Probability Cookbook

Pantelis Sopasakis

January 24, 2019

Contents

Abstract

This document is intended to serve as a collection of important results in general probability theory, stochastic processes, uncertainty quantification, risk measures and a lot more. It can be used for a quick brush up or as a quick reference or cheat sheet for graduate students and researchers in the domain of mathematics and engineering. The readers may find a list of bibliographic references with comments at the end of this document. This is still work in progress, so several important results are still missing.

This document, as well as its future versions, will be available at `https://mathematix.wordpress.com/probability-cookbook`.

1 Probability Theory

1.1 General Probability Theory

1.1.1 Measurable and Probability spaces

1. (σ-algebra). Let X be a nonempty set. A collection $\mathcal{F}$ of subsets of X is called a σ-algebra if (i) $X \in \mathcal{F}$, (i) $A^c \in \mathcal{F}$ whenever $A \in \mathcal{F}$, (ii) if $A_1, \ldots, A_n \in \mathcal{F}$, then $\bigcup_{i=1,\ldots,n} A_i \in \mathcal{F}$. The space X equipped with a σ-algebra $\mathcal{F}$ is called a *measurable space*.

2. (d-system) A collection $\mathcal{D}$ of subsets of X is called a d-system or a Dynkin class if (i) $X \in \mathcal{D}$, (ii) $A \setminus B \in \mathcal{D}$ whenever $A, B \in \mathcal{D}$ and $A \supseteq B$, (iii) $A \in \mathcal{D}$ whenever $A_n \in \mathcal{D}$ and $A_n \uparrow A$ (meaning, $A_k \subseteq A_{k+1}$ and $\bigcup_{k \in \mathbb{N}} A_k = A$).

3. (p-system). A collection of sets $\mathcal{P}$ in X is called a p-system if $A \cap B \in \mathcal{P}$ whenever $A, B \in \mathcal{P}$.

4. A collection of sets is a σ-algebra if and only if it is both a p- and a d-system.

5. (Smallest σ-algebra). Let $\mathcal{H}$ be a collection of sets in X. The smallest collection of sets which contains $\mathcal{H}$ and is a σ-algebra exists and is denoted by $\sigma(\mathcal{H})$.

6. (Monotone class theorem). If a d-system $\mathcal{D}$ contains a p-system $\mathcal{P}$, then is also contains $\sigma(\mathcal{P})$.

7. (Borel σ-algebra). On $\mathbb{R}$, the σ-algebra $\sigma(\{(a,b); a < b\})$ is called the Borel σ-algebra on $\mathbb{R}$ which we denote by $\mathcal{B}_{\mathbb{R}}$. For topological spaces (X, τ), the Borel σ-algebra is defined as $\mathcal{B}_X = \sigma(\tau)$, i.e., it is the smallest σ-algebra which contains all open sets. $\mathcal{B}_{\mathbb{R}}$ is generated by:
   i. The open intervals (a, b)
   ii. The closed intervals $[a, b]$
   iii. All sets of the form $[a, b)$ or $(a, b]$
   iv. Open rays (a, ∞) or $(-\infty, a)$
   v. Closed rays $[a, \infty)$ or $(-\infty, a]$

8. (Measure). A function $\mu : \mathcal{F} \to [0, +\infty]$ is called a measure if for every sequence of disjoint sets A_n from $\mathcal{F}$, $\mu(\bigcup_n A_n) = \sum_n \mu(A_n)$.

9. (Properties of measures). The following hold:
   i. (Empty set is negligible). $\mu(\varnothing) = 0$ [Indeed, $\mu(A) = \mu(A \cup \varnothing) = \mu(A) + \mu(\varnothing)$ for all $A \in \mathcal{F}$]
   ii. (Monotonicity). $A \subseteq B$ implies $\mu(A) \leq \mu(B)$ [Indeed, $\mu(B) = \mu(A \cup (B \setminus A))$]
   iii. (Boole's inequality). For all $A_n \in \mathcal{F}$, $\mu(\bigcup_n A_n) \leq \sum_n \mu(A_n)$
   iv. (Sequential continuity). If $A_n \uparrow A$, then $\mu(A_n) \uparrow \mu(A)$.

10. (Equality of measures). Let μ, ν be two measures on a measurable space $(X, \mathcal{F})$ and let $\mathcal{G}$ be a p-system generating $\mathcal{F}$. If $\mu(A) = \nu(A)$ for all $A \in \mathcal{G}$, then $\mu(B) = \nu(B)$ for all $B \in \mathcal{F}$. As presented in #7 above, p-systems are often available and have simple forms.

11. (Completeness). A measure space $(X, \mathcal{F}, \mu)$ is called *complete* if the following holds:

$$A \in \mathcal{F}, \mu(A) = 0, B \subseteq A \Rightarrow B \in \mathcal{F}.$$

    Of course, by the monotonicity property in #9–iii, if $(X, \mathcal{F}, \mu)$ is a complete measure space then $\mu(B) = 0$.

12. (Completion). Let $(X, \mathcal{F}, \mu)$ be a measure space and define the set of *negligible sets* of μ as $Z_\mu = \{N \subseteq X : \exists N' \supseteq N, N' \in \mathcal{F} \text{ s.t. } \mu(N') = 0\}$. Let $\mathcal{F}'$ be the σ-algebra generated by $\mathcal{F} \cup Z_\mu$. Then

i. Every $B \in \mathcal{F}'$ can be written as $B = A \cup N$ with $A \in \mathcal{F}$ and $N \in Z_\mu$

ii. Define $\mu'(A \cup N) = \mu(A)$; this is a measure on $(X, \mathcal{F}')$ which renders the space $(X, \mathcal{F}', \mu')$ complete.

13. (Lebesgue measure on $\mathbb{R}$ and $\mathbb{R}^n$). It suffices to define the *Lebesgue measure* on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$ on the p-system $\{(a,b), a < b\}$; it is $\lambda((a,b)) = b - a$. This extends to a measure on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$. Likewise, the collection of n-dimensional rectangles $\{(a_1, b_1) \times \ldots \times (a_n, b_n)\}$ is a p-system which generates $\mathcal{B}_{\mathbb{R}^n}$; the Lebesgue measure on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ is $\lambda(\prod_{i=1}^n (a_i, b_i)) = \prod_{i=1}^n (b_i - a_i)$.

14. (Lebesgue measurable sets). The completion of the Lebesgue measure defines the class of Lebesgue-measurable sets.

15. (Negligible boundary). If a set $C \subseteq \mathbb{R}^n$ has a boundary whose Lebesgue measure is 0, then C is Lebesgue measurable.

16. (Independent events). Let E_1, E_2 be two events from $(\Omega, \mathcal{F}, \mathrm{P})$; we say that E_1 and E_2 are *independent* if $\mathrm{P}[E_1 \cap E_2] = \mathrm{P}[E_1]\mathrm{P}[E_2]$.

17. (Independent σ-algebras). We say that two σ-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ on Ω are independent if for any $E_1 \in \mathcal{F}_1$ and $E_2 \in \mathcal{F}_2$, E_1 and E_2 are independent. Note that $E_1 \cap E_2$ is a member of the σ algebra $E_1 \wedge E_2$.

18. (Atom). Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. A set $A \in \mathcal{F}$ is called an atom if $\mu(A) > 0$ and for every $B \subset A$ with $\mu(B) < \mu(A)$ it is $\mu(B) = 0$. A space without atoms is called non-atomic[1].

1.1.2 Random variables

1. (Measurable function). A function $f : (X, \mathcal{F}) \to (Y, \mathcal{G})$ (between two measurable spaces) is called *measurable* if $f^{-1}(G) \in \mathcal{F}$ for all $G \in \mathcal{G}$ (i.e., if it inverts all measurable sets to measurable ones).

2. (Measurability test). Let $\mathcal{F}, \mathcal{G}$ be σ-algebras on the nonempty sets X and Y. Let $\mathcal{G}'$ be a p-system which generates $\mathcal{G}$. A function $f : (X, \mathcal{F}) \to (Y, \mathcal{G})$ is measurable if and only if $f^{-1}(G') \in \mathcal{F}$ for all $G' \in \mathcal{G}'$ (it suffices to check the measurability condition on a p-system).

3. (σ-algebra generated by f). Let $f : (X, \mathcal{F}) \to (Y, \mathcal{G})$ (between two measurable spaces) be a measurable function. The set
$$\sigma(f) := \{f^{-1}(B) \mid B \in \mathcal{G}\},$$
is a sub-σ-algebra of $\mathcal{F}$ and is called the σ-algebra generated by f.

4. (Preservation of measurability). Let $f, g : \Omega \to \mathbb{R}$ be two measurable functions on $(\Omega, \mathcal{F})$. Then, the functions $h_1(x) = f(x) + g(x)$, $h_2(x) = f(x) - g(x)$, $h_3(x) = \max\{f(x), g(x)\}$, $h_4(x) = \min(f(x), g(x))$, $h_5(x) = f(x)g(x)$ are measurable. For all $\alpha \in \mathbb{R}$, $h_6(x) = \alpha f(x)$ is measurable.

5. (Measurability of supermum/infimum). Let $(f_n)_n$ be a sequence of real-valued measurable functions. Then $\sup_n f_n$ and $\inf_n f_n$ are measurable.

6. (Sub/sup-level sets) Let $f : (X, \mathcal{F}) \to \mathbb{R}$. The following are equivalent:

   i. f is measurable,

   ii. Its *sub-level sets*, that is sets of the form $\operatorname{lev}_{\leq \alpha} f := \{x \in X : f(x) \leq \alpha\}$ are measurable,

   iii. Its *sup-level sets*, that is sets of the form $\operatorname{lev}_{\geq \alpha} f := \{x \in X : f(x) \geq \alpha\}$ are measurable.

7. (Random variable). A real-valued random variable $X : (\Omega, \mathcal{F}, \mathrm{P}) \to (\mathbb{R}, \mathcal{B}_\mathbb{R})$ is a measurable function X from a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ to $\mathbb{R}$, equipped with the Borel σ-algebra, that is, for every Borel set B, $X^{-1}(B) \in \mathcal{F}$.

8. Every nonnegative (real-valued) random variable X on $(\mathbb{R}_+, \mathcal{B}_{\mathbb{R}_+})$ is written as
$$X(\omega) = \int_0^{+\infty} 1_{X(\omega) \geq t} \,\mathrm{d}t.$$

[1] A special class of spaces with (only) atoms are the discrete probability spaces where $\mathcal{F}$ is generated by a discrete — often finite — set of events. Several results in measure theory require that the space be non-atomic. However, we may often prove these results for discrete or finite spaces.

9. (Increasing functions). Every increasing function $f : \mathbb{R} \to \overline{\mathbb{R}}$ is Borel-measurable.

10. (Semi-continuous functions). Every lower semi-continuous function $X : \Omega \to \mathbb{R}$ (where Ω is assumed to be equipped with a topology) is Borel-measurable.

11. (Push-forward measure) [3]. Given measurable spaces $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$, a measurable mapping $f : X \to Y$ and a (probability) measure μ on $(\mathcal{X}, \mathcal{F})$, the *push-forward* of μ is defined to be a measure $f(\mu)$ on $(\mathcal{Y}, \mathcal{G})$ given by
$$(f_* \mu)(B) = \mu(f^{-1}(B)) = \mu(\{\omega \mid f(\omega) \in B\}),$$
for $B \in \mathcal{G}$.

12. (Change of variables). Let F be a random variable on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and $F_*\mathrm{P}$ is the push-forward measure. random variable X is integrable with respect to the push-forward measure $F_*\mathrm{P}$ if and only if $X \circ F$ is P-integrable. Then, the integrals coincide
$$\int X \mathrm{d}(F_*\mathrm{P}) = \int (X \circ F) \mathrm{dP}.$$

13. (Measures from random variables). Let X be a random variable on $(\Omega, \mathcal{F}, \mathrm{P})$. We may use X to define the following measure
$$\nu(A) = \int_A X \mathrm{dP},$$
defined for $A \in \mathcal{F}$. This is a positive measure which for short we denote as $\nu = X\mathrm{P}$ and it satisfies:
$$\int_A Y \mathrm{d}\nu = \int_A XY \mathrm{dP},$$
for all random variables Y.

14. (Compositions). Let $f : (X, \mathcal{F}_X) \to (Y, \mathcal{F}_Y)$ and $g : (Y, \mathcal{F}_Y) \to (Z, \mathcal{F}_Z)$ be two measurable functions. Then, the function $h : (X, \mathcal{F}_X) \ni x \mapsto h(x) := f(g(x)) \in (Z, \mathcal{F}_Z)$ is measurable.

15. (Simple function; definition). A simple function is one of the form
$$f(x) = \sum_{k=1}^{n} \alpha_k 1_{A_k}(x),$$
where 1_{A_k} is the characteristic function of a measurable set A_k, that is
$$1_{A_k} = \begin{cases} 1, & \text{if } x \in A_k \\ 0, & \text{otherwise} \end{cases}$$

16. (Characterization of measurability). A function $f : (X, \mathcal{F}) \to \mathbb{R}$ is $\mathcal{F}$-measurable if and only if it is the point-wise limit of a sequence of simple functions. A function $f : (X, \mathcal{F}) \to \mathbb{R}_+$ is $\mathcal{F}$-measurable if and only if it is the point-wise limit of an increasing sequence of simple functions.

17. (Continuity and measurability). Every continuous function $f : (X, \mathcal{F}) \to \overline{\mathbb{R}}$ is Borel-measurable.

18. (Monotone class of functions). Let M be a collection of functions $f : (X, \mathcal{F}) \to \overline{\mathbb{R}}$; let M_+ be all positive functions in M and M_b all bounded functions in M. We say that M is a *monotone class* of functions if (i) $1 \in M$, (ii) if $f, g \in M_b$ and $a, b \in \mathbb{R}$, then $af + bg \in M$ and (iii) if $(f_n)_n \subseteq M_+$ and $f_n \uparrow f$, then $f \in M$.

19. (Monotone class theorem for functions). Let M be a monotone class of functions on $(X, \mathcal{F})$. Suppose that $\mathcal{F}$ is generated by some p-system $\mathcal{C}$, $1_A \in M$ for all $A \in \mathcal{C}$. Then, M includes all positive $\mathcal{F}$-measurable functions and all bounded $\mathcal{F}$-measurable functions.

20. (Simple function approximation theorem). Let X be an extended-real-valued Lebesgue-measurable function defined on a Lebesgue measurable set E. Then there exists a sequence $\{\phi_k\}_{k \in \mathbb{N}}$ of simple functions[2] on E such that

[2] A simple function is a finite linear combination of indicator functions of measurable sets, that is, simple functions are written as $\phi(x) = \sum_{i=1}^{n} \alpha_i 1_{A_i}(x)$.

i. $\phi_k \to X$, point-wise on E

ii. $|\phi_k| \leq |X|$ on E for all $k \in \mathbb{N}$

If $X \geq 0$ then there exists a sequence of point-wise increasing simple functions with these properties.

21. (Simple function approximation trick). Let f be a real-valued measurable function, $f : (\Omega, \mathcal{F}, \mathrm{P}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. Define

$$\phi_k(x) = \begin{cases} \frac{j-1}{2^k}, & \frac{j-1}{2^k} \leq f(x) < \frac{j}{2^k} \\ k, & f(x) \geq k \end{cases}$$

Then,

i. The sets $\{x : f(x) \geq k\}$ and $\{x : \frac{j-1}{2^k} \leq f(x) < \frac{j}{2^k}\}$ are measurable because f is measurable

ii. ϕ_k are measurable for all $k \in \mathbb{N}$

iii. $\phi_k(x) \leq \phi_{k+1}(x)$ for all $k \in \mathbb{N}$ and for all $x \in \Omega$

iv. Let $E \subseteq \Omega$ so that $\sup_{x \in E} f(x) \leq M$. Then $\sup_{x \in \Omega} |f(x) - \phi_k(x)| \leq {}^1\!/_{2^k}$ for all $k \geq M$

1.1.3 Limits

Limits of sequences of events

1. (Nested sequences and probabilities). Let $(E_n)_n$ be a non-increasing sequence of events ($E_n \supseteq E_{n+1}$ for all $n \in \mathbb{N}$). Then $\lim_n \mathrm{P}[E_n]$ exists and

$$\mathrm{P}\left[\bigcap_n E_n\right] = \lim_n \mathrm{P}[E_n].$$

If $(E_n)_n$ is a nondecreasing sequence ($E_n \subseteq E_{n+1}$ for all $n \in \mathbb{N}$), then

$$\mathrm{P}\left[\bigcup_n E_n\right] = \lim_n \mathrm{P}[E_n].$$

2. (Limits inferior). For a sequence of events E_n, the *limit inferior* of $(E_n)_n$ is denoted by $\liminf_n E_n$ and is defined as

$$\liminf_n E_n = \bigcup_{n \in \mathbb{N}} \bigcap_{m \geq n} E_n = \{x : \ x \in E_n \text{ for all but finitely many } n \in \mathbb{N}\}.$$

3. (Limit superior). The *limit superior* of $(E_n)_n$, $\limsup_n E_n$, is

$$\limsup_n E_n = \bigcap_{n \in \mathbb{N}} \bigcup_{m \geq n} E_n = \{x : \ x \in E_n \text{ infinitely often}\}.$$

4. (Limits of complements). The limit (super/inferior) of a sequence of complements is the complement of the limit

$$\liminf_n E_n^c = (\limsup_n E_n)^c,$$
$$\limsup_n E_n^c = (\liminf_n E_n)^c.$$

5. (Relationship between limits). It is

$$\liminf_n E_n \subseteq \limsup_n E_n.$$

6. (Probabilities of $\liminf E_n$ and $\limsup E_n$). The sets $\liminf_n E_n$ and $\limsup_n E_n$ are measurable and

$$\mathrm{P}[\liminf_n E_n] \leq \liminf_n \mathrm{P}[E_n] \leq \limsup_n \mathrm{P}[E_n] \leq \mathrm{P}[\limsup_n E_n].$$

7. (A result reminiscent of Baire's category theorem). Let $(E_n)_n$ be a sequence of almost sure events. Then $\mathrm{P}[\cap_n E_n] = 1$.

8. (Borel-Cantelli lemma). Let $(E_n)_n$ be a sequence of events over $(\Omega, \mathcal{F}, \mathrm{P})$. The following hold
   i. If $\sum_{n=1}^{\infty} \mathrm{P}[E_n] < \infty$, then $\mathrm{P}[\limsup_n E_n] = 0$
   ii. If $(E_n)_n$ are independent events such that $\sum_{n=1}^{\infty} \mathrm{P}[E_n] = \infty$, then $\mathrm{P}[\limsup_n E_n] = 1$.

9. (Corollary: Borel 0-1 law). If $(E_n)_n$ is a sequence of independent events, then $\mathrm{P}[\limsup_n E_n] \in \{0, 1\}$ (according to the summability of $(\mathrm{P}[E_n])_n$).

10. (Kochen-Stoone lemma). Let $(E_n)_n$ be a sequence of events. Then,

$$\mathrm{P}[\limsup_n E_n] \geq \limsup_n \frac{\left(\sum_{k=1}^n \mathrm{P}[A_k]\right)^2}{\sum_{k=1}^n \sum_{j=1}^n \mathrm{P}[A_k \cap A_j]}$$

11. (Corollary of Kochen-Stoone's lemma). If for $i \neq j$, E_i and E_j are either independent or $\mathrm{P}[E_i \cap E_j] \leq \mathrm{P}[E_i]\mathrm{P}[E_j]$ and $\sum_{n=1}^{\infty} \mathrm{P}[E_n] = \infty$, then $\mathrm{P}[\limsup_n E_n] = 1$.

Limits of sequences of random variables

1. (Lebesgue's monotone convergence theorem). Let $(f_n)_n$ be an increasing sequence of nonnegative Borel functions and let $f := \lim_n f_n$ (in the sense $f_n \to f$ point-wise a.e.). Then $\mathbb{E}[f_n] \uparrow \mathbb{E}[f]$.

2. (Lebesgue's Dominated Convergence Theorem). Let X_n be real-valued RVs over $(\Omega, \mathcal{F}, \mathrm{P})$. Suppose that X_n converges point-wise to X and is *dominated* by a $Y \in \mathcal{L}_1(\Omega, \mathcal{F}, \mathrm{P})$, that is $|X_n| \leq Y$ P-a.s for all $n \in \mathbb{N}$. Then, $X \in \mathcal{L}_1(\Omega, \mathcal{F}, \mathrm{P})$ and

$$\lim_n \mathbb{E}[|X_n - X|] = 0,$$

which implies

$$\lim_n \mathbb{E}[X_n] = \mathbb{E}[X].$$

3. (Dominated convergence in $\mathcal{L}^p$). For $p \in [1, \infty)$ and a sequence of random variables $X_k : (\Omega, \mathcal{F}, \mathrm{P}) \to \overline{\mathbb{R}}$, assume that $X_k \to X$ almost everywhere ($X(\omega) = \lim_k X_k(\omega)$ P-a.e.) and there is $Y \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ so that $X_k \leq Y$. Then,
   i. $X_k \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ for all $k \in \mathbb{N}$,
   ii. $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$
   iii. $X_k \to X$ in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$, that is $\lim_k \|X_k - X\|_p = 0$.

4. (Consequence of the dominated convergence theorem) [8]. Let $\{E_k\}_{k=1}^{\infty}$ be a collection of disjoint events and let $E = \bigcup_k E_k$. Then,

$$\int_E f = \sum_{k=1}^{\infty} \int_{E_k} f.$$

5. (Bounded convergence). If $X_k \to X$ almost surely and $\sup_k |X_k| \leq b$ for some constant $b > 0$, then $\mathbb{E}[X_k] \to \mathbb{E}[X]$ and $\mathbb{E}[|X|] \leq b$.

6. (Fatou's lemma). Let $X_n \geq 0$ be a sequence of random variables. Then,

$$\mathbb{E}[\liminf_n X_n] \leq \liminf_n \mathbb{E}[X_n].$$

7. (Fatou's lemma with varying measures). For a sequence of nonnegative random variables $X_n \geq 0$ over $(\Omega, \mathcal{F}, \mathrm{P})$, and a sequence of (probability) measures μ_n which converge strongly to a (probability) measure μ (that is, $\mu_n(A) \to \mu(A)$ for all $A \in \mathcal{F}$), we have

$$\mathbb{E}_\mu[\liminf_n X_n] \leq \liminf_n \mathbb{E}_{\mu_n}[X_n]$$

8. (Reverse Fatou's lemma). Let $X_n \geq 0$ be a sequence of nonnegative random variables over $(\Omega, \mathcal{F}, \mathrm{P})$ and assume there is a $Y \in \mathcal{L}_1(\Omega, \mathcal{F}, \mathrm{P})$ so that $X_n \leq Y$. Then
$$\limsup_n \mathbb{E}[X_n] \leq \mathbb{E}[\limsup_n X_n]$$
9. (Integrable lower bound). Let X_n be a sequence of random variables over $(\Omega, \mathcal{F}, \mathrm{P})$. Suppose, there exists a $Y \geq 0$ such that $X_n \geq -Y$ for all $n \in \mathbb{N}$. Then,
$$\mathbb{E}[\liminf_n X_n] \leq \liminf_n \mathbb{E}[X_n].$$
10. (Beppo Levi's Theorem). Let X_k be a sequence of nonnegative random variables on $(\Omega, \mathcal{F}, \mathrm{P})$ with $0 \leq X_1 \leq X_2 \leq \ldots$. Let $X(\omega) = \lim_{k\to\infty} X_k(\omega)$. Then X is a random variable and
$$\lim_{k\to\infty} \mathbb{E}[X_k] = \mathbb{E}[\lim_{k\to\infty} X_k].$$
11. (Beppo Levi's Theorem for series). Let X_k be a sequence of nonnegative integrable random variables on $(\Omega, \mathcal{F}, \mathrm{P})$ and let $Y_k = \sum_{j=0}^{k} X_k$. Assume that $\sum_{k=1}^{\infty} \mathbb{E}[Y_k]$ converges. Then Y_k satisfies the conditions of the BL theorem and
$$\sum_{k=1}^{\infty} \mathbb{E}[Y_k] = \mathbb{E}\left[\sum_{k=1}^{\infty} Y_k\right].$$
12. (Uniform integrability – definition) [7]. A collection $\{X_k\}_{k\in T}$ is said to be *uniformly integrable* if $\sup_{t\in T} \mathbb{E}[|X_t| 1_{|X_t|>x}] \to 0$ as $x \to \infty$.
13. (Constant absolutely integrable sequences as uniformly integrable) [7]. The sequence $\{Y\}_{t\in T}$ with $\mathbb{E}[|Y|] < \infty$ is uniformly integrable.
14. (Uniform boundedness in $\mathcal{L}^p$, $p > 1$, implies uniform integrability). If $\{X_t\}_{t\in T}$ is uniformly bounded in $\mathcal{L}^p$, $p > 1$ (that is, $\mathbb{E}[|X_k|^p] < c$ for some $c > 0$), then it is uniformly integrable.
15. (Convergence under uniform integrability) [7]. If $X_k \to X$ a.s. and $\{X_k\}_k$ is uniformly integrable then
    i. $\mathbb{E}[X] < \infty$
    ii. $\mathbb{E}[X_k] \to \mathbb{E}[X]$
    iii. $\mathbb{E}|X_k - X| \to 0$

1.1.4 The Radon-Nikodym Theorem

1. (Absolute continuity). Let $(\mathcal{X}, \mathcal{G})$ be a measurable space and μ and ν two measures on it. We say that ν is *absolutely continuous* with respect to μ if for all $A \in \mathcal{G}$, $\nu(A) = 0$ whenever $\mu(A) = 0$. We denote this by $\nu \ll \mu$.
2. (Radon-Nikodym). Let $(\mathcal{X}, \mathcal{G})$ be a measurable space, let ν be a *σ-finite* measure on $(\mathcal{X}, \mathcal{G})$ which is absolutely continuous with respect to a measure μ on $(\mathcal{X}, \mathcal{G})$. Then, there is a measurable function $f : \mathcal{X} \to [0, \infty)$ such that for all $A \in \mathcal{G}$
$$\nu(A) = \int_A f \mathrm{d}\mu.$$
   This function is denoted by $f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.
3. (Linearity). Let ν, μ and λ be σ-finite measures on $(\mathcal{X}, \mathcal{G})$ and $\nu \ll \lambda$, $\mu \ll \lambda$. Then
$$\frac{\mathrm{d}(\nu+\mu)}{\mathrm{d}\lambda} = \frac{\mathrm{d}\nu}{\mathrm{d}\lambda} + \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}, \ \lambda\text{-a.e.}$$
4. (Chain rule). If $\nu \ll \mu \ll \lambda$,
$$\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}, \ \lambda\text{-a.e.}$$

5. (Inverse). If $\nu \ll \mu$ and $\mu \ll \nu$, then

$$\frac{\mathrm{d}\mu}{\mathrm{d}\nu} = \left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)^{-1}, \ \nu\text{-a.e.}$$

6. (Change of measure). If $\mu \ll \lambda$ and g is a μ-integrable function, then

$$\int_{\mathcal{X}} g\mathrm{d}\mu = \int_{\mathcal{X}} g\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\mathrm{d}\lambda.$$

7. (Change of variables in integration). This was addressed using the push-forward.

$$\mathbb{E}[g(X)] = \int g \circ X\mathrm{dP} = \int_{\mathbb{R}} g\,\mathrm{d}(X_*\mathrm{P}).$$

If the measure $\mathrm{d}(X_*\mathrm{P})$ is absolutely continuous with respect to the Lebesgue measure μ (on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, then, the Radon-Nikodym derivative $f_X := \frac{\mathrm{d}(X_*\mathrm{P})}{\mathrm{d}\mu}$, where $f_X : \mathbb{R} \to \mathbb{R}$ exists. Then

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g\,\mathrm{d}(X_*\mathrm{P}) = \int_{\mathbb{R}} gf_X\,\mathrm{d}\mu = \int_{\mathbb{R}} g(\tau)f_X(\tau)\,\mathrm{d}\tau.$$

This is known as the *law of the unconscious statistician* (LotUS).

