathbf{x}_0), \dots, \nabla f(\mathbf{x}_{k-1})\} \quad \forall k = 1, 2, \dots \quad (14.1)$

Note

Gradient descent is a first order method

1. Linear Optimization

1.1. Polyhedra

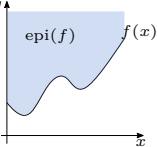
Definition 14.2 Polyhedron: Is a set $P \in \mathbb{R}^n$ that can be described by the *finite* intersection of m closed *half spaces*??:

$$P = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} \leq \mathbf{b}\} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}_j \mathbf{x} \leq b_j, j = 1, \dots, m\}$$
$$\mathbf{A} \in \mathbb{R}^{m \times n} \qquad \mathbf{b} \in \mathbb{R}^m \qquad (14.2)$$

1.1.1. Polyhedral Function

Definition 14.3 Epigraph/Subgraph **epi(f):**

The epigraph of a function $f \in \mathbb{R}^n \mapsto \mathbb{R}$ is defined as the set of point that lie above its graph:

$$\text{epi}(f) := \{(\mathbf{x}, y) \in \mathbb{R}^n \mid y \geq f(\mathbf{x})\} \subseteq \mathbb{R}^{n+1} \quad (14.3)$$


Definition 14.4 Polyhedral Function: A function f is *polyhedral* if its epigraph $\text{epi}(f)$ ^[def. 14.3] is a polyhedral set ^[def. 14.2]:

$$f \text{ is polyhedral} \iff \text{epi}(f) \text{ is polyhedral} \quad (14.4)$$

2. Lagrangian Optimization Theory

Add: derivation of lagrange function

Definition 14.5 (Primal) Constraint Optimization:

Given an optimization problem with domain $\Omega \subseteq \mathbb{R}^d$:

$$\begin{aligned} \min_{\mathbf{w} \in \Omega} f(\mathbf{w}) \\ \text{s.t.} \quad & g_i(\mathbf{w}) \leq 0 & 1 \leq i \leq k \\ & h_j(\mathbf{w}) = 0 & 1 \leq j \leq m \end{aligned}$$

Definition 14.6 Lagrange Function:

$$\mathcal{L}(\alpha, \beta, \mathbf{w}) := f(\mathbf{w}) + \alpha \mathbf{g}(\mathbf{w}) + \beta \mathbf{h}(\mathbf{w}) \quad (14.5)$$

Extremal Conditions

$$\begin{aligned} \nabla \mathcal{L}(\mathbf{x}) &\stackrel{!}{=} 0 && \text{Extremal point } \mathbf{x}^* \\ \frac{\partial}{\partial \beta} \mathcal{L}(\mathbf{x}) &= h(\mathbf{x}) \stackrel{!}{=} 0 && \text{Constraint satisfaction} \end{aligned}$$

For the inequality constraints $g(\mathbf{x}) \leq 0$ we distinguish two situations:

Case I : $g(\mathbf{x}^*) < 0$ switch const. off
Case II : $g(\mathbf{x}^*) \geq 0$ optimize using active eq. constr.

$$\frac{\partial}{\partial \alpha} \mathcal{L}(\mathbf{x}) = g(\mathbf{x}) \stackrel{!}{=} 0 \qquad \text{Constraint satisfaction}$$

Definition 14.7 Lagrangian Dual Problem: Is given by:

$$\begin{aligned} \text{Find} \quad & \max \theta(\alpha, \beta) = \inf_{\mathbf{w} \in \Omega} \mathcal{L}(\mathbf{w}, \alpha, \beta) \\ \text{s.t.} \quad & \alpha_i \geq 0 && 1 \leq i \leq k \end{aligned}$$

Solution Strategy

- Find the extremal point \mathbf{w}^* of $\mathcal{L}(\mathbf{w}, \alpha, \beta)$:
$$\left. \frac{\partial \mathcal{L}}{\partial \mathbf{w}} \right|_{\mathbf{w}=\mathbf{w}^*} \stackrel{!}{=} 0 \quad (14.6)$$
- Insert \mathbf{w}^* into \mathcal{L} and find the extremal point β^* of the resulting dual Lagrangian $\theta(\alpha, \beta)$ for the active constraints:
$$\left. \frac{\partial \theta}{\partial \beta} \right|_{\beta=\beta^*} \stackrel{!}{=} 0 \quad (14.7)$$
- Calculate the solution $\mathbf{w}^*(\beta^*)$ of the constraint minimization problem.

Value of the Problem

Value of the problem: the value $\theta(\alpha^*, \beta^*)$ is called the value of problem (α^*, β^*) .

Theorem 14.1 Upper Bound Dual Cost: Let $\mathbf{w} \in \Omega$ be a feasible solution of the primal problem ^[def. 14.5] and (α, β) a *feasible solution* of the respective dual problem ^[def. 14.7]. Then it holds that:

$$f(\mathbf{w}) \geq \theta(\alpha, \beta) \quad (14.8)$$

Proof 14.1:

$$\begin{aligned} \theta(\alpha, \beta) &= \inf_{\mathbf{u} \in \Omega} \mathcal{L}(\mathbf{u}, \alpha, \beta) \leq \mathcal{L}(\mathbf{w}, \alpha, \beta) \\ &= f(\mathbf{w}) + \sum_{i=1}^k \underbrace{\alpha_i}_{\geq 0} g_i(\mathbf{w}) + \sum_{j=1}^m \underbrace{\beta_j}_{=0} h_j(\mathbf{w}) \\ &\leq f(\mathbf{w}) \end{aligned}$$

Corollary 14.1 Duality Gap Corollary: The value of the dual problem is upper bounded by the value of the primal problem:

$$\sup \{\theta(\alpha, \beta) : \alpha \geq 0\} \leq \inf \{f(\mathbf{w}) : \mathbf{g}(\mathbf{w}) \leq 0, \mathbf{h}(\mathbf{w}) = 0\} \quad (14.9)$$

Theorem 14.2 Optimality: The triple $(\mathbf{w}^*, \alpha^*, \beta^*)$ is a saddle point of the Lagrangian function for the primal problem, if and only if its components are optimal solutions of the primal and dual problems and if there is no duality gap, that is, the primal and dual problems having the same value:

$$f(\mathbf{w}^*) = \theta(\alpha^*, \beta^*) \quad (14.10)$$

Definition 14.8 Convex Optimization: Given: a *convex* function f and a *convex set* S solve:

$$\begin{aligned} \min f(\mathbf{x}) \\ \text{s.t.} \quad \mathbf{x} \in S \end{aligned} \quad (14.11)$$

Often S is specified using linear inequalities:

$$\text{e.g.} \quad S = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{Ax} \leq \mathbf{b}\}$$

Theorem 14.3 Strong Duality: Given an convex optimization problem:

$$\begin{aligned} \min_{\mathbf{w} \in \Omega} f(\mathbf{w}) \\ \text{s.t.} \quad & g_i(\mathbf{w}) \leq 0 & 1 \leq i \leq k \\ & h_j(\mathbf{w}) = 0 & 1 \leq j \leq m \end{aligned}$$

where g_i, h_i can be written as affine functions: $y(\mathbf{w}) = \mathbf{Aw} - \mathbf{b}$.

Then it holds that the *duality gap* is zero and we obtain an optimal solution.

Theorem 14.4 Kuhn-Tucker Conditions: Given an optimization problem with convex domain $\Omega \subseteq \mathbb{R}^d$,

$$\begin{aligned} \min_{\mathbf{w} \in \Omega} f(\mathbf{w}) \\ \text{s.t.} \quad & g_i(\mathbf{w}) \leq 0 & 1 \leq i \leq k \\ & h_j(\mathbf{w}) = 0 & 1 \leq j \leq m \end{aligned}$$

with $f \in C^1$ convex and g_i, h_i affine.

Necessary and sufficient conditions for a normal point \mathbf{w}^* to be an optimum are the existence of α^*, β^* s.t.:

$$\frac{\partial \mathcal{L}(\mathbf{w}, \alpha, \beta)}{\partial \mathbf{w}} \stackrel{!}{=} 0 \qquad \frac{\partial \mathcal{L}(\mathbf{w}^*, \alpha, \beta)}{\partial \beta} \stackrel{!}{=} 0 \quad (14.12)$$

under the conditions that:

- $\forall i_1, \dots, k \quad \alpha_i^* g_i(\mathbf{w}^*) = 0$, s.t.:
 - Inactive Constraint: $g_i(\mathbf{w}^*) < 0 \rightarrow \alpha_i = 0$.
 - Active Constraint:
$$g_i(\mathbf{w}^*) \leq 0 \rightarrow \alpha_i \geq 0 \quad \text{s.t.} \quad \alpha_i^* g_i(\mathbf{w}^*) = 0$$

Consequence

We may become very sparse problems, if a lot of constraints are not active $\iff \alpha_i = 0$.

Only a few points, for which $\alpha_i > 0$ may affect the decision surface.

Combinatorics

Permutations

Definition 15.1 **Permutation:** A n -Permutation is the (re)arrangement of n elements of a set^[def. 1.1] \mathcal{S} of size $n = |\mathcal{S}|$ into a sequence^[def. 2.2] – **order does matter**.

Definition 15.2 **Number of Permutations of a Set** $n!$: Let \mathcal{S} be a set^[def. 1.1] $n = |\mathcal{S}|$ *distinct* objects. The number of permutations of \mathcal{S} is given by:

$$P_n(\mathcal{S}) = n! = \prod_{i=0}^{n-1} (n - i) = n \cdot (n - 1) \cdot (n - 2) \cdot \dots \cdot 1 \tag{15.1}$$

Explanation 15.1. If we have i.e. three distinct elements $\{\bullet, \circ, \color{red}\bullet\}$ For the first element \bullet that we arrange we have three possible choices where to put it. However this reduces the number of possible choices for the second element \circ to only two. Consequently for the last element $\color{red}\bullet$ we have no choice left.



Definition 15.3 **Number of Permutations of a Multiset:** Let \mathcal{S} be a multi set^[def. 1.3] with $n = |\mathcal{S}|$ total and k *distinct* objects. Let n_j be the multiplicity^[def. 1.4] of the member $j \in \{1, \dots, k\}$ of the multiset \mathcal{S} . The permutation of \mathcal{S} is given by:

$$P_{n_1, \dots, n_k}(\mathcal{S}) = \frac{n!}{n_1! \cdot \dots \cdot n_k!} \quad \text{s.t.} \quad \sum_{j=1}^k n_j \leq n \quad k < n \tag{15.2}$$

Note

We need to divide by the permutations as sequence/order does not change if we exchange objects of the same kind (e.g. red ball by red ball) \Rightarrow less possibilities to arrange the elements uniquely.

Picking things from a bag

1. Combinations

Definition 15.4 **k -Combination:** A k -combination of a set \mathcal{S} of *distinct* elements of size $n = \mathcal{S}$ is a subset \mathcal{S}_k (**order does not matter**) of $k = |\mathcal{S}_k|$, *chosen* from \mathcal{S} .

Note

Thus unlike in a permutation we just care about what we pick and not how it ends up beeing arranged.

Definition 15.5 **Number of k -Combinations** $C_{n,k}$: The number of k -combinations of a set \mathcal{S} of size $n = \mathcal{S}$ is given by:

$$C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n - k)!} \tag{15.3}$$

2. Variation

Definition 15.6 **Variation:** A k -variation of a set \mathcal{S} of size $n = \mathcal{S}$ is

- a selection/combination^[def. 15.4] of a subset \mathcal{S}_k (order does not matter) of k -*distinct* elements $k = |\mathcal{S}_k|$, *chosen* from \mathcal{S}
- and an k arrangement/permutation^[def. 15.2] of that subset \mathcal{S}_k (with or without repetition) into a sequence^[def. 2.2]

Definition 15.7 **Number of Variations without repetitions** V_k^n : Let \mathcal{S} be a set^[def. 1.1] $n = |\mathcal{S}|$ *distinct* objects from which we choose k elements. The number of variations of size $k = |\mathcal{S}_k|$ of the set \mathcal{S} *without repetitions* is given by:

$$V_k^n(\mathcal{S}) = \binom{n}{k} k! = \frac{n!}{(n - k)!} \tag{15.4}$$

Note

Sometimes also denotes as P_k^n .

Definition 15.8 **Number of Variations with repetitions** \bar{V}_k^n : Let \mathcal{S} be a set^[def. 1.1] $n = |\mathcal{S}|$ *distinct* objects from which we choose k elements. The number of variations of size $k = |\mathcal{S}_k|$ of the set \mathcal{S} from which we *choose and always return* is given by:

$$\bar{V}_k^n(\mathcal{S}) = n^k \tag{15.5}$$

Stochastics

Definition 15.9 Stochastics:	Is a collective term for the areas of <i>probability theory</i> and <i>statistics</i> .
Definition 15.10 Statistics:	Is concerned with the analysis of data/experiments in order to draw conclusion of the underlying governing models that describe these experiments.
Definition 15.11 Probability:	Is concerned with the quantification of the uncertainty of random experiments by use of statistical models. Hence it is the opposite of statistics.
Definition 15.12 Probability:	Probability is the measure of the likelihood that an event will occur in a Random Experiment. Probability is quantified as a number between 0 and 1, where, loosely speaking, 0 indicates impossibility and 1 indicates certainty.
Improve those definitions, maybe ask on quora/hh	
Note: Stochastics vs. Stochastic	
Stochastics is a noun and is a collective term for the areas of probability theory and statistics, while stochastic is a <i>adjective</i> , describing that a certain phenomena is governed by uncertainty i.e. a process.	
Probability Theory	
Definition 16.1 Probability Space	$W = \{\Omega, \mathcal{F}, \mathbb{P}\}$: Is the unique triple $\{\Omega, \mathcal{F}, \mathbb{P}\}$, where Ω is its sample space, \mathcal{F} its σ -algebra of events, and \mathbb{P} its probability measure.
Definition 16.2 Sample Space Ω :	[example 16.1] Is the set of all possible outcomes (elementary events [cor. 16.5]) of an experiment.
Definition 16.3 Event	[example 16.2] An “event” is a subset of the sample space Ω and is a property which can be observed to hold or not to hold <i>after</i> the experiment is done. Mathematically speaking not every subset of Ω is an event and has an associated probability. Only those subsets of Ω that are part of the corresponding σ -algebra \mathcal{F} are events and have their assigned probability.
Corollary 16.1 : If the outcome ω of an experiment is in the subset A , then the event A is said to “have occurred”.	
Corollary 16.2 Complement Set A^C : is the contrary event of A .	
Corollary 16.3 The Union Set $A \cup B$: Let A, B be two events. The event “ A or B ” is interpreted as the union of both.	
Corollary 16.4 The Intersection Set $A \cap B$: Let A, B be two events. The event “ A and B ” is interpreted as the intersection of both.	
Corollary 16.5 The Elementary Event ω : Is a “singleton”, i.e. a subset $\{\omega\}$ containing a single outcome ω of Ω .	
Corollary 16.6 The Sure Event Ω : Is equal to the sample space as it contains all possible elementary events.	
Corollary 16.7 The Impossible Event \emptyset : The impossible event i.e. nothing is happening is denoted by the empty set.	
Definition 16.4 The Family of All Events $\mathcal{A}/2^\Omega$: The set of all subset of the sample space Ω called family of all events is given by the power set of the sample space $\mathcal{A} = 2^\Omega$ (for finite sample spaces).	

Definition 16.5 Probability $\mathbb{P}(A)$:	Is a number associated with every A , that measures the likelihood of the event to be realized “a priori”. The bigger the number the more likely the event will happen. 1. $0 \leq \mathbb{P}(A) \leq 1$ 2. $\mathbb{P}(\Omega) = 1$ 3. If $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$
Note	We can think of the probability of an event A as the limit of the "frequency" of repeated experiments: $\mathbb{P}(A) = \lim_{n \rightarrow \infty} \frac{\delta_n(A)}{n} \quad \text{where} \quad \delta(A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$
1. Sigma Algebras	
Definition 16.6 Sigma Algebra	[Proof 16.3] σ : A set \mathcal{F} of subsets of Ω is called a σ -algebra on Ω if the following properties apply • $\Omega \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$ • If $A \in \mathcal{F}$ then $\Omega \setminus A = A^C \in \mathcal{F}$: The complementary subset of A is also in Ω . • For all $A_i \in \mathcal{F} : \bigcup_{i=1}^\infty A_i \in \mathcal{F}$
Explanation 16.1 ([def. 16.6]). <i>The σ-algebra determines what events we can measure, it represents all of the possible events of the experiment that we can detect. Thus the sigma algebra is a mathematical construct that tells us how much information we obtain once we conduct some experiment.</i>	
Corollary 16.8 \mathcal{F}_{\min}: $\mathcal{F} = \{\emptyset, \Omega\}$ is the simplest σ -algebra, telling us only if an event happened $\omega \in \Omega$ happened or not but not which one.	
Corollary 16.9 \mathcal{F}_{\max}: $\mathcal{F} = 2^\Omega$ consists of all subsets of Ω and thus corresponds to full information i.e. we know if and which event happened.	
Definition 16.7 Measurable Space (Ω, \mathcal{F}) :	Is the pair of a set and sigma algebra i.e. a sample space and sigma algebra $\{\Omega, \mathcal{F}\}$.
Corollary 16.10 \mathcal{F}-measurable Event $A_i \in \mathcal{F}$:	The measurable events A_i of \mathcal{F} are called <i>\mathcal{F}-measurable or measurable sets</i> .
Definition 16.8 Sigma Algebra generated by a subset of Ω $\sigma(C)$:	[Example 16.4] Let C be a class of subsets of Ω . The σ -algebra generated by C , denoted by $\sigma(C)$, is the <i>smallest</i> sigma algebra \mathcal{F} that included all elements of C .
Definition 16.9 Borel σ-algebra $\mathcal{B}(\mathbb{R})$:	[Example 16.5] The Borel σ -algebra $\mathcal{B}(\mathbb{R})$ is the smallest σ -algebra containing all open intervals in \mathbb{R} . The sets in contained in $\mathcal{B}(\mathbb{R})$ are called Borel sets. The extension to the multi-dimensional case, $\mathcal{B}(\mathbb{R}^n)$, is straightforward. For all real numbers $a, b \in \mathbb{R}$, $\mathcal{B}(\mathbb{R})$ contains various sets.
Why do we need Borel Sets	
So far we only looked at atomic events ω , with the help of sigma algebras we are now able to measure continuous events s.a. $[0, 1]$.	
Definition 16.10 Borel Set:	
Corollary 16.11 Generating Borel σ-Algebra [Proof 16.1]: The Borel σ -algebra of \mathbb{R} is generated by intervals of the form $(-\infty, a]$, where $a \in \mathbb{Q}$ (\mathbb{Q} =rationals).	

