Math Appendix

Logic

Set Theory

Definition 2.1 Collection/Multiset: Is a set-like object in which multiplicity matters (order does not).
I.e. $\{1, 1, 2, 3\} \neq \{1, 2, 3\}$

Definition 2.2 Cardinality $|S|$: Is the number of elements that are contained in a set.

Definition 2.3 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 $2^{|S|}$.

Example 2.1 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 3.1 Index Set: Is a set**??** 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.

$$f : A \mapsto S \qquad A \in \mathbb{N} \tag{3.1}$$

Definition 3.2 Sequence $(a_n)_{n\in A}$: is an by an index set A *enumerated* collection[def. 2.1] of objects in which repetitions are allowed and *order does matter*.

Definition 3.3 Series: is an infinite ordered set of terms combined together by addition.

1. Types of Sequences

1.1. Arithmetic Sequence

Definition 3.4 Arithmetic Sequence: Is a sequence where the *difference* between two consecutive terms constant i.e. $(2, 4, 6, 8, 10, 12, \ldots)$.

$$t_n = t_0 + nd \qquad d \text{ :difference between two terms} \tag{3.2}$$

1.2. Geometric Sequence

Definition 3.5 Geometric Sequence: Is a sequence where the *ratio* between two consecutive terms constant i.e. $(2, 4, 8, 16, 32, \ldots)$.

$$t_n = t_0 \cdot r^n \qquad r \text{ :ratio between two terms} \tag{3.3}$$

Calculus and Analysis

Definition 4.1 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 4.2 Discriminant: $\delta = b^2 - 4ac$

Definition 4.3 Solution to [def. 4.1]:
$$x_{\pm} = \frac{-b \pm \sqrt{\delta}}{2a} \quad \text{or} \quad x_{\pm} = \frac{1}{2}\left(-p \pm \sqrt{p^2 - 4q}\right)$$

Theorem 4.1
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:
$$F(X) = \int_a^x f(t)\,\mathrm{d}t \tag{4.1}$$
Then it follows:
$$F'(x) = f(x) \qquad \forall x \in (a, b) \tag{4.2}$$

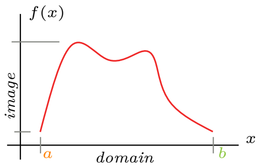
Theorem 4.2
Second Fundamental Theorem of Calculus: Let f be a real-valued function on a closed interval $[a, b]$ and F an antiderivative of f in $[a, b]$: $F'(x) = f(x)$, then it follows if f is Riemann integrable on $[a, b]$:
$$\int_a^b f(t)\,\mathrm{d}t = F(b) - F(a) \iff \int_a^x \frac{\partial}{\partial x} F(t)\,\mathrm{d}t = F(x) \tag{4.3}$$

Definition 4.4 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 4.5
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 4.6 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:
$$\{y \in \mathcal{Y} | y = f(x), \quad \forall x \in \mathcal{X}\} := f[\mathcal{X}] \tag{4.4}$$
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 4.7 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:
$$f^{-1}(A) = \{x \subseteq X : f(x) \subseteq A\} \tag{4.5}$$

Example 4.1 :
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:
$$f[A] = \{y \in \mathcal{Y} | y = f(x), \quad \forall x \in A\} \tag{4.6}$$

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 4.8 (strictly) Increasing Functions:
A function f is called monotonically increasing/ increasing/non-decreasing if:
$$x \leqslant y \iff f(x) \leqslant f(y) \qquad \forall x, y \in \mathtt{dom}(f) \tag{4.7}$$
And **strictly increasing** if:
$$x < y \iff f(x) < f(y) \qquad \forall x, y \in \mathtt{dom}(f) \tag{4.8}$$

Definition 4.9 (strictly) Decreasing Functions:
A function f is called monotonically decreasing/decreasing or non-increasing if:
$$x \geqslant y \iff f(x) \geqslant f(y) \qquad \forall x, y \in \mathtt{dom}(f) \tag{4.9}$$
And *strictly* decreasing if:
$$x > y \iff f(x) > f(y) \qquad \forall x, y \in \mathtt{dom}(f) \tag{4.10}$$

Definition 4.10 Monotonic Function: A function f is called monotonic iff either f is increasing or decreasing.

Definition 4.11 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 4.1 Linearity of Differentiation: The derivative of **any** linear combination of functions equals the same linear combination of the derivatives of the functions:
$$\frac{\mathrm{d}}{\mathrm{d}x}\left(af(x) + bg(x)\right) = a\frac{\mathrm{d}}{\mathrm{d}x}f(x) + b\frac{\mathrm{d}}{\mathrm{d}x}g(x) \qquad a, b \in \mathbb{R} \tag{4.11}$$

Definition 4.12 Quadratic Function:
A function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is quadratic if it can be written in the form:
$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\intercal \boldsymbol{x} + c \tag{4.12}$$

1. Continuity and Smoothness

Definition 4.13 Continuous Function:

Definition 4.14 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:
$$f \in \mathcal{C}^k(\mathcal{X}) \iff \exists f', f'', \dots, f^{(k)} \text{ continuous} \tag{4.13}$$

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 $\mathcal{C}^{k-1}$.
- $\mathcal{C}^m(\mathcal{X}) \subset \mathcal{C}^{m-1}, \dots \mathcal{C}^1 \subset \mathcal{C}^0$
- Continuity is implied by the differentiability of all **derivatives** of up to order $k - 1$.

Corollary 4.2 Smooth Function $\mathcal{C}^\infty$: Is a function $f : \mathcal{X} \to \mathcal{Y}$ that has derivatives infinitely many times differerntiable.
$$f \in \mathcal{C}^\infty(\mathcal{X}) \iff f', f'', \dots, f^{(\infty)} \tag{4.14}$$

Corollary 4.3 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:
$$f \in \mathcal{C}^1(\mathcal{X}) \iff f' \text{ continuous} \tag{4.15}$$

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

Definition 4.15 Lipschitz Continuity: A Lipschitz continuous function is a function f whose rate of change is bound by a Lipschitz Contant L:
$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \qquad \forall \boldsymbol{x}, \boldsymbol{y}, \quad L > 0 \tag{4.16}$$

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 $\Rightarrow$ tells us something about robustness.

Definition 4.16 Lipschitz Continuous Gradient:
A *continuously differentiable* function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has *L-Lipschitz continuous gradient* if it satisfies:
$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\| \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathtt{dom}(f), \quad L > 0 \tag{4.17}$$
if $f \in \mathcal{C}^2$, this is equivalent to:
$$\nabla^2 f(\boldsymbol{x}) \preccurlyeq L\boldsymbol{I} \qquad \forall \boldsymbol{x} \in \mathtt{dom}(f), \quad L > 0 \tag{4.18}$$

Lemma 4.1 Descent Lemma: If a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has *Lipschitz continuous gradient* eq. (4.17) over its domain, then it holds that:
$$|f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^\intercal(\boldsymbol{x} - \boldsymbol{y})| \leqslant \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \tag{4.19}$$

Note

If f is twice differentiable then the largest eigenvalue of the Hessian ([def. 5.5]) of f is uniformly upper bounded by L

Proof. lemma 4.1 for $\mathcal{C}^1$ functions:
Let $g(t) \equiv f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))$ from the FToC (theorem 4.2) 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:
$$\begin{aligned}
&|f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^\intercal(\boldsymbol{x} - \boldsymbol{y})| \\
\overset{\text{Chain. R FToC}}{=} &\left|\int_0^1 \nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))^\intercal(\boldsymbol{x} - \boldsymbol{y})\,\mathrm{d}t - \nabla f(\boldsymbol{y})^\intercal(\boldsymbol{x} - \boldsymbol{y})\right| \\
= &\left|\int_0^1 \left(\nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - \nabla f(\boldsymbol{y})\right)^\intercal(\boldsymbol{x} - \boldsymbol{y})\,\mathrm{d}t\right| \\
= &\left|\int_0^1 \left(\nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - \nabla f(\boldsymbol{y})\right)^\intercal(\boldsymbol{x} - \boldsymbol{y})\,\mathrm{d}t\right| \\
\overset{\text{C.S.}}{\leqslant} &\left|\int_0^1 \|\nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - \nabla f(\boldsymbol{y})\| \cdot \|\boldsymbol{x} - \boldsymbol{y}\|\,\mathrm{d}t\right| \\
\overset{\text{eq. (4.17)}}{=} &\left|\int_0^1 L\|\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}) - \boldsymbol{y}\| \cdot \|\boldsymbol{x} - \boldsymbol{y}\|\,\mathrm{d}t\right| \\
= &\left|L\|\boldsymbol{x} - \boldsymbol{y}\|^2 \int_0^1 t\,\mathrm{d}t\right| = \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2
\end{aligned}$$
□

Proof. lemma 4.1 for $\mathcal{C}^2$ functions:
$$f(\boldsymbol{y}) \overset{\text{Taylor}}{=} f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\intercal(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^\intercal \nabla^2 f(z)(\boldsymbol{y} - \boldsymbol{x})$$
Now we plug in $\nabla^2 f(\boldsymbol{x})$ and recover eq. (4.20):
$$f(\boldsymbol{y}) \leqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\intercal(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^\intercal L(\boldsymbol{y} - \boldsymbol{x})$$
□

Definition 4.17 L-Smoothness: A L-smooth function is a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ that satisfies:
$$f(\boldsymbol{x}) \leqslant f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^\intercal(\boldsymbol{x} - \boldsymbol{y}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$$
with $\qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathtt{dom}(f), \quad L > 0 \quad (4.20)$
If f is a twice differentiable this is equivalent to:
$$\nabla^2 f(\boldsymbol{x}) \preccurlyeq L\boldsymbol{I} \qquad L > 0 \tag{4.21}$$

Theorem 4.3
L-Smoothness of convex functions: A *convex* and L-Smooth function ([def. 4.17]) has a Lipschitz continuous gradient (eq. (4.17)) thus it holds that:
$$f(\boldsymbol{x}) \leqslant f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^\intercal(\boldsymbol{x} - \boldsymbol{y}) \leqslant \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \tag{4.22}$$

Proof. theorem 4.3:
With the definition of convexity for a differentiable function (eq. (4.25)) it follows
$$f(x) - f(y) + \nabla f(y)^\intercal(x - y) \geqslant 0$$
$$\Rightarrow |f(x) - f(y) + \nabla f(y)^\intercal(x - y)|$$
$$\overset{\text{if eq. (4.25)}}{=} f(x) - f(y) + \nabla f(y)^\intercal(x - y)$$
with lemma 4.1 and [def. 4.17] it follows theorem 4.3 □

Corollary 4.4 : 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 4.18 Convex Functions:
A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if it satisfies:
$$f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda)f(y) \qquad \begin{array}{l}\forall x, y \in \mathtt{dom}(f) \\ \forall \lambda \in [0, 1]\end{array} \tag{4.23}$$

include figure from tikz/convexity

Definition 4.19 Concave Functions:
A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if it satisfies:
$$f(\lambda x + (1 - \lambda)y) \geqslant \lambda f(x) + (1 - \lambda)f(y) \qquad \begin{array}{l}\forall x, y \in \mathtt{dom}(f) \\ \forall \lambda \in [0, 1]\end{array} \tag{4.24}$$

Corollary 4.5 Convexity → global minimima: Convexity implies that all local minima (if they exist) are global minima.

Definition 4.20 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:
$$f(x) \geqslant f(y) + \nabla f(y)^\intercal(x - y) \qquad \forall x, y \in \mathtt{dom}(f) \tag{4.25}$$
If f is a twice differentiable function this is equivalent to:
$$\nabla^2 f(x) \geqslant 0 \qquad \forall x, y \in \mathtt{dom}(f) \tag{4.26}$$

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

Corollary 4.6 Strict Convexity → Uniqueness:
Strict convexity implies a unique minimizer $\iff$ at most one global minimum.

Corollary 4.7 : 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 $\mathcal{X}$:
$$f''(x) \geqslant 0 \qquad \forall x \in \mathcal{X} \tag{4.27}$$

Definition 4.21 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:

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

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

$$f(y) \geqslant f(x) + \nabla f(x)^\intercal (y - x) + \frac{\mu}{2}\|y - x\|_2^2 \tag{4.28}$$

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

$$\nabla^2 f(x) \geqslant \mu \boldsymbol{I} \qquad \forall x, y \in \mathcal{X} \quad \mu > 0 \tag{4.29}$$

Corollary 4.8 Strong Convexity implies Strict Convexity:
https://math.stackexchange.com/questions/2090991/proof-for-strongly-convex-function-is-strictly-convex

Property 4.1 :

$$f(\boldsymbol{y}) \leqslant f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^\intercal (\boldsymbol{x} - \boldsymbol{y}) + \frac{1}{2\mu}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \tag{4.30}$$

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. (4.29) analogously to **Proof** eq. (4.21) □

Note

If f is twice differentiable then the smallest eigenvalue of the Hessian ([def. 5.5]) of f is uniformly lower bounded by μ
Hence strong convexity can be considered as the analogous to smoothness

Example 4.2 Quadratic Function: A quadratic function eq. (4.12) is convex if:

$$\nabla^2_{\boldsymbol{x}} \text{eq. (4.12)} = \boldsymbol{A} \geqslant 0 \tag{4.31}$$

Corollary 4.9 :
Strong convexity $\Rightarrow$ Strict convexity $\Rightarrow$ Convexity

2.1. Properties that preserve convexity

Property 4.2 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 4.3 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 4.4 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.
$\Rightarrow$**Geometrically**: its graph remains unchanged after reflection about the y-axis.

$$f(-x) = f(x) \tag{4.32}$$

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

$$f(-x) = -f(x) \tag{4.33}$$

Theorem 4.4 Rules:
Let f be even and f odd respectively.
$g =: f \cdot f$ is even $\qquad g =: f \cdot f$ is even
$g =: f \cdot f$ is odd $\qquad$ the same holds for division

Examples

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

x-**Shift**: $f(x - c) \Rightarrow$ shift to the right
$f(x + c) \Rightarrow$ shift to the left (4.34)
y-**Shift**: $f(x) \pm c \Rightarrow$ shift up/down (4.35)

Proof. **eq. (4.34)** $f(x_n - c)$ we take the x-value at x_n but take the y-value at $x_o := x_n - c$
$\Rightarrow$ we shift the function to x_n. □

Euler's formula

$$e^{\pm ix} = \cos x \pm i \sin x \tag{4.36}$$

Euler's Identity

$$e^{\pm i} = -1 \tag{4.37}$$

Note

$$e^{n} = 1 \Leftrightarrow n = \mathrm{i}\,2\pi k, \qquad k \in \mathbb{N} \tag{4.38}$$

Corollary 4.10 Every norm is a convex function: By using definition [def. 4.18] 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 4.22 Taylor Expansion:

$$T_n(x) = \sum_{i=0}^{n} \frac{1}{n!} f^{(i)}(x_0) \cdot (x - x_0)^{(i)} \tag{4.39}$$

$$= f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 + \mathcal{O}(x^3) \tag{4.40}$$

Definition 4.23 Incremental Taylor:
Goal: evaluate $T_n(x)$ (eq. (4.40)) at the point $x_0 + \Delta x$ in order to propagate the function $f(x)$ by $h = \Delta x$:

$$T_n(x_0 \pm h) = \sum_{i=0}^{n} \frac{h^i}{n!} f^{(i)}(x_0) i^{-1} \tag{4.41}$$

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

Note

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

Definition 4.24 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

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^\intercal H (\boldsymbol{x} - \boldsymbol{x}_0) \tag{4.42}$$

Definition 4.25 Argmax: The argmax of a function defined on a set D is given by:

$$\arg\max_{x \in D} f(x) = \{x | f(x) \geqslant f(y), \forall y \in D\} \tag{4.43}$$

Definition 4.26 Argmin: The argmin of a function defined on a set D is given by:

$$\arg\min_{x \in D} f(x) = \{x | f(x) \leqslant f(y), \forall y \in D\} \tag{4.44}$$

Corollary 4.11 Relationship $\arg\min \leftrightarrow \arg\max$:

$$\arg\min_{x \in D} f(x) = \arg\max_{x \in D} -f(x) \tag{4.45}$$

Property 4.5 Argmax Identities:

1. **Shifting**:
$$\forall \lambda \text{ const} \quad \arg\max f(x) = \arg\max f(x) + \lambda \tag{4.46}$$
2. **Positive Scaling**:
$$\forall \lambda > 0 \text{ const} \quad \arg\max f(x) = \arg\max \lambda f(x) \tag{4.47}$$
3. **Negative Scaling**:
$$\forall \lambda < 0 \text{ const} \quad \arg\max f(x) = \arg\min \lambda f(x) \tag{4.48}$$
4. **Positive Functions**:
$$\forall \arg\max f(x) > 0, \forall x \in \mathtt{dom}(f)$$
$$\arg\max f(x) = \arg\min \frac{1}{f(x)} \tag{4.49}$$
5. **Stricly Monotonic Functions:** for all strictly monotonic increasing functions ([def. 4.8]) g it holds that:
$$\arg\max g(f(x)) = \arg\max f(x) \tag{4.50}$$