1.1.5 Probability distribution

1. (Probability distribution). Let $X : (\Omega, \mathcal{F}, \mathrm{P}) \to (Y, \mathcal{G})$ be a random variable. The measure

$$F_X(A) = \mathrm{P}[X \in A] = \mathrm{P}[\{\omega \in \Omega \mid X(\omega) \in A\}] = \mathrm{P}[X^{-1}A] = (X_*\mathrm{P})(A),$$

is called the *probability distribution* of X and it is a measure . Note that for all $A \in \mathcal{G}$, $X^{-1}A \in \mathcal{F}$ since X is measurable.

2. (Probability distribution of real-valued random variables). The *probability distribution* or *cumulative distribution function* of a random variable X on a space $\mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ is $F_X(x) = \mathrm{P}[X \leq x]$ for $x \in \mathbb{R}$. The inverse cumulative distribution of X is $F_X^{-1}(p)$ for $p \in [0,1]$ is defined as $F_X^{-1} = \inf\{x \in \mathbb{R} : F_X(x) \geq p\}$.

3. (Push-forward). The probability distribution of a random variable X with values in $(\mathcal{X}, \mathcal{G})$, is the push-forward measure $X_*\mathrm{P}$ on $(\mathcal{X}, \mathcal{G})$ which is a probability measure on $(\mathcal{X}, \mathcal{G})$ with $X_*\mathrm{P} = \mathrm{P}X^{-1}$.

4. (Associated p-system). We associate with $F_X : \mathbb{R} \to [0,1]$ the measure μ which is defined on the p-system $\{(-\infty, x]\}_{x\in\mathbb{R}}$ as $\mu((-\infty, x]) = F_X(x)$.

5. (Properties of the cumulative and the inverse cumulative distributions). The notation $X \sim Y$ means that X and Y have the same cumulative distribution, that is $F_X = F_Y$.
   i. If $Y \sim U[0,1]$, then $F_X^{-1}(Y) \sim X$.
   ii. F_X is càdlàg
   iii. $x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$
   iv. $\mathrm{P}[X > x] = 1 - F_X(x)$
   v. $\mathrm{P}[\{x_1 < X \leq x_2\}] = F_X(x_2) - F_X(x_1)$
   vi. $\lim_{x\to-\infty} F_X(x) = 0$, $\lim_{x\to\infty} F_X(x) = 1$
   vii. $F_X^{-1}(F_X(x)) \leq x$
   viii. $F_X(F_X^{-1}(p)) \geq p$
   ix. $F_X^{-1}(p) \leq x \Leftrightarrow p \leq F_X(x)$

1.1.6 Probability density function

1. (Definition). The probability density function f_X of a random variable $X : (\Omega, \mathcal{F}, \mathrm{P}) \to (\mathcal{X}, \mathcal{G})$ with respect to a measure μ on $(\mathcal{X}, \mathcal{G})$ is the Radon-Nikodym derivative

$$f_X = \frac{\mathrm{d}(X_*\mathrm{P})}{\mathrm{d}\mu},$$

which exists provided that $X_*\mathrm{P} \ll \mu$, and f_X is measurable and μ-integrable. Then,

$$\mathrm{P}[X \in A] = \int_{X^{-1}A} \mathrm{dP} = \int_\Omega 1_{X^{-1}A}\mathrm{dP} = \int_\Omega (1_A \circ X)\mathrm{dP} = \int_A \mathrm{d}(X_*\mathrm{P}) = \int_A f_X \mathrm{d}\mu.$$

2. (Probability distribution). If X is a real-valued random variable and its range ($\mathbb{R}$) is taken with the Borel σ-algebra, then

$$\mathrm{P}[X \leq x] = \int_{(-\infty,x]} X\mathrm{dP} = \int_{\{\omega\in\Omega : X(\omega)\leq x\}} \mathrm{dP} = \int_{-\infty}^{x} f_X \mathrm{d}\mu$$

Note that the first integral is written with a slight abuse of notation as the integration with respect to P is carried out over the set $\{\omega \in \Omega : X(\omega) \leq x\}$; The first integral can be understood as shorthand notation for the second integral.

3. (Expectation). Let a real-valued random variable X have probability density f_X. Let ι be the identity function $\iota : x \mapsto x$ on Ω. Then

$$\mathbb{E}[X] = \int_\Omega X\mathrm{dP} = \int_\Omega (\iota \circ X)\mathrm{dP} = \int_{\mathbb{R}} \iota \mathrm{d}(X_*\mathrm{P}) = \int_{\mathbb{R}} \iota(x) f_X(x)\mathrm{d}\mu = \int_{\mathbb{R}} x f_X(x)\mathrm{d}x.$$

4. (Distribution of transformation). Let $g : \mathbb{R} \to \mathbb{R}$ be a strictly increasing function. Let X be a real-valued random variable with probability density function f_X and let $Y(\omega) = g(X(\omega))$ be another random variable. Then

$$\begin{aligned} F_Y(y) &= F_X(g^{-1}(y)), \\ f_Y(y) &= f_X(g^{-1}(y))\frac{\partial g^{-1}(y)}{\partial y}. \end{aligned}$$

5. (Expectation of transformation). Let X be a real-valued random variable on $(\Omega, \mathcal{F}, \mathrm{P})$ with probability density function f_X and let $Y(\omega) = g(X(\omega))$ be another random variable. Then

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} f_X(\tau)g(\tau)\mathrm{d}\tau.$$

If $\Omega = \{\omega_i\}_{i=1}^n$, $\mathcal{F} = 2^\Omega$ and $\mathrm{P}[\{\omega\}_i] = p_i$, then

$$\mathbb{E}[Y] = \sum_{i=1}^{n} p_i g(X(\omega_i)).$$

See also: law of the unconscious statistician.

1.1.7 Decomposition of measures

Does a density function always exist? The answer is negative, but Lebesgue's decomposition theorem offers some further insight.

1. (Singular measures). Let $(\Omega, \mathcal{F})$ be a measurable space and μ, ν be two measures defined thereon. These are called *singular* if there are $A, B \in \mathcal{F}$ so that
   i. $A \cup B = \Omega$,
   ii. $A \cap B = \varnothing$,
   iii. $\mu(B') = 0$ for all $B' \in \mathcal{F}$ with $B' \subseteq B$,

iv. $\nu(A') = 0$ for all $A' \in \mathcal{F}$ with $A' \subseteq A$.

2. (Discrete measure on $\mathbb{R}$). A measure μ on $\mathbb{R}$ equipped with the Lebesgue σ-algebra, is said to be discrete if there is a (possibly finite) sequence of elements $\{s_k\}_{k\in\mathbb{N}}$, so that

$$\mu(\mathbb{R} \setminus \bigcup_{k\in\mathbb{N}} \{s_k\}) = 0.$$

3. (Lebesgue's decomposition Theorem). For every two σ-finite signed measures μ and ν on a measurable space $(\Omega, \mathcal{F})$, there exist two σ-finite signed measures ν_0 and ν_1 on $(\Omega, \mathcal{F})$ such that

   i. $\nu = \nu_0 + \nu_1$

   ii. $\nu_0 \ll \mu$

   iii. $\nu_1 \perp \mu$

   and ν_0 and ν_1 are uniquely determined by ν and μ.

4. (Lebesgue's decomposition Theorem — Corollary). Consider the space $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ and let μ be the Lebesgue measure. Any probability measure ν on this space can be written as

$$\nu = \nu_{\mathrm{ac}} + \nu_{\mathrm{sc}} + \nu_{\mathrm{d}},$$

   where $\nu_{\mathrm{ac}} \ll \mu$ (which is easily understood via the Radon-Nikodym Theorem), ν_{sc} is singular continuous (wrt μ) and ν_{d} is a discrete measure.

1.1.8 $\mathcal{L}^p$ spaces

1. (p-norm). Let X be a real-valued random variable on $(\Omega, \mathcal{F}, \mathrm{P})$. For $p \in [1, \infty)$ define the p-norm of X as

$$\|X\|_p = \mathbb{E}[|X|^p]^{1/p}.$$

2. ($\mathfrak{L}^p$ spaces). Define $\mathfrak{L}^p(\Omega, \mathcal{F}, \mathrm{P}) = \{X : \Omega \to \mathbb{R},\ \text{measurable}, \|X\|_p < \infty\}$ and equip this space with the addition and scalar multiplication operations $(X+Y)(\omega) = X(\omega)+Y(\omega)$ and $(\alpha X)(\omega) = \alpha X(\omega)$. This becomes a semi-normed space[3].

3. ($\mathcal{L}^p$ spaces). Define $\mathcal{N}(\Omega, \mathcal{F}, \mathrm{P}) = \{X : \Omega \to \mathbb{R},\ \text{measurable}, X = 0 \text{ a.s.}\}$; this is the kernel of $\|\cdot\|_p$. Then, define $\mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P}) = \mathfrak{L}^p(\Omega, \mathcal{F}, \mathrm{P})/\mathcal{N}$. This is a normed space where for $X \in \mathfrak{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ and $[X] = X + \mathcal{N} \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ we have $\|[X]\|_p := \|X\|_p$.

4. (∞-norm, $\mathfrak{L}_\infty$ and $\mathcal{L}_\infty$). The infinity norm is defined as

$$\|X\|_\infty = \operatorname{esssup}|X| = \inf\{\lambda \in \mathbb{R} : \mathrm{P}[|X| > \lambda] = 0\},$$

   or equivalently

$$\|X\|_\infty = \inf\{\lambda \in \mathbb{R} : |X| \leq \lambda,\ \text{P-a.s.}\}.$$

   The spaces $\mathfrak{L}_\infty(\Omega, \mathcal{F}, \mathrm{P})$ and $\mathcal{L}_\infty(\Omega, \mathcal{F}, \mathrm{P})$ are defined similarly.

5. ($\mathcal{L}_\infty(\Omega, \mathcal{F}, \mathrm{P})$ as a limit). If there is a $p' \in [1, \infty)$ such that $X \in \mathcal{L}_\infty \cap \mathcal{L}_{p'}$, then

$$\|X\|_\infty = \lim_{p\to\infty} \|X\|_p.$$

6. ($\mathcal{L}_2$ is a Hilbert space). $\mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ is the only Hilbert $\mathcal{L}^p$ space with inner product

$$\langle X, Y \rangle = \mathbb{E}[XY].$$

[3] $\|X\| = 0$ does not imply that $X = 0$, but instead that $X = 0$ almost surely. However, $\|\cdot\|_p$ is absolutely homogeneous, sub-additive and nonnegative

1.1.9 Product spaces

1. (Product σ-algebra). Let $\{X_a\}_{a\in A}$ be an indexed collection of nonempty sets; define $X = \prod_{a\in A} X_a$ and $\pi_a : X = (x_a)_{a\in A} \mapsto x_a \in X_a$. Let $\mathcal{F}_a$ be a σ-algebra on X_a. We define the product σ-algebra as
$$\bigotimes_{a\in A} \mathcal{F}_a := \sigma\left(\{\pi_a^{-1}(E_a); a \in A, E_a \in \mathcal{F}_a\}\right)$$
This is the smallest σ-algebra on the product space which renders all projections measurable (compare to the definition of the *product topology* which is the smallest topology on the product space which renders the projections *continuous*).

2. (Measurability of epigraphs). Let $f : (X, \mathcal{F}) \to \overline{\mathbb{R}}$ be a measurable proper function. Its epigraph, that is the set $\operatorname{epi} f := \{(x, \alpha) \in X \times \mathbb{R} \mid f(x) \leq \alpha\}$ and its hypo-graph, that is the set $\operatorname{hyp} f := \{(x, \alpha) \in X \times \mathbb{R} \mid f(x) \geq \alpha\}$ are measurable in the product measure space $(X \times \mathbb{R}, \mathcal{F} \otimes \mathcal{B}_{\mathbb{R}})$.

3. (Measurability of graph). The graph of a measurable function $f : (X, \mathcal{F}, \mu) \to \mathbb{R}$ is a Lebesgue-measurable set with Lebesgue measure zero.

4. (Countable product of σ-algebras). If A is countable, the product σ-algebra is generated by the products of measurable sets $\{\prod_{a\in A} E_a; E_a \in \mathcal{F}_a\}$.

5. (Product measures). Let $(\mathcal{X}, \mathcal{F}, \mu)$ and $(\mathcal{Y}, \mathcal{G}, \nu)$ be two measure spaces. The product space $\mathcal{X} \times \mathcal{Y}$ becomes a measurable space with the σ-algebra $\mathcal{F} \otimes \mathcal{G}$. Let $E_x \in \mathcal{F}$ and $E_y \in \mathcal{G}$; then $E_x \times E_y \in \mathcal{F} \otimes \mathcal{G}$. We define a measure $\mu \times \nu$ on $(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \otimes \mathcal{G})$ with
$$(\mu \times \nu)(E_x \times E_y) = \mu(E_x)\nu(E_y).$$

6. Let $E \in \mathcal{F} \otimes \mathcal{G}$ and define $E_x = \{y \in \mathcal{Y} : (x, y) \in E\}$ and $E_y = \{x \in \mathcal{X} : (x, y) \in E\}$. Then, $E_x \in \mathcal{F}$ for all $x \in \mathcal{X}$, $E_y \in \mathcal{G}$ for all $y \in \mathcal{Y}$.

7. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be an $\mathcal{F} \otimes \mathcal{G}$-measurable function. Then, $f(x, \cdot)$ is $\mathcal{G}$-measurable for all $x \in \mathcal{X}$ and $f(\cdot, y)$ is $\mathcal{F}$-measurable for all $y \in \mathcal{Y}$.

8. Let $(\mathcal{X}, \mathcal{F}, \mu)$ and $(\mathcal{Y}, \mathcal{G}, \nu)$ be two σ-finite measure spaces. For $E \in \mathcal{F} \otimes \mathcal{G}$, the mappings $\mathcal{X} \ni x \mapsto \nu(E_x) \in \mathbb{R}$ and $\mathcal{Y} \ni y \mapsto \mu(E_y)$ are measurable and
$$(\mu \times \nu)(E) = \int \nu(E_x)\mathrm{d}\mu(x) = \int \mu(E_y)\mathrm{d}\nu(x)$$

9. (Tonelli's Theorem). Let $h : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ be an $\mathcal{F} \otimes \mathcal{G}$-measurable function. Let
$$f(x) = \int_{\mathcal{Y}} h(x, y)\mathrm{d}\nu(y), \;\; g(y) = \int_{\mathcal{X}} h(x, y)\mathrm{d}\mu(x).$$
Then, f and g are measurable and
$$\int_{\mathcal{X}} f\mathrm{d}\mu = \int_{\mathcal{Y}} g\mathrm{d}\nu = \int_{\mathcal{X}\times\mathcal{Y}} g\mathrm{d}(\mu \times \nu).$$

10. (Fubini's Theorem). Let $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be an $\mathcal{F} \otimes \mathcal{G}$-measurable function and
$$\int_{\mathcal{X}} \int_{\mathcal{Y}} h(x, y)\mathrm{d}\nu(y)\mathrm{d}\mu(x) < \infty.$$
Then, $h \in \mathcal{L}_1(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \otimes \mathcal{G}, \mu \times \nu)$ and
$$\int_{\mathcal{X}} \int_{\mathcal{Y}} h(x, y)\mathrm{d}\nu(y)\mathrm{d}\mu(x) = \int_{\mathcal{Y}} \int_{\mathcal{X}} h(x, y)\mathrm{d}\mu(x)\mathrm{d}\nu(y) = \int_{\mathcal{X}\times\mathcal{Y}} h\mathrm{d}(\mu \times \nu)$$

11. (Consequence of Fubini's theorem). Let X be a nonnegative random variable. Let $E = \{(\omega, x) : 0 \leq x \leq X(\omega)\}$. Then, $X(\omega) = \int_0^\infty 1_E(\omega, x)\mathrm{d}x$.
$$\begin{aligned}\mathbb{E}[X] = \int_\Omega X\mathrm{dP} &= \int_\Omega \int_0^\infty 1_E(\omega, x)\mathrm{d}x\mathrm{dP}\\ &= \int_0^\infty \int_\Omega 1_E(\omega, x)\mathrm{dP}\mathrm{d}x\\ &= \int_0^\infty \mathrm{P}[X \geq x]\mathrm{d}x.\end{aligned}$$

1.1.10 Transition Kernels

1. (Definition). Let $(\mathcal{X}, \mathcal{F})$, $(\mathcal{Y}, \mathcal{G})$ be two measurable spaces and let $K : \mathcal{G} \times \mathcal{X} \to [0,1]$. K is called a *(probability) transition kernel* if
   i. $f_B(x) := K(B,x)$ is $\mathcal{F}$-measurable for every $B \in \mathcal{G}$,
   ii. $\mu_x(B) := K(B,x)$ is a measure on $(\mathcal{Y}, \mathcal{G})$ for every $x \in \mathcal{X}$.
2. (Markov kernel). A kernel $K : \mathcal{G} \times \mathcal{X} \to [0,1]$ is called a *Markov kernel* if $K(\mathcal{Y}, x) = 1$ for all $x \in \mathcal{X}$
3. (Existence of transition kernels). Let μ be a finite measure on $(\mathcal{X}, \mathcal{F})$ and $k : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be measurable in the product σ-algebra $\mathcal{F} \otimes \mathcal{G}$ and has the property $\int_{\mathcal{Y}} k(x,y)\nu(\mathrm{d}y) = 1$. Then the mapping $K : \mathcal{X} \times \mathcal{G} \to [0,1]$ given by
$$K(B,x) = \int_B k(x,y)\mu(\mathrm{d}y),$$
is a probability transition kernel.
4. (Measure on product space via a kernel). Let $(\mathcal{X}, \mathcal{F})$, $(\mathcal{Y}, \mathcal{G})$ be two measurable spaces and let $K : \mathcal{G} \times \mathcal{X} \to [0,1]$ be a transition kernel. For $A \in \mathcal{F}$ and $B \in \mathcal{G}$ define
$$\mu(A \times B) = \int_A K(B,x)\mathrm{dP}(x).$$
This extends to a unique measure on the product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \otimes \mathcal{G})$.

1.1.11 Law invariance

1. (Equality in distribution). Let X, Y be two real-valued random variables on $(\Omega, \mathcal{F}, \mathrm{P})$. We say that X and Y are equal in distribution, and we denote $X \overset{\mathrm{d}}{\sim} Y$, if X and Y have equal probability distribution functions, that is $F_X(s) = F_Y(s)$ for all s.
2. (Equal in distribution, nowhere equal). Let $\Omega = \{-1, 1\}$, $\mathcal{F} = 2^\Omega$, $\mathrm{P}[\{\omega_i\}] = \frac{1}{2}$. Let $X(\omega) = \omega$ and $Y(\omega) = -X(\omega)$. These two variables have the same distribution, but are nowhere equal.
3. (Equal in distribution, almost nowhere equal). Take $X \sim \mathcal{N}(0,1)$ and $Y = -X$. These two random variables are almost nowhere equal, but have the same distribution.
4. The following are equivalent:
   i. $X \overset{\mathrm{d}}{\sim} Y$
   ii. $\mathbb{E}[e^{-rX}] = \mathbb{E}[e^{-rY}]$ for all $r > 0$
   iii. $\mathbb{E}[f(X)] = \mathbb{E}[f(Y)]$ for all bounded continuous functions
   iv. $\mathbb{E}[f(X)] = \mathbb{E}[f(Y)]$ for all bounded Borel functions
   v. $\mathbb{E}[f(X)] = \mathbb{E}[f(Y)]$ for all positive Borel functions

1.1.12 Expectation

1. (Definition) Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and X be a random variable. Then, the expected value of X is denoted by $\mathbb{E}[X]$ and is defined as the Lebesgue integral
$$\mathbb{E}[X] = \int_\Omega X\mathrm{dP}$$
2. Because of item 8 in Sec. 1.1.2, for $X \geq 0$ nonnegative
$$\begin{aligned}\mathbb{E}[X] &= \int_0^{+\infty} X\mathrm{dP}\\ &= \int_0^{+\infty}\int_0^{+\infty} 1_{X\geq t}\mathrm{d}t\mathrm{dP}\\ &= \int_0^{+\infty}\int_0^{+\infty} 1_{X\geq t}\mathrm{dP}\mathrm{d}t\end{aligned}$$

and we use the fact that

$$\int_0^{+\infty} 1_{X>t}\mathrm{dP} = \mathrm{P}[X > t],$$

so

$$\mathbb{E}[X] = \int_0^{\infty} \mathrm{P}[X > t]\mathrm{d}t.$$

The function $S(t) = \mathrm{P}[X > t] = 1 - \mathrm{P}[X \leq t]$ is called the *survival function* of X, or its *tail distribution* or *exceedance.*

3. (Expectation in terms of PDF). Let X be a real-valued continuous random variable with PDF f_X. Then,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x.$$

4. (Expectation in terms of CDF). Let X be a real-valued random variable. Then,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\mathrm{d}F(x).$$

5. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and X a real-valued random variable thereon. Define

$$f(\tau) = \int_{\Omega} (X - \tau)^2 \mathrm{dP}.$$

Then $\tau = \mathbb{E}[X]$ minimizes f and the minimum value is $\mathrm{Var}[X]$.

6. Let X be a real-valued random variable. Then,

$$\sum_{n=1}^{\infty} \mathrm{P}[|X| \geq n] \leq \mathbb{E}[|X|] \leq 1 + \sum_{n=1}^{\infty} \mathrm{P}[|X| \geq n].$$

It is $\mathbb{E}[|X|] < \infty$ if and only if the above series converges.

7. If X takes positive integer values, then

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} \mathrm{P}[X \geq n]$$

8. (Finite mean, infinite variance). There are several distributions with finite mean and infinite variance — a standard example is the *Pareto distribution.* A random variable X follows the Pareto distribution with parameters $x_m > 0$ and a if it has support $[x_m, \infty)$ and probability distribution

$$\mathrm{P}[X \leq x] = \frac{a x_m^a}{x^{a+1}},$$

for $x \geq x_m$. For $a \leq 1$, X has infinite mean and variance. For $a > 1$, its mean is $\mathbb{E}[X] = \frac{a x_m}{a-1}$ and infinite variance.

9. (Absolutely bounded a.s. $\Leftrightarrow$ Bounded moments) [11]. Let X be a random variable on $(\Omega, \mathcal{F}, \mathrm{P})$. The following are equivalent:

   i. X is almost surely absolutely bounded (i.e., there is $M \geq 0$ such that $\mathrm{P}[|X| \leq M] = 1$)

   ii. $\mathbb{E}[|X|^k] \leq M^k$, for all $k \in \mathbb{N}_{\geq 1}$

10. (A useful formula) [2]. For $q > 0$

$$\mathbb{E}[|X|^q] = \int_0^{\infty} q x^{q-1} \mathrm{P}[|X| > x]\mathrm{d}x.$$

1.2 Conditioning

1.2.1 Conditional Expectation

1. (Conditional Expectation). Let X be a random variable on $(\Omega, \mathcal{F}, \mathrm{P})$ and $\mathcal{H} \subseteq \mathcal{F}$. A *conditional expectation* of X given $\mathcal{H}$ is an $\mathcal{H}$-measurable random variable, denoted as $\mathbb{E}\left[X \mid \mathcal{H}\right]$, with

$$\int_H \mathbb{E}\left[X \mid \mathcal{H}\right] \mathrm{dP} = \int_H X \mathrm{dP},$$

which equivalently can be written as

$$\mathbb{E}[X 1_H] = \mathbb{E}[\mathbb{E}\left[X \mid \mathcal{H}\right] 1_H],$$

for all $H \in \mathcal{H}$.