Definition 16.11 (\mathbb{P})-trivial Sigma Algebra:	is a σ -algebra \mathcal{F} for which each event has a probability of zero or one: $\mathbb{P}(A) \in \{0, 1\} \quad \forall A \in \mathcal{F} \quad (16.1)$
Interpretation	A trivial sigma algebra means that all events are almost surely constant and that there exist no non-trivial information. An example of a trivial sigma algebra is $\mathcal{F}_{\min} = \{\Omega, \emptyset\}$.
2. Measures	
Definition 16.12 Measure μ :	A measure defined on a measurable space $\{\Omega, \mathcal{F}\}$ is a function/map: $\mu : \mathcal{F} \mapsto [0, \infty]$ (16.2) for which holds: • $\mu(\emptyset) = 0$ • countable additivity [def. 16.13]
Definition 16.13 Countable/σ-Additive Function:	Given a function μ defined on a σ -algebra \mathcal{F} . The function μ is said to be countable additive if for every countable sequence of pairwise disjoint elements $(F_i)_{i \geq 1}$ of \mathcal{F} it holds that: $\mu\left(\bigcup_{i=1}^\infty F_i\right) = \sum_{i=1}^\infty \mu(F_i) \quad \text{for all} \quad F_j \cap F_k = \emptyset \quad \forall j \neq k$ (16.3)
Corollary 16.12 Additive Function: A function that satisfies countable additivity, is also additive, meaning that for every $F, G \in \mathcal{F}$ it holds: $F \cap G = \emptyset \implies \mu(F \cup G) = \mu(F) + \mu(G) \quad (16.4)$	
Explanation 16.2. <i>If we take two events that cannot occur simultaneously, then the probability that at least one of the events occurs is just the sum of the measures (probabilities) of the original events.</i>	
Definition 16.14 [Example 16.6] Equivalent Measures $\mu \sim \nu$:	Let μ and ν be two measures defined on a measurable space [def. 16.7] (Ω, \mathcal{F}) . The two measures are said to be equivalent if it holds that: $\mu(A) > 0 \iff \nu(A) > 0 \quad \forall A \subseteq \mathcal{F} \quad (16.5)$ this is equivalent to μ and ν having equivalent null sets: $\mathcal{N}_\mu = \mathcal{N}_\nu \quad \mathcal{N}_\mu = \{A \in \mathcal{A} \mu(A) = 0\} \quad \mathcal{N}_\nu = \{A \in \mathcal{A} \nu(A) = 0\} \quad (16.6)$
Definition 16.15 Measure Space $(\mathcal{F}, \Omega, \underline{\mu})$:	The triplet of sample space, sigma algebra and a measure is called a measure space.
2.1. Borel Measures	
Definition 16.16 Borel Measure:	A Borel Measure is any measure [def. 16.12] μ defined on the Borel σ -algebra [def. 16.9] $\mathcal{B}(\mathbb{R})$.
2.1.1. The Lebesgue Measure	
Definition 16.17 Lebesgue Measure on \mathcal{B} λ :	Is the Borel measure [def. 16.16] defined on the measurable space $\{\mathbb{R}, \mathcal{B}(\mathbb{R})\}$ which assigns for every half-open interval $(a, b]$ interval its length: $\lambda((a, b]) := b - a \quad (16.7)$

Corollary 16.13 Lebesgue Measure of Atomites:	
• The Lebesgue measure of a set containing only one point must be zero: $\lambda(\{a\}) = 0 \quad (16.8)$	
• The Lebesgue measure of a set containing countably many points $A = \{a_1, a_2, \dots, a_n\}$ must be zero: $\lambda(A) + \sum_{i=1}^n \lambda(\{a_i\}) = 0 \quad (16.9)$	
• The Lebesgue measure of a set containing uncountably many points $A = \{a_1, a_2, \dots\}$ can be either zero, positive and finite or infinite.	
3. Probability/Kolomogorov's Axioms 1931	
One problem we are still having is the range of μ , by standardizing the measure we obtain a well defined measure of events.	
Axiom 16.1 Non-negativity: The probability of an event is a non-negative real number: If $A \in \mathcal{F}$ then $\mathbb{P}(A) \geq 0 \quad (16.10)$	
Axiom 16.2 Unitaicity: The probability that at least one of the elementary events in the entire sample space Ω will occur is equal to one: The certain event $\mathbb{P}(\Omega) = 1 \quad (16.11)$	
Axiom 16.3 σ-additivity: If $A_1, A_2, A_3, \dots \in \mathcal{F}$ are mutually disjoint, then: $\mathbb{P}\left(\bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^\infty \mu(A_i) \quad (16.12)$	
Corollary 16.14 : As a consequence of this it follows: $\mathbb{P}(\emptyset) = 0 \quad (16.13)$	
Corollary 16.15 Complementary Probability: $\mathbb{P}(A^C) = 1 - \mathbb{P}(A) \quad \text{with} \quad A^C = \Omega - A \quad (16.14)$	
Definition 16.18 Probability Measure \mathbb{P} : a probability measure is function $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ defined on a σ -algebra \mathcal{F} of a sample space Ω that satisfies the probability axioms.	
4. Conditional Probability	
Definition 16.19 Conditional Probability: Let A, B be events, with $\mathbb{P}(B) \neq 0$. Then the conditional probability of the event A given B is defined as: $\mathbb{P}(A B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad \mathbb{P}(B) \neq 0 \quad (16.15)$	
5. Independent Events	
Theorem 16.1 Independent Events: Let A, B be two events. A and B are said to be independent iff: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \quad \mathbb{P}(A B) = \mathbb{P}(A), \quad \mathbb{P}(B) > 0$ $\mathbb{P}(B A) = \mathbb{P}(B), \quad \mathbb{P}(A) > 0 \quad (16.16)$	
Note The requirement of no impossible events follows from [def. 16.19]	
Corollary 16.16 Pairwise Independent Evenest: A finite set of events $\{A_i\}_{i=1}^n \in \mathcal{A}$ is <i>pairwise independent</i> if every pair of events is independent: $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \mathbb{P}(A_j) \quad i \neq j, \quad \forall i, j \in \mathcal{A} \quad (16.17)$	
Corollary 16.17 Mutal Independent Evenest: A finite set of events $\{A_i\}_{i=1}^n \in \mathcal{A}$ is <i>mutal independent</i> if every event A_j is independent of any intersection of the other events: $\mathbb{P}\left(\bigcap_{i=i}^k B_i\right) = \prod_{i=1}^k \mathbb{P}(B_i) \quad \forall \{B_i\}_{i=1}^k \subseteq \{A_i\}_{i=1}^n \quad k \leq n, \quad \{A_i\}_{i=1}^n \in \mathcal{A} \quad (16.18)$	

6. Product Rule

Law 16.1 Product Rule: Let A, B be two events then the probability of both events occurring simultaneously is given by:

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) \quad (16.19)$$

Law 16.2

Generalized Product Rule/Chain Rule: is the generalization of the product rule?? to n events $\{A_i\}_{i=1}^n$

$$\begin{aligned} \mathbb{P}\left(\bigcap_{i=1}^n E_i\right) &= \prod_{k=1}^n \mathbb{P}\left(E_k \mid \bigcap_{i=1}^{k-1} E_i\right) = \\ &= \mathbb{P}(E_n | E_{n-1} \cap \dots \cap E_1) \cdot \mathbb{P}(E_{n-1} | E_{n-2} \cap \dots \cap E_1) \cdots \\ &\quad \cdots \mathbb{P}(E_3 | E_2 \cap E_1) \mathbb{P}(E_2 | E_1) \mathbb{P}(E_1) \end{aligned} \quad (16.20)$$

7. Law of Total Probability

Definition 16.20 Complete Event Field: A complete event field $\{A_i : i \in I \subseteq \mathbb{N}\}$ is a countable or finite partition of Ω that is the partitions $\{A_i : i \in I \subseteq \mathbb{N}\}$ are a *disjoint union* of the sample space:

$$\bigcup_{i \in I} A_i = \Omega \quad A_i \cap A_j = \emptyset \quad i \neq j, \forall i, j \in I \quad (16.21)$$

Theorem 16.2

Law of Total Probability/Partition Equation: Let $\{A_i : i \in I\}$ be a complete event field^[def. 16.20] then it holds for $B \in \mathcal{B}$:

$$\mathbb{P}(B) = \sum_{i \in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i) \quad (16.22)$$

8. Bayes Theorem

Law 16.3 Bayes Rule: Let A, B be two events s.t. $\mathbb{P}(B) > 0$ then it holds:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \quad \mathbb{P}(B) > 0 \quad (16.23)$$

follows directly from eq. (16.19).

Theorem 16.3 Bayes Theorem: Let $\{A_i : i \in I\}$ be a complete event field^[def. 16.20] and $B \in \mathcal{B}$ a random event s.t. $\mathbb{P}(B) > 0$, then it holds:

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i \in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i)} \quad (16.24)$$

proof ?? 16.2

Distributions on \mathbb{R}

9.1. Distribution Function

Definition 16.21 Distribution Function of \mathbb{P} F : The *distribution function* F induced by a probability measure \mathbb{P} on $(\mathbb{R}, \mathcal{B})$ is the function:

$$F(x) = \mathbb{P}((-\infty, x]) \quad (16.25)$$

Theorem 16.4 : A function F is the distribution function of a (unique) probability on $(\mathbb{R}, \mathcal{B})$ iff:

- F is non-decreasing
- F is right continuous
- $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow +\infty} F(x) = 1$

Corollary 16.18 : A probability \mathbb{P} is uniquely determined by a distribution function F

That is if there exist another probability \mathbb{Q} s.t.

$$G(x) = \mathbb{Q}((-\infty, x])$$

and if $F = G$ then it follows $\mathbb{P} = \mathbb{Q}$.

9.2. Random Variables

A random variable X is a function/map that determines a quantity of interest based on the outcome $\omega \in \Omega$ of a random experiment. Thus X is not really a variable in the classical sense but a variable with respect to the outcome of an experiment. Its value is determined in two steps:

- ① The outcome of an experiment is a random quantity $\omega \in \Omega$
- ② The outcome ω determines (possibly various) quantities of interests \iff *random variables*

Thus a random variable X , defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a mapping from Ω into another space \mathcal{E} , usually $\mathcal{E} = \mathbb{R}$ or $\mathcal{E} = \mathbb{R}^n$:

$$X : \Omega \mapsto \mathcal{E} \quad \omega \mapsto X(\omega)$$

Let now $E \in \mathcal{E}$ be a quantity of interest, in order to quantify its probability we need to map it back to the original sample space Ω :

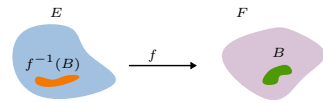
Probability for an event in Ω

$$\mathbb{P}_X(E) = \mathbb{P}(\{\omega : X(\omega) \in E\}) = \mathbb{P}(X \in E) = \mathbb{P}(X^{-1}(E))$$

Probability for an event in E

Definition 16.22 \mathcal{E} -measurable function: Let (E, \mathcal{E}) and (F, \mathcal{F}) be two measurable spaces. A function $f : E \mapsto F$ is called measurable (relative to \mathcal{E} and \mathcal{F}) if

$$\forall B \in \mathcal{F} : f^{-1}(B) = \{\omega \in \mathcal{E} : f(\omega) \in B\} \in \mathcal{E} \quad (16.26)$$



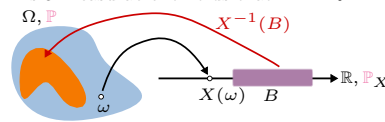
Interpretation

The pre-image^[def. 5.11] of B under f i.e. $f^{-1}(B)$ maps all values of the target space F back to the sample space \mathcal{E} (for all possible $B \in \mathcal{F}$).

Definition 16.23 Random Variable: A real-valued random variable (vector) X , defined on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ is an \mathcal{E} -measurable function mapping, if it maps its sample space Ω into a target space (F, \mathcal{F}) :

$$X : \Omega \mapsto \mathcal{F} \quad (\mathcal{F}^n) \quad (16.27)$$

Since X is \mathcal{E} -measurable it holds that $X^{-1} : \mathcal{F} \mapsto \mathcal{E}$



Corollary 16.19 : Usually $F = \mathbb{R}$, which usually amounts to using the Borel σ -algebra \mathcal{B} of \mathbb{R} .

Corollary 16.20 Random Variables of Borel Sets: Given that we work with Borel σ -algebras then the definition of a random variable is equivalent to (due to ^[cor. 16.11]):

$$\begin{aligned} X^{-1}(B) &= X^{-1}((-\infty, a]) \\ &= \{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{E} \quad \forall a \in \mathbb{R} \end{aligned} \quad (16.28)$$

Definition 16.24

Realization of a Random Variable $x = X(\omega)$: Is the value of a random variable that is actually observed after an experiment has been conducted. In order to avoid confusion lower case letters are used to indicate actual observations/realization of a random variable.

Corollary 16.21 Indicator Functions

An important class of measurable functions that can be used as r.v. are indicator functions:

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} \quad (16.29)$$

We know that a probability measure \mathbb{P} on \mathbb{R} is characterized by the quantities $\mathbb{P}((-\infty, a])$. Thus the quantities.

Corollary 16.22 : Let $(F, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$ and let (E, \mathcal{E}) be an arbitrary measurable space. Let X be a real value function on E .

Then it holds that X is measurable if and only if

$$\begin{aligned} \{X \leq a\} &= \{\omega : X(\omega) \leq a\} = X^{-1}((-\infty, a]) \in \mathcal{E}, \forall a \in \mathbb{R} \\ \text{or} \quad \{X < a\} &\in \mathcal{E}. \end{aligned}$$

Explanation 16.3 (^[cor. 16.22]). A random variable is a function that is measurable if and only if its distribution function is defined.

9.3. The Law of Random Variables

Definition 16.25 Law/Distribution of X $\mathcal{L}(X)$:

Let X be a r.v. on $(\Omega, \mathcal{F}, \mathbb{P})$, with values in (E, \mathcal{E}) , then the *distribution*/law of X is defined as:

$$\mathbb{P} : \mathcal{B} \mapsto [0, 1] \quad (16.30)$$

$$\mathbb{P}^X(B) = \mathbb{P}\{X \in B\} = \mathbb{P}(\omega : X(\omega) \in B) \quad \forall B \in \mathcal{E}$$

Note

- Sometimes \mathbb{P}^X is also called the *image* of \mathbb{P} by X
- The law can also be written as:

$$\mathbb{P}^X(B) = \mathbb{P}(X^{-1}(B)) = (\mathbb{P} \circ X^{-1})(B)$$

Theorem 16.5 : The law/distribution of X is a probability measure \mathbb{P} on (E, \mathcal{E}) .

Definition 16.26

(Cumulative) Distribution Function F_X :

Given a real-valued r.v. then its *cumulative distribution function* is defined as:

$$F_X(x) = \mathbb{P}^X((-\infty, x]) = \mathbb{P}(X \leq x) \quad (16.31)$$

Corollary 16.23 : The distribution of \mathbb{P}^X of a real valued r.v. is entirely characterized by its cumulative distribution function F_X ^[def. 16.33].

Property 16.1:

$$\mathbb{P}(X > x) = 1 - F_X(x) \quad (16.32)$$

Property 16.2:

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) \quad (16.33)$$

9.4. Probability Density Function

Definition 16.27 Continuous Random Variable: Is a r.v. for which a probability density function f_X exists.

Definition 16.28 Probability Density Function: Let X be a r.v. with associated cdf F_X . If F_X is continuously integrable for all $x \in \mathbb{R}$ then X has a *probability density* f_X defined by:

$$F_X(x) = \int_{-\infty}^x f_X(y) dy \quad (16.34)$$

or alternatively:

$$f_X(x) = \lim_{\epsilon \rightarrow 0} \frac{\mathbb{P}(x \leq X \leq x + \epsilon)}{\epsilon} \quad (16.35)$$

Corollary 16.24 $\mathbb{P}(X = b) = 0, \quad \forall b \in \mathbb{R}$:

$$\mathbb{P}(X = b) = \lim_{a \rightarrow b} \mathbb{P}(a < X \leq b) = \lim_{a \rightarrow b} \int_a^b f(x) dx = 0 \quad (16.36)$$

Corollary 16.25 : From ^[cor. 16.24] it follows that the exact borders are not necessary:

$$\begin{aligned} \mathbb{P}(a < X < b) &= \mathbb{P}(a \leq X < b) \\ &= \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b) \end{aligned}$$

Corollary 16.26 :

$$\int_{-\infty}^{\infty} f(x) dx = 1 \quad (16.37)$$

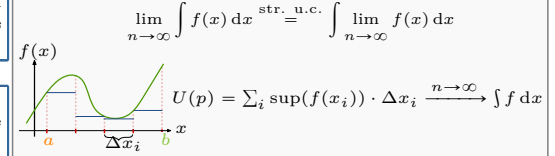
Notes

- Often the cumulative distribution function is referred to as “cdf” or simply *distribution function*.
- Often the probability density function is referred to as “pdf” or simply *density*.

9.5. Lebesgue Integration

Problems of Riemann Integration

- Difficult to extend to higher dimensions – general domains of definitions $f : \Omega \mapsto \mathbb{R}$
- Depends on continuity
- Integration of limit processes require strong uniform convergence in order to integrate limit processes

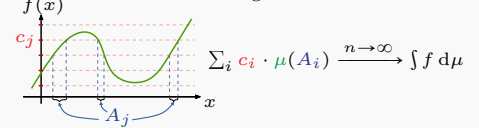


Idea

Partition domain by function values of equal size i.e. values that lie within the same sets/have the same value A_j build up the partitions w.r.t. to the variable x .

Problem: we do not know how big those sets/partitions on the x -axis will be.

Solution: we can use the measure μ of our measure space $(\Omega, \mathcal{A}, \mu)$ in order to obtain the size of our sets $A_j \Rightarrow$ we do not have to care anymore about discontinuities, as we can measure the size of our sets using our measure.



Definition 16.29 Lebesgue Integral:

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n c_i \mu(A_i) = \int_{\Omega} f d\mu \quad f(x) \approx c_i \quad \forall x \in A_i \quad (16.38)$$

Definition 16.30

Simple Functions (Random Variables): A r.v. X is called simple if it takes on only a finite number of values and hence can be written in the form:

$$X = \sum_{i=1}^n a_i \mathbb{1}_{A_i} \quad a_i \in \mathbb{R} \quad \mathcal{A} \ni A_i = \begin{cases} 1 & \text{if } \{X = a_i\} \\ 0 & \text{else} \end{cases} \quad (16.39)$$

9.6. Independent Random Variables

We have seen that two events A and B are independent if knowledge that B has occurred does not change the probability that A will occur theorem 16.1.

For two random variables X, Y we want to know if knowledge of Y leaves the probability of X , to take on certain values unchanged.

Definition 16.31 Independent Random Variables:

Two real valued random variables X and Y are said to be independent iff:

$$\mathbb{P}(X \leq x | Y \leq y) = \mathbb{P}(X \leq x) \quad \forall x, y \in \mathbb{R} \quad (16.40)$$

which amounts to:

$$\begin{aligned} F_{X,Y}(x, y) &= \mathbb{P}(\{X \leq x\} \cap \{Y \leq y\}) = \mathbb{P}(X \leq x, Y \leq y) \\ &= F_X(x) F_Y(y) \quad \forall x, y \in \mathbb{R} \end{aligned} \quad (16.41)$$

or alternatively iff:

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \mathbb{P}(Y \in B) \quad \forall A, B \in \mathcal{B} \quad (16.42)$$

Note
If the joint distribution $F_{X,Y}(x,y)$ can be factorized into two functions of x and y then X and Y are independent.

Definition 16.32
Independent Identically Distributed:

10. Product Rule

Law 16.4 Product Rule: Let X, Y be two random variables then their jo

Law 16.5
Generalized Product Rule/Chain Rule:

11. Change Of Variables Formula

Formula 16.1
(Scalar Discret) Change of Variables: Let X be a discret rv $X \in \mathcal{X}$ with pmf p_X and define $Y \in \mathcal{Y}$ as $Y = g(x)$ s.t. $\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\}$. **Where** g is an arbitrary strictly monotonic (def. 5.14) function.
Let: $\mathcal{X}_y = x_i$ be the set of all $x_i \in \mathcal{X}$ s.t. $y = g(x_i)$.
Then the pmf of Y is given by:
$$p_Y(y) = \sum_{x_i \in \mathcal{X}_y} p_X(x_i) = \sum_{x \in \mathcal{Y}: g(x)=y} p_X(x) \quad (16.43)$$

see proof ?? 16.3

Formula 16.2
(Scalar Continuous) Change of Variables:
Let $X \sim f_X$ be a continuous r.v. and let g be an arbitrary strictly monotonic (def. 5.14) function.
Define a new r.v. Y as
$$\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\} \quad (16.44)$$

then the pdf of Y is given by:
$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = f_X(x) \left| \frac{d}{dy} (g^{-1}(y)) \right| \quad (16.45)$$

$$= f_X(x) \frac{1}{\left| \frac{dy}{dx} \right|} = \frac{f_X(g^{-1}(y))}{\left| \frac{dg}{dx}(g^{-1}(y)) \right|} \quad (16.46)$$

Formula 16.3
(Continuous) Change of Variables:
Let $X = \{X_1, \dots, X_n\} \sim f_X$ be a continuous random vector and let g be an arbitrary strictly monotonic (def. 5.14) function $g: \mathbb{R}^n \mapsto \mathbb{R}^m$
Define a new r.v. Y as
$$\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\} \quad (16.47)$$

and let $h(x) := g(x)^{-1}$ then the pdf of Y is given by:
$$\begin{aligned} f_Y(y) &= f_X(x_1, \dots, x_n) \cdot |J| \\ &= f_X(h_1(y), \dots, h_n(y)) \cdot |J| \\ &= f_X(y) |\det D_x h(x)| \Big|_{x=y} \\ &= f_X(g^{-1}(y)) \left| \det \left(\frac{\partial g}{\partial x} \right) \right|^{-1} \end{aligned} \quad (16.48)$$

where $J = \det D_h$ is the Jacobian (def. 6.6).
See also proof ?? 16.6 and example 16.8

Note
A monotonic function is required in order to satisfy inevitability.

Probability Distributions on \mathbb{R}^n

13. Joint Distribution

Definition 16.33
Joint (Cumulative) Distribution Function $F_{\mathbf{X}}$:
Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random vector in \mathbb{R}^n , then its cumulative distribution function is defined as:
$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) &= \mathbb{P}^X((-\infty, \mathbf{x}]) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}) \\ &= \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) \end{aligned} \quad (16.49)$$

Definition 16.34 Joint Probability Distribution:
Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random vector in \mathbb{R}^n with associated cdf $F_{\mathbf{X}}$. If $F_{\mathbf{X}}$ is continuously integrable for all $\mathbf{x} \in \mathbb{R}$ then \mathbf{X} has a *probability density* f_X defined by:
$$F_X(x) = \int_{-\infty}^{x_n} \dots \int_{-\infty}^{x_1} f_{\mathbf{X}}(y_1, \dots, y_n) dy_1 \dots dy_n \quad (16.50)$$

or alternatively:
$$f_{\mathbf{X}}(\mathbf{x}) = \lim_{\epsilon \rightarrow 0} \frac{\mathbb{P}(x_1 \leq X_1 \leq x_1 + \epsilon, \dots, x_n \leq X_n \leq x_n + \epsilon)}{\epsilon} \quad (16.51)$$

13.1. Marginal Distribution

Definition 16.35 Marginal Distribution:

14. The Expectation

Definition 16.36 Expectation:
$$\mathbb{E}[X] = \int_{\Omega} X(\omega) \mathbb{P}(d\omega) = \int_{\Omega} X d\mathbb{P} \quad (16.52)$$

Corollary 16.27 Expectation of simple r.v.:
If X is a simple (def. 16.30) r.v. its expectation is given by:
$$\mathbb{E}[X] = \sum_{i=1}^n a_i \mathbb{P}(A_i) \quad (16.53)$$

14.1. Properties

14.1.1. Linear Operators

14.1.2. Quadratic Form
Definition 16.37 proof 16.7
Expectation of a Quadratic Form:
Let $\epsilon \in \mathbb{R}^n$ be a random vector with $\mathbb{E}[\epsilon] = \mu$ and $\mathbb{V}[\epsilon] = \Sigma$:
$$\mathbb{E}[\epsilon^T A \epsilon] = \text{tr}(A \Sigma) + \mu^T A \mu \quad (16.54)$$

14.2. The Jensen Inequality

Theorem 16.6 Jensen Inequality: Let X be a random variable and g some function, then it holds:
$$\begin{aligned} g(\mathbb{E}[X]) &\leq \mathbb{E}[g(X)] & \text{if } g \text{ is convex} & \text{[def. 5.24]} \\ g(\mathbb{E}[X]) &\geq \mathbb{E}[g(X)] & \text{if } g \text{ is concave} & \text{[def. 5.25]} \end{aligned} \quad (16.55)$$

14.3. Law of the Unconscious Statistician

Law 16.6 Law of the Unconscious Statistician:
Let $X \in \mathcal{X}, Y \in \mathcal{Y}$ be random variables where Y is defined as:
 $\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\}$
then the expectation of Y can be calculated in terms of X :
$$\mathbb{E}_Y[y] = \mathbb{E}_X[g(x)] \quad (16.56)$$

Consequence

Hence if we p_X we do not have to first calculate p_Y in order to calculate $\mathbb{E}_Y[y]$.

