Probability Cookbook

Pantelis Sopasakis

January 8, 2018

Contents

Abstract

This document is intended to serve as a collection of important results in general probability theory and it can be used for a quick brush up or as a quick reference or cheat sheet, but not as primary tutorial material.

1 General Probability Theory

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

    i. f is measurable,

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

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

5. (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, $X^{-1}(B) \in \mathcal{F}$.

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

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

[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 nonatomic. However, we may often prove these results for discrete or finite spaces.

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

9. (Pushforward 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 *pushforward* 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 $B \in \mathcal{G}$.

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

$$\int X \mathrm{d}(F_*\mathrm{P}) = \int (X \circ F) \mathrm{dP}.$$

11. (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.

12. (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.

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

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

15. (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 $f \in M$.

16. (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.

17. (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

    i. $\phi_k \to X$, pointwise 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 pointwise increasing simple functions with these properties.

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

1.3 Limits

1.3.1 Limits of sequences of events

1. (Nested sequences and probabilities). Let $(E_n)_n$ be a nonincreasing 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 $\mathrm{P}[\limsup_n E_n] = 1$.

1.3.2 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 $f_n \to f$ pointwise 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 pointwise 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) [6]. 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) [5]. 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) [5]. 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) [5]. 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. $\mathbb{E}|X_k - X| \to 0$

1.4 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

$$\frac{\mathrm{d}(\nu + \mu)}{\mathrm{d}\lambda} = \frac{\nu}{\mathrm{d}\lambda} + \frac{\nu}{\mathrm{d}\lambda}, \quad \lambda\text{-a.e.}$$

4. (Chain rule). If $\nu \ll \mu \ll \lambda$,

$$\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}, \quad \lambda\text{-a.e.}$$

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

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

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 pushforward.

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

   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

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

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

1.5 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. (Pushforward). The probability distribution of a random variable X with values in $(\mathcal{X}, \mathcal{G})$, is the pushforward 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. 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. $F_X^{-1}(p) \leq x \Leftrightarrow p \leq F_X(x)$

1.6 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 $Y(\omega) = g(X(\omega))$. Then

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

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.

1.8 $\mathcal{L}^p$ 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 seminormed 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 $\|[X]\|_p := \|X\|_p$.

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

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

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 hypograph, 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, $X(\omega) = \int_0^\infty 1_E(\omega, x) \mathrm{d}x$.

$$\begin{aligned} \mathbb{E}[X] = \int_\Omega X \mathrm{dP} &= \int_\Omega \int_0^\infty 1_E(\omega, x) \mathrm{d}x \mathrm{dP} \\ &= \int_0^\infty \int_\Omega 1_E(\omega, x) \mathrm{dP} \mathrm{d}x \\ &= \int_0^\infty \mathrm{P}[X \geq x] \mathrm{d}x. \end{aligned}$$

1.10 Transition Kernels

1. (Definition). Let $(X, \mathcal{F})$, $(Y, \mathcal{G})$ be two measurable spaces and let $K : X \times \mathcal{G} \to \overline{\mathbb{R}}$. K is called a *transition kernel* if

   i. $f_B(x) := K(x, B)$ is $\mathcal{F}$-measurable for every $B \in \mathcal{G}$,

   ii. $\mu_x(B) := K(x, B)$ is a measure on $(Y, \mathcal{G})$ for every $x \in X$.

2. (Existence of transition kernels). Let μ be a finite measure on $(X, \mathcal{F})$ and $k : X \times Y \to \mathbb{R}_+$ be measurable in the product σ-algebra $\mathcal{F} \otimes \mathcal{G}$. Then,

$$K(x, B) = \int_B k(x, y) \mu(\mathrm{d}y),$$

   is a transition kernel.

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

2 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 6 in Sec. 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. 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 $\mathrm{Var}[X]$.

4. 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.

5. If X takes positive integer values, then

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

6. (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.

