(gp)11Gaussian Processes (GP)section.1 (gp)@cref(gp)@cref[section][1][]1[1][1][]1

Probabilistic Artificial Intelligence

Gaussian Processes (GP)

1. Gaussian Linear Regression

Math Appendix

Logic

Set Theory

Definition 3.1 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 3.2 Empty Set $\{\}/\varnothing$**:** is the unique set having no elements/cardinality[def. 3.4] zero.

Definition 3.3 Multiset/Bag: Is a set-like object in which multiplicity matters, that is we can have multiple elements of the same type.
I.e. $\{1, 1, 2, 3\} \neq \{1, 2, 3\}$

Definition 3.4 Cardinality $|S|$**:** Is the number of elements that are contained in a set.

Definition 3.5 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 $2^{|S|}$.

Definition 3.6 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.

1. Number Sets

1.1. The Real Numbers $\mathbb{R}$

1.1.1. Intervals

Definition 3.7 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:

$$[a, b] = \{x \in \mathbb{R} \,|\, a \leqslant x \leqslant b\} \tag{3.1}$$

Definition 3.8 Open Interval (a, b)**:** The open interval of a and b is the set of all real numbers that are within a and b:

$$(a, b) = \{x \in \mathbb{R} \,|\, a < x \leqslant <\} \tag{3.2}$$

1.2. The Rational Numbers $\mathbb{Q}$

Example 3.1 Power Set/Cardinality of $S = \{x, y, z\}$**:** The subsets of S are:
$\{\varnothing\}$, $\{x\}$, $\{y\}$, $\{z\}$, $\{x, y\}$, $\{x, z\}$, $\{y, z\}$, $\{x, y, z\}$ and hence the power set of S is $\mathcal{P}(S) = \{\{\varnothing\}, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}\}$ with a cardinality of $|S| = 2^3 = 8$.

Sequences&Series

Definition 4.1 Index Set: Is a set[def. 3.1] 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.

$$f : A \mapsto S \qquad A \in \mathbb{N} \tag{4.1}$$

Definition 4.2 Sequence $(a_n)_{n \in A}$**:** is an by an index set A *enumerated* multiset[def. 3.3] (repetitions are allowed) of objects in which *order does matter*.

Definition 4.3 Series: is an infinite ordered set of terms combined together by addition.

1. Types of Sequences

1.1. Arithmetic Sequence

Definition 4.4 Arithmetic Sequence: Is a sequence where the *difference* between two consecutive terms constant i.e. $(2, 4, 6, 8, 10, 12, \ldots)$.

$$t_n = t_0 + nd \qquad d \text{ :difference between two terms} \tag{4.2}$$

1.2. Geometric Sequence

Definition 4.5 Geometric Sequence: Is a sequence where the *ratio* between two consecutive terms constant i.e. $(2, 4, 8, 16, 32, \ldots)$.

$$t_n = t_0 \cdot r^n \qquad r \text{ :ratio between two terms} \tag{4.3}$$

Calculus and Analysis

Definition 5.1 Quadratic Formula: $ax^2 + bx + c = 0$ or in reduced form:
$x^2 + px + q = 0$ with $p = b/a$ and $q = c/a$

Definition 5.2 Discriminant: $\delta = b^2 - 4ac$

Definition 5.3 Solution to [def. 5.1]:

$$x_{\pm} = \frac{-b \pm \sqrt{\delta}}{2a} \qquad \textbf{or} \qquad x_{\pm} = \frac{1}{2}\left(-p \pm \sqrt{p^2 - 4q}\right)$$

Theorem 5.1
Fist 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:

$$F(X) = \int_a^x f(t)\,\mathrm{d}t \tag{5.1}$$

Then it follows:

$$F'(x) = f(x) \qquad \forall x \in (a, b) \tag{5.2}$$

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 $[a, b]$:

$$\int_a^b f(t)\,\mathrm{d}t = F(b) - F(a) \iff \int_a^x \frac{\partial}{\partial x} F(t)\,\mathrm{d}t = F(x) \tag{5.3}$$

Definition 5.4 Domain of a function $\mathtt{dom}(\cdot)$:
Given a function $f : \mathcal{X} \to \mathcal{Y}$, the set of all possible input values $\mathcal{X}$ is called the domain of f – $\mathtt{dom}(f)$.

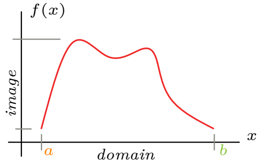
Definition 5.5
Codomain/target set of a function $\mathtt{codom}(\cdot)$:
Given a function $f : \mathcal{X} \to \mathcal{Y}$, the codaomain of that function is the set $\mathcal{Y}$ into which all of the output of the function is **constrained** to fall.

Definition 5.6 Image (Range) of a function: $f[\cdot]$

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

$$\{y \in \mathcal{Y} | y = f(x), \quad \forall x \in \mathcal{X}\} := f[\mathcal{X}] \tag{5.4}$$

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



Definition 5.7 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:

$$f^{-1}(A) = \{x \subseteq X : f(x) \subseteq A\} \tag{5.5}$$

Example 5.1 :
Given $\qquad f : \mathbb{R} \to \mathbb{R}$
defined by $\qquad f : x \mapsto x^2 \iff f(x) = x^2$
$\mathtt{dom}(f) = \mathbb{R}$, $\mathtt{codom}(f) = \mathbb{R}$ **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:

$$f[A] = \{y \in \mathcal{Y} | y = f(x), \quad \forall x \in A\} \tag{5.6}$$

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 5.8 (strictly) Increasing Functions:
A function f is called **monotonically increasing/ increasing/non-decreasing** if:

$$x \leqslant y \iff f(x) \leqslant f(y) \qquad \forall x, y \in \mathtt{dom}(f) \tag{5.7}$$

And **strictly increasing** if:

$$x < y \iff f(x) < f(y) \qquad \forall x, y \in \mathtt{dom}(f) \tag{5.8}$$

Definition 5.9 (strictly) Decreasing Functions:
A function f is called monotonically decreasing/decreasing or non-increasing if:

$$x \geqslant y \iff f(x) \geqslant f(y) \qquad \forall x, y \in \mathtt{dom}(f) \tag{5.9}$$

And *strictly* decreasing if:

$$x > y \iff f(x) > f(y) \qquad \forall x, y \in \mathtt{dom}(f) \tag{5.10}$$

Definition 5.10 Monotonic Function: A function f is called monotonic iff either f is increasing or decreasing.

Definition 5.11 Linear Function:
A function $L : \mathbb{R}^n \mapsto \mathbb{R}^m$ is linear if and only if:

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

Corollary 5.1 Linearity of Differentiation: The derivative of **any** linear combination of functions equals the same linear combination of the derivatives of the functions:

$$\frac{\mathrm{d}}{\mathrm{d}x}(af(x) + bg(x)) = a\frac{\mathrm{d}}{\mathrm{d}x}f(x) + b\frac{\mathrm{d}}{\mathrm{d}x}g(x) \qquad a, b \in \mathbb{R} \tag{5.11}$$

Definition 5.12 Quadratic Function:
A function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is quadratic if it can be written in the form:

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x} + b^{\mathsf{T}}\boldsymbol{x} + c \tag{5.12}$$

1. Continuity and Smoothness

Definition 5.13 Continuous Function:

Definition 5.14 Smoothness of a Function $\mathcal{C}^k$: **Given** a function $f : \mathcal{X} \to \mathcal{Y}$, the function is said to be of class k if it is differentiable up to order k **and** continuous, on its entire domain:

$$f \in \mathcal{C}^k(\mathcal{X}) \iff \exists f', f'', \ldots, f^{(k)} \text{ continuous} \tag{5.13}$$

Note

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

Corollary 5.2 Smooth Function $\mathcal{C}^\infty$: Is a function $f : \mathcal{X} \to \mathcal{Y}$ that has derivatives infinitely many times differerntiable.

$$f \in \mathcal{C}^\infty(\mathcal{X}) \iff f', f'', \ldots, f^{(\infty)} \tag{5.14}$$

Corollary 5.3 Continuously Differentiable Function $\mathcal{C}^1$:
Is the class of functions that consists of all differentiable functions whose derivative is continuous.
Hence a function $f : \mathcal{X} \to \mathcal{Y}$ of the class must satisfy:

$$f \in \mathcal{C}^1(\mathcal{X}) \iff f' \text{ continuous} \tag{5.15}$$

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

Definition 5.15 Lipschitz Continuity: A Lipschitz continuous function is a function f whose rate of change is bound by a Lipschitz Contant L:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \qquad \forall \boldsymbol{x}, \boldsymbol{y}, \quad L > 0 \tag{5.16}$$

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 ⇒ tells us something about robustness.

Definition 5.16 Lipschitz Continuous Gradient:
A *continuously differentiable* function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has *L-Lipschitz continuous gradient* if it satisfies:

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\| \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathtt{dom}(f), \quad L > 0 \tag{5.17}$$

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

$$\nabla^2 f(\boldsymbol{x}) \preccurlyeq L\boldsymbol{I} \qquad \forall \boldsymbol{x} \in \mathtt{dom}(f), \quad L > 0 \tag{5.18}$$

Lemma 5.1 Descent Lemma: If a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has *Lipschitz continuous gradient* eq. (5.17) over its domain, then it holds that:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{y})| \leqslant \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \tag{5.19}$$

Note

If f is twice differentiable then the largest eigenvalue of the Hessian ([def. 6.5]) of f is uniformly upper bounded by L

Proof. lemma 5.1 for $\mathcal{C}^1$ functions:
Let $g(t) \equiv f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))$ from the FToC (theorem 5.2) we know that:

$$\int_0^1 g'(t)\,\mathrm{d}t = g(1) - g(0) = f(\boldsymbol{x}) - f(\boldsymbol{y})$$

It then follows from the reverse:

$$\begin{aligned}
&|f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{y})| \\
\overset{\substack{\text{Chain. R} \\ \text{FToC}}}{=} &\left|\int_0^1 \nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{y})\,\mathrm{d}t - \nabla f(\boldsymbol{y})^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{y})\right| \\
= &\left|\int_0^1 \left(\nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - \nabla f(\boldsymbol{y})\right)^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{y})\,\mathrm{d}t\right| \\
= &\left|\int_0^1 \left(\nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - \nabla f(\boldsymbol{y})\right)^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{y})\,\mathrm{d}t\right| \\
\overset{\text{C.S.}}{\leqslant} &\left|\int_0^1 \|\nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - \nabla f(\boldsymbol{y})\| \cdot \|\boldsymbol{x} - \boldsymbol{y}\|\,\mathrm{d}t\right| \\
\overset{\text{eq. (5.17)}}{=} &\left|\int_0^1 L\|\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}) - \boldsymbol{y}\| \cdot \|\boldsymbol{x} - \boldsymbol{y}\|\,\mathrm{d}t\right| \\
= &\left|L\|\boldsymbol{x} - \boldsymbol{y}\|^2 \int_0^1 t\,\mathrm{d}t\right| = \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2
\end{aligned}$$

□

Proof. lemma 5.1 for $\mathcal{C}^2$ functions:

$$f(\boldsymbol{y}) \overset{\text{Taylor}}{=} f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^{\mathsf{T}}\nabla^2 f(z)(\boldsymbol{y} - \boldsymbol{x})$$

Now we plug in $\nabla^2 f(\boldsymbol{x})$ and recover eq. (5.20):

$$f(\boldsymbol{y}) \leqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^{\mathsf{T}}L(\boldsymbol{y} - \boldsymbol{x})$$

□

Definition 5.17 L-Smoothness: A L-smooth function is a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ that satisfies:

$$f(\boldsymbol{x}) \leqslant f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{y}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$$

with $\qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathtt{dom}(f), \quad L > 0 \quad (5.20)$

If f is a twice differentiable this is equivalent to:

$$\nabla^2 f(\boldsymbol{x}) \preccurlyeq L\boldsymbol{I} \qquad L > 0 \tag{5.21}$$

Theorem 5.3
L-Smoothness of convex functions: A *convex* and L-Smooth function ([def. 5.17]) has a Lipschitz continuous gradient (eq. (5.17)) thus it holds that:

$$f(\boldsymbol{x}) \leqslant f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{y}) \leqslant \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \tag{5.22}$$

Proof. theorem 5.3:
With the definition of convexity for a differentiable function (eq. (5.25)) it follows

$$f(x) - f(y) + \nabla f(y)^{\mathsf{T}}(x - y) \geqslant 0$$
$$\Rightarrow |f(x) - f(y) + \nabla f(y)^{\mathsf{T}}(x - y)|$$
$$\overset{\text{if eq. (5.25)}}{=} f(x) - f(y) + \nabla f(y)^{\mathsf{T}}(x - y)$$

with lemma 5.1 and [def. 5.17] it follows theorem 5.3 □

Corollary 5.4 : L-smoothnes is a weaker condition than L-Lipschitz continuous gradients

2. Convexity

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

Definition 5.18 Convex Functions:
A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if it satisfies:

$$f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda)f(y) \qquad \begin{array}{l}\forall x, y \in \mathtt{dom}(f) \\ \forall \lambda \in [0, 1]\end{array} \tag{5.23}$$

include figure from tikz/convexity

Definition 5.19 Concave Functions:
A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if it satisfies:

$$f(\lambda x + (1 - \lambda)y) \geqslant \lambda f(x) + (1 - \lambda)f(y) \qquad \begin{array}{l}\forall x, y \in \mathtt{dom}(f) \\ \forall \lambda \in [0, 1]\end{array} \tag{5.24}$$

Corollary 5.5 Convexity → global minimima: Convexity implies that all local minima (if they exist) are global minima.

Definition 5.20 Stricly Convex Functions:
A function $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex if it satisfies:

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y) \qquad \begin{array}{l}\forall x, y \in \mathtt{dom}(f) \\ \forall \lambda \in [0, 1]\end{array}$$

add plot

If f is a differentiable function this is equivalent to:

$$f(x) \geqslant f(y) + \nabla f(y)^{\mathsf{T}}(x - y) \qquad \forall x, y \in \mathtt{dom}(f) \tag{5.25}$$

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

$$\nabla^2 f(x) \geqslant 0 \qquad \forall x, y \in \mathtt{dom}(f) \tag{5.26}$$

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. (5.25) implies that $f(x)$ is above the tangent $f(x) + \nabla f(x)^{\mathsf{T}}(y - x)$ for all $x, y \in \mathtt{dom}(f)$
- **??** implies that $f(x)$ is flat or curved upwards

Corollary 5.6 Strict Convexity → Uniqueness:
Strict convexity implies a unique minimizer $\iff$ at most one global minimum.

Corollary 5.7 : A twice differentiable function of one variable $f : \mathbb{R} \to \mathbb{R}$ is convex on an interval $\mathcal{X} = [a, b]$ if and only if its second derivative is non-negative on that interval $\mathcal{X}$:

$$f''(x) \geqslant 0 \qquad \forall x \in \mathcal{X} \tag{5.27}$$

Definition 5.21 μ-Strong Convexity:
Let $\mathcal{X}$ be a Banach space over $\mathbb{K} = \mathbb{R}, \mathbb{C}$. A function $f : \mathcal{X} \to \mathbb{R}$ is called strongly convex iff the following equation holds:

$$f\left(tx + (1-t)y\right) \leqslant tf(x) + (1-t)f(y) - \frac{t(1-t)}{2}\mu\|x-y\| \qquad \forall x, y \in \mathcal{X}, \qquad t \in [0,1], \qquad \mu > 0$$

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

$$f(y) \geqslant f(x) + \nabla f(x)^{\mathsf{T}}(y-x) + \frac{\mu}{2}\|y-x\|_2^2 \tag{5.28}$$

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

$$\nabla^2 f(x) \geqslant \mu \boldsymbol{I} \qquad \forall x, y \in \mathcal{X} \quad \mu > 0 \tag{5.29}$$

Corollary 5.8 Strong Convexity implies Strict Convexity:
https://math.stackexchange.com/questions/2090991/proof-for-strongly-convex-function-is-strictly-convex

Property 5.1:

$$f(\boldsymbol{y}) \leqslant f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^{\mathsf{T}}(\boldsymbol{x}-\boldsymbol{y}) + \frac{1}{2\mu}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \tag{5.30}$$

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. eq. (5.29) analogously to **Proof** eq. (5.21) □

Note

If f is twice differentiable then the smallest eigenvalue of the Hessian ([def. 6.5]) 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. (5.12) is convex if:

$$\nabla^2_{\boldsymbol{x}}\text{eq. } (5.12) = \boldsymbol{A} \geqslant 0 \tag{5.31}$$

Corollary 5.9 :
Strong convexity ⇒ Strict convexity ⇒ Convexity

2.1. Properties that preserve convexity

Property 5.2 Non-negative weighted Sums: Let f be a convex function then $g(x)$ is convex as well:

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

Property 5.3 Composition of Affine Mappings: Let f be a convex function then $g(x)$ is convex as well:

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

Property 5.4 Pointwise Maxima: Let f be a convex function then $g(x)$ is convex as well:

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

Functions

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

$$f(-x) = f(x) \tag{5.32}$$

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

$$f(-x) = -f(x) \tag{5.33}$$

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,$\ldots \exp\left(-x^2/2\right)$.
Odd: $\sin x$, $\tan x$, x, x^3, x^5,$\ldots$.

$$x\text{-Shift:} \qquad f(x-c) \Rightarrow \text{shift to the right}$$
$$f(x+c) \Rightarrow \text{shift to the left} \tag{5.34}$$
$$y\text{-Shift:} \qquad f(x) \pm c \Rightarrow \text{shift up/down} \tag{5.35}$$

Proof. **eq. (5.34)** $f(x_n - c)$ we take the x-value at x_n but take the y-value at $x_o := x_n - c$
⇒ we shift the function to x_n. □

Euler's formula

$$e^{\pm ix} = \cos x \pm i \sin x \tag{5.36}$$

Euler's Identity

$$e^{\pm i} = -1 \tag{5.37}$$

Note

$$e^{n} = 1 \Leftrightarrow n = \mathrm{i}\,2\pi k, \qquad k \in \mathbb{N} \tag{5.38}$$

Corollary 5.10 Every norm is a convex function: By using definition [def. 5.18] and the triangular inequality it follows (with the exception of the L0-norm):

$$\|\lambda x + (1-\lambda)y\| \leqslant \lambda\|x\| + (1-\lambda)\|y\|$$

2.2. Taylor Expansion

Definition 5.22 Taylor Expansion:

$$T_n(x) = \sum_{i=0}^{n} \frac{1}{n!} f^{(i)}(x_0) \cdot (x - x_0)^{(i)} \tag{5.39}$$

$$= f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2}f''(x_0)(x-x_0)^2 + \mathcal{O}(x^3) \tag{5.40}$$

Definition 5.23 Incremental Taylor:
Goal: evaluate $T_n(x)$ (eq. (5.40)) at the point $x_0 + \Delta x$ in order to propagate the function $f(x)$ by $h = \Delta x$:

$$T_n(x_0 \pm h) = \sum_{i=0}^{n} \frac{h^i}{n!} f^{(i)}(x_0) i^{-1} \tag{5.41}$$

$$= f(x_0) \pm h f'(x_0) + \frac{h^2}{2} f''(x_0) \pm f'''(x_0)(h)^3 + \mathcal{O}(h^4)$$

Note

If we chose Δx small enough it is sufficient to look only at the first two terms.

Definition 5.24 Multidimensional Taylor: Suppose $X \in \mathbb{R}^n$ is open, $\boldsymbol{x} \in X$, $f : X \mapsto \mathbb{R}$ and $f \in \mathcal{C}^2$ then it holds that

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^{\mathsf{T}} H(\boldsymbol{x} - \boldsymbol{x}_0) \tag{5.42}$$

Definition 5.25 Argmax: The argmax of a function defined on a set D is given by:

$$\operatorname*{arg\,max}_{x \in D} f(x) = \{x | f(x) \geqslant f(y), \forall y \in D\} \tag{5.43}$$

Definition 5.26 Argmin: The argmin of a function defined on a set D is given by:

$$\operatorname*{arg\,min}_{x \in D} f(x) = \{x | f(x) \leqslant f(y), \forall y \in D\} \tag{5.44}$$

Corollary 5.11 Relationship arg min ↔ arg max**:**

$$\operatorname*{arg\,min}_{x \in D} f(x) = \operatorname*{arg\,max}_{x \in D} -f(x) \tag{5.45}$$

Property 5.5 Argmax Identities:
1. **Shifting**:
$$\forall \lambda \text{ const} \qquad \operatorname{arg\,max} f(x) = \operatorname{arg\,max} f(x) + \lambda \tag{5.46}$$
2. **Positive Scaling**:
$$\forall \lambda > 0 \text{ const} \qquad \operatorname{arg\,max} f(x) = \operatorname{arg\,max} \lambda f(x) \tag{5.47}$$
3. **Negative Scaling**:
$$\forall \lambda < 0 \text{ const} \qquad \operatorname{arg\,max} f(x) = \operatorname{arg\,min} \lambda f(x) \tag{5.48}$$
4. **Positive Functions**:
$$\forall \operatorname{arg\,max} f(x) > 0, \forall x \in \mathtt{dom}(f)$$
$$\operatorname{arg\,max} f(x) = \operatorname{arg\,min} \frac{1}{f(x)} \tag{5.49}$$
5. **Stricly Monotonic Functions**: for all strictly monotonic increasing functions ([def. 5.8]) g it holds that:
$$\operatorname{arg\,max} g(f(x)) = \operatorname{arg\,max} f(x) \tag{5.50}$$

Definition 5.27 Max: The maximum of a function f defined on the set D is given by:

$$\max_{x \in D} f(x) = f(x^*) \qquad \text{with} \qquad \forall x^* \in \operatorname*{arg\,max}_{x \in D} f(x) \tag{5.51}$$

Definition 5.28 Min: The minimum of a function f defined on the set D is given by:

$$\min_{x \in D} f(x) = f(x^*) \qquad \text{with} \qquad \forall x^* \in \operatorname*{arg\,min}_{x \in D} f(x) \tag{5.52}$$

Corollary 5.12 Relationship min ↔ max**:**

$$\min_{x \in D} f(x) = -\max_{x \in D} -f(x) \tag{5.53}$$

Property 5.6 Max Identities:
1. **Shifting**:
$$\forall \lambda \text{ const} \qquad \max\left\{f(x) + \lambda\right\} = \lambda + \max f(x) \tag{5.54}$$
2. **Positive Scaling**:
$$\forall \lambda > 0 \text{ const} \qquad \max \lambda f(x) = \lambda \max f(x) \tag{5.55}$$
3. **Negative Scaling**:
$$\forall \lambda < 0 \text{ const} \qquad \max \lambda f(x) = \lambda \min f(x) \tag{5.56}$$
4. **Positive Functions**:
$$\forall \operatorname{arg\,max} f(x) > 0, \forall x \in \mathtt{dom}(f) \qquad \max \frac{1}{f(x)} = \frac{1}{\min f(x)} \tag{5.57}$$
5. **Stricly Monotonic Functions**: for all strictly monotonic increasing functions ([def. 5.8]) g it holds that:
$$\max g(f(x)) = g\left(\max f(x)\right) \tag{5.58}$$

Definition 5.29 Supremum: The supremum of a function defined on a set D is given by:

$$\sup_{x \in D} f(x) = \{y | y \geqslant f(x), \forall x \in D\} = \min_{y | y \geqslant f(x), \forall x \in D} y \tag{5.59}$$

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

Definition 5.30 Infinmum: The infinmum of a function defined on a set D is given by:

$$\inf_{x \in D} f(x) = \{y | y \leqslant f(x), \forall x \in D\} = \max_{y | y \leqslant f(x), \forall x \in D} y \tag{5.60}$$

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

Corollary 5.13 Relationship sup ↔ inf**:**

$$\in_{x \in D} f(x) = -\sup_{x \in D} -f(x) \tag{5.61}$$

Note

The supremum/infinmum 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 ⇒ there exists no argmax/argmin ⇒ need to bound the functions from above/below ⟺ infinmum/supremum.