14.4. Properties

14.5. Law of Iterated Expectation (LIE)

Law 16.7 [proof 16.8]
Law of Iterated Expectation (LIE):
$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X|Y]] \quad (16.57)$$

14.6. Hoeffdings Bound

Definition 16.38 Hoeffdings Bound:
Let $\mathbf{X} = \{X_i\}_{i=1}^n$ be i.i.d. random variables strictly bounded by the interval $[a, b]$ then it holds:
$$\mathbb{P}(|\mu_{\mathbf{X}} - \mathbb{E}[X]| \geq \epsilon) \leq 2 \exp \left(\frac{-2n^2 \epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2} \right) \stackrel{[0,1]}{=} 2e^{-2n\epsilon^2} \quad (16.58)$$

Explanation 16.4. The difference of the expectation from the empirical average to be bigger than ϵ is upper bound in probability.

15. Moment Generating Function (MGF)

Definition 16.39 Moment of Random Variable: The i -th moment of a random variable X is defined as (if it exists):
$$m_i := \mathbb{E}[X^i] \quad (16.59)$$

Definition 16.40 ψ_X
Moment Generating Function (MGF):
$$\psi_X(t) = \mathbb{E}[e^{tX}] \quad t \in \mathbb{R} \quad (16.60)$$

Corollary 16.28 Sum of MGF: The moment generating function of a sum of n independent variables $(X_j)_{1 \leq j \leq n}$ is the product of the moment generating functions of the components:
$$\psi_{S_n}(t) = \psi_{X_1}(t) \dots \psi_{X_n}(t) \quad S_n := X_1 + \dots + X_n \quad (16.61)$$

Corollary 16.29 : The i -th moment of a random variable is the i -th derivative of its associated moment generating function evaluated zero:
$$\mathbb{E}[X^i] = \psi_X^{(i)}(0) \quad (16.62)$$

16. The Characteristic Function

Transforming probability distributions using the Fourier transform is a handy tool in probability in order to obtain properties or solve problems in another space before transforming them back.

Definition 16.41 $\hat{\mu}$
Fourier Transformed Probability Measure:
$$\hat{\mu} = \int e^{i\langle u, x \rangle} \mu(dx) \quad (16.63)$$

Corollary 16.30 : As $e^{i\langle u, x \rangle}$ can be rewritten using formulae. (1.9) and (1.10) it follows:
$$\hat{\mu} = \int \cos(\langle u, x \rangle) \mu(dx) + i \int \sin(\langle u, x \rangle) \mu(dx) \quad (16.64)$$

where $x \mapsto \cos(\langle x, u \rangle)$ and $x \mapsto \sin(\langle x, u \rangle)$ are both bounded and Borel i.e. Lebesgue integrable.

Definition 16.42 Characteristic Function φ_X : Let \mathbf{X} be an \mathbb{R}^n -valued random variable. Its characteristic function φ_X is defined on \mathbb{R}^n as:
$$\begin{aligned} \varphi_{\mathbf{X}}(u) &= \int e^{i\langle u, \mathbf{x} \rangle} \mathbb{P}^X(d\mathbf{x}) = \widehat{\mathbb{P}^X}(u) \\ &= \mathbb{E}[e^{i\langle u, \mathbf{x} \rangle}] \end{aligned} \quad (16.65) \quad (16.66)$$

Corollary 16.31 : The characteristic function φ_X of a distribution always exists as it is equal to the Fourier transform of the probability measure, which always exists.

Note

This is an advantage over the moment generating function.

Theorem 16.7 : Let μ be a probability measure on \mathbb{R}^n . Then $\hat{\mu}$ is a bounded continuous function with $\hat{\mu}(0) = 1$.
add proof

Theorem 16.8 Uniqueness Theorem: The Fourier Transform $\hat{\mu}$ of a probability measure μ on \mathbb{R}^n characterizes μ . That is, if two probability measures on \mathbb{R}^n admit the same Fourier transform, they are equal.
add proof

Corollary 16.32 : Let $\mathbf{X} = (X_1, \dots, X_n)$ be an \mathbb{R}^n -valued random variable. Then the real valued r.v.'s $(X_j)_{1 \leq j \leq n}$ are independent if and only if:

$$\varphi_X(u_1, \dots, u_n) = \prod_{j=1}^n \varphi_{X_j}(u_j) \quad (16.67)$$

Proofs

Proof 16.1: [cor. 16.11]: Let \mathcal{C} denote all open intervals. Since every open set in \mathbb{R} is the countable union of open intervals (def. 1.12), it holds that $\sigma(\mathcal{C})$ is the Borel σ -algebra of \mathbb{R} .
Let \mathcal{D} denote all intervals of the form $(-\infty, a]$, $a \in \mathbb{Q}$.
Let $a, b \in \mathcal{C}$, and let

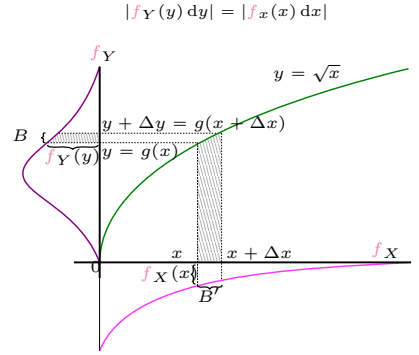
- $(a_n)_{n>1}$ be a sequence of rationals decreasing to a and
- $(b_n)_{n>1}$ be a sequence of rationals increasing strictly to b

 $(a, b) = \cup_{n=1}^{\infty} (a_n, b_n] = \cup_{n=1}^{\infty} ((-\infty, b_n] \cap (-\infty, a_n]^C)$
Thus $\mathcal{C} \subset \sigma(\mathcal{D})$, whence $\sigma(\mathcal{C}) \subset \sigma(\mathcal{D})$ but as each element of \mathcal{D} is a closed subset, $\sigma(\mathcal{D})$ must also be contained in the Borel sets \mathcal{B} with
$$\mathcal{B} = \sigma(\mathcal{C}) \subset \sigma(\mathcal{D}) \subset \mathcal{B}$$

Proof 16.2: theorem 16.3 Plug eq. (16.22) into the denominator and eq. (6.2) into the nominator and then use (def. 16.19):
$$\frac{\mathbb{P}(B|A_j) \mathbb{P}(A_j)}{\sum_{i \in I} \mathbb{P}(B|A_i) \mathbb{P}(A_i)} = \frac{\mathbb{P}(B \cap A_j)}{\mathbb{P}(B)} = \mathbb{P}(A_j|B)$$

Proof 16.3: ??:
$$Y = g(X) \iff \mathbb{P}(Y = y) = \mathbb{P}(x \in \mathcal{X}_y) = p_Y(y)$$

Proof 16.4: ?? (non-formal): The probability contained in a differential area must be invariant under a change of variables that is:



Proof 16.5: ?? from CDF:
$$\mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \begin{cases} \mathbb{P}(X \leq g^{-1}(y)) & \text{if } g \text{ is increas.} \\ \mathbb{P}(X \geq g^{-1}(y)) & \text{if } g \text{ is decreas.} \end{cases}$$

If g is monotonically increasing:

$$F_Y(y) = F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = f_X(x) \cdot \frac{d}{dy} g^{-1}(y)$$

If g is monotonically decreasing:

$$F_Y(y) = 1 - F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = -f_X(x) \cdot \frac{d}{dy} g^{-1}(y)$$

Proof 16.6: ??: Let $B = [x, x + \Delta x]$ and $B' = [y, y + \Delta y] = [g(x), g(x + \Delta x)]$ we know that the probability of equal events is equal:

$y = g(x) \Rightarrow \mathbb{P}(y) = \mathbb{P}(g(x))$ (for disc. rv.)

Now lets consider the probability for the continuous r.v.s:

$$\mathbb{P}(X \in B) = \int_x^{x+\Delta x} f_X(t) dt \xrightarrow{\Delta x \rightarrow 0} |\Delta x \cdot f_X(x)|$$

For y we use Taylor (??)

$$g(x + \Delta x) \stackrel{\text{eq. (5.56)}}{=} g(x) + \frac{dg}{dx} \Delta y \quad \text{for } \Delta x \rightarrow 0$$
$$= y + \Delta y \quad \text{with } \Delta y := \frac{dg}{dx} \cdot \Delta x \quad (16.68)$$

Thus for $\mathbb{P}(Y \in B')$ it follows:

$$\mathbb{P}(X \in B') = \int_y^{y+\Delta y} f_Y(t) dt \xrightarrow{\Delta y \rightarrow 0} |\Delta y \cdot f_Y(y)|$$
$$= \left| \frac{dg}{dx}(x) \Delta x \cdot f_Y(y) \right|$$

Now we simply need to related the surface of the two pdfs:

$B = [x, x + \Delta x]$ same surfaces \propto $[y, y + \Delta y] = B'$

$$\mathbb{P}(Y \in B) = \mathbb{P}(X \in B')$$
$$\stackrel{\Delta y \rightarrow 0}{\iff} |f_Y(y) \cdot \Delta y| = \left| f_Y(y) \cdot \frac{dg}{dx}(x) \Delta x \right| = |f_X(x) \cdot \Delta x|$$
$$f_Y(y) \cdot \left| \frac{dg}{dx}(x) \right| |\Delta x| = f_X(x) \cdot |\Delta x|$$
$$\Rightarrow f_Y(y) = \frac{f_X(x)}{\left| \frac{dg}{dx}(x) \right|} = \frac{f_X(g^{-1}(y))}{\left| \frac{dg}{dx} g^{-1}(y) \right|}$$

Proof 16.7: [def. 16.37]

$$\mathbb{E}[\epsilon^T A \epsilon] \stackrel{\text{eq. (10.54)}}{=} \mathbb{E}[\text{tr}(\epsilon^T A \epsilon)]$$
$$\stackrel{\text{eq. (10.56)}}{=} \mathbb{E}[\text{tr}(A \epsilon \epsilon^T)]$$
$$= \text{tr}(\mathbb{E}[A \epsilon \epsilon^T])$$
$$= \text{tr}(A \mathbb{E}[\epsilon \epsilon^T])$$
$$= \text{tr}(A (\Sigma + \mu \mu^T))$$
$$= \text{tr}(A \Sigma) + \text{tr}(A \mu \mu^T)$$
$$\stackrel{\text{eq. (10.54)}}{=} \text{tr}(A \Sigma) + A \mu \mu^T$$

Proof 16.8: law 16.7

$$\mathbb{E}[X] = \sum_x x \cdot \mathbb{P}_X(x) = \sum_x x \cdot \sum_y \mathbb{P}_{X,Y}(x, y)$$
$$= \sum_x x \cdot \sum_y \mathbb{P}_{X|Y}(x|y) \cdot \mathbb{P}_Y(y)$$
$$= \sum_y \mathbb{P}_Y(y) \cdot \sum_x x \cdot \mathbb{P}_{X|Y}(x|y)$$
$$= \sum_y \mathbb{P}_Y(y) \cdot \mathbb{E}[X|Y] = \mathbb{E}_Y[\mathbb{E}[X|Y]]$$

Examples

- Example 16.1 :**
- Toss of a coin (with head and tail): $\Omega = \{H, T\}$.
 - Two tosses of a coin: $\Omega = \{HH, HT, TH, TT\}$
 - A cubic die: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$
 - The positive integers: $\Omega = \{1, 2, 3, \dots\}$
 - The reals: $\Omega = \{\omega | \omega \in \mathbb{R}\}$
- Example 16.2 :**
- Head in coin toss $A = \{H\}$
 - Odd number in die roll: $A = \{\omega_1, \omega_3, \omega_5, \}$
 - The integers smaller five: $A = \{1, 2, 3, 4\}$
- Example 16.3 :** If the sample space is a die toss $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$, the sample space may be that we are only told whether an even or odd number has been rolled:
- $$\mathcal{F} = \{\emptyset, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}\}$$

Example 16.4 : If we are only interested in the subset $A \in \Omega$ of our experiment, then we can look at the corresponding generating σ -algebra $\sigma(A) = \{\emptyset, A, A^C, \Omega\}$.

- Example 16.5 :**
- open half-lines: $(-\infty, a)$ and (a, ∞) ,
 - union of open half-lines: $(a, b) = (-\infty, a) \cup (b, \infty)$,
 - closed interval: $[a, b] = \overline{(-\infty, a) \cup (b, \infty)}$,
 - closed half-lines: $(-\infty, a] = \bigcup_{n=1}^{\infty} [a - n, a]$ and $[a, \infty) = \bigcup_{n=1}^{\infty} [a, a + n]$,
 - half-open and half-closed $(a, b] = (-\infty, b] \cap (a, \infty)$,
 - every set containing only one real number: $\{a\} = \bigcap_{n=1}^{\infty} (a - \frac{1}{n}, a + \frac{1}{n})$,
 - every set containing finitely many real numbers: $\{a_1, \dots, a_n\} = \bigcup_{k=1}^n \{a_k\}$.

Example 16.6 Equivalent (Probability) Measures:

$$\Omega = \{1, 2, 3\} \quad \mathbb{P}(\{1, 2, 3\}) = \{2/3, 1/6, 1/6\}$$
$$\quad \quad \quad \tilde{\mathbb{P}}(\{1, 2, 3\}) = \{1/3, 1/3, 1/3\}$$

Example 16.7 :

add example fat book p.1280

add example prob th book 4

Example 16.8 ??: Let $X, Y \stackrel{\text{ind.}}{\sim} \mathcal{N}(0, 1)$.

Question: proof that:

$$U = X + Y \quad V = X - 1$$

are indepdent and normally distributed:

$$h(u, v) = \begin{cases} h_1(u, v) = \frac{u+v}{2} \\ h_2(u, v) = \frac{u-v}{2} \end{cases} \quad J = \det \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} = -\frac{1}{2}$$
$$f_{U,V} = f_{X,Y}(\underline{x}, y) \cdot \frac{1}{2}$$
$$\stackrel{\text{indp.}}{=} f_X(\underline{x}) \cdot f_Y(y)$$
$$= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$
$$= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\left\{ \left(\frac{u+v}{2} \right)^2 + \left(\frac{u-v}{2} \right)^2 / 2 \right\}}$$
$$= \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{u^2}{4}} \cdot \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{v^2}{4}}$$

Thus U, V are independent r.v. distributed as $\mathcal{N}(0, 2)$.

Statistics

Delete/Move the following stuff appropriately

The probability that a discreet random variable x is equal to some value $\bar{x} \in \mathcal{X}$ is:

$$\mathbb{P}_x(\bar{x}) = \mathbb{P}(x = \bar{x})$$

addapt

Definition 17.1 Almost Surely \mathbb{P} -(a.s.):

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An event $\omega \in \mathcal{F}$ happens almost surely iff

$$\mathbb{P}(\omega) = 1 \quad \iff \quad \omega \text{ happens a.s.} \quad (17.1)$$

Definition 17.2 Probability Mass Function (PMF):

Definition 17.3 Discrete Random Variable (DVR): The set of possible values \bar{x} of \mathcal{X} is countable of finite.

$$\mathcal{X} = \{0, 1, 2, 3, 4, \dots, 8\} \quad \mathcal{X} = \mathbb{N} \quad (17.2)$$

Definition 17.4 Probability Density Function (PDF): Is real function $f : \mathbb{R}^n \rightarrow [0, \infty)$ that satisfies:

Non-negativity: $f(x) \geq 0, \quad \forall x \in \mathbb{R}^n \quad (17.3)$

Normalization: $\int_{-\infty}^{\infty} f(x) dx \stackrel{!}{=} 1 \quad (17.4)$

Must be integrable (17.5)

Note: why do we need probability density functions

A continuous random variable X can realise an infinite count of real number values within its support B (as there are an infinitude of points in a line segment).

Thus we have an infinitude of values whose sum of probabilities must equal one.

Thus these probabilities must each be zero otherwise we would obtain a probability of ∞ . As we can not work with zero probabilities we use the next best thing, infinitesimal probabilities (defined as a limit).

We say they are almost surely equal to zero:

$$\mathbb{P}(X = x) = 0 \quad \text{a.s.}$$

To have a sensible measure of the magnitude of these infinitesimal quantities, we use the concept of probability density, which yields a probability mass when integrated over an interval.

Definition 17.5 Continuous Random Variable (CRV): A real random variable (rv) X is said to be (absolutely) continuous if there exists a pdf (^{def. 17.4}) f_X s.t. for any subset $B \subset \mathbb{R}$ it holds:

$$\mathbb{P}(X \in B) = \int_B f_X(x) dx \quad (17.6)$$

Property 17.1 Zero Probability: If X is a continuous rv (^{def. 17.5}), then:

$$\mathbb{P}(X = a) = 0 \quad \forall a \in \mathbb{R} \quad (17.7)$$

Property 17.2 Open vs. Closed Intervals: For any real numbers a and b , with $a < b$ it holds:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a < X < b) \quad (17.8)$$

\iff including or not the bounds of an interval does not modify the probability of a continuous rv.

Note

Changing the value of a function at finitely many points has no effect on the value of a definite integral.

Corollary 17.1 : In particular for any real numbers a and b with $a < b$, letting $B = [a, b]$ we obtain:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx$$

Proof 17.1: Property 17.1:

$$\mathbb{P}(X = a) = \lim_{\Delta x \rightarrow 0} \mathbb{P}(X \in [a, a + \Delta x])$$
$$= \lim_{\Delta x \rightarrow 0} \int_a^{a+\Delta x} f_X(x) dx = 0$$

Proof 17.2: Property 17.2:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a < X < b) = \int_a^b f_X(x) dx$$

Definition 17.6 Support of a probability density function: The support of the density of a pdf $f_X(\cdot)$ is the set of values of the random variable X s.t. its pdf is non-zero:

$$\text{supp}(\cdot) f_X := \{x \in \mathcal{X} | f_X(x) > 0\} \quad (17.9)$$

Note: this is not a rigorous definition.

Theorem 17.1 RVs are defined by a PDFs: A probability density function f_X completely determines the distribution of a continuous real-valued random variable X .

Corollary 17.2 Identically Distributed: From theorem 17.1 it follows that to RV X and Y that have exactly the same pdf follow the same distribution. We say X and Y are **identically distributed**.

0.1. Cumulative Distribution Fuction

Definition 17.7 Cumulative distribution function (CDF): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. The (cumulative) distribution function of a real-valued random variable X is the function given by:

$$F_X(x) = \mathbb{P}(X \leq x) \quad \forall x \in \mathbb{R}$$

Property 17.3: Monotonically Increasing $x \leq y \iff F_X(x) \leq F_X(y) \quad \forall x, y \in \mathbb{R} \quad (17.10)$

Upper Limit $\lim_{x \rightarrow \infty} F_X(x) = 1 \quad (17.11)$

Lower Limit $\lim_{x \rightarrow -\infty} F_X(x) = 0 \quad (17.12)$

Definition 17.8 CDF of a discreet rv X: Let X be discreet rv with pdf \mathbb{P}_X , then the CDF of X is given by:

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{t=-\infty}^x \mathbb{P}_X(t)$$

Definition 17.9 CDF of a continuous rv X: Let X be continuous rv with pdf f_X , then the CDF of X is given by:

$$F_X(x) = \int_{-\infty}^x f_X(t) dt \quad \iff \quad \frac{\partial F_X(x)}{\partial x} = f_X(x)$$

Lemma 17.1 Probability Interval: Let X be a continuous rv with pdf f_X and cumulative distribution function F_X , then it holds that:

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) \quad (17.13)$$

Proof 17.3: [^{def. 17.9}]:

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \in (-\infty, x)) = \int_{-\infty}^x f_X(t) dt$$

Proof 17.4: lemma 17.1:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a)$$

or by the fundamental theorem of calculus (theorem 5.2):

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(t) dt = \int_a^b \frac{\partial F_X(t)}{\partial t} dt = [F_X(t)]_a^b$$

Theorem 17.2 A continuous rv is fully characterized by its CDF: A cumulative distribution function completely determines the distribution of a continuous real-valued random variable.