7. (Absolutely bounded a.s. $\Leftrightarrow$ Bounded moments) [9]. 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 $k \in \mathbb{N}_{\geq 1}$

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

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

3 Conditioning

3.1 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 $\mathbb{E}[X \mid \mathcal{H}]$, with

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

which equivalently can be written as

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

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

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

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

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

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

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

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 pushforward $Y_*\mathrm{P}$). 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). $X \leq Y \Rightarrow \mathbb{E}[X \mid \mathcal{H}] \leq \mathbb{E}[Y \mid \mathcal{H}]$
   ii. (Positivity). $X \geq 0 \Rightarrow \mathbb{E}[X \mid \mathcal{H}] \geq 0$ [Set $Y = 0$ in 9i].
   iii. (Linearity). For $a, b \in \mathbb{R}$, $\mathbb{E}[aX + bY \mid \mathcal{H}] = a\mathbb{E}[X \mid \mathcal{H}] + b\mathbb{E}[Y \mid \mathcal{H}]$
   iv. (Monotone convergence). $X_n \geq 0$, $X_n \uparrow X$ implies $\mathbb{E}[X_n \mid \mathcal{H}] \uparrow \mathbb{E}[X \mid \mathcal{H}]$
   v. (Fatou's lemma). For $X_n \geq 0$, $\mathbb{E}[\liminf_n X_n \mid \mathcal{H}] \leq \liminf_n \mathbb{E}[X_n \mid \mathcal{H}]$
   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). $X_n \to X$ (pointwise) 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

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

3.2 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 $\mathrm{P}[H] > 0$.

4 Inequalities on Probability Spaces

4.1 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 $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$.

4.2 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 $X 1_{X>0}$ 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

$$\mathrm{P}[E_1 \vee \ldots \vee E_n] \geq \frac{(\sum_{i=1}^{n} \mathrm{P}[E_i])^2}{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{P}[E_i \wedge E_j]}$$

4.3 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

$$\mathrm{P}[\max_{1\leq i\leq n} |S_i| \geq 3\alpha] \leq \max_{1\leq i\leq n} \mathrm{P}[|S_i| \geq \alpha].$$

5 Convergence of random processes

5.1 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 semicontinuous f bounded from above
iv. $\liminf_k \mathbb{E}_{\mu_k} f \geq \mathbb{E}_\mu f$ for every lower semicontinuous 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).

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

8. (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.

5.3 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).

5. (Almost surely convergent subsequence). If $X_n \overset{p}{\to} X$, 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 $\mathrm{P}[\limsup_n \{|X_n| > \epsilon\}] = 1$.

$$\mathcal{L}_\infty \implies \mathcal{L}_p \implies \mathcal{L}_{p'},\ p' \in [1,p)$$

$$\Downarrow$$

$$\text{a.s.} \implies \text{p} \implies \text{d} \implies \varphi\,(\text{pointwise})$$

$$\Downarrow$$

$$\text{a.s.}\,(\overset{d}{\sim})$$

Figure 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 $Y \overset{d}{\sim} X$).

5.4 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 $X_n = n1_{[0,1/n]}$. 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

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

   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 $t < 2$.

5.5 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)[6]. 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 $\varphi_k \to \varphi$ pointwise. 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).[7] 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)[8]. 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 $n \to \infty$.

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

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

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

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.

5.7 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

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

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)[9]:

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

then,

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

6 Stochastic Processes

6.1 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 random variables on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$.

2. (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.

[9] 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.

3. (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 $(\mathcal{F}_t)_{t \in \mathbf{T}}$.

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

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

6.2 Martingales

1. (Martingale). 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 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) 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.

   (c) 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.

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

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

4. (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)matringale, then X is a (sub)matringale too.

5. (Martingale stopping).

6. (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 $\mathcal{F}_\infty = \sigma(\mathcal{F}_n, n \geq 0)$.

7. (Kolmogorov's submartingale inequality). Let $\{X_k\}_k$ be a nonnegative submartingale. 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 supermartingale. Then, for $n \in \mathbb{N}_{>0}$ and $\alpha > 0$, $\mathrm{P}[\cup_{k \geq n}\{Z_k \geq \alpha\}] \leq \mathbb{E}[Z_n]/\alpha$.

8. (Azuma-Hoeffding inequality for martingales with bounded differences). Let $(X_i)_i$ be a martingale or a supermartingale 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 submartingale,

$$\mathrm{P}[X_N - X_0 \leq -t] \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{N} c_i^2}\right)$$

6.3 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 $V \in \mathcal{G}_t^\infty$, $\mathbb{E}[V \mid \mathcal{F}_t] \in \sigma X_t$.

7 Information Theory

7.1 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).

7.2 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[10]

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

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

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

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). $d_{\mathrm{TV}}(\mathrm{P}, \mathrm{P}') \leq \sqrt{\tfrac{1}{2}\mathrm{D}_{\mathrm{KL}}(\mathrm{P}\|\mathrm{P}')}$

8 Theory of Risk

8.1 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 $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ 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 equivariance). 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 $\rho^* : \mathcal{Z}^* \to \overline{\mathbb{R}}$ 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 $\rho^{**} : \mathcal{Z}^{**} \to \overline{\mathbb{R}}$

$$\rho^{**}(Z) = \sup_{Y \in \mathcal{Z}^*} \left\{\langle Z, Y\rangle - \rho^*(Y)\right\}.$$

4. (Dual representation). Let $\mathcal{Z} = \mathcal{L}^p(\Omega, \mathcal{F}, \mathrm{P})$ with $p \in [1, \infty)$. If ρ is lower semicontinuous, 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.