Definition 5.31 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.

$$y(t) = f(x(t), t) \xrightarrow[\forall \tau]{\text{time-invariance}} y(t-\tau) = f(x(t-\tau), t) \tag{5.62}$$

Definition 5.32 Inverse Function $g = f^{-1}$:
A function g is the inverse function of the function $f : A \subset \mathbb{R} \to B \subset \mathbb{R}$ if

$$f(g(x)) = x \qquad \forall x \in \mathtt{dom}(g) \tag{5.63}$$

and

$$g(f(u)) = u \qquad \forall u \in \mathtt{dom}(f) \tag{5.64}$$

Property 5.7
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 5.14 Inverse functions and strict monotonicity: If a function f is strictly monotonic [def. 5.10] on its entire domain, then it is one-to-one and therefore has an inverse function.

3. Special Functions

3.1. The Gamma Function

Definition 5.33 The gamma function $\Gamma(\alpha)$: Is extension of the factorial function (**??**) to the real and complex numbers (with a positive real part):

$$\Gamma(z) = \int_0^\infty x^{z-1}\mathrm{e}^{-x}\,\mathrm{d}x \qquad \Re(z) > 0 \tag{5.65}$$

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

Differential Calculus

Definition 6.1 Critical/Stationary Point: Given a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, that is differentiable at a point $\boldsymbol{x}_0$ then it is called a critical point if the functions derivative vanishes at that point:

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

Definition 6.2 Second Derivative $\frac{\partial^2}{\partial x_i \partial x_j}$:

Corollary 6.1 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. 6.1] then it follows that:

- $f''(x) > 0 \iff$ $f'(x+\epsilon) > 0$ slope points uphill; $f'(x-\epsilon) < 0$ slope points downhill; $f(x)$ is a local minimum
- $f''(x) < 0 \iff$ $f'(x+\epsilon) > 0$ slope points downhill; $f'(x-\epsilon) < 0$ slope points uphill; $f(x)$ is a local maximum

$\epsilon > 0$ sufficiently small enough

Definition 6.3 Gradient: Given $f : n \mapsto \mathbb{R}$ its gradient is defined as:

$$\operatorname{grad}_{\boldsymbol{x}}(f) = \nabla_{\boldsymbol{x}} f := \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \tag{6.1}$$

Definition 6.4 Jacobi Matrix: Given a vector valued function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ its derivative/Jacobian is defined as:

$$\boldsymbol{J}(\boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{x}) = \boldsymbol{D}\boldsymbol{f} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}(\boldsymbol{x}) = \frac{\partial(f_1, \dots, f_m)}{\partial(x_1, \dots, x_n)}(\boldsymbol{x}) = \tag{6.2}$$

$$= \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{x}) & \frac{\partial f_1}{\partial x_2}(\boldsymbol{x}) \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{x}) \\ \frac{\partial f_2}{\partial x_1}(\boldsymbol{x}) & \ddots & \frac{\partial f_2}{\partial x_n}(\boldsymbol{x}) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{x}) & \frac{\partial f_m}{\partial x_2}(\boldsymbol{x}) \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{x}) \end{bmatrix}$$

Theorem 6.1
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:

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

Definition 6.5 Hessian Matrix: Given a function $f : \mathbb{R} \mapsto \mathbb{R}^n$ its Hessian$\in \mathbb{R}^{n \times n}$ is defined as:

$$\boldsymbol{H}(\boldsymbol{f})(\boldsymbol{x}) = \boldsymbol{H}_{\boldsymbol{f}}(\boldsymbol{x}) = \boldsymbol{J}(\nabla \boldsymbol{f}(\boldsymbol{x}))^T \tag{6.3}$$

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

and it corresponds to the Jacobian of the Gradient.
Due to the differentiability and theorem 6.1 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 $\{(\lambda_1, \boldsymbol{v}_1), \dots, \lambda_n, \boldsymbol{v}_n\}$.
Not let $\boldsymbol{d}$ be a directional unit vector then the second derivative in that direction is given by:

$$\boldsymbol{d}^\intercal \boldsymbol{H} \boldsymbol{d} \qquad \iff \qquad \boldsymbol{d}^\intercal \sum_{i=1}^{n} \lambda_i \boldsymbol{v}_i \qquad \overset{\text{if } \boldsymbol{d}=\boldsymbol{v}_j}{\iff} \qquad \boldsymbol{d}^\intercal \lambda_j \boldsymbol{v}_j$$

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

Corollary 6.3 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 $\boldsymbol{x}$ [def. 6.1] then it follows that:

- If $\boldsymbol{H}$ is p.d $\iff \forall \lambda_i > 0 \in \boldsymbol{H} \quad \to \quad f(\boldsymbol{x})$ is a local min.
- If $\boldsymbol{H}$ is n.d $\iff \forall \lambda_i < 0 \in \boldsymbol{H} \quad \to \quad f(\boldsymbol{x})$ is a local max.
- If $\exists \lambda_i > 0 \in \boldsymbol{H}$ and $\exists \lambda_i < 0 \in \boldsymbol{H}$ then $\boldsymbol{x}$ is a local maximum in one cross section of f but a local minimum in another
- If $\exists \lambda_i = 0 \in \boldsymbol{H}$ 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 $\boldsymbol{H}$ is positive definite for a minima $\boldsymbol{x}^*$ of a *quadratic* function f then this point must be a global minimum of that function.

Integral Calculus

Theorem 7.1 Important Integral Properties:

Addition
$$\int_a^b f(x)\,\mathrm{d}x = \int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x \tag{7.1}$$

Reflection
$$\int_a^b f(x)\,\mathrm{d}x = -\int_b^a f(x)\,\mathrm{d}x \tag{7.2}$$

Translation
$$\int_a^b f(x)\,\mathrm{d}x \overset{u:=x\pm c}{=} \int_{a\pm c}^{b\pm c} f(x \mp c)\,\mathrm{d}x \tag{7.3}$$

f **Odd**
$$\int_{-a}^{a} f(x)\,\mathrm{d}x = 0 \tag{7.4}$$

f **Even**
$$\int_{-a}^{a} f(x)\,\mathrm{d}x = 2\int_0^a f(x)\,\mathrm{d}x \tag{7.5}$$

Proof. **eqs. (7.4) and (7.5)**

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

□

Linear Algebra

Given a matrix $\boldsymbol{A} \in \mathbb{K}^{m,n}$

Rank: $\mathfrak{rank}(\boldsymbol{A}) = \dim(\mathfrak{R}(\boldsymbol{A})$
of a matrix is the dimension of the vector space generated (or spanned) by its columns/rows.
Span/Linear Hull: $\operatorname{span}(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n) =$
$\{\lambda_1\boldsymbol{v}_1, \lambda_2\boldsymbol{v}_2, \ldots, \lambda_n\boldsymbol{v}_n)\} = \{\boldsymbol{v} \mid \boldsymbol{v} = \sum_{i=1}^{n} \lambda_i\boldsymbol{v}_i), \lambda_i \in \mathbb{R}\}$
Is the set of vectors tha can be expressed as a linear combination of the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$.
Note these vectors may be linearly independent.
Generatring Set: Is the set of vectors which span the $\mathbb{R}^n$ that is: $\operatorname{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m) = \mathbb{R}^n$.
e.g. $(4, 0)^\top, (0, 5)^\top$ span the $\mathbb{R}^n$.
Basis $\mathfrak{B}$: A lin. indep. generating set of the $\mathbb{R}^n$ is called basis of the $\mathbb{R}^n$.
The unit vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ build a standard basis of the $\mathbb{R}^n$
Vector Space
Image/Range: $\mathfrak{R}(\boldsymbol{A}) := \{\boldsymbol{A}x \mid x \in \mathbb{K}^n\} \subset \mathbb{K}^n$
Null-Space/Kernel: $\mathtt{N} := \{z \in \mathbb{K}^n \mid \boldsymbol{A}z = 0\}$
Dimension theorem:

Theorem 8.1 Rank-Nullity theorem: For any $\boldsymbol{A} \in \mathbb{Q}^{m\times n}$
$$n = \dim(\mathtt{N}[\boldsymbol{A}]) + \dim(\mathfrak{R}[\boldsymbol{A}])$$

From orthogonality it follows $x \in \mathfrak{R}(\boldsymbol{A}),\ y \in \mathtt{N}(\boldsymbol{A}) \Rightarrow x^\top y = 0$.

1. Transformations

1.1. Affine Transformations

Definition 8.1 Affine Transfromation/Map:
Let $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ then:
$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{x} + b \tag{8.1}$$
is called an affine transformation of $\boldsymbol{x}$.

2. Eigenvalues and Vectors

Formula 8.1 Eigenvalues of a 2x2 matrix: Given a 2x2-matrix $\boldsymbol{A}$ its eigenvalues can be calculated by:
$$\{\lambda_1, \lambda_2\} \in \frac{\mathtt{tr}(\boldsymbol{A}) \pm \sqrt{\mathtt{tr}(\boldsymbol{A})^2 - 4\det(\boldsymbol{A})}}{2} \tag{8.2}$$
with $\quad \mathtt{tr}(\boldsymbol{A}) = a + d \qquad \det(\boldsymbol{A}) = ad - bc$

3. Special Kind of Vectors

Definition 8.2 Orthogonal Vectors: Let $\mathcal{Y}$ be an inner-product space[def. 8.14]. A set of vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n, \ldots\} \in \mathcal{Y}$ is called *orthogonal* iff:
$$\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = 0 \qquad \forall i \neq \tag{8.3}$$

Definition 8.3 Orthonormal Vectors: Let $\mathcal{Y}$ be an inner-product space[def. 8.14]. A set of vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n, \ldots\} \in \mathcal{Y}$ is called *orthonormal* iff:
$$\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad \forall i, j \tag{8.4}$$

4. Special Kind of Matrices

Definition 8.4 Orthogonal Matrix: A real valued square matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is said to be orthogonal if its row vectors (and respectively its column vectors) build an orthonormal basis:
$$\langle \boldsymbol{q}_{:i}, \boldsymbol{q}_{:j} \rangle = \delta_{ij} \qquad \text{and} \qquad \langle \boldsymbol{q}_{i:}, \boldsymbol{q}_{j:} \rangle = \delta_{ij} \tag{8.5}$$
This is exactly true if the inverse of $\boldsymbol{Q}$ equals its transpose:
$$\boldsymbol{Q}^{-1} = \boldsymbol{Q}^\intercal \qquad \Longleftrightarrow \qquad \boldsymbol{Q}\boldsymbol{Q}^\intercal = \boldsymbol{Q}^\intercal\boldsymbol{Q} = \boldsymbol{I} \tag{8.6}$$

Definition 8.5 Unitary/Hermitian Matrices:
$$\boldsymbol{A} = \boldsymbol{A}^{\mathsf{H}} \tag{8.7}$$

4.1. Properties of Matrices

4.1.1. Eigendecomposition

Definition 8.6 Eigendecomposition $\quad \boldsymbol{A} = \boldsymbol{Q}\Lambda\boldsymbol{Q}^{-1}$:

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

Definition 8.7 Square Root:

5. Spaces and Measures

Definition 8.8 Bilinear Form/Functional:
Is a mapping $a : \mathcal{Y} \times \mathcal{Y} \mapsto F$ on a field of scalars $F \subseteq \mathbb{K}$, $K = \mathbb{R}$ or $\mathbb{C}$ that satisfies:
$$a(\alpha u + \beta v, w) = \alpha a(u, w) + \beta a(v, w)$$
$$a(u, \alpha v + \beta w) = \alpha a(u, v) + \beta a(u, w)$$
$$\forall u, v, w \in \mathcal{Y}, \qquad \forall \alpha, \beta \in \mathbb{K}$$
Thus: a is linear w.r.t. each argument.

Definition 8.9 Symmetric bilinear form: A bilinear form a on $\mathcal{Y}$ is symmetric if and only if:
$$a(u, v) = a(v, u) \qquad \forall u, v \in \mathcal{Y}$$

Definition 8.10 Positive (semi) definite bilinear form: A symmetric bilinear form a on a vector space $\mathcal{Y}$ over a field F is positive defintie if and only if:
$$a(u, u) > 0 \qquad \forall u \in \mathcal{Y}\backslash\{0\} \tag{8.8}$$
$$\text{And positive semidefinte} \iff \geqslant \tag{8.9}$$

Corollary 8.1 Matrix induced Bilinear Form:
For finite dimensional inner product spaces $\mathcal{X} \in \mathbb{K}^n$ any *symmetric* matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ induces a bilinear form:
$$a(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\intercal\boldsymbol{A}\boldsymbol{x}' = (\boldsymbol{A}\boldsymbol{x}')\boldsymbol{x},$$

Definition 8.11 Positive (semi) definite Matrix $\succ$:
A matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is positive defintie if and only if:
$$\boldsymbol{x}^\intercal\boldsymbol{A}\boldsymbol{x} > 0 \iff \boldsymbol{A} \succ \qquad \forall \boldsymbol{x} \in \mathbb{R}^n\backslash\{0\} \tag{8.10}$$
$$\text{And positive semidefinte} \iff \succcurlyeq \tag{8.11}$$

Corollary 8.2
Eigenvalues of positive (semi) definite matrix:
A positive definite matrix is a *symmetric matrix* where every eigenvalue is *strictly* positive and positive semi definite if every eigenvalue is *positive*.
$$\forall \lambda_i \in \operatorname{eigenv}(\boldsymbol{A}) > 0 \tag{8.12}$$
$$\text{And positive semidefinte} \iff \geqslant \tag{8.13}$$

Proof. corollary 8.2 (for real matrices):
Let $\boldsymbol{v}$ be an eigenvector of $\boldsymbol{A}$ then it follows:
$$0 \overset{\text{corollary 8.2}}{<} \boldsymbol{v}^\intercal\boldsymbol{A}\boldsymbol{v} = \boldsymbol{v}^\intercal\lambda\boldsymbol{v} = \|\boldsymbol{v}\|\lambda$$
□

Corollary 8.3 Positive Definiteness and Determinant: The determinant of a positive definite matrix is always positive. **Thus** a positive definite matrix is always *nonsingular*

Definition 8.12 Negative (semi) definite Matrix $\prec$:
A matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is negative defintie if and only if:
$$\boldsymbol{x}^\intercal\boldsymbol{A}\boldsymbol{x} < 0 \iff \boldsymbol{A} \prec 0 \qquad \forall \boldsymbol{x} \in \mathbb{R}^n\backslash\{0\} \tag{8.14}$$
$$\text{And negative semidefinte} \iff \preccurlyeq \tag{8.15}$$

Theorem 8.2 Sylvester's criterion: Let $\boldsymbol{A}$ be *symmetric/Hermitian* matrix and denote by $\boldsymbol{A}^{(k)}$ the $k \times k$ upper left sub-matrix of $\boldsymbol{A}$.
Then it holds that:
- $$\boldsymbol{A} \succ 0 \iff \det\left(\boldsymbol{A}^k\right) > 0 \qquad k = 1, \ldots, n \tag{8.16}$$
- $$\boldsymbol{A} \prec 0 \iff (-1)^k \det\left(\boldsymbol{A}^k\right) > 0 \qquad k = 1, \ldots, n \tag{8.17}$$
- $\boldsymbol{A}$ is indefinite if the first $\det\left(\boldsymbol{A}^k\right)$ that breaks both of the previous patterns is on the wrong side.
- Sylvester's criterion is inconclusive ($\boldsymbol{A}$ can be anything of the previous three) if the first $\det\left(\boldsymbol{A}^k\right)$ that breaks both patterns is 0.

6. Inner Products

Definition 8.13 Inner Product: Let $\mathcal{Y}$ be a vector space over a field $F \in \mathbb{K}$ of scalars. An inner product on $\mathcal{Y}$ is a map:
$$\langle\cdot,\cdot\rangle : \mathcal{Y} \times \mathcal{Y} \mapsto F \subseteq \mathbb{K} \qquad K = \mathbb{R} \text{ or } \mathbb{C} \tag{8.18}$$
that satisfies: $\quad \forall x, y, z \in \mathcal{Y}, \qquad \alpha, \beta \in F$
1. (Conjugate) Stmmetry: $\langle x, y\rangle = \overline{\langle x, y\rangle}$.
2. Linearity in the first argument:
$$\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$$
3. Positve-definiteness:
$$\langle x, x\rangle \geqslant 0 : x = 0 \iff \langle x, x\rangle = 0$$

Definition 8.14 Inner Product Space $(\mathcal{Y}, \langle\cdot,\cdot\rangle_{\mathcal{Y}})$: Let $F \in \mathbb{K}$ be a field of scalars.
An inner product space $\mathcal{Y}$ is a vetor space over a field F together with an an **inner product** $\langle\cdot,\cdot\rangle_{\mathcal{Y}})$.

Corollary 8.4 Inner product$\mapsto$S.p.d. Bilinear Form:
Let $\mathcal{Y}$ be a vector space over a field $F \in \mathbb{K}$ of scalar.
An **inner product** on $\mathcal{Y}$ is a positive definite symmetric bilinear form on $\mathcal{Y}$.

Example: scalar prodct

Let $a(u, v) = u^\intercal\boldsymbol{I}v$ 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 defintion $\langle\boldsymbol{x}, \boldsymbol{x}\rangle \geqslant 0$, whereas bilinear forms must not.

Definition 8.15 Norm $\|\cdot\|_{\mathcal{Y}}$: A norm measures the **size** of its argument.
Formally let $\mathcal{Y}$ be a vector space over a field F, a norm on $\mathcal{Y}$ is a map:
$$\|\cdot\|_{\mathcal{Y}} : \mathcal{Y} \mapsto \mathbb{R}_+ \tag{8.19}$$
that satisfies: $\quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{Y}, \qquad \alpha \in F \subseteq \mathbb{K} \quad K = \mathbb{R}$ or $\mathbb{C}$
1. Definitness: $\|\boldsymbol{x}\|_{\mathcal{Y}} = 0 \iff \boldsymbol{x} = 0$.
2. Homogenity: $\|\alpha\boldsymbol{x}\|_{\mathcal{Y}} = |\alpha|\|\boldsymbol{x}\|_{\mathcal{Y}}$
3. Triangular Inequality: $\|\boldsymbol{x} + \boldsymbol{y}\|_{\mathcal{Y}} \leqslant \|\boldsymbol{x}\|_{\mathcal{Y}} + \|\boldsymbol{y}\|_{\mathcal{Y}}$

Meaning: Triangular Inequality

States that for any triangle, the sum of the lengths of any two sides must be greater than or equal to the length of the remaining side.

Corollary 8.5 Reverse Triangular Inequality:
$$-\|\boldsymbol{x} - \boldsymbol{y}\|_{\mathcal{Y}} \leqslant \|\boldsymbol{x}\|_{\mathcal{Y}} - \|\boldsymbol{y}\|_{\mathcal{Y}} \leqslant \|\boldsymbol{x} - \boldsymbol{y}\|_{\mathcal{Y}}$$
$$\text{resp.} \qquad \Big|\|\boldsymbol{x}\|_{\mathcal{Y}} - \|\boldsymbol{y}\|_{\mathcal{Y}}\Big| \leqslant \|\boldsymbol{x} - \boldsymbol{y}\|_{\mathcal{Y}}$$

Semi-norm

add

Corollary 8.6 Normed vector space: Is a vector space $\mathcal{Y}$ over a field F, on which a norm $\|\cdot\|_{\mathcal{Y}}$ can be defined.

Corollary 8.7 Inner product induced norm $\langle\cdot,\cdot\rangle_{\mathcal{Y}} \rightarrow \|\cdot\|_{\mathcal{Y}}$: Every inner product $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$ induces a norm of the form:
$$\|\boldsymbol{x}\|_{\mathcal{Y}} = \sqrt{\langle\boldsymbol{x}, \boldsymbol{x}\rangle} \qquad \boldsymbol{x} \in \mathcal{Y}$$
Thus We can define function spaces by their associated norm $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ and inner product spaces lead to normed vector spaces and vice versa.

Corollary 8.8 Energy Norm: A *s.p.d.* bilinear form $a : \mathcal{Y} \times \mathcal{Y} \mapsto F$ induces an energy norm:
$$\|\boldsymbol{x}\|_a := (a(\boldsymbol{x}, \boldsymbol{x}))^{\frac{1}{2}} = \sqrt{a(\boldsymbol{x}, \boldsymbol{x})} \qquad \boldsymbol{x} \in \mathcal{Y}$$

Definition 8.16 Distance Function/Measure: Is measuring the **distance** between two things.
Formally: on a set S is a mapping:
$$d(\cdot,\cdot) : S \times S \mapsto \mathbb{R}_+$$
that satisfies: $\quad \forall x, y, z \in S$
1. ?: $d(x, x) = 0$
2. Symmetry: $d(x, y) = d(y, x)$
3. Triangular Identiy: $d(x, z) \leqslant d(x, y) + d(y, z)$

Definition 8.17 Metric: Is a distance measure that additonally satisfies: $\quad \forall x, y \in S$
identity of indiscernibles : $\quad d(x, y) = 0 \iff x = y$

Corollary 8.9 Metric→Norm: Every norm $\|\cdot\|_{\mathcal{Y}}$ on a vector space $\mathcal{Y}$ over a field F induces a metric by:
$$d(x, y) = \|x - y\|_{\mathcal{Y}} \qquad \forall x, y \in \mathcal{Y}$$
metric induced by norms additionally satisfy: $\forall x, y \in \mathcal{Y}, \quad \alpha \in F \subseteq \mathbb{K} \quad K = \mathbb{R}$ or $\mathbb{C}$
1. Homogenity/Scaling: $d(\alpha x, \alpha y)_{\mathcal{Y}} = |\alpha| d(x, y)_{\mathcal{Y}}$
2. 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 $\mathcal{Y}$ satisfies the properties then it induces a norm of the form:
$$\|\boldsymbol{x}\|_{\mathcal{Y}} := d(\boldsymbol{x}, 0)_{\mathcal{Y}}$$

Note

Similarity measure is a much weaker notion than a metric as triangular inequality does not have to hold.
Hence: If a is similar to b and b is similar to c it does not imply that a is similar to c.

Note

(bilinear form$\xrightarrow{\text{induces}}$)
inner product$\xrightarrow{\text{induces}}$norm$\xrightarrow{\text{induces}}$metric.