Definition 4.27 Max: The maximum of a function f defined on the set D is given by:

$$\max_{x \in D} f(x) = f(x^*) \quad \text{with} \quad \forall x^* \in \arg\max_{x \in D} f(x) \tag{4.51}$$

Definition 4.28 Min: The minimum of a function f defined on the set D is given by:

$$\min_{x \in D} f(x) = f(x^*) \quad \text{with} \quad \forall x^* \in \arg\min_{x \in D} f(x) \tag{4.52}$$

Corollary 4.12 Relationship $\min \leftrightarrow \max$:

$$\min_{x \in D} f(x) = -\max_{x \in D} -f(x) \tag{4.53}$$

Property 4.6 Max Identities:

1. **Shifting**:
$$\forall \lambda \text{ const} \quad \max\{f(x) + \lambda\} = \lambda + \max f(x) \tag{4.54}$$
2. **Positive Scaling**:
$$\forall \lambda > 0 \text{ const} \quad \max \lambda f(x) = \lambda \max f(x) \tag{4.55}$$
3. **Negative Scaling**:
$$\forall \lambda < 0 \text{ const} \quad \max \lambda f(x) = \lambda \min f(x) \tag{4.56}$$
4. **Positive Functions**:
$$\forall \arg\max f(x) > 0, \forall x \in \mathtt{dom}(f) \quad \max \frac{1}{f(x)} = \frac{1}{\min f(x)} \tag{4.57}$$
5. **Stricly Monotonic Functions**: for all strictly monotonic increasing functions ([def. 4.8]) g it holds that:
$$\max g(f(x)) = g(\max f(x)) \tag{4.58}$$

Definition 4.29 Supremum: The supremum of a function defined on a set D is given by:

$$\sup_{x \in D} f(x) = \{y | y \geqslant f(x), \forall x \in D\} = \min_{y | y \geqslant f(x), \forall x \in D} y \tag{4.59}$$

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

Definition 4.30 Infinmum: The infinmum of a function defined on a set D is given by:

$$\inf_{x \in D} f(x) = \{y | y \leqslant f(x), \forall x \in D\} = \max_{y | y \leqslant f(x), \forall x \in D} y \tag{4.60}$$

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

Corollary 4.13 Relationship $\sup \leftrightarrow \inf$:

$$\in_{x \in D} f(x) = -\sup_{x \in D} -f(x) \tag{4.61}$$

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 $x \Rightarrow$ there exists no argmax/argmin $\Rightarrow$ need to bound the functions from above/below $\iff$ infinmum/supremum.

Definition 4.31 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.

$$y(t) = f(x(t), t) \xrightarrow[\forall \tau]{\text{time-invariance}} y(t - \tau) = f(x(t - \tau), t) \tag{4.62}$$

Definition 4.32 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

$$f(g(x)) = x \qquad \forall x \in \mathtt{dom}(g) \tag{4.63}$$

and

$$g(f(u)) = u \qquad \forall u \in \mathtt{dom}(f) \tag{4.64}$$

Property 4.7
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 4.5 The Existence of an Inverse Function:
A function has an inverse function if and only if it is one-to-one.

Corollary 4.14 Inverse functions and strict monotonicity: If a function f is strictly monotonic [def. 4.10] 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 4.33 The gamma function $\Gamma(\alpha)$: Is extension of the factorial function (**??**) to the real and complex numbers (with a positive real part):

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx \qquad \Re(z) > 0 \tag{4.65}$$

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

Differential Calculus

Definition 5.1 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 5.2 Second Derivative $\frac{\partial^2}{\partial x_i \partial x_j}$**:**

Corollary 5.1 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. 5.1] 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 5.3 Gradient: Given $f : n \mapsto \mathbb{R}$ its gradient is defined as:

$$\operatorname{grad}_{\boldsymbol{x}}(f) = \nabla_{\boldsymbol{x}} f := \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \dots & \frac{\partial f}{\partial x_n} \end{bmatrix} \tag{5.1}$$

Definition 5.4 Jacobi Matrix: Given a vector valued function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ its derivative/Jacobian is defined as:

$$\boldsymbol{J}(\boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{J}_f(\boldsymbol{x}) = \boldsymbol{D}\boldsymbol{f} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}(\boldsymbol{x}) = \frac{\partial(f_1, \dots, f_m)}{\partial(x_1, \dots, x_n)}(\boldsymbol{x}) = \tag{5.2}$$

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

Theorem 5.1
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 5.5 Hessian Matrix:
Given a function $f : \mathbb{R} \mapsto \mathbb{R}^n$ its Hessian$\in \mathbb{R}^{n \times n}$ is defined as:

$$\boldsymbol{H}(\boldsymbol{f})(\boldsymbol{x}) = \boldsymbol{H}_f(\boldsymbol{x}) = \boldsymbol{J}(\nabla \boldsymbol{f}(\boldsymbol{x}))^T \tag{5.3}$$

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

and it corresponds to the Jacobian of the Gradient.
Due to the differentiability and theorem 5.1 it follows that the Hessian is (if it exists):

- Symmetric
- Real

Corollary 5.2 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:

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

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

Corollary 5.3 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. 5.1] 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 6.1 Important Integral Properties:

Addition $$\int_a^b f(x)\,\mathrm{d}x = \int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x \tag{6.1}$$

Reflection $$\int_a^b f(x)\,\mathrm{d}x = -\int_b^a f(x)\,\mathrm{d}x \tag{6.2}$$

Translation $$\int_a^b f(x)\,\mathrm{d}x \overset{u:=x\pm c}{=} \int_{a\pm c}^{b\pm c} f(x\mp c)\,\mathrm{d}x \tag{6.3}$$

f **Odd** $$\int_{-a}^{a} f(x)\,\mathrm{d}x = 0 \tag{6.4}$$

f **Even** $$\int_{-a}^{a} f(x)\,\mathrm{d}x = 2\int_0^a f(x)\,\mathrm{d}x \tag{6.5}$$

Proof. **eqs. (6.4) and (6.5)**

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

$\square$

Linear Algebra

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

Rank: $\operatorname{rank}(\boldsymbol{A}) = \dim(\mathfrak{R}(\boldsymbol{A})$
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: $\mathbb{N} := \{z \in \mathbb{K}^n \mid \boldsymbol{A}z = 0\}$
Dimension theorem:

Theorem 7.1 Rank-Nullity theorem: For any $\boldsymbol{A} \in \mathbb{Q}^{m \times n}$
$$n = \dim(\mathbb{N}[\boldsymbol{A}]) + \dim(\mathfrak{R}[\boldsymbol{A}])$$

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

1. Eigenvalues and Vectors

Formula 7.1 Eigenvalues of a 2x2 matrix: Given a 2x2-matrix $\boldsymbol{A}$ its eigenvalues can be calculated by:
$$\{\lambda_1, \lambda_2\} \in \frac{\operatorname{tr}(\boldsymbol{A}) \pm \sqrt{\operatorname{tr}(\boldsymbol{A})^2 - 4\det(\boldsymbol{A})}}{2} \tag{7.1}$$
with $\quad \operatorname{tr}(\boldsymbol{A}) = a + d \qquad \det(\boldsymbol{A}) = ad - bc$

2. Special Kind of Matrices

Definition 7.1 Hermitian Matrices:
$$\boldsymbol{A} = \boldsymbol{A}^{\mathsf{H}} \tag{7.2}$$

3. Spaces and Measures

Definition 7.2 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:
$$a(\alpha u + \beta v, w) = \alpha a(u, w) + \beta a(v, w)$$
$$a(u, \alpha v + \beta w) = \alpha a(u, v) + \beta a(u, w)$$
$$\forall u, v, w \in \mathcal{Y}, \quad \forall \alpha, \beta \in \mathbb{K}$$
Thus: a is linear w.r.t. each argument.

Definition 7.3 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 7.4 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:
$$a(u, u) > 0 \qquad \forall u \in \mathcal{Y} \backslash \{0\} \tag{7.3}$$
And positive semidefinte $\iff \geqslant$ (7.4)

Corollary 7.1 Matrix induced Bilinear Form: For finite dimensional inner product spaces $\mathcal{X} \in \mathbb{K}^n$ any 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 7.5 Positive (semi) definite Matrix $\succ$: A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is positive defintie if and only if:
$$\boldsymbol{x}^\intercal \boldsymbol{A} \boldsymbol{x} > 0 \iff \boldsymbol{A} \succ \quad \forall \boldsymbol{x} \in \mathbb{R}^n \backslash \{0\} \tag{7.5}$$
And positive semidefinte $\iff \succcurlyeq$ (7.6)

Corollary 7.2 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*.
$$\forall \lambda_i \in \operatorname{eigenv}(\boldsymbol{A}) > 0 \tag{7.7}$$
And positive semidefinte $\iff \geqslant$ (7.8)

Proof. corollary 7.2 (for real matrices):
Let $\boldsymbol{v}$ be an eigenvector of $\boldsymbol{A}$ then it follows:
$$0 \overset{\text{corollary 7.2}}{<} \boldsymbol{v}^\intercal \boldsymbol{A} \boldsymbol{v} = \boldsymbol{v}^\intercal \lambda \boldsymbol{v} = \|\boldsymbol{v}\| \lambda$$
□

Corollary 7.3 Positive Definiteness and Determinant: The determinant of a positive definite matrix is always positive. **Thus** a positive definite matrix is always *nonsingular*

Definition 7.6 Negative (semi) definite Matrix *prec*: A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is negative defintie if and only if:
$$\boldsymbol{x}^\intercal \boldsymbol{A} \boldsymbol{x} < 0 \iff \boldsymbol{A} < 0 \quad \forall \boldsymbol{x} \in \mathbb{R}^n \backslash \{0\} \tag{7.9}$$
And negative semidefinte $\iff \leqslant$ (7.10)

Theorem 7.2 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:
- $$\boldsymbol{A} > 0 \iff \det\left(\boldsymbol{A}^k\right) > 0 \quad k = 1, \ldots, n \tag{7.11}$$
- $$\boldsymbol{A} < 0 \iff (-1)^k \det\left(\boldsymbol{A}^k\right) > 0 \quad k = 1, \ldots, n \tag{7.12}$$
- $\boldsymbol{A}$ 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.

4. Inner Products

Definition 7.7 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:
$$\langle \cdot, \cdot \rangle : \mathcal{Y} \times \mathcal{Y} \mapsto F \subseteq \mathbb{K} \qquad K = \mathbb{R} \text{ or } \mathbb{C} \tag{7.13}$$
that satisfies: $\quad \forall x, y, z \in \mathcal{Y}, \quad \alpha, \beta \in F$
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 7.8 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 7.4 Inner product↦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 7.9 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:
$$\|\cdot\|_{\mathcal{Y}} : \mathcal{Y} \mapsto \mathbb{R}_+ \tag{7.14}$$
that satisfies: $\quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{Y}, \quad \alpha \in F \subseteq \mathbb{K} \quad K = \mathbb{R}$ 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 7.5 Reverse Triangular Inequality:
$$-\|\boldsymbol{x} - \boldsymbol{y}\|_{\mathcal{Y}} \leqslant \|\boldsymbol{x}\|_{\mathcal{Y}} - \|\boldsymbol{y}\|_{\mathcal{Y}} \leqslant \|\boldsymbol{x} - \boldsymbol{y}\|_{\mathcal{Y}}$$
resp. $$\left| \|\boldsymbol{x}\|_{\mathcal{Y}} - \|\boldsymbol{y}\|_{\mathcal{Y}} \right| \leqslant \|\boldsymbol{x} - \boldsymbol{y}\|_{\mathcal{Y}}$$

Semi-norm

add

Corollary 7.6 Normed vector space: Is a vector space $\mathcal{Y}$ over a field F, on which a norm $\|\cdot\|_{\mathcal{Y}}$ can be defined.

Corollary 7.7 Inner product induced norm $\langle \cdot, \cdot \rangle_{\mathcal{Y}} \to \|\cdot\|_{\mathcal{Y}}$: 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 7.8 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 7.10 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 7.11 Metric: Is a distance measure that additonally satisfies: $\quad \forall x, y \in S$
identity of indiscernibles : $d(x, y) = 0 \iff x = y$

Corollary 7.9 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.

5. Vector Algebra

5.1. Planes

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

6. 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

Definition 8.1 Affine Transformation/Map:

Corollary 8.1 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:

$$y = \phi(x) = \frac{d-c}{b-a}(x-a) + c \tag{8.1}$$

Proof. **corollary 8.1** 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. 4.11]), that is the output scales the same way as the input scales.
Thus it follows:

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

□

Trigonometry

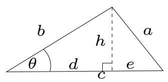
Law 8.1 Law of Cosine: relates the side of a triangle to the cosine of its angles.

$$a^2 = b^2 + c^2 - 2bc\cos\theta_{b,c} \tag{8.2}$$

More general for vectors it holds:

$$\|\boldsymbol{x} - \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\|\boldsymbol{x}\|\|\boldsymbol{y}\|\cos\theta_{\boldsymbol{x},\boldsymbol{y}} \tag{8.3}$$

Proof. eq. (8.2):
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. (8.3):

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

□

Law 8.2 Pythagorean theorem: special case of **??** for right triangle:

$$a^2 = b^2 + c^2 \tag{8.4}$$

Euler's formula

$$e^{\pm ix} = \cos x \pm i \sin x \tag{8.5}$$

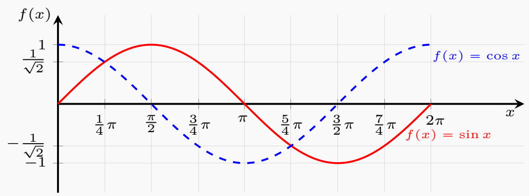
Euler's Identity

$$e^{\pm i} = -1 \tag{8.6}$$

Note

$$e^{n} = 1 \Leftrightarrow n = \mathrm{i}\, 2\pi k, \qquad k \in \mathbb{N} \tag{8.7}$$

Sine and Cosine



$$\cos x \overset{(8.5)}{=} \frac{1}{2}\left[e^{ix} + e^{-ix}\right] \tag{8.8}$$

$$\sin x \overset{(8.5)}{=} \frac{1}{2i}\left[e^{ix} - e^{-ix}\right] = -\frac{i}{2}\left[e^{ix} - e^{-ix}\right] \tag{8.9}$$

Sinh and Cosh

$$\cosh x \overset{(8.5)}{=} \frac{1}{2}\left[e^{x} + e^{-x}\right] = \cos(\mathrm{i}\, x) \tag{8.10}$$

$$\sinh x \overset{(8.5)}{=} \frac{1}{2}\left[e^{x} - e^{-x}\right] = -\mathrm{i}\sin(\mathrm{i}\, x) \tag{8.11}$$

Note

$$e^{x} = \cosh x + \sinh x \qquad e^{-x} = \cosh x - \sinh x \tag{8.12}$$

Note

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

Theorem 8.1 Addition Theorems:

$$\sin(\alpha \pm \beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta \tag{8.13}$$

$$\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta \tag{8.14}$$

Werner Formulas

$$\sin\alpha\cos\beta = \frac{1}{2}\Big[\sin(\alpha+\beta) + \sin(\alpha-\beta)\Big] \tag{8.15}$$

$$\sin\alpha\sin\beta = \frac{1}{2}\Big[\cos(\alpha-\beta) - \cos(\alpha+\beta)\Big] \tag{8.16}$$

$$\cos\alpha\cos\beta = \frac{1}{2}\Big[\cos(\alpha+\beta) + \cos(\alpha-\beta)\Big] \tag{8.17}$$

Note

Using theorem 8.1 if follows:

$$\cos(\alpha \pm \pi) = -\cos\alpha \qquad \text{and} \qquad \sin(\alpha \pm \pi) = -\sin\alpha \tag{8.18}$$

Topology

Numerics

Definition 10.1 Partition Π: Given an interval $[0,T]$ a sequence of values $0 < t_0 < \cdots < t_n < T$ is called a partition $\Pi(t_0,\dots,t_n)$ of this interval.