2. (Uniqueness). All versions of a conditional expectation, $\mathbb{E}\left[X \mid \mathcal{H}\right]$, differ only on a set of measure zero[4].

3. (Equivalent definition). It is equivalent to define the conditional expectation of X, conditioned by a σ-algebra $\mathcal{H}$ as a random variable $\mathbb{E}\left[X \mid \mathcal{H}\right]$ with the property

$$\mathbb{E}[XZ] = \mathbb{E}[\mathbb{E}\left[X \mid \mathcal{H}\right] Z],$$

for all $\mathcal{H}$-measurable random variables Z.

4. (Best estimator). Assuming $\mathbb{E}[Y^2] < \infty$, the best estimator of Y given X is $\mathbb{E}[Y \mid X]$

5. (Radon-Nikodym definition). The conditional expectation as introduced above, is the Radon-Nikodym derivative

$$\mathbb{E}\left[X \mid \mathcal{H}\right] = \frac{\mathrm{d}\mu_{\mathcal{H}}^X}{\mathrm{dP}_{\mathcal{H}}},$$

where $\mu_{\mathcal{H}}^X : \mathcal{H} \to [0, \infty]$ is the measure induced by X restricted on $\mathcal{H}$, that is $\mu_{\mathcal{H}}^X : H \mapsto \int_H X \mathrm{dP}$. This is absolutely continuous with respect to P. The measure $\mathrm{P}_{\mathcal{H}}$ is the restriction of P on $\mathcal{H}$.

6. (Conditional expectation wrt random variable). Let X, Y be random variables on $(\Omega, \mathcal{F}, \mathrm{P})$. The conditional expectation of X given Y is $\mathbb{E}[X \mid Y] := \mathbb{E}[X \mid \sigma(Y)]$, where $\sigma(Y)$ is the σ-algebra generated by Y, that is $\sigma(Y) = Y^{-1}(\mathcal{F}) = \{Y^{-1}(B); B \in \mathcal{F}\}$.

7. (Conditional expectation using the push-forward $Y_*\mathrm{P}$). Let X be an integrable random variable on $(\Omega, \mathcal{F}, \mathrm{P})$. Then, there is a $Y_*\mathrm{P}$-unique random variable $\mathbb{E}[X \mid Y]$

$$\int_{Y^{-1}(B)} X \mathrm{dP} = \int_B \mathbb{E}[X \mid Y] \mathrm{d}(Y_*\mathrm{P}).$$

8. (Conditioning by an event). The conditional expectation $\mathbb{E}[X \mid H]$, conditioned by an event $H \in \mathcal{F}$ is given by

$$\mathbb{E}[X \mid H] = \frac{1}{\mathrm{P}[H]} \int_H X \mathrm{dP} = \frac{1}{\mathrm{P}[H]} \mathbb{E}[X 1_H].$$

9. (Properties of conditional expectations). The conditional expectation has the following properties:
   i. (Monotonicity). $X \leq Y \Rightarrow \mathbb{E}\left[X \mid \mathcal{H}\right] \leq \mathbb{E}\left[Y \mid \mathcal{H}\right]$
   ii. (Positivity). $X \geq 0 \Rightarrow \mathbb{E}\left[X \mid \mathcal{H}\right] \geq 0$ [Set $Y = 0$ in 9i].
   iii. (Linearity). For $a, b \in \mathbb{R}$, $\mathbb{E}\left[aX + bY \mid \mathcal{H}\right] = a\mathbb{E}\left[X \mid \mathcal{H}\right] + b\mathbb{E}\left[Y \mid \mathcal{H}\right]$
   iv. (Monotone convergence). $X_n \geq 0$, $X_n \uparrow X$ implies $\mathbb{E}\left[X_n \mid \mathcal{H}\right] \uparrow \mathbb{E}\left[X \mid \mathcal{H}\right]$
   v. (Fatou's lemma). For $X_n \geq 0$, $\mathbb{E}\left[\liminf_n X_n \mid \mathcal{H}\right] \leq \liminf_n \mathbb{E}\left[X_n \mid \mathcal{H}\right]$
   vi. (Reverse Fatou's lemma).

[4] R. Durrett, "Probability: Theory and Examples," 2013, Available at: https://services.math.duke.edu/~rtd/PTE/PTE4_1.pdf

vii. (Dominated convergence theorem). $X_n \to X$ (point-wise) and $|X_n| \leq Y$ P-a.s. where Y is integrable. Then, $\mathbb{E}[X \mid \mathcal{H}]$ is integrable and

$$\mathbb{E}[X_n \mid \mathcal{H}] \to \mathbb{E}[X \mid \mathcal{H}].$$

viii. (Jensen's inequality). Let $X \in \mathcal{L}_1(\Omega, \mathcal{F}, \mathrm{P})$, $f : \mathbb{R} \to \mathbb{R}$ convex. Then

$$f(\mathbb{E}[X \mid \mathcal{H}]) \leq \mathbb{E}[f(X) \mid \mathcal{H}].$$

ix. (Law of total expectation). For any σ-algebra $\mathcal{H} \subseteq \mathcal{F}$,

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}]] = \mathbb{E}[X].$$

x. (Tower property). For two σ-algebras $\mathcal{H}_1$ and $\mathcal{H}_2$ with $\mathcal{H}_1 \subseteq \mathcal{H}_2$,

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}_1] \mid \mathcal{H}_2] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{H}_2] \mid \mathcal{H}_1] = \mathbb{E}[X \mid \mathcal{H}_1].$$

xi. (Tower property with X being $\mathcal{H}_i$-measurable). Let $\mathcal{H}_1 \subseteq \mathcal{H}_2$ be two σ-algebras. If X is $\mathcal{H}_1$-measurable, then it is also $\mathcal{H}_2$-measurable.

xii. If X is $\mathcal{H}$-measurable then

$$\mathbb{E}[X \mid \mathcal{H}] = X.$$

1.2.2 Conditional Probability

1. (Conditional probability). Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $\mathcal{H}$ be a sub-σ-algebra of $\mathcal{F}$. We define $\mathrm{P}_{\mathcal{H}}$ as an operator so that for all $H \in \mathcal{H}$

$$\mathrm{P}_{\mathcal{H}}[H] = \mathbb{E}_{\mathcal{F}} 1_H.$$

2. (Conditional probability given an event). For $E, H \in \mathcal{F}$, $\mathrm{P}[E \cap H] = \mathrm{P}[H]\mathrm{P}_H[E]$. This is uniquely defined provided that $\mathrm{P}[H] > 0$.

1.2.3 Construction of probability spaces

1. (Inonescu-Tulcea's Theorem).

2. (Kolmogorov's Extension Theorem). Let T denote a time interval, $k \in \mathbb{N}$ and let $\nu_{t_1,\ldots,t_k}$ be probability measures on $\mathbb{R}^{nk}$ such that

   a) For all permutations π on $\{1, \ldots, k\}$ it holds that

   $$\nu_{\pi(t_1),\ldots,\pi(t_k)}(F_1 \times \cdots \times F_k) = \nu_{t_1,\ldots,t_k}(F_{\pi^{-1}(1)} \times \cdots \times F_{\pi^{-1}(k)}),$$

   b) For all $m \in \mathbb{N}$ the following holds

   $$\nu_{t_1,\ldots,t_k}(F_1 \times \cdots \times F_k) = \nu_{t_1,\ldots,t_k,t_{k+1},\ldots,t_{k+1}}(F_1 \times \cdots \times F_k \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n).$$

   Then, there exists a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and a stochastic process $(X_t)_t$ on Ω such that

   $$\nu_{t_1,\ldots,t_k}(F_1 \times \cdots \times F_k) = \mathrm{P}[X_{t_1} \in F_1, \ldots, X_{t_k} \in F_k],$$

   for all Borel sets F_i, $i = 1, \ldots, k$.

1.3 Inequalities on Probability Spaces

1.3.1 Inequalities on $\mathcal{L}^p$ spaces

1. (Hölder's inequality). If $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$, $Y \in \mathcal{L}_q(\Omega, \mathcal{F}, \mathrm{P})$ (where p, q are conjugate exponents), then $XY \in \mathcal{L}_1(\Omega, \mathcal{F}, \mathrm{P})$ and

$$\mathbb{E}[|XY|] = \|XY\|_1 \leq \|X\|_p \|Y\|_q.$$

2. (Cauchy-Schwarz inequality). This is Hölder's inequality with $p = q = 2$:

$$\|XY\|_1 \leq \|X\|_2 \|Y\|_2.$$

3. (Minkowski inequality). If $X, Y \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ ($p \in [1, \infty]$), then $X + Y \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ and $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$.

1.3.2 Generic inequalities involving probabilities or expectations

1. (Lyapunov's inequality). Let $0 < s < t$. Then
$$(\mathbb{E}[|X|^s])^{1/s} \leq (\mathbb{E}[|X|^t])^{1/t}.$$

2. (Markov's inequality). Let $X \geq 0$, integrable. For all $t > 0$,
$$\mathrm{P}[X > t] \leq \frac{\mathbb{E}[X]}{t}.$$

3. (Chebyshev's inequality). Let X have finite expectation μ and finite variance σ^2. Then
$$\mathrm{P}[|X-\mu| \geq t] \leq \frac{\sigma^2}{t^2}.$$

4. (Generalized Markov's inequality). Let X be a real-valued random variable and $f : \mathbb{R} \to \mathbb{R}_+$ be an increasing function. Then, for all $b \in \mathbb{R}$,
$$\mathrm{P}[X > b] \leq \frac{1}{f(b)}\mathbb{E}[f(X)]$$

5. (Gaussian tail inequality). Let $X \sim N(0,1)$. Then,
$$\mathrm{P}[|X| > \epsilon] \leq \frac{2e^{-\epsilon^2/2}}{\epsilon}.$$

6. (Hoeffding's lemma). Let $a \leq X \leq b$ be an RV with finite expectation $\mu = \mathbb{E}[X]$. Then
$$\mathbb{E}[e^{tX}] \leq e^{t\mu}e^{\frac{t^2(b-a)^2}{8}}.$$

7. (Corollary of Hoeffding's lemma). Let X be such that e^{tX} is integrable for $t \geq 0$. Then
$$\mathrm{P}[X > \epsilon] \leq \inf_{t\geq 0} e^{-t\epsilon}\mathbb{E}[e^{tX}].$$

8. (Jensen's inequality). Let $X \in \mathcal{L}_1(\Omega, \mathcal{F}, \mathrm{P})$, $f : \mathbb{R} \to \mathbb{R}$ convex. Then
$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

9. (Paley-Zygmund). Let $Z \geq 0$ be a random variable with finite variance. Then,
$$\mathrm{P}[Z > \theta\mathbb{E}[Z]] \geq (1-\theta)^2\frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]},$$
and this bound can be improved (using the Cauchy-Schwartz inequality) as
$$\mathrm{P}[Z > \theta\mathbb{E}[Z]] \geq (1-\theta)^2\frac{\mathbb{E}[Z]^2}{\mathrm{Var}[Z] + (1-\theta)^2\mathbb{E}[Z^2]},$$

10. Let $X \geq 0$ and $\mathbb{E}[X^2] < \infty$. We apply the Cauchy-Schwarz inequality to $X1_{X>0}$ and obtain
$$\mathrm{P}[X > 0] \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

11. (Dvoretzky-Kiefer-Wolfowitz inequality). Let $X_1, \ldots, X_n$ be iid random variables (samples) with cumulative distribution F. Let F_n be the associated empirical distribution
$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \leq x},$$
Then,
$$\mathrm{P}[\sup_{x\in\mathbb{R}}(F_n(x) - F(x)) > \epsilon] \leq e^{-2n\epsilon^2},$$
for every $\epsilon \geq \sqrt{\frac{1}{2n}\ln 2}$.

12. (Chung-Erdős inequality). Let $E_1, \ldots, E_n \in \mathcal{F}$ and $\mathrm{P}[E_i] > 0$ for some i. Then
$$\mathrm{P}[E_1 \vee \ldots \vee E_n] \geq \frac{(\sum_{i=1}^n \mathrm{P}[E_i])^2}{\sum_{i=1}^n \sum_{j=1}^n \mathrm{P}[E_i \wedge E_j]}$$

1.3.3 Involving sums or averages

1. (Hoeffding's inequality for sums #1). Let $X_1, X_2, \ldots, X_n$ be independent random variables in $[0, 1]$. Define
$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$
Then,
$$\mathrm{P}[\bar{X} - \mathbb{E}[\bar{X}] \geq t] \leq e^{-2nt^2}.$$

2. (Hoeffding's inequality for sums #2). Let $X_1, X_2, \ldots, X_n$ be independent random variables and $X_i \in [a_i, b_i]$. Let $\bar{X}$ be as above and let $r_i = b_i - a_i$. Then
$$\mathrm{P}[\bar{X} - \mathbb{E}[\bar{X}] \geq t] \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n r_i^2}\right),$$
and
$$\mathrm{P}[|\bar{X} - \mathbb{E}[\bar{X}]| \geq t] \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n r_i^2}\right).$$

3. (Kolmogorov's inequality). Let X_k, $k = 1, \ldots, N$ be independent random variables on $(\Omega, \mathcal{F}, \mathrm{P})$ with mean 0 and variances σ_k^2. Let $S_k = X_1 + X_2 + \ldots + X_k$. For all $\epsilon > 0$,
$$\mathrm{P}[\max_{1\leq k\leq n} |S_k| > \epsilon] \leq \frac{1}{\epsilon^2}\sum_{k=1}^n \sigma_k^2.$$

4. (Gaussian tail inequality for averages). Let $X_1, \ldots, X_n \sim \mathcal{N}(0, 1)$ and let $\bar{X}_n := n^{-1}\sum_{i=1}^n X_i$. Then $\bar{X}_n \sim \mathcal{N}(0, n^{-1})$ and
$$\mathrm{P}[|\bar{X}_n| > \epsilon] \leq \frac{2e^{-n\epsilon^2/2}}{\sqrt{n}\epsilon}.$$

5. (Etemadi's inequality). Let $X_1, \ldots, X_n$ be independent real-valued random variables and $\alpha \geq 0$. Let $S_n = X_1 + \ldots + X_n$. Then
$$\mathrm{P}[\max_{1\leq i\leq n} |S_i| \geq 3\alpha] \leq \max_{1\leq i\leq n} \mathrm{P}[|S_i| \geq \alpha].$$

1.4 Convergence of random processes

1.4.1 Convergence of measures

1. (Strong convergence). Let $\{\mu_k\}_{k\in\mathbb{N}}$ be a sequence of measures defined on a measurable space $(\mathcal{X}, \mathcal{G})$. We say that the sequence converges strongly to a measure μ if
$$\lim_k \mu_k(A) = \mu(A),$$
for all $A \in \mathcal{G}$.

2. (Total variation convergence). The total variation distance between two measures μ and ν on a measurable space $(\mathcal{X}, \mathcal{G})$ is defined as
$$\begin{aligned} d_{\mathrm{TV}}(\mu, \nu) &= \|\mu - \nu\|_{\mathrm{TV}} \\ &:= \sup\left\{\int_{\mathcal{X}} f\mathrm{d}\mu - \int_{\mathcal{X}} f\mathrm{d}\nu,\ f : \mathcal{X} \to [-1, 1] \text{ measurable}\right\} \\ &= 2\sup_{A\in\mathcal{G}} |\mu(A) - \nu(A)| \end{aligned}$$
A sequence of measures $\{\mu_k\}_{k\in\mathbb{N}}$ converges in the total variation to a measure μ if $d_{\mathrm{TV}}(\mu_k(A) - \mu(A)) \to 0$ as $k \to \infty$ for all $A \in \mathcal{G}$.

3. (Weak convergence). The sequence of measures $\{\mu_k\}_{k\in\mathbb{N}}$ is said to converge in the weak sense, denoted by $\mu_k \rightharpoonup \mu$, if any of the conditions of the *Portmanteau Theorem* hold; these are

i. $\mathbb{E}_{\mu_k} f \to \mathbb{E}_\mu f$ for all bounded continuous functions f

ii. $\mathbb{E}_{\mu_k} f \to \mathbb{E}_\mu f$ for all bounded Lipschitz functions f

iii. $\limsup_k \mathbb{E}_{\mu_k} f \leq \mathbb{E}_\mu f$ for every upper semi-continuous f bounded from above

iv. $\liminf_k \mathbb{E}_{\mu_k} f \geq \mathbb{E}_\mu f$ for every lower semi-continuous f bounded from below

v. $\limsup \mu_k(C) \leq \mu(C)$ for all closed set $C \subseteq \mathcal{X}$

vi. $\liminf \mu_k(O) \geq \mu(O)$ for all open set $O \subseteq \mathcal{X}$

4. (Tightness). A sequence of measures $(\mu_n)_n$ is called *tight* if for every $\epsilon > 0$ there is a compact set K so that $\mu_n(K) > 1 - \epsilon$ for all $n \in \mathbb{N}$.

5. (Prokhorov's Theorem). If $(\mu_n)_n$ is tight, then every subsequence of it has a further subsequence which is weakly convergent.

6. (Lévy-Prokhorov distance). Let (X, d) be a metric space and let $\mathcal{B}_X$ be the Borel σ-algebra which makes $(X, \mathcal{B}_X)$ a measurable space. Let $\mathrm{P}(X)$ be the space of all probability measures on $(X, \mathcal{B}_X)$. For all $A \subseteq X$ we define

$$A^\epsilon := \{p \in X \mid \exists q \in A, d(p,q) < \epsilon\} = \bigcup_{p \in A} B_\epsilon(p),$$

where $B_\epsilon(p)$ is an open ball centered at p with radius ϵ.

The Lévy-Prokhorov distance is a mapping $\pi : \mathrm{P}(X) \times \mathrm{P}(X) \to [0,1]$ between two probability measures μ and ν defined as

$$\pi(\mu, \nu) := \inf\{\epsilon > 0 \mid \mu(A) \leq \nu(A^\epsilon) + \epsilon, \nu(A) \leq \mu(A^\epsilon) + \epsilon, \forall A \in \mathcal{B}_X\}.$$

7. (Metrizability of weak convergence). If (X, d) is a separable metric space, then convergence of a sequence of measures in the Lévy-Prokhorov distance is equivalent to weak convergence.

8. (Separability of (P_X, π)). The space (P_X, π) is separable if and only if (X, d) is separable.

9. (Skorokhod's representation theorem). Let $(\mu_n)_n$ be a sequence of probability measures on a metric measurable space $(S, \mathcal{H})$ such that $\mu_n \to \mu$ weakly. Suppose that the support of μ is separable[5]. Then, there exist random variables $(X_n)_n$ and X on a common probability space such that the distribution of X_n is μ_n, the distribution of X is μ and $X_n \to X$ almost surely.

10. (Strong $\not\Rightarrow$ TV).

1.4.2 Almost sure convergence

1. (Almost sure convergence). A sequence of random variables $(X_n)_n$ is said to converge *almost surely* if the sequence $(X_n(\omega))_n$ converges (somewhere) for almost every ω. It converges almost surely to X if $\lim_n X_n(\omega) = X(\omega)$ for almost every ω.

2. (Uniqueness almost surely). If $X_n \to X$ a.s. and $X_n \to Y$ a.s., then $X = Y$ a.s.

3. (Characterization of a.s. convergence). The sequence $(X_n)_n$ converges a.s. to X if and only if for every $\epsilon > 0$

$$\sum_{n \in \mathbb{N}} 1_{(\epsilon, \infty)} \circ |X_n - X| < \infty.$$

4. (Characterization of a.s. convergence *a là* Borel-Cantelli #1). The sequence $(X_n)_n$ converges a.s. to X if for every $\epsilon > 0$

$$\sum_{n \in \mathbb{N}} \mathrm{P}[|X_n - X| > \epsilon] < \infty.$$

[5] The support of a measure μ on $(\Omega, \mathcal{F}, \mathrm{P})$ which is equipped with a topology τ is the set of $\omega \in \Omega$ for which every open neighbourhood N_ω of ω has a positive measure: $\mathrm{supp}(\mu) = \{\omega \in \Omega : \mu(N_x) > 0, \text{ for all } N_\omega \in \tau, N_\omega \ni \omega\}$.

5. (Characterization of a.s. convergence *a là* Borel-Cantelli #2). The sequence $(X_n)_n$ converges a.s. to X if there is a decreasing sequence $(\epsilon_n)_n$ converging to 0 so that
$$\sum_{n\in\mathbb{N}} \mathrm{P}[|X_n - X| > \epsilon_n] < \infty.$$

6. (Cauchy criterion). The sequence $(X_n)_n$ is convergent almost surely if and only if $\lim_{m,n\to\infty} |X_n - X_m| \to 0$ almost surely.

7. (Kolmogorov's three-series theorem). Let $(X_n)_n$ be a sequence of independent random variables. The random series $\sum_n X_n$ converges almost surely in $\mathbb{R}$ if and only if the following conditions hold for some $\epsilon > 0$:
   i. $\sum_n \mathrm{P}[|X_n| > \epsilon]$ converges
   ii. Let $Y_n = X_n 1_{|X_n|\leq\epsilon}$. Then, $\sum_n \mathbb{E}[Y_n]$ converges
   iii. $\sum_n \operatorname{var} Y_n$ converges

8. (Consequence of Kolmogorov's three-series theorem: Random harmonic series). Let X_n be a sequence of independent random variables taking the values $\{-1, 1\}$, each with probability $1/2$. Then, $\sum_n \frac{X_n}{n}$ converges almost surely. This is proven using Kolmogorov's three-series theorem with $\epsilon = 2$.[6]

9. (Kolmogorov's two-series theorem). Let $(X_n)_n$ be a sequence of independent random variables with expected values $\mathbb{E}[X_n] = \mu_n$ and variances $\operatorname{var}[X_n] = \sigma_n^2$ such that $\sum_n \mu_n$ and $\sum_n \sigma_n^2$ converge in $\mathbb{R}$. Then, $\sum_n X_n$ converges in $\mathbb{R}$ almost surely.

10. (Continuous mapping theorem). Let $X_n \overset{a.s.}{\to} X$ and g be a (almost everywhere) continuous mapping. Then $g(X_n) \overset{a.s.}{\to} g(X)$.

11. (Topological (non) characterization). The concept of almost sure convergence does not come from a topology on the space of random variables. This means there is no topology on the space of random variables such that the almost surely convergent sequences are exactly the converging sequences with respect to that topology. In particular, there is no metric of almost sure convergence.

1.4.3 Convergence in probability

1. (Convergence in probability). We say that the stochastic process $(X_n)_n$ converges to a random variable X in probability if for every $\epsilon > 0$,
$$\lim_n \mathrm{P}[|X_n - X| > \epsilon] = 0.$$
   We denote $X_n \overset{p}{\to} X$.

2. (Continuous mapping theorem). Let $X_n \overset{p}{\to} X$ and g be a (almost everywhere) continuous mapping. Then $g(X_n) \overset{p}{\to} g(X)$.

3. (Metrizability). Convergence in probability defines a topology which is metrizable via the *Ky Fan metric*
$$d(X, Y) = \inf\{\epsilon > 0 \mid \mathrm{P}[|X - Y| > \epsilon] \leq \epsilon\} = \mathbb{E}[\min(|X - Y|, 1)].$$

4. (Metrizability #2). The sequence X_n converges to 0 in probability if and only if
$$\mathbb{E}\left[\frac{|X_n|}{1 + |X_n|}\right] \to 0.$$
   The functional
$$d(X, Y) := \mathbb{E}\left[\frac{|X - Y|}{1 + |X - Y|}\right]$$
   is a metric that induces the convergence in probability (provided we identify two random variables as equal if they are almost everywhere equal).

[6] More on random harmonic series can be found in B. Schmuland, "Random Harmonic Series," American Mathematical Monthly 110: 407–416, 2003, doi:10.2307/3647827. There the author explains that the pdf of $\sum_n \frac{X_n}{n}$ evaluated at 2 differs from $1/8$ by less than 10^{-42}.

5. (Almost surely convergent subsequence). If $X_n \xrightarrow{p} X$, then there exists a subsequence of $(X_n)_n$, $(X_{k_n})_n$ which converges almost surely to X.

6. (Sum of independent variables). Let $(X_n)_n$ be a sequence of independent random variables and let $(S_n)_n$ be a sequence defined as $S_n = X_1 + \ldots + X_n$. Then S_n converges almost surely if and only if it converges in probability.

7. (Convergence of pairs). If $X_n \to X$ in probability and $Y_n \to Y$ in probability, then $(X_n, Y_n) \to (X, Y)$ in probability.

8. (Almost surely $\Rightarrow$ in probability). If a sequence of random variables $\{X_k\}_k$ converges almost surely, it converges in probability to the same limit.

9. (In probability $\not\Rightarrow$ almost surely). There are sequences which converge in probability but not almost surely. Here is an example: Let $(X_n)_n$ be a sequence of independent random variables on $\Omega = \mathbb{N}$ with $X_n = 1$ with probability $1/n$ and 0 with probability $1 - 1/n$. Then, for any $\epsilon > 0$ it is $\mathrm{P}[|X_n| > \epsilon] = \frac{1}{n} \to 0$, but by the second Borel-Cantelli lemma since $\sum_{n=1}^{\infty} \mathrm{P}[|X_n| > \epsilon]$ (and the events $\{|X_n| > \epsilon\}$ are independent), we have $\mathrm{P}[\limsup_n \{|X_n| > \epsilon\}] = 1$.

1.4.4 Convergence in $\mathcal{L}^p$

1. (Convergence in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$). We say that X_k converges to X in $\mathcal{L}^p$ if $X, X_k \in \mathcal{L}^p$ for all $k \in \mathbb{N}$ and $\|X_k - X\|_p \to 0$.

