Probability theory

Kalle Heinonen

This is my self study on propability theory assuming knowledge on measure theory on the taking a course on Lebesgue measure and integration. Based on `https://math.aalto.fi/~kkytola/files_KK/ProbaTh2019/ProbaTh-2019.pdf` notes by Kalle Kytölä and Probability with martingales by David Williams. Much more compact version but with scratched solutions to the exercises. If you want to read study this, please do no look at the solutions. Main things written out is just for my own reality check as a student of mainly topology. In general assume $0 \notin \mathbb{N}$ in this text even though it is not my usual personal convention.

Contents

1 Basics in terms of measures

- Let Ω be a sample space. An **outcome** ω of a random experiment represents a single realization of randomness involved. Omega is then realized as space of all outcomes.

- An **event** is a subset $E \subset \Omega$ of possible outcomes. Te event E is said to occur if the randomly realized outcome $\omega \in \Omega$ belongs to E. Generally we cannot allow every subset of Ω as an event, but a suitable collection $\mathcal{F}$ of subsets for consistant rules (measurability issues).
- To each event E we can assign a **probability** (measure) $\mathbf{P}[E]$ of the event, which has valeus in $[0,1]$. We will call the triple $(\Omega, \mathcal{F}, \mathbf{P})$ a propability space (as a measure space).
- A **random variable** is a suitable (measurable) function $X \colon \Omega \to S$ taking $\omega \mapsto X(\omega) \in S$. One may think of this as choosing the outcome as random and determining at some value $X(\omega)$.
- The **expected value** of a real valueted $X \colon \Omega \to S \subset \mathbb{R}$ represents the average of possible values of X over all randomness, weighted according to probabilityies $\mathbf{P}$. In the sense of Lebesgue integration $\mathbf{E}[X] = \int_\Omega X(\omega) d\mathbf{P}(\omega)$.

The following set theoretical interpretations are used in probability

- The whole sample space Ω is a sure event.
- The empty set $\emptyset$ is impossible event.
- Intersection of two events is and.
- Union is or.
- Compliment $E^c = \Omega / E$ is "event E does not occur".
- $E_1 \subset E_2$ "Occurance of E_1 implies E_2."

These things seem to mimic the definition of a sigma algebra. Why sigma algebra? Because we want to sum over propabilities and countable summing happens in nature (toss a coin as many times as u like). Meaning $\Omega \in \mathcal{F}$, if $E \in \mathcal{F}$, then $E^c \in \mathcal{F}$ and $\cup_{\mathbb{N}} E_n \in \mathcal{F}$ for countable unions. The countability lets us sum over probabilities when viewed as measures. The definition of a sigma algebra implies that empty set is contained in there and countable intersection is there. Hence sigma algebra is stable under all countable set operations.

Lemma 1. *Intersection of of sigma algebras $\cap_{i \in I} \mathcal{F}_i$ is a sigma algebra.*

Proof is trivial. This leads to a corollary that sigma algebra generated by any $\mathcal{C} \subset \mathcal{P}(\Sigma)$ is well defines and unique. Defining $\sigma(\mathcal{C}) := \bigcap_{\mathcal{F}, \mathcal{C} \subset \mathcal{F}} \mathcal{F}$ one can easily check that it suffices.

One can check that Borel sigma algebra $\mathcal{B}(\mathbb{R}^d)$ on $\mathbb{R}^d$ can be generated by multiple different types of collections of boxes.

Intuition why we want to look at generators is that some times we want to examine less events for the full information.

1.1 Measure theoretic definitions

Definition 1. A **measurable** space is a pair $(S, \mathcal{F})$ where S is a set and $\mathcal{F}$ a sigma algebra on S.

Definition 2. A (positive) **measure** μ on a measure space $(S, \mathcal{F})$ is a function $\mu\colon \mathcal{F} \to \mathbb{R} \cup \{+\infty\}$ such that $\mu(\emptyset) = 0$ and for disjoint countable collection $\mu(\bigcup_{i=1}^{\infty} A_n) = \sum_{i=1}^{\infty} \mu(A_n)$. Measure is a **probability measure** if in addition $\mu(S) = 1$. Often instead of symbnol μ, one chooses $\mathbf{P}$. A **measure space** is measurable space equipped with a measure.

From now on **a propability space** will be the triple $(\Omega, \mathcal{F}, \mathbf{P})$. Often the underlying sigma-algebra is assumed known and omitted with presense of a measure. The event Ω is a **sure event**: it contains all possible outcomes. By definition also $\mathbf{P}(\Omega) = 1$. Conversly if $\mathbf{P}(E) = 1$ does not impply $E = \Omega$ since it depends on the measure. If $E \subsetneq \Omega$, then it is an **almost sure** event. Often stated $\mathbf{P}-$almost sure(ly) in contrast within measure theory $\mu-$almost everywhere.

Example: Let Ω be a finite set with the sigma algebra being the powerset. Then $\mathbf{P}(E) = \#E/\#\Omega$ defines a probability measure called the **(discrete) uniform probability measure**.

Exercises (Truncation of measures and conditioning of probability measures): Let μ be a measure on $(S, \mathcal{F})$ and $B \in \mathcal{F}$. Show that $A \mapsto \mu(A \cap B)$ is a measure. Sol: Empty set condition is trivial. If (A_n) is a disjoint collection, then so is $(A_n \cap B)$ so additivity holds.

Let $\mathbf{P}$ be a probability measure on $(\Omega, \mathcal{F})$ and $B \in \mathcal{F}$ an even for which $\mathbf{P}(B) > 0$. Show that the conditional probability $A \mapsto \mathbf{P}(A|B) := \mathbf{P}(A \cap B)/\mathbf{P}(B)$ is a probability measure on $(\Omega, \mathcal{F})$. Sol: Again empty case is trivial, and by monotonicity of measures it evaluates on $[0, 1]$. The full event also clearly becomes one.

If we take $\Omega = \mathbb{R}$ and truncate by $[0, 1]$ we get a propability measure from the Lebesgue measure of $\mathbb{R}$. We call this the **uniform probability measure** on the unit interval. This generalizes to any interval, but you have to then normalize by the length.

1.2 Probability distributions on countable spaces

Many probabilistic models convern the sets $\mathbb{N}$ and $\mathbb{Z}$. We can characterize probability measures in an intuitive way using mass functions. For rest of this subsection assume Ω is non-empty countable. Therefore we have an enumeration $\Omega = \{\omega_1, \omega_2, \dots\}$. We can sum over this enumeration $\sum_{\omega\in\Omega} a(\omega) := \sum_j a(\omega_j)$. If terms $a(\omega_j) \geq 0$ are non-negative the resulting sum doesn't depend on enumeration.

Definition 3. A **probability mass function** (p.m.f) on Ω is a function $p\colon \Omega \to [0, 1]$ for which $\sum_{\omega\in\Omega} p(\omega) = 1$ and each term is non-negative.

To each probablity mass function we can associate a propability measure $\mathbf{P}(E) = \sum_{\omega\in E} p(\omega)$ for all $E \subset \Omega$. Conversly to a propability measure $(\Omega, \mathcal{P}(\Omega))$ it is natural to associate masses of singleton events $p(\omega) = \mathbf{P}(\{\omega\})$ for all $\omega \in \Omega$.

Example: (Poisson distribution) Let $\lambda > 0$ and recall $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$, hence we can define $p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$ as a probability mass function on $\mathbb{N}$. Then as above this defines a probability measure on $\mathbb{N}$.

(Geometric distribution) let $0 < q < 1$ and using the series $\sum_{i=0}^{\infty} r^k = 1/(1-r)$ with $r = 1-q$, it is easy to see that $p(k) = (1-q)^{k-1}q$ for $k = 1, 2, 3, \dots$ is a probability mass function on $\mathbb{N} \setminus 0$. The geometric distribution with parameter q is the probability measure on $\mathbb{N} \setminus 0$ with the above probability mass function.