7. Vector Algebra

7.1. Planes

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

8. Derivatives

$$\frac{\partial}{\partial\mathbf{x}}(\mathbf{b}^\top\mathbf{x}) = \frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{b}) = \mathbf{b}$$
$$\frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{x}) = 2\mathbf{x}$$
$$\frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{A}\mathbf{x}) = (\mathbf{A}^\top + \mathbf{A})\mathbf{x} \qquad \frac{\partial}{\partial\mathbf{x}}(\mathbf{b}^\top\mathbf{A}\mathbf{x}) = \mathbf{A}^\top\mathbf{b}$$
$$\frac{\partial}{\partial\mathbf{X}}(\mathbf{c}^\top\mathbf{X}\mathbf{b}) = \mathbf{c}\mathbf{b}^\top \qquad \frac{\partial}{\partial\mathbf{x}}(\|\mathbf{x} - \mathbf{b}\|_2) = \frac{\mathbf{x}-\mathbf{b}}{\|\mathbf{x}-\mathbf{b}\|_2}$$
$$\frac{\partial}{\partial\mathbf{x}}(\|\mathbf{x}\|_2^2) = \frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{x}) = 2\mathbf{x} \qquad \frac{\partial}{\partial\mathbf{X}}(\|\mathbf{X}\|_F^2) = 2\mathbf{X}$$
$$\frac{\partial}{\partial\mathbf{x}}||\mathbf{x}||_1 = \frac{\mathbf{x}}{|\mathbf{x}|}$$
$$\frac{\partial}{\partial\mathbf{x}}(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2) = \mathbf{2}(\mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b}) \qquad \frac{\partial}{\partial\mathbf{X}}(|\mathbf{X}|) = |\mathbf{X}| \cdot \mathbf{X}^{-1}$$
$$\frac{\partial}{\partial x}(\mathbf{Y}^{-1}) = -\mathbf{Y}^{-1}\frac{\partial\mathbf{Y}}{\partial x}\mathbf{Y}^{-1}$$

Geometry

Corollary 9.1 Affine Transformation in 1D: Given: numbers $x \in \hat{\Omega}$ with $\hat{\Omega} = [a, b]$
The affine transformation of $\phi : \hat{\Omega} \to \Omega$ with $y \in \Omega = [c, d]$ is defined by:

$$y = \phi(x) = \frac{d-c}{b-a}(x-a) + c \tag{9.1}$$

Proof. **corollary 9.1** By [def. 8.1] we want a function $f : [a, b] \to [c, d]$ that satisfies:

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

additionally $f(x)$ has to be a linear function ([def. 5.11]), that is the output scales the same way as the input scales.
Thus it follows:

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

□

Trigonometry

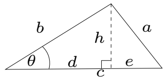
Law 9.1 Law of Cosine: relates the side of a triangle to the cosine of its angles.

$$a^2 = b^2 + c^2 - 2bc\cos\theta_{b,c} \tag{9.2}$$

More general for vectors it holds:

$$\|\boldsymbol{x} - \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\|\boldsymbol{x}\|\|\boldsymbol{y}\|\cos\theta_{\boldsymbol{x},\boldsymbol{y}} \tag{9.3}$$

Proof. eq. (9.2):
We know: $\sin\theta = \frac{h}{b} \Rightarrow \underline{h}$ and $\cos\theta = \frac{d}{b} \Rightarrow d$
Thus $\underline{e} = c - d = c - b\cos\theta \Rightarrow a^2 = \underline{e}^2 + \underline{h}^2 \Rightarrow a$ □



Proof. eq. (9.3):

$$\begin{aligned}\|\boldsymbol{x} - \boldsymbol{y}\|^2 &= (\boldsymbol{x} - \boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{y}) \\ &= \boldsymbol{x} \cdot \boldsymbol{x} - 2\boldsymbol{x} \cdot \boldsymbol{y} + \boldsymbol{y} \cdot \boldsymbol{y} \\ &= \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\left(\|\boldsymbol{x}\|\|\boldsymbol{y}\|\cos\theta\right)\end{aligned}$$

□

Law 9.2 Pythagorean theorem: special case of **??** for right triangle:

$$a^2 = b^2 + c^2 \tag{9.4}$$

Formula 9.1 Euler's Formula:

$$e^{\pm ix} = \cos x \pm i \sin x \tag{9.5}$$

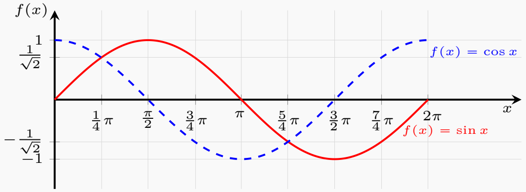
Formula 9.2 Euler's Identity:

$$e^{\pm i} = -1 \tag{9.6}$$

Note

$$e^{n} = 1 \Leftrightarrow n = \mathrm{i}\, 2\pi k, \qquad k \in \mathbb{N} \tag{9.7}$$

Sine and Cosine



$$\cos x \overset{(5.36)}{=} \frac{1}{2}\left[e^{ix} + e^{-ix}\right] \tag{9.8}$$

$$\sin x \overset{(5.36)}{=} \frac{1}{2i}\left[e^{ix} - e^{-ix}\right] = -\frac{i}{2}\left[e^{ix} - e^{-ix}\right] \tag{9.9}$$

Sinh and Cosh

$$\cosh x \overset{(5.36)}{=} \frac{1}{2}\left[e^{x} + e^{-x}\right] = \cos(\mathrm{i}\, x) \tag{9.10}$$

$$\sinh x \overset{(5.36)}{=} \frac{1}{2}\left[e^{x} - e^{-x}\right] = -\mathrm{i}\sin(\mathrm{i}\, x) \tag{9.11}$$

Note

$$e^{x} = \cosh x + \sinh x \qquad e^{-x} = \cosh x - \sinh x \tag{9.12}$$

Note

- $\cosh x$ is strictly positive.
- $\sinh x = 0$ has a unique root at $x = 0$.

Theorem 9.1 Addition Theorems:

$$\sin(\alpha \pm \beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta \tag{9.13}$$

$$\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta \tag{9.14}$$

Werner Formulas

$$\sin\alpha\cos\beta = \frac{1}{2}\left[\sin(\alpha+\beta) + \sin(\alpha-\beta)\right] \tag{9.15}$$

$$\sin\alpha\sin\beta = \frac{1}{2}\left[\cos(\alpha-\beta) - \cos(\alpha+\beta)\right] \tag{9.16}$$

$$\cos\alpha\cos\beta = \frac{1}{2}\left[\cos(\alpha+\beta) + \cos(\alpha-\beta)\right] \tag{9.17}$$

Note

Using theorem 9.1 if follows:

$$\cos(\alpha \pm \pi) = -\cos\alpha \qquad \text{and} \qquad \sin(\alpha \pm \pi) = -\sin\alpha \tag{9.18}$$

Topology

Numerics

1. Machine Arithmetic's

1.1. Machine Numbers

Definition 11.1 Institute of Electrical and Electronics Engineers (IEEE): Is a engineering associations that defines a standard on how computers should treat machine numbers in order to have certain guarantees.

Definition 11.2 Machine/Floating Point Numbers $\mathbb{F}$: Computers are only capable to represent a *finite, discrete* set of the real numbers $\mathbb{F} \subset \mathbb{R}$

1.1.1. Floating Point Arithmetic's $x\tilde{\Omega}y = \mathbf{fl}(x\Omega y)$

Corollary 11.1 Closure: Machine numbers $\mathbb{F}$ are not *closed*[def. 3.6] under basic arithmetic operations:

$$\mathbb{F}\,\Omega\,\mathbb{F} \mapsto \not{\mathbb{F}} \qquad \Omega = \{+,-,*,/\} \tag{11.1}$$

Note

Corollary 11.1 provides a problem as the computer can only represent floating point number $\mathbb{F}$.

Definition 11.3 Floating Point Operation $\tilde{\Omega}$: Is a basic arithmetic operation that obtains a number $x \in \mathbb{F}$ by applying a function rd:

$$\mathbb{F}\,\tilde{\Omega}\,\mathbb{F} \mapsto \mathbb{F} \qquad \begin{aligned}\tilde{\Omega} &:= \mathrm{rd} \circ \Omega \\ \Omega &= \{+,-,*,/\}\end{aligned} \tag{11.2}$$

Definition 11.4 Rounding Function rd: Given a real number $x \in \mathbb{R}$ the rounding function replaces it by the nearest machine number $\tilde{x} \in \mathbb{F}$. If this is ambiguous (there are two possibilities), then it takes the larger one:

$$\mathrm{rd}: \begin{cases} \mathbb{R} \mapsto \mathbb{F} \\ x \mapsto \max \arg\min_{\tilde{x}\in\mathbb{F}} |x - \tilde{x}| \end{cases} \tag{11.3}$$

Consequence

Basic arithmetic rules such as associativity do no longer hold for operations such as addition and subtraction.

Axiom 11.1 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:

$$\begin{aligned} x\tilde{\Omega}y &= (x\Omega y)(1+\delta) && \Omega = \{+,-,*,/\} \\ \tilde{f}(x) &= f(x)(1+\delta) && f \in \{\exp, \sin, \cos, \log, \ldots\} \end{aligned} \tag{11.4}$$

with $|\delta| < \text{EPS}$

Explanation 11.1 (axiom 11.1). *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 $x\Omega y$.*

Finish

Definition 11.5 Overflow: Result is bigger then the biggest representable floating point number.

Definition 11.6 Underflow: Result is smaller then the smaller representable floating point number i.e. to close to zero.

1.2. Roundoff Errors

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. 11.7] 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 11.7 Log sum Exponential:

$$\mathrm{LogSumExp}(x_1, \ldots, x_n) := \log\left(\sum_{i=1}^n \mathrm{e}^{x_i}\right) \tag{11.5}$$

Formula 11.1 Log-Sum-Exp Trick:

$$\log\left(\sum_{i=1}^n \mathrm{e}^{x_i}\right) = a + \log\sum_{i=1}^n \mathrm{e}^{x_i - a} \qquad a := \max_{i\in\{1,\ldots,n\}} x_i \tag{11.6}$$

Explanation 11.2 (formula 11.1). *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.*

Proof.

$$\begin{aligned} \mathrm{LSE} &= \log\Big(\sum_{i=1}^n e^{x_i}\Big) = \log\Big(\sum_{i=1}^n e^{x_i - a} e^{a}\Big) \\ &= \log\Big(e^{a} \sum_{i=1}^n e^{x_i - a}\Big) = \log\Big(\sum_{i=1}^n e^{x_i - a}\Big) + \log(e^{a}) \\ &= \log\Big(\sum_{i=1}^n e^{x_i - a}\Big) + a \end{aligned}$$

□

Definition 11.8 Partition Π: Given an interval $[0, T]$ a sequence of values $0 < t_0 < \cdots < t_n < T$ is called a partition $\Pi(t_0, \ldots, t_n)$ of this interval.

1.3. Convention for iterative methods

Definition 11.9 Linear/Exponential Convergence: A sequence $\{\boldsymbol{x}^{(k)}\}_k \in \mathbb{R}^n$ converges linearly to $\boldsymbol{x}^*$ if in the asymptotic limit $k \to \infty$ it satisfies:

$$\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\right\| \leqslant \rho \left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\| \qquad \rho \in (0,1), \forall k \in \mathbb{N}_0 \tag{11.7}$$

Exponetial Convergence

Linear convergence is sometimes called exponential convergence. This is due to the fact that:

1. We often have expressions of the form:
$$\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\right\| \leqslant \underbrace{(1-\alpha)}_{:=\rho} \left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\|$$
2. and that $(1-\alpha) = \exp(-\alpha)$ from which follows that:
$$\text{eq. (11.8)} \iff \left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\right\| \leqslant \mathrm{e}^{-\alpha} \left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\|$$

Definition 11.10 Rate of Convergence: Is a way to measure the rate of convergence of a sequence $\{\boldsymbol{x}^{(k)}\}_k \in \mathbb{R}^n$ to a value to $\boldsymbol{x}^*$. Let $\rho \in [0,1]$ be the *rate of convergence* and define:

$$\lim_{k\mapsto\infty} \frac{\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\right\|}{\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\|} = \rho \tag{11.8}$$

- $\rho = 1 \iff$ Sublinear Rate i.e. slower than linear
- $\rho =\in (0,1) \iff$ Linear Rate
- $\rho = 0 \iff$ Superlinear Rate i.e. faster then linear

Definition 11.11 Convergence of order p: In order to distinguish *superlinear convergence* we define the order of convergence.
A sequence $\{\boldsymbol{x}^{(k)}\}_k \in \mathbb{R}^n$ converges superlinear with order $p \in \{2, \ldots\}$ to $\boldsymbol{x}^*$ if it satisfies:

$$\lim_{k\mapsto\infty} \frac{\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\right\|}{\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\|^p} = C \qquad C < 1 \tag{11.9}$$

Does this even exist/check if this is ture

Definition 11.12 Exponential Convergence: A sequence $\{\boldsymbol{x}^{(k)}\}_k \in \mathbb{R}^n$ converges exponentially with rate ρ to $\boldsymbol{x}^*$ if in the asymptotic limit $k \to \infty$ it satisfies:

$$\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\right\| \leqslant \rho^k \left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\| \qquad \rho < 1 \tag{11.10}$$

1.4. Convention for discritization methods

2. Numerical Quadrature

Definition 11.13 Order of a Quadrature Rule: The order of a quadrature rule $\mathcal{Q}_n : \mathcal{C}^0([a,b]) \to \mathbb{R}$ is defined as:

$$\mathrm{order}(\mathcal{Q}_n) := \max\left\{n \in \mathbb{N}_0 : \mathcal{Q}_n(p) = \in_a^b p(t)\,\mathrm{d}t \quad \forall p \in \mathcal{P}_n\right\} + 1 \tag{11.11}$$

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.

2.1. Composite Quadrature

Definition 11.14 Composite Quadrature: **Given** a mesh $\mathcal{M} = \{a = x_0 < x_1 < \ldots < x_m = b\}$ apply a Q.R. $\mathcal{Q}_n$ to each of the mesh cells $I_j := [x_{j-1}, x_j] \quad \forall j = 1, \ldots, m \mathrel{\hat{=}}$ p.w. Quadrature:

$$\int_a^b f(t)\,\mathrm{d}t = \sum_{j=1}^m \int_{x_{j-1}}^{x_j} f(t)\,\mathrm{d}t = \sum_{j=1}^m \mathcal{Q}_n(f_{I_j}) \tag{11.12}$$

Lemma 11.1 Error of Composite quadrature Rules: **Given** a function $f \in \mathcal{C}^k([a,b])$ with integration domain:

$$\sum_{i=1}^m h_i = |b - a| \qquad \text{for } \mathcal{M} = \{x_j\}_{j=1}^m$$

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:

$$\begin{aligned} \epsilon_n(f) &= \mathcal{O}\left(n^{-\min\{k,q\}}\right) = \mathcal{O}\left(h_{\mathcal{M}}^{\min\{k,q\}}\right) && \text{for } n \to \infty \\ &\overset{\text{corollary 5.2}}{=} \mathcal{O}\left(n^{-q}\right) = \mathcal{O}\left(h_{\mathcal{M}}^{q}\right) && \text{with } h_{\mathcal{M}} = \frac{1}{n} \end{aligned} \tag{11.13}$$

Definition 11.15 Complexity W: Is the number of function evaluations $\mathrel{\hat{=}}$ number of quadrature points.

$$W(\mathcal{Q}(f)_n) = \#\text{f-eval} \mathrel{\hat{=}} n \tag{11.14}$$

Lemma 11.2 Error-Complexity $W(\epsilon_n(f))$: Relates the complexity to the quadrature error.
Assuming and quadrature error of the form :

$$\epsilon_n(f) = \mathcal{O}(n^{-q}) \iff \epsilon_n(f) = cn^{-q} \qquad c \in \mathbb{R}_+$$

the error complexity is algebraic (**??**) and is given by:

$$W(\epsilon_n(f)) = \mathcal{O}(\epsilon_n^{1/q}) = \mathcal{O}\left(\sqrt[q]{\epsilon_n}\right) \tag{11.15}$$

Proof. lemma 11.2: **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?

$$\frac{c \cdot n_n^q}{c \cdot n_o^q} = \frac{1}{\epsilon_n} \quad \Rightarrow \quad n_n = n_o \cdot \sqrt[q]{\epsilon_n} = \mathcal{O}(\sqrt[q]{\epsilon_n}) \tag{11.16}$$

□

Optimization

Definition 12.1 Fist Order Method: A first-order method is an algorithm that chooses the k-th iterate in

$$\boldsymbol{x}_0 + \mathrm{span}\{\nabla f(\boldsymbol{x}_0), \ldots \nabla f(\boldsymbol{x}_{k-1})\} \qquad \forall k = 1, 2, \ldots \tag{12.1}$$

Note

Gradient descent is a first order method

1. Lagrangian Optimization Theory

Add: derivation of lagragne function

Definition 12.2 (Primal) Constraint Optimization: **Given** an optimization problem with domain $\Omega \subseteq \mathbb{R}^d$:

$$\begin{aligned} &\min_{\boldsymbol{w}\in\Omega} f(\boldsymbol{w}) \\ \textbf{s.t.} \quad & g_i(\boldsymbol{w}) \leqslant 0 && 1 \leqslant i \leqslant k \\ & h_j(\boldsymbol{w}) = 0 && 1 \leqslant j \leqslant m \end{aligned}$$

Definition 12.3 Lagrange Function:

$$\mathscr{L}(\alpha, \beta, \boldsymbol{w}) := f(\boldsymbol{w}) + \alpha \boldsymbol{g}(\boldsymbol{w}) + \beta \boldsymbol{h}(\boldsymbol{w}) \tag{12.2}$$

Extremal Conditions

$$\begin{aligned} &\nabla \mathscr{L}(\boldsymbol{x}) \overset{!}{=} 0 && \text{Extremal point } \boldsymbol{x}^* \\ &\frac{\partial}{\partial\beta} \mathscr{L}(\boldsymbol{x}) = h(\boldsymbol{x}) \overset{!}{=} 0 && \text{Constraint satififsfaction} \end{aligned}$$

For the inequality constraints $g(\boldsymbol{x}) \leqslant 0$ we distinguish two situations:

Case I : $g(\boldsymbol{x}^*) < 0$ switch const. off

Case II : $g(\boldsymbol{x}^*) \geqslant 0$ optimze using active eq. constr.

$$\frac{\partial}{\partial\alpha} \mathscr{L}(\boldsymbol{x}) = g(\boldsymbol{x}) \overset{!}{=} 0 \qquad \text{Constraint satififsfaction}$$

Definition 12.4 Lagrangian Dual Problem: Is given by:

$$\begin{aligned} \text{Find} \quad & \max \theta(\alpha, \beta) = \inf_{\boldsymbol{w}\in\Omega} \mathscr{L}(\boldsymbol{w}, \alpha, \beta) \\ \textbf{s.t.} \quad & \alpha_i \geqslant 0 \qquad\qquad 1 \leqslant i \leqslant k \end{aligned}$$

Solution Strategy

1. Find the extremal point $\boldsymbol{w}^*$ of $\mathscr{L}(\boldsymbol{w}, \alpha, \beta)$:
$$\left.\frac{\partial \mathscr{L}}{\partial \boldsymbol{w}}\right|_{\boldsymbol{w}=\boldsymbol{w}^*} \overset{!}{=} 0 \tag{12.3}$$
2. Insert $\boldsymbol{w}^*$ into $\mathscr{L}$ and find the extremal point β^* of the resulting dual Lagrangian $\theta(\alpha, \beta)$ for the active constraints:
$$\left.\frac{\partial \theta}{\partial \beta}\right|_{\beta=\beta^*} \overset{!}{=} 0 \tag{12.4}$$
3. Calculate the solution $\boldsymbol{w}^*(\beta^*)$ 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 12.1 Upper Bound Dual Cost: Let $\boldsymbol{w} \in \Omega$ be a feasible solution of the primal problem [def. 12.2] and (α, β) a feasible solution of the respective dual problem [def. 12.4]. Then it holds that:

$$f(\boldsymbol{w}) \geqslant \theta(\alpha, \beta) \tag{12.5}$$

Proof.

$$\begin{aligned} \theta(\alpha, \beta) &= \inf_{\boldsymbol{u}\in\Omega} \mathscr{L}(\boldsymbol{u}, \alpha, \beta) \leqslant \mathscr{L}(\boldsymbol{w}, \alpha, \beta) \\ &= f(\boldsymbol{w}) + \sum_{i=1}^k \underbrace{\alpha_i}_{\geqslant 0} \underbrace{g_i(\boldsymbol{w})}_{\leqslant 0} + \sum_{j=}^m \beta_j \underbrace{h_j(\boldsymbol{w})}_{=0} \\ &\leqslant f(\boldsymbol{w}) \end{aligned}$$

□

Corollary 12.1 Duality Gap Corollary: The value of the dual problem is upper bounded by the value of the primal problem:

$$\sup\{\theta(\alpha, \beta) : \alpha \geqslant 0\} \leqslant \inf\{f(\boldsymbol{w}) : \boldsymbol{g}(\boldsymbol{w}) \leqslant 0, \boldsymbol{h}(\boldsymbol{w}) = 0\} \tag{12.6}$$

Theorem 12.2 Optimality: The triple $(\boldsymbol{w}^*, \alpha^*, \beta^*)$ 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:

$$f(\boldsymbol{w}^*) = \theta(\alpha^*, \beta^*) \tag{12.7}$$

Definition 12.5 Convex Optimization: Given: a convex function f and a convex set S solve:

$$\begin{aligned} &\min f(\boldsymbol{x}) \\ &\text{s.t.} \quad \boldsymbol{x} \in S \end{aligned} \tag{12.8}$$

Often S is specified using linear inequalities:

$$\text{e.g.} \qquad S = \left\{ \boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{A}\boldsymbol{x} \leqslant b \right\}$$

Theorem 12.3 Strong Duality: Given an convex optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{w} \in \Omega} f(\boldsymbol{w}) \\ \textbf{s.t.} \qquad &g_i(\boldsymbol{w}) \leqslant 0 \qquad 1 \leqslant i \leqslant k \\ &h_j(\boldsymbol{w}) = 0 \qquad 1 \leqslant j \leqslant m \end{aligned}$$

where g_i, h_i can be written as affine functions: $y(\boldsymbol{w}) = \boldsymbol{A}\boldsymbol{w} - b$.

Then it holds that the duality gap is zero and we obtain an optimal solution.

Theorem 12.4 Kuhn-Tucker Conditions: Given an optimization problem with convex domain $\Omega \subseteq \mathbb{R}^d$,

$$\begin{aligned} &\min_{\boldsymbol{w} \in \Omega} f(\boldsymbol{w}) \\ \textbf{s.t.} \qquad &g_i(\boldsymbol{w}) \leqslant 0 \qquad 1 \leqslant i \leqslant k \\ &h_j(\boldsymbol{w}) = 0 \qquad 1 \leqslant j \leqslant m \end{aligned}$$

with $f \in C^1$ convex and g_i, h_i affine.
Necessary and sufficient conditions for a normal point $\boldsymbol{w}^*$ to be an optimum are the existence of α^*, β^* s.t.:

$$\frac{\partial \mathscr{L}(\boldsymbol{w}, \alpha, \beta)}{\partial \boldsymbol{w}} \overset{!}{=} 0 \qquad \frac{\partial \mathscr{L}(\boldsymbol{w}^*, \alpha, \beta)}{\partial \beta} \overset{!}{=} 0 \tag{12.9}$$

under the condtions that:

- $\forall i_1, \ldots, k \qquad \alpha_i^* g_i(\boldsymbol{w}^*) = 0$, s.t.:
  - Inactive Constraint: $g_i(\boldsymbol{w}^*) < 0 \rightarrow \alpha_i = 0$.
  - Active Constraint:
    $g_i(\boldsymbol{w}^*) \nless 0 \rightarrow \alpha_i \geqslant 0 \quad \text{s.t.} \quad \alpha_i^* g_i(\boldsymbol{w}^*) = 0$

Consequence

We may become very sparce problems, if a lot of constraints are not actice $\iff$ $\alpha_i = 0$.
Only a few points, for which $\alpha_i > 0$ may affact the decision surface.

Stochastics

Definition 12.6 Stochastics: Is a collective term for the areas of *probability theory* and *statistics.*

Definition 12.7 Statistics: Is concerned with the analysis of data/experiments in order to draw conclusion of the underlying governing models that describe these experiments.

Definition 12.8 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 12.9 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 quorahho

Note: Stochastics vs. Stochastic

Stochasticss is a noun and is a collective term for the areas of probability theory and statistics, while stochastic is a *adjective*, describing that a certain phenomena is governed by uncertainty i.e. a process.

Probability Theory

Definition 13.1 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 13.2 Sample Space Ω: Is the set of all possible outcomes (elementary events corollary 13.5) of an experiment see example 13.1

Definition 13.3 Event A**:**
An "event" is a subset of the sample space Ω and is a property which can be observed to hold or not to hold *after* the experiment is done (example 13.2).
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 13.1 : If the outcome ω of an experiment is in the subset A, then the event A is said to "have occured".

Corollary 13.2 Complement Set A^{C}**:**
is the contrary event of A.

Corollary 13.3 The Union Set $A \cup B$**:**
Let A, B be to evenest. The event "A or B" is interpreted as the union of both.

Corollary 13.4 The Intersection Set $A \cap B$**:**
Let A, B be to evenest. The event "A and B" is interpreted as the intersection of both.

Corollary 13.5 The Elementary Event ω**:**
Is a "singleton", i.e. a subset $\{\omega\}$ containing a single outcome ω of Ω.