2. (Convergence $\mathcal{L}_1$ under uniform integrability). If $X_n \to X$ in probability and $(X_n)_n$ is uniformly integrable, then $X_n \to X$ in $\mathcal{L}_1$.

3. (In $\mathcal{L}_s(\Omega, \mathcal{F}, \mathrm{P}) \Rightarrow$ in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$, for $s > p \geq 1$).

4. (Scheffé's theorem). Let $X_n \in \mathcal{L}_1$, $X \in \mathcal{L}_1$ and $X_n \to X$ almost surely. The following are equivalent:
   i. $\mathbb{E}[|X_n|] \to \mathbb{E}[|X|]$,
   ii. $\mathbb{E}[|X_n - X|] \to 0$.

5. (Convergence in $\mathcal{L}^p$ for all $p \in [1, \infty)$ but not in $\mathcal{L}_\infty$). Let X be a random variable on $\Omega = \mathbb{N}$ which follows the Poisson distribution ($\mathrm{P}[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}$, $\lambda > 0$). Define the sequence $X_k = 1_{\{X=k\}}$. Then $\|X_k\|_\infty = 1$.

6. (Vitali's theorem). Suppose that $X_n \in \mathcal{L}^p$, $p \in [1, \infty)$ and $X_n \to X$ in probability. The following are equivalent
   i. $\{X_n\}_n$ is uniformly integrable
   ii. $X_n \to X$ in $\mathcal{L}^p$
   iii. $\mathbb{E}[|X_n|^p] \to \mathbb{E}[|X|^p]$

7. (In $\mathcal{L}^p \Rightarrow$ in probability). If $(X_n)_n$ converges to X in $\mathcal{L}^p$, for any $p \in [1, \infty]$ it also converges to X in probability.

8. (Almost surely $\not\Rightarrow$ in $\mathcal{L}^p$). On $([0, 1], \mathcal{B}_{[0,1]}, \lambda)$ take $X_n = n 1_{[0,1/n]}$. Then, for all $p \in [1, \infty]$ we have $\|X_n\|_p = 1$, but the sequence converges almost surely to 0.

9. (In $\mathcal{L}^p$, $p \in [1, 2) \not\Rightarrow$ In $\mathcal{L}^p$, for $p \geq 2$). Let $\Omega = \mathbb{N}$ and Z_k^p be a sequence of random variables with parameter p and

$$\begin{aligned} \mathrm{P}[Z_k^p = n] &= pn, \\ \mathrm{P}[Z_k^p = 0] &= 1 - pn. \end{aligned}$$

   Let $X = 0$ and X_k be defined as

$$X_k = Z_k^{p_k}$$

   where $p_k = {}^1\!/_{k^2 \ln k}$. Then $\mathbb{E}[|X_k|^t] = {}^{k^{t-2}}\!/_{\ln k}$. We have $\mathbb{E}[|X_k|^t] \to 0$ if and only if $t < 2$.

$$\begin{tikzcd}
\mathcal{L}_\infty \arrow[r, Rightarrow] & \mathcal{L}_p \arrow[r, Rightarrow] & \mathcal{L}_{p'},\, p' \in [1,p) \arrow[d, Rightarrow] & & \\
& \text{a.s.} \arrow[r, Rightarrow] & \text{p} \arrow[r, Rightarrow] & \text{d} \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \varphi\,(\text{pointwise}) \\
& & & \text{a.s.}\,(\overset{d}{\sim}) &
\end{tikzcd}$$

Figure 1.1: Illustration of the relationships among different modes of convergence of random variables. Convergence in $\mathcal{L}_\infty$ implies convergence in $\mathcal{L}^p$ for all $p \in [1, \infty)$ which in turn implies convergence in $\mathcal{L}_{p'}$ for all $1 \leq p' \leq p$ which implies convergence in probability which implies convergence in distribution which implies convergence of the characteristic functions (Lévy's continuity theorem). Convergence in distribution implies almost convergence of a sequence of RVs $\{Y_k\}_k$ which have the same distribution as $\{X_k\}_k$ ($Y_k \overset{d}{\sim} X_k$ and $Y \overset{d}{\sim} X$).

1.4.5 Convergence in distribution

1. (Convergence in distribution). The sequence of random variables $\{X_n\}_n$ with distributions $\{\mu_n\}_n$ is said to converge in distribution of X if $\{\mu_n\}_n$ converges weakly to μ, the distribution of X.

2. (Slutsky's theorem). Let $X_k \to X$ in distribution and $Y_n \to c$ in probability, where c is a constant. Then,

   i. $X_n + Y_n \to X + c$ in distribution

   ii. $X_n Y_n \to cX$ in distribution

   iii. $X_n / Y_n \to X/c$ in distribution, provided that $c \neq 0$, $Y_n \neq 0$.

3. (Almost sure convergence). If $X_n \to X$ in distribution, we may find a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and random variables Y and $(Y_n)_n$ so that Y_n is equal in distribution to X_n, Y is equal in distribution to X and $Y_n \to Y$ almost surely.

4. (Lévy's continuity theorem)[7]. Let $\{X_k\}_k$ be a sequence of random variables with characteristic functions $\varphi_k(t)$ and let X be a random variable with characteristic function $\varphi(t)$. It X_k converges to X in distribution then $\varphi_k \to \varphi$ point-wise. Conversely, if $\varphi_k \to \varphi$ and φ is continuous at 0, then φ is the characteristic function of a random variable X and $X_k \to X$ in distribution. the

5. (Scheffé's theorem for density functions).[8] Let P_n and P have densities f_n and f with respect to a measure μ. If $f_n \to f$ μ-a.s., then $\mathrm{P}_n \to \mathrm{P}$ in the total variation metric and, as a result, $\mathrm{P}_n \to \mathrm{P}$ weakly.

6. (Continuous mapping theorem). For a (almost everywhere) continuous function g, if the sequence $\{X_k\}_k$ converges in distribution to X, then $\{g(X_k)\}_k$ converges in distribution to $g(X)$.

7. (Convergence in probability $\Rightarrow$ in distribution). If $\{X_k\}_k$ converges in probability, then it converges in distribution to the same limit.

8. (In distribution $\not\Rightarrow$ in probability). There are sequences which converge in distribution, but not in probability. For example: On the space $([0,1], \mathcal{B}_{[0,1]}, \lambda)$, let $X_{2n}(\omega) = \omega$ and $X_{2n-1}(\omega) = 1 - \omega$. Then all X_k have the same distribution, but the sequence does not converge in probability. As a second example, the sequence $X_n = X$ where X follows the Bernoulli distribution with parameter $\frac{1}{2}$, converges in distribution to $1 - X$, but not in probability.

9. (Polya-Cantelli lemma)[9]. If $X_n \to X$ in distribution, F_n are the distribution function of X_n and X has the *continuous* distribution function F, then $\|F_n - F\|_\infty := \sup_x |F_n(x) - F(x)| \to 0$ as $n \to \infty$.

[7] Lecture notes 6.436J/15.085J by MIT, Available online at https://goo.gl/7ZaHW9.

[8] S. Sagitov, "Weak Convergence of Probability Measures," 2013, Available at: https://goo.gl/m4Qi5i

[9] Lecture notes of M. Banerjee, Available at: http://dept.stat.lsa.umich.edu/~moulib/ch2.pdf

10. (Delta method). Let X be a real-valued random variable and X_n be a sequence of real-valued random variables with $n^c(X_n - \theta) \to X$, in distribution for some $c > 0$. Let $g : \mathbb{R} \to \mathbb{R}$ be function which is differentiable at θ. Then, $n^c(g(X_n) - g(\theta)) \to g'(\theta)X$.[10]

1.4.6 Tail events and 0-1 Laws

1. (Simple 0-1 law). Let $\{E_n\}$ be a sequence of independent events. Then $\mathrm{P}[\limsup_n E_n] \in \{0, 1\}$.

2. (Unions of σ-algebras). Let $\mathcal{F}_1$, $\mathcal{F}_2$ be two σ-algebras on a nonempty set X. The σ-algebra generated by the sets $E_1 \cup E_2$ with $E_1 \in \mathcal{F}_1$ and $E_2 \in \mathcal{F}_2$ is denoted by $\mathcal{F}_1 \vee \mathcal{F}_2$

3. (Tail σ-algebra). Let $(\mathcal{F}_n)_n$ be a sequence of sub-σ-algebras of $\mathcal{F}$. The σ-algebra $T_n := \bigvee_{m>n} \mathcal{F}_m$ encodes the information about the future after n and $T = \bigcap_n T_n$ is the *tail σ-algebra* which encodes the information of the end of time.

4. (Events in the tail σ-algebra). For a process $(E_n)_n$ be a sequence of events. The associated tail σ-algebra T is $\bigcap_n \sigma(\{E_k\}_{k \geq n})$. The event $\limsup_n E_n$ is in T.

5. (Kolmogorov's zero-one law). Let $(\mathcal{F}_n)_n$ be a sequence of *independent* σ-algebras on a nonempty set X and let T be the tail σ-algebra. We equip $(X, \mathcal{F})$ with a probability measure P. For every $H \in T$, $\mathrm{P}(H) \in \{0, 1\}$.

6. (Counterpart of the Borel-Cantelli lemma). Let $\{E_n\}_{n \in \mathbb{N}}$ be a nested increasing sequence of events in $(\Omega, \mathcal{F}, \mathrm{P})$, that is $E_k \subseteq E_{k+1}$ and let E_k^c denote the complement of E_k. Infinitely many E_k occur with probability 1 if and only if there is an increasing sequence $t_k \in \mathbb{N}$ such that

$$\sum_k \mathrm{P}[A_{t_{k+1}} \mid A_{t_k}^c] = \infty.$$

7. (Lévy's zero-one law). Let $\mathfrak{F} = \{\mathcal{F}_k\}_{k \in \mathbb{N}}$ be any filtration of $\mathcal{F}$ on $(\Omega, \mathcal{F}, \mathrm{P})$ and $X \in \mathcal{L}_1(\Omega, \mathcal{F}, \mathrm{P})$. Let $\mathcal{F}_\infty$ be the minimum σ-algebra generated by $\mathfrak{F}$. Then

$$\mathbb{E}[X \mid \mathcal{F}_k] \to \mathbb{E}[X \mid \mathcal{F}_\infty],$$

both in $\mathcal{L}_1(\Omega, \mathcal{F}, \mathrm{P})$ and P-a.s.

1.4.7 Laws of large numbers and CLTs

1. (Weak law of large numbers). Also known as Bernoulli's theorem. Let $\{X_k\}_k$ be a sequence of independent identically distributed random variables, each having a finite mean $\mathbb{E}[X_k] = \mu$ and finite variance σ^2. Define $\bar{X}_k = {}^1\!/_k(X_1 + \ldots + X_k)$. Then $\bar{X}_k \to \mu$ in probability.

2. (Strong law of large numbers). Let $\{X_k\}_k$ and $\bar{X}_k$ be as above. Then $\bar{X}_k \to \mu$ almost surely.

3. (Uniform law of large numbers). Let $f(x, \theta)$ be a function defined over $\theta \in \Theta$. For fixed θ and a random process $\{X_k\}_k$ define $Z_k^\theta := f(X_k, \theta)$. Let $\{Z_k^\theta\}_k$ be a sequence of independent and identically distributed random variables, such that the sample mean converges in probability to $\mathbb{E}[f(X, \theta)]$. Suppose that (i) Θ is compact, (ii) f is continuous in θ for almost all x and measurable with respect to x for each θ, (iii) there is a function g such that $\mathbb{E}[g(X)] < \infty$ and $\|f(x, \theta)\| \leq g(x)$ for all $\theta \in \Theta$. Then, $\mathbb{E}[f(X, \theta)]$ is continuous in θ and

$$\sup_{\theta \in \Theta} \left\| \bar{Z}_k^\theta - \mathbb{E}[f(X, \theta)] \right\| \xrightarrow{a.s.} 0$$

4. (Lindeberg-Lévy central limit theorem). Let $\{X_k\}_k$ be iid, finite mean and variance and $\bar{X}_k$ as above. Then

$$\frac{\bar{X}_k - \mu}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

where $\mathcal{N}(0, \sigma^2)$ is the normal distribution is zero mean and variance σ^2 (See Section 1.5.2).

[10] See the lecture notes at http://personal.psu.edu/drh20/asymp/fall2006/lectures/, Chap. 5. The proof makes use of Taylor's first-order expansion and Slutsky's theorem.

5. (Lyapunov central limit theorem). Let $\{X_k\}_k$ be a sequence of independent random variables with $\mathbb{E}[X_k] = \mu_k$ and finite variance σ_k^2. Define $s_k^2 = \sum_{i=1}^k \sigma_i^2$. If for some $\delta > 0$, the following condition holds (Lyapunov's condition)[11]:

$$\lim_{k\to\infty} \frac{1}{s_k^{2+\delta}} \sum_{i=1}^{k} \mathbb{E}\left[|X_i - \mu_i|^{2+\delta}\right] = 0,$$

then,

$$\frac{1}{s_k} \sum_{i=1}^{k} (X_i - \mu_i) \xrightarrow{d} \mathcal{N}(0,1).$$

6. (Law od iterated logarithm). Let $(X_k)_{t\in\mathbb{N}}$ be independent identically distributed random variables and let $S_k := X_1 + \ldots + X_k$. Then,

$$\limsup_k \frac{S_k}{\sqrt{2k \ln\ln k}} = 1,$$

and convergence is in the almost sure sense.

7. (Delta method). Let $(X_n)_n$ be a sequence of random variables satisfying

$$\sqrt{n}(X_n - \theta) \overset{d}{\to} \mathcal{N}(0, \sigma^2).$$

Let $g : \mathbb{R} \to \mathbb{R}$ be a function which is differentiable at θ and $g'(\theta) \neq 0$. Then,

$$\sqrt{n}(g(X_n) - g(\theta)) \overset{d}{\to} \mathcal{N}(0, \sigma^2[g'(\theta)]^2)$$

8. (Glivenko-Cantelli Theorem). Let $X_1, X_2, \ldots, X_N$ be independent and identically distributed real-valued random variables with a common cumulative distribution function $F(x)$. The *empirical distribution* of $X_1, X_2, \ldots, X_N$ is

$$F_N(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{[X_i,+\infty)}(x).$$

Then,

$$\|F_N - F\|_\infty := \sum_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0,$$

almost surely as $N \to \infty$[12].

9. (Dvoretzky–Kiefer–Wolfowitz inequality). This inequality is a stronger result compared to the Glivenko-Cantelli Theorem above as it gives the rate of convergence of the empirical distribution to the actual one. Let $X_1, \ldots, X_N$ and F, F_N be as above. Then,

$$\mathrm{P}[\|F_N - F\|_\infty > \epsilon] \leq e^{-2n\epsilon^2},$$

for $\epsilon \geq \sqrt{\frac{1}{2n} \ln 2}$ and

$$\mathrm{P}[\|F_N - F\|_\infty > \epsilon] \leq 2e^{-2n\epsilon^2},$$

for $\epsilon > 0$.

[11] In practice it is usually easiest to check Lyapunov's condition for $\delta = 1$. If a sequence of random variables satisfies Lyapunov's condition, then it also satisfies Lindeberg's condition. The converse implication, however, does not hold.

[12] The almost sure pointwise convergence of F_N to F follows from the strong law of large numbers. This is, therefore, a stronger result.

1.5 Standard Distributions

1.5.1 Uniform distribution

1. (Definition). A random variable $X : \Omega \to [a, b]$, $a < b$, is said to follow the uniform distribution, denoted as $X \sim U(a, b)$, if its cumulative distribution function is

$$F_X(x) = \mathrm{P}[X \leq x] = \begin{cases} 0, & \text{for } x < a \\ {}^{x-a}\!/_{b-a}, & \text{for } a \leq x < b \\ 1, & \text{for } x \geq b \end{cases}$$

The probability density function of X is

$$f_X(x) = \begin{cases} {}^{1}\!/_{b-a}, & \text{for } x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

The distribution $U(0, 1)$ is called the *standard uniform distribution.*

2. (Characteristics). For $X \sim U(a, b)$, the expectation of X is $\mathbb{E}[X] = \frac{a+b}{2}$, and its variance is $\mathrm{Var}[X] = {}^{1}\!/_{12}(b - a)^2$.

3. (Probability integral transform). Let X be a real-valued random variable which has a continuous distribution with cumulative distribution function F_X. Then, the random variable $Y = F_X(X)$, that is, $Y(\omega) = F_X(X(\omega))$ follows the standard uniform distribution.

4. (Inverse probability integral transform). Let $Y \sim U(0, 1)$ and let X be a random variable with cumulative distribution function F_X. Then $F_X^{-1}(Y)$ has the same distribution as X.

1.5.2 Normal distribution

1. (Definition: univariate case). For a real-valued random variable X, the PDF f_X of the normal distribution on $\mathbb{R}$ with parameters μ and σ is given by

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

and cumulative distribution function

$$F_X(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right).$$

We write $X \sim \mathcal{N}(\mu, \sigma)$ to denote that X follows the normal distribution with parameters μ and σ.

2. (Characteristics). The mean, median and mode of $\mathcal{N}(\mu, \sigma)$ are equal to μ. Its variance is σ^2 and its MGF is $M_X(z) = \exp(\mu z + \sigma^2 z^2/2)$

3. (Sum and scalar product of normals). Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2$ be two independent random variables. Then $X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$. If X and Y are jointly normally distributed random variables, then $X + Y$ is normally distributed and $E[X + Y] = \mu_X + \mu_Y$ and $\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y}$, where ρ is the correlation between X and Y. For any $\alpha \in \mathbb{R}$, $\alpha X \sim \mathcal{N}(\alpha\mu_X, \alpha^2\sigma_X^2)$.

4. (Multivariate normal distribution). The multivariate variant of the normal distribution, denoted by $\mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n\times n}$, symmetric positive semi-definite, is supported on $\mu + \mathrm{im}(\Sigma)$ and PDF

$$p_X(x) = |2\pi\Sigma|^{-1/2} \exp\left(-\tfrac{1}{2}(x-\mu)^\top \Sigma (x-\mu)\right).$$

5. (Isserlis' theorem – high-order moments of multivariate normal). Let $X = (X_1, \ldots, X_{2n})$ follow the multivariate normal distribution with zero mean and covariances $\Sigma_{i,j} = \mathrm{cov}(X_i, X_j)$. Then,

$$\mathbb{E}[X_1 X_2 \cdots X_{2n}] = \sum \prod \mathbb{E}[X_i X_j] = \sum \prod \mathrm{cov}(X_i, X_j),$$

and

$$\mathbb{E}[X_1 X_2 \cdots X_{2n-1}] = 0.$$

6. (Linear transformation). Let $X \sim \mathcal{N}(\mu, \Sigma)$, $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n\times n}$ and $Y = AX + b$ for constant A and b. Then $Y \sim \mathcal{N}(A\mu + b, A\Sigma A')$.
7. (Conditioning). Let X_1, X_2 be two random variables with values in $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$ respectively. Suppose that
$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right).$$
Then, $\mathbb{E}[X_1 \mid X_2 = x_2] \sim \mathcal{N}(\mu(x_2), \Sigma)$ with
$$\begin{aligned} \mu(x_2) &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \\ \Sigma &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$
8. (χ^2 distribution). If X is an n-dimensional random variable and $X \sim \mathcal{N}(\mu, \Sigma)$, then
$$(X - \mu)^\top \Sigma^{-1}(X - \mu)$$
follows the $\chi^2(n)$ distribution.

1.5.3 Binomial distribution

1. (Definition). For $n \in \mathbb{N}$ and $p \in [0, 1]$, we say that a random variable $X : \Omega \to \{0, \ldots, n\}$ follows the Binomial distribution, we denote $X \sim B(n, p)$, if its probability mass function is
$$\mathrm{P}[X = k] = \binom{n}{k} p^k (1 - p)^{n-k},$$
where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
2. (Characteristics). If $X \sim B(n, p)$, then $\mathbb{E}[X] = np$, the median of X is either $\lfloor np \rfloor$ or $\lceil np \rceil$, its variance is $\mathrm{Var}[X] = np(1 - p)$ and moment generating function (MGF) of X is $M_X(z) = (1 - p + pe^z)^n$. The cumulative probability function of X is given in terms of the regularized incomplete beta function
$$\mathrm{P}[X \leq x] = I_{1-p}(n - k, k + 1)$$
3. (Sum of independent Binomials). Let $X \sim B(n, p)$ and $Y \sim B(m, p)$ be independent random variables. Then, $X + Y$ follows the Binomial distribution $B(n + m, p)$.
4. (Binomial sum variance inequality). Let $X \sim B(n, p)$ and $Y \sim B(n', p')$ be two random variables, not necessarily independent. Let $S = X + Y$. Then,
$$\mathrm{Var}[S] \leq \mathbb{E}[S]\frac{1 - \mathbb{E}[S]}{n + n'}.$$
5. (De Moivre-Laplace theorem). For large n and k in the neighbourhood of np, it is
$$\binom{n}{k} p^k (1 - p)^{n-k} \approxeq \frac{1}{\sqrt{2\pi np(1 - p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}.$$
As an example, consider the experiment of tossing n coins a large number of times and observing the number of heads each time. Then, as n grows large, the shape of the distribution of the number of heads approaches that of the normal distribution.

1.5.4 Poisson distribution

1. (Definition). We say that a random variable $X : \Omega \to \mathbb{N}$ follows the Poisson distribution with parameter λ if its probability mass function is
$$\mathrm{P}[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}$$
We denote $X \sim \mathrm{Poisson}(\lambda)$.

2. (Characteristics). If $X \sim \mathrm{Poisson}(\lambda)$, then $\mathbb{E}[X] = \lambda$, $\mathrm{Var}[X] = \lambda$ and the moment generating function (MGF) of X is $M_X(z) = \exp(\lambda(e^z - 1))$. The median of X is in the interval $[\lambda - \ln 2, \lambda + {}^1\!/_3$.

3. (Sum of independent Poissons). Let X_i, $i \in \mathbb{N}_{[1,N]}$, be independent random variables with $X_i \sim \mathrm{Poisson}(\lambda_i)$. Define the random variable $S = X_1 + \ldots + X_N$ and $\lambda = \lambda_1 + \ldots + \lambda_N$, then $S \sim \mathrm{Poisson}(\lambda)$.

4. (Raikov's theorem). If the sum of two independent nonnegative random variables, X and Y, follows the Poisson distribution, then both X and Y follow the Poisson distribution.

5. (Law of rare events/Poisson limit theorem). Let $(p_n)_{n \in \mathbb{N}}$ be a sequence of numbers in $[0, 1]$ such that np_n converges to a limit λ. Then

$$\underbrace{\lim_{n \to \infty} \binom{n}{k} p_n^k (1 - p_n)^{n-k}}_{\text{Binomial PMF}} = \underbrace{e^{-\lambda} \frac{\lambda^k}{k!}}_{\text{Poisson PMF}}.$$

6. (Large λ). For large values of λ, e.g., $\lambda > 10^3$, the normal distribution $\mathcal{N}(\lambda, \lambda)$, is considered a good approximation to $\mathrm{Poisson}(\lambda)$.

2 Multivariate distributions

2.1 Multivariate random variables

1. (Multivariate CDF). The cdf of a random variable $X : (\Omega, \mathcal{F}, \mathrm{P}) \to \mathbb{R}^d$ is the function

$$\begin{aligned} F_X(x_1, \ldots, x_d) &= \mathrm{P}[X_1 \leq x_1, \ldots, X_d \leq x_d] \\ &= \mathrm{P}\left[\bigcap_{i=1,\ldots,d} \{X_i \leq x_i\}\right] \end{aligned}$$

2. (Multivariate CDF properties). The cdf F_X of a random variable $X : (\Omega, \mathcal{F}, \mathrm{P}) \to \mathbb{R}^d$ has the following properties
   i. It is monotonically decreasing with respect to each variable
   ii. It is right-continuous with respect to each variable
   iii. $0 \leq F_X(x_1, \ldots, x_d) \leq 1$ for all $x_1, \ldots, x_d \in \mathbb{R}$
   iv. $\lim_{x_1,\ldots,x_d \to \infty} F_X(x_1, \ldots, x_d) = 1$
   v. $\lim_{x_i \to -\infty} F_X(x_1, \ldots, x_d) = 0$ for all $i \in \mathbb{N}_{[1,d]}$

3. (Expectation and variance). The expectation of a random variable $X : (\Omega, \mathcal{F}, \mathrm{P}) \to \mathbb{R}^d$, $X = (X_1, X_2, \ldots, X_d)$ is defined as

$$\mathbb{E}[X] = (\mathbb{E}[X_1], \mathbb{E}[X_2], \ldots, \mathbb{E}[X_d])$$

   And the *variance-covariance matrix* of X is

$$\mathrm{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top],$$

   which is the multivariate counterpart of variance.

2.2 Copulas

2.2.1 Sklar's theorem

1. (Definition). Let X be an d-dimensional random variable, with $X(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_d(\omega)$ with continuous marginal CDFs $F_{X_i}(x) = \mathrm{P}[X_i \leq x]$. By the probability integral transform, the random variable $U = (U_1, \ldots, U_d)$ defined as

$$U_i = F_{X_i}(X_i),$$

   for $i = 1, \ldots, d$, has uniformly distributed marginals, that is, $U_i \sim U(0,1)$. The *copula* of X is the joint cumulative distribution function of U, that is

$$\begin{aligned} C(u_1, \ldots, u_d) &= \mathrm{P}[U_1 \leq u_1, \ldots, U_d \leq u_d] \\ &= \mathrm{P}[X_1 \leq F_{X_1}^{-1}(u_1), \ldots, X_d \leq F_{X_d}^{-1}(u_d)]. \end{aligned}$$

   A d-dimensional copula is a function $C : [0,1]^d \to [0,1]$ which is a joint cumulative distribution function of a d-dimensional random variable on the $[0,1]^d$ with uniform marginals.