(Binomial distribution) Let $n \in \mathbb{N}$ and $q \in (0,1)$. Using the binomial formula

$$\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = (a+b)^n$$

with $a = q$ and $b = 1-q$ it is easy to see that the function p given by

$$p(k) = \binom{n}{k} q^k (1-q)^{n-k} \quad \text{for } k \in \{0, 1, \dots, n-1, n\} \tag{II.5}$$

is a probability mass function on the finite set $\{0, 1, \dots, n-1, n\}$.

The binomial distribution with parameters n and q is the probability measure on the finite set $\{0, 1, \dots, n-1, n\}$ with the above probability mass function.

1.3 Recalling measure theoretic facts

Lemma 2. *Let μ be a measure on $(S, \mathcal{F})$. Then*

- *μ is monotonic, meaning $A \subset B \Rightarrow \mu(A) \le \mu(B)$.*
- *(Monotone convergence) Let $A_1 \subset A_2 \subset \cdots$ be increasing sequence on measurable sets. Then the limit $A_n \uparrow A = \bigcup_{j=1}^{\infty} A_j$ of sets constitute the increasing limit $\mu(A_n) \uparrow \mu(A)$. Not that for general measures theres no downwards monotone convergence.*
- *Countable subadditivity for non disjoint measurable sets $\mu(\bigcup_n A_n) \le \sum_n \mu(A_n)$.*

So for probabilty this means that for a sequence of events $E_1, E_2, \dots$, $\mathbf{P}(\bigcup_{j=1}^{\infty} E_j) \le \sum_{j=1}^{\infty} \mathbf{P}(E_j)$. So probability that at least one event in a sequence occurs can not exceed the sum of probabilities of the events in the sequence.

Excercise: Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Show that for any event $E \in \mathcal{F}$ we have $\mathbf{P}(E^c) = 1 - \mathbf{P}(E)$. Sol: write $\Omega = E^c \cup E$ which is a disjoint union. Then applying $\mathbf{P}$, we get $1 = \mathbf{P}(E) + \mathbf{P}(E^c)$.

For any two events $\mathbf{P}(E \cup E') = \mathbf{P}(E) + \mathbf{P}(E') - \mathbf{P}(E_1 \cap E_2)$. Sol: Write as disjoint union so $E \cup E' = E \cup (E \setminus (E \cap E'))$. Thus $\mathbf{P}(E \cup E') = \mathbf{P}(E) + \mathbf{P}(E \setminus (E \cap E'))$. Also we can write $E' = E \setminus (E \cap E') \cup (E \cap E')$ which is again disjoint, so $\mathbf{P}(E') = \mathbf{P}(E \setminus (E \cap E')) + \mathbf{P}(E \cap E')$. Plugging in we get that $\mathbf{P}(E \cup E') = \mathbf{P}(E) + \mathbf{P}(E \setminus (E \cap E')) = \mathbf{P}(E) + \mathbf{P}(E') - \mathbf{P}(E \cap E')$.

Counter example to downwards going monotone sequence. Take the set $\mathbb{N} \setminus 0$ with the counting measure on the power set sigma-algebra. Take $A_n = \{n, n+1, n+2, \cdots\}$ so $A_1 \supset A_2 \supset \cdots$ so the intersection limit is empty. But the sequence of counting measures $\mu(A_n) = +\infty$ does not tend to $\mu(A) = 0$. For the Lebesgue measure on $\mathbb{R}$ we can take the intervals $[n, \infty)$ and deduce similarly.

However for probability measures theres a catch. Also the decreasing sequences converge monotonely in measure. Proof: Let $E_1 \supset E_2 \supset \cdots$ be decreasing sequence of events with limit $\bigcap_{i=1}^{\infty} E_i$. Now $\mathbf{P}(\bigcap_{i=1}^{n} E_i) \leq \mathbf{P}(\bigcap_{i=1}^{n-1} E_i)$ so it forms a decreasing sequence of reals where the ambient most set has measure 1. Therefore the monotone convergence theorem applies since our starting point is not infinite measure so a decreasing sequence bounded below by 0.

1.4 Identifying and consutrcting measures

Does a measure with desired properties even exist (can we construct it)? How can we check that two measures are the same? What does one need to know to distinquish measures. The first question is more in the realm of pure measure theory, when do functions from set algebras extend to measures (Carathéodory extension theorem). The latter is relevant in stochastic and statistical reasoning.

Collection of following type are used for purposes of identifying measures.

Definition 4. ($\pi-$system) A follection $\mathcal{F}$ of subsets of S is called a **pi-system** if the following holds:

$$A, B \in \mathcal{F} \Rightarrow A \cap B \in \mathcal{F}.$$

Notably sigma-algebras are pi-systems, but clearly not all conversely.

Example: The set $\mathcal{F}(\mathbb{R}) := \{(-\infty, x] | x \in \mathbb{R}\}$ is clearly a pi-system. But with previous remarks, this system also generates the Borel sigma-algebra on $\mathbb{R}$. It is one of the simplest pi-systems and will be used over and over again especially when dealing with real valued random variables.

(Dynkin's indentification) Let $\mathbf{P}_1, \mathbf{P}_2$ be probability measures on the measurable space $(\Omega, \mathcal{F})$. Assume that $\mathcal{J}$ is a pi-system on Ω for which $\sigma(\mathcal{J}) = \mathcal{F}$ (sigma-algebra generated by $\mathcal{J}$). Then the following statements are equivalent

1. $\mathbf{P}_1(E) = \mathbf{P}_2(E)$ for all $E \in \mathcal{J}$
2. the two probability measures are equal, $\mathbf{P}_1 = \mathbf{P}_2$.

Clearly 2 implies 1 but the other way requires some machinery.

Definition 5. (D-system) A collection $\mathcal{D}$ of subsets of S is said to be a $d-$system on S if it satisfies the following:

(D1) $S \in \mathcal{D}$

(Dd) if $A, B \in \mathcal{D}$ and $A \subset B$, then $B \setminus A \in \mathcal{D}$.

(D↑) if A_n is increasing sequence indexed by $\mathbb{N}$ in $\mathcal{D}$, then the limit $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{D}$.

Theorem 1. *A collection $\mathcal{F}$ of subsets of S is a sigma-algebra iff it is a $d-$system and a pi-system.*

Proof. It is clear that a sigma algebra is a $d-$system and a pi-system. Suppose converse for $\mathcal{F}$. Clearly $\Omega \in \mathcal{F}$ and $\mathcal{F}$ is closed under set complement. We are left to obverser countable unions. For two sets A_1 and A_2 we have that $A_1^c \cap A_2^c \in \mathcal{F}$ by axiom Dd and being a pi-system. Hence $A_1 \cup A_2 = S \backslash (A_1^c \cap A_2^c)$ by de Morgan's laws. By induction one can deduce finite unions. Let (A_n) be a countably infinite family and define the sequence $B_n = A_1 \cup \dots A_{n-1} \cup A_n$. By D↑ we have $\bigcup_{n\in\mathbb{N}} B_n \in \mathcal{D}$, but this union is precicely the union $\bigcup_{n\in\mathbb{N}} A_n$. □

Definition 6. ($D-$system generated by a collection) The $d-$system $d(\mathcal{J})$ generated by a collection of subsets $\mathcal{J}$ of S is the smallset $d-$system that contains $\mathcal{J}$.

One can check just like for generated sigma-algebras that this is well-defined.