1. Key figures

1.1. The Expectation

Definition 17.10 Expectation (disc. case):

$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{x} \in \mathcal{X}} \bar{x} \mathbb{P}_x(\bar{x}) \quad (17.14)$$

Definition 17.11 Expectation (cont. case):

$$\mathbb{E}_x[x] := \int_{\bar{x} \in \mathcal{X}} \bar{x} f_x(\bar{x}) d\bar{x} \quad (17.15)$$

Law 17.1 Expectation of independent variables:

$$\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y] \quad (17.16)$$

Property 17.4 Translation and scaling: If $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^n$ are random vectors, and $a, b, a \in \mathbb{R}^n$ are constants then it holds:

$$\mathbb{E}[a + b\mathbf{X} + c\mathbf{Y}] = a + b\mathbb{E}[\mathbf{X}] + c\mathbb{E}[\mathbf{Y}] \quad (17.17)$$

Thus \mathbb{E} is a **linear** operator (^[def. 5.15]).

Note: Expectation of the expectation

The expectation of a r.v. X is a constant hence with Property 17.6 it follows:

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \quad (17.18)$$

Property 17.5 Matrix×Expectation: If $\mathbf{X} \in \mathbb{R}^n$ is a random vector and $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times m}$ are constant matrices then it holds:

$$\mathbb{E}[\mathbf{A}\mathbf{X}\mathbf{B}] = \mathbf{A}\mathbb{E}[(\mathbf{X}\mathbf{B})] = \mathbf{A}\mathbb{E}[\mathbf{X}] \mathbf{B} \quad (17.19)$$

Proof 17.5: eq. (17.24):

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{p}_{X,Y}(x, y)xy \\ &\stackrel{??}{=} \sum_{x \in \mathcal{X}} \mathbf{p}_X(x)x \sum_{y \in \mathcal{Y}} \mathbf{p}_Y(y)y = \mathbb{E}[X] \mathbb{E}[Y]\end{aligned}$$

Definition 17.12 Autocorrelation/Crosscorrelation $\gamma(t_1, t_2)$: Describes the covariance (def. 17.16) between the two values of a stochastic process $(\mathbf{X}_t)_{t \in T}$ at different time points t_1 and t_2 .
 $\gamma(t_1, t_2) = \text{Cov}[\mathbf{X}_{t_1}, \mathbf{X}_{t_2}] = \mathbb{E}[(\mathbf{X}_{t_1} - \mu_{t_1})(\mathbf{X}_{t_2} - \mu_{t_2})]$ (17.20)

For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance:

$$\gamma(t, t) = \text{Cov}[\mathbf{X}_t, \mathbf{X}_t] \stackrel{\text{eq. (17.35)}}{=} \mathbb{V}[\mathbf{X}_t] \quad (17.21)$$

Notes

- Hence the autocorrelation describes the correlation of a function or signal with itself at a previous time point.
- Given a random time dependent variable $\mathbf{x}(t)$ the autocorrelation function $\gamma(t, t - \tau)$ describes how similar the time translated function $\mathbf{x}(t - \tau)$ and the original function $\mathbf{x}(t)$ are.
- If there exists some relation between the values of the time series that is non-random then the autocorrelation is non-zero.
- The autocorrelation is maximized/most similar for no translation $\tau = 0$ at all.

2. Key Figures

2.1. The Expectation

more to prob theory maybe

Definition 17.13 Expectation (disc. case):

$$\mu_X := \mathbb{E}_x[X] := \sum_{\mathbf{x} \in \mathcal{X}} \mathbf{x} \mathbf{p}_x(\mathbf{x}) \quad (17.22)$$

Definition 17.14 Expectation (cont. case):

$$\mathbb{E}_x[x] := \int_{\mathbf{x} \in \mathcal{X}} \mathbf{x} f_x(\mathbf{x}) d\mathbf{x} \quad (17.23)$$

Law 17.2 Expectation of independent variables:

$$\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y] \quad (17.24)$$

Property 17.6 Translation and scaling: If $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^n$ are random vectors, and $a, b, c \in \mathbb{R}^n$ are constants then it holds:

$$\mathbb{E}[a + b\mathbf{X} + c\mathbf{Y}] = a + b\mathbb{E}[\mathbf{X}] + c\mathbb{E}[\mathbf{Y}] \quad (17.25)$$

Thus \mathbb{E} is a linear operator^[def. 5.15].

Property 17.7

Affine Transformation of the Expectation:

If $\mathbf{X} \in \mathbb{R}^n$ is a random vector, $\mathbf{A} \in \mathbb{R}^{m \times n}$ a constant matrix and $b \in \mathbb{R}^m$ then it holds:

$$\mathbb{E}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mu + b \quad (17.26)$$

Note: Expectation of the expectation

The expectation of a r.v. X is a constant hence with Property 17.6 it follows:

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \quad (17.27)$$

Property 17.8 Matrix×Expectation: If $\mathbf{X} \in \mathbb{R}^n$ is a random vector and $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$ are constant matrices then it holds:

$$\mathbb{E}[\mathbf{A}\mathbf{X}\mathbf{B}] = \mathbf{A}\mathbb{E}[\mathbf{X}\mathbf{B}] = \mathbf{A}\mathbb{E}[\mathbf{X}]\mathbf{B} \quad (17.28)$$

Proof 17.6: eq. (17.24):

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{p}_{X,Y}(x, y)xy \\ &\stackrel{??}{=} \sum_{x \in \mathcal{X}} \mathbf{p}_X(x)x \sum_{y \in \mathcal{Y}} \mathbf{p}_Y(y)y = \mathbb{E}[X] \mathbb{E}[Y]\end{aligned}$$

2.2. The Variance

Definition 17.15 Variance $\mathbb{V}[X]$: The variance of a random variable X is the expected value of the squared deviation from the expectation of X ($\mu = \mathbb{E}[X]$). It is a measure of how much the actual values of a random variable X fluctuate around its expected value $\mathbb{E}[X]$ and is defined by:

$$\mathbb{V}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2] \stackrel{\text{see ?? 17.7}}{=} \mathbb{E}[X^2] - \mathbb{E}[X]^2 \quad (17.29)$$

2.2.1. Properties

Property 17.9 Variance of a Constant: If $a \in \mathbb{R}$ is a constant then it follows that its expected value is deterministic \Rightarrow we have no uncertainty \Rightarrow no variance:

$$\mathbb{V}[a] = 0 \quad \text{with} \quad a \in \mathbb{R} \quad (17.30)$$

see shift and scaling for proof ?? 17.8

Property 17.10 Shifting and Scaling:

$$\mathbb{V}[a + bX] = a^2 \sigma^2 \quad \text{with} \quad a \in \mathbb{R} \quad (17.31)$$

see ?? 17.8

Property 17.11

[proof 17.9]

Affine Transformation of the Variance:

If $\mathbf{X} \in \mathbb{R}^n$ is a random vector, $\mathbf{A} \in \mathbb{R}^{m \times n}$ a constant matrix and $b \in \mathbb{R}^m$ then it holds:

$$\mathbb{V}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mathbb{V}[\mathbf{X}]\mathbf{A}^T \quad (17.32)$$

Definition 17.16 Covariance: The Covariance is a measure of how much two or more random variables vary linearly with each other.

$$\begin{aligned}\text{Cov}[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned} \quad (17.33)$$

see ?? 17.10

Definition 17.17 Covariance Matrix: The variance of a k -dimensional random vector $\mathbf{X} = (X_1 \dots X_k)$ is given by a p.s.d. eq. (10.109) matrix called Covariance Matrix. The Covariance is a measure of how much two or more random variables vary linearly with each other and the Variance on the diagonal is again a measure of how much a variable varies:

$$\begin{aligned}\mathbb{V}[\mathbf{X}] &:= \Sigma(\mathbf{X}) := \text{Cov}[\mathbf{X}, \mathbf{X}] := \\ &= \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] \\ &= \mathbb{E}[\mathbf{X}\mathbf{X}^T] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^T \in [-\infty, \infty]\end{aligned} \quad (17.34)$$

$$\begin{aligned}&= \begin{bmatrix} \mathbb{V}[X_1] & \dots & \text{Cov}[X_1, X_k] \\ \vdots & \ddots & \vdots \\ \text{Cov}[X_k, X_1] & \dots & \mathbb{V}[X_k] \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \dots & \mathbb{E}[(X_1 - \mu_1)(X_k - \mu_k)] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[(X_k - \mu_k)(X_1 - \mu_1)] & \dots & \mathbb{E}[(X_k - \mu_k)(X_k - \mu_k)] \end{bmatrix}\end{aligned}$$

Note: Covariance and Variance

The variance is a special case of the covariance in which two variables are identical:

$$\text{Cov}[X, X] = \mathbb{V}[X] \equiv \sigma^2(X) \equiv \sigma_X^2 \quad (17.35)$$

add <http://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/>

Property 17.12 Translation and Scaling:

$$\text{Cov}(a + bX, c + dY) = bd\text{Cov}(X, Y) \quad (17.36)$$

Property 17.13

Affine Transformation of the Covariance:

If $\mathbf{X} \in \mathbb{R}^n$ is a random vector, $\mathbf{A} \in \mathbb{R}^{m \times n}$ a constant matrix and $b \in \mathbb{R}^m$ then it holds:

$$\text{Cov}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mathbb{V}[\mathbf{X}]\mathbf{A}^T = \mathbf{A}\Sigma(\mathbf{X})\mathbf{A}^T \quad (17.37)$$

Definition 17.18 Correlation Coefficient: Is the standardized version of the covariance:

$$\begin{aligned}\text{Corr}[\mathbf{X}] &:= \frac{\text{Cov}[\mathbf{X}]}{\sigma_{X_1} \dots \sigma_{X_k}} \in [-1, 1] \\ &= \begin{cases} +1 & \text{if } Y = aX + b \text{ with } a > 0, b \in \mathbb{R} \\ -1 & \text{if } Y = aX + b \text{ with } a < 0, b \in \mathbb{R} \end{cases}\end{aligned} \quad (17.38)$$

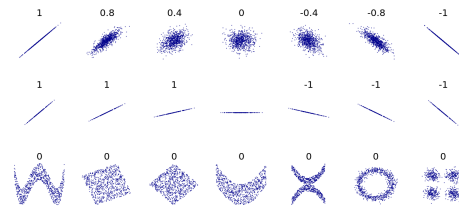


Figure 3: Several sets of (x, y) points, with their correlation coefficient

Law 17.3 Translation and Scaling:

$$\text{Corr}(a + bX, c + dY) = \text{sign}(b)\text{sign}(d)\text{Cov}(X, Y) \quad (17.39)$$

Note

- The correlation/covariance reflects the noisiness and direction of a linear relationship (top row fig. 3), but not the slope of that relationship (middle row fig. 3) nor many aspects of nonlinear relationships (bottom row)
- The set in the center of fig. 3 has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.
- Zero covariance/correlation $\text{Cov}(X, Y) = \text{Corr}(X, Y) = 0$ implies that there does not exist a linear relationship between the random variables X and Y .

Difference Covariance&Correlation

1. Variance is affected by scaling and covariance not ?? and law 17.3.
2. Correlation is dimensionless, whereas the unit of the covariance is obtained by the product of the units of the two RV variables.

Law 17.4 Covariance of independent RVs: The covariance/correlation of two independent variable's (??) is zero:

$$\begin{aligned}\text{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &\stackrel{\text{eq. (17.24)}}{=} \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0\end{aligned}$$

Zero covariance/correlation \Rightarrow independence

$\text{Cov}(X, Y) = \text{Corr}(X, Y) = 0 \Rightarrow \mathbf{p}_{X,Y}(x, y) = \mathbf{p}_X(x)\mathbf{p}_Y(y)$
For example: let $X \sim \mathcal{U}([-1, 1])$ and let $Y = X^2$.

1. Clearly X and Y are dependent
2. But the covariance/correlation between X and Y is non-zero:

$$\begin{aligned}\text{Cov}(X, Y) &= \text{Cov}(X, X^2) = \mathbb{E}[X \cdot X^2] - \mathbb{E}[X]\mathbb{E}[X^2] \\ &= \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] \stackrel{??}{=} 0 - 0 \cdot \mathbb{E}[X^2] \\ &\stackrel{??}{=} 0\end{aligned}$$
 \Rightarrow the relationship between Y and X must be non-linear.

Definition 17.19 Quantile: Are specific values q_α in the range^[def. 5.10] of a random variable X that are defined as the value for which the cumulative probability is less then q_α with probability $\alpha \in (0, 1)$:

$$q_\alpha : \mathbb{P}(X \leq x) = F_X(q_\alpha) = \alpha \xrightarrow{F \text{ invert.}} q_\alpha = F_X^{-1}(\alpha) \quad (17.40)$$

add figure

3. Proofs

Proof 17.7: eq. (17.29)

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] \\ &\stackrel{\text{Property 17.6}}{=} \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mu^2\end{aligned}$$

Proof 17.8: Property 17.10

$$\begin{aligned}\mathbb{V}[a + bX] &= \mathbb{E}[(a + bX - \mathbb{E}[a + bX])^2] \\ &= \mathbb{E}[(a + bX - a - b\mathbb{E}[X])^2] \\ &= \mathbb{E}[(bX - b\mathbb{E}[X])^2] \\ &= \mathbb{E}[b^2(X - \mathbb{E}[X])^2] \\ &= b^2\mathbb{E}[(X - \mathbb{E}[X])^2] = b^2\sigma^2\end{aligned}$$

Proof 17.9: Property 17.11

$$\begin{aligned}\mathbb{V}(\mathbf{A}\mathbf{X} + b) &= \mathbb{E}[(\mathbf{A}\mathbf{X} - \mathbb{E}[\mathbf{A}\mathbf{X}])^2] + 0 = \\ &= \mathbb{E}[(\mathbf{A}\mathbf{X} - \mathbb{E}[\mathbf{A}\mathbf{X}])(\mathbf{A}\mathbf{X} - \mathbb{E}[\mathbf{A}\mathbf{X}])^T] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T \mathbf{A}^T] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T \mathbf{A}^T] \\ &= \mathbf{A}\mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] \mathbf{A}^T = \mathbf{A}\mathbb{V}[\mathbf{X}]\mathbf{A}^T\end{aligned}$$

Proof 17.10: eq. (17.33)

$$\begin{aligned}\text{Cov}[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

Discrete Distributions

Definition 17.20 Multivariate Distribution: the variate refers to the number of input variables i.e. a m-variate distribution has m-input variables whereas a uni-variate distribution has only one.

Dimensional vs. Multivariate

The dimension refers to the number of dimensions we need to embed the function. If the variables of a function are independent than the dimension is the same as the number of inputs but the number of input variables can also be less.

4.1. Bernoulli Distribution

Bern(p)

Definition 17.21 Bernoulli Trial: Is a random experiment with exactly two possible outcomes, success (1) and failure (0), in which the probability of success/failure is constant in every trial i.e. independent trials.

Definition 17.22 Bernoulli Distribution $X \sim \text{Bern}(\mathbf{p})$: X is a binary variable i.e. can only attain the values 0 (failure) or 1 (success) with a parameter \mathbf{p} that signifies the success probability:

$$\mathbf{p}(x; \mathbf{p}) = \begin{cases} \mathbf{p} & \text{for } x = 1 \\ 1 - \mathbf{p} & \text{for } x = 0 \end{cases} \iff \begin{cases} \mathbb{P}(X = 1) = \mathbf{p} \\ \mathbb{P}(X = 0) = 1 - \mathbf{p} \end{cases} \\ = \mathbf{p}^x \cdot (1 - \mathbf{p})^{1-x} \quad \text{for } x \in \{0, 1\}$$

$$\mathbb{E}[X] = \mathbf{p} \quad (17.41) \quad \mathbb{V}[X] = \mathbf{p}(1 - \mathbf{p}) \quad (17.42)$$

4.2. Multinoulli/Categorical Distribution

Cat(\mathbf{n}, \mathbf{p})

Definition 17.23 Multinoulli/Categorical Distribution $X \sim \text{Cat}(\mathbf{p})$: Is the generalization of the Bernoulli distribution?? to a sample space^[def. 16.2] of k individual items $\{c_1, \dots, c_k\}$ with probabilities $\mathbf{p} = \{p_1, \dots, p_k\}$:

$$p(x = c_i | \mathbf{p}) = p_i \iff p(x | \mathbf{p}) = \prod_i p_i^{\delta[x=c_i]} \\ \sum_{j=1}^k p_j = 1 \quad p_j \in [0, 1] \quad \forall j = 1, \dots, k \quad (17.43) \\ \mathbb{E}[X] = \mathbf{p} \quad \mathbb{V}[X]_{i,j} = \Sigma_{i,j} = \begin{cases} p_i(1 - p_i) & \text{if } i = j \\ -p_i p_j & \text{if } i \neq j \end{cases}$$

Corollary 17.3 One-hot encoded Categorical Distribution: If we encode the k categories by a sparse vectors^[def. 10.70] with norm one:

$$\mathbb{B}_r^n = \left\{ \mathbf{x} \in \{0, 1\}^n : \mathbf{x}^\top \mathbf{x} = \sum_{i=1}^n x_i = 1 \right\} \\ \text{s.t.} \quad \mathbf{x}_j = \mathbf{e}_j \iff \mathbf{x} = \mathbf{c}_j \\ \text{then we can rewrite ?? as:} \\ p(\mathbf{x} | \mathbf{p}) = \prod_i \mathbf{x}_i \cdot p_i \quad \sum_{j=1}^k p_j = 1 \quad (17.44)$$

4.3. Binomial Distribution

B(\mathbf{n}, \mathbf{p})

Definition 17.24 Binomial Coefficient: The binomial coefficient occurs inside the binomial distribution?? and signifies the different combinations/order that x out of n successes can happen.

Definition 17.25 Binomial Distribution [proof ??]: Models the probability of exactly X success given a fixed number n -Bernoulli experiments??, where the probability of success of a single experiment is given by \mathbf{p} :

$$p(x) = \binom{n}{x} \mathbf{p}^x (1 - \mathbf{p})^{n-x} \quad \begin{array}{l} n : \text{nb. of repetitions} \\ x : \text{nb. of successes} \\ \mathbf{p} : \text{probability of success} \end{array} \\ \mathbb{E}[X] = n\mathbf{p} \quad (17.45) \quad \mathbb{V}[X] = n\mathbf{p}(1 - \mathbf{p}) \quad (17.46)$$

Note: Binomial Coefficient

The Binomial Coefficient corresponds to the permutation of two classes and not the variations as it seems from the formula.

Lets consider a box of n balls consisting of black and white balls. If we want to know the probability of drawing first x white and then $n - x$ black balls we can simply calculate:

$$\underbrace{(\mathbf{p} \cdots \mathbf{p})}_{x\text{-times}} \cdot \underbrace{(q \cdots q)}_{n-x\text{-times}} = \mathbf{p}^x \mathbf{q}^{n-x}$$

4.4. Geometric Distribution

Geom(p)

Definition 17.26 Geometric Distribution Geom(p): Models the probability of the number X of Bernoulli trials?? until the first success

$$p(x) = \mathbf{p}(1 - \mathbf{p})^{x-1} \quad \begin{array}{l} x : \text{nb. of repetitions until first success} \\ \mathbf{p} : \text{success probability of single Bernoulli experiment} \end{array}$$

$$F(x) = \sum_{i=1}^x \mathbf{p}(1 - \mathbf{p})^{i-1} \stackrel{\text{eq. (2.4)}}{=} 1 - (1 - \mathbf{p})^x \\ \mathbb{E}[X] = \frac{1}{\mathbf{p}} \quad (17.47) \quad \mathbb{V}[X] = \frac{1 - \mathbf{p}}{\mathbf{p}^2} \quad (17.48)$$

Notes

- $\mathbb{E}[X]$ is the mean waiting time until the first success
- the number of trials x in order to have at least one success with a probability of $\mathbf{p}(x)$:

$$x \geq \frac{\mathbf{p}(x)}{1 - \mathbf{p}}$$

- $\log(1 - \mathbf{p}) \approx -\mathbf{p}$ for small \mathbf{p}

4.5. Poisson Distribution

Pois(λ)

Definition 17.27 Poisson Distribution: Is an extension of the binomial distribution, where the realization x of the random variable X may attain values in $\mathbb{Z}_{\geq 0}$. It expresses the probability of a given number of events X occurring in a fixed interval if those events occur independently of the time since the last event.

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!} \quad \begin{array}{l} \lambda > 0 \\ x \in \mathbb{Z}_{\geq 0} \end{array} \quad (17.49)$$

Event Rate λ : describes the average number of events in a single interval.

$$\mathbb{E}[X] = \lambda \quad (17.50) \quad \mathbb{V}[X] = \lambda \quad (17.51)$$

Continuous Distributions

5.1. Uniform Distribution

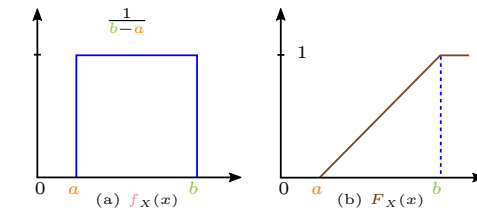
$\mathcal{U}(a, b)$

Definition 17.28 Uniform Distribution $\mathcal{U}(a, b)$: Is probability distribution, where all intervals of the same length on the distribution's support^[def. 17.6] $\text{supp}(\mathcal{U}[a, b]) = [a, b]$ are equally probable/likely.

$$f(x) = \frac{1}{b - a} \mathbb{1}_{x \in [a, b]} = \begin{cases} \frac{1}{b - a} = \text{const} & \text{if } a \leq x \leq b \\ 0 & \text{else} \end{cases} \quad (17.52)$$

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x - a}{b - a} & \text{if } a \leq x \leq b \\ 1 & x > b \end{cases} \quad (17.53)$$

$$\mathbb{E}[X] = \frac{a + b}{2} \quad \mathbb{V}(X) = \frac{(b - a)^2}{12} \quad (17.54)$$



5.2. Exponential Distribution

$\exp(\lambda)$

Definition 17.29 Exponential Distribution $X \sim \exp(\lambda)$: Is the continuous analogue to the geometric distribution ?? It describes the probability $f(x; \lambda)$ that a continuous Poisson process (i.e., a process in which events occur continuously and independently at a constant average rate) will succeed/change state after a time interval x .