0.1. Convention for iterative methods

Definition 10.2 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:

$$\left\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^*\right\| \leqslant \rho\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\right\| \qquad \rho\in(0,1),\forall k\in\mathbb{N}_0 \tag{10.1}$$

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:
$$\text{eq. (10.2)} \iff \left\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^*\right\| \leqslant e^{-\alpha}\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\right\|$$

Definition 10.3 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:

$$\lim_{k\mapsto\infty}\frac{\left\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^*\right\|}{\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\right\|}=\rho \tag{10.2}$$

- $\rho=1 \iff$ 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 10.4 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 $p\in\{2,\dots\}$ to $\boldsymbol{x}^*$ if it satisfies:

$$\lim_{k\mapsto\infty}\frac{\left\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^*\right\|}{\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\right\|^p}=C \qquad C<1 \tag{10.3}$$

Does this even exist/check if this is ture

Definition 10.5 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:

$$\left\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^*\right\|\leqslant\rho^k\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\right\| \qquad \rho<1 \tag{10.4}$$

0.2. Convention for discritization methods

1. Numerical Quadrature

Definition 10.6 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:

$$\text{order}(\mathcal{Q}_n):=\max\left\{n\in\mathbb{N}_0:\mathcal{Q}_n(p)=\int_a^b p(t)\,\mathrm{d}t\quad\forall p\in\mathcal{P}_n\right\}+1 \tag{10.5}$$

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.

1.1. Composite Quadrature

Definition 10.7 Composite Quadrature:
Given a mesh $\mathcal{M}=\{a=x_0<x_1<\dots<x_m=b\}$ apply a Q.R. $\mathcal{Q}_n$ to each of the mesh cells $I_j:=[x_{j-1},x_j]$ $\forall j=1,\dots,m$ ≙ p.w. Quadrature:

$$\int_a^b f(t)\,\mathrm{d}t=\sum_{j=1}^m\int_{x_{j-1}}^{x_j}f(t)\,\mathrm{d}t=\sum_{j=1}^m\mathcal{Q}_n(f_{I_j}) \tag{10.6}$$

Lemma 10.1 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:

$$\epsilon_n(f)=\mathcal{O}\left(n^{-\min\{k,q\}}\right)=\mathcal{O}\left(h_\mathcal{M}^{\min\{k,q\}}\right) \qquad \text{for } n\to\infty$$
$$\overset{\text{corollary 4.2}}{=}\mathcal{O}\left(n^{-q}\right)=\mathcal{O}\left(h^q_\mathcal{M}\right) \qquad \text{with } h_\mathcal{M}=\frac{1}{n} \tag{10.7}$$

Definition 10.8 Complexity W: Is the number of function evaluations ≙ number of quadrature points.

$$W(\mathcal{Q}(f)_n)=\#\text{f-eval}\triangleq n \tag{10.8}$$

Lemma 10.2 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:

$$W(\epsilon_n(f))=\mathcal{O}(\epsilon_n^{1/q})=\mathcal{O}\left(\sqrt[q]{\epsilon_n}\right) \tag{10.9}$$

Proof. lemma 10.2: **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?

$$\frac{c\cdot n_n^q}{c\cdot n_o^q}=\frac{1}{\epsilon_n} \quad\Rightarrow\quad n_n=n_o\cdot\sqrt[q]{\epsilon_n}=\mathcal{O}(\sqrt[q]{\epsilon_n}) \tag{10.10}$$

□

Optimization

Definition 11.1 Fist Order Method: A first-order method is an algorithm that chooses the k-th iterate in

$$\boldsymbol{x}_0+\text{span}\{\nabla f(\boldsymbol{x}_0),\dots\nabla f(\boldsymbol{x}_{k-1})\} \qquad \forall k=1,2,\dots \tag{11.1}$$

Note

Gradient descent is a first order method

1. Lagrangian Optimization Theory

Add: derivation of lagrange function

Definition 11.2 (Primal) Constraint Optimization:
Given an optimization problem with domain $\Omega\subseteq\mathbb{R}^d$:

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

Definition 11.3 Lagrange Function:

$$\mathcal{L}(\alpha,\beta,\boldsymbol{w}):=f(\boldsymbol{w})+\alpha\boldsymbol{g}(\boldsymbol{w})+\beta\boldsymbol{h}(\boldsymbol{w}) \tag{11.2}$$

Extremal Conditions

$$\nabla\mathcal{L}(\boldsymbol{x})\overset{!}{=}0 \qquad \text{Extremal point } \boldsymbol{x}^*$$
$$\frac{\partial}{\partial\beta}\mathcal{L}(\boldsymbol{x})=h(\boldsymbol{x})\overset{!}{=}0 \qquad \text{Constraint satifisfaction}$$

For the inequality constraints $g(\boldsymbol{x})\leqslant0$ we distinguish two situations:

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

Case II : $g(\boldsymbol{x}^*)\geqslant0$ optimze using active eq. constr.

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

Definition 11.4 Lagrangian Dual Problem: Is given by:

$$\begin{aligned}\text{Find}\quad & \max\theta(\alpha,\beta)=\inf_{\boldsymbol{w}\in\Omega}\mathcal{L}(\boldsymbol{w},\alpha,\beta)\\ \textbf{s.t.}\quad & \alpha_i\geqslant0 \qquad 1\leqslant i\leqslant k\end{aligned}$$

Solution Strategy

1. Find the extremal point $\boldsymbol{w}^*$ of $\mathcal{L}(\boldsymbol{w},\alpha,\beta)$:
$$\left.\frac{\partial\mathcal{L}}{\partial\boldsymbol{w}}\right|_{\boldsymbol{w}=\boldsymbol{w}^*}\overset{!}{=}0 \tag{11.3}$$
2. Insert $\boldsymbol{w}^*$ into $\mathcal{L}$ and find the extremal point β^* of the resulting dual Lagrangian $\theta(\alpha,\beta)$ for the active constraints:
$$\left.\frac{\partial\theta}{\partial\beta}\right|_{\beta=\beta^*}\overset{!}{=}0 \tag{11.4}$$
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 11.1 Upper Bound Dual Cost: Let $\boldsymbol{w}\in\Omega$ be a feasible solution of the primal problem [def. 11.2] and (α,β) a feasible solution of the respective dual problem [def. 11.4].
Then it holds that:

$$f(\boldsymbol{w})\geqslant\theta(\alpha,\beta) \tag{11.5}$$

Proof.

$$\begin{aligned}\theta(\alpha,\beta)&=\inf_{\boldsymbol{u}\in\Omega}\mathcal{L}(\boldsymbol{u},\alpha,\beta)\leqslant\mathcal{L}(\boldsymbol{w},\alpha,\beta)\\ &=f(\boldsymbol{w})+\sum_{i=1}^k\underbrace{\alpha_i}_{\geqslant0}\underbrace{g_i(\boldsymbol{w})}_{\leqslant0}+\sum_{j=}^m\beta_j\underbrace{h_j(\boldsymbol{w})}_{=0}\\ &\leqslant f(\boldsymbol{w})\end{aligned}$$

□

Corollary 11.1 Duality Gap Corollary: The value of the dual problem is upper bounded by the value of the primal problem:

$$\sup\{\theta(\alpha,\beta):\alpha\geqslant0\}\leqslant\inf\{f(\boldsymbol{w}):\boldsymbol{g}(\boldsymbol{w})\leqslant0,\boldsymbol{h}(\boldsymbol{w})=0\} \tag{11.6}$$

Theorem 11.2 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:

$$f(\boldsymbol{w}^*)=\theta(\alpha^*,\beta^*) \tag{11.7}$$

Definition 11.5 Convex Optimization: **Given**: a convex function f and a convex set S solve:

$$\begin{aligned}&\min f(\boldsymbol{x})\\ \text{s.t.}\quad&\boldsymbol{x}\in S\end{aligned} \tag{11.8}$$

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 11.3 Strong Duality: Given an convex optimization problem:

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

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

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

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.:

$$\frac{\partial\mathcal{L}(\boldsymbol{w},\alpha,\beta)}{\partial\boldsymbol{w}}\overset{!}{=}0 \qquad \frac{\partial\mathcal{L}(\boldsymbol{w}^*,\alpha,\beta)}{\partial\beta}\overset{!}{=}0 \tag{11.9}$$

under the condtions that:

- $\forall i_1,\dots,k$ $\quad\alpha_i^*g_i(\boldsymbol{w}^*)=0$, s.t.:
  - Inactive Constraint: $g_i(\boldsymbol{w}^*)<0\to\alpha_i=0$.
  - Active Constraint: $g_i(\boldsymbol{w}^*)\nless0\to\alpha_i\geqslant0$ s.t. $\alpha_i^*g_i(\boldsymbol{w}^*)=0$

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

Definition 11.7 Statistics: Is concerned with the analysis of data/experiments in order to draw conclusion of the underlying governing models that describe these experiments.

Definition 11.8 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 11.9 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

1. Foundations

Definition 12.1 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 12.2 Sample Space Ω: Is the set of all possible outcomes (elementary events corollary 12.5) of an experiment see example 12.1

Definition 12.3 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 12.2).
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 12.1 : If the outcome ω of an experiment is in the subset A, then the event A is said to "have occured".

Corollary 12.2 Complement Set A^{C}: is the contrary event of A.

Corollary 12.3 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 12.4 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 12.5 The Elementary Event ω: Is a "singleton", i.e. a subset $\{\omega\}$ containing a single outcome ω of Ω.

Corollary 12.6 The Sure Event Ω: Is equal to the sample space as it contains all possible elementary events.

Corollary 12.7 The Impossible Event $\varnothing$: The impossible event i.e. nothing is happening is denoted by the empty set.

Definition 12.4 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 12.5 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:

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

1.1. Sigma Algebras

Definition 12.6 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 $\Omega \backslash A = A^{C} \in \mathcal{F}$:
  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 12.3.

Corollary 12.8 $\mathcal{F}_{\min}$: $\mathcal{F} = \{\varnothing, \Omega\}$ is the simplest σ-algebra, telling us only if an event happened $\omega \in \Omega$ happened or not but not which one.

Corollary 12.9 $\mathcal{F}_{\max}$: $\mathcal{F} = 2^{\Omega}$ consists of all subsets of Ω and thus corresponds to full information i.e. we know if and which event happened.

Definition 12.7 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 12.10 $\mathcal{F}$-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 12.8 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 12.4.

Definition 12.9 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 12.5.

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 12.11 : 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 at the end of the section.

Definition 12.10 ($\mathbb{P}$)-trivial Sigma Algebra: is a σ-algebra $\mathcal{F}$ for which each event has a probability of zero or one:

$$\mathbb{P}(A) \in \{0, 1\} \qquad \forall A \in \mathcal{F} \tag{12.1}$$

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

1.2. Measures

Definition 12.11 Measure μ: A measure defined on a measurable space $\{\Omega, \mathcal{F}\}$ is a function/map:

$$\mu : \mathcal{F} \mapsto [0, \infty] \tag{12.2}$$

for which holds:
- $\mu(\varnothing) = 0$
- countable additivity [def. 12.12]

Definition 12.12 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:

$$\mu\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} \mu(F_i) \quad \text{for all} \quad F_j \cap F_k = \varnothing \quad \forall j \neq k \tag{12.3}$$

Corollary 12.12 Additive Function: A function that satisfies countable additivity, is also additive, meaning that for every $F, G \in \mathcal{F}$ it holds:

$$\mu(F \cup G) = \mu(F) + \mu(G) \iff F \cap G = \varnothing \tag{12.4}$$

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 12.13 Equivalent Measures $\mu \sim \nu$: Let μ and ν be two measures defined on a measurable space[def. 12.7] $(\Omega, \mathcal{F})$. The two measures are said to be equivalent if it holds that:

$$\mu(A) > 0 \iff \nu(A) > 0 \qquad \forall A \subseteq \mathcal{F} \tag{12.5}$$

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

$$\mathcal{N}_\mu = \mathcal{N}_\nu \qquad \begin{array}{l} \mathcal{N}_\mu = \{A \in \mathcal{A} | \mu(A) = 0\} \\ \mathcal{N}_\nu = \{A \in \mathcal{A} | \nu(A) = 0\} \end{array} \tag{12.6}$$

see example 12.6

Definition 12.14 Measure Space $\{\mathcal{F}, \Omega, \mu\}$: The triplet of sample space, sigma algebra and a measure is called a measure space.

Definition 12.15 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:

$$\lambda([a, b]) = b - a \tag{12.7}$$

Corollary 12.13 Lebesgue Measure of Atomitcs:

- The Lebesgue measure of a set containing only one point must be zero:
$$\lambda(\{a\}) = 0 \tag{12.8}$$
- The Lebesgue measure of a set containing countably many points $A = \{a_1, a_2 \ldots, a_n\}$ must be zero:
$$\lambda(A) + \sum_{i=1}^{n} \lambda(\{a_i\}) = 0 \tag{12.9}$$
- 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.

1.3. 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 12.1 Non-negativity: The probability of an event is a non-negative real number:

$$\text{If } A \in \mathcal{F} \qquad \text{then} \qquad \mathbb{P}(A) \geqslant 0 \tag{12.10}$$

Axiom 12.2 Unitairity: The probability that at least one of the elementary events in the entire sample space Ω will occur is equal to one:

$$\text{The certain event} \qquad \mathbb{P}(\Omega) = 1 \tag{12.11}$$

Axiom 12.3 σ-additivity: If $A_1, A_2, A_3, \ldots \in \mathcal{F}$ are mutually disjoint, then:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \tag{12.12}$$

Corollary 12.14 : As a consequence of this it follows:

$$\mathbb{P}(\varnothing) = 0 \tag{12.13}$$

Corollary 12.15 Complementary Probability:

$$\mathbb{P}(A^{C}) = 1 - \mathbb{P}(A) \qquad \text{with} \qquad A^{C} = \Omega - A \tag{12.14}$$

Definition 12.16 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.

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

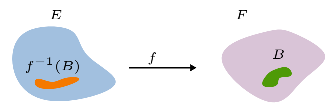
$$X : \Omega \mapsto \mathcal{E} \qquad \omega \mapsto X(\omega)$$

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 12.17 $\mathcal{E}$-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

$$\forall B \in \mathcal{F} : \qquad f^{-1}(B) = \{\omega \in \mathcal{E} : f(\omega) \in B\} \in \mathcal{E} \tag{12.15}$$



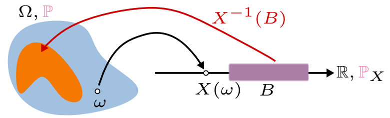
Interpretation

The pre-image[def. 4.7] 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 12.18 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 $(F, \mathcal{F})$:

$$X : \Omega \mapsto \mathcal{F} \quad (\mathcal{F}^n) \tag{12.16}$$

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



Corollary 12.16 : Usually $F = \mathbb{R}$, which usually amounts to using the Borel σ-algebra $\mathcal{B}$ of $\mathbb{R}$.

Corollary 12.17 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 12.11):

$$\begin{aligned} X^{-1}(B) &= X^{-1}((-\infty, a]) \\ &= \{\omega \in \Omega : X(\omega) \leqslant a\} \in \mathcal{E} \quad \forall a \in \mathbb{R} \end{aligned} \tag{12.17}$$

Definition 12.19 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 12.18 Indicator Functions $I_A(\omega)$: An important class of measurable functions that can be used as r.v. are indicator functions:

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} \tag{12.18}$$

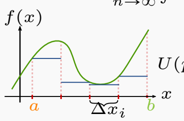
3. Lebesgue Integration

Problems of Riemann Integration

verify and deepen knowledge about that

- 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

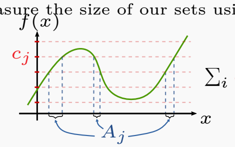
$$\lim_{n\to\infty} \int f(x)\,dx \overset{\text{str. u.c.}}{=} \int \lim_{n\to\infty} f(x)\,dx$$



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

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.



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

Definition 12.20 Lebesgue Integral:

$$\lim_{n\to\infty} \sum_{i=1}^{n} c_i \mu(A_i) = \int_\Omega f\,d\mu \qquad \begin{array}{l} f(x) \approx c_i \\ \forall x \in A_i \end{array} \tag{12.19}$$

Definition 12.21 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:

$$X = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i} \qquad a_i \in \mathbb{R} \qquad \mathcal{A} \ni A_i = \begin{cases} 1 & \text{if } \{X = a_i\} \\ 0 & \text{else} \end{cases} \tag{12.20}$$

Definition 12.22 Expectation:

$$\mathbb{E}[X] = \int_\Omega X(\omega)\mathbb{P}(d\omega) = \int_\Omega X\,d\mathbb{P} \tag{12.21}$$

Corollary 12.19 Expectation of simple r.v.: If X is a simple r.v. its *expectation* is given by:

$$\mathbb{E}[X] = \sum_{i=1}^{n} a_i \mathbb{P}(A_i) \tag{12.22}$$

Proofs

Proof. corollary 12.11: Let $\mathcal{C}$ denote all open intervals. Since every open set in $\mathbb{R}$ is the countable union of open intervals, it holds that $\sigma(\mathcal{C})$=the Borel σ-algebra of $\mathbb{R}$. Now let $\mathcal{D}$ denote all intervals of the form $(-\infty, a]$, where $a \in \mathbb{Q}$. Let $a, b \in \mathcal{C}$, and let

add

See book □

Examples

Example 12.1 :
- 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: $\Omega = \{1, 2, 3, \ldots\}$
- The reals: $\Omega = \{\omega | \omega \in \mathbb{R}\}$

Example 12.2 :
- 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 12.3 : 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 12.4 : 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 $\sigma(A) = \left\{\varnothing, A, A^C, \Omega\right\}$.