Lemma 3. *(Dynkin) Suppose that $\mathcal{J}$ is a pi-system on S. Then $d(\mathcal{J}) = \sigma(\mathcal{J})$.*

Proof. Since any sigma-algebra is a $d-$system $d(\mathcal{J}) \subset \sigma(\mathcal{J})$. Other inclusion is done in two steps. First we show that whenever $B \in d(\mathcal{J})$ and $C \in \mathcal{J}$, we have that $B \cap C \in d(\mathcal{J})$. Define the collection of sets with this property

$$\mathcal{D}_1 = \{B \in d(\mathcal{J}) | B \cap C \in d(\mathcal{J}), \forall C \in \mathcal{J}\}.$$

We then wish to show that $\mathcal{D}_1 = d(\mathcal{J})$. By construction $\mathcal{D}_1 \subset d(\mathcal{J})$. Since $\mathcal{J}$ is a pi-system, $\mathcal{J} \subset \mathcal{D}_1$. Since $d(\mathcal{J})$ is the smallest $d-$system containing $\mathcal{J}$, by showing $\mathcal{D}_1$ is a $d-$system, we get that $\mathcal{D}_1 = d(\mathcal{J})$. Firstly D1 holds since $S \cap C = C \in \mathcal{J} \subset d(\mathcal{J})$. For Dd take $A, B \in \mathcal{D}_1, A \subset B, C \in \mathcal{J}$. Then $(B \backslash A) \cap C = (B \cap C) \backslash (A \cap C) \in d(\mathcal{J})$ since $d(\mathcal{J})$ is a $d-$system. For D↑ take an increasing family $(A_n) \subset \mathcal{D}_1$. Thus

$$\left(\bigcup_{n\in\mathbb{N}} A_n \right) \cap C = \left(\bigcup_{n\in\mathbb{N}} A_n \cap C \right) \in d(\mathcal{J})$$

by again $d(\mathcal{J})$ being a $d-$system. Hence we conclude $\mathcal{D}_1 = d(\mathcal{J})$. Secondly we want to show that $A, B \in d(\mathcal{J})$ implies $A \cap B \in d(\mathcal{J})$. Define

$$\mathcal{D}_2 = \{A \in d(\mathcal{J}) | A \cap B \in d(\mathcal{J}), \forall B \in d(\mathcal{J})\}.$$

We wish to do similar argument as above to show $\mathcal{D}_2 = d(\mathcal{J})$, but the steps are basically identical. In conclusion $d(\pi)$ is a pi-system, therefore it is a sigma-algebra. □

Proof of non trivial part of Dynkin's identification theorem: Define the colletion

$$\mathcal{D} = \{E \in \mathcal{F} | \mathbf{P}_1(E) = \mathbf{P}_2(E)\}$$

of the events where the equality holds. We must then show that $\mathcal{D} = \mathcal{F}$. First we show that $\mathcal{D}$ is a $d-$system on Ω. The property D1 is satisfied by definition of probability measure. Let $A, B \in \mathcal{D}$ and $A \subset B$. Writing $B = A \cup (B \setminus A)$ one gets $\mathbf{P}_i(B) = \mathbf{P}_i(A) + \mathbf{P}_i(B \setminus A)$ so $\mathbf{P}_i(B \setminus A) = \mathbf{P}_i(B) - \mathbf{P}_i(A)$ for $i = 1, 2$. Thus

$$\mathbf{P}_1(B \setminus A) = \mathbf{P}_1(B) - \mathbf{P}_1(A) = \mathbf{P}_2(B) - \mathbf{P}_2(A) = \mathbf{P}_2(B \setminus A).$$

Lastly for D↑ take an increasing sequence $(A_n) \subset \mathcal{D}$ with limit A. Then

$$\mathbf{P}_1(A) = \lim \uparrow \mathbf{P}_1(A_n) = \lim \uparrow \mathbf{P}_2(A_n) = \mathbf{P}_2(A).$$

Thus $\mathcal{D}$ is a $d-$system and by generative property $d(\mathcal{J}) \subset \mathcal{D}$. By Dynkin's lemma $d(\mathcal{J}) = \sigma(\mathcal{J})$ and $\sigma(\mathcal{J}) = \mathcal{F}$ by assumption, so $d(\mathcal{J}) = \sigma(\mathcal{J})$. Therefore $\mathcal{F} \subset \mathcal{D}$. Therefore $\mathbf{P}_1(E) = \mathbf{P}_2(E)$ for all events $E \in \mathcal{F}$. □

In the followin we consider a probability measure on $\mathbb{R}$. Typically such a measure could appear as the law of real-valued random variable as we will discuss later. In order to have a suitable notion for such common situations, denote a probability measure by ν instead of $\mathbf{P}$.

Definition 7. (Cumulative distribution function) If ν is a probability measure on the Borel sets of $\mathbb{R}$, the the **cumulative distribution function** (c.d.f) of ν is the function $F \colon \mathbb{R} \to [0,1]$ defined by $F(x) := \nu((-\infty, x])$.

An imporant application of Dynkin's identification is the following.

Corollary 1. *If ν_1, ν_2 are probability measure on Borel sets of $\mathbb{R}$, and F_1, F_2 the respective c.d.f's, then following are equivalent:*

1. $F_1 = F_2$

2. $\nu_1 = \nu_2$.

Proof. Assuming the measures are equal, we get 1 by definition. If $F_1 = F_2$ consider the pi-system $\mathcal{J}(\mathbb{R})$ on Borel sets considered previously of half opens $(-\infty, x]$. Since by definition the c.d.f's agree on the $\pi-$system, by Dynkin identification they agree on the whole Borel sigma-algebra since $\sigma(\mathcal{J}(\mathbb{R})) = \mathcal{B}(\mathbb{R})$. □

This means that cumulative distribution functions characterize probability measures on $\mathcal{B}(\mathbb{R})$. Next we ask what kind of functions qualify as c.d.f's.

Theorem 2. *(Properties of c.d.f's) If $F \colon \mathbb{R} \to [0,1]$ is a c.d.f of a probability measure ν on $\mathcal{B}(\mathbb{R})$, then it satisfies the following properties:*

(a) *F is increasing function $x \leq y \Rightarrow F(x) \leq F(y)$*

(b) *F is right-continuous $x_n \downarrow x$ as $n \to \infty$, then $F(x_n) \downarrow F(x)$*

(c) $\lim_{x \to +\infty} F(x) = 1$, $\lim_{x \to -\infty} F(x) = 0$.

Proof. (a) follows from monotonicity of measures. Next let (x_n) be downwards monotonic, then $(-\infty, x_n]$ is decreasing with limit $(-\infty, x]$. For probability measures we have downward monotonic convergence, so (b) holds. Part (c) follows from upwards and downwards monotinic convergence where the other limit is empty and other limit is the entire real line. □

2 Random variables

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. The key idea of a random variable is the following two step procedure by which randomness is thought to have effect.

1. "Chance determines the random outcome $\omega \in \Omega$."
2. "The outcome ω determines various qunatities of interest." (random variables)

Therefore a random variable will be a function of the form $X \colon \Omega \to S'$ where S' is a suitable set of possible values of the quantity of interest. We want this function to behave sufficiently well so we can talk about probabilities with which the quantity assumes certain values. So whanever $A' \subset S'$ is a reasonable enough subset of the possible values, the set of outcomes ω for which $X(\omega) \in A'$ should constitute an event, .i.e, $\{\omega \in \Omega | X(\omega) \in A'\} \in \mathcal{F}$. This is the preimage $X^{-1}(A')$ and in measure theoretic terms, we want X to be measurable map (preimages of measurable sets are measurable). Often we abuse notation and denote $\{X \in A'\} \subset \Omega$ with a probabilistic interpretation "the value of our (random) quantity of interest X lies in A'" of the event becomes apparent at a glance.

Examples:

Random numbers: Target space of random variable real line with Borel sets.

Random vectors: Target space of random variable real $n-$space with Borel sets.

Random matrices: Target space of random variable real $n \times m-$matrices with Borel sets.

Random graphs: Set S' of some graphs with a suitably chosen sigma-algebra (often just the power set $\mathcal{P}(S')$).

Usually both the sample Ω space and target space S' are atleast topological spaces so we can equip them with Borel sigma-algebras. Then we call a Boreal measurable functions **Borel functions**.

Example that is not just a simple Euclidean space but a topological space: Brownian motion is one of the most important stochastic processes: it is a continuous real valued Markov process in continuous time, which is used to model many things from termal motion of microscopic particles to stock prices. Mathematically the Bronian motion on the unit time interval is a random variables taking values in $C([0,1])$ with sup-norm and Borel sigma-algebra.

The law of random variable

Suppose that $X \colon \Omega \to S'$ is a random variable. Then there is a probability measure on S' which describes how the values of the random variable are distributed.