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (17.55)$$

$$F(x; \lambda) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (17.56)$$

$$\mathbb{E}[X] = \frac{1}{\lambda} \quad \mathbb{V}(X) = \frac{1}{\lambda^2} \quad (17.57)$$

5.3. Laplace Distribution

Definition 17.30 Laplace Distribution:

$$\text{Laplace Distribution} \quad f(\mathbf{x}; \mu, \sigma) = \frac{1}{2\sigma} \exp\left(-\frac{|\mathbf{x} - \mu|}{\sigma}\right) \quad (17.58)$$

5.4. The Normal Distribution

$\mathcal{N}(\mu, \sigma)$

Definition 17.31 Normal Distribution $\mathbf{X} \sim \mathcal{N}(\mu, \sigma^2)$: Is a symmetric distribution where the population parameters μ, σ^2 are equal to the expectation and variance of the distribution:

$$\mathbb{E}[X] = \mu \quad \mathbb{V}(X) = \sigma^2 \quad (17.59)$$

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right\} \quad (17.60)$$

$$F(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left\{-\frac{1}{2}\left(\frac{u - \mu}{\sigma}\right)^2\right\} du \quad (17.61)$$

$$\varphi_X(u) = \exp\left\{iu\mu - \frac{u^2\sigma^2}{2}\right\} \quad (17.62)$$

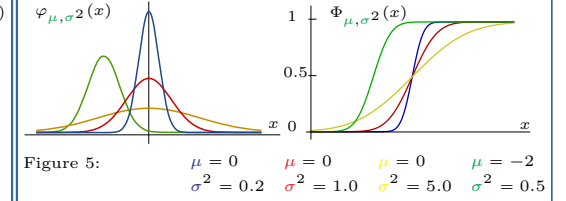


Figure 5: $\mu = 0, \sigma^2 = 0.2$ (green), $\mu = 0, \sigma^2 = 1.0$ (red), $\mu = 0, \sigma^2 = 5.0$ (yellow), $\mu = -2, \sigma^2 = 0.5$ (blue)

Property 17.14: $\mathbb{P}_X(\mu - \sigma \leq x \leq \mu + \sigma) = 0.66$

Property 17.15: $\mathbb{P}_X(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = 0.95$

5.5. The Standard Normal distribution

$\mathcal{N}(0, 1)$

Historic Problem: the cumulative distribution ?? does not have an analytical solution and numerical integration was not always computationally so easy. So how should people calculate the probability of x falling into certain ranges $\mathbb{P}(x \in [a, b])$?

Solution: use a standardized form/set of parameters (by convention) $\mathcal{N}_{0,1}$ and tabulate many different values for its cumulative distribution $\Phi(x)$ s.t. we can transform all families of Normal Distributions into the standardized version $\mathcal{N}(\mu, \sigma^2) \xrightarrow{z} \mathcal{N}(0, 1)$ and look up the value in its table.

Definition 17.32

Standard Normal Distribution $\mathbf{X} \sim \mathcal{N}(0, 1)$:

$$\mathbb{E}[X] = 0 \quad \mathbb{V}(X) = 1 \quad (17.63)$$

$$f(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad (17.64)$$

$$F(x; 0, 1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}u^2} du \quad (17.65)$$

$$x \in \mathbb{R} \quad \text{or} \quad -\infty < x < \infty \\ \psi_X(u) = e^{-\frac{u^2}{2}} \quad \varphi_X(u) = e^{-\frac{u^2}{2}} \quad (17.66)$$

Corollary 17.4

Standard Normal Distribution Notation: As the standard normal distribution is so commonly used people often use the letter Z in order to denote its the standard normal distribution and its α -quantile^[def. 17.19] is then denoted by:

$$z_\alpha = \Phi^{-1}(\alpha) \quad \alpha \in (0, 1) \quad (17.67)$$

5.5.1. Calculating Probabilities

Property 17.16 Symmetry: Let $z > 0$

$$\mathbb{P}(Z \leq z) = \Phi(z) \quad (17.68)$$

$$\mathbb{P}(Z \leq -z) = \Phi(-z) = 1 - \Phi(z) \quad (17.69)$$

$$\mathbb{P}(-a \leq Z \leq b) = \Phi(b) - \Phi(-a) = \Phi(b) - (1 - \Phi(a)) \\ \stackrel{a=b=z}{=} 2\Phi(z) - 1 \quad (17.70)$$

5.5.2. Linear Transformations of Normal Dist.

Proposition 17.1 Linear Transformation [proof ??]: Let X be a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then the linear transformed r.v. Y given by the *affine transformation* $Y = a + bX$ with $a \in \mathbb{R}, b \in \mathbb{R}_+$ follows:

$$Y \sim \mathcal{N}\left(a + b\mu, b^2\sigma^2\right) \iff f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right) \quad (17.71)$$

Proposition 17.2 Standardization [proof ??]: Let X be a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then there exists a linear transformation $Z = a + bX$ s.t. Z is a standard normally distributed random variable:

$$X \sim \mathcal{N}(\mu, \sigma^2) \xrightarrow{Z = \frac{X-\mu}{\sigma}} Z \sim \mathcal{N}(0, 1) \quad (17.72)$$

Note
If we know how many standard deviations our distribution is away from our target value then we can characterize it fully by the standard normal distribution.

Proposition 17.3 Standardization of the CDF: [proof ??] Let $F_X(X)$ be the cumulative distribution function of a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then the cumulative distribution function $\Phi_Z(z)$ of the standardized random normal variable $Z \sim \mathcal{N}(0, 1)$ is related to $F_X(X)$ by:

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \quad (17.73)$$

6. The Multivariate Normal distribution

Definition 17.33 Multivariate Normal/Gaussian: An \mathbb{R}^n -valued random variable $\mathbf{X} = (X_1 \dots X_n)$ is *Multivariate Gaussian/Normal* if every linear combination of its components is a (one-dimensional) Gaussian:

$$\exists \mu, \sigma : \mathcal{L}\left(\sum_{i=1}^n \alpha_i X_i\right) = \mathcal{N}(\mu, \sigma^2) \quad \forall \alpha_i \in \mathbb{R} \quad (17.74)$$

(possible degenerated $\mathcal{N}(0, 0)$ for $\forall \alpha_j = 0$)

Note

- Joint vs. multivariate:** a joint normal distribution can be a multivariate normal distribution or a product of univariate normal distributions **but**
- Multivariate refers to the number of variables that are placed as inputs to a function.

Definition 17.34 Multivariate Normal distribution $\mathbf{X} \sim \mathcal{N}_k(\mu, \Sigma)$: A k -dimensional random vector $\mathbf{X} = (X_1 \dots X_n)^\top$ with $\mu = (\mathbb{E}[\mathbf{x}_1] \dots \mathbb{E}[\mathbf{x}_k])^\top$ and $k \times k$ **p.s.d.** covariance matrix: $\Sigma := \mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^\top] = [\text{Cov}[\mathbf{x}_i, \mathbf{x}_j], 1 \leq i, j \leq k]$ follows a k -dim multivariate normal/Gaussian distribution if its law^(def. 16.25) satisfies:

$$f_{\mathbf{X}}(X_1, \dots, X_k) = \mathcal{N}(\mu, \Sigma) \quad (17.75)$$

$$= \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{X} - \mu)^\top \Sigma^{-1}(\mathbf{X} - \mu)\right)$$

Normalisation

$$\varphi_{\mathbf{X}}(\mathbf{u}) = \exp\left\{i\mathbf{u}^\top \mu - \frac{1}{2}\mathbf{u}^\top \Sigma \mathbf{u}\right\} \quad (17.76)$$

6.1. Joint Gaussian Distributions

Definition 17.35 Jointly Gaussian Random Variables: Two random variables X, Y both scalars or vectors, are said to be **jointly Gaussian** if the joint vector random variable $\mathbf{Z} = [X \ Y]^\top$ is again a GRV.

Property 17.17 proof ??
Joint Independent Gaussian Random Variables: Let X_1, \dots, X_n be \mathbb{R} -valued *independent* random variables with laws $\mathcal{N}(\mu_i, \sigma_i^2)$. Then the law of $\mathbf{X} = (X_1 \dots X_n)$ is a (multivariate) Gaussian distribution $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ with:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_n^2 \end{bmatrix} \quad \text{and} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_n \end{bmatrix} \quad (17.77)$$

Corollary 17.5 Quadratic Form:
If \mathbf{x} and \mathbf{y} are both independent GRVs $\mathbf{x} \sim \mathcal{N}(\mu_x, \Sigma_x)$ $\mathbf{y} \sim \mathcal{N}(\mu_y, \Sigma_y)$ then they are jointly Gaussian?? given by:

$$\mathbb{P}(\mathbf{x}, \mathbf{y}) = \mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{y}) \quad (17.78)$$

$$\propto \exp\left(-\frac{1}{2}\left\{(\mathbf{x} - \mu_x)^\top \Sigma_x^{-1}(\mathbf{x} - \mu_x) + (\mathbf{y} - \mu_y)^\top \Sigma_y^{-1}(\mathbf{y} - \mu_y)\right\}\right)$$

$$= \exp\left(-\frac{1}{2}\left[(\mathbf{x} - \mu_x)^\top \quad (\mathbf{y} - \mu_y)^\top\right] \begin{bmatrix} \Sigma_x^{-1} & 0 \\ 0 & \Sigma_y^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{x} - \mu_x \\ \mathbf{y} - \mu_y \end{bmatrix}\right)$$

$$\triangleq \exp\left(-\frac{1}{2}(\mathbf{z} - \mu_z)^\top \Sigma_z^{-1}(\mathbf{z} - \mu_z)\right)$$

Property 17.18 Marginal Distribution of Multivariate Gaussian: Let $\mathbf{X} = (X_1 \dots X_n)^\top \sim \mathcal{N}(\mu, \Sigma)$ be an \mathbb{R}^n valued Gaussian and let $V = \{1, 2, \dots, n\}$ be the index set of its variables. The k -variate marginal distribution of the Gaussian indexed by a subset of the variables:

$$\mathbf{A} = \{i_1, \dots, i_k\} \quad i_j \in V \quad (17.79)$$

is given by:

$$\mathbf{X} = (X_{i_1} \dots X_{i_k})^\top \sim \mathcal{N}(\mu_A, \Sigma_{AA}) \quad (17.80)$$

$$\Sigma = \begin{bmatrix} \sigma_{i_1, i_1}^2 & \sigma_{i_1, i_k}^2 \\ & \ddots \\ \sigma_{i_k, i_1}^2 & \sigma_{i_k, i_k}^2 \end{bmatrix} \quad \text{and} \quad \mu = \begin{bmatrix} \mu_{i_1} \\ \mu_{i_2} \\ \mu_{i_k} \end{bmatrix}$$

6.2. Conditional Gaussian Distributions

Property 17.19 Conditional Gaussian Distribution: Let $\mathbf{X} = (X_1 \dots X_n)^\top \sim \mathcal{N}(\mu, \Sigma)$ be an \mathbb{R}^n valued Gaussian and let $V = \{1, 2, \dots, n\}$ be the index set of its variables. Suppose we take two disjoint subsets of V :

$$\mathbf{A} = \{i_1, \dots, i_k\} \quad \mathbf{B} = \{j_1, \dots, j_m\} \quad i_l, j_l' \in V$$

then the conditional distribution of the random vector \mathbf{X}_A , conditioned on \mathbf{X}_B given by $\mathbb{P}(\mathbf{X}_A | \mathbf{X}_B = \mathbf{x}_B)$ is:

$$\mathbf{X}_A = (X_{i_1} \dots X_{i_k})^\top \sim \mathcal{N}(\mu_{A|B}, \Sigma_{A|B}) \quad (17.81)$$

$$\begin{bmatrix} \mu_{A|B} = \mu_A + \Sigma_{AB} \Sigma_{BB}^{-1} (\mathbf{x}_B - \mu_B) \\ \Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA} \end{bmatrix}$$

Note
Can be proofed using the matrix inversion lemma but is a very tedious computation.

Corollary 17.6 Conditional Distribution of Joint Gaussian's: Let \mathbf{X} and \mathbf{Y} be jointly Gaussian random vectors:

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{B} \end{bmatrix}\right) \quad (17.82)$$

then the *marginal* distribution of \mathbf{x} conditioned on \mathbf{y} can be written as:

$$X \sim \mathcal{N}(\mu_{X|Y}, \Sigma_{X|Y})$$

$$\begin{bmatrix} \mu_{X|Y} = \mu_X + \mathbf{C}\mathbf{B}^{-1}(\mathbf{y} - \mu_Y) \\ \Sigma_{X|Y} = \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\top \end{bmatrix} \quad (17.83)$$

add proofs

6.3. Transformations

Property 17.20 Multiples of Gaussian's **A_x:** Let $\mathbf{X} = (X_1 \dots X_n)^\top \sim \mathcal{N}(\mu, \Sigma)$ be an \mathbb{R}^n valued Gaussian and let $\mathbf{A} \in \mathbb{R}^{d \times n}$ then it follows:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} \in \mathbb{R}^d \quad \mathbf{Y} \sim \mathcal{N}(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^\top) \quad (17.84)$$

Property 17.21 Affine Transformation of GRVs: Let $\mathbf{y} \in \mathbb{R}^n$ be GRV, $\mathbf{A} \in \mathbb{R}^{d \times n}$, $\mathbf{b} \in \mathbb{R}^d$ and let \mathbf{x} be defined by the **affine transformation**^[def. 10.45]:

$$\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{b} \quad \mathbf{A} \in \mathbb{R}^{d \times n}, \mathbf{b} \in \mathbb{R}^d$$

Then \mathbf{x} is a GRV (see ??).

Property 17.22 Linear Combination of jointly GRVs: Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$ two jointly GRVs, and let \mathbf{z} be defined as:

$$\mathbf{z} = \mathbf{A}_x \mathbf{x} + \mathbf{A}_y \mathbf{y} \quad \mathbf{A}_x \in \mathbb{R}^{d \times n}, \mathbf{A}_y \in \mathbb{R}^{d \times m}$$

Then \mathbf{z} is GRV (see ??).

Definition 17.36 Gaussian Noise: Is statistical noise having a probability density function (PDF) equal to that of the normal/Gaussian distribution.

6.4. Gamma Distribution $\Gamma(x, \alpha, \beta)$

Definition 17.37 Gamma Distribution $X \sim \Gamma(x, \alpha, \beta)$: Is a widely used distribution that is related to the exponential distribution, Erlang distribution, and chi-squared distribution as well as Normal distribution:

$$f(x; \alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \quad (17.85)$$

$$\Gamma(\alpha) \stackrel{\text{eq. (5.81)}}{=} \int_0^\infty t^{\alpha-1} e^{-t} dt \quad (17.86)$$

with $\alpha, \beta \in \mathbb{R}_{>0}$

6.5. Chi-Square Distribution

6.6. Student's t-distribution

Definition 17.38 Student' t-distribution:

add

6.7. Delta Distribution

Definition 17.39 The delta function $\delta(\mathbf{x})$: The delta/dirac function $\delta(\mathbf{x})$ is defined by:

$$\int_{\mathbb{R}} \delta(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = f(0)$$

for any integrable function f on \mathbb{R} .

Or alternatively by:

$$\delta(x - x_0) = \lim_{\sigma \rightarrow 0} \mathcal{N}(x|x_0, \sigma) \quad (17.87)$$

$$\approx \infty \mathbb{1}_{\{x=x_0\}} \quad (17.88)$$

Property 17.23 Properties of δ :

- Normalization:** The delta function integrates to 1:

$$\int_{\mathbb{R}} \delta(x) dx = \int_{\mathbb{R}} \delta(x) \cdot c_1(x) dx = c_1(0) = 1$$

where $c_1(x) = 1$ is the constant function of value 1.

- Shifting:**

$$\int_{\mathbb{R}} \delta(x - x_0) f(x) dx = f(x_0) \quad (17.89)$$

- Symmetry:**

$$\int_{\mathbb{R}} \delta(-x) f(x) dx = f(0)$$

- Scaling:**

$$\int_{\mathbb{R}} \delta(\alpha x) f(x) dx = \frac{1}{|\alpha|} f(0)$$

Note

- In mathematical terms δ is not a function but a **generalized function**.
- We may regard $\delta(x - x_0)$ as a density with all its probability mass centered at the single point x_0 .
- Using a box/indicator function s.t. its surface is one and its width goes to zero, instead of a normal distribution ?? would be a non-differentiable/discrete form of the dirac measure.

Definition 17.40 Heaviside Step Function:

$$H(x) := \frac{d}{dx} \max\{x, 0\} \quad x \in \mathbb{R}_{\neq 0} \quad (17.90)$$

or alternatively:

$$H(x) := \int_{-\infty}^x \delta(s) ds \quad (17.91)$$

Proofs

Proof 17.11 Definition ??: Consider a sequence of n random $\{X_i\}_{i=1}^n$ Bernoulli experiments?? with success probability p . Define the r.v. Y_n to be the sum of the n Bernoulli variables:

$$Y_n = \sum_{i=1}^n X_i \quad n \in \mathbb{N}$$

i.e. the total number of successes. Now let's calculate the probability density function f_n of Y_n . First let $(x_1 \dots x_n) \in \{0, 1\}^n$ and let $y = \sum_{i=1}^n x_i$ a bit string of zeros and ones, with one occurring y times.

$$\mathbb{P}((X_1, X_2, \dots, X_n) = (x_1, x_2, \dots, x_n)) = (\underbrace{p \dots p}_y) \cdot (\underbrace{q \dots q}_{n-y \text{ times}}) = p^y (1-p)^{n-y}$$

However we need to take into account that there exists further realization $\mathbf{X} = \mathbf{x}$, that correspond to different orders of the elements in our two classes $\{0, 1\}$ which leads to

$$\frac{n!}{y!(n-y)!} = \binom{n}{y}: \quad f_n(y) = \binom{n}{y} p^y (1-p)^{n-y} \quad y \in \{0, 1, \dots, n\}$$

Proof 17.12: ??: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$F_Y(y) \stackrel{y \geq 0}{=} \mathbb{P}_Y(Y \leq y) = \mathbb{P}(a + bX \leq y) = \mathbb{P}_X\left(X \leq \frac{y-a}{b}\right) = F_X\left(\frac{y-a}{b}\right)$$

$$F_Y(y) \stackrel{y \leq 0}{=} \mathbb{P}_Y(Y \leq y) = \mathbb{P}(a + bX \leq y) = \mathbb{P}_X\left(X \geq \frac{y-a}{b}\right) = 1 - F_X\left(\frac{y-a}{b}\right)$$

Differentiating both expressions w.r.t. y leads to:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \begin{cases} \frac{1}{b} \frac{dF_X\left(\frac{y-a}{b}\right)}{d\frac{y-a}{b}} & \text{if } y \geq 0 \\ \frac{1}{-b} \frac{dF_X\left(\frac{y-a}{b}\right)}{d\frac{y-a}{b}} & \text{if } y < 0 \end{cases} = \frac{1}{|b|} f_X(x) \left(\frac{y-a}{b}\right)$$

??).

in order to prove that $Y \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$ we simply plug f_X in the previous expression:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma|b|}} \exp\left\{-\frac{1}{2}\left(\frac{\frac{y-a}{b} - \mu}{\sigma}\right)^2\right\} = \frac{1}{\sqrt{2\pi\sigma|b|}} \exp\left\{-\frac{1}{2}\left(\frac{y - (a + b\mu)}{\sigma|b|}\right)^2\right\}$$

Proof 17.13: ??: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$Z := \frac{X - \mu}{\sigma} = \frac{1}{\sigma} X - \frac{\mu}{\sigma} = aX + b \quad \text{with } a = \frac{1}{\sigma}, b = -\frac{\mu}{\sigma}$$

$$\sim \mathcal{N}(a\mu + b, a^2\sigma^2) \sim \mathcal{N}\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{\sigma^2}{\sigma^2}\right) \sim \mathcal{N}(0, 1)$$

Proof 17.14: ??: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$F_X(x) = \mathbb{P}(X \leq x) \stackrel{-\mu}{=} \mathbb{P}\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = \mathbb{P}\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

Proof 17.15: ?? scalar case
Let $y \sim \text{p}(y) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$ and define $\mathbf{x} = ay + b$ $a \in \mathbb{R}_+, b \in \mathbb{R}$
Using the Change of variables formula it follows:
 $\text{p}_x(\bar{x}) \stackrel{\text{eq. (16.46)}}{=} \frac{\text{p}_y(\bar{y})}{\left|\frac{dx}{dy}\right|} \left[\begin{array}{c} \bar{y}(\bar{x}) \\ \left|\frac{dx}{dy}\right| = a \end{array} \right]$
 $\bar{y} \stackrel{\bar{x}-b}{=} \frac{1}{a} \frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{1}{2\sigma^2} \left(\frac{\bar{x}-b}{a} - \mu\right)^2\right)$
 $= \frac{1}{\sqrt{2\pi a^2 \mu^2}} \exp\left(-\frac{1}{2\sigma^2 a^2} \underbrace{(\bar{x}-b-a\mu)^2}_{\mu_x}\right)$