Corollary 13.6 The Sure Event Ω**:**
Is equal to the sample space as it contains all possible elementary events.

Corollary 13.7 The Impossible Event $\varnothing$**:**
The impossible event i.e. nothing is happening is denoted by the empty set.

Definition 13.4 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 13.5 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. $0 \leqslant \mathbb{P}(A) \leqslant 1$
2. $\mathbb{P}(\Omega) = 1$
3. If $A \cap B = \varnothing$ 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:

$$\mathbb{P}(A) = \lim_{n\to\infty} \frac{\delta(A)}{n} \qquad \text{where} \qquad \delta(A) = \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ if } \omega \notin A \end{cases}$$

0.1. Sigma Algebras

Definition 13.6 Sigma Algebra 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 $\varnothing \in \mathcal{F}$
- If $A \in \mathcal{F}$ then $\Omega \backslash A = A^{\mathrm{C}} \in \mathcal{F}$:
  The complementary subset of A is also in Ω.
- For all $A_i \in \mathcal{F} : \bigcup_{i=1} A_i \in \mathcal{F}$

See example 13.3.

Corollary 13.8 $\mathcal{F}_{\min}$: $\mathcal{F} = \{\varnothing, \Omega\}$ is the simplest σ-algebra, telling us only if an event happened $\omega \in \Omega$ happened or not but not which one.

Corollary 13.9 $\mathcal{F}_{\max}$: $\mathcal{F} = 2^{\Omega}$ consists of all subsets of Ω and thus corresponds to full information i.e. we know if and which event happened.

Definition 13.7 Measurable Space $\{\Omega, \mathcal{F}\}$**:**
Is the pair of a set and sigma algebra i.e. a sample space and sigma algebra $\{\Omega, \mathcal{F}\}$.

Corollary 13.10 $\mathcal{F}$-measurable Event: The elements $A_i \in \mathcal{F}$ are called *measurable sets* or *$\mathcal{F}$-measurable.*

Interpretation

The σ-algebra represents all of possible events of the experiment that we can detect.
Thus we call the sets in $\mathcal{F}$ measurable sets/events.
The sigma algebra is the mathematical construct that tells us how much information we obtain once we conduct some experiment.

Definition 13.8
Sigma Algebra generated by a subset of Ω $\sigma(\mathcal{C})$**:**
Let $\mathcal{C}$ be a class of subsets of Ω. The σ-algebra generated by $\mathcal{C}$, denoted by $\sigma(\mathcal{C})$, is the *smallest* sigma algebra $\mathcal{F}$ that included all elements of $\mathcal{C}$ see example 13.4.

Definition 13.9 Borel σ-algebra $\mathcal{B}(\mathbb{R})$**:** 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 see example 13.5.

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. $[0, 1]$.

Corollary 13.11 : The Borel σ-algebra of $\mathbb{R}$ is generated by intervals of the form $(-\infty, a]$, where $a \in \mathbb{Q}$ ($\mathbb{Q}$ =rationals). See proof section 13.

Definition 13.10 ($\mathbb{P}$)-trivial Sigma Algebra:
is a σ-algebra $\mathcal{F}$ for which each event has a probability of zero or one:

$$\mathbb{P}(A) \in \{0, 1\} \qquad \forall A \in \mathcal{F} \tag{13.1}$$

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 $\mathcal{F}_{\min} = \{\Omega, \varnothing\}$.

0.2. Measures

Definition 13.11 Measure μ**:**
A measure defined on a measurable space $\{\Omega, \mathcal{F}\}$ is a function/map:

$$\mu : \mathcal{F} \mapsto [0, \infty] \tag{13.2}$$

for which holds:

- $\mu(\varnothing) = 0$
- countable additivity [def. 13.12]

Definition 13.12 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 $(F_i)_{i\geqslant 1}$ of $\mathcal{F}$ it holds that:

$$\mu\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} \mu(F_i) \quad \text{for all} \quad F_j \cap F_k = \varnothing \quad \forall j \neq k \tag{13.3}$$

Corollary 13.12 Additive Function: A function that satisfies countable additivity, is also additive, meaning that for every $F, G \in \mathcal{F}$ it holds:

$$\mu(F \cup G) = \mu(F) + \mu(G) \quad \Longleftrightarrow \quad F \cap G = \varnothing \tag{13.4}$$

Intuition

If we take two event that cannot occur simultaneously, then the probability that at least one vent occurs is just the sum of the measure (probabilities) of the original events.

Definition 13.13 Equivalent Measures $\mu \sim \nu$**:**
Let μ and ν be two measures defined on a measurable space[def. 13.7] $(\Omega, \mathcal{F})$. The two measures are said to be equivalent if it holds that:

$$\mu(A) > 0 \iff \nu(A) > 0 \qquad \forall A \subseteq \mathcal{F} \tag{13.5}$$

this is equivalent to μ and ν having equivalent null sets:

$$\mathcal{N}_\mu = \mathcal{N}_\nu \qquad \begin{aligned} \mathcal{N}_\mu &= \{A \in \mathcal{A} | \mu(A) = 0\} \\ \mathcal{N}_\nu &= \{A \in \mathcal{A} | \nu(A) = 0\} \end{aligned} \tag{13.6}$$

see example 13.6

Definition 13.14 Measure Space $\{\mathcal{F}, \Omega, \mu\}$**:**
The triplet of sample space, sigma algebra and a measure is called a measure space.

Definition 13.15 Lebesgue Measure on $\mathcal{B}$ λ**:**
Is the measure defined on the measurable space $\{\mathbb{R}, \mathcal{B}(\mathbb{R})\}$ which assigns the measure of each interval to be its length:

$$\lambda([a, b]) = b - a \tag{13.7}$$

Corollary 13.13 Lebesgue Measure of Atomitcs:

- The Lebesgue measure of a set containing only one point must be zero:
$$\lambda(\{a\}) = 0 \tag{13.8}$$
- The Lebesgue measure of a set containing countably many points $A = \{a_1, a_2 \ldots, a_n\}$ must be zero:
$$\lambda(A) + \sum_{i=1}^{n} \lambda(\{a_i\}) = 0 \tag{13.9}$$
- The Lebesgue measure of a set containing uncountably many points $A = \{a_1, a_2 \ldots, \}$ can be either zero, positive and finite or infinite.

0.3. 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 13.1 Non-negativity: The probability of an event is a non-negative real number:

$$\text{If } A \in \mathcal{F} \qquad \text{then} \qquad \mathbb{P}(A) \geqslant 0 \tag{13.10}$$

Axiom 13.2 Unitairity: The probability that at least one of the elementary events in the entire sample space Ω will occur is equal to one:

$$\text{The certain event} \qquad \mathbb{P}(\Omega) = 1 \tag{13.11}$$

Axiom 13.3 σ-additivity: If $A_1, A_2, A_3, \ldots \in \mathcal{F}$ are mutually disjoint, then:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \tag{13.12}$$

Corollary 13.14 : As a consequence of this it follows:

$$\mathbb{P}(\varnothing) = 0 \tag{13.13}$$

Corollary 13.15 Complementary Probability:

$$\mathbb{P}(A^{\mathrm{C}}) = 1 - \mathbb{P}(A) \qquad \text{with} \qquad A^{\mathrm{C}} = \Omega - A \tag{13.14}$$

Definition 13.16 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.

1. Conditional Probability

Definition 13.17 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:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \qquad \mathbb{P}(B) \neq 0 \tag{13.15}$$

2. Independent Events

Theorem 13.1
Independent Events: Let A, B be two events. A and B are said to be independent iffy:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \qquad \begin{aligned} \mathbb{P}(A|B) &= \mathbb{P}(A), \quad \mathbb{P}(B) > 0 \\ \mathbb{P}(B|A) &= \mathbb{P}(B), \quad \mathbb{P}(A) > 0 \end{aligned} \tag{13.16}$$

Note

The requirement of no impossible events follows from [def. 13.17]

Corollary 13.16 Pairwise Independent Evenest:
A finite set of events $\{A_i\}_{i=1}^{n} \in \mathcal{A}$ is *pairwise independent* if every pair of events is independent:

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \cap \mathbb{P}(A_j) \quad i \neq j, \quad \forall i, j \in \mathcal{A} \tag{13.17}$$

Corollary 13.17 Mutal Independent Evenest:
A finite set of events $\{A_i\}_{i=1}^{n} \in \mathcal{A}$ is *mutal independent* if every event A_j is independent of any intersection of the other events:

$$\mathbb{P}\left(\bigcap_{i=i}^{k} B_i\right) = \prod_{i=1}^{k} \mathbb{P}(B_i) \qquad \begin{aligned} &\forall \{B_i\}_{i=1}^{k} \subseteq \{A_i\}_{i=1}^{n} \\ &k \leqslant n, \qquad \{A_i\}_{i=1}^{n} \in \mathcal{A} \end{aligned} \tag{13.18}$$

3. Product Rule

Law 13.1 Product Rule: Let A, B be two events then the probability of both events occurring simultaneously is given by:

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) \tag{13.19}$$

4. Law of Total Probability

Definition 13.18 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* the sample space:

$$\bigcup_{i\in I} A_i = \Omega \qquad A_i \cap A_j = \varnothing \qquad i \neq j, \forall i, j \in I \tag{13.20}$$

Theorem 13.2
Law of Total Probability/Partition Equation:
Let $\{A_i : i \in I\}$ be a complete event field[def. 13.18] then it holds for $B \in \mathcal{B}$:

$$\mathbb{P}(B) = \sum_{i\in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i) \tag{13.21}$$

5. Bayes Theorem

Law 13.2 Bayes Rule: Let A, B be two events s.t. $\mathbb{P}(B) > 0$ then it holds:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \qquad \mathbb{P}(B) > 0 \tag{13.22}$$

follows directly fromeq. (13.19).

Theorem 13.3 Bayes Theorem: Let $\{A_i : i \in I\}$ be a complete event field[def. 13.18] and $B \in \mathcal{B}$ a random event s.t. $\mathbb{P}(B) > 0$, then it holds:

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i\in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i)} \tag{13.23}$$

proof section 13

Distributions on $\mathbb{R}$

6.1. Distribution Function

Definition 13.19 Distribution Function of $\mathbb{P}$ F**:** The *distribution function* F induced by a a probability $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B})$ is the function:

$$F(x) = \mathbb{P}((\infty, x]) \tag{13.24}$$

Theorem 13.4 : 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
- $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$

Corollary 13.18 : 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 $\mathbb{P} = \mathbb{Q}$.

6.2. Random Variables

A random variable X is a quantity that is not a variable in the classical sense but a variable with respect to the outcome of an experiment. Thus it is actually not a variable but a function/map.

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 $\mathcal{E} = \mathbb{R}^n$:

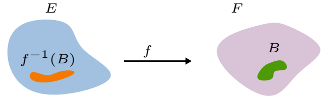
$$X : \Omega \mapsto \mathcal{E} \qquad\qquad \omega \mapsto X(\omega)$$

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:

$$\underbrace{\mathbb{P}_X(E)}_{\text{Probability for an event in } E} = \mathbb{P}(\{\omega : X(\omega) \in E\}) = \mathbb{P}(X \in E) = \overbrace{\mathbb{P}\left(X^{-1}(E)\right)}^{\text{Probability for an event in } \Omega}$$

Definition 13.20 $\mathcal{E}$-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

$$\forall B \in \mathcal{F} : \qquad f^{-1}(B) = \{\omega \in \mathcal{E} : f(\omega) \in B\} \in \mathcal{E} \tag{13.25}$$



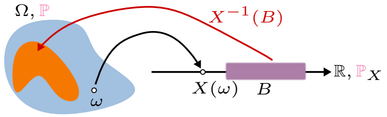
Interpretation

The pre-image[def. 5.7] 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 13.21 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 $(F, \mathcal{F})$:

$$X : \Omega \mapsto \mathcal{F} \quad (\mathcal{F}^n) \tag{13.26}$$

Since X is $\mathcal{E}$-measurable it holds that $X^{-1} : \mathcal{F} \mapsto \mathcal{E}$



Corollary 13.19 : Usually $F = \mathbb{R}$, which usually amounts to using the Borel σ-algebra $\mathcal{B}$ of $\mathbb{R}$.

Corollary 13.20 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 corollary 13.11):

$$\begin{aligned} X^{-1}(B) &= X^{-1}((-\infty, a]) \\ &= \{\omega \in \Omega : X(\omega) \leqslant a\} \in \mathcal{E} \quad \forall a \in \mathbb{R} \end{aligned} \tag{13.27}$$

Definition 13.22
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 13.21 Indicator Functions $I_A(\omega)$**:** An important class of measurable functions that can be used as r.v. are indicator functions:

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} \tag{13.28}$$

We know that a probability measure $\mathbb{P}$ on $\mathbb{R}$ is characterized by the quantities $\mathbb{P}((-\infty, a])$. Thus the quantities.

Corollary 13.22 : Let $(F, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$ and let $(E, \mathcal{E})$ and arbitrary measurable space. Let X be a real value function on E.

Then it holds that X is measurable *if and only if*

$\{X \leqslant a\} = \{\omega : X(\omega) \leqslant a\} = X^{-1}((-\infty, a]) \in \mathcal{E}$, each $a \in \mathbb{R}$

or $\{X < a\} \in \mathcal{E}$.

Explanation 13.1 (corollary 13.22). *A random variable is a function that is measurable if and only if its distribution function is defined.*

6.3. The Law of Random Variables

Definition 13.23 Law/Distribution of X: Let X be a r.v. on $\{\Omega, \mathcal{F}, \mathbb{P}\}$, with values in $\underline{(E, \mathcal{E})}$, then the *distribution/law* of X is defined as:

$$\mathbb{P} : \mathcal{B} \mapsto [0, 1] \tag{13.29}$$

$$\mathbb{P}^X(B) = \mathbb{P}\{X \in B\} = \mathbb{P}(\omega : X(\omega) \in B) \qquad \forall b \in \mathcal{E}$$

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 13.5 : The law/distribution of X is a probability measure $\mathbb{P}$ on $\underline{(E, \mathcal{E})}$.

Definition 13.24
(Cumulative) Distribution Function F_X**:** Given a real-valued r.v. then its *cumulative distribution function* is defined as:

$$F_X(x) = \mathbb{P}^X((-\infty, x]) = \mathbb{P}(X \leqslant x) \tag{13.30}$$

Corollary 13.23 : The distribution of $\mathbb{P}^X$ of a real valued r.v. is entirely characterized by its cumulative distribution function F_X[def. 13.31].

Property 13.1:

$$\mathbb{P}(X > x) = 1 - F_X(x) \tag{13.31}$$

Property 13.2: Probability of $X \in [a, b]$

$$\mathbb{P}(a < X \leqslant B) = F_X(b) - F_X(a) \tag{13.32}$$

6.4. Probability Density Function

Definition 13.25 Continuous Random Variable: Is a r.v. for which a probability density function f_X exists.

Definition 13.26 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:

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y \tag{13.33}$$

or alternatively:

$$f_X(x) = \lim_{\epsilon\to0} \frac{\mathbb{P}(x \leqslant X \leqslant x + \epsilon)}{\epsilon} \tag{13.34}$$

Corollary 13.24 $\mathbb{P}(X = b) = 0, \quad \forall b \in \mathbb{R}$**:**

$$\mathbb{P}(X = b) = \lim_{a\to b} \mathbb{P}(a < X \leqslant b) = \lim_{a\to b} \int_a^b f(x) = 0 \tag{13.35}$$

Corollary 13.25 corollary 13.24: From corollary 13.24 it follows that the exact borders are not necessary:

$$\begin{aligned} \mathbb{P}(a < X < b) &= \mathbb{P}(a \leqslant X < b) \\ &= \mathbb{P}(a < X \leqslant b) = \mathbb{P}(a \leqslant X << b) \end{aligned}$$

Corollary 13.26 :

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1 \tag{13.36}$$

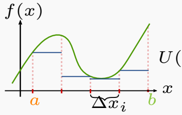
Notes

- 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.*

6.5. 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
$$\lim_{n\to\infty} \int f(x)\,\mathrm{d}x \overset{\text{str. u.c.}}{=} \int \lim_{n\to\infty} f(x)\,\mathrm{d}x$$



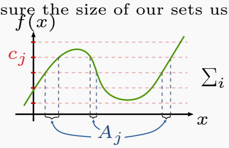
$$U(p) = \sum_i \sup(f(x_i)) \cdot \Delta x_i \xrightarrow{n\to\infty} \int f\,\mathrm{d}x$$

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.



$$\sum_i c_i \cdot \mu(A_i) \xrightarrow{n\to\infty} \int f\,\mathrm{d}\mu$$

Definition 13.27 Lebesgue Integral:

$$\lim_{n\to\infty} \sum_{i=1}^{n} c_i\mu(A_i) = \int_\Omega f\,\mathrm{d}\mu \qquad \begin{aligned} &f(x) \approx c_i \\ &\forall x \in A_i \end{aligned} \tag{13.37}$$

Definition 13.28
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:

$$X = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i} \qquad a_i \in \mathbb{R} \qquad \mathcal{A} \ni A_i = \begin{cases} 1 & \text{if } \{X = a_i\} \\ 0 & \text{else} \end{cases} \tag{13.38}$$

7. 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 13.1.

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 13.29 Independent Random Variables: Two real valued random variables X and Y are said to be independent iff:

$$\mathbb{P}(X \leqslant x|Y \leqslant y) = \mathbb{P}(X \leqslant x) \qquad \forall x, y \in \mathbb{R} \tag{13.39}$$

which amounts to:

$$\begin{aligned} F_{X,Y}(x, y) &= \mathbb{P}(\{X \leqslant x\} \cap \{Y \leqslant y\}) = \mathbb{P}(X \leqslant x, Y \leqslant y) \\ &= F_X(x)F_Y(y) \quad \forall x, y \in \mathbb{R} \end{aligned} \tag{13.40}$$

or alternatively iff:

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \qquad \forall A, B \in \mathcal{B} \tag{13.41}$$

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 13.30
Independent Identically Distributed:

8. Change Of Variables Formula

Formula 13.1
(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. 5.10]) 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:

$$p_Y(y) = \sum_{x_i\in\mathcal{X}_y} p_X(x_i) = \sum_{x\in\mathcal{Y}:g(x)=y} p_X(x) \tag{13.42}$$

see proof section 13

Formula 13.2
(Scalar Continuous) Change of Variables:

Let $X \sim f_X$ be a continuous r.v. and let g be an arbitrary strictly monotonic[def. 5.10] function.

Define a new r.v. Y as

$$\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\} \tag{13.43}$$

then the pdf of Y is given by:

$$f_Y(y) = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = f_X(x)\left|\frac{\mathrm{d}}{\mathrm{d}y}\left(g^{-1}(y)\right)\right| \tag{13.44}$$

$$= f_X(x)\frac{1}{\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|} = \frac{f_X(g^{-1}(y))}{\left|\frac{\mathrm{d}g}{\mathrm{d}x}(g^{-1}(y))\right|} \tag{13.45}$$

Formula 13.3
(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. 5.10] function

$$g : \mathbb{R}^n \mapsto \mathbb{R}^m$$

Define a new r.v. Y as

$$\mathcal{Y} = \{\boldsymbol{y} | \boldsymbol{y} = g(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X}\} \tag{13.46}$$

and let $h(\boldsymbol{x}) := g(\boldsymbol{x})^{-1}$ then the pdf of Y is given by:

$$\begin{aligned} f_Y(\boldsymbol{y}) &= f_X(x_1, \dots, x_n) \cdot |J| \\ &= f_X(h_1(\boldsymbol{y}), \dots, h_n(\boldsymbol{y})) \cdot |J| \\ &= f_X(\boldsymbol{y}) |\det D_{\boldsymbol{x}} h(\boldsymbol{x})| \Big|_{\boldsymbol{x}=\boldsymbol{y}} \\ &= f_X(g^{-1}(\boldsymbol{y})) \left| \det \left(\frac{\partial g}{\partial \boldsymbol{x}} \right) \right|^{-1} \end{aligned} \tag{13.47}$$

where $J = \det Dh$ is the Jaccobian[def. 6.4].
See also proof section 13 and example 13.8

Note

A monotonic function is required in order to satisfy inevitability.

Probability Distributions on $\mathbb{R}^n$

10. Joint Distribution

Definition 13.31
Joint (Cumulative) Distribution Function $F_{\boldsymbol{X}}$:
Let $\boldsymbol{X} = (X_1 \cdots\cdots X_n)$ be a random vector in $\mathbb{R}^n$, then its *cumulative distribution function* is defined as:

$$\begin{aligned} F_{\boldsymbol{X}}(\boldsymbol{x}) &= \mathbb{P}^X((-\infty, \boldsymbol{x}]) = \mathbb{P}(\boldsymbol{X} \leqslant \boldsymbol{x}) \\ &= \mathbb{P}(X_1 \leqslant x_1, \dots X_n \leqslant x_n) \end{aligned} \tag{13.48}$$

Definition 13.32 Joint Probability Distribution:
Let $\boldsymbol{X} = (X_1 \cdots\cdots X_n)$ be a random vector in $\mathbb{R}^n$ with associated cdf $F_{\boldsymbol{X}}$. If $F_{\boldsymbol{X}}$ is continuously integrable for all $\boldsymbol{x} \in \mathbb{R}$ then $\boldsymbol{X}$ has a *probability density* f_X defined by:

$$F_X(x) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f_{\boldsymbol{X}}(y_1, \dots, y_n) \, \mathrm{d}y_1 \, \mathrm{d}y_n \tag{13.49}$$

or alternatively:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(x_1 \leqslant X_1 \leqslant x_1 + \epsilon, \dots, x_n \leqslant X_n \leqslant x_n + \epsilon)}{\epsilon} \tag{13.50}$$

10.1. Marginal Distribution

Definition 13.33 Marginal Distribution:

11. The Expectation

Definition 13.34 Expectation:

$$\mathbb{E}[X] = \int_\Omega X(\omega) \mathbb{P}(\mathrm{d}\omega) = \int_\Omega X \, \mathrm{d}\mathbb{P} \tag{13.51}$$

Corollary 13.27 Expectation of simple r.v.:
If X is a simple[def. 13.28] r.v. its *expectation* is given by:

$$\mathbb{E}[X] = \sum_{i=1}^n a_i \mathbb{P}(A_i) \tag{13.52}$$

12. Moment Generating Function (MGF)

Definition 13.35 Moment of Random Variable: The i-th moment of a random variable X is defined as (if it exists):

$$m_i := \mathbb{E}\left[X^i\right] \tag{13.53}$$

Definition 13.36 ψ_X
Moment Generating Function (MGF):

$$\psi_X(t) = \mathbb{E}\left[e^{tX}\right] \qquad t \in \mathbb{R} \tag{13.54}$$

Corollary 13.28 Sum of MGF: The moment generating function of a sum of n independent variables $(X_j)_{1 \leqslant j \leqslant n}$ is the product of the moment generating functions of the components:

$$\psi_{S_n}(t) = \psi_{X_1}(t) \cdots \psi_{X_n}(t) \qquad S_n := X_1 + ldotsX_n \tag{13.55}$$

Corollary 13.29 : The i-th moment of a random variable is the i-th derivative of its associated moment generating function evaluated zero:

$$\mathbb{E}\left[X^i\right] = \psi_X^{(i)}(0) \tag{13.56}$$

13. 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 13.37 $\hat{\mu}$
Fourier Transformed Probability Measure:

$$\hat{\mu} = \int e^{i\langle u, x \rangle} \mu(\mathrm{d}x) \tag{13.57}$$

Corollary 13.30 : As $e^{i\langle u, x \rangle}$ can be rewritten using formulaeqs. (9.5) and (9.6) it follows:

$$\hat{\mu} = \int \cos(\langle u, x \rangle) \mu(\mathrm{d}x) + i \int \sin(\langle u, x \rangle) \mu(\mathrm{d}x) \tag{13.58}$$