Definition 8. (The law of random variable) The **law** (or the **distribution**) of the random variable $X \colon \Omega \to S'$ is the probability measure P_X on $(S', \mathcal{F}')$ defined by $P_X(A') := \mathbf{P}(X^{-1}(A'))$, for $A' \in \mathcal{F}'$.

To verify that this is a probability measure it is clear that $P_X(S') = 1, P_X(\emptyset) = 0$. Since X is measurable it preserves measurable sets and preimage preserves unions and disjoint sets. Therefore a probability measure. This is also called the pushforward measure in measure theory.

Again with slight abuse of notation one can write $P_X[A'] = \mathbf{P}[X \in A']$.

2.1 Indicator random variables

Constant functins are trivial examples of random variables. They have only one possible value. Arguably the next simples example of a random variable would be one whicih assumes one of the two possible values (depending on the outcote). It is convinient ot take 0 and 1. Then we speak of the indicator function $\mathbb{I}_E$ which is 1 on $\omega \in E$ and 0 on $\omega \in E^c$.

Exercise: $\mathbb{I}_E$ is measurable iff $E \in \mathcal{F}$. Since $\{1\}$ is a Borel set $\mathbb{I}_E^{-1}(1) = E$ has to be measurable if $\mathbb{I}_E$ is measurable.

Whence $\mathbb{I}_E$ is measurable we call it **indicator random variable**. Clearly $\mathbb{I}_{E\cap E'} = \mathbb{I}_E \cdot \mathbb{I}'_E$ and for disjoint E, E', $\mathbb{I}_{E\cup E'} = \mathbb{I}_E + \mathbb{I}_{E'}$.

2.2 Constructing random variables

At this stage one might therefore still worry that perhaps the requirements of a measurable function are too stringent for any interesting examples to exist. . . Fortunately, this is not the case: almost all functions that you ever encounter turn out to be measurable.

A very easy case

If $\mathcal{F} = \mathcal{P}(\Omega)$ i.e. all subsets are events, then every function out of Ω to some other measure space is measurable.

The above situation is reasonable usually when Ω is countable. We want to see common non trivial random variables in uncountable cases as well.

Proposition 1. *Composition of measurable maps are measurable.*

For a practical verification of measurability we only have to look at a generating set of a sigma-algebra.

Lemma 4. *Let $\mathcal{J}' \subset \mathcal{P}(S')$ be a collection of subsets that generate sigma-algebra $\mathcal{F}'$. Then a function $f\colon (S, \mathcal{F}) \to (S', \mathcal{F}')$ is measurable iff $f^{-1}C' \in \mathcal{F}$ for all $C' \in \mathcal{J}'$.*

Proof. The condition is clearly necessary for measurability. Assume hence that $f^{-1}C' \in \mathcal{F}$ for all $C' \in \mathcal{J}'$ and show f is measurable. Define

$$\mathcal{G}' = \{G' \in \mathcal{F}' | f^{-1}(G') \in \mathcal{F}\}$$

the collection of "good" subsets G' of S' whose preimages are measurable. By assumption $\mathcal{J}' \subset \mathcal{G}'$. Since

$$f^{-1}(S') = S, f^{-1}((G')^c) = (f^{-1}(G'))^{-1}, f^{-1}\left(\bigcup_{n=1}^{\infty} G'_n\right) = \bigcup_{n=1}^{\infty} f^{-1}(G'_n)$$

and $\mathcal{F}$ a sigma-algebra on S, we see that collection $\mathcal{G}'$ is a sigma-algebra on S'.

The fact that sigma-algebra $\mathcal{G}'$ contains the collection $\mathcal{J}'$, means that $\sigma(\mathcal{J}') \subset \mathcal{G}'$. We assumed that $\sigma(\mathcal{J}') = \mathcal{F}'$ meaning $\mathcal{F}' \subset \mathcal{G}'$. Thus every $\mathcal{F}'-$measurable subset is $\mathcal{G}'$ measurable. This means that every $A' \in \mathcal{F}'$ has the property that $f^{-1}A' \in \mathcal{F}$ so f is measurable. □

Corollary 2. *Let X, X' be topological spaces with Borel sigma-algebra. Then a continuous function $f\colon X \to X'$ is Borel-measurable.*

Proof. Since the open sets generate entire $\mathcal{B}(X')$ and we know continuity by definition preserves opennes under preimage. Then by previous lemma f is measurable. □

Corollary 3. *A function $f\colon S \to \mathbb{R}$ (Borel sets on $\mathbb{R}$) is measurable iff for all $c \in \mathbb{R}$, $\{f \leq c\}$.*

Proof. The sets $(-\infty, c]$ generate $\mathcal{B}(\mathbb{R})$. Apply the previous lemma. □

In general a real valued measurable map will be just called a function.

Proposition 2. *Real valued random variables are preserved under pointwise addition, multiplication and $\mathbb{R}-$scalar multiplication.*

Given a sequence $(f_n\colon S \to \mathbb{R})_n$ of measurable function we can define pointwise supremum $(\sup_n f_n)(s)$ and infimum, $(\inf_n f_n)(s)$ over $\mathbb{N}$ with values posisbly in $[-\infty, \infty]$. We equip the extended real line with a topology of a closed interval so that there is a homeomorphism given by $\tan\colon [-\pi/2, \pi/2] \to [-\infty, \infty]$ with sigma-algebra of Borel sets. Similaryly we can define pointwise $\limsup, \liminf$.

Proposition 3. *Pointwise* $\sup, \inf, \limsup, \liminf$ *of a sequence of measurable extended real valued functions is measurable.*

Proof. We can write $\{\sup_n f_n \leq c\} = \bigcap_n \{f_n \leq c\}$ over all $\mathbb{N}$ for the sup case. Infimum case is similar. For limit superior $\limsup_n f_n(s) = \inf_n(\sup_{k \geq n} f_n(s))$ and for inferior similar type form and argue from measurability of sup and inf from first part. □

Corollary 4. *The pointwise limits of measurable functions are measurable.*

Proof. Write limits with a limit superior to / limit inferior. □

Let us visit the example of repeated coin tossing. The previous operations let us construct rather nontrivial random variables starting from very basic ones.

Example: Take $\Omega = \{H, T\}^{\mathbb{N}}$ as the sample space, let $\mathcal{F}$ be the sigma-algebra generated by the events $E_j := \{\omega \in \Omega | \omega(j) = H\} = j$th coin toss are heads. Then the indicator $\mathbb{I}_{E_j}(\omega) = 1$ if $\omega(j) = H$ and 0 if $\omega(j) = T$. Since E_j is an event, the indicator in this case is measurable, i.e. a random variable. By pointwise summing and scalar multiplication properties of measurable functions, the relative frequency of heads $X_n(\omega) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}_{E_j}(\omega)$ is measurable. Also the upper and lower limits $L^+(\omega) = \limsup_n X_n(\omega), L^-(\omega) \liminf_n X_n(\omega)$ are random variables. Knowing that these are random variables, do we learn something interesting about events? For example take $r \in [0, 1]$, we should hope to be abled to form the event "relative frequencies of heads tend to r" $= \{\lim_{n\to\infty} X_n = r\}$ which in more careful notation meants $\{\omega \in \Omega | \lim_{n\to\infty} X_n(\omega) = r\} \subset \Omega$. Now we can write $\{\lim_{n\to\infty} X_n = r\} = \{L^+ = r\} \cap \{L^- = r\}$ where the both sets in the intersection are measurable as $(L^{\pm})^{-1}(r) = \{L^{\pm} = r\}$, hence the limit is also measurable. Once we introduce a probability measure it will be thus meaningful to talk about relative frequency approaching a particular limit.

The above can be done tought without talking about random variables. But this method has an advantage. Consider a slightly modified "relative frequencies have a limit" $= \{\exists \lim_{n\to\infty} X_n\}$. To show that it is an even we notice that $= \{\exists \lim_{n\to\infty} X_n\} = \{L^+ - L^- = 0\}$ which can easily be concluded measurable. It would be more cubersome to conclude this by direct manipulation events of using just countable set operations.