Example 12.5 :
- open half-lines: $(-\infty, a)$ and (a, ∞),
- union of open half-lines: $(a, b) = (-\infty, a) \cup (b, \infty)$,
- closed interval: $[a, b] = \overline{(-\infty, \cup a) \cup (b, \infty)}$,
- 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: $\{a_1, \ldots, a_n\} = \bigcup_{k=1}^{n} a_k$.

Example 12.6 Equivalent (Probability) Measures:

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

Example 12.7 :

add example fat book p.1286

add example prob th book 4

Combinatorics

0.1. Permutations

Definition 13.1 Permutation $n!$: Given a set?? $\mathcal{S}$ of n distinct objects, into how many distinct sequences/orders can we arrange/permutate those distinct objects

$$P(\mathcal{S}) = n! \qquad \Longleftrightarrow \qquad P(\mathcal{S}) = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 1 \tag{13.1}$$

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:

$$P(\mathcal{S}) = \frac{n!}{n_1! \cdot \ldots \cdot n_k} \qquad \text{s.t.} \qquad \sum_{i=1}^{k} n_i \leqslant n \tag{13.2}$$

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

$$\mathbb{P}(\omega) = 1 \qquad \Longleftrightarrow \qquad \omega \text{ happens a.s.} \tag{14.1}$$

Definition 14.2 Probability Mass Function (PMF):

Definition 14.3 Discrete Random Variable (DVR): The set of possible values $\bar{x}$ of $\mathcal{X}$ is countable of finite.

$$\mathcal{X} = \{0, 1, 2, 3, 4, \ldots, 8\} \qquad \mathcal{X} = \mathbb{N} \tag{14.2}$$

Definition 14.4 Probability Density Function (PDF): Is real function $f : \mathbb{R}^n \to [0, \infty)$ that satisfies:

Non-negativity: $$f(x) \geqslant 0, \quad \forall x \in \mathbb{R}^n \tag{14.3}$$

Normalization: $$\int_{-\infty}^{\infty} f(x)\,dx \overset{!}{=} 1 \tag{14.4}$$

Must be integrable (14.5)

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:

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

To have a sensible measure of the magnitude of these infinitesimal quantities, we use the concept of probability density, which yields a probability mass when integrated over an interval.

Definition 14.5 Continuous Random Variable (CRV): A real random variable (rrv) X is said to be (absolutely) continuous if there exists a pdf ([def. 14.4]) f_X s.t. for any subset $B \subset \mathbb{R}$ it holds:

$$\mathbb{P}(X \in B) = \int_B f_X(x)\,dx \tag{14.6}$$

Property 14.1 Zero Probability: If X is a continuous rrv ([def. 14.5]), then:

$$\mathbb{P}(X = a) = 0 \qquad \forall a \in \mathbb{R} \tag{14.7}$$

Property 14.2 Open vs. Closed Intervals: For any real numbers a and b, with $a < b$ it holds:

$$\begin{aligned} \mathbb{P}(a \leqslant X \leqslant b) &= \mathbb{P}(a \leqslant X < b) = \mathbb{P}(a < X \leqslant b) \\ &= \mathbb{P}(a < X < b) \end{aligned} \tag{14.8}$$

$\Longleftrightarrow$ 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 14.1 : 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 14.1:

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

□

Proof. Property 14.2:

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

□

Definition 14.6 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:

$$\operatorname{supp}(()f_X) := \{x \in \mathcal{X} | f(x) > 0\} \tag{14.9}$$

Note: this is not a rigorous definition.

Theorem 14.1 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 14.2 Identically Distributed: From theorem 14.1 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 14.7 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:

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

Property 14.3 :

Monotonically Increasing $$x \leqslant y \iff \mathbb{F}_X(x) \leqslant \mathbb{F}_X(y) \quad \forall x, y \in \mathbb{R} \tag{14.10}$$

Upper Limit $$\lim_{x\to\infty} \mathbb{F}_X(x) = 1 \tag{14.11}$$

Lower Limit $$\lim_{x\to-\infty} \mathbb{F}_X(x) = 0 \tag{14.12}$$

Definition 14.8 CDF of a discret rv X: Let X be discret rv with pdf p_X, then the CDF of X is given by:

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

Definition 14.9 CDF of a continuous rv X: Let X be continuous rv with pdf f_X, then the CDF of X is given by:

$$\mathbb{F}_X(x) = \int_{-\infty}^{x} f_X(t)\,dt \qquad \Longleftrightarrow \qquad \frac{\partial \mathbb{F}_X(x)}{\partial x} = f_X(x)$$

Lemma 14.1 Probability Interval: Let X be a continuous rrv with pdf f_X and cumulative distribution function $\mathbb{F}_X$, then it holds that:

$$\mathbb{P}(a \leqslant X \leqslant b) = \mathbb{F}_X(b) - \mathbb{F}_X(a) \tag{14.13}$$

Proof. [def. 14.9]:

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

□

Proof. lemma 14.1:
$$\mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(X \leqslant b) - \mathbb{P}(X \leqslant a)$$
or by the fundamental theorem of calculus (theorem 4.2):
$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_X(t)\,\mathrm{d}t = \int_a^b \frac{\partial \mathbb{F}_X(t)}{\partial t}\,\mathrm{d}t = [\mathbb{F}_X(t)]|_a^b$$
□

Theorem 14.2 A continuous rv is fully characterized by its CDF: A cumulative distribution function completely determines the distribution of a continuous real-valued random variable.

Theorem 14.3
(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. 4.10]) 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:
$$p_Y(y) = \sum_{x_i \in \mathcal{X}_y} p_X(x_i) = \sum_{x \in \mathcal{Y}: g(x)=y} p_X(x) \qquad (14.14)$$

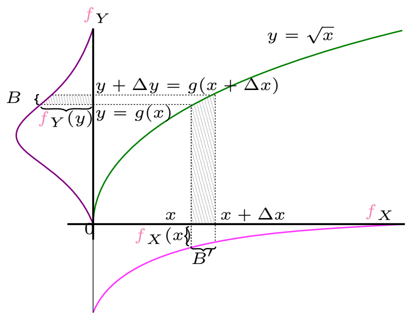
Proof. theorem 14.3:
$$Y = g(X) \iff \mathbb{P}(Y = y) = \mathbb{P}(x \in \mathcal{X}_y) = p_Y(y)$$
□

Theorem 14.4
(Scalar Continuous) Change of Variables: Let X be a continuous rv $X \in \mathcal{X}$ with pdf f_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. 4.10]) function.
Then the pdf of Y is given by:
$$f_Y(y) = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = f_X(x)\left|\frac{\mathrm{d}}{\mathrm{d}y}\left(g^{-1}(y)\right)\right| \qquad (14.15)$$
$$= f_X(x)\frac{1}{\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|} = \frac{f_X(g^{-1}(y))}{\left|\frac{\mathrm{d}g}{\mathrm{d}x}(g^{-1}(y))\right|} \qquad (14.16)$$

Theorem 14.5
(Continuous) Change of Variables: Let X be a continuous rv $X \in \mathcal{X}$ with pdf f_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. 4.10]) function.
Then the pdf of Y is given by:
$$f_Y(\boldsymbol{y}) = f_X(\boldsymbol{x})\left|\det\left(\frac{\partial g}{\partial \boldsymbol{x}}\right)\right|^{-1} \qquad (14.17)$$
$$= f_X(g^{-1}(\boldsymbol{y}))\left|\det\left(\frac{\partial g}{\partial \boldsymbol{x}}\right)\right|^{-1}$$
Where $\frac{\partial g}{\partial \boldsymbol{x}}$ is the Jacobian ([def. 5.4]).

Note
A monotonic function is required in order to satisfy inevitability.

Proof. theorem 14.4 (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. theorem 14.4 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:
$$\mathbb{F}_Y(y) = \mathbb{F}_X(g^{-1}(y))$$
$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{F}_X(g^{-1}(y)) = f_X(x)\cdot\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)$$
If g is monotonically decreasing:
$$\mathbb{F}_Y(y) = 1 - \mathbb{F}_X(g^{-1}(y))$$
$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{F}_X(g^{-1}(y)) = -f_X(x)\cdot\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)$$
□

Proof. theorem 14.4: **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:
$$y = g(x) \quad\Rightarrow\quad \mathbb{P}(y) = \mathbb{P}(g(x)) \text{ (for disc. rv.)}$$
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 (??)
$$g(x + \Delta x) \overset{\text{eq. (4.40)}}{=} g(x) + \frac{\mathrm{d}g}{\mathrm{d}x}\Delta y \qquad \text{for } \Delta x \to 0$$
$$= y + \Delta y \qquad \text{with } \Delta y := \frac{\mathrm{d}g}{\mathrm{d}x}\cdot\Delta x \qquad (14.18)$$
Thus for $\mathbb{P}(Y \in B')$ it follows:
$$\mathbb{P}(X \in B') = \int_y^{y+\Delta y} f_Y(t)\,\mathrm{d}t \xrightarrow{\Delta y \to 0} |\Delta y \cdot f_Y(y)|$$
$$= \left|\frac{\mathrm{d}g}{\mathrm{d}x}(x)\Delta x \cdot f_Y(y)\right|$$
Now we simply need to related the surface of the two pdfs:
$$B = [x, x + \Delta x] \overset{\text{same surfaces}}{\propto} [y, y + \Delta y] = B'$$
$$\mathbb{P}(Y \in B) = \mathbb{P}(X \in B')$$
$$\overset{\Delta y \to 0}{\iff} |f_Y(y)\cdot\Delta y| = \left|f_Y(y)\cdot\frac{\mathrm{d}g}{\mathrm{d}x}(x)\Delta x\right| = |f_X(x)\cdot\Delta x|$$
$$f_Y(y)\left|\cdot\frac{\mathrm{d}g}{\mathrm{d}x}(x)\right||\Delta x| = f_X(x)\cdot|\Delta x|$$
$$\Rightarrow f_Y(y) = \frac{f_X(x)}{\left|\frac{\mathrm{d}g}{\mathrm{d}x}(x)\right|} = \frac{f_X(g^{-1}(y))}{\left|\frac{\mathrm{d}g}{\mathrm{d}x}g^{-1}(y)\right|}$$
□

Rules of Probability

Definition 14.10 Marginalization/Sum Rule:
Given: $p_{x,y}(\bar{x}, \bar{y})$
$$p_x(\bar{x}) := \sum_{\bar{y} \in \mathcal{Y}} p_{x,y}(\bar{x}, \bar{y}) \qquad (14.19)$$

Definition 14.11 Conditioning:
Given: p_{xy}
$$p_{xy}(x | y = \bar{y}) := \frac{p_{xy}(x, y = \bar{y})}{p_y(y = \bar{y})}$$
if $p_y(\bar{y}) \neq 0 \qquad (14.20)$

Definition 14.12 Product Rule: follows directly from eq. (14.20)
$$p(x, y) = p(y|x)p_x(x) = p(x|y)p(y) \qquad (14.21)$$

Theorem 14.6 Total Probability Theorem: Given: $p_{x,y}(\bar{x}, \bar{y})$ with eq. (14.19) and eq. (14.21) it follows:
$$p_x(\bar{x}) \overset{\text{eq. (14.19)}}{=} \sum_{\bar{y} \in \mathcal{Y}} p_{x,y}(\bar{x}, \bar{y})$$
$$\overset{\text{eq. (14.21)}}{=} \sum_{\bar{y} \in \mathcal{Y}} p_{x|y}(\bar{x}|\bar{y})\,p_y(\bar{y}) \qquad (14.22)$$

Definition 14.13 Independence: Two random variables x and y are said to be **independent** if:
$$p(x|y) = p(x) \overset{eq.\ (14.20)}{\iff} p(x, y) = p(x)\,p(y) \qquad (14.23)$$

Corollary 14.3 eq. (14.23):
$$p(x|y) = p(x) \overset{\text{implies}}{\iff} p(y|x) = p(y) \qquad (14.24)$$

add mutual independence

0.2. Conditional PDF

Let x, y, z be R.V. (which themselves may be collections of random variables)

Definition 14.14 Marginalization:
$$p_{x|z}(\bar{x}|\bar{z}) = \sum_{\bar{y} \in \mathcal{Y}} p_{xy|z}(\bar{x}, \bar{y}|\bar{z}) \qquad (14.25)$$

Definition 14.15 Conditioning:
$$p_{x|yz}(\bar{x}|\bar{y}, \bar{z}) = \frac{p_{xy|z}(\bar{x}, \bar{y}|\bar{z})}{p_{y|z}(\bar{y}|\bar{z})} \qquad (14.26)$$

Definition 14.16 Product Rule: follows directly from eq. (14.26)
$$p_{xyz}(\bar{x}, \bar{y}|\bar{z}) = p_{x|yz}(\bar{x}|\bar{y}, \bar{z})\,p_{y|z}(\bar{y}|\bar{z}) \qquad (14.27)$$

Note
z basically parameterizes the pdf.

Definition 14.17 Conditional Independence: Two random variables x and y are said to be conditionally independent on z if
$$p(x|y, z) = p(x|z) \overset{eq.\ (14.26)}{=} p(x, y|z) = p(x|z)p(y|z) \qquad (14.28)$$
Hence, knowledge of z makes x and y independent.

Note
Conditional independence does not imply $p(x, y) = p(x)p(y)$

Rule 14.1 (Bayes' Rule). **Given**: *the prior* $p(X)$ *and the liklihood* $p(Y|X)$*, we can find the posterior by:*
$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(Y, X)}{\mathbb{P}(Y)} = \frac{\mathbb{P}(X)\mathbb{P}(Y|X)}{\mathbb{P}(Y)}$$
$$\overset{\text{normalization}}{=} \frac{\mathbb{P}(X)\mathbb{P}(Y|X)}{\sum_{X=x}\mathbb{P}(X = x)\mathbb{P}(Y|X = x)}$$
$$\textit{Posterior} = \frac{\textit{Prior}\cdot\textit{Liklihood}}{\textit{Normalization}}$$

Proof. Equation (14.25)
$$p_{x|z}(\bar{x}|\bar{z}) \overset{eq.\ (14.20)}{=} \frac{p_{xz}(\bar{x}, \bar{z})}{p_z(\bar{z})} \overset{eq.\ (14.19)}{=} \frac{\sum_{y \in \mathcal{Y}} p_{xyz}(\bar{x}, \bar{y}, \bar{z})}{p_z(\bar{z})}$$
$$\overset{eq.\ (14.21)}{=} \frac{\sum_{y \in \mathcal{Y}} p_{xy|z}(\bar{x}, \bar{y}|\bar{z})\cancel{p_z(\bar{z})}}{\cancel{p_z(\bar{z})}}$$
□

Proof. Equation (14.26)
$$p_{x|yz}(\bar{x}|\bar{y}, \bar{z}) \overset{eq.\ (14.20)}{=} \frac{p_{xyz}(\bar{x}, \bar{y}, \bar{z})}{p_{yz}(\bar{y}, \bar{z})}$$
$$\overset{eq.\ (14.21)}{=} \frac{p_{xy|z}(\bar{x}, \bar{y}|\bar{z})\cancel{p_z(\bar{z})}}{p_{y|z}(\bar{y}|\bar{z})\cancel{p_z(\bar{z})}}$$
□

Proof. Equation (14.24)
$$p(y|x) \overset{eq.\ (14.20)}{=} \frac{p(x, y)}{p(x)} \overset{p(x,y)=p(x)p(y)}{=} p(y)$$
□

Proof. Equation (14.28)
$$p(x|y, z) = p(x|z)$$
$$p(x|y, z) = \frac{p(x, y, z)}{p(y, z)} = \frac{p(x, y|z)\cancel{p(z)}}{p(y|z)\cancel{p(z)}} = \frac{p(x, y|z)}{p(y|z)}$$
$$\Rightarrow p(x, y|z) = p(x|z)p(y|z)$$
□

1. Key figures

1.1. The Expectation

Definition 14.18 Expectation (disc. case):
$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{x} \in \mathcal{X}} \bar{x}\,p_x(\bar{x}) \qquad (14.29)$$

Definition 14.19 Expectation (cont. case):
$$\mathbb{E}_x[x] := \int_{\bar{x} \in \mathcal{X}} \bar{x}\,f_x(\bar{x})\,\mathrm{d}\bar{x} \qquad (14.30)$$

Law 14.1 Expectation of independent variables:
$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] \qquad (14.31)$$

Property 14.4 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:
$$\mathbb{E}[a + b\boldsymbol{X} + c\boldsymbol{Y}] = a + b\mathbb{E}[\boldsymbol{X}] + c\mathbb{E}[\boldsymbol{Y}] \qquad (14.32)$$
Thus $\mathbb{E}$ is a linear operator ([def. 4.11]).