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 13.38 Characteristic Function φ_X**:** Let $\boldsymbol{X}$ be an $\mathbb{R}^n$-valued random variable. Its characteristic function $\varphi_{\boldsymbol{X}}$ is defined on $\mathbb{R}^n$ as:

$$\varphi_{\boldsymbol{X}}(u) = \int e^{i\langle \boldsymbol{u}, \boldsymbol{x} \rangle} \mathbb{P}^X(\mathrm{d}\boldsymbol{x}) = \widehat{\mathbb{P}^X}(\boldsymbol{u}) \tag{13.59}$$

$$= \mathbb{E}\left[e^{i\langle \boldsymbol{u}, \boldsymbol{x} \rangle}\right] \tag{13.60}$$

Corollary 13.31 : 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 13.6 : 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 13.7 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 13.32 : Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be an $\mathbb{R}^n$-valued random variable. Then the real valued r.v.'s $(X_j)_{1 \leqslant j \leqslant n}$ are independent if and only if:

$$\varphi_X(u_1, \dots, u_n) = \prod_{j=1}^n \varphi_{X_j}(u_j) \tag{13.61}$$

Proofs

Proof. corollary 13.11: Let $\mathcal{C}$ denote all open intervals.
Since every open set in $\mathbb{R}$ is the countable union of open intervals[def. 3.8], 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

$$(a, b) = \cup_{n=1}^\infty (a_n, b_n] = \cup_{n=1}^\infty \left((-\infty, b_n] \cap (-\infty, a_n]^{\mathrm{C}} \right)$$

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

$$\mathcal{B} = \sigma(\mathcal{C}) \subset \sigma((D) \subset \mathcal{B}$$

□

Proof. theorem 13.3 Plug eq. (13.21) into the denominator and eq. (13.19) into the nominator and then use [def. 13.17]:

$$\frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i \in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i)} = \frac{\mathbb{P}(B \cap A_j)}{\mathbb{P}(B)} = \mathbb{P}(A_j|B)$$

□

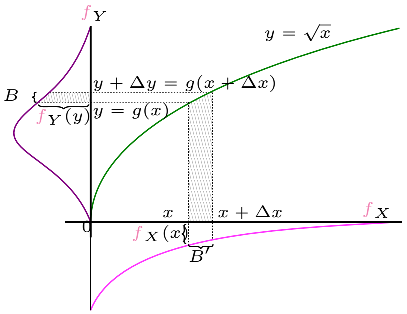
Proof. formula 13.1:

$$Y = g(X) \qquad \Longleftrightarrow \qquad \mathbb{P}(Y = y) = \mathbb{P}(x \in \mathcal{X}_y) = \mathrm{p}_Y(y)$$

□

Proof. formula 13.2 (non-formal): The probability contained in a differential area must be invariant under a change of variables that is:

$$|f_Y(y) \, \mathrm{d}y| = |f_x(x) \, \mathrm{d}x|$$



□

Proof. formula 13.2 from CDF:

$$\mathbb{P}(Y \leqslant y) = \mathbb{P}(g(X) \leqslant y) = \begin{cases} \mathbb{P}(X \leqslant g^{-1}(y)) & \text{if } g \text{ is increas.} \\ \mathbb{P}(X \geqslant g^{-1}(y)) & \text{if } g \text{ is decreas.} \end{cases}$$

If g is monotonically increasing:

$$F_Y(y) = F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) = f_X(x) \cdot \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)$$

If g is monotonically decreasing:

$$F_Y(y) = 1 - F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) = -f_X(x) \cdot \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)$$

□

Proof. formula 13.2: **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:

$$y = g(x) \qquad \Rightarrow \qquad \mathbb{P}(y) = \mathbb{P}(g(x)) \text{ (for disc. rv.)}$$

Now lets consider the probability for the continuous r.v.s:

$$\mathbb{P}(X \in B) = \int_x^{x+\Delta x} f_X(t) \, \mathrm{d}t \xrightarrow{\Delta x \to 0} |\Delta x \cdot f_x(x)|$$

For y we use Taylor (**??**)

$$\begin{aligned} g(x + \Delta x) &\overset{\text{eq. (5.40)}}{=} g(x) + \frac{\mathrm{d}g}{\mathrm{d}x} \Delta y \qquad \text{for } \Delta x \to 0 \\ &= \quad y + \Delta y \qquad \text{with } \Delta y := \frac{\mathrm{d}g}{\mathrm{d}x} \cdot \Delta x \end{aligned} \tag{13.62}$$

Thus for $\mathbb{P}(Y \in B')$ it follows:

$$\begin{aligned} \mathbb{P}(X \in B') &= \int_y^{y+\Delta y} f_Y(t) \, \mathrm{d}t \xrightarrow{\Delta y \to 0} |\Delta y \cdot f_Y(y)| \\ &= \left| \frac{\mathrm{d}g}{\mathrm{d}x}(x) \Delta x \cdot f_Y(y) \right| \end{aligned}$$

Now we simply need to related the surface of the two pdfs:

$$B = [x, x + \Delta x] \overset{\text{same surfaces}}{\propto} [y, y + \Delta y] = B'$$

$$\mathbb{P}(Y \in B) = \mathbb{P}(X \in B')$$

$$\overset{\Delta y \to 0}{\Longleftrightarrow} |f_Y(y) \cdot \Delta y| = \left| f_Y(y) \cdot \frac{\mathrm{d}g}{\mathrm{d}x}(x) \Delta x \right| = |f_X(x) \cdot \Delta x|$$

$$f_Y(y) \left| \cdot \frac{\mathrm{d}g}{\mathrm{d}x}(x) \right| |\Delta x| = f_X(x) \cdot |\Delta x|$$

$$\Rightarrow f_Y(y) = \frac{f_X(x)}{\left| \frac{\mathrm{d}g}{\mathrm{d}x}(x) \right|} = \frac{f_X(g^{-1}(y))}{\left| \frac{\mathrm{d}g}{\mathrm{d}x} g^{-1}(y) \right|}$$

□

Examples

Example 13.1 :

- 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 13.2 :

- 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 13.3 : 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:

$$\mathcal{F} = \{\varnothing, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}\}$$

Example 13.4 : If we are only interested in the subset-set $A \in \Omega$ of our experiment, then we can look at the corresponding generating σ-algebra $\sigma(A) = \left\{\varnothing, A, A^{\mathrm{C}}, \Omega\right\}$.

Example 13.5 :

- open half-lines: $(-\infty, a)$ and (a, ∞),
- union of open half-lines: $\underline{(a, b) = (-\infty, a) \cup} (b, \infty)$,
- closed interval: $[a, b] = \overline{(-\infty, \cup a) \cup (b, \infty)}$,
- 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 $(a, b] = (-\infty, b] \cup (a, \infty)$,
- 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 13.6 Equivalent (Probability) Measures:

$$\Omega = \{1, 2, 3\} \qquad \begin{aligned} \mathbb{P}(\{1, 2, 3\}) &= \{2/3, 1/6, 1/6\} \\ \tilde{\mathbb{P}}(\{1, 2, 3\}) &= \{1/3, 1/3, 1/3\} \end{aligned}$$

Example 13.7 :

add example fat book p.1286

add example prob th book 4

Example 13.8 formula 13.2: Let $X, Y \overset{\text{ind.}}{\sim} \mathcal{N}(0,1)$.
Question: proof that:

$$U = X + Y \qquad\qquad V = X - 1$$

are indepdent and normally distributed:

$$h(u,v) = \begin{cases} h_1(u,v) = \underline{\frac{u+v}{2}} \\ h_2(u,v) = \underline{\frac{u-v}{2}} \end{cases} \qquad J = \det\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} = -\frac{1}{2}$$

$$\begin{aligned} f_{U,V} &= f_{X,Y}(\underline{x}, \underline{y}) \cdot \frac{1}{2} \\ &\overset{\text{indp.}}{=} f_X(\underline{x}) \cdot f_X(\underline{y}) \\ &= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\underline{x}^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\underline{y}^2}{2}} \\ &= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\left\{\left(\underline{\frac{u+v}{2}}\right)^2 + \left(\underline{\frac{u-v}{2}}\right)^2 \Big/ 2\right\}} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{u^2}{4}} \cdot \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{v^2}{4}} \end{aligned}$$

Thus U, V are independent r.v. distributed as $\mathcal{N}(0,2)$.

Combinatorics

0.1. Permutations

Definition 14.1 Permutation $n!$: Given a set[def. 3.1] $\mathcal{S}$ of n distinct objects, into how many distinct sequences/orders can we arrange/permutate those distinct objects

$$P(\mathcal{S}) = n! \quad \iff \quad P(\mathcal{S}) = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 1 \tag{14.1}$$

If there exists multiple n_j objects of the same kind within $\mathcal{S}$ with $j \in 1, \ldots, n-1$ then we need to divide by those permutations:

$$P(\mathcal{S}) = \frac{n!}{n_1! \cdot \ldots \cdot n_k} \quad \text{s.t.} \quad \sum_{i=1}^{k} n_i \leqslant n \tag{14.2}$$

Note

This is because the sequence/order does not change if we exchange objects of the same kind (e.g. red ball by red ball).

Statistics

The probability that a discret random variable x is equal to some value $\bar{x} \in \mathcal{X}$ is:

$$p_x(\bar{x}) = \mathbb{P}(x = \bar{x})$$

addapet

Definition 15.1 Almost Surely (a.s.): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An event $\omega \in \mathcal{F}$ happens almost surely iff

$$\mathbb{P}(\omega) = 1 \quad \iff \quad \omega \text{ happens a.s.} \tag{15.1}$$

Definition 15.2 Probability Mass Function (PMF):

Definition 15.3 Discrete Random Variable (DVR): The set of possible values $\bar{x}$ of $\mathcal{X}$ is countable of finite.

$$\mathcal{X} = \{0,1,2,3,4,\ldots,8\} \qquad \mathcal{X} = \mathbb{N} \tag{15.2}$$

Definition 15.4 Probability Density Function (PDF): Is real function $f : \mathbb{R}^n \to [0, \infty)$ that satisfies:

Non-negativity: $$f(x) \geqslant 0, \quad \forall x \in \mathbb{R}^n \tag{15.3}$$

Normalization: $$\int_{-\infty}^{\infty} f(x)\,dx \overset{!}{=} 1 \tag{15.4}$$

Must be integrable (15.5)

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:

$$\mathbb{P}(X = x) = 0 \qquad \textbf{a.s.}$$

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 15.5 Continuous Random Variable (CRV): A real random variable (rrv) X is said to be (absolutely) continuous if there exists a pdf ([def. 15.4]) f_X s.t. for any subset $B \subset \mathbb{R}$ it holds:

$$\mathbb{P}(X \in B) = \int_B f_X(x)\,dx \tag{15.6}$$

Property 15.1 Zero Probability: If X is a continuous rrv ([def. 15.5]), then:

$$\mathbb{P}(X = a) = 0 \qquad \forall a \in \mathbb{R} \tag{15.7}$$

Property 15.2 Open vs. Closed Intervals: For any real numbers a and b, with $a < b$ it holds:

$$\begin{aligned} \mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(a \leqslant X < b) &= \mathbb{P}(a < X \leqslant b) \\ &= \mathbb{P}(a < X < b) \end{aligned} \tag{15.8}$$

$\iff$ including or not the bounds of an interval does not modify the probability of a continuous rrv.

Note

Changing the value of a function at finitely many points has no effect on the value of a definite integral.

Corollary 15.1 : In particular for any real numbers a and b with $a < b$, letting $B = [a, b]$ we obtain:

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_x(x)\,dx$$

Proof. Property 15.1:

$$\begin{aligned} \mathbb{P}(X = a) &= \lim_{\Delta x \to 0} \mathbb{P}(X \in [a, a + \Delta x]) \\ &= \lim_{\Delta x \to 0} \int_a^{a+\Delta x} f_X(x)\,dx = 0 \end{aligned}$$

□

Proof. Property 15.2:

$$\begin{aligned} \mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(a \leqslant X < b) &= \mathbb{P}(a < X \leqslant b) \\ &= \mathbb{P}(a < X < b) = \int_a^b f_X(x)\,dx \end{aligned}$$

□

Definition 15.6 Support of a probability density function: The support of the density of a pdf $f_X(.)$ is the set of values of the random variable X s.t. its pdf is non-zero:

$$\text{supp}(()f_X) := \{x \in \mathcal{X} | f(x) > 0\} \tag{15.9}$$

Note: this is not a rigorous definition.

Theorem 15.1 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 15.2 Identically Distributed: From theorem 15.1 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 Fucntion

Definition 15.7 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:

$$F_X(x) = \mathbb{P}(X \leqslant x) \qquad \forall x \in \mathbb{R}$$

Property 15.3:

Monotonically Increasing $$x \leqslant y \iff F_X(x) \leqslant F_X(y) \quad \forall x, y \in \mathbb{R} \tag{15.10}$$

Upper Limit $$\lim_{x \to \infty} F_X(x) = 1 \tag{15.11}$$

Lower Limit $$\lim_{x \to -\infty} F_X(x) = 0 \tag{15.12}$$

Definition 15.8 CDF of a discret rv X: Let X be discret rv with pdf p_X, then the CDF of X is given by:

$$F_X(x) = \mathbb{P}(X \leqslant x) = \sum_{t=-\infty}^{x} p_X(t)$$

Definition 15.9 CDF of a continuous rv X: Let X be continuous rv with pdf f_X, then the CDF of X is given by:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt \iff \frac{\partial F_X(x)}{\partial x} = f_X(x)$$

Lemma 15.1 Probability Interval: Let X be a continuous rrv with pdf f_X and cumulative distribution function F_X, then it holds that:

$$\mathbb{P}(a \leqslant X \leqslant b) = F_X(b) - F_X(a) \tag{15.13}$$

Proof. [def. 15.9]:

$$F_X(x) = \mathbb{P}(X \leqslant x) = \mathbb{P}(X \in (-\infty, x)) = \int_{-\infty}^{x} f_X(t)\,dt$$

□

Proof. lemma 15.1:

$$\mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(X \leqslant b) - \mathbb{P}(X \leqslant a)$$

or by the fundamental theorem of calculus (theorem 5.2):

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_X(t)\,dt = \int_a^b \frac{\partial F_X(t)}{\partial t}\,dt = [F_X(t)]|_a^b$$

□

Theorem 15.2 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 15.10 Expectation (disc. case):

$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{\boldsymbol{x}} \in \mathcal{X}} \bar{\boldsymbol{x}} p_x(\bar{\boldsymbol{x}}) \tag{15.14}$$

Definition 15.11 Expectation (cont. case):

$$\mathbb{E}_x[x] := \int_{\bar{\boldsymbol{x}} \in \mathcal{X}} \bar{\boldsymbol{x}} f_x(\bar{\boldsymbol{x}})\,d\bar{\boldsymbol{x}} \tag{15.15}$$

Law 15.1 Expectation of independent variables:

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] \tag{15.16}$$

Property 15.4 Translation and scaling: If $\boldsymbol{X} \in \mathbb{R}^n$ and $\boldsymbol{Y} \in \mathbb{R}^n$ are random vectors, and $a, b, a \in \mathbb{R}^n$ are constants then it holds:

$$\mathbb{E}[a + b\boldsymbol{X} + c\boldsymbol{Y}] = a + b\mathbb{E}[\boldsymbol{X}] + c\mathbb{E}[\boldsymbol{Y}] \tag{15.17}$$

Thus $\mathbb{E}$ is a linear operator ([def. 5.11]).

Note: Expectation of the expectation

The expectation of a r.v. X is a constant hence with Property 15.6 it follows:

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \tag{15.18}$$

Property 15.5 Matrix×Expectation: If $\boldsymbol{X} \in \mathbb{R}^n$ is a randomn vector and $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{B} \in \mathbb{R}^{n \times m}$ are constant matrices then it holds:

$$\mathbb{E}[\boldsymbol{AXB}] = \boldsymbol{A}\mathbb{E}[(\boldsymbol{XB})] = \boldsymbol{A}\mathbb{E}[\boldsymbol{X}]\,\boldsymbol{B} \tag{15.19}$$

Proof. eq. (15.27):

$$\begin{aligned} \mathbb{E}[XY] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x,y)xy \\ &\overset{??}{=} \sum_{x \in \mathcal{X}} p_X(x)x \sum_{y \in \mathcal{Y}} p_Y(y)y = \mathbb{E}[X]\,\mathbb{E}[Y] \end{aligned}$$

□

Law 15.2 of the Unconscious Statistician: Let X be a random variable $X \in \mathcal{X}$ and define $Y \in \mathcal{Y}$ as $Y = g(x)$ s.t. $\mathcal{Y} = \{y | y = g(x), \forall x \in \mathcal{X}\}$, then Y is a random variable with expectation:

$$\mathbb{E}_Y[y] = \sum_{y \in \mathcal{Y}} y p_Y(y) = \sum_{x \in \mathcal{X}} g(x) p_x(x) \quad \text{or integral for CRV} \tag{15.20}$$

Consequence

Hence if we p_X we do not have to first calculate p_Y in order to calculate $\mathbb{E}_Y[y]$.

Theorem 15.3 Jensen's Inequality: If X is a random variable and f is a convex function, then it holds that:

$$f(\mathbb{E}[X]) \leqslant \mathbb{E}[f(X)] \tag{15.21}$$

on the contrary if f is a concave function it follows:

$$f(\mathbb{E}[X]) \geqslant \mathbb{E}[f(X)] \tag{15.22}$$

Definition 15.12 Autocorrelation/Crosscorelation $\gamma(t_1, t_2)$: Describes the covariance ([def. 15.16]) between the two values of a stochastic process $(\boldsymbol{X}_t)_{t \in T}$ at different time points t_1 and t_2.

$$\gamma(t_1, t_2) = \text{Cov}[\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}] = \mathbb{E}[(\boldsymbol{X}_{t_1} - \mu_{t_1})(\boldsymbol{X}_{t_2} - \mu_{t_2})] \tag{15.23}$$

For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance:

$$\gamma(t,t) = \text{Cov}[\boldsymbol{X}_t, \boldsymbol{X}_t] \overset{\text{eq. (15.41)}}{=} = \mathbb{V}[\boldsymbol{X}_t] \tag{15.24}$$

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 $\boldsymbol{x}(t)$ the autocorrelation function $\gamma(t, t - \tau)$ describes how *similar* the time translated function $\boldsymbol{x}(t - \tau)$ and the original function $\boldsymbol{x}(t)$ are.
- If there exists some relation between the values of the time series that is non-random then the autocorrelation is nonzero.
- The autocorrelation is maximized/most similar for no translation $\tau = 0$ at all.

2. Key Figures

2.1. The Expectation

Definition 15.13 Expectation (disc. case):

$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{\boldsymbol{x}}\in\mathcal{X}} \bar{\boldsymbol{x}}\mathrm{p}_x(\bar{\boldsymbol{x}}) \tag{15.25}$$

Definition 15.14 Expectation (cont. case):

$$\mathbb{E}_x[x] := \int_{\bar{\boldsymbol{x}}\in\mathcal{X}} \bar{\boldsymbol{x}} f_x(\bar{\boldsymbol{x}})\,\mathrm{d}\bar{\boldsymbol{x}} \tag{15.26}$$

Law 15.3 Expectation of independent variables:

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \tag{15.27}$$

Property 15.6 Translation and scaling: If $\boldsymbol{X} \in \mathbb{R}^n$ and $\boldsymbol{Y} \in \mathbb{R}^n$ are random vectors, and $a, b, a \in \mathbb{R}^n$ are constants then it holds:

$$\mathbb{E}[a + b\boldsymbol{X} + c\boldsymbol{Y}] = a + b\mathbb{E}[\boldsymbol{X}] + c\mathbb{E}[\boldsymbol{Y}] \tag{15.28}$$

Thus $\mathbb{E}$ is a linear operator[def. 5.11].

Property 15.7
Affine Transformation of the Expectation:
If $\boldsymbol{X} \in \mathbb{R}^n$ is a randomn vector, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ a constant matrix and $b \in \mathbb{R}^n$ then it holds:

$$\mathbb{E}[\boldsymbol{AX} + b] = \boldsymbol{A}\mu + \boldsymbol{b} \tag{15.29}$$

Note: Expectation of the expectation

The expectation of a r.v. X is a constant hence with Property 15.6 it follows:

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \tag{15.30}$$

Property 15.8 Matrix×Expectation: If $\boldsymbol{X} \in \mathbb{R}^n$ is a randomn vector and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{B} \in \mathbb{R}^{n\times m}$ are constant matrices then it holds:

$$\mathbb{E}[\boldsymbol{AXB}] = \boldsymbol{A}\mathbb{E}[(\boldsymbol{XB})] = \boldsymbol{A}\mathbb{E}[\boldsymbol{X}]\boldsymbol{B} \tag{15.31}$$

Proof. eq. (15.27):

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathrm{p}_{X,Y}(x,y)xy \\ &\overset{??}{=} \sum_{x\in\mathcal{X}} \mathrm{p}_X(x)x \sum_{y\in\mathcal{Y}} \mathrm{p}_Y(y)y = \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

□

Law 15.4 of the Unconscious Statistician: Let X be a random variable $X \in \mathcal{X}$ and define $Y \in \mathcal{Y}$ as $Y = g(x)$ s.t. $\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\}$, then Y is a random variable with expectation:

$$\mathbb{E}_Y[y] = \sum_{y\in\mathcal{Y}} y\mathrm{p}_Y(y) = \sum_{x\in\mathcal{X}} g(x)\mathrm{p}_x(x) \qquad \text{or integral for CRV} \tag{15.32}$$

Consequence

Hence if we p_X we do not have to first calculate p_Y in order to calculate $\mathbb{E}_Y[y]$.

Theorem 15.4 Jensen's Inequality: If X is a random variable and f is a convex function, then it holds that:

$$f(\mathbb{E}[X]) \leqslant \mathbb{E}[f(X)] \tag{15.33}$$

on the contrary if f is a concave function it follows:

$$f(\mathbb{E}[X]) \geqslant \mathbb{E}[f(X)] \tag{15.34}$$

2.2. The Variance

Definition 15.15 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 executed value $\mathbb{E}[X]$ and is defined by:

$$\mathbb{V}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \overset{\text{see section 3}}{=} \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \tag{15.35}$$

2.2.1. Properties

Property 15.9 Variance of a Constant: If $a \in \mathbb{R}$ is a constant then it follows that its expected value is deterministic ⇒ we have no uncertainty ⇒ no variance:

$$\mathbb{V}[a] = 0 \qquad \textbf{with} \qquad a \in \mathbb{R} \tag{15.36}$$

see shift and scaling for proof section 3

Property 15.10 Shifting and Scaling:

$$\mathbb{V}[a + bX] = a^2\sigma^2 \qquad \textbf{with} \qquad a \in \mathbb{R} \tag{15.37}$$

see section 3

Property 15.11
Affine Transformation of the Variance:
If $\boldsymbol{X} \in \mathbb{R}^n$ is a randomn vector, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ a constant matrix and $b \in \mathbb{R}^n$ then it holds:

$$\mathbb{V}[\boldsymbol{AX} + b] = \boldsymbol{A}\mathbb{V}[\boldsymbol{X}]\boldsymbol{A}^\intercal \tag{15.38}$$

see section 3.

Definition 15.16 Covariance: The Covariance is a measure of how much two or more random variables vary linearly with each other.