2.3 Simple random variables

Just like in measure theory a simple random variable is a $\mathbb{R}-$linear combinations of indicator random variables. So indicator functions over measurable sets. We can always choose a linear combination so that the indicator function sets are all disjoint measurable sets. Sometimes choosing this minimal combination is neat, but othertimes we might not want to insist disjointness. As we saw previously that pointwise limit of measurable functions is measurable. An important converse is also true, any measurable function can be obtained as a limit of simple functions.

Let $(S, \mathcal{F})$ be a measurable space, denote set of measurable functions $f\colon S \to [0, +\infty]$ by $m\mathcal{F}^+$ in sense of Borel sigma-algebra on the target.

Lemma 5. *(Approximation of non-negative measurable functions). Let $f \in m\mathcal{F}^+$. There there exists a sequence (f_n) of non-negative simple functions such that $f_n \uparrow f$ pointwise as $n \to \infty$.*

We will construct certain "staircases" which become more tiny and longer to prove this lemma. We define them as follows

$$s_n : [0, +\infty] \to [0, n]$$

$$s_n(x) = \begin{cases} 0 & \text{if } 0 \le x \le \frac{1}{2^n} \\ \frac{1}{2^n} & \text{if } \frac{1}{2^n} < x \le \frac{2}{2^n} \\ \frac{2}{2^n} & \text{if } \frac{2}{2^n} < x \le \frac{3}{2^n} \\ \vdots & \\ \frac{n2^{n-1}}{2^n} & \text{if } \frac{n2^{n-1}}{2^n} < x \le n \\ n & \text{if } n < x. \end{cases}$$

Lemma 6. *(Properties of staircases)*

(a) The staircase functions s_n are simple, Borel-measurable and left-continuous.

(b) For every $x \in [0, +\infty]$, we have $s_n(x) \uparrow x$ as $n \to \infty$.

Proof. By definition the image of each s_n is $\{j/2^n | j = 0, \cdots, n/2^n\}$ and preimage of each point is an interval which are Borel measurable simple function. Left continuity follows from the fact that we can pick any x and there is close enough $x' < x$ such that $s_n(x') = s_n(x)$ are on the same interval.

For the other part consider case finite and infinite. If $x = +\infty$, then $s_n(x) = n$ which has limit $+\infty$ as n grows. Take $x < +\infty$ and large enough n such that $x < n$. Then $|s_n(x) - x| \leq 1/2^n \to 0$ as n grows and the sequence is monotone. $\square$

Proof of the approximation lemma: Take $f \in m\mathcal{F}^+$ and define a monotone sequence by $f_n := s_n \circ f$ of measurable functions. Now we know that $f_n(x) = s_n(f(x)) \uparrow f(x)$ by the previous lemma as n grows large. T

In the proof it was not really imporant if the staircases were constructed left or right-continuous. Where left-continuity is actually convinient is the proof of monotone convergence theorem. Given $g_1, g_2, \dots \in m\mathcal{F}^+$ such that $g_n \uparrow g$ pointwise as n grows, we may construct simple approximations $g_n^r = s_r \circ g_n$ where $g_n^r \uparrow g_n$ pointwise as well as $g^r = s_r \circ g$ with $g^r \uparrow g$ as r grows large pointwise. Therefore by left continuity $g_n^r(x) = s_r(g_n(x)) \uparrow s_r(g(x)) = g^r(x)$ as n grows large.

3 Information generated by random variables.

Probability theory offers an important interpretation of sigma-algebras: they describe information. Firstly the elementary notion of independence of events is generalized to the notion of independence of information. Another common and fruitful use of information is conditional expected value, which represents the best estimate of a random number given some (partial) information about it. Finally, stochastic processes are random time-dependent phenomena, and it is often relevant to model how information accumulates as we observe the phenomenon over a period of time — the mathematical notion suitable for this is refining collections of sigma-algebras indexed by time known as filtrations. There would be yet other con- texts, but it is in fact useful to interpret all sigma-algebras as describing information, and relate the notion of measurability of functions to this interpretation. We start with an informal description of these ideas and proceed to definitions and properties.

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space with indexed collection of real valued random variables $Y_\gamma \colon \Omega \to \mathbb{R}$, $\gamma \in \Gamma$ (real valued for concreteness). To understand what is the stored information, recall the two step procedure by which randomness is thought to arise:

1. "Chance determines the random outcomes $\omega \in \Omega$"

2. "The outcome ω determiens the values $Y_\gamma(\omega)$ of quantities of the interest Y_γ".

Hence the motimation about information is then: "If you do not know the outcome ω of all randomness, but someone tells you the values of the quantities of interest Y_γ for all $\gamma \in \Gamma$, then for which events $E \in \mathcal{F}$ are you abled to decide wheter E occurs or not?" By this formulation the information contained in the collection (Y_γ) of random variables is some collection of events. Name those events whose occurance can be decided based on the random variables. Evindently any event of type $\{Y_\gamma \in A'\}$ concerning the value of any one of the random variables Y_γ can be decided, and thus belongs to the collection. Hoever, in deciding about events you are furthermore allowed to use logical reasoning (e.g, set theoretic interpretation like in the beginning), so the collection of events that can be decided should be a sigma-algebra itself. This motivates the following.

3.1 Definition of sigma-algebra generated by random variables

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $(Y_\gamma)_{\gamma\in\Gamma}$ collection of real random variables.

Definition 9. (Sigma-algebra generated by random variables) The sigma-algebra generated by a collection of random variables $(Y_\gamma)_{\gamma\in\Gamma}$ is the smallest sigma-algebra $\mathcal{J}$ such that for each $\gamma \in \Gamma$, Y_γ is $\mathcal{J}-$measurable. We denote the generated sigma-algebra by $\mathcal{J} = \sigma((Y_\gamma)_{\gamma\in\Gamma})$.

One can easily show the existance and uniqueness like we have shown previously. Since by definition each Y_γ is $\mathcal{F}-$measurable, $\sigma((Y_\gamma)_{\gamma\in\Gamma}) \subset \mathcal{F}$. According to the information interpretation, $\mathcal{F}$ represents "full information" (all events in our space), so no amount of random variables could contain more information than that. Using the same $\sigma(\cdots)$ notation for two types of generation will be always clear from context which type of generation it is.

An already interesting case is collection of one random variable. Then we denote the sigma-algebra generated by $\sigma(Y)$ for $Y\colon \Omega \to \mathbb{R}$.

Exercise: Show that the sigma-algebra $\mathcal{J} = \sigma(Y)$ coincides with $\sigma(Y^{-1}(\mathcal{B}(\mathbb{R}))$ where $Y^{-1}(\mathcal{B}(\mathbb{R})) = \{Y^{-1}(B)|B \in \mathcal{B}(\mathbb{R})\}$. Show that in fact $\sigma(Y) = Y^{-1}(\mathcal{B}(\mathbb{R}))$.

Sol: Take any $A \in \sigma(Y)$, then it can be written as countable union or intersection of preimages of borel sets. Since preimage preserves unions and intersections, we can just talk about preimages of any Borel set and vise versa. Because Borel sets are a sigma-algebra, the preimage is also a sigma algebra. Thus $\sigma(Y) \subset Y^{-1}(\mathcal{B}(\mathbb{R}))$ by being smallest. But we also know that preimage of any borel set is in $\sigma(Y)$ so equality also must hold.

Exercise: (A pi-system to generate the sigma-algebra generated by a random number). Let Y be a random variable as previously and $\mathcal{J}(\mathbb{R}) = \{(-\infty, x]|x \in \mathbb{R}\}$ the pi-system that generates Borel sets. Define $\mathcal{J} = Y^{-1}(\mathcal{J}(\mathbb{R})) := \{Y^{-1}((-\infty, x])|x \in \mathbb{R}\}$. Show that $\mathcal{J}$ is a pi-system that generates $\sigma(Y)$.