Note: Expectation of the expectation
The expectation of a r.v. X is a constant hence with Property 14.9 it follows:
$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \qquad (14.33)$$

Property 14.5 Matrix×Expectation: If $\boldsymbol{X} \in \mathbb{R}^n$ is a random vector and $\boldsymbol{A} \in \mathbb{R}^{m \times n}, \boldsymbol{B} \in \mathbb{R}^{n \times m}$ are constant matrices then it holds:
$$\mathbb{E}[\boldsymbol{AXB}] = \boldsymbol{A}\mathbb{E}[(\boldsymbol{XB})] = \boldsymbol{A}\mathbb{E}[\boldsymbol{X}]\,\boldsymbol{B} \qquad (14.34)$$

Proof. eq. (14.51):
$$\mathbb{E}[XY] = \sum_{x \in \mathcal{X}}\sum_{y \in \mathcal{Y}} p_{X,Y}(x, y)xy$$
$$\overset{\text{[def. 14.13]}}{=} \sum_{x \in \mathcal{X}} p_X(x)x \sum_{y \in \mathcal{Y}} p_Y(y)y = \mathbb{E}[X]\,\mathbb{E}[Y]$$
□

Law 14.2 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:

$$\mathbb{E}_Y[y] = \sum_{y\in\mathcal{Y}} y p_Y(y) = \sum_{x\in\mathcal{X}} g(x)p_x(x) \quad \text{or integral for CRV} \tag{14.35}$$

Consequence

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

Theorem 14.7 Jensen's Inequality: If X is a random variable and f is a convex function, then it holds that:

$$f(\mathbb{E}[X]) \leqslant \mathbb{E}[f(X)] \tag{14.36}$$

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

$$f(\mathbb{E}[X]) \geqslant \mathbb{E}[f(X)] \tag{14.37}$$

1.2. The Variance

Definition 14.20 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:

$$\mathbb{V}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \tag{14.38}$$

Proof. eq. (14.58)

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right]$$
$$\overset{\text{Property 14.9}}{=} \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}\left[X^2\right] - \mu^2$$

□

Property 14.6 Variance of a Constant: If $a \in \mathbb{R}$ is a constant then it follows that its expected value is deterministic $\Rightarrow$ we have no uncertainty $\Rightarrow$ no variance:

$$\mathbb{V}[a] = 0 \qquad \textbf{with} \qquad a \in \mathbb{R} \tag{14.39}$$

Property 14.7 Affine Transformation: 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:

$$\mathbb{V}[\boldsymbol{A}\boldsymbol{X} + b] = \boldsymbol{A}\mathbb{V}[X]\boldsymbol{A}^{\intercal} \tag{14.40}$$

Proof. Property 14.13

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

□

Definition 14.21 Covariance: The Covariance is a measure of how much two or more random variables vary linearly with each other.

$$\begin{aligned}
\mathrm{Cov}[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned} \tag{14.41}$$

Proof. eq. (14.62)

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

□

Definition 14.22 Covariance Matrix: The variance of a k-dimensional random vector $\boldsymbol{X} = (X_1 \ \ldots \ X_k)$ is given by the 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:

$$\begin{aligned}
\mathbb{V}[\boldsymbol{X}] :=& \Sigma(\boldsymbol{X}) := \mathrm{Cov}[\boldsymbol{X}, \boldsymbol{X}] := \\
=& \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^{\intercal}\right] \\
=& \mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^{\intercal}\right] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^{\intercal} \in [-\infty, \infty]
\end{aligned} \tag{14.42}$$

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

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

Note: Covariance and Variance

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

$$\mathrm{Cov}[X, X] = \mathbb{V}[X] \equiv \sigma^2(X) \equiv \sigma_X^2 \tag{14.43}$$

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

Property 14.8 Translation and Scaling:

$$\mathrm{Cov}(a + bX, c + dY) = bd\mathrm{Cov}(X, Y) \tag{14.44}$$

Definition 14.23 Correlation Coefficient: Is the standardized version of the covariance:

$$\begin{aligned}
\mathrm{Corr}[\boldsymbol{X}] : &= \frac{\mathrm{Cov}[\boldsymbol{X}]}{\sigma_{X_1} \cdots \sigma_{X_k}} \in [-1, 1] \\
&= \begin{cases} +1 & \text{if } Y = aX + b \text{ with } a > 0, b \in \mathbb{R} \\ -1 & \text{if } Y = aX + b \text{ with } a < 0, b \in \mathbb{R} \end{cases}
\end{aligned} \tag{14.45}$$

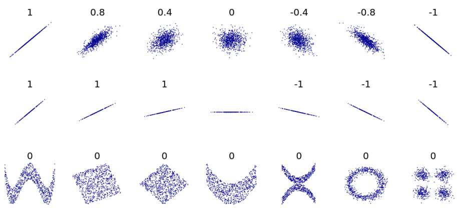


Figure 1: Several sets of (x, y) points, with their correlation coefficient

Law 14.3 Translation and Scaling:

$$\mathrm{Corr}(a + bX, c + dY) = \mathrm{sign}(b)\mathrm{sign}(d)\mathrm{Cov}(X, Y) \tag{14.46}$$

Note

- The correlation/covariance reflects the noisiness and direction of a linear relationship (top row fig. 2), **but** not the slope of that relationship (middle row fig. 2) nor many aspects of nonlinear relationships (bottom row)
- The set in the center of fig. 2 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 14.7.
2. Correlation is dimensionless, whereas the unit of the covariance is obtained by the product of the units of the two RV variables.

Law 14.4 Covariance of independent RVs: The covariance/correlation of two independent variable's ([def. 14.13]) is zero:

$$\begin{aligned}
\mathrm{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
&\overset{\text{eq. (14.51)}}{=} \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0
\end{aligned}$$

Zero covariance/correlation$\nRightarrow$ independence

$$\mathrm{Cov}(X, Y) = \mathrm{Corr}(X, Y) = 0 \nRightarrow p_{X,Y}(x, y) = p_X(x)p_Y(y)$$

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:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathrm{Cov}(X, X^2) = \mathbb{E}\left[X \cdot X^2\right] - \mathbb{E}[X]\mathbb{E}\left[X^2\right] \\
&= \mathbb{E}\left[X^3\right] - \mathbb{E}[X]\mathbb{E}\left[X^2\right] \underset{\text{eq. (14.78)}}{\overset{\text{eq. (14.88)}}{=}} 0 - 0 \cdot \mathbb{E}\left[X^2\right]
\end{aligned}$$

$\Rightarrow$ the relationship between Y and X must be non-linear.

Definition 14.24 Autocorrelation/Crosscorrelation $\gamma(t_1, t_2)$: Describes the covariance ([def. 14.28]) between the two values of a stochastic process $(\boldsymbol{X}_t)_{t\in T}$ at different time points t_1 and t_2.

$$\gamma(t_1, t_2) = \mathrm{Cov}\left[\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}\right] = \mathbb{E}\left[\left(\boldsymbol{X}_{t_1} - \mu_{t_1}\right)\left(\boldsymbol{X}_{t_2} - \mu_{t_2}\right)\right] \tag{14.47}$$

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

$$\gamma(t, t) = \mathrm{Cov}\left[\boldsymbol{X}_t, \boldsymbol{X}_t\right] \overset{\text{eq. (14.64)}}{=} \mathbb{V}[\boldsymbol{X}_t] \tag{14.48}$$

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 autocorrelation is maximized/most similar for no translation $\tau = 0$ at all.

2. Key Figures

2.1. The Expectation

Definition 14.25 Expectation (disc. case):

$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{\boldsymbol{x}}\in\mathcal{X}} \bar{\boldsymbol{x}} p_x(\bar{\boldsymbol{x}}) \tag{14.49}$$

Definition 14.26 Expectation (cont. case):

$$\mathbb{E}_x[x] := \int_{\bar{\boldsymbol{x}}\in\mathcal{X}} \bar{\boldsymbol{x}} f_x(\bar{\boldsymbol{x}})\,\mathrm{d}\bar{\boldsymbol{x}} \tag{14.50}$$

Law 14.5 Expectation of independent variables:

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \tag{14.51}$$

Property 14.9 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:

$$\mathbb{E}[a + b\boldsymbol{X} + c\boldsymbol{Y}] = a + b\mathbb{E}[\boldsymbol{X}] + c\mathbb{E}[\boldsymbol{Y}] \tag{14.52}$$

Thus $\mathbb{E}$ is a linear operator[def. 4.11].

Note: Expectation of the expectation

The expectation of a r.v. X is a constant hence with Property 14.9 it follows:

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \tag{14.53}$$

Property 14.10 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:

$$\mathbb{E}[\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}] = \boldsymbol{A}\mathbb{E}[(\boldsymbol{X}\boldsymbol{B})] = \boldsymbol{A}\mathbb{E}[\boldsymbol{X}]\boldsymbol{B} \tag{14.54}$$

Proof. eq. (14.51):

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{X,Y}(x, y)xy \\
&\overset{\text{[def. 14.13]}}{=} \sum_{x\in\mathcal{X}} p_X(x)x \sum_{y\in\mathcal{Y}} p_Y(y)y = \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}$$

□

Law 14.6 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:

$$\mathbb{E}_Y[y] = \sum_{y\in\mathcal{Y}} y p_Y(y) = \sum_{x\in\mathcal{X}} g(x)p_x(x) \quad \text{or integral for CRV} \tag{14.55}$$

Consequence

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

Theorem 14.8 Jensen's Inequality: If X is a random variable and f is a convex function, then it holds that:

$$f(\mathbb{E}[X]) \leqslant \mathbb{E}[f(X)] \tag{14.56}$$

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

$$f(\mathbb{E}[X]) \geqslant \mathbb{E}[f(X)] \tag{14.57}$$

2.2. The Variance

Definition 14.27 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:

$$\mathbb{V}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \overset{\text{see section 3}}{=} \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \tag{14.58}$$

2.2.1. Properties

Property 14.11 Variance of a Constant: If $a \in \mathbb{R}$ is a constant then it follows that its expected value is deterministic $\Rightarrow$ we have no uncertainty $\Rightarrow$ no variance:

$$\mathbb{V}[a] = 0 \qquad \textbf{with} \qquad a \in \mathbb{R} \tag{14.59}$$

see shift and scaling for proof section 3

Property 14.12 Shifting and Scaling:

$$\mathbb{V}[a+bX] = a^2\sigma^2 \qquad \text{with} \qquad a \in \mathbb{R} \tag{14.60}$$

see section 3

Property 14.13 Affine Transformation: 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:

$$\mathbb{V}[\boldsymbol{AX}+b] = \boldsymbol{A}\mathbb{V}[\boldsymbol{X}]\boldsymbol{A}^\intercal \tag{14.61}$$

Proof. Property 14.13

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

□

Definition 14.28 Covariance: The Covariance is a measure of how much two or more random variables vary linearly with each other.

$$\begin{aligned}
\mathrm{Cov}[X,Y] &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned} \tag{14.62}$$

Proof. eq. (14.62)

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

□

Definition 14.29 Covariance Matrix: The variance of a k-dimensional random vector $\boldsymbol{X} = (X_1 \ \ldots \ X_k)$ is given by the 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:

$$\begin{aligned}
\mathbb{V}[\boldsymbol{X}] :=& \Sigma(\boldsymbol{X}) := \mathrm{Cov}[\boldsymbol{X},\boldsymbol{X}] := \\
=& \mathbb{E}\left[(\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}])(\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}])^\intercal\right] \\
=& \mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^\intercal\right] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\intercal \in [-\infty,\infty]
\end{aligned} \tag{14.63}$$

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

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

Note: Covariance and Variance

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

$$\mathrm{Cov}[X,X] = \mathbb{V}[X] \equiv \sigma^2(X) \equiv \sigma_X^2 \tag{14.64}$$

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

Property 14.14 Translation and Scaling:

$$\mathrm{Cov}(a+bX, c+dY) = bd\mathrm{Cov}(X,Y) \tag{14.65}$$

Definition 14.30 Correlation Coefficient: Is the standardized version of the covariance:

$$\begin{aligned}
\mathrm{Corr}[\boldsymbol{X}] :&= \frac{\mathrm{Cov}[\boldsymbol{X}]}{\sigma_{X_1}\cdots\sigma_{X_k}} \in [-1,1] \\
&= \begin{cases} +1 & \text{if } Y = aX+b \text{ with } a>0, b\in\mathbb{R} \\ -1 & \text{if } Y = aX+b \text{ with } a<0, b\in\mathbb{R} \end{cases}
\end{aligned} \tag{14.66}$$

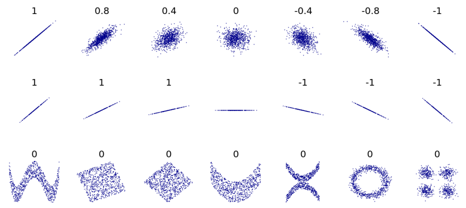


Figure 2: Several sets of (x, y) points, with their correlation coefficient

Law 14.7 Translation and Scaling:

$$\mathrm{Corr}(a+bX, c+dY) = \mathrm{sign}(b)\mathrm{sign}(d)\mathrm{Cov}(X,Y) \tag{14.67}$$

Note

- The correlation/covariance reflects the noisiness and direction of a linear relationship (top row fig. 2), **but** not the slope of that relationship (middle row fig. 2) nor many aspects of nonlinear relationships (bottom row)
- The set in the center of fig. 2 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 14.7.
2. Correlation is dimensionless, whereas the unit of the covariance is obtained by the product of the units of the two RV variables.

Law 14.8 Covariance of independent RVs: The covariance/correlation of two independent variable's ([def. 14.13]) is zero:

$$\begin{aligned}
\mathrm{Cov}[X,Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
&\overset{\text{eq. (14.51)}}{=} \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0
\end{aligned}$$

Zero covariance/correlation $\not\Rightarrow$ independence

$$\mathrm{Cov}(X,Y) = \mathrm{Corr}(X,Y) = 0 \not\Rightarrow p_{X,Y}(x,y) = p_X(x)p_Y(y)$$

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:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathrm{Cov}(X,X^2) = \mathbb{E}\left[X\cdot X^2\right] - \mathbb{E}[X]\mathbb{E}\left[X^2\right] \\
&= \mathbb{E}\left[X^3\right] - \mathbb{E}[X]\mathbb{E}\left[X^2\right] \underset{\text{eq. (14.78)}}{\overset{\text{eq. (14.88)}}{=}} 0 - 0\cdot\mathbb{E}\left[X^2\right]
\end{aligned}$$

$\Rightarrow$ the relationship between Y and X must be non-linear.

Definition 14.31 Quantile: Are specific values q_α in the range[def. 4.6] of a random variable X that are defined as the value for which the cumulative probability is less then $\alpha \in (0,1)$:

$$q_\alpha : \mathbb{P}(X \leqslant x) = \mathbb{F}_X(q_\alpha) = \alpha \quad \xrightarrow{\mathbb{F}\text{ invert.}} \quad q_\alpha = \mathbb{F}_X^{-1}(\alpha) \tag{14.68}$$

add figure

3. Proofs

Proof. eq. (14.58)

$$\begin{aligned}
\mathbb{V}[X] &= \mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\
&\overset{\text{Property 14.9}}{=} \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}\left[X^2\right] - \mu^2
\end{aligned}$$

□

Proof. Property 14.12

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

□

Discrete Distributions

4.1. Bernoulli Distribution $\mathrm{Bern}(p)$

Definition 14.32 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 14.33 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:

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

$$\mathbb{E}[X] = p \qquad (14.69) \qquad \mathbb{V}[X] = p(1 - p) \qquad (14.70)$$

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

Definition 14.34 Binomial Distribution: Models the probability of exactly X success given a fixed number *n-Bernoulli experiments*[def. 14.32], 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

$$\mathbb{E}[X] = np \qquad (14.71) \qquad \mathbb{V}[X] = np(1 - p) \qquad (14.72)$$

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:

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

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 14.35 Geometric Distribution $\mathrm{Geom}(p)$: Models the probability of the number X of Bernoulli trials[def. 14.32] *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*

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

$$\mathbb{E}[X] = \frac{1}{p} \qquad (14.73) \qquad \mathbb{V}[X] = \frac{1 - p}{p^2} \qquad (14.74)$$

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 14.36 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.