2. (Sklar's theorem). Every multivariate CDF, $H(x_1, \ldots, x_d) = \mathrm{P}[X_1 \leq x_1, \ldots, X_d \leq x_d]$, can be expressed in terms of its marginals, $F_{X_i}(x) = \mathrm{P}[X_i \leq x]$, and a copula $C : [0,1]^d \to [0,1]$, that is

$$H(x_1, \ldots, x_d) = C(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)).$$

If the multivariate distribution has a PDF h, then there is a function c called the density copula and

$$h(x_1, \ldots, x_d) = c(F_{X_1}(x_1), \ldots, F_{X_d}(x_d))f_1(x_1) \cdot \ldots \cdot f_d(x_d).$$

Conversely, given a copula $C : [0,1]^d \to [0,1]$ and marginal distributions F_{X_i}, there is a d-dimensional CDF as described above.

3. (Characterization). A function $C : [0,1]^d \to [0,1]$ is a copula if and only if it satisfies the following properties
   i. For every $j \in \mathbb{N}_{[1,d]}$, $C(1, \ldots, 1, t, 1, \ldots, 1) = t$
   ii. C is isotonic (order preserving), that is, $C(u) \leq C(u')$ whenever $u \leq u'$ in the sense $u_i \leq u_i'$ for all $i \in \mathbb{N}_{[0,d]}$
   iii. C is d-nondecreasing, that is, for every hyperrectange B, the dC-volume B is nonnegative, that is
   $$\int_B \mathrm{d}C \geq 0,$$
   where dC is treated as a measure.

4. (Characterization of two-dimensional copulas). A two-dimensional copula $C : [0,1]^2 \to [0,1]$ satisfies the following properties
   i for every $u \in [0,1]$, $C(u,0) = C(0,u) = 0$
   ii for every $u \in [0,1]$, $C(u,1) = C(1,u) = u$
   iii for all $u, u', v, v' \in [0,1]$ with $u \leq u'$ and $v \leq v'$
   $$C(u',v') - C(u',v) - C(u,v') + C(u,v) \geq 0$$

5. (Properties of copulas). A copula $C : [0,1]^d \to [0,1]$ possesses the following properties
   i. $C(u_1, \ldots, u_d) = 0$ if there is an $i_0 \in \mathbb{N}_{[1,d]}$ so that $u_{i_0} = 0$
   ii. C is nonexpansive in the following sense
   $$|C(u) - C(v)| \leq \sum_{i=1}^{d} |u_i - v_i|$$

6. (Fréchet-Hoeffding copula bounds). For any copula $C : [0,1]^d \to [0,1]$,
   $$W(u_1, \ldots, u_d) \leq C(u_1, \ldots, u_d) \leq M(u_1, \ldots, u_d),$$
   where
   $$W(u_1, \ldots, u_d) := \max\{0, 1 - d + \textstyle\sum_{i=1}^{d} u_i\},$$
   and
   $$M(u_1, \ldots, u_d) = \min\{u_1, \ldots, u_d\}.$$
   The upper bound is sharp, M is always a copula and equality is attained for comonotone random variables.

2.2.2 Examples of copulas

1. (Independence copula). The independence copula $\Pi_d(u_1, \ldots, u_d) = u_1 u_2 \cdots u_d$, which is associated with random variables with independent marginals and uniformly distributed.

2. (Comonotonicity copula). The copula $M_d(u_1, \ldots, u_d) = \min\{u_1, \ldots, u_d\}$ which is associated with a random variable $U = (U_1, \ldots, U_d)$ where U_i are uniformly distributed and $U_1 = U_2 = \ldots = U_d$ almost surely.

3. (Counter-monotonicity copula in 2D). We define the copula $W_2(u_1, u_2) = \max\{u_1 + u_2 - 1, 0\}$, which is associated with a $U = (U_1, U_2)$ where U_i are uniformly distributed on I and $U_1 = 1 - U_2$ almost surely.

4. (Fréchet-Hoeffding bounds). For every d-dimensional copula C_d and $u \in [0,1]^d$, we have

$$W_d(u) \leq C_d(u) \leq M_d(u),$$

where W_d is the d-dimensional variant of the counter-monotonicity copula, W_2 shown above. This is defined as

$$W_d(u) = \max\left\{\sum_{i=1}^{d} u_i - d + 1, 0\right\}.$$

For $d > 2$, W_d is not a copula. The above bounds are tight, that is,

$$\inf_{C:d\text{-dim. copula}} C(u) = W_d(u) \qquad \sup_{C:d\text{-dim. copula}} C(u) = M_d(u)$$

5. (Convex combinations of copulas). Suppose that

   i. U and U' are two d-dimensional random variables on $(\Omega, \mathcal{F}, \mathrm{P})$, distributed with copulas C and C' respectively.

   ii. Z is a random variable which follows the Bernoulli distribution with $\mathrm{P}[Z = 1] = \alpha$ and $\mathrm{P}[Z = 2] = 1 - \alpha$.

   iii. We define two functions $\sigma, \sigma' \mathbb{R} \to \{0, 1\}$. Define $\sigma(x) = \delta_1(x)$, that is, $\sigma(1) = 1$ and $\sigma(x) = 0$ for $x \neq 1$. Similarly, define $\sigma'(x) = \delta_2(x)$.

   Then, the random variable

$$\bar{U} = \sigma(Z)U + \sigma' U',$$

   is distributed according to the copula

$$\bar{C} = \alpha C + (1 - \alpha)C'.$$

6. (Fréchet-Mardia family of copulas). Define the d-dimensional Fréchet-Mardia copula as

$$C_d^{\mathrm{FM}} = \lambda \Pi_d + (1 - \lambda) M_d,$$

   for $\lambda \in [0, 1]$.

3 Stochastic Processes

3.1 General

1. (Stochastic process). Let $\mathbf{T} \subseteq \overline{\mathbb{R}}$ (e.g., $T = \mathbb{N}$ or $T = \overline{\mathbb{R}}$). A random process is a sequence/net $(X_n)_{n\in\mathbf{T}}$ of (real-valued) random variables on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$.

2. (Version). Let $T = [0, \infty)$ be a time index set and $(X_t)_t$, $(Y_t)_t$ be two stochastic processes on $(\Omega, \mathcal{F}, \mathrm{P})$. We say that $(X_t)_t$ is a version of $(Y_t)_t$ if
$$X_t = Y_t, \text{P-a.s. for all } t \in T,$$
that is, $\mathrm{P}[\{\omega \mid X_t(\omega) = Y_t(\omega)\}] = 1$ for all $t \in T$.

3. (Centered). Let $(X_t)_t$ be a real-valued stochastic process with $t \in [a, b]$. We say that $(X_t)_t$ is centered if $\mathbb{E}[X_t] = 0$ for all $t \in [a, b]$.

4. (Mean-square continuous). Let $(X_t)_t$ be a real-valued stochastic process with $t \in [a, b]$. We say that $(X_t)_t$ is mean-square continuous if
$$\lim_{\epsilon\to 0} \mathbb{E}[(X_{t+\epsilon} - X_t)^2] = 0,$$
for all $t \in [a, b]$.

5. (Auto-correlation function). Let $(X_t)_{t\in T}$ be a stochastic process. Define the function $R_X : T \times T \to \mathbb{R}$ as
$$R_X(s, t) = \mathbb{E}[X_s X_t].$$
This function is called the auto-correlation function of $(X_t)_t$.

6. (Mean-square continuity criterion). A stochastic process $(X_t)_{t\in[a,b]}$ is mean-square continuous if and only if its auto-correlation function, R_X, is continuous on $[a, b] \times [a, b]$.

7. (Filtrations). A filtration is an increasing sequence of sub-σ-algebras of $\mathcal{F}$. The space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbf{T}}, \mathrm{P})$ is called a filtered probability space. The filtration $\mathcal{F}_t = \sigma(\{X_s; s \in \mathbf{T}, s \leq t\})$ is called the filtration *generated by* $(X_n)_{n\in\mathbf{T}}$. We say that $(X_n)_n$ is adapted to a filtration $(\mathcal{F}_n)_n$ if for all $n \in \mathbf{T}$, X_n is $\mathcal{F}_n$-measurable.

8. (Stopping times). Let $(\mathcal{F}_n)_n$ be a filtration on $(\Omega, \mathcal{F}, \mathrm{P})$ and define $\overline{\mathbf{T}} := \mathbf{T} \cup \{+\infty\}$. A random variable $T : \Omega \to \overline{\mathbf{T}}$ is called a stopping time if
$$\{\omega \mid T(\omega) \leq t\} \in \mathcal{F}_t,$$
for all $t \in \mathbf{T}$. This is equivalent to requiring that the process $Z_t = 1_{T\leq t}$ is adapted to $(\mathcal{F}_t)_{t\in\mathbf{T}}$.

9. (Wald's first identity)[1]. Let $(X_k)_{k\in\mathbb{N}}$ be a sequence of iid random variables with common finite mean, $\mathbb{E}[|X_i|] < \infty$. Let τ be a stopping time with $\mathbb{E}[\tau] < \infty$. Then,
$$\mathbb{E}[X_1 + \ldots + X_\tau] = \mathbb{E}[\tau]\mathbb{E}[X_1].$$

10. (Wald's second identity). Let $(X_k)_{k\in\mathbb{N}}$ be a sequence of iid random variables with zero mean and common finite variance $\sigma^2 = \mathbb{E}[X_i^2] < \infty$. Let τ be a stopping time with $\mathbb{E}[\tau] < \infty$. Then,
$$\mathbb{E}[(X_1 + \ldots + X_\tau)^2] = \sigma^2\mathbb{E}[\tau].$$

[1] Details and proofs for the three identities of Wald can be found in the lecture notes of S. Lalley (Statistics 381).

11. (Wald's third identity). Let $(X_k)_{k\in\mathbb{N}}$ be a sequence of nonnegative iid random variables with mean $\mathbb{E}[X_k] = 1$. Let τ be a bounded stopping time with $\mathbb{E}[\tau] < \infty$. Then,

$$\mathbb{E}\prod_{i=1}^{T} X_i = 1.$$

12. (A useful property). For any stochastic process $(X_n)_{n\in\mathbb{N}}$, we have

$$\mathrm{P}\left(\max_{i\leq k}|X_i| > \epsilon\right) = \mathrm{P}\left(\sum_{i=0}^{k} X_i^2 \cdot 1_{\{|X_i|>\epsilon\}} > \epsilon^2\right).$$

13. (Kolmogorov's continuity theorem). Let $(X_t)_t$ be an $\mathbb{R}^n$-valued stochastic process on $(\Omega, \mathcal{F}, \mathrm{P})$. Suppose that $(X_t)_t$ is such that for all $t > 0$ there are positive constants α, β, L such that

$$\mathbb{E}[\|X_\tau - X_{\tau'}\|^\alpha] \leq L|\tau' - \tau|^{1+\beta},$$

for $\tau \geq 0$ and $\tau' \leq t$. Then, there is a continuous version of X.

14. (càdlàg function). Let M, d be a metric space and $E \subseteq \mathbb{R}$. A function $f : E \to M$ is called a càdlàg (continue à droite, limite à gauche) function if for every $t \in E$,

    i. $\lim_{s\to t^-} f(s)$ exists

    ii. $\lim_{s\to t^+} f(s)$ exists and is equal to $f(t)$,

    that is, f is right-continuous and the limit from the left exists.

3.2 Martingales

1. (Martingale — discrete time). A random process $(X_n)_n$ is called a *martingale* if $\mathbb{E}[|X_n|] < \infty$ and $\mathbb{E}[X_{n+1} \mid X_1, \ldots, X_n] = X_n$.

2. (Martingale — continuous time). A random process $(X_t)_{t\geq 0}$ on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathrm{P})$ is called a *martingale* if (i) it is adapted to $(\mathcal{F}_t)_{t\geq 0}$, (ii) for every $t \geq 0$, $\mathbb{E}[|X_t|] < \infty$, (iii) for all $s, t \geq 0$ with $s < t$ and all $F \in \mathcal{F}_s$, $\mathbb{E}[1_F(Y_t - Y_s)] = 0$, or, equivalently, $Y_s = \mathbb{E}[Y_t \mid \mathcal{F}_s]$.

3. (Martingale examples). The following are common examples of martingales:

    a) Let $(X_n)_n$ be a sequence of iid random variables with mean $\mathbb{E}[X_n] = \mu$. Then $Y_n = \sum_{i=1}^{n}(X_i - \mu)$ is a martingale.

    b) Let $(X_n)_n$ be a sequence of iid random variables with mean $\mathbb{E}[X_n] = 1$ and finite variance. Define a sequence of random variables with $Y_0 = 0$ and $Y_n = X_0X_1 \cdot \ldots \cdot X_n$. Then, by the Cauchy-Schwartz inequality, Y_n is a martingale.

    c) If $(X_n)_n$ is a sequence of iid random variables with mean 1, then $Y_n = \prod_{i=1}^{n} X_i$ is a martingale.

    d) If $(X_n)_n$ is a sequence of random variables with finite expectation and $\mathbb{E}[X_n \mid X_1, \ldots, X_{n-1}] = 0$, then $Y_n = \sum_{i=0}^{n} X_i$ is a martingale.

    e) (The classical martingale). The fortune of a gambler is a martingale in a fair game.

4. (Sub- and super-martingales). A random process $(X_n)_n$ is called a *super-martingale* if $\mathbb{E}[|X_n|] < \infty$ and $\mathbb{E}[X_{n+1} \mid X_1, \ldots, X_n] \leq X_n$. Likewise, it is a *sub-martingale* if $\mathbb{E}[|X_n|] < \infty$ and $\mathbb{E}[X_{n+1} \mid X_1, \ldots, X_n] \geq X_n$.

5. (Stopping time). Let $\{Z_k\}_k$ be a random process and T a stopping time. Define $X_k(\omega) = Z_{k\wedge T(\omega)}$, that is

$$X_k(\omega) = \begin{cases} Z_k(\omega), & \text{if } k \leq T(\omega) \\ Z_{T(\omega)}(\omega), & \text{otherwise} \end{cases}$$

    If Z is a (sub-) martingale, then X is a (sub-) martingale too.

6. (Stopped martingales are martingales). Let $(X_n)_n$ be a martingale. Let τ be a stopping time. Then $\tilde{X}_n = X_{n\wedge\tau}$ is a martingale.

7. (Doob's optional stopping theorem). Let $(X_n)_n$ be a super-martingale and T be a stopping time. Then X_T is integrable and $\mathbb{E}[X_T] \leq \mathbb{E}[X_0]$ in each of the following cases
   i. T is bounded
   ii. X is bounded and T is almost surely finite
   iii. $E[T] < \infty$ and $(X_n)_n$ has (surely) bounded differences, i.e., there is an $M > 0$ such that
   $$|X_n(\omega) - X_{n-1}(\omega)| \leq M,$$
   for all $n \in \mathbb{N}$ and $\omega \in \Omega$
   iv. $X_n \geq 0$ for all n and T is almost surely finite

8. (Optional stopping theorem, version 2). Let $(X_t)_t$ be a martingale on $(\Omega, \mathcal{F}, \mathrm{P})$ subject to a filtration $\mathfrak{F} = (\mathcal{F}_t)_t$ and let τ be a stopping time. Assume that one of the following holds
   i. τ is almost surely bounded, that is, there is a $\bar{\tau} \geq 0$, so that $\tau(\omega) \leq \bar{\tau}$ for P-almost all ω[2]
   ii. $\mathbb{E}[\tau]$ is finite and $\mathbb{E}[|X_k - X_k| \mid \mathcal{F}_k]$ is almost surely bounded, uniformly in k,
   iii. $|X_{\min(t,\tau)}|$ is almost surely bounded,

   Then X_τ is almost surely a well-defined random variable and
   $$\mathbb{E}[X_\tau] = \mathbb{E}[X_0].$$
   If X is assumed to be a super-martingale, then
   $$\mathbb{E}[X_\tau] \leq \mathbb{E}[X_0].$$
   If X is assumed to be a sub-martingale, then
   $$\mathbb{E}[X_\tau] \geq \mathbb{E}[X_0].$$

9. (Optional stopping theorem, more general version). Let $(X_t)_t$ be a martingale on $(\Omega, \mathcal{F}, \mathrm{P})$ subject to a filtration $\mathfrak{F} = (\mathcal{F}_t)_t$ and let τ be a stopping time. Suppose that X is uniformly integrable (then, it has a well-defined limit, X_∞ so we may define $\bar{X}_\tau = X_\tau 1_{\tau<\infty} + X_\infty 1_{\tau=\infty}$). Let $\tau' \leq \tau$ be two stopping times. Then,
   $$\mathbb{E}[X_\tau \mid \mathcal{F}_{\tau'}] = X_{\tau'}.$$

10. (Almost sure martingale convergence). Let $(X_n)_n$ be a martingale which is uniformly bounded in $\mathcal{L}_1$, i.e., $\sup_n \mathbb{E}[|X_n|] < \infty$. Then, there is a $X \in \mathcal{L}_1(\mathcal{F}_\infty)$, so that $X_n \to X$ a.s., where $\mathcal{F}_\infty = \sigma(\mathcal{F}_n, n \geq 0)$.

11. (Kolmogorov's sub-martingale inequality). Let $\{X_k\}_k$ be a nonnegative sub-martingale. Then, for $n \in \mathbb{N}_{>0}$ and $\alpha > 0$,
   $$\mathrm{P}\left[\max_{k=1,\ldots,n} X_k \geq \alpha\right] \leq \frac{\mathbb{E}[X_n]}{\alpha}.$$
   i. (Corollary 1). Let $\{X_k\}_k$ be a nonnegative martingale. Then $\mathrm{P}[\sup_{k\geq 1} X_k \leq \alpha] \leq \mathbb{E}[X_1]/\alpha$ for $\alpha > 0$.
   ii. (Corollary 2). Let $\{X_k\}_k$ be a martingale with $\mathbb{E}[X_k^2] < \infty$ for all $k \in \mathbb{N}_{>0}$. Then, $\mathrm{P}[\max_{k=1,\ldots,n} |X_k| \geq \alpha] \leq \mathbb{E}[X_n^2]/\alpha$ for all $n \in \mathbb{N}_{\geq 2}$ and $\alpha > 0$.
   iii. (Corollary 3). Let $\{X_k\}_k$ be a nonnegative super-martingale. Then, for $n \in \mathbb{N}_{>0}$ and $\alpha > 0$, $\mathrm{P}[\cup_{k\geq n}\{Z_k \geq \alpha\}] \leq \mathbb{E}[Z_n]/\alpha$.

12. (Azuma-Hoeffding inequality for martingales with bounded differences). Let $(X_i)_i$ be a martingale or a super-martingale and $|X_k - X_{k-1}| < c_k$ almost surely. Then for all $N \in \mathbb{N}$ and $t \in \mathbb{R}$,
   $$\mathrm{P}[X_N - X_0 \geq t] \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^N c_i^2}\right)$$
   If $(X_i)_i$ is a sub-martingale,
   $$\mathrm{P}[X_N - X_0 \leq -t] \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^N c_i^2}\right)$$

[2] This is a strong condition which is often not satisfied in practice. However, for fixed $N \in \mathbb{N}$, $\tau \wedge N$ is a stopping time. We often apply the optional stopping theorem for the bounded stopping time $\tau \wedge N$ and take $N \to \infty$.

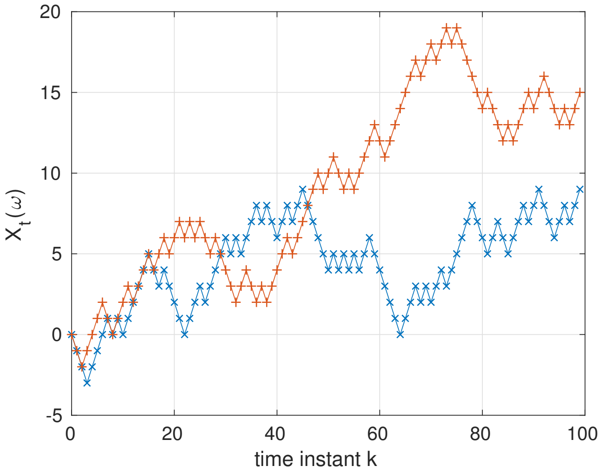

Figure 3.1: Random walk: two different paths, $(X_t(\omega_1))_t$ and $(X_t(\omega_2))_t$.

13. (Martingale inequalities). Let $(X_t)_{t\geq 0}$ be a càdlàg martingale and $t>0$. Define $X_t^* = \sup_{s\leq t}|X_s|$. Then, for every $t>0$
    i. for $\alpha>0$, $\mathrm{P}[X_t^*\geq\alpha]\leq\frac{\|X_t\|_1}{\alpha}$
    ii. for $p>1$, $\|X_t^*\|_p\leq\frac{p}{p-1}\|X_t\|_p$
14. (Nonnegative submartingale inequalities). Let $(X_t)_{t\geq 0}$ be a nonnegative càdlàg submartingale and $t>0$. Define $X_t^* = \sup_{s\leq t}|X_s|$. Then, for every $t>0$
    i. for $\alpha>0$, $\mathrm{P}[X_t^*\geq\alpha]\leq\frac{\|X_t\|_1}{\alpha}$
    ii. for $p>1$, $\|X_t^*\|_p\leq\frac{p}{p-1}\|X_t\|_p$

3.3 Random walk

1. (One-dimensional random walk). Take a sequence of independent random variables $(Z_t)_{t\in\mathbb{N}}$ that take values in $\{-1,1\}$, each with probability $1/2$. Define a random process $(X_t)_{t\in\mathbb{N}}$ with $X_0=0$ and $X_t=\sum_{i=1}^t Z_i$. The process $(X_t)_{t\in\mathbb{N}}$ is called a (one-dimensional) (simple) random walk on $\mathbb{Z}$. A random walk is shown in Figure 3.1.

   More generally, a random walk can be such that $p := \mathrm{P}[Z_{t+1}-Z_t=1]\neq\mathrm{P}[Z_{t+1}-Z_t=-1]$.

2. (Characteristics). The expectation of random walk $(X_t)_{t\in\mathbb{N}}$ with $p=1/2$ is $\mathbb{E}[X_t]=0$ and its variance is $\mathbb{E}[X_t^2]=t$. Additionally,
$$\lim_t\frac{\mathbb{E}[X_t]}{\sqrt{t}}=\sqrt{\frac{2}{\pi}}.$$

3. (Distribution of X_t). Let $(X_t)_t$ be a one-dimensional random walk with $p := \mathrm{P}[X_{t+1}-X_t=1]$ and define $q := 1-p$. For $t\in\mathbb{N}$, $t\geq 1$, random variables X_t take values on $\mathcal{X}_t := \{-t,-t+2,\ldots,t-2,t\}$ and their distribution is given by
$$\mathrm{P}[X_t=m]=\binom{t}{\frac{1}{2}(t+m)}p^{\frac{1}{2}(t+m)}q^{\frac{1}{2}(t-m)},$$
   for all $m\in\mathcal{X}_t$ This is the Binomial distribution on $\mathcal{X}_t$ with parameter p (See definition in Section 1.5.3).

4. (Maximum of random walk). Let X_t be a simple symmetric random walk (with $p=0.5$ and define $M_t=\max_{t'\leq t}X_{t'}$. Then, $M_0=0$, the support of M_t is $\{0,1,\ldots,t\}$ and
$$\mathrm{P}[M_t=m]=\mathrm{P}[X_t=m]+\mathrm{P}[X_t=m+1]=\binom{t}{\lfloor\frac{t+m+1}{2}\rfloor}2^{-t}$$

5. (Infinite often visits). Almost surely, the one-dimensional simple random walk visits every integer $n \in \mathbb{N}$ infinitely often.

6. (As a Markov chain). The one-dimensional random walk can be seen as a Markov chain with states in $\mathbb{Z}$ and $\mathrm{P}[X_{k+1} = i + i \mid X_k = i] = p$ and $\mathrm{P}[X_{k+1} = i - i \mid X_k = i] = 1 - p$.

7. (Probability to reach upper bound before lower bound). Let $(X_n)_n$ be a simple random walk starting at $x \in \mathbb{Z}$, that is, $X_0 = x$. Let $a < x < b$ for some $a, b \in \mathbb{Z}$. Let $\tau_a = \inf\{n \in \mathbb{Z} \mid X_n = a\}$ and $\tau_b = \inf\{n \in \mathbb{Z} \mid X_n = b\}$. Then
$$\mathrm{P}[\tau_a < \tau_b] = \frac{x-a}{b-a}.$$

8. (Average time to exit interval). Let $(X_k)_k$ be a random walk with $X_0 = x$. Suppose $x \in [a, b]$ with $a, b \in \mathbb{Z}$. Define the stopping time $\tau = \inf\{n \in \mathbb{N} \mid X_n \in \{a, b\}\}$ (therefore, $X_\tau \in \{a, b\}$). Define the stochastic process
$$Y_n = n + (X_n - a)(b - X_n).$$
Let $\mathcal{F}_n$ be the sigma algebra generated by $(X_k)_{k=0}^n$. Then,
   a) τ is an almost surely bounded stopping time
   b) $(Y_n)_n$ is an $(\mathcal{F}_n)_n$-martingale
   c) $Y_0 = (x - a)(b - x)$ and $Y_\tau = \tau$
   d) From the (general) optional stopping theorem, $\mathbb{E}[\tau] = \mathbb{E}[Y_\tau] = \mathbb{E}[Y_0] = (x - a)(b - x)$

9. (Gaussian random walk). Take a sequence of independent random variables Z_t with $Z_t \sim \mathcal{N}(\mu, \sigma^2)$. The random process $(X_t)_t$ with $X_0 = 0$ and $X_t = Z_1 + \ldots + Z_t$ is called Gaussian random walk. It is $X_t \sim \mathcal{N}(t\mu, t\sigma^2)$.