Sol: It is evidently clear that since preimages preserve intersections, $\mathcal{J}$ is a pi-system. Since $\sigma(Y)$ is the preimage of the entire Borel sigma algebra and $\mathcal{J}(\mathbb{R}) \subset \mathcal{B}(\mathbb{R})$ then $\mathcal{J} \subset \sigma(Y) \Rightarrow \sigma(\mathcal{J}) \subset \sigma(Y)$. Write again $\sigma(Y) = \sigma(Y^{-1}(\mathcal{B}(\mathbb{R})))$ because $Y^{-1}(\mathcal{B}(\mathbb{R})) \subset \sigma(\mathcal{J})$, then $\sigma(Y) = \sigma(Y^{-1}(\mathcal{B}(\mathbb{R}))) \subset \sigma(\mathcal{J})$.

3.2 Doob's representation theorem

Definition of information contained in random variables may seem abstract. The following theorem offers a nice interpretation for the information contained in a realvalued random variable.

Theorem 3. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $Y, Z\colon \Omega \to \mathbb{R}$ two real valued random variables. Then Z is $\sigma(Y)-$measurable iff there exists a Borel function $f\colon \mathbb{R} \to \mathbb{R}$ such that $Z = f(Y)$.*

This theorem gives the following interpretation: to say that Z is measurable with respect to information contained in Y means that the value of Z could be obtained from the value of Y by applying some deterministic function f.

In the proof one uses "the monotone class theorem".

Definition 10. A collection $\mathcal{H}$ of bounded functions from S to $\mathbb{R}$ is said to be a **monotone class** if it satisfies the following conditions:

(MC-1) Constant $1 \in \mathcal{H}$

(MC-$\mathbb{R}$) $\mathcal{H}$ is a $\mathbb{R}-$vectorspace with function addition.

(MX-$\uparrow$) If $f_1, f_2, \dots \in \mathcal{H}$ is an increasing sequence of non-negatice functions in $\mathcal{H}$ such that the pointwise limit f is a bounded function, then $f \in \mathcal{H}$.

Theorem 4. *(Monotone class theorem) Suppose that $\mathcal{H}$ is a monotone class of bounded functions from S to $\mathbb{R}$. Let $\mathcal{J}$ be a pi-system on S. If $\mathcal{H}$ contains the indicator function $\mathbb{I}_A$ of every $A \in \mathcal{J}$, then $\mathcal{H}$ contains all bounded $\sigma(\mathcal{J})/\mathcal{B}-$measurable functions.*

Proof. Let $\mathcal{H}$ be our monotone class of bounded functions. Define $\mathcal{D}$ as the collection of subsets $A \subset S$ whose indicator function $\mathbb{I}_A \in \mathcal{H}$. Now we show that $\mathcal{D}$ is a $d-$system. The constant $\mathcal{H} \ni 1 = \mathbb{I}_S$ implies $S \in D$. Let $A \subset B, A, B \in \mathcal{D}$ and consider $\mathbb{I}_{B \setminus A} = \mathbb{I}_{A^c}\mathbb{I}_B \in \mathcal{H}$ because $\mathbb{I}_{A^c} = 1 - \mathbb{I}_A \in \mathcal{H}$. Now consider increasing sequence $A_1, A_2, \dots$, this gives an induced monotone sequence of indicators, whose limit is in $\mathcal{H}$.

Now assume as in the statement that $\mathcal{J}$ is a pi-system such that $\mathcal{H}$ contains the indicator function of each member of $\mathcal{J}$. Thus $\mathcal{J} \subset \mathcal{D}$ and by $\mathcal{D}$ being a $d-$system and Dynkin's lemma, $\sigma(\mathcal{J}) = d(\mathcal{J}) \subset \mathcal{D}$. I.e. every indicator function of sets in $\sigma(\mathcal{J})$ are in $\mathcal{D}$. Thus $\sigma(\mathcal{J})$ is also in the monotone class $\mathcal{H}$. Thus a bounded simple $\sigma(\mathcal{J})/\mathcal{B}-$measurable f can be obtained from $\mathbb{R}-$linear combinations simple functions in $\sigma(\mathcal{J})$. Thus $\mathcal{H}$ contains all such functions.

Let $f\colon S \to \mathbb{R}$ be non-negative $\sigma(\mathcal{J})/\mathcal{B}-$measurable bounded function. By the approximation by simples lemma, f can be obtained by upwards monotone sequence of simples pointwise. We observer above that such simples are in $\mathcal{H}$ so thus $f \in \mathcal{H}$.

Finally write a general $\sigma(\mathcal{J})/\mathcal{B}-$measurable function $f = f_+ - f_-$, where $f_+ = \max(f, 0), f_- = \max(-f, 0)$. As this is a $\mathbb{R}-$linear combination, $f \in \mathcal{H}$. □

Proof of Doob's representation theorem: Assume that $Z = f \circ Y$ for some Borel function $f\colon \mathbb{R} \to \mathbb{R}$ then composition of measurables preserves measurability because Y is by definition $\sigma(Y)-$measurable. To show the other direction we first assume that Z is bounded and then the full generality.

Bounded Z: Let $\mathcal{H}$ define the collection of bounded functions $Z\colon \Omega \to \mathbb{R}$ which can be written as $Z = f \circ Y$ for some Borel function f. Out goal is to show that

$\mathcal{H}$ contains all $\sigma(Y)-$measurable functions. Denote $\mathcal{J} = \sigma(Y)$. We first check that atleast the indicator of any $E \in \mathcal{J}$ is in $\mathcal{H}$. From an exercise we could write $\sigma(Y) = Y^{-1}(\mathcal{B})$ i.e. any set can be written as a preimage of Borel sets. Then $\mathbb{I}_E = \mathbb{I}_{Y^{-1}B}$. It is well known that $\mathbb{I}_B \colon \mathbb{R} \to \mathbb{R}$ is bounded Borel and $\mathbb{I}_E = \mathbb{I}_B \circ Y$.

Next we show that $\mathcal{H}$ is a monotone class. Trivially $1 \in \mathcal{H}$. Take $Z, Z' \in \mathcal{H}$ and look at so $Z = f \circ Y, Z' = f' \circ Y$. The function $f + f'$ is Borel and $(f + f') \circ Y = f \circ Y + f' \circ Y = Z + Z'$ hence $Z + Z \in \mathcal{H}$. For scalar multiplication just multiply the Borel function by the scalar for similar argument. Lastly suppose that $Z_n \uparrow Z, 0 \leq Z_n \leq K$ for a constant $K < \infty$ as n large with $Z_n \in \mathcal{H}$ for all n. So we can write $Z_n = f_n \circ Y$ for Borel f_n's. Define $f = \limsup_n f_n$, then f is Borel measurable and bounded between 0 and K. We can assume this bound on f because we can truncate f_n's by $\tilde{f}_n(y) = \begin{cases} K & \text{if } f_n(y) > K \\ f_n(y) & \text{if } 0 \leq f_n(y) \leq K \\ 0 & \text{if } f_n(y) < 0 \end{cases}$ so still $Z_n = \tilde{f}_n \circ Y$. This is because image of Y may not be entire $\mathbb{R}$ so we can ignore elements not in the image. Thus it follows that $Z = \lim_{n\to\infty} Z_n = \lim_{n\to\infty} f_n \circ Y = \limsup_n f_n \circ Y = f \circ Y$. Thus $\mathcal{H}$ must be a monotone class. The monotone class theorem then implies that $\mathcal{H}$ contains all bounded $\sigma(Y)-$measurable functions.

Unbounded Z: Let $\mathbb{Z} \colon \Omega \to \mathbb{R}$ now be unbounded. In that case apply $\arctan \colon \mathbb{R} \to (-\pi/2, \pi/2)$ and get the bounded random variable $\tilde{Z} = \arctan \circ Z$. By the composition property this is $\sigma(Y)-$measurable since arctan is Borel. From the first part of the proof we know that $\tilde{Z} = \tilde{f} \circ Y$ for bounded $\tilde{f}$. Then $Z = \tan \circ \tilde{Z} = \tan \circ \tilde{f} \circ Y$ and choosing $f = \tan \circ \tilde{f}$ as our Borel function, the proof is complete. □

4 Independence

As the first application on sigma-algebras gathering probabilistic information we look at independence.