Hence $x \sim \mathcal{N}(\mu_x, \sigma_x^2) = \mathcal{N}(a\mu + b, a^2\sigma^2)$

Note

We can also verify that we have calculated the right mean and variance by:
 $\mathbb{E}[x] = \mathbb{E}[ay + b] = a\mathbb{E}[y] + b = a\mu + b$
 $\mathbb{V}[x] = \mathbb{V}[ay + b] = a^2\mathbb{V}[y] = a^2\sigma^2$

Proof 17.16: ??
 $\text{p}_{\mathbf{X}}(\mathbf{u}) = \prod_{i=1}^n \text{p}_{X_i}(u_i)$
 $= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right)$
 $\varphi_{\mathbf{X}}(\mathbf{u}) = \exp\left\{iu_1\mu_1 - \frac{1}{2}\sigma_1u_1^2\right\} \cdots \exp\left\{iu_n\mu_n - \frac{1}{2}\sigma_nu_n^2\right\}$
 $= \exp\left\{i\sum_i^n u_n\mu_n - \frac{1}{2}\sum_i^n \sigma_nu_n^2\right\} = \exp\left\{i\mathbf{u}^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{u}\boldsymbol{\Sigma}\mathbf{u}\right\}$

Proof 17.17: ??
From ?? it follows immediately that \mathbf{z} is GRV $\mathbf{z} \sim \mathcal{N}(\mu_z, \Sigma_z)$ with:
 $\mathbf{z} = \mathbf{A}\boldsymbol{\xi}$ with $\mathbf{A} = \begin{bmatrix} \mathbf{A}_x & \mathbf{A}_y \end{bmatrix}$ and $\boldsymbol{\xi} = (\mathbf{x} \ \mathbf{y})$
Knowing that \mathbf{z} is a GRV it is sufficient to calculate μ_z and Σ_z in order to characterize its distribution:
 $\mathbb{E}[\mathbf{z}] = \mathbb{E}[\mathbf{A}_x x + \mathbf{A}_y y] = \mathbf{A}_x \mu_x + \mathbf{A}_y \mu_y$
 $\mathbb{V}[\mathbf{z}] = \mathbb{V}[\mathbf{A}\boldsymbol{\xi}] \stackrel{??}{=} \mathbf{A}\mathbb{V}[\boldsymbol{\xi}]\mathbf{A}^\top$
 $= \begin{bmatrix} \mathbf{A}_x & \mathbf{A}_y \end{bmatrix} \begin{bmatrix} \mathbb{V}[x] & \text{Cov}[x, y] \\ \text{Cov}[y, x] & \mathbb{V}[y] \end{bmatrix} \begin{bmatrix} \mathbf{A}_x & \mathbf{A}_y \end{bmatrix}^\top$
 $= \begin{bmatrix} \mathbf{A}_x & \mathbf{A}_y \end{bmatrix} \begin{bmatrix} \mathbb{V}[x] & \text{Cov}[x, y] \\ \text{Cov}[y, x] & \mathbb{V}[y] \end{bmatrix} \begin{bmatrix} \mathbf{A}_x^\top \\ \mathbf{A}_y^\top \end{bmatrix}$
 $= \mathbf{A}_x \mathbb{V}[x] \mathbf{A}_x^\top + \mathbf{A}_y \mathbb{V}[y] \mathbf{A}_y^\top$
 $+ \underbrace{\mathbf{A}_y \text{Cov}[y, x] \mathbf{A}_x^\top}_{=0\text{by independence}} + \underbrace{\mathbf{A}_x \text{Cov}[x, y] \mathbf{A}_y^\top}_{=0\text{by independence}}$
 $= \mathbf{A}_x \Sigma_x \mathbf{A}_x^\top + \mathbf{A}_y \Sigma_y \mathbf{A}_y^\top$

Note

Can also be proofed by using the normal definition of [def. 17.15] and tedious computations.

Proof 17.18: ?? If $\mathbf{x} = c_i$ i.e. the outcome c_i has occurred then it follows:
 $\prod_j^k \text{p}_i^{\delta[x=c_i]} = \text{p}_1^0 \cdots \text{p}_i^1 \cdots \text{p}_k^0 = 1 \cdots \text{p}_i \cdots 1 = p(\mathbf{x} = c_i | \text{p})$

Sampling Methods

1. Sampling Random Numbers

Most math libraries have uniform **random number generator (RNG)** i.e. functions to generate uniformly distributed random numbers $U \sim \mathcal{U}[a, b]$ (??).

Furthermore repeated calls to these RNG are independent, that is:

$$\begin{aligned} \mathbb{P}_{U_1, U_2}(u_1, u_2) &\stackrel{??}{=} \mathbb{P}_{U_1}(u_1) \cdot \mathbb{P}_{U_2}(u_2) \\ &= \begin{cases} 1 & \text{if } u_1, u_2 \in [a, b] \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Question: using samples $\{u_1, \dots, u_n\}$ of these CRVs with uniform distribution, how can we create random numbers with arbitrary discrete or continuous PDFs?

2. Inverse-transform Technique

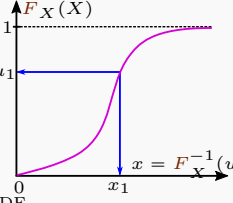
Idea

Can make use of section 1 and the fact that CDF are increasing functions ([def. 5.12]). **Advantage:**

- Simple to implement
- All discrete distributions can be generated via inverse-transform technique

Drawback:

- Not all continuous distributions can be integrated/have closed form solution for their CDF. E.g. Normal-, Gamma-, Beta-distribution.



2.1. Continuous Case

Definition 18.1 One Continuous Variable: Given: a desired continuous pdf f_X and uniformly distributed rn $\{u_1, u_2, \dots\}$:

1. Integrate the desired pdf f_X in order to obtain the desired cdf F_X :

$$F_X(x) = \int_{-\infty}^x f_X(t) dt \quad (18.1)$$

2. Set $F_X(X) \stackrel{!}{=} U$ on the range of X with $U \sim \mathcal{U}[0, 1]$.

3. Invert this equation/find the inverse $F_X^{-1}(U)$ i.e. solve:

$$U = F_X(X) = F_X \left(\underbrace{F_X^{-1}(U)}_X \right) \quad (18.2)$$

4. Plug in the uniformly distributed rn:

$$x_i = F_X^{-1}(u_i) \quad \text{s.t.} \quad x_i \sim f_X \quad (18.3)$$

Definition 18.2 Multiple Continuous Variable:

Given: a pdf of multiple rvs $f_{X,Y}$:

1. Use the product rule (??) in order to decompose $f_{X,Y}$:

$$f_{X,Y} = f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y) \quad (18.4)$$

2. Use ?? to first get a rv for y of $Y \sim f_Y(y)$.

3. Then with this fixed y use ?? again to get a value for x of $X \sim f_{X|Y}(x|y)$.

Proof 18.1: ??:

Claim: if U is a uniform rv on $[0, 1]$ then $F_X^{-1}(U)$ has F_X as its CDF.

Assume that F_X is strictly increasing ([def. 5.12]).

Then for any $u \in [0, 1]$ there must exist a **unique** x s.t. $F_X(x) = u$.

Thus F_X must be invertible and we may write $x = \underline{F_X^{-1}(u)}$.

Now let a arbitrary:

$$F_X(a) = \mathbb{P}(x \leq a) = \mathbb{P}(F_X^{-1}(U) \leq a)$$

Since F_X is strictly increasing:

$$\begin{aligned} \mathbb{P}(F_X^{-1}(U) \leq a) &= \mathbb{P}(U \leq F_X(a)) \\ &\stackrel{??}{=} \int_0^{F_X(a)} 1 dt = F_X(a) \end{aligned}$$

Note

Strictly speaking we may not assume that a CDF is **strictly** increasing but we as all CDFs are weakly increasing ([def. 5.12])

we may always define an auxiliary function by its infimum:

$$\hat{F}_X^{-1} := \inf \{x | F_X(x) \geq 0\} \quad u \in [0, 1] \quad (18.5)$$

2.2. Discret Case

Idea

Given: a desired $U \sim \mathcal{U}[0, 1]$ and discrete pmf p_X s.t. $\mathbb{P}(X = x_i) = p_X(x_i)$ and uniformly distributed rn $\{u_1, u_2, \dots\}$.
Goal: given a uniformly distributed rn u determine k s.t.:

$$\begin{aligned} \sum_{i=1}^{k-1} p_X(x_i) < U \leq \sum_{i=1}^k p_X(x_i) &\iff F_X(x_{k-1}) < u \leq F_X(x_k) \end{aligned} \quad (18.6)$$

and return x_k .

Definition 18.3 One Discret Variable:

1. Compute the CDF of p_X ([def. 17.8])

$$F_X(x) = \sum_{t=-\infty}^x p_X(t) \quad (18.7)$$

2. Given the uniformly distributed rn $\{u_i\}_{i=1}^n$ find k^i ($\hat{=}$ inversion) s.t.:

$$F_X(x_{k(i)-1}) < u_i \leq F_X(x_{k(i)}) \quad \forall u_i \quad (18.8)$$

Proof 18.2: ??: First of all notice that we can always solve for an unique x_k .

Ask: why are Discret CRV always strictly increasing/unique?

Given a fixed x_k determine the values of u for which:

$$F_X(x_{k-1}) < u \leq F_X(x_k) \quad (18.9)$$

Now observe that:

$$\begin{aligned} u &\leq F_X(x_k) = F_X(x_{k-1}) + p_X(x_k) \\ \implies F_X(x_{k-1}) < u &\leq F_X(x_{k-1}) + p_X(x_k) \end{aligned}$$

The probability of U being in $(F_X(x_{k-1}), F_X(x_k)]$ is:

$$\begin{aligned} \mathbb{P}(U \in [F_X(x_{k-1}), F_X(x_k)]) &= \int_{F_X(x_{k-1})}^{F_X(x_k)} p_U(t) dt \\ &= \int_{F_X(x_{k-1})}^{F_X(x_k)} 1 dt = \int_{F_X(x_{k-1})}^{F_X(x_{k-1}) + p_X(x_k)} 1 dt = p_X(x_k) \end{aligned}$$

Hence the random variable $x_k \in \mathcal{X}$ has the pdf p_X .

Definition 18.4

Multiple Continuous Variables (Option 1):

Given: a pdf of multiple rvs $p_{X,Y}$:

1. Use the product rule (??) in order to decompose $p_{X,Y}$:

$$p_{X,Y} = p_{X,Y}(x, y) = p_{X|Y}(x|y) p_Y(y) \quad (18.10)$$

2. Use ?? to first get a rv for y of $Y \sim p_Y(y)$.

3. Then with this fixed y use ?? again to get a value for x of $X \sim p_{X|Y}(x|y)$.

Definition 18.5

Multiple Continuous Variables (Option 2):

Note: this only works if \mathcal{X} and \mathcal{Y} are finite.

Given: a pdf of multiple rvs $p_{X,Y}$ let $N_x = |\mathcal{X}|$ and $N_y = |\mathcal{Y}|$ the number of elements in \mathcal{X} and \mathcal{Y} .

Define

$$\begin{aligned} p_Z(1) &= p_{X,Y}(1, 1), p_Z(2) = p_{X,Y}(1, 2), \dots \\ \dots, p_Z(N_x \cdot N_y) &= p_{X,Y}(N_x, N_y) \end{aligned}$$

Then simply apply ?? to the auxiliary pdf p_Z .

1. Use the product rule (??) in order to decompose $f_{X,Y}$:

$$f_{X,Y} = f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y) \quad (18.11)$$

2. Use ?? to first get a rv for y of $Y \sim f_Y(y)$.

3. Then with this fixed y use ?? again to get a value for x of $X \sim f_{X|Y}(x|y)$.

also examples see comment in code text

3. Monte Carlo Methods

3.1. Monte Carlo (MC) Integration

Integration methods s.a. Simpson integration ([def. 13.34]) suffer heavily from the curse of dimensionality. An n-order ([def. 13.31]) quadrature scheme \mathcal{Q}_n in 1-dimension is usually of order n/d in d-dimensions.

Idea estimate an integral stochastically by drawing sample from some distribution.

Definition 18.6 Monte Carlo Integration:

$$3 + 4 \quad (18.12)$$

3.2. Rejection Sampling

3.3. Importance Sampling

Descriptive Statistics

1. Populations and Distributions

Definition 19.1 Population $\{x_i\}_{i=1}^N$:
Is the entire set of entities from which we can draw sample.

Definition 19.2 Families of Probability Distributions p_θ :
Are probability distributions that vary only by a set of hyper parameters θ ??.

Definition 19.3 Population/Statistical Parameter θ :
Are the parameters defining families of probability distributions??

Explanation 19.1 (Definition ??). *Such hyper parameters are often characterized by populations following a certain family of distributions with the help of a statistic. Hence they are called populations or statistical parameters.*

1.1. Characteristics of Populations

Definition 19.4 Population Mean: Given a population $\{x_i\}_{i=1}^N$ of size N its variance is defined as:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad (19.1)$$

Definition 19.5 Population Variance: Given a population $\{x_i\}_{i=1}^N$ of size N its variance is defined as: $\{x_i\}_{i=1}^N$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad (19.2)$$

Note
The population variance and mean are equally to the mean derived from the true distribution of the population.

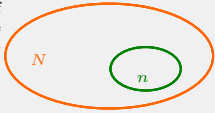
2. Sample Statistics

Definition 19.6 (Sample) Statistic: A statistic is a measurable function T that assigns a **single** value t to a sample of random variables or population:
$$t: \mathbb{R}^n \mapsto \mathbb{R} \quad t = T(X_1, \dots, X_n)$$

E.g. T could be the mean, variance,...

Definition 19.7 Degrees of freedom of a Statistic: Is the number of values in the final calculation of a statistic that are free to vary.

Note
The function itself is independent of the sample's distribution; that is, the function can be stated before realization of the data.



3. Point and Interval Estimation

Assume a population X with a given sample $\{x_i\}_{i=1}^n$ follows some family of distributions:
$$X \sim p_X(\cdot; \theta) \quad (19.3)$$

how can we estimate the correct value of the parameter θ or some function of that parameter $\tau(\theta)$?

3.1. Point Estimates

Definition 19.8 (Point) Estimator $\hat{\theta}$:
Is a statistic?? that tries estimates an unknown parameter θ of an underlying family of distributions?? for a given sample $\{\mathbf{x}_i\}_{i=1}^n$ of that distribution:
$$\hat{\theta} = t(\mathbf{x}_1, \dots, \mathbf{x}_n) \quad (19.4)$$

Note
The other kind of estimators are interval estimators which do not calculate a statistic **but** an interval of plausible values of an unknown population parameter θ .
The most prevalent forms of interval estimation are:
• Confidence intervals (frequentist method).
• Credible intervals (Bayesian method).

3.1.1. Empirical Mean

Definition 19.9 Sample/Empirical Mean \bar{x} :
The sample mean is an estimate/statistic of the population mean?? and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$\bar{x} = \hat{\mu}_X = \frac{1}{n} \sum_{i=1}^n x_i \quad (19.5)$$

Corollary 19.1 [proof ??]
Unbiased Sample Mean:
The sample mean estimator is unbiased:
$$\mathbb{E}[\hat{\mu}_X] = \mu \quad (19.6)$$

Corollary 19.2 [Proof ??]
Variance of the Sample Mean:
The variance of the sample mean estimator is given by:
$$\mathbb{V}[\hat{\mu}_X] = \frac{1}{n} \sigma_X^2 \quad (19.7)$$

3.1.2. Empirical Variance

Definition 19.10 Biased Sample Variance:
The sample variance is an estimate/statistic of the population variance?? and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$s_n^2 = \hat{\sigma}_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad (19.8)$$

Definition 19.11 (Unbiased) Sample Variance: [proof ??]
The unbiased form of the sample variance?? is given by:
$$s^2 = \hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2 \quad (19.9)$$

Definition 19.12 Bessel's Correction: The factor $\frac{n}{n-1}$ (19.10)
is called Bessel's correction. Multiplying the uncorrected population variance ?? by this term yields an unbiased estimated of the variance.

Attention:
• The Bessel correction holds for the variance but not for the standard deviation.
• Usually only the unbiased variance is used and sometimes also denoted by s_n^2

3.2. Interval Estimates

Definition 19.13 Interval Estimator $\hat{\theta}$:
Is an estimator that tries to bound an unknown parameter θ of an underlying family of distributions?? for a given sample $\{\mathbf{x}_i\}_{i=1}^n$ of that distribution.
Let $\hat{\theta} \in \Theta$ and define two point statistics?? g and h then an interval estimate is defined as:
$$\mathbb{P}(L_n < \theta < U_n) = \gamma \quad \forall \theta \in \Theta \quad L_n = g(\mathbf{x}_1, \dots, \mathbf{x}_n) \quad \gamma \in [0, 1] \quad U_n = h(\mathbf{x}_1, \dots, \mathbf{x}_n) \quad (19.11)$$

Statistical Tests

4. Parametric Hypothesis Testing

Definition 19.14 Parametric Hypothesis Testing:
Hypothesis testing is a statistical procedure in which a hypothesis is tested based on sampled data X_1, \dots, X_n .

4.1. Null Hypothesis

Definition 19.15 Null Hypothesis H_0 :
A null hypothesis H_0 is an *assumption* on a population?? parameter?? θ :

$$H_0: \theta = \theta_0 \quad (19.12)$$

Note
Often, a null hypothesis cannot be verified, but can only be falsified.

Definition 19.16 Alternative Hypothesis H_A/H_1 :
The alternative hypothesis H_1 is an *assumption* on a population?? parameter?? θ that is opposite to the null hypothesis.

$$H_A: \theta \begin{cases} > \theta_0 & \text{(one-sided)} \\ < \theta_0 & \text{(one-sided)} \\ \neq \theta_0 & \text{(two-sided)} \end{cases} \quad (19.13)$$

4.2. Test Statistic

The decision on the hypothesis test is based on a sample from the population $X(n) = \{X_1, \dots, X_n\}$ however the decision is usually not based on single sample but a sample statistic?? as this is easier to use.

Definition 19.17 Test Statistic/Testing Parameter T :
Is a sample statistic?? used for hypothesis tests in order to give evidence for or against a hypothesis:
$$t_n = T(D_n) = T(\{X_1, \dots, X_n\}) \quad (19.14)$$

4.3. Sampling Distribution

Definition 19.18 $T_{\theta_0}(t)$
Null Distribution/Sampling Distribution under H_0 :
Let $D_n = \{X_1, \dots, X_n\}$ be a random sample from the true population p_{pop} and let $T(D_n)$ be a test statistic of that sample.
The probability distribution of the test statistic under the assumption that the null hypothesis is true is called *sampling distribution*:
$$t \sim T_{\theta_0} = T(t|H_0 \text{ true}) \quad X_i \sim p_{\text{pop}} \quad (19.15)$$

4.4. The Critical Region

Given a sample $D_n = \{X_1, \dots, X_n\}$ of the true population p_{pop} how should we decide whether the null hypothesis should be rejected or not?
Idea: let \mathcal{T} be the set of all possible values that the sample statistic T can map to. Now let's split \mathcal{T} in two disjoint sets \mathcal{T}_0 and \mathcal{T}_1 :
$$\mathcal{T} = \mathcal{T}_0 \cup \mathcal{T}_1 \quad \mathcal{T}_0 \cap \mathcal{T}_1 = \emptyset$$

• if $t_n = T(X_n) \in \mathcal{T}_0$ we accept the null hypothesis H_0
• if $t_n = T(X_n) \in \mathcal{T}_1$ we reject the null hypothesis for H_1

Definition 19.19 Critical/Rejection Region \mathcal{T}_1 :
Is the set of all values of the test statistic?? t_n that causes us to reject the Null Hypothesis in favor of the alternative hypothesis H_A :
$$K = \mathcal{T}_1 = \{T: H_0 \text{ rejected}\} \quad (19.16)$$

Definition 19.20 Acceptance Region \mathcal{T}_0 :
Is the region where we accept the null hypothesis H_0 .
$$\mathcal{T}_0 = \{T: H_0 \text{ accepted}\} \quad (19.17)$$

Definition 19.21 Critical Value c :
Is the value of the *critical region* $c \in \mathcal{T}_1$ which is closest to the *region of acceptance*??:

4.5. Type I&II Errors

Definition 19.22 False Positive **Type I Error:**
Is the rejection of the null hypothesis H_0 , even-tough it is true
$$\text{Test rejects } H_0|H_0 \text{ true} \iff t_n \in \mathcal{T}_1|H_0 \text{ true} \quad (19.18)$$

Definition 19.23 False Negative **Type II Error:**
Is the acceptance of a null hypothesis H_0 , even-tough its false:
$$\text{Test accepts } H_0|H_A \text{ true} \iff t_n \in \mathcal{T}_0|H_A \text{ true} \quad (19.19)$$

Types of Errors

Decision	H_0 true	H_0 false	
Accept	TN	Type II (FN)	
Reject	Type I (FP)	TP	

4.6. Statistical Significance & Power

Question: how should we choose the split $\{\mathcal{T}_0, \mathcal{T}_1\}$?
The bigger we choose Θ_1 (and thus the smaller Θ_0) the more likely it is to accept the alternative.
Idea: take the position of the adversary and choose Θ_1 so small that $\theta \in \Theta_1$ has only a small *probability* of occurring.