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!} \qquad \begin{matrix} \lambda > 0 \\ x \in \mathbb{Z}_{\geqslant 0} \end{matrix} \qquad (14.75)$$

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

$$\mathbb{E}[X] = \lambda \qquad (14.76) \qquad \mathbb{V}[X] = \lambda \qquad (14.77)$$

Continuous Distributions

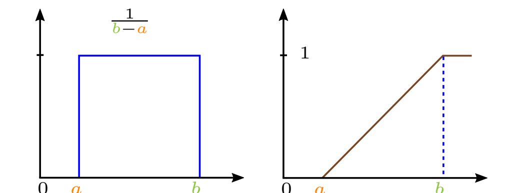
5.1. Uniform Distribution $\mathcal{U}(a, b)$

Definition 14.37 Uniform Distribution $\mathcal{U}(a, b)$: Is probability distribution, where all intervals of the **same** length on the distribution's support ([def. 14.6]) $\mathrm{supp}(\mathcal{U}[a, b]) = [a, b]$ are equally probable/likely.

$$f(x) = \frac{1}{b - a} \mathbb{1}_{x \in [a;b)} = \begin{cases} \frac{1}{b-a} = \text{const} & \text{if} \quad a \leqslant x \leqslant b \\ 0 & \text{else} \end{cases} \qquad (14.78)$$

$$\mathbb{F}(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & \text{if} \quad a \leqslant x \leqslant b \\ 1 & x > b \end{cases} \qquad (14.79)$$

$$\mathbb{E}[X] = \frac{a + b}{2} \qquad \mathbb{V}(X) = \frac{(b - a)^2}{12} \qquad (14.80)$$



(a) $f_X(x)$ (b) $\mathbb{F}_X(x)$

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

Definition 14.38 Exponential Distribution $X \sim \exp(\lambda)$: Is the continuous analogue to the geometric distribution [def. 14.35].
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.

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if} \quad x \geqslant 0 \\ 0 & x < 0 \end{cases} \qquad (14.81)$$

$$\mathbb{F}(x; \lambda) = \begin{cases} 1 - e^{-\lambda x} & \text{if} \quad x \geqslant 0 \\ 0 & x < 0 \end{cases} \qquad (14.82)$$

$$\mathbb{E}[X] = \frac{1}{\lambda} \qquad \mathbb{V}(X) = \frac{1}{\lambda^2} \qquad (14.83)$$

5.3. Laplace Distribution

Definition 14.39 Laplace Distribution:

Laplace Distibution
$$f(\boldsymbol{x}; \mu, \sigma) = \frac{1}{2\sigma} \exp\left(-\frac{|\boldsymbol{x} - \mu|}{\sigma}\right) \qquad (14.84)$$

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

Definition 14.40 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:

$$\mathbb{E}[X] = \mu \qquad \mathbb{V}(X) = \sigma^2 \qquad (14.85)$$

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right\} \qquad (14.86)$$

$$\mathbb{F}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{1}{2}\left(\frac{u - \mu}{\sigma}\right)^2\right\} du \qquad (14.87)$$
$$x \in \mathbb{R} \quad \text{or} \quad -\infty < x < \infty$$

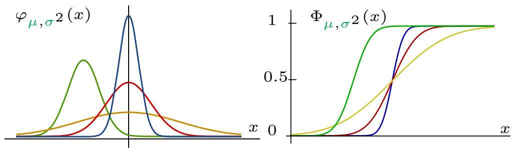


Figure 4: $\mu = 0$, $\sigma^2 = 0.2$; $\mu = 0$, $\sigma^2 = 1.0$; $\mu = 0$, $\sigma^2 = 5.0$; $\mu = -2$, $\sigma^2 = 0.5$

Property 14.15 : $\mathbb{P}_X(\mu - \sigma \leqslant x \leqslant \mu - \sigma) = 0.66$

Property 14.16 : $\mathbb{P}_X(\mu - 2\sigma \leqslant x \leqslant \mu - 2\sigma) = 0.95$

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

Historic Problem: the cumulative distribution eq. (14.87) 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 14.41
Standard Normal Distribution $\boldsymbol{X} \sim \mathcal{N}(0, 1)$:

$$\mathbb{E}[X] = 0 \qquad \mathbb{V}(X) = 1 \qquad (14.88)$$

$$f(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad (14.89)$$

$$\mathbb{F}(x; 0, 1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}u^2} du \qquad (14.90)$$
$$x \in \mathbb{R} \quad \text{or} \quad -\infty < x < \infty$$

Corollary 14.4
Standard Normal Distribution Notation: As the standard normal distribution is so commonly used people often use the letter Z in order to denote its the *standard* normal distribution and its α-quantile[def. 14.31] is then denoted by:

$$z_\alpha = \Phi^{-1}(\alpha) \qquad \alpha \in (0, 1) \qquad (14.91)$$

5.5.1. Calculating Probabilities

Property 14.17 Symmetry: Let $z > 0$

$$\mathbb{P}(Z \leqslant z) = \Phi(z) \qquad (14.92)$$
$$\mathbb{P}(Z \leqslant -z) = \Phi(-z) = 1 - \Phi(z) \qquad (14.93)$$
$$\mathbb{P}(-a \leqslant Z \leqslant b) = \Phi(b) - \Phi(-a) = \Phi(b) - (1 - \Phi(a))$$
$$\overset{a=b=z}{=} 2\Phi(z) - 1 \qquad (14.94)$$

5.5.2. Linear Transformations of Normal Dist.

Proposition 14.1 Linear 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:

$$Y \sim \mathcal{N}\left(a + b\mu, b^2\sigma^2\right) \iff f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y - a}{b}\right) \qquad (14.95)$$

section 2

Proposition 14.2 Standardization: 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:

$$X \sim \mathcal{N}(\mu, \sigma^2) \xrightarrow{Z = \frac{X - \mu}{\sigma}} Z \sim \mathcal{N}(0, 1) \qquad (14.96)$$

section 2

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 14.3 Standardization 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:

$$F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) \qquad (14.97)$$

section 2

6. The Multivariate Normal distribution

Definition 14.42
Multivariate Normal distribution $\boldsymbol{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$:
The k-multivariate Normal distribution of:
$\boldsymbol{X} = (x_1 \ \dots \ x_k)^\intercal$ a k-dimensional random vector with:
$\boldsymbol{\mu} = (\mathbb{E}[\boldsymbol{x}_1] \ \dots \ \mathbb{E}[\boldsymbol{x}_k])^\intercal$ a k-dim mean vector
and $k \times k$ **p.s.d.**covariance matrix:

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

is given by:

$$f_{\boldsymbol{X}}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k) = \tag{14.98}$$
$$= \underbrace{\frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}}}_{\text{Normalisation}} \exp\left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^\intercal \Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right)$$

Definition 14.43 Jointly Gaussian Random Variables:
Two random variables $\boldsymbol{x}$, $\boldsymbol{y}$ both scalars or vectors, are said to be jointly Gaussian if the joint vector random variable $[\boldsymbol{x} \quad \boldsymbol{y}]^\intercal$ is again a GRV.

Corollary 14.5 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. 14.43]).

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y}) \tag{14.99}$$
$$\propto \exp\left(-\frac{1}{2}\left\{(\boldsymbol{x}-\mu_x)^\intercal \Sigma_x^{-1}(\boldsymbol{x}-\mu_x) + (\boldsymbol{y}-\mu_y)^\intercal \Sigma_y^{-1}(\boldsymbol{y}-\mu_y)\right\}\right)$$
$$= \exp\left(-\frac{1}{2}[(\boldsymbol{x}-\mu_x)^\intercal \quad (\boldsymbol{y}-\mu_y)^\intercal]\begin{bmatrix} 0 & \Sigma_x^{-1} \\ \Sigma_y^{-1} & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{x}-\mu_x \\ \boldsymbol{y}-\mu_y \end{bmatrix}\right)$$

Property 14.18 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:

$$\boxed{x \sim \mathcal{N}(\mu_x, \sigma_x^2) = \mathcal{N}(a\mu + b, a^2\sigma^2)} \tag{14.100}$$

Property 14.19 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 2).

Property 14.20 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 2).

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:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k^2 \end{bmatrix} \quad \text{and} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} \tag{14.101}$$

eq. (14.98) decomposed s.t. $x_1, \dots, x_k$ become mutal independent (??):

$$p(\boldsymbol{X}) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right) \tag{14.102}$$

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

Definition 14.44 Gamma Distribution $X \sim \Gamma(x, \alpha, \beta)$:
Is a widely used distribution that is related to the exponential distribution, Erlang distribution, and chi-squared distribution as well as Normal distribution:

$$f(x; \alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{if} \quad x > 0 \\ 0 & \quad x \leqslant 0 \end{cases} \tag{14.103}$$
$$\Gamma(\alpha) \overset{\text{eq. (4.65)}}{=} \int_0^\infty t^{\alpha-1} e^{-t}\, dt \tag{14.104}$$

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

6.2. Delta Distribution

Definition 14.45 The delta function $\delta(\boldsymbol{x})$:
The delta/dirac function $\delta(\boldsymbol{x})$ is defined by:

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

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

$$\delta(x - x_0) = \lim_{\sigma \to 0} \mathcal{N}(x|x_0, \sigma) \tag{14.105}$$
$$\approx \infty \mathbb{1}_{\{x = x_0\}} \tag{14.106}$$

Property 14.21 Properties of δ:

- **Normalization**: The delta function integrates to 1:
$$\int_{\mathbb{R}} \delta(x)\, dx = \int_{\mathbb{R}} \delta(x) \cdot c_1(x)\, dx = c_1(0) = 1$$
where $c_1(x) = 1$ is the constant function of value 1.
- **Shifting**:
$$\int_{\mathbb{R}} \delta(x - x_0) f(x)\, dx = f(x_0) \tag{14.107}$$
- **Symmetry**: $\int_{\mathbb{R}} \delta(-x) f(x)\, dx = f(0)$
- **Scaling**: $\int_{\mathbb{R}} \delta(\alpha x) f(x)\, dx = \frac{1}{|\alpha|} f(0)$

Note

- In mathematical terms δ is not a function but a **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. (14.105) would be a non-differentiable/discret form of the dirac measure.

Proofs

Proof. proposition 14.1: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$\mathbb{F}_Y(y) \overset{y \geqslant 0}{=} \mathbb{P}_Y(Y \leqslant y) = \mathbb{P}(a + bX \leqslant y) = \mathbb{P}_X\left(X \leqslant \frac{y-a}{b}\right)$$
$$= \mathbb{F}_X\left(\frac{y-a}{b}\right)$$
$$\mathbb{F}_Y(y) \overset{y \leqslant 0}{=} \mathbb{P}_Y(Y \leqslant y) = \mathbb{P}(a + bX \leqslant y) = \mathbb{P}_X\left(X \geqslant \frac{y-a}{b}\right)$$
$$= 1 - \mathbb{F}_X\left(\frac{y-a}{b}\right)$$

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

$$f_Y(y) = \frac{d\mathbb{F}_Y(y)}{dy} = \begin{cases} \frac{1}{b}\frac{d\mathbb{F}_X\left(\frac{y-a}{b}\right)}{dy} \\ \frac{1}{-b}\frac{d\mathbb{F}_X\left(\frac{y-a}{b}\right)}{dy} \end{cases} = \frac{1}{|b|} f_X(x)\left(\frac{y-a}{b}\right)$$

eq. (14.95)).
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 14.2: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$Z := \frac{X - \mu}{\sigma} = \frac{1}{std}X - \frac{\mu}{\sigma} = aX + b \quad \text{with } a = \frac{1}{\sigma}, b = -\frac{\mu}{\sigma}$$
$$\overset{\text{eq. (14.95)}}{\sim} \mathcal{N}\left(a\mu + b, a^2\sigma^2\right) \sim \mathcal{N}\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{\sigma^2}{\sigma^2}\right) \sim \mathcal{N}(0, 1)$$

□

Proof. proposition 14.3: 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 14.19 scalar case
Let $y \sim p(y) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$ and define $\boldsymbol{x} = a\boldsymbol{y} + b \quad a \in \mathbb{R}_+, b \in \mathbb{R}$
Using the Change of variables formula it follows:

$$p_{\boldsymbol{x}}(\bar{x}) \overset{??}{=} \frac{p_y(\bar{y})}{\left|\frac{dx}{dy}\right|} \overset{\bar{y} = \frac{\bar{x}-b}{a}}{=} \frac{1}{a}\frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{1}{2\sigma^2}\left(\overbrace{\frac{\bar{x}-b}{a}}^{\bar{y}(\bar{x})} - \mu\right)^2\right)$$
$$= \frac{1}{\sqrt{2\pi a^2\mu^2}} \exp\left(-\frac{1}{2\sigma^2 a^2}\left(\bar{x} \underbrace{-b - a\mu}_{\mu_x}\right)^2\right)$$

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

□

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 14.20
From Property 14.19 it follows immediately that $\boldsymbol{z}$ is GRV $\boldsymbol{z} \sim \mathcal{N}(\mu_z, \Sigma_z)$ with:

$$\boldsymbol{z} = \boldsymbol{A}\xi \quad \textbf{with} \quad \boldsymbol{A} = [\boldsymbol{A}_x \quad \boldsymbol{A}_y] \text{ and } \xi = (\boldsymbol{x} \quad \boldsymbol{y})$$

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

$$\mathbb{E}[\boldsymbol{z}] = \mathbb{E}[\boldsymbol{A}_x x + \boldsymbol{A}_y y] = \boldsymbol{A}_x\mu_x + \boldsymbol{A}_y\mu_y$$
$$\mathbb{V}[\boldsymbol{z}] = \mathbb{V}[\boldsymbol{A}\xi] \overset{\text{Property 14.13}}{=} \boldsymbol{A}\mathbb{V}[\xi]\boldsymbol{A}^\intercal$$
$$= [\boldsymbol{A}_x \quad \boldsymbol{A}_y]\begin{bmatrix} \mathbb{V}[x] & \mathrm{Cov}[x, y] \\ \mathrm{Cov}[y, x] & \mathbb{V}[y] \end{bmatrix}[\boldsymbol{A}_x \quad \boldsymbol{A}_y]^\intercal$$
$$= [\boldsymbol{A}_x \quad \boldsymbol{A}_y]\begin{bmatrix} \mathbb{V}[x] & \mathrm{Cov}[x, y] \\ \mathrm{Cov}[y, x] & \mathbb{V}[y] \end{bmatrix}\begin{bmatrix} \boldsymbol{A}_x^\intercal \\ \boldsymbol{A}_y^\intercal \end{bmatrix}$$
$$= \boldsymbol{A}_x\mathbb{V}[x]\boldsymbol{A}_x^\intercal + \boldsymbol{A}_y\mathbb{V}[y]\boldsymbol{A}_y^\intercal$$
$$+ \underbrace{\boldsymbol{A}_y\mathrm{Cov}[y, x]\boldsymbol{A}_x^\intercal}_{=0\text{by independence}} + \underbrace{\boldsymbol{A}_x\mathrm{Cov}[x, y]\boldsymbol{A}_y^\intercal}_{=0\text{by independence}}$$
$$= \boldsymbol{A}_x\Sigma_x\boldsymbol{A}_x^\intercal + \boldsymbol{A}_y\Sigma_y\boldsymbol{A}_y^\intercal$$

□

Note

Can also be proofed by using the normal definition of [def. 14.27] and tedious computations.

7. Sampling Random Numbers

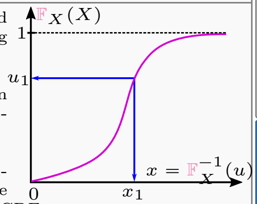
Most math libraries have uniform **random number generator (RNG)** i.e. functions to generate uniformly distributed random numbers $U \sim \mathcal{U}[a,b]$ (eq. (14.78)).
Furthermore repeated calls to these RNG are independent, that is:

$$p_{U_1,U_2}(u_1,u_2) \overset{\text{eq. (14.23)}}{=} p_{U_1}(u_1)\cdot p_{U_2}(u_2) = \begin{cases}1 & \text{if } u_1,u_2 \in [a,b]\\ 0 & \text{otherwise}\end{cases}$$

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

8. Inverse-transform Technique

Idea



Can make use of section 1 and the fact that CDF are increasing functions ([def. 4.8]). **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.

8.1. Continuous Case

Definition 14.46 One Continuous Variable: **Given**: a desired continuous pdf f_X and uniformly distributed rn $\{u_1,u_2,\dots\}$:

1. Integrate the desired pdf f_X in order to obtain the desired cdf $\mathbb{F}_X$:
$$\mathbb{F}_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \qquad (14.108)$$
2. Set $\mathbb{F}_X(X) \overset{!}{=} U$ on the range of X with $U \sim \mathcal{U}[0,1]$.
3. Invert this equation/find the inverse $\mathbb{F}_X^{-1}(U)$ i.e. solve:
$$U = \mathbb{F}_X(X) = \mathbb{F}_X\Big(\underbrace{\mathbb{F}_X^{-1}(U)}_{X}\Big) \qquad (14.109)$$
4. Plug in the uniformly distributed rn:
$$x_i = \mathbb{F}_X^{-1}(u_i) \quad \textbf{s.t.} \quad x_i \sim f_X \qquad (14.110)$$

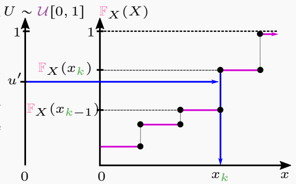
Definition 14.47 Multiple Continuous Variable:
Given: a pdf of multiple rvs $f_{X,Y}$:
1. Use the product rule (eq. (14.21)) in order to decompose $f_{X,Y}$:
$$f_{X,Y} = f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) \qquad (14.111)$$
2. Use [def. 14.48] to first get a rv for y of $Y \sim f_Y(y)$.
3. Then with this fixed y use [def. 14.48] again to get a value for x of $X \sim f_{X|Y}(x|y)$.