Recall the conditional probability $\mathbf{P}(A|B) = \frac{\mathbf{P}(A\cap B)}{\mathbf{P}(B)}$. This follows from the interpretation that if Ω is our universal set, but we assume event B happened, we change the universal perspective to B. Now the probability of A in the universe where B happens is $\mathbf{P}(A \cap B)$. Now in a Bayes theorem view $\mathbf{P}(B) = |B|/|\Omega|$ interpreted as mass of the whole universe divides the smaller mass of a certain event happening should be the probability. Then in a sense $\mathbf{P}(A|B) = |A \cap B|/|B|$. Multiplying by $(1/|\Omega|)/(1/|\Omega|)$ we get

$$|A \cap B|/|B| = \frac{|A \cap B|}{|\Omega|} / \frac{|B|}{|\Omega|} = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

This should be taken as a heuristic, but in the measure theoretic foundations one can think of it as following. When restricting to a smaller sample space B, we need $\mathbf{P}_B(B) = 1 = \mathbf{P}(B)/\mathbf{P}(B)$. Now $\mathbf{P}(A|B) := \mathbf{P}_B(A) := \mathbf{P}_B(A \cap B) = \text{scaling factor} \cdot \mathbf{P}(A \cap B)$. But this scaling factor must be $1/\mathbf{P}(B)$ to stay consistant with $\mathbf{P}_B(B) = 1$.

If $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$, then $\mathbf{P}(A|B) = \mathbf{P}(A)$. We interpret this as saying that the knowledge of the occurrence of B does not reveal anything that could be used to improve our estimate about the occurrence of A, and we therefore consider the event A independent of the event B.

4.1 Independence formulated in language of sigma-algebras

Take a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Let $\mathcal{G}_i \subset \mathcal{F}$ be a indexed collection $j \in J$ of sigma-algebras

Definition 11. The collection $(\mathcal{G}_j)$ is **independent** if for any distinct $j_1, \dots, j_n \in J$ and events $A_{j_1} \in \mathcal{G}_{j_1}, \dots, A_{j_n} \in \mathcal{G}_{j_n}$ we have

$$\mathbf{P}(A_{j_1} \cap \cdots \cap A_{j_n}) = \mathbf{P}(A_{j_1}) \cdots \mathbf{P}(A_{j_N}).$$

Ofcourse whilst the collection of sigma-algebas may be infinite, we formula is stated for all finite strings.

To go from the abstract case to a more concrete case, take collection of random variables (X_j). We need to consider what does it mean for random variables to be independent. For this purpose we consider the sigma-algebra generate the collection.

Definition 12. Collcetion of random variables $(X_j)_{j\in J}$ is **independent** if the sigma-algebra $(\sigma(X_j))_{j\in J}$ is independent.

Finally let $(E_j) \subset \mathcal{F}$ be a collection of events, to consider what it means for events to be independent, we consider the indicator random variables $\mathbb{I}_{E_j} : \Omega \to \mathbb{R}$.

Definition 13. Collection of events $(E_j)_{j\in J}$ are **independent** if the collection of random variables $(\mathbb{I}_{E_j})_{j\in J}$ is independent.

Remark: (The elementary notion of independence of events) the sigma-algebra $\sigma(\mathbb{I}_E) = \{E, E^c, \Omega, \emptyset\}$ for any event $E \in \mathcal{F}$. Thus events of a collection (E_j) are indepedent iff $(\{E_j, E_j, \Omega, \emptyset\})_{j\in J}$ is an independent collection. In in the spirit of the definition of independece this equates to

$$\mathbf{P}(E^{\star}_{j_1} \cap \cdots \cap E^{\star}_{j_n}) = \mathbf{P}(E^{\star}_{j_1}) \cdots \mathbf{P}(E^{\star}_{j_n})$$

where $\star$ denotes either complement or not. Notice that we do not need to check cases containing the entire sample space or empty set. In the familiar case of two events E_1, E_2 this means $\mathbf{P}(E_1 \cap E_2) = \mathbf{P}(E_1)\mathbf{P}(E_2)$ (*).

Exercise: Show that $\mathbf{P}(E_1 \cap E_2^c) = \mathbf{P}(E_1)\mathbf{P}(E_2^c)$ from (*). Sol:

$$\mathbf{P}(E_1 \cap E_2^c) = \mathbf{P}(E_1) - \mathbf{P}(E_1 \cap E_2) = \mathbf{P}(E_1)(1 - \mathbf{P}(E_2)) = \mathbf{P}(E_1)\mathbf{P}(E_2^c).$$

Let $X, Y \colon \Omega \to \mathbb{R}$ two random variables on $(\Omega, \mathcal{F}, \mathbf{P})$.

(a) Show that if X, Y are independent, then $f(X)$ and $g(Y)$ are independent for any Borel-measurable functions $f, g \colon \mathbb{R} \to \mathbb{R}$.

(b) Show that if $f(X), g(Y)$ are independent for all Borel functions $f, g \colon \mathbb{R} \to \mathbb{R}$, then X, Y are independent.

Sol: (a) Keep in mind that for real valued random variables we can think of the sigma-algeras they generate as just preimage over Borel sets. Let $\sigma(X), \sigma(Y)$ be independent. Then taking any two sets in the Borel set form, $(f(X))^{-1}(B) \in \sigma(f(X)), (g(Y))^{-1}(B') \in \sigma(g(X))$. Since preimage is contravariant, $(f(X))^{-1}(B) = X^{-1}(f^{-1}(B))$ and same for the other variable. So since f, g are Borel the preimage sets are borel and X, Y independent preimages of Borel sets,

$$\mathbf{P}(X^{-1}(f^{-1}(B)) \cap Y^{-1}(g^{-1}(B'))) = \mathbf{P}(X^{-1}(f^{-1}(B)))\mathbf{P}(Y^{-1}(g^{-1}(B'))). \ (*)$$

(b) If we assume $(*)$ holds for any Borel functions, then pick the Identity functions so we have the claim.

Assume that X, Y are real valued random variables such that $\mathbf{P}(X+Y = 42) = 1$. Is it possible that X and Y are independent. Sol: Assume that X, Y are indepedent. We know that $X \overset{a.s.}{=} 42-Y$. Thus $\mathbf{P}(\{X \leq c\} \cap \{X \leq c\}) = \mathbf{P}(\{X \leq c\} \cap \{42-Y \leq c\}) = \mathbf{P}(X \leq c)^2$. This implies $\mathbf{P}(X \leq c) = 0$ or 1. Same logic also implies that $\mathbf{P}(Y \leq c) = 0$ or 1. Because the sets of type we used generate both sigma-algebras, one can deduce from subadditivity and monotonicity that the previous result holds for all measurable sets in their respectful sigma-algebras. So it is possible in this case for example constant random variables.

Let $X, Y : \Omega \to \mathbb{N}$ be two independent random variables with

$$\mathsf{P}[X = j] = \mathsf{P}[Y = j] = \frac{1}{2^j} \qquad \text{for all } j \in \mathbb{N} = \{1, 2, \ldots\}.$$

(a) Show that $\mathsf{P}[Y > n] = \frac{1}{2^n}$ for any $n \in \mathbb{N}$.

Calculate the following probabilities

$$\begin{array}{llll} \text{(b):} & \mathsf{P}[X = Y] & \text{(c):} & \mathsf{P}[\min(X, Y) \leq k], \text{ where } k \in \mathbb{N} \\ \text{(d):} & \mathsf{P}[Y > X] & \text{(e):} & \mathsf{P}[X > kY], \text{ where } k \in \mathbb{N} \\ \text{(f):} & \mathsf{P}[X \text{ divides } Y] & & \end{array}$$

Sol: (a) $\mathbf{P}[Y > n] = \mathbf{P}(\cup_{i>n}(X = i))$. Now these are disjoint since preimages preserverve disjointness so $\mathbf{P}(\cup_{i>n}(X = i)) = \sum_{i=n+1}^{\infty} \frac{1}{2^i} = \frac{(1/2^{n+1})}{(1/2)}$ by geometric series limit.