Definition 19.24 (Statistical) Significance α :
A study's defined significance level α denotes the probability to incur a *Type I Error*??:
$$\mathbb{P}(t_n \in \mathcal{T}_1|H_0 \text{ true}) = \mathbb{P}(\text{test rejects } H_0|H_0 \text{ true}) \leq \alpha \quad (19.20)$$

Definition 19.25 Probability Type II Error β :
A test probability to for a *false negative*?? is defined as:
$$\beta(t_n) = \mathbb{P}(t_n \in \mathcal{T}_0|H_1 \text{ true}) = \mathbb{P}(\text{test accepts } H_0|H_1 \text{ true}) \quad (19.21)$$

Definition 19.26 (Statistical) Power $1 - \beta$:
A study's power $1 - \beta$ denotes a tests probability for a *true positive*:
$$1 - \beta(t_n) = \mathbb{P}(t_n \in \mathcal{T}_1|H_1 \text{ true}) = \mathbb{P}(\text{test rejects } H_0|H_1 \text{ true}) \quad (19.22) \quad (19.23)$$

Corollary 19.3 Types of Split:
The Critical region is chosen s.t. we incur a Type I Error with probability less than α , which corresponds to the type of the test??:

$$\mathbb{P}(c_2 \leq X \leq c_1) \leq \alpha \quad \text{two-sided}$$

or
$$\mathbb{P}(c_2 \leq X) \leq \frac{\alpha}{2} \quad \text{and} \quad \mathbb{P}(X \leq c_1) \leq \frac{\alpha}{2}$$

$$\mathbb{P}(c_2 \leq X) \leq \alpha \quad \text{one-sided}$$

$$\mathbb{P}(X \leq c_1) \leq \alpha \quad \text{one-sided}$$

	Truth		
Decision \	H_0 true	H_0 false	
H_0 accept	$1 - \alpha$	$1 - \beta$	
H_0 rejected	α	β	

4.7. P-Value

Definition 19.27 P-Value p :
Given a test statistic $t_n = T(X_1, \dots, X_n)$ the p-value $p \in [0, 1]$ is the smallest value s.t. we reject the null hypothesis:
$$p := \inf \{\alpha | t_n \in \mathcal{T}_1\} \quad t_n = T(X_1, \dots, X_n) \quad (19.24)$$

Explanation 19.2.
• The smaller the p-value the less likely is an observed statistic t_n and thus the higher is the evidence against a null hypothesis.
• A null hypothesis has to be rejected if the p-value is bigger than the chosen significance niveau α .

5. Conducting Hypothesis Tests

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- Select an appropriate test statistic?? T .
- Define the null hypothesis H_0 and the alternative hypothesis H_1 for T .
- Find the sampling distribution?? $T_{\theta_0}(t)$ for T , given H_0 true.
- Chose the significance level α
- Evaluate the test statistic $t_n = T(X_1, \dots, X_n)$ for the sampled data.
- Determine the p-value p .
- Make a decision (accept or reject H_0)

5.1. Tests for Normally Distributed Data

Let us consider an i.i.d. sample of observations $\{x_i\}_{i=1}^n$, of a normally distributed population $X_{\text{pop}} \sim \mathcal{N}(\mu, \sigma^2)$. From ??? it follows that the *mean of the sample* is distributed as:

$$\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$$

thus the mean of the sample \bar{X}_n should equal the mean μ of the population. We now want to test the null hypothesis:

$$H_0 : \mu = \mu_0 \iff \bar{X}_n \sim \mathcal{N}(\mu_0, \sigma^2/n) \tag{19.25}$$

This is obviously only likely if the realization \bar{x}_n is close to μ_0 .

5.1.1. Z-Test σ known

Definition 19.28 Z-Test:

For a realization of Z with $\{x_i\}_{i=1}^n$ and mean \bar{x}_n :

$$z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}$$

we *reject the null hypothesis* $H_0 : \mu = \mu_0$ for the alternative H_A for significance niveau?? α if:

$$\begin{aligned} |z| \geq z_{1-\frac{\alpha}{2}} &\iff z \leq z_{\frac{\alpha}{2}} \vee z \geq z_{1-\frac{\alpha}{2}} \\ &\iff z \in \mathcal{T}_1 = \left(-\infty, -z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}}, \infty\right) \\ z \geq z_{1-\alpha} &\iff z \in \mathcal{T}_1 = [z_{1-\alpha}, \infty) \\ z \leq z_{\alpha} = -z_{1-\alpha} &\iff z \in \mathcal{T}_1 = (-\infty, -z_{\alpha}] = (\infty, -z_{1-\alpha}] \end{aligned} \tag{19.26}$$

Notes

- Recall from [def. 17.19] and ?? that:
 z_{α} i.e. $\alpha=0.05$ $z_{0.05} = \Phi^{-1}(\alpha) \iff \mathbb{P}(Z \leq z_{0.05}) = 0.05$
- $|z| \geq z_{1-\frac{\alpha}{2}}$ which stands for:
 $\mathbb{P}(Z \leq z_{0.05}) + \mathbb{P}(Z \geq z_{0.95}) = \mathbb{P}(Z \leq -z_{1-0.05}) + \mathbb{P}(Z \geq z_{0.95}) = \mathbb{P}(|Z| \geq z_{0.95})$
can be rewritten as:
 $z \geq z_{1-\frac{\alpha}{2}} \vee -z \geq z_{1-\frac{\alpha}{2}} \iff z \leq -z_{1-\frac{\alpha}{2}} = z_{\frac{\alpha}{2}}$
- One usually goes over to the standard normal distribution ?? and thus test how far one is away from zero mean \Rightarrow Z-test.
- We thus inquire a Type I error with probability α and should be small i.e. 1%.

5.1.2. t-Test σ unknown

In reality we usually do not know the true σ of the whole data set and thus calculate it over our sample. This however increases uncertainty and thus our sample does no longer follow a normal distribution but a **t-distribution** wiht $n-1$ degrees of freedom:

$$T \sim t_{n-1} \tag{19.27}$$

Definition 19.29 t-Test:

For a realization of T with $\{x_i\}_{i=1}^n$ and mean \bar{x}_n :

$$t = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}$$

we *reject the null hypothesis* $H_0 : \mu = \mu_0$ for the alternative H_A if:

$$\begin{aligned} |t| \geq t_{n-1, 1-\frac{\alpha}{2}} &\iff t \in \mathcal{T}_1 = \left(-\infty, -t_{n-1, 1-\frac{\alpha}{2}}\right] \cup \left[t_{n-1, 1-\frac{\alpha}{2}}, \infty\right) \\ t \geq t_{n-1, 1-\alpha} &\iff t \in \mathcal{T}_1 = [t_{n-1, 1-\alpha}, \infty) \\ t \leq t_{n-1, \alpha} = -t_{n-1, 1-\alpha} &\iff t \in \mathcal{T}_1 = (-\infty, -t_{n-1, \alpha}] = (\infty, -t_{n-1, 1-\alpha}] \end{aligned}$$

Notes

- The t-distribution has fatter tails as the normal distribution \Rightarrow rare event become more likely
- For $n \rightarrow \infty$ the t-distribution goes over into the normal distribution
- The t-distribution gains a degree of foredoom for each sample and loses one for each parameter we are interested in \Rightarrow n -samples and we are interested in one parameter μ .

5.2. Confidence Intervals

Now we are interested in the opposite of the critical region?? namely the region of plausible values.

Definition 19.30 Confidence Interval I :

Let $D_n = \{X_1, \dots, X_n\}$ be a *sample* of observations and T_n a sample statistic of that sample. The confidence interval is defined as:

$$I(D_n) = \{\theta_0 : T_n(D_n) \in \mathcal{T}_0\} = \{\theta_0 : H_0 \text{ is not rejected}\} \tag{19.28}$$

Corollary 19.4 : The confidence interval captures the unknown parameter θ with probability $1 - \alpha$:

$$\mathbb{P}_{\theta}(\theta \in I(D_n)) = \mathbb{P}(T_n(D_n) \in \mathcal{T}_0) = 1 - \alpha \tag{19.29}$$

add page 91 confidence intervals z-test and t-test

6. Inferential Statistics

Goal of Inference

- 1
- 2
- What is a good guess of the parameters of my model?
- How do I quantify my uncertainty in the guess?

7. Examples

Example 19.1 ??: Let x be uniformly distributed on $[0, 1]$ (??) with pmf $\mathsf{p}_X(x)$ then it follows:
 $\frac{dy}{dx} = \frac{1}{\mathsf{p}_Y(y)} \Rightarrow dx = dy \mathsf{p}_Y(y) \Rightarrow x = \int_{-\infty}^y \mathsf{p}_Y(t) dt = F_Y(x)$

Example 19.2 ??: Let

add <https://www.youtube.com/watch?v=WUUhTVIRagg>

Example 19.3 Family of Distributions: The family of normal distribution \mathcal{N} has two parameters $\{\mu, \sigma^2\}$

Example 19.4 Test Statistic: Lets assume the test statistic follows a normal distribution:
 $T \sim \mathcal{N}(\mu; 1)$

however we are unsure about the population parameter?? $\theta = \mu$ but assume its equal to θ_0 thus the null-and alternative hypothesis are:
 $H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$

Example 19.5 Binomialtest:
Given: a manufacturer claims that a maximum of 10% of its delivered components are substandard goods.
In a sample of size $n = 20$ we find $x = 5$ goods that do not fulfill the standard and are skeptical that what the manufacture claims is true, so we want to test:
 $H_0 : \mathsf{p} = \mathsf{p}_0 = 0.1 \qquad \text{vs.} \qquad H_A : \mathsf{p} > 0.1$

We model the number of number of defective goods using the binomial distribution??

$$X \sim \mathcal{B}(n, \mathsf{p}), n = 20 \quad \mathbb{P}(X \geq x) = \sum_{k=x}^n \binom{n}{k} \mathsf{p}^k (1 - \mathsf{p})^{n-k} \\ \sim \mathcal{T}(n, \mathsf{p})$$

from this we find:
 $\mathbb{P}_{\mathsf{p}_0}(X \geq 4) = 1 - \mathbb{P}_{\mathsf{p}_0}(X \leq 3) = 0.13$
 $\mathbb{P}_{\mathsf{p}_0}(X \geq 5) = 1 - \mathbb{P}_{\mathsf{p}_0}(X \leq 4) = 0.04 \leq \alpha$
thus the probability that equal 5 or more then 5 parts out of the 20 are rejects is less then 4%.
 \Rightarrow throw away null hypothesis for the 5% niveau in favor to the alternative.
 \Rightarrow the 5% significance niveau is given by $K = \{5, 6, \dots, 20\}$

Note

If $x < n/2$ it is faster to calculate $\mathbb{P}(X \geq x) = 1 - \mathbb{P}(X \leq x - 1)$

8. Proofs

Proof 19.1: ??:

$$\mathbb{E}[\hat{\mu}_X] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^n x_i\right] = \frac{1}{n} \mathbb{E}\left[\underbrace{\mu + \dots + \mu}_{1, \dots, n}\right]$$

Proof 19.2: ??:

$$\mathbb{V}[\hat{\mu}_X] = \mathbb{V}\left[\frac{1}{n} \sum_{i=1}^n x_i\right] \stackrel{\text{Property 17.10}}{=} \frac{1}{n^2} \mathbb{V}\left[\sum_{i=1}^n x_i\right] \\ \frac{1}{n^2} n \mathbb{V}[X] = \frac{1}{n} \sigma^2$$

Proof 19.3: definition ??:

$$\begin{aligned} \mathbb{E}[\hat{\sigma}_X^2] &= \mathbb{E}\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2\right] \\ &= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^n (x_i^2 - 2x_i \bar{x} + \bar{x}^2)\right] \\ &= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x}^2\right] \\ &= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^n x_i^2 - 2n\bar{x} \cdot n\bar{x} + n\bar{x}^2\right] \\ &= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^n x_i^2 - n\bar{x}^2\right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n \mathbb{E}[x_i^2] - n\mathbb{E}[\bar{x}^2]\right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (\sigma^2 + \mu^2) - n\mathbb{E}[\bar{x}^2]\right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (\sigma^2 + \mu^2) - n\left(\frac{1}{n}\sigma^2 + \mu^2\right)\right] \\ &= \frac{1}{n-1} \left[(n\sigma^2 + n\mu^2) - (\sigma^2 + n\mu^2)\right] \\ &= \frac{1}{n-1} [n\sigma^2 - \sigma^2] = \frac{1}{n-1} [(n-1)\sigma^2] = \sigma^2 \end{aligned}$$

Stochastic Calculus

Stochastic Processes

Definition 20.1 Random/Stochastic Process An (\mathbb{R}^d -valued) stochastic process is a collection of (\mathbb{R}^d -valued) random variables X_t on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The index set \mathcal{T} is usually representing time and can be either an interval $[t_1, t_2]$ or a discrete set $\{t_1, t_2, \dots\}$. Therefore, the random process X can be written as a function: $X : \mathcal{T} \subseteq \mathbb{R}_+ \times \Omega \mapsto \mathbb{R}^d \iff (t, \omega) \mapsto X(t, \omega) \quad (20.1)$
Definition 20.2 Sample path/Trajector/Realization: Is the <i>stochastic/noise signal</i> $r(\cdot, \omega)$ on the index set ^[def. 2.1] \mathcal{T} , that we obtain be sampling ω from Ω .
Notation Even though the r.v. X is a function of two variables, most books omit the argument of the sample space $X(t, \omega) := X(t)$
Corollary 20.1 Strictly Positive Stochastic Processes: A stochastic process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$ is called strictly positive if it satisfies: $X_t > 0 \quad \text{\textcolor{red}{P-a.s.}} \quad \forall t \in \mathcal{T} \quad (20.2)$
Definition 20.3 Random/Stochastic Chain is a collection of random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ ^[def. 16.1] . The random variables are ordered by an associated index set ^[def. 2.1] \mathcal{T} and take values in the same mathematical <i>discrete state space</i> ?? S , which must be measurable w.r.t. some σ -algebra ^[def. 16.6] Σ . Therefore for a given probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a measurable space (S, Σ) , the random <i>chain</i> X is a collection of S -valued random variables that can be written as: $X : \mathcal{T} \times \Omega \mapsto S \iff (t, \omega) \mapsto X(t, \omega) \quad (20.3)$
Definition 20.4 Index/Parameter Set \mathcal{T} : Usually represents time and can be either an interval $[t_1, t_2]$ or a discrete set $\{t_1, t_2, \dots\}$.
Definition 20.5 State Space S : Is the range/possible values of the random variables of a stochastic process?? and must be measurable ^[def. 16.7] w.r.t. some σ -algebra Σ .
Sample-vs. State Space Sample space ^[def. 16.2] hints that we are working with probabilities i.e. probability measures will be defined on our sample space. State space is used in dynamics, it implies that there is a time progression, and that our system will be in different states at time progresses.
Definition 20.6 Sample path/Trajector/Realization: Is the <i>stochastic/noise signal</i> $r(\cdot, \omega)$ on the index set \mathcal{T} , that we obtain be sampling ω from Ω .
Notation Even though the r.v. X is a function of two variables, most books omit the argument of the sample space $X(t, \omega) := X(t)$
1.1. Filtrations Definition 20.7 Filtration A collection $\{\mathcal{F}_t\}_{t \geq 0}$ of sub σ -algebras ^[def. 16.6] $\{\mathcal{F}_t\}_{t \geq 0} \in \mathcal{F}$ is called filtration if it is <i>increasing</i> : $\mathcal{F}_s \subseteq \mathcal{F}_t \quad \forall s \leq t \quad (20.4)$
Explanation 20.1 (Definition ??). <i>A filtration describes the flow of information i.e. with time we learn more information.</i>
Definition 20.8 Filtered Probability Space A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ together with a filtration $\{\mathcal{F}_t\}_{t \geq 0}$ is called a <i>filtered probability space</i> .

Definition 20.9 Adapted Process: A stochastic process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$ is called adapted <i>to a</i> filtration \mathbb{F} if: $X_t \text{ is } \mathcal{F}_t\text{-measurable} \quad \forall t \quad (20.5)$ That is the value of X_t is observable at time t
Definition 20.10 Predictable Process: A stochastic process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$ is called predictable <i>w.r.t. a</i> filtration \mathbb{F} if: $X_t \text{ is } \mathcal{F}_{t-1}\text{-measurable} \quad \forall t \quad (20.6)$ That is the value of X_t is known at time $t - 1$
Note The price of a stock will usually be adapted since date k prices are known at date k . On the other hand the interest rate of a bank account is usually already known at the beginning $k - 1$, s.t. the interest rate r_t ought to be \mathcal{F}_{k-1} measurable, i.e. the process $r = (r_k)_{k=1, \dots, T}$ should be predictable.
Corollary 20.2 : The amount of information of an adapted random process is increasing see ??.
2. Martingales Definition 20.11 Martingales: A stochastic process $X(t)$ is a martingale on a <i>filtered probability space</i> $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$ if the following conditions hold: <ol style="list-style-type: none">Given $s \leq t$ the best prediction of $X(t)$, with a filtration $\{\mathcal{F}_s\}$ is the current expected value: $\forall s \leq t \quad \mathbb{E}[X(t) \mathcal{F}_s] = X(s) \quad \text{a.s.} \quad (20.7)$The expectation is finite: $\mathbb{E}[X(t)] < \infty \quad \forall t \geq 0 \quad X(t) \text{ is } \{\mathcal{F}_t\}_{t \geq 0} \text{ adapted} \quad (20.8)$
Interpretation <ul style="list-style-type: none">For any \mathcal{F}_s-adapted process the best prediction of $X(t)$ is the currently known value $X(s)$ i.e. if $\mathcal{F}_s = \mathcal{F}_{t-1}$ then the best prediction is $X(t - 1)$A martingale models fair games of limited information.
Definition 20.12 Auto Covariance Describes the covariance ^[def. 17.16] between two values of a stochastic process $(\mathbf{X}_t)_{t \in \mathcal{T}}$ at different time points t_1 and t_2 . $\gamma(t_1, t_2) = \text{Cov}[\mathbf{X}_{t_1}, \mathbf{X}_{t_2}] = \mathbb{E}[(\mathbf{X}_{t_1} - \mu_{t_1})(\mathbf{X}_{t_2} - \mu_{t_2})] \quad (20.9)$ For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance: $\gamma(t, t) = \text{Cov}[\mathbf{X}_t, \mathbf{X}_t] \stackrel{\text{eq. (17.35)}}{=} \mathbb{V}[\mathbf{X}_t] \quad (20.10)$
Notes <ul style="list-style-type: none">Hence the autocorrelation describes the correlation of a function or signal with itself at a previous time point.Given a random time dependent variable $\mathbf{x}(t)$ the autocorrelation function $\gamma(t, t - \tau)$ describes how <i>similar</i> the time translated function $\mathbf{x}(t - \tau)$ and the original function $\mathbf{x}(t)$ are.If there exists some relation between the values of the time series that is non-random, then the autocorrelation is non-zero.The auto covariance is maximized/most similar for no translation $\tau = 0$ at all.
Definition 20.13 Auto Correlation Is the scaled version of the auto-covariance??: $\rho(t_2 - t_1) = \frac{\text{Cov}[\mathbf{X}_{t_1}, \mathbf{X}_{t_2}]}{\sigma_{X_{t_1}} \sigma_{X_{t_2}}} = \frac{\mathbb{E}[(\mathbf{X}_{t_1} - \mu_{t_1})(\mathbf{X}_{t_2} - \mu_{t_2})]}{\sigma_{X_{t_1}} \sigma_{X_{t_2}}} \quad (20.11)$
3. Different kinds of Processes

3.1. Markov Process Definition 20.14 Markov Process: A continuous-time stochastic process $X(t), t \in T$, is called a Markov process if for any finite parameter set $\{t_i : t_i < t_{i+1}\} \in T$ it holds: $\mathbb{P}(X(t_{n+1}) \in B X(t_1), \dots, X(t_n)) = \mathbb{P}(X(t_{n+1}) \in B X(t_n))$ it thus follows for the <i>transition probability</i> – the probability of $X(t)$ lying in the set B at time t , given the value x of the process at time s : $\mathbb{P}(s, x, t, B) = P(X(t) \in B X(s) = x) \quad 0 \leq s < t \quad (20.12)$
Interpretation In order to predict the future only the current/last value counts.
Corollary 20.3 Transition Density: The transition probability of a continuous distribution \mathbf{p} can be calculated via: $\mathbb{P}(s, x, t, B) = \int_B \mathbf{p}(s, x, t, y) \, dy \quad (20.13)$
3.2. Gaussian Process Definition 20.15 Gaussian Process: Is a stochastic process $X(t)$ where the random variables follow a Gaussian distribution: $X(t) \sim \mathcal{N}(\mu(t), \sigma^2(t)) \quad \forall t \in T \quad (20.14)$
3.3. Diffusions Definition 20.16 Diffusion: Is a Markov Process?? for which it holds that: $\mu(t, X(t)) = \lim_{t \rightarrow 0} \frac{1}{\Delta t} \mathbb{E}[X(t + \Delta t) - X(t) X(t)] \quad (20.15)$ $\sigma^2(t, X(t)) = \lim_{t \rightarrow 0} \frac{1}{\Delta t} \mathbb{E}[(X(t + \Delta t) - X(t))^2 X(t)] \quad (20.16)$ <ul style="list-style-type: none">$\mu(t, X(t))$ is called drift$\sigma^2(t, X(t))$ is called diffusion coefficient
Interpretation There exist not discontinuities for the trajectories.
3.4. Brownian Motion/Wiener Process Definition 20.17 d-dim standard Brownian Motion/Wiener Process: Is an \mathbb{R}^d valued <i>stochastic process</i> ?? $(W_t)_{t \in \mathcal{T}}$ starting at $\mathbf{x}_0 \in \mathbb{R}^d$ that satisfies: <ol style="list-style-type: none">Normal Independent Increments: the increments are <i>normally distributed independent random variables</i>: $W(t_i) - W(t_{i-1}) \sim \mathcal{N}(0, (t_i - t_{i-1}) \mathbf{1}_{d \times d}) \quad \forall i \in \{1, \dots, T\} \quad (20.17)$Stationary increments: $W(t + \Delta t) - W(t)$ is independent of $t \in \mathcal{T}$Continuity: for <i>a.e.</i> $\omega \in \Omega$, the function $t \mapsto W_t(\omega)$ is continuous $\lim_{t \rightarrow 0} \frac{\mathbb{P}(W(t + \Delta t) - W(t) \geq \delta)}{\Delta t} = 0 \quad \forall \delta > 0 \quad (20.18)$Start $W(0) := W_0 = 0 \quad \text{a.s.} \quad (20.19)$
Notation <ul style="list-style-type: none">In many source the Brownian motion is a synonym for the standard Brownian Motion and it is the same as the Wiener process.However in some sources the Wiener process is the standard Brownian Motion, while the Brownian motion denotes a general form $\alpha W(t) + \beta$.