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned} \tag{15.39}$$

see section 3

Definition 15.17 Covariance Matrix: The variance of a k-dimensional random vector $\boldsymbol{X} = \begin{pmatrix} X_1 & \dots & X_k\end{pmatrix}$ is given by a p.s.d. eq. (8.11) 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:

$$\begin{aligned}\mathbb{V}[\boldsymbol{X}] :=& \Sigma(\boldsymbol{X}) := \mathrm{Cov}[\boldsymbol{X},\boldsymbol{X}] := \\ =&\mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^\intercal\right] \\ =&\mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^\intercal\right] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\intercal \in [-\infty, \infty]\end{aligned} \tag{15.40}$$

$$= \begin{bmatrix} \mathbb{V}[X_1] & \cdots & \mathrm{Cov}[X_1, X_k] \\ \vdots & \ddots & \\ \mathrm{Cov}[X_k, X_1] & \cdots & \mathbb{V}[X_k] \end{bmatrix}$$

$$= \begin{bmatrix} \mathbb{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \cdots & \mathbb{E}[(X_1 - \mu_1)(X_k - \mu_k)] \\ \vdots & \ddots & \\ \mathbb{E}[(X_k - \mu_k)(X_1 - \mu_1)] & \cdots & \mathbb{E}[(X_k - \mu_k)(X_k - \mu_k)] \end{bmatrix}$$

Note: Covariance and Variance

The variance is a special case of the covariance in which two variables are identical:

$$\mathrm{Cov}[X,X] = \mathbb{V}[X] \equiv \sigma^2(X) \equiv \sigma_X^2 \tag{15.41}$$

add http://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/

Property 15.12 Translation and Scaling:

$$\mathrm{Cov}(a + bX, c + dY) = bd\mathrm{Cov}(X,Y) \tag{15.42}$$

Property 15.13
Affine Transformation of the Covariance:
If $\boldsymbol{X} \in \mathbb{R}^n$ is a randomn vector, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ a constant matrix and $b \in \mathbb{R}^n$ then it holds:

$$\mathrm{Cov}[\boldsymbol{AX} + b] = \boldsymbol{A}\mathbb{V}[\boldsymbol{X}]\boldsymbol{A}^\intercal = \boldsymbol{A}\Sigma(\boldsymbol{X})\boldsymbol{A}^\intercal \tag{15.43}$$

Definition 15.18 Correlation Coefficient: Is the standardized version of the covariance:

$$\begin{aligned}\mathrm{Corr}[\boldsymbol{X}] :=& \frac{\mathrm{Cov}[\boldsymbol{X}]}{\sigma_{X_1}\cdots\sigma_{X_k}} \in [-1, 1] \\ =& \begin{cases} +1 & \text{if } Y = aX + b \text{ with } a > 0, b \in \mathbb{R} \\ -1 & \text{if } Y = aX + b \text{ with } a < 0, b \in \mathbb{R}\end{cases}\end{aligned} \tag{15.44}$$

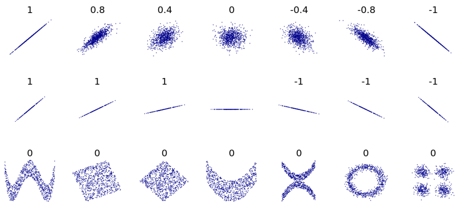


Figure 1: Several sets of (x, y) points, with their correlation coefficient

Law 15.5 Translation and Scaling:

$$\mathrm{Corr}(a + bX, c + dY) = \mathrm{sign}(b)\mathrm{sign}(d)\mathrm{Cov}(X,Y) \tag{15.45}$$

Note

- The correlation/covariance reflects the noisiness and direction of a linear relationship (top row fig. 1), **but** not the slope of that relationship (middle row fig. 1) nor many aspects of nonlinear relationships (bottom row)
- The set in the center of fig. 1 has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.
- Zero covariance/correlation $\mathrm{Cov}(X,Y) = \mathrm{Corr}(X,Y) = 0$ 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 15.5.
2. Correlation is dimensionless, whereas the unit of the covariance is obtained by the product of the units of the two RV variables.

Law 15.6 Covariance of independent RVs: The covariance/correlation of two independent variable's (**??**) is zero:

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &\overset{\text{eq. (15.27)}}{=} = \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0\end{aligned}$$

Zero covariance/correlation⇏ independence

$$\mathrm{Cov}(X,Y) = \mathrm{Corr}(X,Y) = 0 \nRightarrow \mathrm{p}_{X,Y}(x,y) = \mathrm{p}_X(x)\mathrm{p}_Y(y)$$

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:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{Cov}(X, X^2) = \mathbb{E}\left[X \cdot X^2\right] - \mathbb{E}[X]\mathbb{E}\left[X^2\right] \\ &= \mathbb{E}\left[X^3\right] - \mathbb{E}[X]\mathbb{E}\left[X^2\right] \underset{\text{eq. (15.56)}}{\overset{\text{eq. (15.67)}}{=}} 0 - 0 \cdot \mathbb{E}\left[X^2\right]\end{aligned}$$

⇒ the relationship between Y and X must be non-linear.

Definition 15.19 Quantile: Are specific values q_α in the range[def. 5.6] of a random variable X that are defined as the value for which the cumulative probability is less then $\alpha \in (0,1)$:

$$q_\alpha : \mathbb{P}(X \leqslant x) = F_X(q_\alpha) = \alpha \quad \xrightarrow{F\text{ invert.}} \quad q_\alpha = F_X^{-1}(\alpha) \tag{15.46}$$

add figure

3. Proofs

Proof. eq. (15.35)

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\ &\overset{\text{Property 15.6}}{=} \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}\left[X^2\right] - \mu^2\end{aligned}$$

□

Proof. Property 15.10

$$\begin{aligned}\mathbb{V}[a + bX] &= \mathbb{E}\left[(a + bX - \mathbb{E}[a + bX])^2\right] \\ &= \mathbb{E}\left[\left(\cancel{a} + bX - \cancel{a} - b\mathbb{E}[X]\right)^2\right] \\ &= \mathbb{E}\left[(bX - b\mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[b^2(X - \mathbb{E}[X])^2\right] \\ &= b^2\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = b^2\sigma^2\end{aligned}$$

□

Proof. Property 15.11

$$\begin{aligned}\mathbb{V}(\boldsymbol{AX} + b) &= \mathbb{E}\left[(\boldsymbol{AX} - \mathbb{E}[\boldsymbol{XA}])^2\right] + 0 = \\ &=\mathbb{E}\left[(\boldsymbol{AX} - \mathbb{E}[\boldsymbol{AX}])(\boldsymbol{AX} - \mathbb{E}[\boldsymbol{AX}])^\intercal\right] \\ &=\mathbb{E}\left[\boldsymbol{A}(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{A}(\boldsymbol{X} - (\mathbb{E}[\boldsymbol{X}]))^\intercal\right] \\ &=\mathbb{E}\left[\boldsymbol{A}(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - (\mathbb{E}[\boldsymbol{X}])^\intercal\boldsymbol{A}^\intercal\right] \\ &=\boldsymbol{A}\mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - (\mathbb{E}[\boldsymbol{X}])^\intercal\right]\boldsymbol{A}^\intercal = \boldsymbol{A}\mathbb{V}[\boldsymbol{X}]\boldsymbol{A}^\intercal\end{aligned}$$

□

Proof. eq. (15.39)

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

□

Discrete Distributions

Definition 15.20 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 $\mathrm{Bern}(p)$

Definition 15.21 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 15.22 Bernoullidistribution $X \sim \mathrm{Bern}(p)$: X is a binary variable i.e. can only attain the values 0 (failure) or 1 (success) with a parameter p that signifies the success probability:

$$p(x;p) = \begin{cases} p & \text{for } x = 1 \\ 1-p & \text{for } x = 0 \end{cases} \iff \begin{cases} \mathbb{P}(X=1) = p \\ \mathbb{P}(X=0) = 1-p \end{cases}$$
$$= p^x \cdot (1-p)^{1-x} \quad \text{for } x \in \{0,1\}$$

$$\mathbb{E}[X] = p \qquad (15.47) \qquad \mathbb{V}[X] = p(1-p) \qquad (15.48)$$

4.2. Binomial Distribution $\mathcal{B}(n,p)$

Definition 15.23 Binomial Distribution: Models the probability of exactly X success given a fixed number *n-Bernoulli experiments*[def. 15.21], where the probability of success of a single experiment is given by p:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

n :nb. of repetitions
x :nb. of successes
p :probability of success

$$\mathbb{E}[X] = np \qquad (15.49) \qquad \mathbb{V}[X] = np(1-p) \qquad (15.50)$$

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:

$$\underbrace{(p\cdots p)}_{\text{x-times}} \cdot \underbrace{(q\cdots q)}_{n-x\text{-times}} = p^x q^{n-x}$$

But there exists obviously further realization $X = x$, that correspond to permutations of the n-drawn balls.
There exist two classes of $n_1 = x$-white and $n_2 = (n-x)$ black balls s.t.

$$P(n;n_1,n_2) = \frac{n!}{x!(n-x)!} = \binom{n}{x}$$

4.3. Geometric Distribution $\mathrm{Geom}(p)$

Definition 15.24 Geometric Distribution$\mathrm{Geom}(p)$: Models the probability of the number X of Bernoulli trials[def. 15.21] *until the first success*

$$p(x) = p(1-p)^{x-1}$$

x :nb. of repetitions *until first success*
p :success probability *of single Bernoulli experiment*

$$F(x) = \sum_{i=1}^{x} p(1-p)^{i-1} \overset{??}{=} 1-(1-p)^x$$

$$\mathbb{E}[X] = \frac{1}{p} \qquad (15.51) \qquad \mathbb{V}[X] = \frac{1-p}{p^2} \qquad (15.52)$$

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 $p(x)$:
$$x \geqslant \frac{p(x)}{1-p}$$
- $\log(1-p) \approx -p$ for small p

4.4. Poisson Distribution $\mathrm{Pois}(\lambda)$

Definition 15.25 Poisson Distribution: Is an extension of the binomial distribution, where the realization x of the random variable X may attain values in $\mathbb{Z}_{\geqslant 0}$.
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.

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad \begin{matrix} \lambda > 0 \\ x \in \mathbb{Z}_{\geqslant 0} \end{matrix} \qquad (15.53)$$

Event Rate λ: describes the average number of events in a single interval.

$$\mathbb{E}[X] = \lambda \qquad (15.54) \qquad \mathbb{V}[X] = \lambda \qquad (15.55)$$

Continuous Distributions

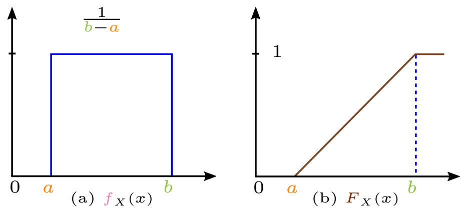
5.1. Uniform Distribution $\mathcal{U}(a,b)$

Definition 15.26 Uniform Distribution $\mathcal{U}(a,b)$: Is probability distribution, where all intervals of the **same** length on the distribution's support ([def. 15.6]) $\mathrm{supp}(\mathcal{U}[a,b]) = [a,b]$ are equally probable/likely.

$$f(x) = \frac{1}{b-a}\mathbb{1}_{x\in[a;b)} = \begin{cases} \frac{1}{b-a} = \text{const} & \text{if} \quad a \leqslant x \leqslant b \\ 0 & \quad\;\;\, \text{else} \end{cases} \qquad (15.56)$$

$$F(x) = \begin{cases} 0 & & x < a \\ \frac{x-a}{b-a} & \text{if} & a \leqslant x \leqslant b \\ 1 & & x > b \end{cases} \qquad (15.57)$$

$$\mathbb{E}[X] = \frac{a+b}{2} \qquad \mathbb{V}(X) = \frac{(b-a)^2}{12} \qquad (15.58)$$



(a) $f_X(x)$ (b) $F_X(x)$

5.2. Exponential Distribution $\exp(\lambda)$

Definition 15.27 Exponential Distribution $X \sim \exp(\lambda)$: Is the continuous analogue to the geometric distribution [def. 15.24].

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.

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if} \quad x \geqslant 0 \\ 0 & \quad\;\;\, x < 0 \end{cases} \qquad (15.59)$$

$$F(x;\lambda) = \begin{cases} 1-e^{-\lambda x} & \text{if} \quad x \geqslant 0 \\ 0 & \quad\;\;\, x < 0 \end{cases} \qquad (15.60)$$

$$\mathbb{E}[X] = \frac{1}{\lambda} \qquad \mathbb{V}(X) = \frac{1}{\lambda^2} \qquad (15.61)$$

5.3. Laplace Distribution

Definition 15.28 Laplace Distribution:

Laplace Distibution $\qquad f(\boldsymbol{x};\mu,\sigma) = \frac{1}{2\sigma}\exp\left(-\frac{|\boldsymbol{x}-\mu|}{\sigma}\right)$ (15.62)

5.4. The Normal Distribution $\mathcal{N}(\mu,\sigma)$

Definition 15.29 Normal Distribution $\boldsymbol{X} \sim \mathcal{N}(\mu,\sigma^2)$: Is a symmetric distribution where the population parameters μ, σ^2 are equal to the expectation and variance of the distribution:

$$\mathbb{E}[X] = \mu \qquad \mathbb{V}(X) = \sigma^2 \qquad (15.63)$$

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \qquad (15.64)$$

$$F(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left\{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right\}du \qquad (15.65)$$

$$x \in \mathbb{R} \quad \text{or} \quad -\infty < x < \infty$$

$$\varphi_X(u) = \exp\left\{iu\mu - \frac{u^2\sigma^2}{2}\right\} \qquad (15.66)$$

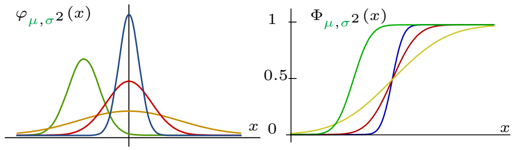


Figure 3: $\mu = 0$, $\sigma^2 = 0.2$ $\quad\mu = 0$, $\sigma^2 = 1.0$ $\quad\mu = 0$, $\sigma^2 = 5.0$ $\quad\mu = -2$, $\sigma^2 = 0.5$

Property 15.14: $\mathbb{P}_X(\mu-\sigma \leqslant x \leqslant \mu-\sigma) = 0.66$

Property 15.15: $\mathbb{P}_X(\mu-2\sigma \leqslant x \leqslant \mu-2\sigma) = 0.95$

5.5. The Standard Normal distribution $\mathcal{N}(0,1)$

Historic Problem: the cumulative distribution eq. (15.65) 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 15.30
Standard Normal Distribution $\boldsymbol{X} \sim \mathcal{N}(0,1)$:

$$\mathbb{E}[X] = 0 \qquad \mathbb{V}(X) = 1 \qquad (15.67)$$

$$f(x;0,1) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} \qquad (15.68)$$

$$F(x;0,1) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{1}{2}u^2}\,du \qquad (15.69)$$

$$x \in \mathbb{R} \quad \text{or} \quad -\infty < x < \infty$$

$$\psi_X(u) = e^{\frac{u^2}{2}} \qquad \varphi_X(u) = e^{-\frac{u^2}{2}} \qquad (15.70)$$

Corollary 15.3
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. 15.19] is then denoted by:

$$z_\alpha = \Phi^{-1}(\alpha) \qquad \alpha \in (0,1) \qquad (15.71)$$

5.5.1. Calculating Probabilities

Property 15.16 Symmetry: Let $z > 0$

$$\mathbb{P}(Z \leqslant z) = \Phi(z) \qquad (15.72)$$
$$\mathbb{P}(Z \leqslant -z) = \Phi(-z) = 1-\Phi(z) \qquad (15.73)$$
$$\mathbb{P}(-a \leqslant Z \leqslant b) = \Phi(b) - \Phi(-a) = \Phi(b) - (1-\Phi(a))$$
$$\overset{a=b=z}{=} 2\Phi(z) - 1 \qquad (15.74)$$

5.5.2. Linear Transformations of Normal Dist.

Proposition 15.1Linear Transformation: Let X be a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then the linear transformed r.v. $Y = a + bX$ is distributed as:

$$Y \sim \mathcal{N}\left(a + b\mu, b^2\sigma^2\right) \iff f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right) \tag{15.75}$$

section 1

Proposition 15.2Standardization: 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:

$$X \sim \mathcal{N}(\mu, \sigma^2) \xrightarrow{Z = \frac{X-\mu}{\sigma}} Z \sim \mathcal{N}(0, 1) \tag{15.76}$$

section 1

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 15.3Standardization of the CDF: 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:

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{15.77}$$

section 1

6. The Multivariate Normal distribution

Definition 15.31
Multivariate Normal distribution $\boldsymbol{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$:
The k-multivariate Normal distribution of:

$\boldsymbol{X} = \begin{pmatrix} x_1 & \dots & x_k \end{pmatrix}^\intercal$ a k-dimensional random vector with:

$\boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}[\boldsymbol{x}_1] & \dots & \mathbb{E}[\boldsymbol{x}_k] \end{pmatrix}^\intercal$ a k-dim mean vector

and $k \times k$ **p.s.d**.covariance matrix:

$$\boldsymbol{\Sigma} := \mathbb{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\intercal] = [\mathrm{Cov}[\boldsymbol{x}_i, \boldsymbol{x}_j], 1 \leqslant i, j \leqslant k]$$

is given by:

$$f_{\boldsymbol{X}}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k) = \tag{15.78}$$

$$= \underbrace{\frac{1}{\sqrt{(2\pi)^k \det(\boldsymbol{\Sigma})}}}_{\text{Normalisation}} \exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\right)$$

Definition 15.32 Jointly Gaussian Random Variables: Two random variables $\boldsymbol{U}$, $\boldsymbol{V}$ both scalars or vectors, are said to be jointly Gaussian if the joint vector random variable $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{V} \end{bmatrix}^\intercal$ is again a GRV.

Corollary 15.4 Jointly GRV of GRVs: If $\boldsymbol{x}$ and $\boldsymbol{y}$ are both independent GRVs $\boldsymbol{x} \sim \mathcal{N}(\mu_x, \Sigma_x)$, $\boldsymbol{y} \sim \mathcal{N}(\mu_y, \Sigma_y)$, **then** they are jointly Gaussian ([def. 15.32]).

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{p}(x)\mathrm{p}(y) \tag{15.79}$$

$$\propto \exp\left(-\frac{1}{2}\left\{(\boldsymbol{x} - \boldsymbol{\mu}_x)^\intercal \boldsymbol{\Sigma}_x^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_x) + (\boldsymbol{y} - \boldsymbol{\mu}_y)^\intercal \boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y)\right\}\right)$$

$$= \exp\left(-\frac{1}{2}\begin{bmatrix}(\boldsymbol{x} - \boldsymbol{\mu}_x)^\intercal & (\boldsymbol{y} - \boldsymbol{\mu}_y)^\intercal\end{bmatrix}\begin{bmatrix} 0 & \boldsymbol{\Sigma}_x^{-1} \\ \boldsymbol{\Sigma}_y^{-1} & 0 \end{bmatrix}\begin{bmatrix}\boldsymbol{x} - \boldsymbol{\mu}_x \\ \boldsymbol{y} - \boldsymbol{\mu}_y\end{bmatrix}\right)$$

Property 15.17 Scalar Affine Transformation of GRVs: **Let** $\boldsymbol{y} \in \mathbb{R}^n$ be GRV, $a \in \mathbb{R}_+$, $b \in \mathbb{R}$ and let $\boldsymbol{x}$ be defined by the affine transformation ([def. 8.1]):

$$\boldsymbol{x} = a\boldsymbol{y} + b \qquad a \in \mathbb{R}_+, b \in \mathbb{R}^d$$

Then $\boldsymbol{x}$ is a GRV with:

$$\boxed{x \sim \mathcal{N}(\mu_x, \sigma_x^2) = \mathcal{N}(a\mu + b, a^2\sigma^2)} \tag{15.80}$$

Property 15.18 Affine Transformation of GRVs: Let $\boldsymbol{y} \in \mathbb{R}^n$ be GRV, $\boldsymbol{A} \in \mathbb{R}^{d \times n}$, $b \in \mathbb{R}^d$ and let $\boldsymbol{x}$ be defined by the affine transformation[def. 8.1]:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{y} + b \qquad \boldsymbol{A} \in \mathbb{R}^{d \times n}, b \in \mathbb{R}^d$$

Then $\boldsymbol{x}$ is a GRV (see Section 1).

Property 15.19 Linear Combination of jointly GRVs: Let $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^m$ two jointly GRVs, and let $\boldsymbol{z}$ be defined as:

$$\boldsymbol{z} = \boldsymbol{A}_x\boldsymbol{x} + \boldsymbol{A}_y\boldsymbol{y} \qquad \boldsymbol{A}_x \in \mathbb{R}^{d \times n}, \boldsymbol{A}_x \in \mathbb{R}^{d \times m}$$

Then $\boldsymbol{z}$ is GRV (see Section 1).

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.

Diagonal Covariance Matrix

For i.i.d. data the covariance matrix becomes diagonal:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k^2 \end{bmatrix} \quad \text{and} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} \tag{15.81}$$

eq. (15.78) decomposed s.t. $x_1, \dots, x_k$ become mutal independent (**??**):

$$\mathrm{p}(\boldsymbol{X}) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right) \tag{15.82}$$

6.1. Gamma Distribution $\Gamma(x, \alpha, \beta)$

Definition 15.33 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:

$$\mathrm{f}(x; \alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\mathrm{e}^{-\beta x} & \text{if} \quad x > 0 \\ 0 & \phantom{\text{if}} \quad x \leqslant 0 \end{cases} \tag{15.83}$$

$$\Gamma(\alpha) \overset{\text{eq. (5.65)}}{=} \int_0^\infty t^{\alpha-1}\mathrm{e}^{-t}\,\mathrm{d}t \tag{15.84}$$

with $\alpha, \beta \in \mathbb{R}_{>0}$

7. Student's t-distribution

Definition 15.34 Student' t-distribution:

add

7.1. Delta Distribution

Definition 15.35 The delta function $\delta(\boldsymbol{x})$: The delta/dirac function $\delta(\boldsymbol{x})$ is defined by:

$$\int_{\mathbb{R}} \delta(\boldsymbol{x}) f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = f(0)$$

for any integrable function f on $\mathbb{R}$.
Or alternativly by:

$$\delta(x - x_0) = \lim_{\sigma \to 0} \mathcal{N}(x | x_0, \sigma) \tag{15.85}$$

$$\approx \infty \mathbb{1}_{\{x = x_0\}} \tag{15.86}$$

Property 15.20 Properties of δ:

- **Normalization**: The delta function integrates to 1:
$$\int_{\mathbb{R}} \delta(x)\,\mathrm{d}x = \int_{\mathbb{R}} \delta(x) \cdot c_1(x)\,\mathrm{d}x = c_1(0) = 1$$
where $c_1(x) = 1$ is the constant function of value 1.
- **Shifting**:
$$\int_{\mathbb{R}} \delta(x - x_0) f(x)\,\mathrm{d}x = f(x_0) \tag{15.87}$$
- **Symmetry**: $\int_{\mathbb{R}} \delta(-x) f(x)\,\mathrm{d}x = f(0)$
- **Scaling**: $\int_{\mathbb{R}} \delta(\alpha x) f(x)\,\mathrm{d}x = \frac{1}{|\alpha|} f(0)$

Note

- In mathematical terms δ is not a function but a **gernalized function**.
- We may regard $\delta(x - x_0)$ as a density with all its probability mass centered at the signle point x_0.
- Using a box/indicator function s.t. its surface is one and its width goes to zero, instead of a normaldistribution eq. (15.85) would be a non-differentiable/discret form of the dirac measure.