Proof. [def. 14.48]:
Claim: if U is a uniform rv on $[0,1]$ then $\mathbb{F}_X^{-1}(U)$ has $\mathbb{F}_X$ as its CDF.
Assume that $\mathbb{F}_X$ is strictly increasing ([def. 4.8]).
Then for any $u \in [0,1]$ there must exist a **unique** x s.t. $\mathbb{F}_X(x) = u$.
Thus $\mathbb{F}_X$ must be invertible and we may write $x = \underline{\mathbb{F}_X^{-1}(u)}$.
Now let a arbitrary:
$$\mathbb{F}_X(a) = \mathbb{P}(\underline{x} \leqslant a) = \mathbb{P}(\mathbb{F}_X^{-1}(U) \leqslant a)$$
Since $\mathbb{F}_X$ is strictly increasing:
$$\mathbb{P}\left(\mathbb{F}_X^{-1}(U) \leqslant a\right) = \mathbb{P}(U \leqslant \mathbb{F}_X(a)) \overset{\text{eq. (14.78)}}{=} \int_0^{\mathbb{F}_X(a)} 1\,\mathrm{d}t = \mathbb{F}_X(a)$$
□

Note
Strictly speaking we may not assume that a CDF is strictly increasing but we as all CDFs are weakly increasing ([def. 4.8]) we may always define an auxiliary function by its infimum:
$$\hat{\mathbb{F}}_X^{-1} := \inf\{x|\mathbb{F}_X(X) \geqslant 0\} \qquad u \in [0,1] \qquad (14.112)$$

8.2. 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 $\{u_1,u_2,\dots\}$.
Goal: given a uniformly distributed rn u determine k s.t.:
$$\sum_{i=1}^{k-1} < U \leqslant \sum_{i=1}^{k} \iff \mathbb{F}_X(x_{k-1}) < u \leqslant \mathbb{F}_X(x_k) \qquad (14.113)$$
and return x_k.

Definition 14.48 One Discret Variable:
1. Compute the CDF of p_X ([def. 14.8])
$$\mathbb{F}_X(x) = \sum_{t=-\infty}^{x} p_X(t) \qquad (14.114)$$
2. Given the uniformly distributed rn $\{u_i\}_{i=1}^n$ find k^i (≙ inversion) s.t.:
$$\mathbb{F}_X\left(x_{k(i)-1}\right) < u_i \leqslant \mathbb{F}_X\left(x_{k(i)}\right) \qquad \forall u_i \qquad (14.115)$$

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:
$$\mathbb{F}_X(x_{k-1}) < u \leqslant \mathbb{F}_X(x_k) \qquad (14.116)$$
Now observe that:
$$u \leqslant \mathbb{F}_X(x_k) = \mathbb{F}_X(x_{k-1}) + p_X(x_k)$$
$$\Rightarrow \mathbb{F}_X(x_{k-1}) < u \leqslant \mathbb{F}_X(x_{k-1}) + p_X(x_k)$$
The probability of U being in $(\mathbb{F}_X(x_{k-1}), \mathbb{F}_X(x_k)]$ is:
$$\mathbb{P}\left(U \in [\mathbb{F}_X(x_{k-1}), \mathbb{F}_X(x_k)]\right) = \int_{\mathbb{F}_X(x_{k-1})}^{\mathbb{F}_X(x_k)} p_U(t)\,\mathrm{d}t$$
$$= \int_{\mathbb{F}_X(x_{k-1})}^{\mathbb{F}_X(x_k)} 1\,\mathrm{d}t = \int_{\mathbb{F}_X(x_{k-1})}^{\mathbb{F}_X(x_{k-1})+p_X(x_k)} 1\,\mathrm{d}t = p_X(x_k)$$
Hence the random variable $x_k \in \mathcal{X}$ has the pdf p_X. □

Definition 14.49
Multiple Continuous Variables (Option 1):
Given: a pdf of multiple rvs $p_{X,Y}$:
1. Use the product rule (eq. (14.21)) in order to decompose $p_{X,Y}$:
$$p_{X,Y} = p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y) \qquad (14.117)$$
2. Use **??** to first get a rv for y of $Y \sim p_Y(y)$.
3. Then with this fixed y use **??** again to get a value for x of $X \sim p_{X|Y}(x|y)$.

Definition 14.50
Multiple Continuous Variables (Option 2):
Note: this only works if $\mathcal{X}$ and $\mathcal{Y}$ are finite.
Given: a pdf of multiple rvs $p_{X,Y}$ **let** $N_x = |\mathcal{X}|$ and $N_y = |\mathcal{Y}|$ the number of elements in $\mathcal{X}$ and $\mathcal{Y}$.

Define $\quad p_Z(1) = p_{X,Y}(1,1), p_Z(2) = p_{X,Y}(1,2), \dots$
$\quad \dots, p_Z(N_x \cdot N_y) = p_{X,Y}(N_x, N_y)$

Then simply apply **??** to the auxillary pdf p_Z
1. Use the product rule (eq. (14.21)) in order to decompose $f_{X,Y}$:
$$f_{X,Y} = f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) \qquad (14.118)$$
2. Use [def. 14.48] to first get a rv for y of $Y \sim f_Y(y)$.
3. Then with this fixed y use [def. 14.48] again to get a value for x of $X \sim f_{X|Y}(x|y)$.

nice examples see comment in code text

9. Descriptive Statistics

9.1. Population Parameters

Definition 14.51 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 14.52 Population Mean: Given a population $\{x_i\}_{i=1}^N$ of size N its variance is defined as:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{14.119}$$

Definition 14.53 Population Variance: Given a population $\{x_i\}_{i=1}^N$ of size N its variance is defined as: $\{x_i\}_{i=1}^N$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2 \tag{14.120}$$

Note

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

9.2. Sample Estimates

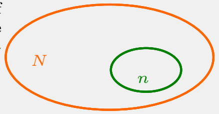
Definition 14.54 (Sample) Statistic: A statistc is a measuarble function f that assigns a **single** value F to a sample of random variables or population:

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

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

$$\hat{\theta} = h(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \tag{14.121}$$

is a statistic/randomn variable used to estimate a true (population) parameter θ[def. 14.51].

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

9.2.1. Empirical Mean

Definition 14.57 Sample/Empirical Mean $\bar{x}$: The sample mean is an estimate/statistic of the population mean[def. 14.52] and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$\bar{x} = \hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{14.122}$$

Corollary 14.6 Expectation: The sample mean estimator is unbiased (see section 13):

$$\mathbb{E}\left[\hat{\mu}_X\right] = \mu \tag{14.123}$$

Corollary 14.7 Variance: For the variance of the sample mean estimator it holds (see section 13):

$$\mathbb{V}\left[\hat{\mu}_X\right] = \frac{1}{n}\sigma_X^2 \tag{14.124}$$

9.2.2. Empirical Variance

Definition 14.58 Biased Sample Variance: The sample mean is an estimate/statistic of the population variance[def. 14.53] and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$s_n^2 = \hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2 \tag{14.125}$$

Definition 14.59 (Unbiased) Sample Variance:

$$s^2 = \hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2 \tag{14.126}$$

see section 13

Definition 14.60 Bessel's Correction: The factor

$$\frac{n}{n-1} \tag{14.127}$$

as multiplying the uncorrected population varianceeq. (14.125) 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
□

10. Statistical Tests

Definition 14.61 Null Hypothesis: A Null Hypothesis H_0 is usually a commonly accepted fact/view/base hypothesis that researchers try to nullify or disprove.

$$H_0 : \theta = \theta_0 \tag{14.128}$$

Definition 14.62 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.

$$H_A : \theta \begin{cases} > \theta_0 & \text{(one-sided)} \\ < \theta_0 & \text{(one-sided)} \\ \neq \theta_0 & \text{(two-sided)} \end{cases} \tag{14.129}$$

Definition 14.63 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.

$$\Theta = \Theta_0 \cup \Theta_1 \qquad \Theta_0 \cap \Theta_1 = \emptyset$$

$$\text{Null Hypothesis} \qquad H_0 : \theta \in \Theta_0 \tag{14.130}$$

$$\text{Alternative Hypothesis} \qquad H_A : \theta \in \Theta_1 \tag{14.131}$$

10.1. Type I&II Errors

Definition 14.64 Type I Error: Is the rejection of a Null Hypothesis, even-tough its true (also known as a "false positive").

Definition 14.65 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 14.66 Critical Value c: Value from which on the Null-hypothesis H_0 gets rejected.

Definition 14.67 Statistical significance α: A study's defined significance level, denoted by α, is the **probability** of the study rejecting the null hypothesis, given that the null hypothesis were true (Type I Error).

Definition 14.68 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.

$$\boxed{K_\alpha \in \Theta : \mathbb{P}(\text{Type I Error}) \leqslant \alpha} \tag{14.132}$$

$$\begin{aligned} & \mathbb{P}(c_2 \leqslant X \leqslant c_1) \leqslant \alpha && \text{two-sided} \\ \text{or} \quad & \mathbb{P}(c_2 \leqslant X) \leqslant \frac{\alpha}{2} \quad \text{and} \quad \mathbb{P}(X \leqslant c_1) \leqslant \frac{\alpha}{2} \\ & \mathbb{P}(c_2 \leqslant X) \leqslant \alpha && \text{one-sided} \\ & \mathbb{P}(X \leqslant c_1) \leqslant \alpha && \text{one-sided} \end{aligned}$$

Definition 14.69 Acceptance Region: Is the region where we accept the null hypothesis H_0.

Note

see example 14.3.

10.2. Normally Distributed Data

Let us consider a sample of $\{x_i\}_{i=1}^n$ i.i.d. observations, that follow a normal distribution $x_i \sim \mathcal{N}(\mu, \sigma^2)$.

10.2.1. Z-Test — σ known

10.2.2. t-Test — σ unknown

11. 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?

12. Examples

Example 14.1 Theorem 14.4: Let x be uniformly distributed on $[0,1]$ ([def. 14.37]) with pmf $p_X(x)$ then it follows:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{p_Y(y)} \Rightarrow \mathrm{d}x = \mathrm{d}y\, p_y(y) \Rightarrow x = \int_{-\infty}^{y} p_y(t)\,\mathrm{d}t = \mathbb{F}_Y(x)$$

Example 14.2 Theorem 14.4: Let

add https://www.youtube.com/watch?v=WUUb7VIRzgg

Example 14.3 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. 14.34]

$$X \sim \mathcal{B}(n, p), n = 20 \qquad \mathbb{P}(X \geqslant x) = \sum_{k=x}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

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 $\mathbb{P}(X \geqslant x) = 1 - \mathbb{P}(X \leqslant x-1)$

13. Proofs

Proof. corollary 14.6:

$$\mathbb{E}[\hat{\mu}_X] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\mathbb{E}[\underbrace{\mu + \cdots + \mu}_{1,\ldots,n}]$$

□

Proof. corollary 14.7:

$$\mathbb{V}[\hat{\mu}_X] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \overset{\text{Property 14.12}}{=} \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} x_i\right]$$
$$\frac{1}{n^2} n \mathbb{V}[X] = \frac{1}{n}\sigma^2$$

□

Proof. definition 14.59:

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}_X^2\right] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\cdot n\bar{x} + n\bar{x}^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\mathbb{E}\left[x_i^2\right] - n\mathbb{E}\left[\bar{x}^2\right]\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\mathbb{E}\left[\bar{x}^2\right]\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{1}{n}\sigma^2 + \mu^2\right)\right] \\
&= \frac{1}{n-1}\left[(n\sigma^2 + n\mu^2) - \left(\sigma^2 + n\mu^2\right)\right] \\
&= \frac{1}{n-1}\left[n\sigma^2 - \sigma^2\right] = \frac{1}{n-1}\left[(n-1)\sigma^2\right] = \sigma^2
\end{aligned}$$

□

Stochastic Calculus

Stochastic Processes

Definition 15.1
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:
$$X : \mathbb{R} \times \Omega \mapsto \mathbb{R} \qquad \iff \qquad (t, \omega) \mapsto X(t, \omega) \tag{15.1}$$

Definition 15.2 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 15.3 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*:
$$\mathcal{F}_s \subseteq \mathcal{F}_t \qquad \forall s \leqslant t \tag{15.2}$$

Definition 15.4 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 15.5 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 15.6
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 15.1 : The amount of information of an adapted random process is increasing see example 15.1.

Definition 15.7 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:
$$\forall s \leqslant t \qquad \mathbb{E}[X(t)|\mathcal{F}_s] = X(s) \quad \text{a.s.} \tag{15.3}$$
2. The expectation is finite:
$$\mathbb{E}[|X(t)|] < \infty \quad \forall t \geqslant 0 \qquad X(t) \text{ is } \{\mathcal{F}_t\}_{t \geqslant 0} \text{ adapted} \tag{15.4}$$

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 15.8 Auto Covariance $\gamma(t_2 - t_1)$:
Describes the covariance[def. 14.28] between two values of a stochastic process $(\boldsymbol{X}_t)_{t \in \mathcal{T}}$ at different time points t_1 and t_2.
$$\gamma(t_1, t_2) = \operatorname{Cov}[\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}] = \mathbb{E}[(\boldsymbol{X}_{t_1} - \mu_{t_1})(\boldsymbol{X}_{t_2} - \mu_{t_2})] \tag{15.5}$$
For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance:
$$\gamma(t, t) = \operatorname{Cov}[\boldsymbol{X}_t, \boldsymbol{X}_t] \overset{\text{eq. } (14.64)}{=} \mathbb{V}[\boldsymbol{X}_t] \tag{15.6}$$

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 15.9 Auto Correlation $\rho(t_2 - t_1)$:
Is the scaled version of the auto-covariance[def. 15.8]:
$$\rho(t_2 - t_1) = \operatorname{Corr}[\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}] \tag{15.7}$$
$$= \frac{\operatorname{Cov}[\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}]}{\sigma_{X_{t_1}} \sigma_{X_{t_2}}} = \frac{\mathbb{E}[(\boldsymbol{X}_{t_1} - \mu_{t_1})(\boldsymbol{X}_{t_2} - \mu_{t_2})]}{\sigma_{X_{t_1}} \sigma_{X_{t_2}}}$$

1. Different kinds of Processes

1.1. Markov Process

Definition 15.10 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:
$$\mathbb{P}(X(t_{n+1}) \in B | X(t_1), \dots, X(t_n)) = \mathbb{P}(X(t_{n+1}) \in B | X(t_n))$$
it thus follows for the *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:
$$\mathbb{P}(s, x, t, B) = P(X(t) \in B | X(s) = x) \qquad 0 \leqslant s < t \tag{15.8}$$

Interpretation

In order to predict the future only the current/last value counts.

Corollary 15.2 Transition Density: The transition probability of a continuous distribution p can be calculated via:
$$\mathbb{P}(s, x, t, B) = \int_B \mathrm{p}(s, x, t, y)\, \mathrm{d}y \tag{15.9}$$

1.2. Gaussian Process

Definition 15.11 Gaussian Process: Is a stochastic process $X(t)$ where the random variables follow a Gaussian distribution:
$$X(t) \sim \mathcal{N}\left(\mu(t), \sigma^2(t)\right) \qquad \forall t \in T \tag{15.10}$$

1.3. Diffusions

Definition 15.12 Diffusion: Is a Markov Process[def. 15.10] for which it holds that:
$$\mu(t, X(t)) = \lim_{t \to 0} \frac{1}{\Delta t} \mathbb{E}[X(t + \Delta t) - X(t) | X(t)] \tag{15.11}$$
$$\sigma^2(t, X(t)) = \lim_{t \to 0} \frac{1}{\Delta t} \mathbb{E}\left[(X(t + \Delta t) - X(t))^2 \,|X(t)\right] \tag{15.12}$$
See ??/eq. (15.12) for simple proof of eq. (15.11)/??.