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

7. (Translation equivariance condition). If for every $Y \in \mathfrak{A}$ it is $\mathbb{E}[Y]=1$, then and only then, ρ is translation equivariant.

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. $\rho(X) := \lambda_1\rho_1(X)+\lambda_2\rho_2(X)$, $\lambda_1,\lambda_2\in\mathbb{R}$ not both equal to 0

   ii. $\rho(X)=\max\{\rho_1(X),\rho_2(X)\}$

10. (Subdifferentiability). If $\rho:\mathcal{Z}\to\mathbb{R}$ is real valued, convex and monotone, then it is continuous and subdifferentiable on $\mathcal{Z}$.

11. (Subdifferentials of risk measures). Let $\rho:\mathcal{L}^p(\Omega,\mathcal{F},\mathrm{P})\to\overline{\mathbb{R}}$, $p\in[1,\infty)$, be convex and lower semicontinuous. 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[11] 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[12]
$$\phi'(\bar{x};h)=\sup_{Y\in\partial\rho(\bar{Z})}\langle Y,f'(\bar{x};h)\rangle$$

14. (Subdifferentiability 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 subdifferentiable at $\bar{x}$ and
$$\partial\phi(\bar{x})=\mathrm{cl}\bigcup_{\substack{Y\in\partial\rho(\bar{Z})\\ F'\in\partial F(\bar{x})}}\langle Y,F'\rangle$$

15. (Continuity equivalences). Let $\rho:\mathcal{Z}\to\overline{\mathbb{R}}$ be a convex, monotone, translation equivariant risk measure and $\mathcal{Z}=\mathcal{L}^p(\Omega,\mathcal{F},\mathrm{P})$. The following are equivalent[13]:

   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 equivariant 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

[11] 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.

[12] 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 $\langle Y,f'(\bar{x};h)\rangle=\int_\Omega Y(\omega)f'(\bar{x};h)(\omega)\mathrm{d}\mathrm{P}(\omega)$.

[13] 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.

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)[14]. Let $\mathcal{L}^\Phi$ denote an Orlicz space[15]. 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 nonatomic 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).

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 semicontinuous, law invariant, coherent risk measure. We say that $\hat{\rho}$ is regular if there is a proper, lower semicontinuous, 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 nonatomic probability space. Let ρ be a proper, lower semicontinuous, 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 nonatomic 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, $\rho(\mathbb{E}[X \mid \mathcal{H}]) \leq \rho(X)$, 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

$$\inf_{\chi \in \mathfrak{M}_X} \rho(F_\chi) = \rho\left(\inf_{x \in X} F(x)\right).$$

25. (Interchangeability principle).

8.2 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 equivariant. 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 equivariant, but nonconvex and not subadditive[16].

[14] 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.

[15] 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 Luxemburg 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 $K > 0$.

[16] 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.

4. (Average Value-at-Risk). The Average Value-at-Risk is defined as[17]

$$\mathrm{AV@R}_\alpha[X] = \inf_{t\in\mathbb{R}} t + {}^1\!/\!_\alpha \mathbb{E}[X-t]_+.$$

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 equivariant and positively homogeneous risk measure. It is monotone if $p = 1$, $(\Omega, \mathcal{F}, \mathrm{P})$ is nonatomic 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 equivariant and positively homogeneous risk measure. It is monotone if $p = 1$, $(\Omega, \mathcal{F}, \mathrm{P})$ is nonatomic and $c \in [0, 1]$.

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 equivariant, 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[18]

$$\mathrm{EV@R}_{1-\alpha}[X] = \inf_{t>0}\{\tfrac{1}{t}\ln(M_X(t)/\alpha\}.$$

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 $\mathrm{AV@R}_\alpha$)[19]. 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 semicontinuous, monotone, convex and positive homogeneous. Then

$$\rho(X) = \inf_t \{t + \phi(X-t)\},$$

is a coherent risk measure[20].

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

[18] 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 Cachy distribution does not define an MGF).

[19] 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.

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

9 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 Kallnberg. *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. 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`.

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

7. 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/`.

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

9. 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.

10. 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.