Proofs

Proof. proposition 15.1: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$F_Y(y) \overset{y>0}{=} \mathbb{P}_Y(Y \leqslant y) = \mathbb{P}(a + bX \leqslant y) = \mathbb{P}_X\left(X \leqslant \frac{y-a}{b}\right)$$

$$= F_X\left(\frac{y-a}{b}\right)$$

$$F_Y(y) \overset{y<0}{=} \mathbb{P}_Y(Y \leqslant y) = \mathbb{P}(a + bX \leqslant y) = \mathbb{P}_X\left(X \geqslant \frac{y-a}{b}\right)$$

$$= 1 - F_X\left(\frac{y-a}{b}\right)$$

Differentiating both expressions w.r.t. y leads to:

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = \begin{cases} \frac{1}{b}\frac{\mathrm{d}F_X\left(\frac{y-a}{b}\right)}{\mathrm{d}y} \\ \frac{1}{-b}\frac{\mathrm{d}F_X\left(\frac{y-a}{b}\right)}{\mathrm{d}y} \end{cases} = \frac{1}{|b|} f_X(x)\left(\frac{y-a}{b}\right)$$

eq. (15.75)).
in order to prove that $Y \sim \mathcal{N}\left(a + b\mu, b^2\sigma^2\right)$ 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. proposition 15.2: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$Z := \frac{X-\mu}{\sigma} = \frac{1}{std}X - \frac{\mu}{\sigma} = aX + b \quad \text{with } a = \frac{1}{\sigma}, b = -\frac{\mu}{\sigma}$$

$$\overset{\text{eq. (15.75)}}{\sim} \mathcal{N}\left(a\mu + b, a^2\sigma^2\right) \sim \mathcal{N}\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{\sigma^2}{\sigma^2}\right) \sim \mathcal{N}(0, 1)$$

□

Proof. proposition 15.3: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$F_X(x) = \mathbb{P}(X \leqslant x) \overset{\substack{-\mu \\ \div\sigma}}{=} \mathbb{P}\left(\frac{X-\mu}{\sigma} \leqslant \frac{x-\mu}{\sigma}\right) \mathbb{P}\left(Z \leqslant \frac{x-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{x-\mu}{\sigma}\right)$$

□

Proof. Property 15.18 scalar case

Let $y \sim \mathrm{p}(y) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$ and define $\boldsymbol{x} = ay + b \qquad a \in \mathbb{R}_+, b \in \mathbb{R}$
Using the Change of variables formula it follows:

$$\mathrm{p}_x(\bar{x}) \overset{??}{=} \frac{\mathrm{p}_y(\bar{y})}{|\frac{\mathrm{d}x}{\mathrm{d}y}|} \overset{\bar{y} = \frac{\bar{x}-b}{a}}{=} \frac{1}{a}\frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{1}{2\sigma^2}\left(\overbrace{\frac{\bar{x} - b}{a}}^{\bar{y}(\bar{x})} - \mu\right)^2\right)$$

$$= \frac{1}{\sqrt{2\pi a^2\mu^2}} \exp\left(-\frac{1}{2\sigma^2 a^2}\left(\bar{x}\underbrace{-b - a\mu}_{\mu_x}\right)^2\right)$$

Hence $\quad x \sim \mathcal{N}(\mu_x, \sigma_x^2) = \mathcal{N}(a\mu + b, a^2\sigma^2)$

□

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. Property 15.19
From Property 15.18 it follows immediately that $\boldsymbol{z}$ is GRV $\boldsymbol{z} \sim \mathcal{N}(\mu_z, \Sigma_z)$ with:

$$\boldsymbol{z} = \boldsymbol{A}\xi \quad \textbf{with} \quad \boldsymbol{A} = \begin{bmatrix}\boldsymbol{A}_x & \boldsymbol{A}_y\end{bmatrix} \text{ and } \xi = \begin{pmatrix}\boldsymbol{x} & \boldsymbol{y}\end{pmatrix}$$

Knowing that $\boldsymbol{z}$ is a GRV it is sufficient to calculate μ_z and Σ_z in order to characterize its distribution:

$$\mathbb{E}[\boldsymbol{z}] = \mathbb{E}\left[\boldsymbol{A}_x x + \boldsymbol{A}_y y\right] = \boldsymbol{A}_x\mu_x + \boldsymbol{A}_y\mu_y$$

$$\mathbb{V}[\boldsymbol{z}] = \mathbb{V}[\boldsymbol{A}\xi] \overset{??}{=} \boldsymbol{A}\mathbb{V}[\xi]\boldsymbol{A}^\intercal$$
$$= \begin{bmatrix}\boldsymbol{A}_x & \boldsymbol{A}_y\end{bmatrix}\begin{bmatrix}\mathbb{V}[x] & \mathrm{Cov}[x, y] \\ \mathrm{Cov}[y, x] & \mathbb{V}[y]\end{bmatrix}\begin{bmatrix}\boldsymbol{A}_x & \boldsymbol{A}_y\end{bmatrix}^\intercal$$
$$= \begin{bmatrix}\boldsymbol{A}_x & \boldsymbol{A}_y\end{bmatrix}\begin{bmatrix}\mathbb{V}[x] & \mathrm{Cov}[x, y] \\ \mathrm{Cov}[y, x] & \mathbb{V}[y]\end{bmatrix}\begin{bmatrix}\boldsymbol{A}_x^\intercal \\ \boldsymbol{A}_y^\intercal\end{bmatrix}$$
$$= \boldsymbol{A}_x\mathbb{V}[x]\boldsymbol{A}_x^\intercal + \boldsymbol{A}_y\mathbb{V}[y]\boldsymbol{A}_y^\intercal$$
$$+ \underbrace{\boldsymbol{A}_y\mathrm{Cov}[y, x]\boldsymbol{A}_x^\intercal}_{=0\text{by independence}} + \underbrace{\boldsymbol{A}_x\mathrm{Cov}[x, y]\boldsymbol{A}_y^\intercal}_{=0\text{by independence}}$$
$$= \boldsymbol{A}_x\Sigma_x\boldsymbol{A}_x^\intercal + \boldsymbol{A}_y\Sigma_y\boldsymbol{A}_y^\intercal$$

□

Note

Can also be proofed by using the normal definition of [def. 15.15] and tedious computations.

8. Sampling Random Numbers

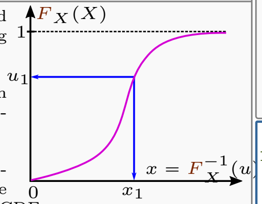
Most math libraries have uniform **random number generator** (**RNG**) i.e. functions to generate uniformly distributed random numbers $U \sim \mathcal{U}[a, b]$ (eq. (15.56)).
Furthermore repeated calls to these RNG are independent, that is:

$$\mathrm{p}_{U_1,U_2}(u_1, u_2) \overset{??}{=} \mathrm{p}_{U_1}(u_1) \cdot \mathrm{p}_{U_2}(u_2) = \begin{cases} 1 & \text{if } u_1, u_2 \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

Question: using samples $\{u_1, \ldots, u_n\}$ of these CRVs with uniform distribution, how can we create random numbers with arbitrary discreet or continuous PDFs?

9. Inverse-transform Technique

Idea



Can make use of section 1 and the fact that CDF are increasing functions ([def. 5.8]). **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.

9.1. Continuous Case

Definition 15.36 One Continuous Variable: Given: a desired continuous pdf f_X and uniformly distributed rn $\{u_1, u_2, \ldots\}$:

1. Integrate the desired pdf f_X in order to obtain the desired cdf F_X:
$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \qquad (15.88)$$
2. Set $F_X(X) \overset{!}{=} U$ 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:
$$U = F_X(X) = F_X\Big(\underbrace{F_X^{-1}(U)}_{X}\Big) \qquad (15.89)$$
4. Plug in the uniformly distributed rn:
$$x_i = F_X^{-1}(u_i) \qquad \textbf{s.t.} \qquad x_i \sim f_X \qquad (15.90)$$

Definition 15.37 Multiple Continuous Variable:
Given: a pdf of multiple rvs $f_{X,Y}$:

1. Use the product rule (**??**) in order to decompose $f_{X,Y}$:
$$f_{X,Y} = f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y) \qquad (15.91)$$
2. Use [def. 15.38] to first get a rv for y of $Y \sim f_Y(y)$.
3. Then with this fixed y use [def. 15.38] again to get a value for x of $X \sim f_{X|Y}(x|y)$.

Proof. [def. 15.38]:
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. 5.8]).
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:

$$F_X(a) = \mathbb{P}(\underline{x} \leqslant a) = \mathbb{P}(F_X^{-1}(U) \leqslant a)$$

Since F_X is strictly increasing:

$$\mathbb{P}\left(F_X^{-1}(U) \leqslant a\right) = \mathbb{P}(U \leqslant F_X(a)) \overset{\text{eq. (15.56)}}{=} \int_0^{F_X(a)} 1\,\mathrm{d}t = F_X(a)$$

□

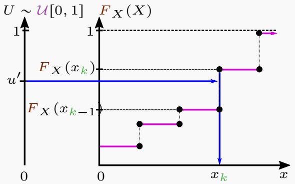
Note

Strictly speaking we may not assume that a CDF is strictly increasing but we as all CDFs are weakly increasing ([def. 5.8]) we may always define an auxiliary function by its infinimum:

$$\hat{F}_X^{-1} := \inf\{x | F_X(X) \geqslant 0\} \qquad u \in [0, 1] \qquad (15.92)$$

9.2. Discret Case

Idea



Given: a desired discret pmf p_X s.t. $\mathbb{P}(X = x_i) = p_X(x_i)$ and uniformly distributed rn $\{u_1, u_2, \ldots\}$.
Goal: given a uniformly distributed rn u determine k s.t.:

$$\sum_{i=1}^{k-1} < U \leqslant \sum_{i=1}^{k} \iff F_X(x_{k-1}) < u \leqslant F_X(x_k) \qquad (15.93)$$

and return x_k.

Definition 15.38 One Discret Variable:

1. Compute the CDF of p_X ([def. 15.8])
$$F_X(x) = \sum_{t=-\infty}^{x} \mathrm{p}_X(t) \qquad (15.94)$$
2. Given the uniformly distributed rn $\{u_i\}_{i=1}^{n}$ find k^i (≙ inversion) s.t.:
$$F_X\left(x_{k(i)-1}\right) < u_i \leqslant F_X\left(x_{k(i)}\right) \qquad \forall u_i \qquad (15.95)$$

Proof. **??**: 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:

$$F_X(x_{k-1}) < u \leqslant F_X(x_k) \qquad (15.96)$$

Now observe that:

$$u \leqslant F_X(x_k) = F_X(x_{k-1}) + \mathrm{p}_X(x_k)$$
$$\Rightarrow F_X(x_{k-1}) < u \leqslant F_X(x_{k-1}) + \mathrm{p}_X(x_k)$$

The probability of U being in $(F_X(x_{k-1}), F_X(x_k)]$ is:

$$\mathbb{P}(U \in [F_X(x_{k-1}), F_X(x_k)]) = \int_{F_X(x_{k-1})}^{F_X(x_k)} \mathrm{p}_U(t)\,\mathrm{d}t$$
$$= \int_{F_X(x_{k-1})}^{F_X(x_k)} 1\,\mathrm{d}t = \int_{F_X(x_{k-1})}^{F_X(x_{k-1})+\mathrm{p}_X(x_k)} 1\,\mathrm{d}t = \mathrm{p}_X(x_k)$$

Hence the random variable $x_k \in \mathcal{X}$ has the pdf p_X. □

Definition 15.39
Multiple Continuous Variables (Option 1):
Given: a pdf of multiple rvs $\mathrm{p}_{X,Y}$:

1. Use the product rule (**??**) in order to decompose $\mathrm{p}_{X,Y}$:
$$\mathrm{p}_{X,Y} = \mathrm{p}_{X,Y}(x, y) = \mathrm{p}_{X|Y}(x|y)\mathrm{p}_Y(y) \qquad (15.97)$$
2. Use **??** to first get a rv for y of $Y \sim \mathrm{p}_Y(y)$.
3. Then with this fixed y use **??** again to get a value for x of $X \sim \mathrm{p}_{X|Y}(x|y)$.

Definition 15.40
Multiple Continuous Variables (Option 2):
Note: this only works if $\mathcal{X}$ and $\mathcal{Y}$ are finite.
Given: a pdf of multiple rvs $\mathrm{p}_{X,Y}$ **let** $N_x = |\mathcal{X}|$ and $N_y = |\mathcal{Y}|$ the number of elements in $\mathcal{X}$ and $\mathcal{Y}$.

Define
$$\mathrm{p}_Z(1) = \mathrm{p}_{X,Y}(1, 1), \mathrm{p}_Z(2) = \mathrm{p}_{X,Y}(1, 2), \ldots$$
$$\ldots, \mathrm{p}_Z(N_x \cdot N_y) = \mathrm{p}_{X,Y}(N_x, N_y)$$

Then simply apply **??** to the auxillary pdf p_Z

1. Use the product rule (**??**) in order to decompose $f_{X,Y}$:
$$f_{X,Y} = f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y) \qquad (15.98)$$
2. Use [def. 15.38] to first get a rv for y of $Y \sim f_Y(y)$.
3. Then with this fixed y use [def. 15.38] again to get a value for x of $X \sim f_{X|Y}(x|y)$.

nice examples see comment in code text

10. Descriptive Statistics

10.1. Population Parameters

Definition 15.41 Population/Statistical Parameter: Are parameters defining families of probability distributions and thus characteristics of population following such distributions i.e. the normal distribution has two parameters $\left\{\mu, \sigma^2\right\}$

Definition 15.42 Population Mean: Given a population $\{x_i\}_{i=1}^N$ of size N its variance is defined as:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{15.99}$$

Definition 15.43 Population Variance: Given a population $\{x_i\}_{i=1}^N$ of size N its variance is defined as: $\{x_i\}_{i=1}^N$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2 \tag{15.100}$$

Note

The population variance and mean are equally to the mean derived from the true distribution of the population.

10.2. Sample Estimates

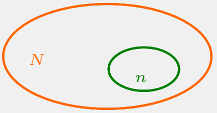
Definition 15.44 (Sample) Statistic: A statistc is a measuarble function f that assigns a **single** value F to a sample of random variables or population:

$f : \mathbb{R}^n \mapsto \mathbb{R} \qquad F = f(X_1, \dots, X_n)$

E.g. F could be the mean, variance,...

Note

The function itself is independent of the sample's distribution; that is, the function can be stated before realization of the data.



Definition 15.45 (Point) Estimator $\hat{\theta} = \hat{\theta}(\boldsymbol{X})$:
Given: n-samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \sim \boldsymbol{X}$ an estimator

$$\hat{\theta} = h(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \tag{15.101}$$

is a statistic/randomn variable used to estimate a true (population) parameter θ[def. 15.41].

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).

Definition 15.46 Degrees of freedom of a Statistic: Is the number of values in the final calculation of a statistic that are free to vary.

10.2.1. Empirical Mean

Definition 15.47 Sample/Empirical Mean $\bar{x}$: The sample mean is an estimate/statistic of the population mean[def. 15.42] and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$\bar{x} = \hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{15.102}$$

Corollary 15.5 Expectation: The sample mean estimator is unbiased (see section 14):

$$\mathbb{E}\left[\hat{\mu}_X\right] = \mu \tag{15.103}$$

Corollary 15.6 Variance: For the variance of the sample mean estimator it holds (see section 14):

$$\mathbb{V}\left[\hat{\mu}_X\right] = \frac{1}{n}\sigma_X^2 \tag{15.104}$$

10.2.2. Empirical Variance

Definition 15.48 Biased Sample Variance: The sample mean is an estimate/statistic of the population variance[def. 15.43] and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$s_n^2 = \hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2 \tag{15.105}$$

Definition 15.49 (Unbiased) Sample Variance:

$$s^2 = \hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2 \tag{15.106}$$

see section 14

Definition 15.50 Bessel's Correction: The factor

$$\frac{n}{n-1} \tag{15.107}$$

as multiplying the uncorrected population varianceeq. (15.105) by this term yields an unbiased estimated of the variance (not the standard deviation). The reason for this is that are

Attention: Usually only unbiased variance is used and also sometimes denoted by s_n^2

Proof.
finish this
□

11. Statistical Tests

Definition 15.51 Null Hypothesis: A Null Hypothesis H_0 is usually a commonly accepted fact/view/base hypothesis that researchers try to nullify or disprove.

$$H_0 : \theta = \theta_0 \tag{15.108}$$

Definition 15.52 Alternative Hypothesis: The Alternative Hypothesis H_A/H_1 is the opposite of the Null Hypotheses/contradicts it and is what we try to test against the Null Hypothesis.

$$H_A : \theta \begin{cases} > \theta_0 & \text{(one-sided)} \\ < \theta_0 & \text{(one-sided)} \\ \neq \theta_0 & \text{(two-sided)} \end{cases} \tag{15.109}$$

Definition 15.53 Testing Parameters:
Given: a parameter θ that we want to test.
Let Θ be the set of all possible values that θ can achieve.
We now split Θ in two disjunct sets Θ_0 and Θ_1.

$$\Theta = \Theta_0 \cup \Theta_1 \qquad \Theta_0 \cap \Theta_1 = \varnothing$$

Null Hypothesis $\qquad H_0 : \theta \in \Theta_0 \qquad$ (15.110)
Alternative Hypothesis $\qquad H_A : \theta \in \Theta_1 \qquad$ (15.111)

11.1. Type I&II Errors

Definition 15.54 Type I Error: Is the rejection of a Null Hypothesis, even-tough its true (also known as a "false positive").

Definition 15.55 Type II Error: Is the acceptance of a Null Hypothesis, even-tough its false (also known as a "false negative").

Decision	H_0 **true**	H_0 **false**
Accept	**TN**	**Type II (FN)**
Reject	**Type I (FP)**	**TP**

Definition 15.56 Critical Value c: Value from which on the Null-hypothesis H_0 gets rejected.

Definition 15.57 Statistical significance α: A study's defined significance level, denoted α, is the **probability** of the study rejecting the null hypothesis, given that the null hypothesis were true (Type I Error).

Definition 15.58 Critical Region K_α: Is the set of all values that causes us to reject the Null Hypothesis in favor for the Alternative Hypothesis H_A.
The Critical region is usually chosen s.t. we incur a Type I Error with probability less than α.

$$\boxed{K_\alpha \in \Theta : \mathbb{P}(\text{Type I Error}) \leqslant \alpha} \tag{15.112}$$

$\mathbb{P}(c_2 \leqslant X \leqslant c_1) \leqslant \alpha \qquad$ two-sided
or $\quad \mathbb{P}(c_2 \leqslant X) \leqslant \frac{\alpha}{2} \quad$ and $\quad \mathbb{P}(X \leqslant c_1) \leqslant \frac{\alpha}{2}$
$\mathbb{P}(c_2 \leqslant X) \leqslant \alpha \qquad$ one-sided
$\mathbb{P}(X \leqslant c_1) \leqslant \alpha \qquad$ one-sided

Definition 15.59 Acceptance Region: Is the region where we accept the null hypothesis H_0.

Note

see example 15.3.

11.2. Normally Distributed Data

Let us consider a sample of $\{x_i\}_{i=1}^n$ i.i.d. observations, that follow a normal distribution $x_i \sim \mathcal{N}(\mu, \sigma^2)$.

11.2.1. Z-Test $\qquad \sigma$ known

11.2.2. t-Test $\qquad \sigma$ unknown

12. Inferential Statistics

Goal of Inference

① What is a good guess of the parameters of my model?
② How do I quantify my uncertainty in the guess?

13. Examples

Example 15.1 ??: Let x be uniformly distributed on $[0,1]$ ([def. 15.26]) with pmf $p_X(x)$ then it follows:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{p_Y(y)} \Rightarrow \mathrm{d}x = \mathrm{d}y p_y(y) \Rightarrow x = \int_{-\infty}^{y} p_y(t)\,\mathrm{d}t = F_Y(x)$$

Example 15.2 ??: Let

add https://www.youtube.com/watch?v=WUUb7VIRzgg

Example 15.3 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 the what the manufacture claims is true, so we want to test:

$$H_0 : p = p_0 = 0.1 \qquad \text{vs.} \qquad H_A : p > 0.1$$

We model the number of number of defective goods using the binomial distribution[def. 15.23]

$$X \sim \mathcal{B}(n,p), n = 20 \qquad \mathbb{P}(X \geqslant x) = \sum_{k=x}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

from this we find:

$$\mathbb{P}_{p_0}(X \geqslant 4) = 1 - \mathbb{P}_{p_0}(X \leqslant 3) = 0.13$$
$$\mathbb{P}_{p_0}(X \geqslant 4) = 1 - \mathbb{P}_{p_0}(X \leqslant 3) = 0.04 \leqslant \alpha$$

thus the probability that equal 5 or more then 5 parts out of the 20 are rejects is less then 4%.
⇒ throw away null hypothesis for the 5% niveau in favor to the alternative.
⇒ the 5% significance niveau is given by $K = \{5, 6, \ldots, 20\}$

Note

If $x < n/2$ it is faster to calculate $\mathbb{P}(X \geqslant x) = 1 - \mathbb{P}(X \leqslant x-1)$

14. Proofs

Proof. corollary 15.5:

$$\mathbb{E}[\hat{\mu}_X] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\mathbb{E}[\underbrace{\mu + \cdots + \mu}_{1,\ldots,n}]$$

□

Proof. corollary 15.6:

$$\mathbb{V}[\hat{\mu}_X] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \overset{\text{Property 15.10}}{=} \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} x_i\right]$$
$$\frac{1}{n^2} n \mathbb{V}[X] = \frac{1}{n}\sigma^2$$

□

Proof. definition 15.49:

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}_X^2\right] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\cdot n\bar{x} + n\bar{x}^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\mathbb{E}\left[x_i^2\right] - n\mathbb{E}\left[\bar{x}^2\right]\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\mathbb{E}\left[\bar{x}^2\right]\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{1}{n}\sigma^2 + \mu^2\right)\right] \\
&= \frac{1}{n-1}\left[(n\sigma^2 + n\mu^2) - \left(\sigma^2 + n\mu^2\right)\right] \\
&= \frac{1}{n-1}\left[n\sigma^2 - \sigma^2\right] = \frac{1}{n-1}\left[(n-1)\sigma^2\right] = \sigma^2
\end{aligned}$$

□

Stochastic Calculus

Stochastic Processes

Definition 16.1
Random/Stochastic Process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$**:**
is a collection of random variables 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:

$$X : \mathbb{R} \times \Omega \mapsto \mathbb{R} \qquad \Longleftrightarrow \qquad (t, \omega) \mapsto X(t, \omega) \tag{16.1}$$

Definition 16.2 Sample path/Trajector/Realization: Is the *stochastic/noise signal* $r(\cdot, \omega)$ 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 $X(t, \omega) := X(t)$

Definition 16.3 Filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t\geqslant 0}$**:**
A collection $\{\mathcal{F}_t\}_{t\geqslant 0}$ of sub σ-*algebras* $\{\mathcal{F}_t\}_{t\geqslant 0} \in \mathcal{F}$ is called filtration if is *increasing*:

$$\mathcal{F}_s \subseteq \mathcal{F}_t \qquad \forall s \leqslant t \tag{16.2}$$

Definition 16.4 Adapted Process: A stochastic process $\{X_t : 0 \leqslant t \leqslant \infty\}$ is called adapted to a filtration $\mathbb{F}$ if, X_t is $\mathcal{F}_t$-measurable, i.e. observable at time t.

Definition 16.5 Predictable Process: A stochastic process $\{X_t : 0 \leqslant t \leqslant \infty\}$ is called predictable w.r.t. a filtration $\mathbb{F}$ if, X_t is $\{\mathcal{F}_{t-1}\}$-measurable, i.e. 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.

Definition 16.6
Filtered Probability Space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant 0}, \mathbb{P})$**:**
A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ together with a filtration $\{\mathcal{F}_t\}_{t\geqslant 0}$ is called a *filtered probability* space.

Corollary 16.1 : The amount of information of an adapted random process is increasing see example 16.1.