- $\mu(t, X(t))$ 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 15.13
d**-dim standard Brownian Motion/Wienner Process:**
Is an $\mathbb{R}^d$ valued *stochastic process*[def. 15.1] $(W_t)_{t \in \mathcal{T}}$ starting at $\boldsymbol{x}_0 \in \mathbb{R}^d$ that satisfies:

1. **Normal Independent Increments:** the increments are *normally distributed independent random variables*:
$$W(t_i) - W(t_{i-1}) \sim \mathcal{N}(0, (t_i - t_{i-1})\mathbb{1}_{d \times d}) \qquad \forall i \in \{1, \dots, T\} \tag{15.13}$$
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
$$\lim_{t \to 0} \frac{\mathbb{P}(|W(t + \Delta t) - W(t)| \geqslant \delta)}{\Delta t} = 0 \qquad \forall \delta > 0 \tag{15.14}$$
4. **Start**
$$W(0) := W_0 = 0 \qquad a.s. \tag{15.15}$$

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 15.3 $W_t \sim \mathcal{N}(0, \sigma)$:
The random variable W_t follows the $\mathcal{N}(0, \sigma)$ law
$$\mathbb{E}[W(t)] = \mu = 0 \tag{15.16}$$
$$\mathbb{V}[W(t)] = \mathbb{E}\left[W^2(t)\right] = \sigma^2 = t \tag{15.17}$$
See section 5

1.4.1. Properties of the Wienner Process

Property 15.1 Non-Differentiable Trajectories:
The sample paths of a Brownian motion are not differentiable:
$$\frac{\mathrm{d}W(t)}{t} = \lim_{t \to 0} \mathbb{E}\left[\left(\frac{W(t + \Delta t) - W(t)}{\Delta t}\right)^2\right]$$
$$= \lim_{t \to 0} \frac{\mathbb{E}[W(t + \Delta t) - W(t)]}{\Delta t} = \lim_{t \to 0} \frac{\sigma^2}{\Delta t} = \infty$$
$\xrightarrow{\text{result}}$ cannot use normal calculus anymore
$\xrightarrow{\text{solution}}$ Ito Calculus see section 16.

Property 15.2 Auto covariance Function:
The auto-covariance[def. 15.8] for a Wienner process
$$\mathbb{E}\left[(W(t) - \mu t)(W(t') - \mu t')\right] = \min(t, t') \tag{15.18}$$

Property 15.3 : A standard Brownian motion is a

Quadratic Variation

Definition 15.14 Total Variation: The total variation of a function $f : [a, b] \subset \mathbb{R} \mapsto \mathbb{R}$ is defined as:
$$LV_{[a,b]}(f) = \sup_{\Pi \in \mathcal{S}} \sum_{i=0}^{n_\Pi - 1} |f(x_{i+1}) - f(x_i)| \tag{15.19}$$
$$\mathcal{S} = \left\{\Pi\{x_0, \dots, x_{n_\Pi}\} : \Pi \text{ is a partition }^{\text{[def. 10.1]}} \text{ of } [a, b]\right\}$$
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 15.15
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:
$$QV_{[a,b]}(f) = \sup_{\Pi \in \mathcal{S}} \sum_{i=0}^{n_\Pi - 1} |f(x_{i+1}) - f(x_i)|^2 \tag{15.20}$$
$$\mathcal{S} = \left\{\Pi\{x_0, \dots, x_{n_\Pi}\} : \Pi \text{ is a partition }^{\text{[def. 10.1]}} \text{ of } [a, b]\right\}$$

Corollary 15.4 Bounded (quadratic) Variation:
The (quadratic) variation[def. 15.14] of a function is bounded if it is finite:
$$\exists M \in \mathbb{R}_+ : \quad LV_{[a,b]}(f) \leqslant M \quad \left(QV_{[a,b]}(f) \leqslant M\right) \quad \forall \Pi \in \mathcal{S} \tag{15.21}$$

Theorem 15.1 Variation of Wienner Process: Almost surely the total variation of a Brownian motion over a interval $[0, T]$ is infinite:
$$\mathbb{P}(\omega : LV(W(\omega)) < \infty) = 0 \tag{15.22}$$

Theorem 15.2
Quadratic Variation of standard Brownian Motion:
The quadratic variation of a standard Brownian motion over $[0, T]$ is finite:
$$\lim_{N \to \infty} \sum_{k=1}^{N} \left[W\left(k\frac{T}{N}\right) - W\left((k-1)\frac{T}{N}\right)\right]^2 = T \quad \text{with probability 1} \tag{15.23}$$
See ??

Corollary 15.5 : theorem 15.2 can also be written as:
$$(\mathrm{d}W(t))^2 = \mathrm{d}t \tag{15.24}$$

1.4.2. Lévy's Characterization of BM

Theorem 15.3
d**-dim standard BM/Wienner Process by Paul Lévy:**
An $\mathbb{R}^d$ valued *adapted stochastic process*[def's. 15.1, 15.3] $(W_t)_{t \in \mathcal{T}}$ with the filtration $\{\mathcal{F}_t\}_{t \in \mathbb{R}_+}$, that satisfies:

1. **Start**
$$W(0) := W_0 = 0 \qquad a.s. \tag{15.25}$$
2. **Continuous Martingale:** W_t is an a.s. *continuous* martingale[def. 15.7] w.r.t. the filtration $(\mathcal{F}_t)_{t \in \mathcal{T}}$ under $\mathbb{P}$.
3. **Quadratic Variation:**
$$W_t^2 - t \text{ is also an martingale} \iff QV(W_t) = t \tag{15.26}$$

is a standard Brownian motion[def. 15.18]. Proof see section 5

Further Stochastic Processes

1.4.3. White Noise

undestand script and add

Definition 15.16 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):
$$\mathbb{E}[\epsilon[k]] = 0 \qquad \forall k \in T_{\text{discret}} \tag{15.27}$$
- Zero autocorrelation[def. 15.9] γ i.e. the signals of different times are in no-way correlated:
$$\gamma(\epsilon[k], \epsilon[k + n]) = \mathbb{E}\left[\epsilon[k]\epsilon[k + n]^\intercal\right] = \mathbb{V}[\epsilon[k]]\, \delta_{\text{discret}}[n] \qquad \forall k, n \in T_{\text{discret}} \tag{15.28}$$

With $\qquad \delta_{\text{discret}}[n] := \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{else} \end{cases}$

See proofs

Definition 15.17 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):
$$\mathbb{E}[\epsilon(t)] = 0 \qquad \forall t \in T_{\text{continuous}} \tag{15.29}$$
- Zero autocorrelation[def. 15.9] γ i.e. the signals of different times are in no-way correlated:
$$\gamma(\epsilon(t), \epsilon(t + \tau)) = \mathbb{E}\left[\epsilon(t)\epsilon(t + \tau)^\intercal\right] \tag{15.30}$$
$$\overset{\text{eq. } (14.106)}{=} \mathbb{V}[\epsilon(t)]\, \delta(t - \tau) = \begin{cases} \mathbb{V}[\epsilon(t)] & \text{if } \tau = 0 \\ 0 & \text{else} \end{cases} \qquad \forall t, \tau \in T_{\text{continuous}} \tag{15.31}$$

1.4.4. Generalized Brownian Motion

Definition 15.18 Brownian Motion: Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 15.13], and define:

$$X_t = \mu t + \sigma W_t \qquad t \in \mathbb{R}_+ \qquad \begin{array}{ll} \mu \in \mathbb{R} & : \text{drift parameter} \\ \sigma \in \mathbb{R}_+ & : \text{scale parameter} \end{array} \tag{15.32}$$

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 15.4 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:

$$f_{W(t)}(x) \sim \mathcal{N}(\mu t, \sigma^2 t) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left\{-\frac{(x-\mu t)^2}{2\sigma^2 t}\right\} \tag{15.33}$$

Corollary 15.6 : More generally we may define the process:

$$t \mapsto f(t) + \sigma W_t \tag{15.34}$$

which corresponds to a noisy version of f.

Corollary 15.7
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:

$$\mathrm{d}X(t) = \mu\,\mathrm{d}t + \sigma\,\mathrm{d}W(t) \tag{15.35}$$
$$X(0) = 0 \tag{15.36}$$

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 15.19 Geometric Brownian Motion: Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 15.13] the exponential transform:

$$X(t) = \exp\left(W(t)\right) = \exp\left(\mu t + \sigma W(t)\right) \qquad t \in \mathbb{R}_+ \tag{15.37}$$

is called geometric Brownian motion

Corollary 15.8 Log-normal Returns: For a geometric BM we obtain log-normal returns:

$$\ln\left(\frac{S_t}{S_0}\right) = \mu t + \sigma W(t) \iff \mu t + \sigma W(t) \sim \mathcal{N}(\mu t, \sigma^2 t) \tag{15.38}$$

meaning that the mean and the variance of the process (stock) *log-returns* grow over time linearly.

Corollary 15.9
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:

$$\begin{aligned} \mathrm{d}X(t) &= X(t)\left(\mu\,\mathrm{d}t + \sigma\,\mathrm{d}W(t)\right) \\ &= \mu X(t)\,\mathrm{d}t + \sigma X(t)\,\mathrm{d}W(t) \end{aligned} \tag{15.39}$$
$$X(0) = 0 \tag{15.40}$$

1.4.6. Locally Brownian Motion

Definition 15.20 Locally Brownian Motion: Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 15.13] a local Brownian motion is a stochastic process $X(t)$ that satisfies the SDE:

$$\mathrm{d}X(t) = \mu\left(X(t), t\right)\mathrm{d}t + \sigma\left(X(t), t\right)\mathrm{d}W(t) \tag{15.41}$$

Note

A local Brownian motion is an generalization of a geometric Brownian motion.

1.4.7. Ornstein-Uhlenbeck Process

Definition 15.21 Ornstein-Uhlenbeck Process: Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 15.13] a Ornstein-Uhlenbeck Process or exponentially correlated noise is a stochastic process $X(t)$ that satisfies the SDE:

$$\mathrm{d}X(t) = -aX(t)\,\mathrm{d}t + b\sigma\,\mathrm{d}W(t) \qquad a > 0 \tag{15.42}$$

1.5. Poisson Processes

Definition 15.22 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 15.23 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. 14.36]:
$$\mathbb{P}\left((X_t - X_s) = k\right) = \frac{\lambda(t-s)}{k!}\mathrm{e}^{-\lambda(t-s)} \qquad \begin{array}{l} 0 \leqslant s < t < \infty \\ \forall k \in \mathbb{N} \end{array}$$
3. No correlation of (non-overlapping) increments: $\forall t_0 < t_1 < \cdots < t_n$: the increments are independent
$$X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \ldots, X_{t_n} - X_{t_{n-1}} \tag{15.43}$$

Interpretation

A Poisson Process is a *continuous-time* process with *discrete, positive* realizations in $\in \mathbb{N}_{\geqslant 0}$

Corollary 15.10 Probability of events: Using Taylor in order to expand the Poisson distribution one obtains:

$$\mathbb{P}\left(X_{(t+\Delta t)} - X_t \neq 0\right) = \lambda \Delta t + o(\Delta t^2) \qquad t \text{ small i.e. } t \to 0 \tag{15.44}$$

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 15.24 Differential of a Poisson Process: The differential of a Poisson Process is defined as:

$$\mathrm{d}X_t = \lim_{\Delta t \to \mathrm{d}t}\left(X_{(t+\Delta t)} - X_t\right) \tag{15.45}$$

Property 15.4 Probability of Events for differential: With the definition of the differential and using the previous results from the Taylor expansion it follows:

$$\mathbb{P}(\mathrm{d}X_t = 0) = 1 - \lambda \tag{15.46}$$
$$\mathbb{P}(|\mathrm{d}X_t| = 1) = \lambda \tag{15.47}$$

Proofs

Proof. eq. (15.11):
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:

$$\begin{aligned} \mathbb{E}[x(n)] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N x_i(n)\right] = \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[x_i(n-1) \pm \delta\right] \\ &= \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[x_i(n-1)\right] \\ &\overset{\text{induction}}{=} \mathbb{E}\left[x_{n-1}\right] = \ldots \mathbb{E}[x(0)] = 0 \end{aligned}$$

Thus in expectation the particles goes nowhere. □

Proof. eq. (15.12):
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:

$$\sigma^2 = \mathbb{E}\left[x^2(n)\right] - \mathbb{E}\left[x(n)\right]^2 = \mathbb{E}\left[x^2(n)\right] = \frac{\delta^2}{\Delta x}t \tag{15.48}$$

Thus in expectation the particles goes nowhere. □

Proof. eq. (15.30):

$$\begin{aligned} \gamma(\epsilon[k], \epsilon[k+n]) &= \mathrm{Cov}\left[\epsilon[k], \epsilon[k+1]\right] \\ &= \mathbb{E}\left[\left(\epsilon[k] - \mathbb{E}\left[\epsilon[k]\right]\right)\left(\epsilon[k+n] - \mathbb{E}\left[\epsilon[k+n]\right]\right)^{\intercal}\right] \\ &\overset{\text{eq. (15.27)}}{=} \mathbb{E}\left[\left(\epsilon[k]\right)\left(\epsilon[k+n]\right)\right] \end{aligned}$$
□

Proof. corollary 15.3:
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. (15.15) it follows:

$$B_t - B_s \sim B_{t-s} \qquad B_{t-s} \sim \mathcal{N}(0, t-s)$$
□

Proof. corollary 15.3:

$$W(t) = W(t) - \underbrace{W(0)}_{=0} \sim \mathcal{N}(0, t)$$

$$\Rightarrow \qquad \mathbb{E}[X] = 0 \qquad \mathbb{V}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = t$$
□

Proof. theorem 15.2:

$$\begin{aligned} &\sum_{k=0}^{N-1}\left[W(t_k) - W(t_{k-1})\right]^2 && t_k = k\frac{T}{N} \\ &= \sum_{k=0}^{N-1} X_k^2 && X_k \sim \mathcal{N}\left(0, \frac{T}{N}\right) \\ &= \sum_{k=0}^{N-1} Y_k = n\left(\frac{1}{n}\sum_{k=0}^{N-1} Y_k\right) && \mathbb{E}[Y_k] = \frac{T}{N} \\ &\overset{\text{S.L.L.N}}{=} n\frac{T}{n} = T \end{aligned}$$
□

Proof. theorem 15.3 2:
1. first we need to show eq. (15.3): $\mathbb{E}[W_t|\mathcal{F}_s] = W_s$
Due to the fact that W_t is $\mathcal{F}_t$ measurable i.e. $W_t \in \mathcal{F}_t$ we know that:
$$\mathbb{E}\left[W_t|\mathcal{F}_t\right] = W_t \tag{15.49}$$
$$\begin{aligned} \mathbb{E}\left[W_t|\mathcal{F}_s\right] &= \mathbb{E}\left[W_t - W_s + W_s|\mathcal{F}\right] \\ &= \mathbb{E}\left[W_t - W_s|\mathcal{F}_s\right] + \mathbb{E}\left[W_s|\mathcal{F}_s\right] \\ &\overset{\text{eq. (15.49)}}{=} \mathbb{E}\left[W_t - W_s\right] + W_s \\ &\overset{W_t - W_s \sim \mathcal{N}(0, t-s)}{=} W_s \end{aligned}$$
2. second we need to show eq. (15.4): $\mathbb{E}\left[|X(t)|\right] < \infty$
$$\mathbb{E}\left[|W(t)|\right]^2 \overset{\text{eq. (14.56)}}{\leqslant} \mathbb{E}\left[|W(t)|^2\right] = \mathbb{E}\left[W^2(t)\right] = t \leqslant \infty$$
□

Proof. theorem 15.3 3: $W_t^2 - t$ is a martingale?
Using the binomial formula we can write and adding $W_s - W_s$:

$$W_t^2 = (W_t - W_s)^2 + 2W_s(W_t - W_s) + W_s^2$$

using the expectation:

$$\begin{aligned} \mathbb{E}\left[W_t^2|\mathcal{F}_s\right] &= \mathbb{E}\left[(W_t - W_s)^2|\mathcal{F}_s\right] + \mathbb{E}\left[2W_s(W_t - W_s)|\mathcal{F}_s\right] \\ &\quad + \mathbb{E}\left[W_s^2|\mathcal{F}_s\right] \\ &\overset{\text{eq. (15.49)}}{=} \mathbb{E}\left[(W_t - W_s)^2\right] + 2W_s\mathbb{E}\left[(W_t - W_s)\right] + W_s^2 \\ &\overset{\text{eq. (15.17)}}{=} \mathbb{V}\left[W_t - W_s\right] + 0 + W_s^2 \\ &\quad t - s + W_s^2 \end{aligned}$$

from this it follows that:

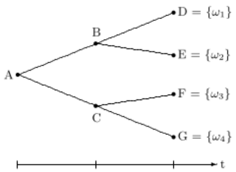
$$\mathbb{E}\left[W_t^2 - t|\mathcal{F}_s\right] = W_s^2 - s$$
□

understand why $\mathbb{E}\left[(W_t - W_s)^2|\mathcal{F}\right] = \mathbb{E}\left[(W_t - W_s)^2\right]$

Examples

Example 15.1 :

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.



$$\mathcal{F} = \begin{cases} \{\varnothing, \Omega\} & t \in [0, T/2) \\ \{\varnothing, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \Omega\} & t \in [T/2, T) \\ \mathcal{F}_{\max} = 2^\Omega & t = T \end{cases} \tag{15.50}$$

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