Corollary 20.4 $W_t \sim \mathcal{N}(0, \sigma)$ [proof ??],[proof ??]: The random variable W_t follows the $\mathcal{N}(0, \sigma)$ law $\mathbb{E}[W(t)] = \mu = 0 \quad (20.20)$ $\mathbb{V}[W(t)] = \mathbb{E}[W^2(t)] = \sigma^2 = t \quad (20.21)$
3.4.1. Properties of the Wiener Process Property 20.1 Non-Differentiable Trajectories: The sample paths of a Brownian motion are not differentiable: $\frac{dW(t)}{dt} = \lim_{t \rightarrow 0} \mathbb{E} \left[\left(\frac{W(t + \Delta t) - W(t)}{\Delta t} \right)^2 \right]$ $= \lim_{t \rightarrow 0} \frac{\mathbb{E}[W(t + \Delta t) - W(t)]}{\Delta t} = \lim_{t \rightarrow 0} \frac{\sigma^2}{\Delta t} = \infty$ <i>result</i> cannot use normal calculus anymore <i>solution</i> \rightarrow Ito Calculus see ??.
Property 20.2 Auto covariance Function: The auto-covariance?? for a Wiener process $\mathbb{E}[(W(t) - \mu(t))(W(t') - \mu(t'))] = \min(t, t') \quad (20.22)$
Property 20.3: A standard Brownian motion is a Quadratic Variation
Definition 20.18 Total Variation: The total variation of a function $f : [a, b] \subset \mathbb{R} \mapsto \mathbb{R}$ is defined as: $LV_{[a, b]}(f) = \sup_{\Pi \in S} \sum_{i=0}^{n_{\Pi}-1} f(x_{i+1}) - f(x_i) \quad (20.23)$ $S = \left\{ \Pi \{x_0, \dots, x_{n_{\Pi}}\} : \Pi \text{ is a partition } \stackrel{\text{[def. 13.12]}}{\text{of}} [a, b] \right\}$ it is a measure of the (one dimensional) length of a function w.r.t. to the y-axis, when moving alone the function. Hence it is a measure of the variation of a function w.r.t. to the y-axis.
Definition 20.19 Total Quadratic Variation/“sum of squares”: The total quadratic variation of a function $f : [a, b] \subset \mathbb{R} \mapsto \mathbb{R}$ is defined as: $QV_{[a, b]}(f) = \sup_{\Pi \in S} \sum_{i=0}^{n_{\Pi}-1} f(x_{i+1}) - f(x_i) ^2 \quad (20.24)$ $S = \left\{ \Pi \{x_0, \dots, x_{n_{\Pi}}\} : \Pi \text{ is a partition } \stackrel{\text{[def. 13.12]}}{\text{of}} [a, b] \right\}$
Corollary 20.5 Bounded (quadratic) Variation: The (quadratic) variation?? of a function is bounded if it is finite: $\exists M \in \mathbb{R}_+ : \quad LV_{[a, b]}(f) \leq M \quad \left(QV_{[a, b]}(f) \leq M \right) \quad \forall \Pi \in S \quad (20.25)$
Theorem 20.1 Variation of Wiener Process: Almost surely the total variation of a Brownian motion over an interval $[0, T]$ is infinite: $\mathbb{P}(\omega : LV(W(\omega)) < \infty) = 0 \quad (20.26)$
Theorem 20.2 [proof ??] Quadratic Variation of standard Brownian Motion: The quadratic variation of a standard Brownian motion over $[0, T]$ is finite: $\lim_{N \rightarrow \infty} \sum_{k=1}^N \left[W\left(k \frac{T}{N}\right) - W\left((k-1) \frac{T}{N}\right) \right]^2 = T$ with probability 1 $(dW(t))^2 = dt \quad (20.27)$
Corollary 20.6 : ?? can also be written as: $(dW(t))^2 = dt \quad (20.28)$

3.4.2. Lévy's Characterization of BM

Theorem 20.3 [proof ??],[proof ??]
d-dim standard BM/Wiener Process by Paul Lévy:
An \mathbb{R}^d valued *adapted stochastic process*???? $(W_t)_{t \in \mathcal{T}}$ with the filtration $\{\mathcal{F}_t\}_{t \in \mathbb{R}_+}$, that satisfies:

- ① **Start**
$$W(0) := W_0 = 0 \quad \text{a.s.} \quad (20.29)$$
- ② **Continuous Martingale:** W_t is an a.s. *continuous* martingale?? w.r.t. the filtration $(\mathcal{F}_t)_{t \in \mathcal{T}}$ under \mathbb{P} .
- ③ **Quadratic Variation:**
$$W_t^2 - t \text{ is also an martingale} \iff QV(W_t) = t \quad (20.30)$$

is a standard Brownian motion??.

Further Stochastic Processes

3.4.3. White Noise

understand script and add

Definition 20.20 Discrete-time white noise: Is a random signal $\{\epsilon_t\}_{t \in T_{\text{discret}}}$ having equal intensity at different frequencies and is defined by:

- Having zero tendencies/expectation (otherwise the signal would not be random):
$$\mathbb{E}[\epsilon * [k]] = 0 \quad \forall k \in T_{\text{discret}} \quad (20.31)$$

- Zero autocorrelation[def. 17.12] γ i.e. the signals of different times are in no-way correlated:
$$\gamma(\epsilon * [k], \epsilon * [k + n]) = \mathbb{E}[\epsilon * [k] \epsilon * [k + n]^T] = \mathbb{V}[\epsilon * [k]] \delta_{\text{discret}}[n] \quad \forall k, n \in T_{\text{discret}} \quad (20.32)$$

With
$$\delta_{\text{discret}}[n] := \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{else} \end{cases}$$

See proofs

Definition 20.21 Continuous-time white noise: Is a random signal $(\epsilon_t)_{t \in T_{\text{continuous}}}$ having equal intensity at different frequencies and is defined by:

- Having zero tendencies/expectation (otherwise the signal would not be random):
$$\mathbb{E}[\epsilon * (t)] = 0 \quad \forall t \in T_{\text{continuous}} \quad (20.33)$$

- Zero autocorrelation[def. 17.12] γ i.e. the signals of different times are in no-way correlated:
$$\gamma(\epsilon * (t), \epsilon * (t + \tau)) = \mathbb{E}[\epsilon * (t) \epsilon * (t + \tau)^T] = \mathbb{V}[\epsilon * (t)] \delta(t - \tau) = \begin{cases} \mathbb{V}[\epsilon * (t)] & \text{if } \tau = 0 \\ 0 & \text{else} \end{cases} \quad \forall t, \tau \in T_{\text{continuous}} \quad (20.35)$$

Definition 20.22 Homoscedastic Noise: Has constant variability for all observations/time-steps:

$$\mathbb{V}[\epsilon_{i,t}] = \sigma^2 \quad \forall t = 1, \dots, T \\ \forall i = 1, \dots, N \quad (20.36)$$

Definition 20.23 Heteroscedastic Noise: Is noise whose variability may vary with each observation/time-step:

$$\mathbb{V}[\epsilon_{i,t}] = \sigma(i, t)^2 \quad \forall t = 1, \dots, T \\ \forall i = 1, \dots, N \quad (20.37)$$

3.4.4. Generalized Brownian Motion

Definition 20.24 Brownian Motion:

Let $\{W_t\}_{t \in \mathbb{R}_+}$ be a standard Brownian motion??, and define:

$$X_t = \mu t + \sigma W_t \quad t \in \mathbb{R}_+ \quad \begin{matrix} \mu \in \mathbb{R} & : & \text{drift parameter} \\ \sigma \in \mathbb{R}_+ & : & \text{scale parameter} \end{matrix} \quad (20.38)$$

then $\{X_t\}_{t \in \mathbb{R}_+}$ is normally distributed with mean μt and variance $t\sigma^2$ $X_t \sim \mathcal{N}(\mu t, \sigma^2 t)$.

Theorem 20.4 Normally Distributed Increments:

If $W(t)$ is a Brownian motion, then $W(t) - W(0)$ is a normal random variable with mean μt and variance $\sigma^2 t$, where $\mu, \sigma \in \mathbb{R}$. From this it follows that $W(t)$ is distributed as:

$$f_{W(t)}(x) \sim \mathcal{N}(\mu t, \sigma^2 t) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left\{-\frac{(x - \mu t)^2}{2\sigma^2 t}\right\} \quad (20.39)$$

Corollary 20.7 : More generally we may define the process:
$$t \mapsto f(t) + \sigma W_t \quad (20.40)$$

which corresponds to a noisy version of f .

Corollary 20.8

Brownian Motion as a Solution of an SDE: A stochastic process X_t follows a BM with drift μ and scale σ if it satisfies the following SDE:

$$\begin{aligned} dX(t) &= \mu dt + \sigma dW(t) \\ X(0) &= 0 \end{aligned} \quad (20.41) \quad (20.42)$$

3.4.5. Geometric Brownian Motion (GBM)

For many processes $X(t)$ it holds that:

- there exists an (exponential) growth
- that the values may not be negative $X(t) \in \mathbb{R}_+$

Definition 20.25 Geometric Brownian Motion:

Let $\{W_t\}_{t \in \mathbb{R}_+}$ be a standard Brownian motion?? the stochastic process $\mathbf{S}_t^1 \triangleq \mathbf{S}^1(t)$ with drift parameter μ and scale σ satisfying the SDE:

$$\begin{aligned} d\mathbf{S}_t^1 &= \mathbf{S}_t^1 (\mu dt + \sigma dW_t) \\ &= \mu \mathbf{S}_t^1 dt + \sigma \mathbf{S}_t^1 dW_t \end{aligned} \quad (20.43)$$

is called geometric Brownian motion and is given by:

$$\mathbf{S}_t^1 = \mathbf{S}_0^1 \exp\left(\sigma W_t + \left(\mu - \frac{1}{2}\sigma^2\right)t\right) \quad t \in \mathbb{R}_+ \quad (20.44)$$

Corollary 20.9 Log-normal Returns:

For a geometric BM we obtain log-normal returns:

$$\begin{aligned} \ln\left(\frac{S_t}{S_0}\right) &= \bar{\mu} t + \sigma W(t) \iff \bar{\mu} t + \sigma W(t) \sim \mathcal{N}(\mu t, \sigma^2 t) \\ \text{with} \quad \bar{\mu} &:= \mu - \frac{1}{2}\sigma^2 \end{aligned} \quad (20.45)$$

3.4.6. Locally Brownian Motion

Definition 20.26 Locally Brownian Motion:

Let $\{W_t\}_{t \in \mathbb{R}_+}$ be a standard Brownian motion?? a local Brownian motion is a stochastic process $X(t)$ that satisfies the SDE:

$$dX(t) = \mu(X(t), t) dt + \sigma(X(t), t) dW(t) \quad (20.46)$$

Note

A local Brownian motion is an generalization of a geometric Brownian motion.

3.4.7. Ornstein-Uhlenbeck Process

Definition 20.27 Ornstein-Uhlenbeck Process:

Let $\{W_t\}_{t \in \mathbb{R}_+}$ be a standard Brownian motion?? a Ornstein-Uhlenbeck Process or exponentially correlated noise is a stochastic process $X(t)$ that satisfies the SDE:

$$dX(t) = -aX(t) dt + b\sigma dW(t) \quad a > 0 \quad (20.47)$$

3.5. Poisson Processes

Definition 20.28 Rare/Extreme Events: Are events that lead to discontinuous in stochastic processes.

Problem

A Brownian motion is not sufficient as model in order to describe extreme events s.a. crashes in financial market time series. Need a model that can describe such discontinuities/jumps.

Definition 20.29 Poisson Process: A Poisson Process with *rate* $\lambda \in \mathbb{R}_{\geq 0}$ is a collection of random variables $X(t)$, $t \in [0, \infty)$ defined on a probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$, having a discrete *state space* $N = \{0, 1, 2, \dots\}$ and satisfies:

- $X_0 = 0$
- The increments follow a Poisson distribution??:
$$\mathbb{P}((X_t - X_s) = k) = \frac{\lambda(t-s)}{k!} e^{-\lambda(t-s)} \quad 0 \leq s < t < \infty \quad \forall k \in \mathbb{N}$$
- No correlation of (non-overlapping) increments:
$$\forall t_0 < t_1 < \dots < t_n : \text{the increments are independent} \\ X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \dots, X_{t_n} - X_{t_{n-1}} \quad (20.48)$$

Interpretation

A Poisson Process is a *continuous-time* process with *discrete, positive* realizations in $\mathbb{N}_{\geq 0}$

Corollary 20.10 Probability of events: Using Taylor in order to expand the Poisson distribution one obtains:

$$\mathbb{P}(X_{(t+\Delta t)} - X_t \neq 0) = \lambda \Delta t + o(\Delta t^2) \quad t \text{ small i.e. } t \rightarrow 0 \quad (20.49)$$

- Thus the probability of an event happening during Δt is proportional to time period and the rate λ
- The probability of two or more events to happen *during* Δt is of order $o(\Delta t^2)$ and thus extremely small (as *Deltat* is small).

Definition 20.30 Differential of a Poisson Process: The differential of a Poisson Process is defined as:

$$dX_t = \lim_{\Delta t \rightarrow dt} (X_{(t+\Delta t)} - X_t) \quad (20.50)$$

Property 20.4 Probability of Events for differential: With the definition of the differential and using the previous results from the Taylor expansion it follows:

$$\begin{aligned} \mathbb{P}(dX_t = 0) &= 1 - \lambda \\ \mathbb{P}(|dX_t| = 1) &= \lambda \end{aligned} \quad (20.51) \quad (20.52)$$

Proofs

Proof 20.1: ??:

Let by δ denote the displacement of a particle at each step, and assume that the particles start at the center i.e. $x(0) = 0$, then we have:

$$\begin{aligned} \mathbb{E}[x(n)] &= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N x_i(n)\right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i(n-1) \pm \delta] \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i(n-1)] \\ &\stackrel{\text{induction}}{=} \mathbb{E}[x_{n-1}] = \dots \mathbb{E}[x(0)] = 0 \end{aligned}$$

Thus in expectation the particles goes nowhere.

Proof 20.2: ??:

Let by δ denote the displacement of a particle at each step, and assume that the particles start at the center i.e. $x(0) = 0$, then we have:

$$\begin{aligned} \mathbb{E}[x(n)^2] &= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N x_i(n)^2\right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i(n-1) \pm \delta]^2 \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i(n-1)^2 \pm 2\delta x_i(n-1) + \delta^2] \\ &\stackrel{\text{ind.}}{=} \mathbb{E}[x_{n-1}^2] + \delta^2 = \mathbb{E}[x_{n-2}^2] + 2\delta^2 = \dots \\ &= \mathbb{E}[x(0)] + n\delta^2 = n\delta^2 \\ \text{as } n &= \frac{\text{time}}{\text{step-size}} = \frac{t}{\Delta x} \text{ it follows:} \end{aligned}$$

$$\sigma^2 = \mathbb{E}[x^2(n)] - \mathbb{E}[x(n)]^2 = \mathbb{E}[x^2(n)] = \frac{\delta^2}{\Delta x} t \quad (20.53)$$

Thus in expectation the particles goes nowhere.

Proof 20.3: ??:

$$\begin{aligned} \gamma(\epsilon * [k], \epsilon * [k + n]) &= \text{Cov}[\epsilon * [k], \epsilon * [k + 1]] \\ &= \mathbb{E}[(\epsilon * [k] - \mathbb{E}[\epsilon * [k]]) (\epsilon * [k + n] - \mathbb{E}[\epsilon * [k + n]])^T] \\ &\stackrel{??}{=} \mathbb{E}[(\epsilon * [k]) (\epsilon * [k + n])] \end{aligned}$$

Proof 20.4: ??:

Since $B_t - B_s$ is the increment over the interval $[s, t]$, it is the same in distribution as the incremeent over the interval $[s - s, t - s] = [0, t - s]$

$$\begin{aligned} \text{Thus} \quad B_t - B_s &\sim B_{t-s} - B_0 \\ \text{but as } B_0 &\text{ is a.s. zero by definition ?? it follows:} \\ B_t - B_s &\sim B_{t-s} \quad B_{t-s} \sim \mathcal{N}(0, t - s) \end{aligned}$$

Proof 20.5: ??:

$$\begin{aligned} W(t) &= W(t) - \underbrace{W(0)}_{=0} \sim \mathcal{N}(0, t) \\ \Rightarrow \quad \mathbb{E}[X] &= 0 \quad \mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = t \end{aligned}$$

Proof 20.6: ??:

$$\begin{aligned} \sum_{k=0}^{N-1} [W(t_k) - W(t_{k-1})]^2 & \quad t_k = k \frac{T}{N} \\ &= \sum_{k=0}^{N-1} X_k^2 \quad X_k \sim \mathcal{N}\left(0, \frac{T}{N}\right) \\ &= \sum_{k=0}^{N-1} Y_k = n \left(\frac{1}{n} \sum_{k=0}^{N-1} Y_k\right) \quad \mathbb{E}[Y_k] = \frac{T}{N} \\ \text{S.L.L.N} \quad \frac{T}{n} &= T \end{aligned}$$

Proof 20.7: ?? ②:

- first we need to show ??:
Due to the fact that W_t is \mathcal{F}_t measurable i.e. $W_t \in \mathcal{F}_t$ we know that:

$$\begin{aligned} \mathbb{E}[W_t | \mathcal{F}_t] &= W_t \quad (20.54) \\ \mathbb{E}[W_t | \mathcal{F}_s] &= \mathbb{E}[W_t - W_s + W_s | \mathcal{F}] \\ &= \mathbb{E}[W_t - W_s | \mathcal{F}_s] + \mathbb{E}[W_s | \mathcal{F}_s] \\ &\stackrel{??}{=} \mathbb{E}[W_t - W_s] + W_s \\ W_t - W_s &\stackrel{??}{=} \mathcal{N}(0, t-s) \quad W_s \end{aligned}$$

- second we need to show ??:
$$\mathbb{E}[|W(t)|^2] \stackrel{??}{\leq} \mathbb{E}[|W(t)|^2] = \mathbb{E}[W^2(t)] = t \leq \infty$$

Proof 20.8: ?? ③: $W_t^2 - t$ is a martingale?
Using the binomial formula we can write and adding $W_s - W_s$:
$$W_t^2 = (W_t - W_s)^2 + 2W_s(W_t - W_s) + W_s^2$$

$$\begin{aligned} \text{using the expectation:} \\ \mathbb{E}[W_t^2 | \mathcal{F}_s] &= \mathbb{E}[(W_t - W_s)^2 | \mathcal{F}_s] + \mathbb{E}[2W_s(W_t - W_s) | \mathcal{F}_s] \\ &\quad + \mathbb{E}[W_s^2 | \mathcal{F}_s] \\ &\stackrel{??}{=} \mathbb{E}[(W_t - W_s)^2] + 2W_s \mathbb{E}[(W_t - W_s)] + W_s^2 \\ &\stackrel{??}{=} \mathbb{V}[W_t - W_s] + 0 + W_s^2 \\ &\quad t - s + W_s^2 \end{aligned}$$

from this it follows that:
$$\mathbb{E}[W_t^2 - t | \mathcal{F}_s] = W_s^2 - s \quad (20.55)$$

understand why $\mathbb{E}[(W_t - W_s)^2 | \mathcal{F}] = \mathbb{E}[(W_t - W_s)^2]$

Examples

Example 20.1 :

Suppose we have a sample space of four elements: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. At time zero, we do not have any information about which ω has been chosen. At time $T/2$ we know whether we have $\{\omega_1, \omega_2\}$ or $\{\omega_3, \omega_4\}$. At time T , we have full information.

t

$$\mathcal{F} = \begin{cases} \{\emptyset, \Omega\} & t \in [0, T/2) \\ \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \Omega\} & t \in [T/2, T) \\ \mathcal{F}_{\max} = 2^\Omega & t = T \end{cases} \quad (20.56)$$

Thus, \mathcal{F}_0 represents initial information whereas \mathcal{F}_∞ represents full information (all we will ever know). Hence, a stochastic process is said to be defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$.

Ito Calculus