Definition 16.7 Martingales: A stochastic process $X(t)$ is a martingale on a *filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant 0}, \mathbb{P})$ if the following conditions hold:

1. Given $s \leqslant t$ the best prediction of $X(t)$, with a filtration $\{\mathcal{F}_s\}$ is the current expected value:
$$\forall s \leqslant t \qquad \mathbb{E}\left[X(t)|\mathcal{F}_s\right] = X(s) \quad \text{a.s.} \tag{16.3}$$
2. The expectation is finite:
$$\mathbb{E}\left[|X(t)|\right] < \infty \quad \forall t \geqslant 0 \qquad X(t) \text{ is } \{\mathcal{F}_t\}_{t\geqslant 0} \text{ adapted} \tag{16.4}$$

Interpretation

- 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 16.8 Auto Covariance $\gamma(t_2 - t_1)$**:**
Describes the covariance[def. 15.16] between two values of a stochastic process $(\boldsymbol{X}_t)_{t\in\mathcal{T}}$ at different time points t_1 and t_2.

$$\gamma(t_1, t_2) = \operatorname{Cov}\left[\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}\right] = \mathbb{E}\left[\left(\boldsymbol{X}_{t_1} - \mu_{t_1}\right)\left(\boldsymbol{X}_{t_2} - \mu_{t_2}\right)\right] \tag{16.5}$$

For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance:

$$\gamma(t, t) = \operatorname{Cov}\left[\boldsymbol{X}_t, \boldsymbol{X}_t\right] \overset{\text{eq. (15.41)}}{=} \mathbb{V}\left[\boldsymbol{X}_t\right] \tag{16.6}$$

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 $\boldsymbol{x}(t)$ the autocorrelation function $\gamma(t, t - \tau)$ describes how *similar* the time translated function $\boldsymbol{x}(t - \tau)$ and the original function $\boldsymbol{x}(t)$ 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 16.9 Auto Correlation $\rho(t_2 - t_1)$**:**
Is the scaled version of the auto-covariance[def. 16.8]:

$$\rho(t_2 - t_1) = \operatorname{Corr}\left[\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}\right] \tag{16.7}$$
$$= \frac{\operatorname{Cov}\left[\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}\right]}{\sigma_{X_{t_1}} \sigma_{X_{t_2}}} = \frac{\mathbb{E}\left[\left(\boldsymbol{X}_{t_1} - \mu_{t_1}\right)\left(\boldsymbol{X}_{t_2} - \mu_{t_2}\right)\right]}{\sigma_{X_{t_1}} \sigma_{X_{t_2}}}$$

1. Different kinds of Processes

1.1. Markov Process

Definition 16.10 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:

$$\mathbb{P}\left(X(t_{n+1}) \in B | X(t_1), \dots, X(t_n)\right) = \mathbb{P}\left(X(t_{n+1}) \in B | X(t_n)\right)$$

it thus follows for the *transition probability* – the probability of $X(t)$ lying in the set B at time t, given the value x of the process at time s:

$$\mathbb{P}(s, x, t, B) = P(X(t) \in B | X(s) = x) \quad 0 \leqslant s < t \tag{16.8}$$

Interpretation

In order to predict the future only the current/last value counts.

Corollary 16.2 Transition Density: The transition probability of a continuous distribution p can be calculated via:

$$\mathbb{P}(s, x, t, B) = \int_B \mathrm{p}(s, x, t, y)\, \mathrm{d}y \tag{16.9}$$

1.2. Gaussian Process

Definition 16.11 Gaussian Process: Is a stochastic process $X(t)$ where the random variables follow a Gaussian distribution:

$$X(t) \sim \mathcal{N}\left(\mu(t), \sigma^2(t)\right) \quad \forall t \in T \tag{16.10}$$

1.3. Diffusions

Definition 16.12 Diffusion: Is a Markov Process[def. 16.10] for which it holds that:

$$\mu(t, X(t)) = \lim_{t\to 0} \frac{1}{\Delta t} \mathbb{E}\left[X(t + \Delta t) - X(t) | X(t)\right] \tag{16.11}$$

$$\sigma^2(t, X(t)) = \lim_{t\to 0} \frac{1}{\Delta t} \mathbb{E}\left[\left(X(t + \Delta t) - X(t)\right)^2 | X(t)\right] \tag{16.12}$$

See **??**/eq. (16.12) for simple proof of eq. (16.11)/**??**.

- $\mu(t, X(t))$ is called **drift**
- $\sigma^2(t, X(t))$ is called **diffusion coefficient**

Interpretation

There exist not discontinuities for the trajectories.

1.4. Brownian Motion/Wienner Process

Definition 16.13
d-**dim standard Brownian Motion/Wienner Process:**
Is an $\mathbb{R}^d$ valued *stochastic process*[def. 16.1] $(W_t)_{t\in\mathcal{T}}$ starting at $\boldsymbol{x}_0 \in \mathbb{R}^d$ that satisfies:

1. **Normal Independent Increments**: the increments are *normally distributed independent random variables*:
$$W(t_i) - W(t_{i-1}) \sim \mathcal{N}\left(0, (t_i - t_{i-1})\mathbb{1}_{d\times d}\right) \qquad \forall i \in \{1, \dots, T\} \tag{16.13}$$
2. **Stationary increments**:
$W(t + \Delta t) - W(t)$ is independent of $t \in \mathcal{T}$
3. **Continuity**: for *a.e.* $\omega \in \Omega$, the function $t \mapsto W_t(\omega)$ is continuous
$$\lim_{t\to 0} \frac{\mathbb{P}\left(|W(t + \Delta t) - W(t)| \geqslant \delta\right)}{\Delta t} = 0 \qquad \forall \delta > 0 \tag{16.14}$$
4. **Start**
$$W(0) := W_0 = 0 \qquad a.s. \tag{16.15}$$

Notation

- In many source the Brownian motion is a synonym for the standard Brownian Motion and it is the same as the Wienner process.
- **However** in some sources the Wienner process is the standard Brownian Motion, while the Brownian motion denotes a general form $\alpha W(t) + \beta$.

Corollary 16.3 $W_t \sim \mathcal{N}(0, \sigma)$**:**
The random variable W_t follows the $\mathcal{N}(0, \sigma)$ law

$$\mathbb{E}\left[W(t)\right] = \mu = 0 \tag{16.16}$$
$$\mathbb{V}\left[W(t)\right] = \mathbb{E}\left[W^2(t)\right] = \sigma^2 = t \tag{16.17}$$

See section 5

1.4.1. Properties of the Wienner Process

Property 16.1 Non-Differentiable Trajectories:
The sample paths of a Brownian motion are not differentiable:

$$\frac{\mathrm{d}W(t)}{t} = \lim_{t\to 0} \mathbb{E}\left[\left(\frac{W(t + \Delta t) - W(t)}{\Delta t}\right)^2\right]$$
$$= \lim_{t\to 0} \frac{\mathbb{E}\left[W(t + \Delta t) - W(t)\right]}{\Delta t} = \lim_{t\to 0} \frac{\sigma^2}{\Delta t} = \infty$$

$\xrightarrow{\text{result}}$ cannot use normal calculus anymore
$\xrightarrow{\text{solution}}$ Ito Calculus see section 17.

Property 16.2 Auto covariance Function:
The auto-covariance[def. 16.8] for a Wienner process

$$\mathbb{E}\left[(W(t) - \mu t)(W(t') - \mu t')\right] = \min(t, t') \tag{16.18}$$

Property 16.3: A standard Brownian motion is a

Quadratic Variation

Definition 16.14 Total Variation: The total variation of a function $f : [a, b] \subset \mathbb{R} \mapsto \mathbb{R}$ is defined as:

$$LV_{[a,b]}(f) = \sup_{\Pi\in\mathcal{S}} \sum_{i=0}^{n_\Pi - 1} |f(x_{i+1}) - f(x_i)| \tag{16.19}$$

$$\mathcal{S} = \left\{\Pi\{x_0, \dots, x_{n_\Pi}\} : \Pi \text{ is a partition }^{[\text{def. } 11.8]} \text{ of } [a, b]\right\}$$

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 16.15
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:

$$QV_{[a,b]}(f) = \sup_{\Pi\in\mathcal{S}} \sum_{i=0}^{n_\Pi - 1} |f(x_{i+1}) - f(x_i)|^2 \tag{16.20}$$

$$\mathcal{S} = \left\{\Pi\{x_0, \dots, x_{n_\Pi}\} : \Pi \text{ is a partition }^{[\text{def. } 11.8]} \text{ of } [a, b]\right\}$$

Corollary 16.4 Bounded (quadratic) Variation:
The (quadratic) variation[def. 16.14] of a function is bounded if it is finite:

$$\exists M \in \mathbb{R}_+ : \quad LV_{[a,b]}(f) \leqslant M \quad \left(QV_{[a,b]}(f) \leqslant M\right) \quad \forall \Pi \in \mathcal{S} \tag{16.21}$$

Theorem 16.1 Variation of Wienner Process: Almost surely the total variation of a Brownian motion over a interval $[0, T]$ is infinite:

$$\mathbb{P}\left(\omega : LV(W(\omega)) < \infty\right) = 0 \tag{16.22}$$

Theorem 16.2
Quadratic Variation of standard Brownian Motion:
The quadratic variation of a standard Brownian motion over $[0, T]$ is finite:

$$\lim_{N\to\infty} \sum_{k=1}^{N} \left[W\left(k\frac{T}{N}\right) - W\left((k-1)\frac{T}{N}\right)\right]^2 = T \quad \text{with probability 1} \tag{16.23}$$

See **??**

Corollary 16.5 : theorem 16.2 can also be written as:

$$(\mathrm{d}W(t))^2 = \mathrm{d}t \tag{16.24}$$

1.4.2. Lévy's Characterization of BM

Theorem 16.3
d-**dim standard BM/Wienner Process by Paul Lévy:**
An $\mathbb{R}^d$ valued *adapted stochastic process*[def's. 16.1, 16.3] $(W_t)_{t\in\mathcal{T}}$ with the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{R}_+}$, that satisfies:

1. **Start**
$$W(0) := W_0 = 0 \qquad a.s. \tag{16.25}$$
2. **Continuous Martingale**: W_t is an a.s. *continuous* martingale[def. 16.7] w.r.t. the filtration $(\mathcal{F}_t)_{t\in\mathcal{T}}$ under $\mathbb{P}$.
3. **Quadratic Variation**:
$$W_t^2 - t \text{ is also an martingale} \qquad \Longleftrightarrow \qquad QV(W_t) = t \tag{16.26}$$

is a standard Brownian motion[def. 16.18]. Proof see section 5

Further Stochastic Processes

1.4.3. White Noise

undestand script and add

Definition 16.16 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):
$$\mathbb{E}\left[\boldsymbol{\epsilon}[k]\right] = 0 \qquad \forall k \in T_{\text{discret}} \tag{16.27}$$
- Zero autocorrelation[def. 16.9] γ i.e. the signals of different times are in no-way correlated:
$$\gamma(\boldsymbol{\epsilon}[k], \boldsymbol{\epsilon}[k+n]) = \mathbb{E}\left[\boldsymbol{\epsilon}[k]\boldsymbol{\epsilon}[k+n]^\intercal\right] = \mathbb{V}\left[\boldsymbol{\epsilon}[k]\right]\delta_{\text{discret}}[n] \qquad \forall k, n \in T_{\text{discret}} \tag{16.28}$$

With $\qquad \delta_{\text{discret}}[n] := \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{else} \end{cases}$

See proofs

Definition 16.17 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):
$$\mathbb{E}\left[\boldsymbol{\epsilon}(t)\right] = 0 \qquad \forall t \in T_{\text{continuous}} \tag{16.29}$$
- Zero autocorrelation[def. 16.9] γ i.e. the signals of different times are in no-way correlated:
$$\gamma(\boldsymbol{\epsilon}(t), \boldsymbol{\epsilon}(t+\tau)) = \mathbb{E}\left[\boldsymbol{\epsilon}(t)\boldsymbol{\epsilon}(t+\tau)^\intercal\right] \tag{16.30}$$
$$\overset{\text{eq. (15.86)}}{=} \mathbb{V}\left[\boldsymbol{\epsilon}(t)\right]\delta(t - \tau) = \begin{cases} \mathbb{V}\left[\boldsymbol{\epsilon}(t)\right] & \text{if } \tau = 0 \\ 0 & \text{else} \end{cases} \qquad \forall t, \tau \in T_{\text{continuous}} \tag{16.31}$$

1.4.4. Generalized Brownian Motion

Definition 16.18 Brownian Motion: Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 16.13], and define:

$$X_t = \mu t + \sigma W_t \qquad t \in \mathbb{R}_+ \qquad \begin{array}{ll}\mu \in \mathbb{R} & \text{: drift parameter} \\ \sigma \in \mathbb{R}_+ & \text{: scale parameter}\end{array} \tag{16.32}$$

then $\{X_t\}_{t\in\mathbb{R}_+}$ is normally distributed with mean μt and variance $t\sigma^2$ $X_t \sim \mathcal{N}\left(\mu t, \sigma^2 t\right)$.

Theorem 16.4 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:

$$f_{W(t)}(x) \sim \mathcal{N}(\mu t, \sigma^2 t) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left\{-\frac{(x-\mu t)^2}{2\sigma^2 t}\right\} \tag{16.33}$$

Corollary 16.6 : More generally we may define the process:

$$t \mapsto f(t) + \sigma W_t \tag{16.34}$$

which corresponds to a noisy version of f.

Corollary 16.7
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:

$$\mathrm{d}X(t) = \mu \,\mathrm{d}t + \sigma \,\mathrm{d}W(t) \tag{16.35}$$
$$X(0) = 0 \tag{16.36}$$

1.4.5. 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 16.19 Geometric Brownian Motion: Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 16.13] the exponential transform:

$$X(t) = \exp\left(W(t)\right) = \exp\left(\mu t + \sigma W(t)\right) \qquad t \in \mathbb{R}_+ \tag{16.37}$$

is called geometric Brownian motion

Corollary 16.8 Log-normal Returns: For a geometric BM we obtain log-normal returns:

$$\ln\left(\frac{S_t}{S_0}\right) = \mu t + \sigma W(t) \iff \mu t + \sigma W(t) \sim \mathcal{N}(\mu t, \sigma^2 t) \tag{16.38}$$

meaning that the mean and the variance of the process (stock) *log-returns* grow over time linearly.

Corollary 16.9
Geometric BM as a Solution of an SDE:
A stochastic process X_t follows a geometric BM with drift μ and scale σ if it satisfies the following SDE:

$$\begin{aligned}\mathrm{d}X(t) &= X(t)\left(\mu \,\mathrm{d}t + \sigma \,\mathrm{d}W(t)\right) \\ &= \mu X(t)\,\mathrm{d}t + \sigma X(t)\,\mathrm{d}W(t) \end{aligned}\tag{16.39}$$
$$X(0) = 0 \tag{16.40}$$

1.4.6. Locally Brownian Motion

Definition 16.20 Locally Brownian Motion: Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 16.13] a local Brownian motion is a stochastic process $X(t)$ that satisfies the SDE:

$$\mathrm{d}X(t) = \mu\left(X(t), t\right)\mathrm{d}t + \sigma\left(X(t), t\right)\mathrm{d}W(t) \tag{16.41}$$

Note

A local Brownian motion is an generalization of a geometric Brownian motion.

1.4.7. Ornstein-Uhlenbeck Process

Definition 16.21 Ornstein-Uhlenbeck Process: Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 16.13] a Ornstein-Uhlenbeck Process or exponentially correlated noise is a stochastic process $X(t)$ that satisfies the SDE:

$$\mathrm{d}X(t) = -aX(t)\,\mathrm{d}t + b\sigma\,\mathrm{d}W(t) \qquad a > 0 \tag{16.42}$$

1.5. Poisson Processes

Definition 16.22 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 16.23 Poisson Process: A Poisson Process with *rate* $\lambda \in \mathbb{R}_{\geqslant 0}$ is a collection of random variables $X(t)$, $t \in [0, \infty)$ defined on a probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant 0}, \mathbb{P})$, having a discrete *state space* $N = \{0, 1, 2, \ldots\}$ and satisfies:

1. $X_0 = 0$
2. The increments follow a Poisson distribution[def. 15.25]:
$$\mathbb{P}\left((X_t - X_s) = k\right) = \frac{\lambda(t-s)}{k!}\mathrm{e}^{-\lambda(t-s)} \qquad \begin{array}{l} 0 \leqslant s < t < \infty \\ \forall k \in \mathbb{N}\end{array}$$
3. No correlation of (non-overlapping) increments:
$$\forall t_0 < t_1 < \cdots < t_n \text{ : the increments are independent}$$
$$X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \ldots, X_{t_n} - X_{t_{n-1}} \tag{16.43}$$

Interpretation

A Poisson Process is a *continuous-time* process with *discrete*, *positive* realizations in $\in \mathbb{N}_{\geqslant 0}$

Corollary 16.10 Probability of events: Using Taylor in order to expand the Poisson distribution one obtains:

$$\mathbb{P}\left(X_{(t+\Delta t)} - X_t \neq 0\right) = \lambda \Delta t + o(\Delta t^2) \qquad t \text{ small i.e. } t \to 0 \tag{16.44}$$

1. Thus the probability of an event happening during Δt is proportional to time period and the rate λ
2. 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 16.24 Differential of a Poisson Process: The differential of a Poisson Process is defined as:

$$\mathrm{d}X_t = \lim_{\Delta t \to \mathrm{d}t}\left(X_{(t+\Delta t)} - X_t\right) \tag{16.45}$$

Property 16.4 Probability of Events for differential: With the definition of the differential and using the previous results from the Taylor expansion it follows:

$$\mathbb{P}\left(\mathrm{d}X_t = 0\right) = 1 - \lambda \tag{16.46}$$
$$\mathbb{P}\left(|\mathrm{d}X_t| = 1\right) = \lambda \tag{16.47}$$

Proofs

Proof. eq. (16.11):
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:

$$\begin{aligned}\mathbb{E}\left[x(n)\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i(n)\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i(n-1) \pm \delta\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i(n-1)\right] \\ &\overset{\text{induction}}{=} \mathbb{E}\left[x_{n-1}\right] = \ldots \mathbb{E}\left[x(0)\right] = 0\end{aligned}$$

Thus in expectation the particles goes nowhere. □

Proof. eq. (16.12):
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:

$$\begin{aligned}\mathbb{E}\left[x(n)^2\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i(n)^2\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i(n-1) \pm \delta\right]^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i(n-1)^2 \pm 2\delta x_i(n-1) + \delta^2\right] \\ &\overset{\text{ind.}}{=} \mathbb{E}\left[x_{n-1}^2\right] + \delta^2 = \mathbb{E}\left[x_{n-2}^2\right] + 2\delta^2 = \ldots \\ &= \mathbb{E}\left[x(0)\right] + n\delta^2 = n\delta^2\end{aligned}$$

as $n = \frac{\text{time}}{\text{step-size}} = \frac{t}{\Delta x}$ it follows:

$$\sigma^2 = \mathbb{E}\left[x^2(n)\right] - \mathbb{E}\left[x(n)\right]^2 = \mathbb{E}\left[x^2(n)\right] = \frac{\delta^2}{\Delta x}t \tag{16.48}$$

Thus in expectation the particles goes nowhere. □

Proof. eq. (16.30):

$$\begin{aligned}\gamma(\epsilon[k], \epsilon[k+n]) &= \mathrm{Cov}\left[\epsilon[k], \epsilon[k+1]\right] \\ &= \mathbb{E}\left[\left(\epsilon[k] - \mathbb{E}\left[\epsilon[k]\right]\right)\left(\epsilon[k+n] - \mathbb{E}\left[\epsilon[k+n]\right]\right)^{\intercal}\right] \\ &\overset{\text{eq. (16.27)}}{=} \mathbb{E}\left[\left(\epsilon[k]\right)\left(\epsilon[k+n]\right)\right]\end{aligned}$$

□

Proof. corollary 16.3:
Since $B_t - B_s$ is the increment over the interval $[s, t]$, it is the same in distribution as the increment over the interval $[s-s, t-s] = [0, t-s]$

$$\text{Thus} \qquad B_t - B_s \sim B_{t-s} - B_0$$

but as B_0 is a.s. zero by definition eq. (16.15) it follows:

$$B_t - B_s \sim B_{t-s} \qquad B_{t-s} \sim \mathcal{N}(0, t-s)$$

□

Proof. corollary 16.3:

$$W(t) = W(t) - \underbrace{W(0)}_{=0} \sim \mathcal{N}(0, t)$$

$$\Rightarrow \qquad \mathbb{E}[X] = 0 \qquad \mathbb{V}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = t$$

□

Proof. theorem 16.2:

$$\begin{aligned}&\sum_{k=0}^{N-1}\left[W(t_k) - W(t_{k-1})\right]^2 && t_k = k\frac{T}{N} \\ &= \sum_{k=0}^{N-1} X_k^2 && X_k \sim \mathcal{N}\left(0, \frac{T}{N}\right) \\ &= \sum_{k=0}^{N-1} Y_k = n\left(\frac{1}{n}\sum_{k=0}^{N-1} Y_k\right) && \mathbb{E}\left[Y_k\right] = \frac{T}{N} \\ &\overset{\text{S.L.L.N}}{=} n\frac{T}{n} = T\end{aligned}$$

□

Proof. theorem 16.3 (2):

1. first we need to show eq. (16.3): $\qquad \mathbb{E}\left[W_t|\mathcal{F}_s\right] = W_s$
   Due to the fact that W_t is $\mathcal{F}_t$ measurable i.e. $W_t \in \mathcal{F}_t$ we know that:
$$\mathbb{E}\left[W_t|\mathcal{F}_t\right] = W_t \tag{16.49}$$
$$\begin{aligned}\mathbb{E}\left[W_t|\mathcal{F}_s\right] &= \mathbb{E}\left[W_t - W_s + W_s|\mathcal{F}\right] \\ &= \mathbb{E}\left[W_t - W_s|\mathcal{F}_s\right] + \mathbb{E}\left[W_s|\mathcal{F}_s\right] \\ &\overset{\text{eq. (16.49)}}{=} \mathbb{E}\left[W_t - W_s\right] + W_s \\ &\overset{W_t - W_s \sim \mathcal{N}(0, t-s)}{=} W_s\end{aligned}$$
2. second we need to show eq. (16.4): $\qquad \mathbb{E}\left[|X(t)|\right] < \infty$
$$\mathbb{E}\left[|W(t)|\right]^2 \overset{\text{eq. (15.33)}}{\leqslant} \mathbb{E}\left[|W(t)|^2\right] = \mathbb{E}\left[W^2(t)\right] = t \leqslant \infty$$

□

Proof. theorem 16.3 (3): $W_t^2 - t$ is a martingale?
Using the binomial formula we can write and adding $W_s - W_s$:

$$W_t^2 = \left(W_t - W_s\right)^2 + 2W_s\left(W_t - W_s\right) + W_s^2$$

using the expectation:

$$\begin{aligned}\mathbb{E}\left[W_t^2|\mathcal{F}_s\right] &= \mathbb{E}\left[\left(W_t - W_s\right)^2|\mathcal{F}_s\right] + \mathbb{E}\left[2W_s\left(W_t - W_s\right)|\mathcal{F}_s\right] \\ &\quad + \mathbb{E}\left[W_s^2|\mathcal{F}_s\right] \\ &\overset{\text{eq. (16.49)}}{=} \mathbb{E}\left[\left(W_t - W_s\right)^2\right] + 2W_s\mathbb{E}\left[\left(W_t - W_s\right)\right] + W_s^2 \\ &\overset{\text{eq. (16.17)}}{=} \mathbb{V}\left[W_t - W_s\right] + 0 + W_s^2 \\ &\quad t - s + W_s^2\end{aligned}$$

from this it follows that:

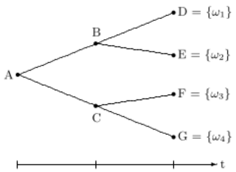
$$\mathbb{E}\left[W_t^2 - t|\mathcal{F}_s\right] = W_s^2 - s$$

□

understand why $\mathbb{E}\left[\left(W_t - W_s\right)^2|\mathcal{F}\right] = \mathbb{E}\left[\left(W_t - W_s\right)^2\right]$

Examples

Example 16.1 :



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.

$$\mathcal{F} = \begin{cases}\{\varnothing, \Omega\} & t \in [0, T/2) \\ \{\varnothing, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \Omega\} & t \in [T/2, T) \\ \mathcal{F}_{\max} = 2^{\Omega} & t = T\end{cases} \tag{16.50}$$

Thus, $\mathcal{F}_0$ represents initial information whereas $\mathcal{F}_\infty$ represents full information (all we will ever know). Hence, a stochastic process is said to be defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant 0}, \mathbb{P})$.

Ito Calculus