3.4 Brownian motion

1. (Definition). A stochastic process $(X_t)_{t \in \mathbb{R}_+}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called Brownian motion if it is continuous and has stationary independent increments. A process $(W_t)_{t \in \mathbb{R}_+}$ is called a Wiener process if it is a Brownian motion with
   i. $W_0 = 0$
   ii. $\mathbb{E}[W_t] = 0$
   iii. $\mathbb{E}[W_t^2] = t$ for all $t \in \mathbb{R}_+$

2. (Definition according to Øksendal). Following Kolmogorov's extension theorem, for given time instants $t_1, \ldots, t_k$ in a time set T (typically $T = [0, \infty)$), define the following proability measure on $\mathbb{R}^{nk}$
$$\nu_{t_1,\ldots,t_k}(F_1 \times \cdots \times F_k) = \int_{F_1 \times \cdots \times F_k} p(t_1, x, x_1)p(t_2 - t_1, x_2, x_1) \cdots p(t_k - t_{k-1}, x_{k-1}, x_k)\mathrm{d}x_1\mathrm{d}x_2 \cdots \mathrm{d}x_k,$$
where p is the function
$$p(t, x, y) = (2\pi t)^{-n/2} \exp\left(-\tfrac{\|x-y\|^2}{2t}\right).$$
Let $(\Omega, \mathcal{F}, \mathrm{P})$ be the associated probability space from Following Kolmogorov's extension theorem and let $(B_t)_t$ be an $\mathbb{R}^n$-valued stochastic process with
$$\mathrm{P}[B_{t_1} \in F_1, \ldots, B_{t_k} \in F_k] = \nu_{t_1,\ldots,t_k}(F_1 \times \cdots \times F_k).$$
Such a process is a (version of) Brownian motion starting at x.

3. (Expectation and covariance). Let $(B_t)_{t \in T}$ be a Brownian motion starting at x. Then $\mathbb{E}[B_t] = x$ for all $t \in T$. Let $t_1, \ldots, t_k$ be time instants and $Z = (B_{t_1}, B_{t_2}, \ldots, B_{t_k})$ be an $\mathbb{R}^{nk}$-valued random variable. Then,
$$\operatorname{cov}[Z] = \begin{bmatrix} t_1 I_n & t_1 I_n & \cdots & t_1 I_n \\ t_1 I_n & t_2 I_n & \cdots & t_2 I_n \\ \vdots & \vdots & \ddots & \vdots \\ t_1 I_n & t_2 I_n & \cdots & t_k I_n \end{bmatrix}$$

Observe that, as a result, we have

$$\mathbb{E}[(B_t - x)(B_s - x)] = n \min(s, t).$$

4. (Existence of continuous version). The Brownian motion satisfies Kolmogorov's continuity theorem with $\alpha = 4$, $\beta = 1$ and $L = n(n+2)$. In particular, $\mathbb{E}[\|X_\tau - X_{\tau'}\|^4] = n(n+2)|\tau' - \tau|^2$.

5. (Zero crossing). Properties:

    i. Define the set of zero crossing times, $Z_0 = \{t \geq 0 \mid B_t = 0\}$. With probability 1, the Lebesgue measure of Z_0 is zero,

    ii. Almost surely, Z_0 is a closed set and has no isolated points,

    iii.

    iv. The Brownian motion crosses the time axis infinitely often in every time interval $(0, t)$ for $t > 0$.

6. (Distribution of maximum). Let X_t be a Brownian motion and $M_t = \max_{s \leq t} X_s$. Then, for all $t > 0$ and $a > 0$,

$$\mathrm{P}[M_t \geq a] = 2\mathrm{P}[X_t \geq a] = 2(1 - \Phi(a/\sqrt{t})).$$

7. (Attainment of maximum 1). Almost surely, the set of times where B_t attains a local maximum is dense in $[0, +\infty)$

8. (Attainment of maximum 2). On any interval, B_t almost surely does not attain the same maximum.

9. (Strict maximum). Almost surely, every local maximum of a Brownian motion is a strict local maximum.

10. (Countability of the set maximum times). Almost surely, the set of times when B_t attains a local maximum is countable.

11. (Maxima are distinct). The local maxima of B_t are almost surely distinct

12. (Nowhere differentiable). For every $\omega \in \Omega$, $t \mapsto X_t(\omega)$ is nowhere differentiable.

13. (Orthogonal transformation). Let B_t be an n-dimensional Brownian motion starting at 0 and U be an orthogonal matrix, $UU^\top = I$. Them, $\tilde{B}_t = UB_t$ is a Brownian motion.

14. (Brownian scaling). Let B_t be an n-dimensional Brownian motion starting at 0 and $c > 0$. Then, $\hat{B}_t = {}^{1}\!/_{c} B_{c^2 t}$ is a Brownian motion.

15. (Time inversion). Let B_t be an n-dimensional Brownian motion starting at 0 and $(\breve{B}_t)_t$ is a process with $\breve{B}_0 = 0$ and $\breve{B}_t = tB_{1/t}$. Then $\breve{B}$ is a Brownian motion.

16. (Integrated Brownian motion). The integral of the one-dimensional Brownian motion starting at 0, $\mathrm{ibm}(t, \omega) := \int_0^t B_s(\omega)\mathrm{d}s$, is a random variable which follows the normal distribution $\mathcal{N}(0, t^3/3)$.

17. (Exit time). Let $(B_t)_t$ be a one-dimensional Brownian motion on $(\Omega, \mathcal{F}, \mathrm{P})$ started at 0 and define $\tau(\omega) = \inf\{t \in T \mid B_t \notin [-a, b]\}$, where $a, b > 0$. This means that τ is the first time when the process leaves the interval $[-a, b]$. Then,

    i. τ is an integrable random variable

    ii. $\mathbb{E}[\tau] = ab$

    iii. $\mathbb{E}[W_\tau] = 0$ and $\mathbb{E}[W_\tau^2] = \mathbb{E}[\tau] = ab$

    iv. $\mathrm{P}[W_\tau = -a] = \frac{b}{a+b}$ and $\mathrm{P}[W_\tau = b] = \frac{a}{a+b}$

Note. We often need to evaluate the expectation of a transformation of the Brownian motion, $Y_t = f(B_t)$. Using the fact that B_t is normally distributed at every t and the law of the unconscious statistician,

$$\mathbb{E}[f(B_t)] = \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{\infty} f(x) e^{-\frac{x^2}{2t}} \mathrm{d}x,$$

provided $f(B_t)$ is integragle. Similarly, we may need to evaluate $\mathbb{E}[\int_0^t f(B_s)\mathrm{d}s] = \int_0^t \mathbb{E}[f(B_s)]\mathrm{d}s = \int_0^t \frac{1}{\sqrt{2\pi s}} \int_{-\infty}^{\infty} f(x) e^{-\frac{x^2}{2s}} \mathrm{d}x\, \mathrm{d}s$ (using Fubini's Theorem).

3.5 Markov processes

1. (Definition). Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in T}, \mathrm{P})$ be a filtered probability space. Let $\{X_t\}_{t\in T}$ be a random process which is adapted to the filtration $\{\mathcal{F}_t\}_{t\in T}$. Let $\{\mathcal{G}_t^0\}_{t\in T}$ be the filtration generated by $\{X_t\}_{t\in T}$ and $\mathcal{G}_t^\infty = \sigma(\{X_u : u \geq t, u \in T\})$. The process is said to be Markovian if for every $t \in T$, the past $\mathcal{F}_t$ and the future $\mathcal{G}_t^\infty$ are conditionally independent given X_t.

2. (Characterization). The following are equivalent
   i. The process $\{X_t\}_t$ is Markovian with state space $(E, \mathcal{E})$
   ii. For every $t \in T$ and $u > t$, and $f \in \mathcal{E}_+$ $\mathbb{E}[f \circ X_u \mid \mathcal{F}_t] = \mathbb{E}[f \circ X_u \mid X_t]$.
   iii. Let E be a p-system generating $\mathcal{E}$. For every $t \in T$ and $u > t$, and $A \in E$ it is $\mathbb{E}[1_A \circ X_u \mid \mathcal{F}_t] = \mathbb{E}[1_A \circ X_u \mid X_t]$.
   iv. For every $t \in T$ and positive $V \in \mathcal{G}_t^\infty$, $\mathbb{E}[V \mid \mathcal{F}_t] = E[V \mid X_t]$.
   v. For every $t \in T$ and positive $V \in \mathcal{G}_t^\infty$, $\mathbb{E}[V \mid \mathcal{F}_t] \in \sigma X_t$.

3. (Markov transition functions). Let $(P_{t,u})_{t,u\in T, t\leq u}$ be a family of Markov transition kernels on $(\Omega, \mathcal{F})$. This is said to be a Markovian transition function if $P_{s,t}P_{t,u} = P_{s,u}$ for all $0 \leq s < t \leq u$.

4. (Chapman-Kolmogorov equation). A Markov process $X = \{X_t\}_{t\in T}$ is said to admit $(P_{t,u})$ as a transition function if
$$\mathbb{E}[f \circ X_u \mid X_t] = (P_{t,u} f) \circ X_t, t < u,$$
for all nonnegative functions f.

5. (Time homogeneity). We call a Markov process *time homogeneous* if it admits a transition function $(P_{t,u})$ which depends only on $t - u$, i.e., there is a Markov kernel with $P_{t,u} = P_{t-u}$.

6. (Martingales from Markov chains). Let $(X_n)_n$ a random process adapted to a filtration $\mathcal{F}$ with state space $(E, \mathcal{E})$. Then $(X_n)_n$ is a Markov chain with transition kernel P with respect to $\mathcal{F}$ if and only if
$$M_n = f \circ X_n - \sum_{m=0}^{n-1} (Pf - f) \circ X_n,$$
is a martingale with respect to $\mathcal{F}$ for every nonnegative $\mathcal{E}$-measurable function f.

3.6 Markov decision processes

3.6.1 General

1. (Definition of MCM). A Markov control model (MCM) is a tuple $(\mathcal{X}, \mathcal{U}, U, Q, c)$ consisting of
   i. two Borel spaces $\mathcal{X}$ and $\mathcal{U}$ called the state and control spaces respectively,
   ii. a multivalued function mapping $U : \mathcal{X} \rightrightarrows \mathcal{U}$ which maps a state $x \in \mathcal{X}$ to a set $U(x) \subseteq \mathcal{U}$ of feasible control actions. Define the graph of U as the set $\mathcal{K} := \operatorname{gph}(U) = \{(x, u) \in \mathcal{X} \times \mathcal{U} \mid u \in U(x)\}$. We assume that $\mathcal{K}$ contains the graph of a measurable (single-valued) function $u : \mathcal{X} \to \mathcal{U}$.
   iii. a transition kernel $Q : \mathcal{B}(\mathcal{X}) \times \mathcal{K} \ni (B, x, u) \mapsto Q(B, x, u) \in [0, 1]$, where (x, u) is such that $x \in \mathcal{X}$, $u \in U(x)$ and $B \in \mathcal{B}(\mathcal{X})$ (See Section 1.1.10).
   iv. A cost function c is a measurable function $c : \mathcal{K} \to \mathbb{R}$.

2. (Example). Let $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ be a Borel space. Let $\{\xi_t\}_t$ be a collection of iid random variables with values in $\mathcal{S}$. Let μ be their common probability distribution. Consider the dynamical system
$$x_{t+1} = F(x_t, u_t, \xi_t).$$
Then, the transition kernel $Q(B, x, u)$ is
$$\begin{aligned} Q(B, x, u) &= \mu(\{\omega \in \mathcal{S} \mid F(x, u, \omega) \in B\}) \\ &= \int_{\mathcal{S}} 1_B(F(x, u, \omega)) \mathrm{d}\mu(\omega) \\ &= \mathbb{E} 1_B(F(x, u, \omega)). \end{aligned}$$

3. (Equivalent representation of Markov control models). For every MCM, there is a Borel space $\mathcal{S}$, a function $F : \mathcal{X} \times \mathcal{U} \times \mathcal{S}$ and an $\mathcal{S}$-valued iid process $\{\xi_t\}_t$ so that

$$x_{t+1} = F(x_t, u_t, \xi_t).$$

4. (Definition of $\mathcal{H}_t$ and $\overline{\mathcal{H}}_t$). Define $\mathcal{H}_0 = \mathcal{X}$ and $\mathcal{H}_t = \mathcal{K}^t \times \mathcal{X}$. $\mathcal{H}_t$ contains elements of the form $h_t = (x_0, u_0, \ldots, x_{t-1}, u_{t-1}, x_t)$ with $u_k \in U(x_k)$. Define also the linear space $\overline{\mathcal{H}}_t = (\mathcal{X} \times \mathcal{U})^t \times \mathcal{X}$ with $\overline{\mathcal{H}}_0 = \mathcal{X}$.

5. (Definition of a policy). A policy is a sequence $\pi = (\pi_0, \pi_1, \ldots)$ of transition kernels $\pi_t : \mathcal{B}(\mathcal{U}) \times \mathcal{H}_t \to [0, 1]$ with

$$\pi_t(A(x_t), h_t) = 1, \text{ for all } h_t \in H_t, t \in \mathbb{N}.$$

6. (The canonical probability space $(\Omega, \mathcal{F}, \mathrm{P})$). Given an MCM $(\mathcal{X}, \mathcal{U}, U, Q, c)$, let $\Omega = \overline{\mathcal{H}}_\infty = \prod_{t=1}^{\infty} \mathcal{X} \times \mathcal{U}$. Ω contains sequences $\omega = (x_0, u_0, x_1, u_1, \ldots)$. Let $\mathcal{F}$ be the corresponding product σ-algebra. Given a probability measure ν on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ (called the initial distribution) and a policy π, according to the Ionescu-Tulcea Theorem, there is a unique probability measure P_ν^π so that for $B \in \mathcal{B}(\mathcal{X})$, $C \in \mathcal{B}(\mathcal{U})$, $h_t \in \mathcal{H}_t$ and $t \in \mathbb{N}$:

   i. $\mathrm{P}_\nu^\pi[x_0 \in B] = \nu(B)$

   ii. $\mathrm{P}_\nu^\pi[u_t \in C \mid h_t] = \pi_t(C, h_t)$

   iii. $\mathrm{P}_\nu^\pi[x_{t+1} \in B \mid h_t, u_t] = Q(B, x_t, u_t)$

   Note that the last condition is a Markov-like property, but it does not imply that x_t is a Markov process.

7. (Markov decision process). A (discrete-time) Markov decision process (MDP) is a tuple $(\Omega, \mathcal{F}, \mathrm{P}_\nu^\pi, \{x_t\}_t)$. In other words, for a given policy π and a given initial distribution ν, an MDP is a stochastic process $\{x_t(\omega)\}_{t \in \mathbb{N}}$ over the canonical probability space $(\Omega, \mathcal{F}, \mathrm{P}_\nu^\pi)$.

8. (Space Φ). We define the space Φ of all transition kernels $\phi : \mathcal{B}(\mathcal{U}) \times \mathcal{X} \to [0, 1]$ with $\phi(\mathcal{U}(x), x) = 1$.

9. (Types of policies). A policy π is called

   i. a *randomized Markov policy* if there are $\phi_t \in \Phi$ so that $\pi_t(\,\cdot \mid h_t) = \phi_t(\,\cdot \mid x_t)$, for $h_t \in \mathcal{H}_t$, $t \in \mathbb{N}$.

   ii. a *randomized stationary policy* if there is a $\phi \in \Phi$ with $\pi_t(\,\cdot \mid h_t) = \phi(\,\cdot \mid x_t)$, for $h_t \in \mathcal{H}_t$, $t \in \mathbb{N}$.

   iii. a *deterministic policy* if there are functions $g_t : \mathcal{H}_t \to \mathcal{U}$ such that for $h_t \in \mathcal{H}_t$, $t \in \mathbb{N}$, $g_t(h_t) \in U(x_t)$ and $\pi_t(\,\cdot \mid h_t)$ is concentrated at $g_t(h_t)$, that is, $\pi_t(C, h_t) = 1_C(g_t(h_t))$ for all $C \in \mathcal{B}(\mathcal{U})$

   iv. a *deterministic Markov policy* if there exist functions $g_t : \mathcal{X} \to \mathcal{U}$, with $g_t(x) \in U(x)$ for all $x \in \mathcal{X}$, such that $\pi_t(\,\cdot\,, h_t)$ is concentrated at $g_t(x_t)$ for all $h_t \in \mathcal{H}_t$, $t \in \mathbb{N}$.

   v. a *deterministic stationary policy* if there is a function $g : \mathcal{X} \to \mathcal{U}$, with $g(x) \in U(x)$ for all $x \in \mathcal{X}$, such that $\pi_t(\,\cdot\,, h_t)$ is concentrated at $g(x_t)$ for all $h_t \in \mathcal{H}_t$, $t \in \mathbb{N}$.

10. (Markovianity of $\{x_t\}_t$). Let ν be an initial distribution. Let $\pi = \{\phi_t\}$ be a randomized Markov policy (see 9-i). Then, $\{x_t\}_t$ is a non-homogeneous Markov process with transition kernels $\{Q(\cdot, \cdot, \phi_t)\}_t$, that is, for $B \in \mathcal{B}(\mathcal{X})$

$$\mathrm{P}_\nu^\pi[x_{t+1} \in B \mid x_0, \ldots, x_t] = \mathrm{P}_\nu^\pi[x_{t+1} \in B \mid x_t] = Q(B, x_t, \phi_t).$$

   If $\pi = \{\phi\}_t$ is a stationary policy, then the above also holds with

$$\mathrm{P}_\nu^\pi[x_{t+1} \in B \mid x_0, \ldots, x_t] = \mathrm{P}_\nu^\pi[x_{t+1} \in B \mid x_t] = Q(B, x_t, \phi),$$

   so $\{x_t\}_t$ is a time-homogeneous Markov process.

3.6.2 Optimal control problems

1. (Statement). The finite-horizon optimal control problem is stated as

$$\begin{aligned}\operatorname*{minimize}_{\pi:\text{policy}} J(\pi,x) &:= \mathbb{E}_x^{\pi}\left[\sum_{t=0}^{N-1} c(x_t,u_t) + c_N(x_N).\right] \\ &= \mathbb{E}^{\pi}\left[\sum_{t=0}^{N-1} c(x_t,u_t) + c_N(x_N)\middle| x_0 = x\right]\end{aligned}$$

2. (DP Theorem). Define the function $J_N(x) = c_N(x)$ and for $t = N-1, N-2, \ldots, 0$,

$$J_t(x) := \min_{U(x)} c(x,u) + \int_{\mathcal{X}} J_{t+1}(\chi) Q(\mathrm{d}\chi, x, u),$$

where the minimization is over control functions $u : \mathcal{X} \to \mathcal{U}$. Suppose that Suppose all J_t are measurable and for all $t \in \mathbb{N}_{[0,N-1]}$ there is a selection $u_t^\star(x) \in U(x)$, $u_t^\star : \mathcal{X} \to \mathcal{U}$ which attains the minimum, that is

$$J_t(x) = c(x, u_t^\star(x)) + \int_{\mathcal{X}} J_{t+1}(\chi) Q(\mathrm{d}\chi, x, u_t^\star(\,\cdot\,)).$$

Then, the deterministic Markov policy $\pi^\star = \{u_0^\star, u_1^\star, \ldots, u_{N-1}^\star\}$ is optimal and the value function $J^\star$ is equal to J_0, that is

$$J^\star(x) = J_0(x) = J(\pi^\star, x)$$

3. (Measurable selection theorem 1). There exists a measurable selection $u_t^\star$ in the above DP theorem, if

   i. (Control constraints). U is compact-valued (i.e., for every x, $U(x)$ is compact)

   ii. (Cost function). $c(x,\,\cdot\,)$ is lower semicontinuous on $U(x)$ for every $x \in \mathcal{X}$

   iii. (Integral). the function $\xi(x,u) = \int_{\mathcal{X}} u(\chi)Q(\mathrm{d}\chi,x,u)$ on K satisfies one of the following conditions:

      i. $\xi(x,\,\cdot\,)$ is lower semi-continuous on $U(x)$ for every $x \in \mathcal{X}$ and every continuous bounded function u on $\mathcal{X}$

      ii. $\xi(x,\,\cdot\,)$ is lower semi-continuous on $U(x)$ for every $x \in \mathcal{X}$ and every measurable bounded function u on $\mathcal{X}$.

4. (Measurable selection theorem 2). There exists a measurable selection $u_t^\star$ in the above DP theorem, if

   i. (Control constraints). U is compact-valued (i.e., for every x, $U(x)$ is compact) and the multi-valued function $x \mapsto U(x)$ is upper semi-continuous

   ii. (Cost function). Function c is lower semicontinuous and bounded below

   iii. (Transition kernel). the transition kernel Q satisfies of the following conditions

      i. it is *weakly continuous*, that is, $\xi(x,u) = \int_X u(\chi)Q(\mathrm{d}\chi,x,u)$ is continuous and bounded on K for every continuous bounded function u on $\mathcal{X}$

      ii. it is *strongly continuous*, that is, ξ is continuous and bounded on K for every measurable bounded function u on $\mathcal{X}$

5. (Measurable selection theorem 3). There exists a measurable selection $u_t^\star$ in the above DP theorem, if

   i. (Cost function). The stage cost c is lower semi-continuous, bouned below and inf-compact on K, that is, for every $x \in \mathcal{X}$ and $r \geq 0$, the set $\{u \in U(x) \mid c(x,u) \leq r\}$ is compact (in other words, c has compact level sets)

   ii. (Transition kernel). Condition 4iii in Measurable Selection Theorem 2 holds.

3.6.3 Linear-Quadratic problems

Work in progress (this section will be updated in the upcoming version). Stay tuned on Twitter for updates (`@isToxic`).

4 Stochastic Differential Equations

4.1 Itô Integral

1. (Class $\mathcal{V}$). Let $(\Omega, \mathcal{F}, \mathfrak{F}, \mathrm{P})$ be a filtered probability space where $\mathfrak{F} = (\mathcal{F}_t)_{t \in T}$ is a filtration, $T = [0, +\infty)$ and $t, t' \in T$ with $t < t'$. We define the class $\mathcal{V} = \mathcal{V}(t, t')$ to be a class of functions $f(t, \omega) : T \times \Omega \to \mathbb{R}$ with
   i. f is $\mathcal{B} \times \mathcal{F}$-measurable where $\mathcal{B}$ is the Borel σ-algebra on T
   ii. $f(t, \omega)$ is $\mathfrak{F}$-adapted
   iii. $\mathbb{E} \int_t^{t'} f(t, \omega)^2 \mathrm{d}t < \infty$

2. (Itô integral for elementary functions). Let $(B_t)_{t \geq 0}$ be the standard Brownian motion on the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathrm{P})$ and ϕ be an elementary function of class $\mathcal{V}(t, t')$, that is,
$$\phi(t, \omega) = \sum_i e_i(\omega) 1_{[t_i, t_{i+1})}(t),$$
where e_i is $\mathcal{F}_{t_i}$-measurable. We define the Itô integral of ϕ from t to t' to be the random variable
$$\int_t^{t'} \phi(t, \omega) \mathrm{d}B_t = \sum_i e_i(\omega)(B_{t_{i+1}}(\omega) - B_{t_i}(\omega))$$

3. (Itô integral on $\mathcal{V}(t, t')$). Let $(B_t)_{t \geq 0}$ be the standard Brownian motion on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathrm{P})$ and $f \in \mathcal{V}(t, t')$. Let ϕ_n be a sequence of elementary functions which converges to f in the following sense:
$$\mathbb{E}\left[\int_t^{t'} (\phi_n(s, \omega) - f(s, \omega))^2 \mathrm{d}s\right] \to 0.$$
Then,
$$\int_t^{t'} f(s, \omega) \mathrm{d}B_s = \lim_{n \to \infty} \int_t^{t'} \phi_n(s, \omega) \mathrm{d}B_s,$$
where the limit is in the $L^2(\Omega, \mathcal{F}, \mathrm{P})$ sense.

4. (Properties of the Itô integral). The Itô integral has the following properties (for all $f, g \in \mathcal{V}(t, t')$)
   i. (Break down). $\int_t^{t'} f(s, \omega) \mathrm{d}B_s = \int_t^{t''} f(s, \omega) \mathrm{d}B_s + \int_{t''}^{t'} f(s, \omega) \mathrm{d}B_s$, for almost all ω
   ii. (Linearity). $\int_t^{t'} (cf + g) \mathrm{d}B_s = c \int_t^{t'} f \mathrm{d}B_s + \int_t^{t'} g \mathrm{d}B_s$, for almost all ω, where c is a constant
   iii. (Zero expectation). $\mathbb{E}\left[\int_t^{t'} f(s, \omega) \mathrm{d}B_s\right] = 0$
   iv. (Measurability). $\int_t^{t'} f(s, \omega) \mathrm{d}B_s$ is $\mathcal{F}_{t'}$-measurable
   v. (Isometry property). $\mathbb{E}[(\int_t^{t'} f(s, \omega) \mathrm{d}B_s)^2] = \mathbb{E}[\int_t^{t'} f(s, \omega)^2 \mathrm{d}s]$
   vi. (Martingale). $M_t(\omega) = \int_0^t f(s, \omega) \mathrm{d}B_s$ is a martingale with respect to the filtration $(\mathcal{F}_t)_t$ and P and $t \mapsto M_t(\omega)$ is almost surely continuous. As a result, Doob's martingale theorem applies, that is
$$\mathrm{P}[|M_t| \geq t] \leq \frac{1}{t^2} \mathbb{E}[\int_0^t f(s, \omega)^2 \mathrm{d}s]$$

5. (Itô process). The stochastic process X_t given by
$$X_t = X_0 + \int_0^t u(s, \omega) \mathrm{d}s + \int_0^t v(s, \omega) \mathrm{d}B_s,$$

where v is an Itô-integrable function with $\mathrm{P}[\int_0^t v^2\mathrm{d}s < \infty, \forall t \geq 0] = 1$, is called an Itô process. Such a process is also written in the following shorter differential form

$$\mathrm{d}X_t = u\mathrm{d}t + v\mathrm{d}B_t.$$

6. (Multi-dimensional Itô process). Let B_t be an m-dimensional Brownian motion and

$$\mathrm{d}X_t = u\mathrm{d}t + V\mathrm{d}B_t$$

where $u(t,\omega) = [u_1(t,\omega)\ \cdots u_n(t,\omega)]^\top$, $V = (V_{i,j}(t,\omega))_{i,j}$ is an n-by-m matrix where $V_{i,j}$ are Itô-integrable functions, $\mathrm{d}B_t = [\mathrm{d}B_{1,t}\ \cdots\ \mathrm{d}B_{m,t}]^\top$ and $X_t = [X_{1,t}\ \cdots\ X_{n,t}]^\top$. In other words,

$$\begin{bmatrix}\mathrm{d}X_{1,t}\\ \mathrm{d}X_{2,t}\\ \vdots\\ \mathrm{d}X_{n,t}\end{bmatrix} = \begin{bmatrix}u_1\\ u_2\\ \vdots\\ u_n\end{bmatrix}\mathrm{d}t + \begin{bmatrix}V_{1,1} & \cdots & V_{1,m}\\ \vdots & \ddots & \vdots\\ V_{n,1} & \cdots & V_{n,m}\end{bmatrix}\begin{bmatrix}\mathrm{d}B_{1,t}\\ \mathrm{d}B_{2,t}\\ \vdots\\ \mathrm{d}B_{m,t}\end{bmatrix}.$$

This is called an n-by-m-dimensional Itô process.

7. (Itô formula — 1-dimensional). Let X_t be an Itô process, $\mathrm{d}X_t = u\mathrm{d}t + v\mathrm{d}B_t$, and let $g \in C^2([0,+\infty) \times \mathbb{R})$. Define $Y_t = g(t, X_t)$. Let us denote the partial derivative of g wrt t by g_t and let, likewise, $g_x(t,x) = \frac{\partial g}{\partial x}g(t,x)$ and $g_{xx}(t,x) = \frac{\partial g}{\partial x}g_x(t,x)$. Then,

$$\mathrm{d}Y_t = g_t(t,X_t)\mathrm{d}t + g_x(t,X_t)\mathrm{d}x + \tfrac{1}{2}g_{xx}(t,X_t)(\mathrm{d}X_t)^2,$$

with the calculus rules $\mathrm{d}t \cdot \mathrm{d}t = \mathrm{d}t \cdot \mathrm{d}B_t = \mathrm{d}B_t \cdot \mathrm{d}t = 0$ and $\mathrm{d}B_t \cdot \mathrm{d}B_t = \mathrm{d}t$.

8. (Itô formula — multi-dimensional). Let B_t be an m-dimensional Brownian motion and

$$\mathrm{d}X_t = u\mathrm{d}t + V\mathrm{d}B_t$$

be an n-by-m-dimensional Itô process. Let g be a $C^2([0,\infty) \times \mathbb{R}^n; \mathbb{R}^p)$ map and $Y_t = g(t, X_t)$. Then,

$$\mathrm{d}Y_{k,t} = \frac{\partial g_k}{\partial t}(t,X_t)\mathrm{d}t + \sum_i \frac{\partial g_k}{\partial x_i}(t,X_t)\mathrm{d}X_i + \tfrac{1}{2}\sum_{i,j}\frac{\partial^2 g_k}{\partial x_i \partial x_j}(t,X_t)\mathrm{d}X_i\mathrm{d}X_j,$$

where $\mathrm{d}B_i\mathrm{d}B_j = \delta_{i,j}\mathrm{d}t$, $\mathrm{d}B_i\mathrm{d}t = \mathrm{d}t\mathrm{d}B_i = 0$.

9. (Examples). Here are a few examples:

   i. Take $g(t,x) = \frac{1}{2}x^2$ and $X_t = B_t$. Then, $\mathrm{d}Y_t = B_t\mathrm{d}B_t + \frac{1}{2}\mathrm{d}t$, which leads to

   $$\tfrac{1}{2}B_t^2 = \int B_s\mathrm{d}B_s + \tfrac{1}{2}t.$$

   ii. By taking $X_t = B_t$ and $g(t,x) = tx$ we obtain

   $$\int_0^t s\mathrm{d}B_s = tB_t - \int_0^t B_s\mathrm{d}s,$$

   where $\int_0^t B_s\mathrm{d}s$ is an integrated Brownian motion.

   iii. By taking $X_t = B_t$ and $g(t,x) = \frac{1}{3}x^3$, we obtain

   $$\int_0^t B_s^2\mathrm{d}B_s = \tfrac{1}{3}B_t^3 - \int_0^t B_s\mathrm{d}s.$$

   iv. (Integration by parts). Let X_t and Y_t be two Itô processes. Then,

   $$\mathrm{d}(X_tY_t) = X_t\mathrm{d}Y_t + Y_t\mathrm{d}X_t + \mathrm{d}X_t\mathrm{d}Y_t,$$

   in other words,

   $$\int_0^t X_s\mathrm{d}Y_s = X_tY_t - X_0Y_0 - \int_0^t Y_s\mathrm{d}X_s - \int_0^t \mathrm{d}X_s\mathrm{d}Y_s$$

v. (Exponential). Let $\theta(t,\omega)$ be an n-dimensional random process with $\theta_i(t,\omega) \in \mathcal{V}([0,T])$ for $i = 1,\dots,n$ with $T \leq \infty$. Define

$$Z_t = \exp\left\{\int_0^t \theta(s,\omega)\mathrm{d}B_s - \tfrac{1}{2}\int_0^t \theta^2(s,\omega)\mathrm{d}s\right\},$$

for $t \in [0,T]$. Then,

$$\mathrm{d}Z_t = Z_t\theta(t,\omega)\mathrm{d}B_t$$

10. (Ito integrals of deterministic functions). Let $g : [0,T] \to \mathbb{R}$ be a Borel-measurable function. For every $t \geq 0$, the random variable

$$X_t(\omega) = \int_0^t g(s)\mathrm{d}B_s(\omega),$$

is normally distributed with zero mean and variance $\mathrm{Var}[X] = \int_0^t g(s)^2\mathrm{d}s$.

11. (Expectation of product of Itô integrals). Let $(X_t)_t$, $(Y_t)_t$ be two stochastic processes on the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathrm{P})$ so that X_t and Y_t are in $\mathcal{V}(0,T)$ for some $T > 0$. Then, for $0 \leq t \leq T$

$$\mathbb{E}\left[\int_0^t X_s\mathrm{d}B_s \int_0^t Y_s\mathrm{d}B_s\right] = \int_0^t \mathbb{E}[X_sY_s]\mathrm{d}s$$

Table 4.1: Application of Itô's formula

X_t	$\mathrm{d}X_t = u\mathrm{d}t + v\mathrm{d}B_t$
B_t	$\mathrm{d}B_t$
B_t^2	$2B_t\mathrm{d}B_t + \mathrm{d}t$
$B_t^2 - t$	$2B_t\mathrm{d}B_t$
B_t^3	$3B_t^2\mathrm{d}B_t + 3B_t\mathrm{d}t$
e^{B_t}	$e^{B_t}\mathrm{d}B_t + \frac{1}{2}e^{B_t}\mathrm{d}t$
$e^{B_t - \frac{1}{2}t}$	$e^{B_t - \frac{1}{2}t}\mathrm{d}B_t$
$e^{\frac{1}{2}t}\sin B_t$	$e^{\frac{1}{2}t}\cos B_t\mathrm{d}B_t$
$e^{\frac{1}{2}t}\cos B_t$	$-e^{\frac{1}{2}t}\sin B_t\mathrm{d}B_t$
$(B_t + t)e^{-B_t - \frac{1}{2}t}$	$(1 - B_t - t)e^{-B_t - \frac{1}{2}t}\mathrm{d}B_t$

4.2 Stochastic differential equations

1. (Ornstein-Uhlenbeck process). The stochastic differential equation

$$\mathrm{d}X_t = \mu X_t\mathrm{d}t + \sigma\mathrm{d}B_t,$$

is called the Ornstein-Uhlenbeck process. Its solution is[1]

$$X_t = e^{\mu t}X_0 + \sigma\int_0^t e^{\mu(t-s)}\mathrm{d}B_s,$$

with expectation $\mathbb{E}[X_t] = e^{\mu t}X_0$ and variance $\mathrm{Var}[X_t] = {}^{\sigma^2}\!/_{2\mu}(e^{2\mu t} - 1)$.

[1] Multiply both sides by $e^{-\mu t}$ and apply It's formula on $\mathrm{d}(e^{-\mu t}X_t)$. To find the variance, use the Itô isometry.

Table 4.2: Stochastic integrals — we denote by B_t the standard Brownian motion with $B_0 = 0$ and $\mathrm{ibm}(t, \omega) := \int_0^t B_s \mathrm{d}s$ is the integrated Brownian motion.

Stochastic Integral	Result	Variance
$\int_0^t \mathrm{d}B_s$	B_t	t
$\int_0^t s\mathrm{d}B_s$	$tB_t - \mathrm{ibm}(t, \omega)$	$\frac{1}{3}t^3$
$\int_0^t B_s \mathrm{d}B_s$	$\frac{1}{2}B_t^2 - \frac{1}{2}t$	$\frac{1}{2}t^2$
$\int_0^t B_s^2 \mathrm{d}B_s$	$\frac{1}{3}B_t^3 - \mathrm{ibm}(t, \omega)$	$3t^3$
$\int_0^t B_s^k \mathrm{d}B_s$	$\frac{1}{k+1}B_t^{k+1} - k\int_0^t B_s^{k-1}\mathrm{d}s$	
$\int_0^t e^{B_s - s/2}\mathrm{d}B_s$	$e^{B_t - t/2} - 1$	$e^t - 1$

5 Information Theory

5.1 Entropy and Conditional Entropy

1. (Self-Information, construction). Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a discrete probability space. A self-information function I must satisfy the following desiderata: (i) if ω_i is sure ($\mathrm{P}[\omega_i] = 1$), then this offers no information, that is $I(\omega_i) = 0$, (ii) if ω_i is not sure, that is $\mathrm{P}[\omega_i] < 1$, then $I(\omega_i) > 0$, (iii) $I(\omega)$ depends on the probability $\mathrm{P}[\omega]$, that is, there is a function f so that $I(\omega) = f(\mathrm{P}[\omega])$ (iv) for two independent events A and B, $I(A \cap B) = I(A) + I(B)$.

2. (Self-information, definition). A definition which satisfied the above desiderata is $I(\omega) = -\log(\mathrm{P}[\omega])$.

3. (Self-information, units). When $\log_2$ is used in the definition, the units of measurement of self-information are the *bits*. If $\ln \equiv \log_e$ is used, the self-information is measures in *nats*. For the decimal logarithm, I is measured in *hartley*.

4. (Entropy, definition). The *entropy* (or Shannon entropy) of a random variable is the expectation of its self-information denoted as $H(X) = \mathbb{E}[I(X)]$, where $I(X)$ is to be interpreted as follows: Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $X : (\Omega, \mathcal{F}, \mathrm{P}) \to \{x_i\}_{i=1}^n$ a finite-valued random variable. Consider the events $E_i = \{\omega \in \Omega \mid X(\omega) = x_i\}$ with self-information $I(E_i)$. Then, $I(X)$ is the random variable $I(X)(\omega) = I(E_{\iota(\omega)})$, where $\iota(\omega)$ is such that $X(\omega) = x_{\iota(\omega)}$.

   The entropy of X is given by

   $$H(X) = -\sum_{i=1}^{n} p_i \log(p_i),$$

   where $p_i = \mathrm{P}[X = x_i]$.

5. (Joint entropy). The *joint entropy* of two random variables X and Y (with values $\{x_i\}_i$ and $\{y_j\}_j$ respectively) is the entropy of the random variable (X, Y) in the product space, that is

   $$H(X, Y) = -\sum_{i,j} p_{ij} \log p_{ij},$$

   where $p_{ij} = \mathrm{P}[X = x_i, Y = y_j]$.

6. (Conditional Entropy).

7. (Mutual information).

5.2 KL divergence

1. (Definition/Discrete spaces). Let $(\Omega, \mathcal{F})$ be a discrete measurable space and P and P' two probability measures on it. The Kullback-Leibler (KL) divergence from P' and P is defined as[1]

   $$\mathrm{D}_{\mathrm{KL}}(\mathrm{P} \| \mathrm{P}') = -\sum_i \mathrm{P}_i \log\left({}^{\mathrm{P}'_i}\!/\!{}_{\mathrm{P}_i}\right) = \sum_i \mathrm{P}_i \log\left({}^{\mathrm{P}_i}\!/\!{}_{\mathrm{P}'_i}\right)$$

2. (Definition/Continuous spaces with PDFs). The KL divergence over a continuous probability space and for two probability measures P and P' with PDFs p and p' respectively is

   $$\mathrm{D}_{\mathrm{KL}}(\mathrm{P} \| \mathrm{P}') = \int_{-\infty}^{\infty} p(x) \log\left({}^{p(x)}\!/\!{}_{p'(x)}\right) \mathrm{d}x$$

[1] Lecture notes by S. Khudanpur available online at https://www.clsp.jhu.edu/~sanjeev/520.447/Spring00/I-divergence-properties.ps

3. (Definition/Continuous spaces). If P is absolutely continuous with respect to P', we define

$$\mathrm{D}_{\mathrm{KL}}(\mathrm{P}\|\mathrm{P}') = \int_{\Omega} \log\left({}^{\mathrm{dP}}/_{\mathrm{dP}'}\right) \mathrm{dP}.$$

4. (Nonnegative). The KL divergence is always nonnegative: $\mathrm{D}_{\mathrm{KL}}(\mathrm{P}\|\mathrm{P}') \geq 0$
5. (Pinsker's inequality). $d_{\mathrm{TV}}(\mathrm{P}, \mathrm{P}') \leq \sqrt{\frac{1}{2}\mathrm{D}_{\mathrm{KL}}(\mathrm{P}\|\mathrm{P}')}$

6 Risk

6.1 Risk measures

1. (Risk measures and coherency). Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $\mathcal{Z} = \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ for $p \in [1, \infty]$. A risk measure $\rho : \mathcal{Z} \to \overline{\mathrm{I\!R}}$ is called coherent if
   i. (Convexity). For $Z, Z' \in \mathcal{Z}$ and $\lambda \in [0,1]$, $\rho(\lambda Z + (1-\lambda)Z') \leq \lambda\rho(Z) + (1+\lambda)\rho(Z')$
   ii. (Monotonicity). For $Z, Z' \in \mathcal{Z}$, $\rho(Z) \leq \rho(Z')$ whenever $Z \leq Z'$ a.s.,
   iii. (Translation equi-variance). For $Z \in \mathcal{Z}$ and $C \in \mathcal{Z}$ with $C(\omega) = c$ for almost all ω (almost surely constant), it is $\rho(C + Z) = c + \rho(Z)$,
   iv. (Positive homogeneity). For $Z \in \mathcal{Z}$ and $\alpha \geq 0$, $\rho(\alpha Z) = \alpha\rho(Z)$.

2. (Conjugate risk measure). With every convex risk measure, we associate the conjugate risk measure $\rho^* : \mathcal{Z}^* \to \overline{\mathrm{I\!R}}$ defined as
$$\rho^*(Y) = \sup_{Z \in \mathcal{Z}} \left\{ \langle Z, Y \rangle - \rho(Z) \right\}.$$

3. (Biconjugate risk measure). With every convex risk measure, we associate the biconjugate risk measure $\rho^{**} : \mathcal{Z}^{**} \to \overline{\mathrm{I\!R}}$
$$\rho^{**}(Z) = \sup_{Y \in \mathcal{Z}^*} \left\{ \langle Z, Y \rangle - \rho^*(Y) \right\}.$$

4. (Dual representation). Let $\mathcal{Z} = \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ with $p \in [1, \infty)$. If ρ is lower semi-continuous, then $\rho = \rho^{**}$. In particular,
$$\rho(Z) = \sup_{Y \in \mathcal{Z}^*} \left\{ \langle Z, Y \rangle - \rho^*(Y) \right\} = \sup_{Y \in \mathfrak{A}} \left\{ \langle Z, Y \rangle - \rho^*(Y) \right\},$$
where $\mathfrak{A} = \operatorname{dom} \rho^*$.

5. (Acceptance set). The set $\mathcal{A}_\rho = \{X \in \mathcal{Z} : \rho(X) \leq 0\}$ is called the acceptance set of ρ. Several properties of ρ can be tested using its acceptance set.

6. (Monotonicity condition). If $Y \geq 0$ (almost surely) for every $Y \in \mathfrak{A}$, then and only then ρ is monotone.

7. (Translation equi-variance condition). If for every $Y \in \mathfrak{A}$ it is $\mathrm{I\!E}[Y] = 1$, then and only then, ρ is translation equi-variant.

8. (Positive homogeneity condition). If ρ is the support function of $\mathfrak{A}$, that is, $\rho(Z) = \sup_{Y \in \mathfrak{A}} \langle Y, Z \rangle$, then and only then it is positively homogeneous. $\mathfrak{A}$ is called the admissibility set of ρ.

9. (Coherency-preserving operations). Let ρ_1, ρ_2 be two risk measures on $\mathcal{Z}$. Then, the following risk measures are coherent
   i. $\rho(X) := \lambda_1\rho_1(X) + \lambda_2\rho_2(X)$, $\lambda_1, \lambda_2 \in \mathrm{I\!R}$ not both equal to 0
   ii. $\rho(X) = \max\{\rho_1(X), \rho_2(X)\}$

10. (Sub-differentiability). If $\rho : \mathcal{Z} \to \mathrm{I\!R}$ is real valued, convex and monotone, then it is continuous and sub-differentiable on $\mathcal{Z}$.

11. (Sub-differentials of risk measures). Let $\rho : \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P}) \to \overline{\mathrm{I\!R}}$, $p \in [1, \infty)$, be convex and lower semi-continuous. Then $\partial\rho(Z) = \arg\max_{Y \in \mathfrak{A}}\{\langle Y, Z \rangle - \rho^*(Z)\}$. If, additionally, ρ is positively homogeneous, then $\partial\rho(Z) = \arg\max_{Y \in \mathfrak{A}} \langle Y, Z \rangle$.

12. (Convexity of $\rho \circ F$). Let $F : \mathbb{R}^n \to \mathcal{Z}$ be a convex mapping[1] and ρ be a convex monotone risk measure. Then $\rho \circ F$ is convex.

13. (Directional differentiability). Let $\mathcal{Z} = \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ with $p \in [1, \infty)$, $F : \mathbb{R}^n \to \mathcal{Z}$ be a convex mapping and $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ be a convex monotone risk measure which is finite-valued and continuous at $\bar{Z} = F(\bar{x})$. Then, $\phi := \rho \circ F$ is directionally differentiable at $\bar{x}$, $\phi'(\bar{x}; h)$ is finite-valued for all $h \in \mathbb{R}^n$ and[2]
$$\phi'(\bar{x}; h) = \sup_{Y \in \partial\rho(\bar{Z})} \langle Y, f'(\bar{x}; h)\rangle$$

14. (Sub-differentiability of $\rho \circ F$). Let $\mathcal{Z}$, F, ρ, $\bar{x}$ and $\bar{Z}$ be as above. Define $\phi = \rho \circ F$. Then ϕ is sub-differentiable at $\bar{x}$ and
$$\partial\phi(\bar{x}) = \operatorname{cl} \bigcup_{\substack{Y \in \partial\rho(\bar{Z}) \\ F' \in \partial F(\bar{x})}} \langle Y, F'\rangle$$

15. (Continuity equivalences). Let $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ be a convex, monotone, translation equi-variant risk measure and $\mathcal{Z} = \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$. The following are equivalent[3]:
    i. ρ is continuous
    ii. ρ is continuous at a $X \in \operatorname{dom} \rho$
    iii. $\operatorname{int} \mathcal{A}_\rho \neq \varnothing$
    iv. ρ is lower semi-continuous and finite-valued ($\operatorname{dom} \rho = \mathcal{Z}$)

16. (Lipschitz continuity wrt infinity norm). Let $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ be a proper, convex, monotone, translation equi-variant risk measure. Then, for all $X, X' \in \operatorname{dom} \rho$
$$|\rho(X) - \rho(X')| \leq \|X - X'\|_\infty.$$

17. (Law invariance). A risk measure ρ is called law invariant if $\rho(Z) = \rho(Z')$ whenever Z and Z' have the same distribution.

18. (Fatou property #1). Let $\rho : \mathcal{L}^\infty \to \overline{\mathbb{R}}$ be a proper convex risk measure. The following are equivalent:
    i. ρ is $\sigma(\mathcal{L}^\infty, \mathcal{L}^1)$-lower semi-continuous
    ii. ρ has the Fatou property, i.e., $\rho(X) \leq \liminf_k \rho(X_k)$ whenever $\{X_k\}$ is essentially uniformly bounded (there is $Z \in \mathcal{L}^\infty$ so that $X_k \leq Z$ for all $k \in \mathbb{N}$) and $X_k \xrightarrow{p} X$.

19. (Law-invariant risk measures have the Fatou property)[4]. Let $\mathcal{L}^\Phi$ denote an Orlicz space[5]. Any proper, (quasi)convex, law-invariant risk measure $\rho : \mathcal{L}^\Phi \to \overline{\mathbb{R}}$ that is norm-lower semi-continuous has the Fatou property if and only if Φ is Δ_2.

20. (Kusuoka representations). Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a non-atomic space and let $\rho : \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P}) \to \overline{\mathbb{R}}$ be a proper lower semi-continuous law-invariant coherent risk measure. Then, there exists a set $\mathfrak{M}$ of probability measures on $[0, 1)$ so that
$$\rho(Z) = \sup_{\mu \in \mathfrak{M}} \int_0^1 \mathrm{AV@R}_{1-\alpha}(Z) \mathrm{d}\mu(\alpha),$$
where $\mathrm{AV@R}_{1-\alpha}$ is the average value-at-risk operator at level $1 - \alpha$ (defined in the next section).

[1] The mapping $F : \mathbb{R}^n \to \mathcal{Z}$ if for every $\lambda \in [0, 1]$ and $x, y \in \mathbb{R}^n$ it is $F(\lambda x + (1 - \lambda)y)(\omega) \leq \lambda F(x)(\omega) + (1 - \lambda)F(y)(\omega)$ for P-almost every ω.

[2] F maps a vector x to random variables, so it is $F(x)(\omega) = f(x, \omega)$. The directional derivative of f with respect to x along a direction h is $f'(\bar{x}; h)$ and it is a random variable. The scalar product here is defined as $\langle Y, f'(\bar{x}; h)\rangle = \int_\Omega Y(\omega) f'(\bar{x}; h)(\omega) \mathrm{d}\mathrm{P}(\omega)$.

[3] For a detailed discussion on continuity properties of risk measures, see D. Filipović and G. Svindland, "Convex risk measures on $\mathcal{L}^p$," Available online at: http://www.math.lmu.de/~filipo/PAPERS/crmlp.pdf.

[4] This result is rather involved. For a detailed presentation refer to the article E. Jouini, W. Schachermayer and N. Touzi, "Law invariant risk measures have the Fatou property," (Chapter) Advances in Mathematical Economics, 2006, Springer Japan.

[5] An Orlicz space is a function space which generalizes the $\mathcal{L}^p$ spaces. A Young function $\Phi : [0, \infty) \to [0, \infty)$ is a convex function with $\lim_{x \to \infty} \Phi(x) \to \infty$ and $\Phi(0) = 0$. Given a Young function Φ and a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, define $L^\Phi(\Omega, \mathcal{F}, \mathrm{P}) = \{X : \Omega \to \mathbb{R}, \text{measurable}, \mathbb{E}[\Phi(|X|)] < \infty\}$ This set is not necessarily a vector space. The vector space spanned by L^Φ is the Orlicz space $\mathcal{L}^\Phi(\Omega, \mathcal{F}, \mathrm{P})$. This space is equipped with the Luxembourg norm $\|X\|_\Phi = \inf\{\lambda > 0 : \mathbb{E}[\Phi(X/\lambda)] \leq 1\}$. We say that Φ has the Δ_2 condition if $\Phi(2t) \leq K\Phi(t)$ for some $K > 0$.

21. (Regularity in spaces with atoms). Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a space with atoms and $(\Omega, \mathcal{H}, \mathrm{P})$ be a uniform probability space so that $(\Omega, \mathcal{F}, \mathrm{P})$ is isomorphic to it. Let $\mathcal{Z} := \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ and $\hat{\mathcal{Z}} := \mathcal{L}^p(\Omega, \mathcal{H}, \mathrm{P})$, $p \in [1, \infty)$. Let $\hat{\rho} : \hat{\mathcal{Z}} \to \overline{\mathbb{R}}$ be a proper, lower semi-continuous, law invariant, coherent risk measure. We say that $\hat{\rho}$ is regular if there is a proper, lower semi-continuous, law invariant, coherent risk measure $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ so that $\rho_{|\hat{\mathcal{Z}}} = \hat{\rho}$.

22. (Zero risk). Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a non-atomic probability space. Let ρ be a proper, lower semi-continuous, coherent, law invariant risk measure. If $Z \in \mathcal{Z}$, $Z \geq 0$ a.s. then $\rho(Z) = 0$ if and only if $Z = 0$ a.s.

23. (Risk under conditioning). Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a non-atomic space and $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ be a proper convex lower semi-continuous law-invariant risk measure. Let $\mathcal{H}$ be a sub-σ-algebra of $\mathcal{F}$. Then, $\rho(\mathbb{E}\left[X \mid \mathcal{H}\right]) \leq \rho(X)$, for all $X \in \mathcal{Z}$ and $\mathbb{E}[X] \leq \rho(X)$.

24. (Interchangeability principle for risk measures). Let $\mathcal{Z} := \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ and $\mathcal{Z}' := \mathcal{L}_{p'}(\Omega, \mathcal{F}, \mathrm{P})$ with $p, p' \in [1, \infty]$. Let $F : \mathbb{R}^n \to \mathcal{Z}$, that is, for $x \in \mathbb{R}^n$, $F(x)$ is a random variable; let $(F(x))(\omega) = f(x, \omega)$. For a set $X \subseteq \mathbb{R}^n$ define $\mathfrak{M}_X := \{\chi \in \mathcal{Z}' : \chi \in X, \text{P-a.s.}\}$. Let $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ be a proper monotone risk measure. For $\chi \in \mathcal{Z}'$ define $F_\chi(\omega) = f(\chi(\omega), \omega)$ Suppose that $\inf_{x \in X} F(x) \in \mathcal{Z}$ and that ρ is continuous at $\inf_{x \in X} F(x)$. Then
$$\inf_{\chi \in \mathfrak{M}_X} \rho(F_\chi) = \rho\left(\inf_{x \in X} F(x)\right).$$

6.2 Popular risk measures

1. (Trivially coherent risk measures). The expectation operator and the essential supremum are coherent risk measures. For $\omega \in \Omega$, define $\rho(X) = X(\omega)$. This is a coherent risk measure, however, it is not law invariant.

2. (Mean-Variance measure). The mean-variance risk measure is defined as $\rho(X) = \mathbb{E}[X] + c\mathrm{Var}[X]$. This risk measure is law invariant, continuous, convex and translation equi-variant. However, it is neither monotone nor positively homogeneous.

3. (Value-at-Risk). The Vale-at-Risk of a random variable X at level α is the $(1-\alpha)$-quantile of X, that is, $\mathrm{V@R}_\alpha[X] = \inf\{t \in \mathbb{R} : \mathrm{P}[X > t] \leq \alpha\}$. $\mathrm{V@R}_\alpha$ is monotone, positively homogeneous and translation equi-variant, but non-convex and not sub-additive[6].

4. (Average Value-at-Risk). The Average Value-at-Risk is defined as[7]
$$\mathrm{AV@R}_\alpha[X] = \inf_{t \in \mathbb{R}} t + {}^1\!/\!_\alpha \mathbb{E}[X - t]_+.$$
This is a coherent law-invariant risk measure.

5. (Mean-Deviation of order p). Let $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$, $p \in [1, \infty)$ and $c \geq 0$. Define
$$\rho(X) = \mathbb{E}[X] + c\mathbb{E}\left[|X - \mathbb{E}[X]|^p\right]^{1/p}$$
This is a convex, translation equi-variant and positively homogeneous risk measure. It is monotone if $p = 1$, $(\Omega, \mathcal{F}, \mathrm{P})$ is non-atomic and $c \in [0, 1/2]$.

6. (Mean-Upper-Semideviation of order p). Let $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$, $p \in [1, \infty)$ and $c \geq 0$. The mapping
$$\rho(X) = \mathbb{E}[X] + c\mathbb{E}\left[[X - \mathbb{E}[X]]_+^p\right]^{1/p}$$
This is a convex, translation equi-variant and positively homogeneous risk measure. It is monotone if $p = 1$, $(\Omega, \mathcal{F}, \mathrm{P})$ is non-atomic and $c \in [0, 1]$.

[6] The Value-at-Risk is convex for certain classes of random variables. See A. I. Kibzun and E. A. Kuznetsov, “Convex Properties of the Quantile Function in Stochastic Programming,” Automation and Remote Control, Vol. 65, No. 2, 2004, pp. 184–192.

[7] We use the notation $[X] = \max\{X, 0\}$. We use the definition of Shapiro et al. Other authors use different definitions such as $\mathrm{AV@R}_\alpha[X] = \inf_{t \in \mathbb{R}} t + {}^1\!/\!_{1-\alpha} \mathbb{E}[X - t]_+$.

7. (Entropic risk measure). Let $\mathcal{Z} = \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$, $p \in [1, \infty]$. For $\gamma > 0$, define the entropic risk measure
$$\rho_\gamma^{\mathrm{ent}}(X) = {}^1\!/\!_\gamma \mathbb{E}[e^{\gamma X}].$$
For $p = \infty$, $\rho_\gamma^{\mathrm{ent}}$ is finite valued and w*-lower-semi-continuous. Moreover, $\rho_\gamma^{\mathrm{ent}}$ is convex, monotone and translation equi-variant, but not positively homogeneous. Furthermore, $\lim_{\gamma \to 0} \rho_\gamma^{\mathrm{ent}}(X) = \mathbb{E}[X]$ and $\lim_{\gamma \to \infty} \rho_\gamma^{\mathrm{ent}}(X) = \operatorname{esssup}[X]$.

8. (Entropic Value-at-Risk). The entropic value-at-risk at level $1 - \alpha$, $\alpha \in (0, 1]$ of a random variable X for which the moment generating function M_X exists is defined as[8]
$$\mathrm{EV@R}_{1-\alpha}[X] = \inf_{t>0} \{ \tfrac{1}{t} \ln(M_X(t)/\alpha \}.$$
The entropic value-at-risk is a coherent risk measure for all $\alpha \in (0, 1]$.

9. (Expectiles). Let $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathrm{P})$ and $\tau \in (0, 1)$. The τ-expectile of X is defined as
$$e_\tau(X) = \operatorname*{argmin}_{t \in \mathbb{R}} \mathbb{E}[\tau[X - t]_+^2 + (1 - \tau)[t - X]_+^2].$$
For all $\tau \in (0, 1)$, e_τ is a coherent risk measure.

10. (Generalizations of $\mathrm{AV@R}_\alpha$)[9]. Let $X \in \mathcal{Z} := \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ and $\phi : \mathcal{Z} \to \mathbb{R}_+$ be a function which is lower semi-continuous, monotone, convex and positive homogeneous. Then
$$\rho(X) = \inf_t \{ t + \phi(X - t) \},$$
is a coherent risk measure[10].

[8] The moment generating function (MGF) M_X of a random variable X is defined as $M_X(z) := \mathbb{E}[e^{zX}]$ for $z \in \mathbb{R}$. Not all random variables have an MGF (e.g., the Cauchy distribution does not define an MGF).

[9] These risk measures were first introduced by Ben-Tal and Teboulle; see for example A. Ben-Tal, M. Teboulle, "An oldnew concept of convex risk measures: an optimized certainty equivalent," Mathematical Finance 17 (2007) 449–476. These measures are discussed in: P. Krokhma, M. Zabarankin and S. Uryasev, "Modeling and optimization of risk," Surveys in Operations Research and Management Science 16 (2011) 49–66.

[10] In the case of $\mathrm{AV@R}_\alpha$, it is $\phi(X) = {}^1\!/\!_\alpha \mathbb{E}[X]_+$ which is indeed convex, monotone and translation equi-variant.

7 Uncertainty Quantification

7.1 Polynomial chaos

7.1.1 The Kosambi-Karhunen-Loève theorem

1. (Kernel function[1]). A kernel function is a symmetric continuous function $K : [a, b] \times [a, b] \to \mathbb{R}$. K is called positive semidefinite if $\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \leq 0$, for all scalars $(c_i)_{i=1}^{n}$ and $(x_i)_{i=1}^{n}$ in $[a, b]$ and all $n \in \mathbb{N}$.

2. (Hilbert-Schmidt integral operator). A kernel K is associated with its Hilbert-Schmidt integral operator which is defined as $T_K : \mathcal{L}^2[a, b] \to \mathcal{L}^2[a, b]$
$$T_K : \mathcal{L}^2[a, b] \ni \phi \mapsto T_K \phi = \int_a^b K(\,\cdot\,, s)\phi(s)\mathrm{d}s$$

3. (Mercer's theorem). Mercer's theorem offers a representation of kernel functions using a basis of $\mathcal{L}^2([a, b])$: Let K be a positive definite kernel. Then, there is an orthonormal basis $(e_i)_i$ of $\mathcal{L}^2([a, b])$ and a sequence of nonnegative coefficients $(\lambda_i)_i$ so that
$$K(s, t) = \sum_{j=1}^{\infty} \lambda_j e_j(s) e_j(t),$$
where convergence is absolute and uniform in $[a, b] \times [a, b]$ and $(e_i)_i$ and $(\lambda_i)_i$ are eigenfunctions and eigenvalues of T_K.

4. (Kosambi-Karhunen-Loève theorem). Let $(X_t)_{t \in T}$, $T = [a, b]$, be a centered, mean-square continuous stochastic process on $(\Omega, \mathcal{F}, \mathrm{P})$ with $X_t \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathrm{P})$ for all $t \in T$. Then, there is a basis $(e_i)_{i \in \mathbb{N}}$ or $\mathcal{L}^2(T)$ such that for all $t \in T$,
$$X_t = \sum_{i=1}^{\infty} \lambda_i e_i(t),$$
in $\mathcal{L}^2([a, b])$, where the random coefficients $\lambda_i \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathrm{P})$ are given by
$$\lambda_i(\omega) = \int_T X_t(\omega) e_i(t)$$
and satisfy $\mathbb{E}[\lambda_i] = 0$ and $\mathbb{E}[\lambda_i \lambda_j] = 0$ for $i \neq j$.

   In particular, $(e_i)_i$ and $(\lambda_i)_i$ are eigenvectors and eigenvalues of the Hilbert-Schmidt integral operator T_{R_X} of the auto-correlation function R_X of the random process, that is, they satisfy the Fredholm integral equation of the second kind
$$T_{R_X} e_i = \lambda_i e_i$$
   or equivalently
$$\int_a^b R_X(s, t) e_i(s) \mathrm{d}s = \lambda_i e_i(t),$$
   for all $i \in \mathbb{N}$.

[1] This should not be confused with *transition kernels*, which we discussed in Section 1.1.10.

5. (Corollary of KKL theorem[2].). Let $(X_t)_{t\in T}$, $T=[a,b]$, be a stochastic process which satisfies the requirements of the Kosambi-Karhunen-Loève theorem. Then, there exists a basis $(e_i)_i$ of $\mathcal{L}^2(T)$ such that for all $t\in T$
$$X_t(\omega)=\sum_{i=1}^{\infty}\sqrt{\lambda_i}\xi_i(\omega)e_i(t),$$
in $\mathcal{L}^2(\Omega,\mathcal{F},\mathrm{P})$, where ξ_i are centered, mutually uncorrelated random variables with unit variance and are given by
$$\xi_i(\omega)={}^1\!/\!\sqrt{\lambda_i}\int_a^b X_\tau(\omega)e_i(\tau)\mathrm{d}\tau.$$

7.1.2 Orthogonal polynomials

1. (Space of polynomials). Let $\mathsf{P}_N(I)$ denote the space of polynomials $\psi:I\to\mathbb{R}$ of degree $N\in\mathbb{N}$.
2. (Weighted $\mathcal{L}^2_w$ space). For two functions $\psi_1,\psi_2:I\to\mathbb{R}$, we define the scalar product
$$\langle\psi_1,\psi_2\rangle_w=\int_I w(\tau)\psi_1(\tau)\psi_2(\tau)\mathrm{d}\tau.$$
and the corresponding norm $\|f\|_2=\langle f,f\rangle_w$. Define the space
$$\mathcal{L}^2_w(I)=\{f:I\to\mathbb{R}\mid \|f\|_2<\infty\}.$$
For a random variable $\Xi:(\Omega,\mathcal{F},\mathrm{P})\to\mathbb{R}$ with PDF p_Ξ, we define
$$\langle\psi_1,\psi_2\rangle_\Xi:=\langle\psi_1,\psi_2\rangle_{p_\Xi}.$$
3. (Orthogonality wrt random variable). Let Ξ be a real-valued random variable with probability density function p_Ξ. Let $\psi_1,\psi_2:\Omega\to\mathbb{R}$ be two polynomials. We say that ψ_1,ψ_2 are orthogonal with respect to (the pdf of) Ξ if $\langle\pi_1,\pi_2\rangle_\Xi=0$.
4. Let $\psi_0,\psi_2,\ldots$, with $\psi_0=1$, be a sequence of orthogonal polynomials. Then,
$$0=\langle\psi_0,\psi_1\rangle_\Xi=\int_{-\infty}^{\infty}\psi_1(s)\underbrace{p_\Xi(s)\mathrm{d}s}_{\mathrm{d}(\Xi_*\mathrm{P})}=\mathbb{E}[\psi_1(\Xi)],$$
by virtue of LotUS. Recursively, $\mathbb{E}[\psi_i(\Xi)]=0$ for all i.
5. (Hermite polynomials). Let Ξ be distributed as $\mathcal{N}(0,1)$. Then, its pdf is $p_\Xi(s)=1$ and the polynomials $\psi_0,\psi_1,\ldots$ are the Hermite polynomials, the first few of which are $H_0(x)=1$, $H_1(x)=x$, $H_2(x)=x^2-1$, $H_3(x)=x^3-3x$. These are orthogonal with respect to Ξ, that is
$$\langle H_i,H_j\rangle_\Xi=\int_{-\infty}^{\infty}H_i(s)H_j(s)e^{-\frac{s^2}{2}}\mathrm{d}s=0,$$
for $i\neq j$.
6. (Legendre polynomials). If $\Xi\sim U([-1,1])$, then $\psi_0,\psi_1,\ldots$ are the Legendre polynomials. If, instead, $\Xi\sim U([-1,1])$, the coefficients of the Legendre polynomials can be modified.
7. (Laguerre polynomials). If the germ is an exponential random variable on $[0,\infty)$, then $\psi_0,\psi_1,\ldots$ are the Laguerre polynomials.
8. (Polynomial projection). Let $\mathcal{M}_N=\{\psi_i\}_{i=0}^N\subseteq\mathsf{P}_N(I)$ be a set of orthogonal polynomials $\psi_i:I\to\mathbb{R}$ with respect to the inner product $\langle\,\cdot\,,\,\cdot\,\rangle_w$. We define the projection operator onto $\mathcal{M}_N$ as
$$P_N:\mathcal{L}^2_w(I)\ni f\mapsto P_Nf:=\sum_{j=0}^{N}\hat{f}_j\psi_j\in\mathsf{P}_N,$$
where
$$\hat{f}_j={}^1\!/\!\|\psi_j\|^2\langle f,\psi_j\rangle_w.$$

[2] For applications of the KKL theorem to decompose stochastic processes, see http://amslaurea.unibo.it/10169/1/Giambartolomei_Giordano_Tesi.pdf

9. (Properties of polynomial projection). It is easy to see that $P_N f = f$ for all $f \in \mathsf{P}_N(I)$, whereas, for $g \perp \mathsf{P}_N(I)$ it is $P_N g = 0$.

10. (Best approximation). For $f \in \mathcal{L}^2_w(I)$,

$$\|f - P_N f\|_w = \inf_{\psi \in \mathsf{P}_N(I)} \|f - \psi\|_w$$

Additionally, the approximation error, $e_f := f - P_N f$ is orthogonal to $\mathsf{P}_N(I)$, that is, $\langle e_f, \psi_i \rangle_w = 0$ for all $i \in \mathbb{N}_{[0,N]}$.

11. (Approximation properties). For $f \in \mathcal{L}^2_w(I)$[3], define $f_N = P_N f$. Then,

$$\lim_{N \to \infty} \|f - f_N\|_w = 0.$$

12. (Strong/weak approximation). An approximation f_N which converges, as $N \to \infty$, to f as the above best approximation (via polynomial projection), is called a strong approximation. An approximation f_N which converges in a weaker sense (e.g., in probability) is called weak.

7.1.3 Generalized polynomial chaos expansions

1. (Polynomial chaos expansions). Given a random variable X, a polynomial chaos expansion is a function f so that for a given random variable Ξ with known distribution, called a "germ," it holds that $X \overset{d}{=} f(\Xi)$ (the notation $\overset{d}{=}$ means that the distributions of the two random variables are equal).

2. (Weak approximation theorem). Let X be a random variable of $\mathcal{L}^2(\Omega, \mathcal{F}, \mathrm{P})$ with cumulative distribution function F_X. Let Ξ be a germ in $\mathcal{L}^{2N}(\Omega, \mathcal{F}, \mathrm{P})$ with distribution F_Ξ and

$$\langle \psi_i, \psi_j \rangle_\Xi = \delta_{i,j} \gamma_i,$$

for all $i, j \in \mathbb{N}_{[0,N]}$. Let

$$X_N = \sum_{j=0}^N \alpha_j \psi_j(\Xi),$$

where

$$\alpha_j = {}^1\!/\!_{\gamma_j} \int_0^1 \psi_j(u) F_X^{-1}(F_\Xi(u)) \mathrm{d}F_\Xi(u),$$

Then, as $N \to \infty$, $X_N \to X$ in probability[4].

3. (Non-intrusive solution via linear regression).

4. (Non-intrusive solution via stochastic projection). Let X be a random variable on $(\Omega, \mathcal{F}, \mathrm{P})$ and $Y = f(X)$. Suppose we have obtained a truncated PC expansion for X. The question uncertainty propagates from X to Y via f; in other words, what is the distribution of Y. For $N \in \mathbb{N}$, let X_N be a PC expansion of X as follows

$$X_N = \sum_{j=0}^N x_j \psi_j(\Xi) = f_N(\Xi)$$

[3] In certain cases, we may derive approximation bounds. For example, for f in the weighted Sobolev space $H^p_w([-1,1]) = \{g : [-1,1] \to \mathbb{R} \mid {}^{\mathrm{d}^i g}\!/\!_{\mathrm{d}\tau^i} \in \mathcal{L}^2_w([-1,1]), i = 0, \ldots, p\}$, equipped with the inner product

$$\langle f, g \rangle_{H^p_w([-1,1])} = \sum_{j=0}^p \left\langle \frac{\mathrm{d}^j g}{\mathrm{d}\tau^j}, \frac{\mathrm{d}^j g}{\mathrm{d}\tau^j} \right\rangle_{\mathcal{L}^2_w([-1,1])}$$

and induced norm $\|f\|_{H^p_w([-1,1])} = \langle f, f \rangle^{1/2}_{H^p_w([-1,1])}$, and with $\mathcal{M}_N$ being the set of Legendre polynomials on $[-1,1]$, we have that there is a constant c, independent of N, so that $\|f - P_N f\| \leq {}^c\!/\!_N \|f\|_{H^p_w([-1,1])}$.

[4] X can be approximated by projecting on the space spanned by the orthogonal polynomials $\mathcal{M} = \{\psi_i\}_{i=0}^N$ leading to an approximation $X_N = \sum_{j=0}^N \alpha_j \psi_j(\Xi)$, where the coefficients α_j are computed by $\alpha_j = \langle X, \psi_j \rangle_\Xi / \gamma_j$. The problem is that the inner product $\langle X, \psi_j \rangle_\Xi$, typically, cannot be evaluated. The trick is that $X \overset{d}{=} F_X^{-1}(U)$, where U is a random variable which is uniformly distributed on $[0,1]$; one such variable is $U = F_\Xi(\Xi)$, therefore, $X \overset{d}{=} F_X^{-1}(F_\Xi(\Xi))$. This leads to the above formula. The integral can be evaluated by quadrature methods.

Let Y_N be the desired approximation (we take the approximation length to be the same and the orthogonal polynomial basis to be also the same), $Y_N = \sum_{j=0}^{N} y_j \psi_j(\Xi)$. It follows that

$$y_k = {}^{1}\!/\!_{\|\psi_k\|_\Xi^2} \cdot \langle \psi_k, \eta \circ f_N \rangle_\Xi = {}^{1}\!/\!_{\|\psi_k\|_\Xi^2} \cdot \int \eta(f_N(u))\psi(u)p_\Xi(u)\mathrm{d}u,$$

and the integral can be evaluated using a quadrature method, or even simple Monte Carlo, that is $\int \eta(f_N(u))\psi(u)p_\Xi(u)\mathrm{d}u \approx N_{\mathrm{mc}}^{-1} \sum_{i=1}^{N_{\mathrm{mc}}} \eta(f_N(u^{(i)}))\psi(u^{(i)})p_\Xi(u^{(i)})$ where $u^{(i)}$ are samples from the distribution of Ξ.

5. (Galerkin projection).

6.

8 Bibliography with comments

Bibliographic references including lecture notes and online resources with some comments:

1. R.G. Gallager. *Stochastic processes: theory for applications.* Cambridge University Press, 2013: A gentle introduction to stochastic processes suitable for engineers who want to eschew the mathematical drudgery. Following a short, but circumspect introduction to probability theory, the author discusses several processes such as Poisson, Gaussian, Markovian and renewal processes. Lastly, the book discusses hypothesis testing, martingales and estimation theory. Without doubt, an excellent introduction to the topic for the uninitiated.

2. Robert L. Wolpert. Probability and measure, 2005. Lecture notes: Lecture notes with a succinct presentation of some very useful results, but without many proofs. Available at https://www2.stat.duke.edu/courses/Spring05/sta205/lec/s05wk07.pdf.

3. Erhan Çinlar. *Probability and Stochastics.* Springer New York, 2011: A fantastic book for one's first steps in probability theory with emphasis on random processes, filtrations, Martingales, stopping times and convergence theorems, Poisson random measures, Lévy and Markovian processes and Brownian motion.

4. Olav Kallenberg. *Foundations of modern probability.* Springer, 1997: The definitive reference for researchers. In its 23 chapters it gives a circumspect overview of probability theory and stochastic processes; ideal for researchers in the field.

5. Onesimo Hernández-Lerma and Jean Bernarde Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria.* Springer, 1996

6. Bernt Øksendal. *Stochastic Differential Equations.* Springer Berlin Heidelberg, sixth edition, 2003: An amazing eye-opening book on stochastic differential equations and their aplications. It offers a very comprehensive presentation of the Brownian motion and Itô's integral. The exercises are an invaluable tool for assimilating the theory.

7. Karl Simgman. Lecture notes on stochastic modeling I, 2009: Lecture notes by K. Sigman, Columbia University, http://www.columbia.edu/~ks20/stochastic-I/stochastic-I.html.

8. David Walnut. Convergence theorems, 2011. Lecture notes: A short compilation of convergence theorems

9. S.R. Srinivasa Varadhan. Lecture notes on limit theorems, 2002: A lot of material on limit theorems starting from general measure theory, to weak convergence results, limits of independent sums, results for dependent processes with emphasis on Markov chains, a comprehensive introduction to martingales, stationary processes and ergodic theorems and some notes on dynamic programming. Available online at https://www.math.nyu.edu/faculty/varadhan/.

10. Zhengyan Lin and Zhidong Bai. *Probability Inequalities.* Springer, 2011: several interesting (elementary and advanced) inequalities on probability spaces.

11. Andrea Ambrosio. Relation between almost surely absolutely bounded random variables and their absolute moments, 2013: A short note at http://planetmath.org/sites/default/files/texpdf/38346.pdf showing that almost surely bounded RVs have all their moments bounded.

12. Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory.* SIAM, second edition, 2014: Excellent book on stochastic programming and the definitive reference for risk measures.

13. Anthony O'Hagan. Polynomial chaos: A tutorial and critique from a statisticians perspective, 2013. Available at http://tonyohagan.co.uk/academic/pdf/Polynomial-chaos.pdf: This article is written in a very intuitive manner, it is easy to follow. It seems that it targets applied scientists and practitioners, rather than mathematicians. It is a good read to understand the basics of polynomial chaos. The author questions certain aspects of polynomial chaos from a statistics standpoint.

14. Dongbin Xiu. *Numerical methods for stochastic computations: a spectral method approach.* Princeton University Press, 2010: A proper theoretical treatise on polynomial chaos and several other topics related to approximations of (multivariate) probability distributions.

15. M.S. Eldred. Recent advances in non-intrusive polynomial chaos and stochastic collocation methods for uncertainty analysis and design. In *50th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, California, USA, 2009

16. Thorsten Schmidt. Coping with copulas, 2006. https://www.researchgate.net/publication/228876267/download

17. Carlo Sempi. Introduction to copulas, 2011. The 33rd Finnish Summer School on Probability Theory and Statistics; available at http://web.abo.fi/fak/mnf/mate/gradschool/summer_school/tammerfors2011/slides_sempi.pdf: a very thorough presentation of copulas along with lots of theoretical results.

18. A. Kaintura, T. Dhaene, and D. Spina. Review of polynomial chaos-based methods for uncertainty quantification in modern integrated circuits. *Electronics*, 7(3):30, 2018: a not very rigorous review, but it offers an overview of basic properties of polynomial chaos expansions

About the author

I was born in Athens, Greece, in 1985. I received a Diploma in Chemical Engineering in 2007 and an MSc with honours in Applied Mathematics in 2009 from NTU Athens. In December 2012, I defended my PhD thesis titled “Modelling and Control of Biological and Physiological Systems” at NTU Athens. In January 2013 I joined the Dynamical Systems, Control and Optimization research unit at IMT Lucca as a post-doctoral Fellow. Afterwards, I worked as a post-doctoral researcher at ESAT, KU Leuven. I am currently a post-doctoral researcher at KIOS Center of Excellence, University of Cyprus. My research focuses on model predictive control and numerical optimization.

Web page: https://alphaville.github.io/