(b) $\mathbf{P}[X = Y] = \mathbf{P}[\cup_j(X = j = Y)] = \mathbf{P}[\cup_j(X = j) \cap (Y = j)] = \sum_{i=1}^{\infty} \frac{1}{2^{2i}} = \sum_{i=1}^{\infty} \frac{1}{4^i} = \frac{(1/4)}{1-(1/4)} = \frac{1 \cdot 4}{4 \cdot 3} = \frac{1}{3}$.

(c) $\mathbf{P}[\min(X, Y) \leq k] = \mathbf{P}[(X \leq k) \cup (Y \leq k)] = 1 - \mathbf{P}((X > k) \cap (Y > k)) = 1 - \frac{1}{2^{2n}} = 1 - \frac{1}{4^n}$.

(d) One can notice that $\mathbf{P}[X = Y] + \mathbf{P}[Y < X] + \mathbf{P}[Y > X] = 1$. The inequality events have the same probability by symmetry so $\mathbf{P}[Y > X] = \frac{1}{3}$ by applying (b).

(e). By disjointness of events $\mathbf{P}[X > kY] = \mathbf{P}[\cup_i((X > ki) \cap (Y = i))] = \sum_{i=1}^{\infty} \mathbf{P}[X > ki]\mathbf{P}[Y = i] = \sum_{i=1}^{\infty} (\frac{1}{2^{k+1}})^i = \frac{(1/2^{k+1})}{1-(1/2^{k+1})} = \frac{1-2^{-(k+1)}}{2^{k+1}}$.

(f) $\mathbf{P}[X \text{ divides } Y] = \mathbf{P}[\cup_{i,j}((Y = ji) \cap (X = i))] = \sum_{i,j} \mathbf{P}[Y = ij]\mathbf{P}[X = i] = \sum_i \frac{1}{2^i} \sum_j \frac{1}{2^{ij}} = \sum_i \frac{1}{2^i} \frac{1}{2^i-1} = \sum_i \frac{1}{4^i-2^i}$.

Notation:

$$\begin{array}{rc}
(\mathcal{G}_j)_{j\in J} & \perp\!\!\!\perp & \text{if the collection } (\mathcal{G}_j)_{j\in J} \text{ of } \sigma\text{-algebras is independent} \\
(X_j)_{j\in J} & \perp\!\!\!\perp & \text{if the collection } (X_j)_{j\in J} \text{ of random variables is independent} \\
(E_j)_{j\in J} & \perp\!\!\!\perp & \text{if the collection } (E_j)_{j\in J} \text{ of events is independent.}
\end{array}$$

In the case of enumerated (countable) collections, we use the notation:

$$\begin{array}{rcc}
\mathcal{G}_1, \mathcal{G}_2, \dots & \perp\!\!\!\perp & \text{if the collection } (\mathcal{G}_j)_{j\in\mathbb{N}} \text{ of } \sigma\text{-algebras is independent} \\
X_1, X_2, \dots & \perp\!\!\!\perp & \text{if the collection } (X_j)_{j\in\mathbb{N}} \text{ of random variables is independent} \\
E_1, E_2, \dots & \perp\!\!\!\perp & \text{if the collection } (E_j)_{j\in\mathbb{N}} \text{ of events is independent.}
\end{array}$$

In the case of collections of just two members, we use notation:

$$\begin{array}{lc}
\mathcal{G}_1 \perp\!\!\!\perp \mathcal{G}_2 & \text{if the collection } (\mathcal{G}_j)_{j\in\{1,2\}} \text{ of } \sigma\text{-algebras is independent} \\
X_1 \perp\!\!\!\perp X_2 & \text{if the collection } (X_j)_{j\in\{1,2\}} \text{ of random variables is independent} \\
E_1 \perp\!\!\!\perp E_2 & \text{if the collection } (E_j)_{j\in\{1,2\}} \text{ of events is independent.}
\end{array}$$

Exercise: (Pairwise independence does not imply independence). Construct an example of three pairwise independent sigma-algebras such that the collectoin they form is not independent.

Sol: Take the space $\Omega = \{HH, TT, HT, TH\}$ of two coin tosses. Take the events $E_1 = \{HH, TT\}, E_2 = \{HH, HT\}, E_3 = \{TT, HT\}$ with corresponding probabilities 1/2 each. Sigma-algebras generates by single events are of the form $\{\emptyset, E_i, E_i^c, \Omega\}$. By previous remark events are independent iff their generated sigma-algebras are independent. Since any two distinct events of the form E_i intersect at a single outcome, the intersection probabilities are 1/4. But also clearly $\mathbf{P}(E_i)\mathbf{P}(E_j) = 1/4$. But for this collection is not an independent triple. Take $\mathbf{P}(E_1 \cap E_2 \cap E_3) = 0 \neq 1/8 = \mathbf{P}(E_1)\mathbf{P}(E_2)\mathbf{P}(E_3)$ so not.

4.2 Verifying independence

One does not usually want to work directly with the definition of independence as it is clunky.

Proposition 4. *Suppose that $\mathcal{J}_1, \mathcal{J}_2$ are pi-systems on ω and let $\mathcal{F}_1 = \sigma(\mathcal{J}_1), \mathcal{F}_2 = \sigma(\mathcal{J}_2)$. The the following conditions are equivalent:*

- $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2$
- *For all $I_1 \in \mathcal{J}_1, I_2 \in \mathcal{J}_2$, we have $\mathbf{P}[I_2 \cap I_2] = \mathbf{P}[I_1]\mathbf{P}[I_2]$.*

Proof. First condition clearly implies the second one so assume the second condition.

Step 1: Let $I \in \mathcal{J}_1$. First we show that $\mathbf{P}[I_1 \cap E_2] = \mathbf{P}[I_1]\mathbf{P}[E_2]$ for all $E_2 \in \mathcal{F}_2$. If $\mathbf{P}[I_1] = 0$, it holds trivially thus assume non-zero probability. Define a new probability measure $\tilde{\mathbf{P}}_2[E_2] = \mathbf{P}[I_1 \cap E_2]/\mathbf{P}[I_1]$ on $(\Omega, \mathcal{F}_2)$. By our assumpiont $\tilde{\mathbf{P}}_2$ agrees with $\mathbf{P}$ on the pi-system $\mathcal{J}_2$. By Dynkin's identification theorem, $\mathbf{P} = \tilde{\mathbf{P}}_2$

on $\mathcal{F}_2$. This shows that $\mathbf{P}[E_2] = \tilde{\mathbf{P}}_2[E_2] = \mathbf{P}[I_1 \cap E_2]/\mathbf{P}[I_1]$ which yields what we wanted.

Step 2: Let $E_2 \in \mathcal{F}_2, E_1 \in \mathcal{F}_1$. Again assume that $\mathbf{P}[E_2] > 0$ and define a new probability measure $\tilde{\mathbf{P}}_1[E_1] = \mathbf{P}[E_1 \cap E_2]/\mathbf{P}[E_2]$ on $(\Omega, \mathcal{F}_1)$. By step 1, $\tilde{\mathbf{P}}_1$ agrees with $\mathbf{P}$ on the pi-system $\mathcal{J}_1$ so again by Dynkin's indentification theorem, on the entire sigma-algebra. Thus like at the end of step 1, we can conclude. □

Typical first example:

Corollary 5. *Suppose that X, Y are real-valued random variables. Then $X \perp\!\!\!\perp Y$ iff for all $x, y \in \mathbb{R}$, $\mathbf{P}[X \leq x, Y \leq y] = \mathbf{P}[X \leq x]\mathbf{P}[Y \leq y]$.*

Proof. Once again remember the pi-system $\mathcal{J}(\mathbb{R}) = \{(-\infty, x] | x \in \mathbb{R}\}$ and let $\mathcal{J}_1 = X^{-1}\mathcal{J}(\mathbb{R})$. In a previous exercise we showed that $\sigma(\mathcal{J}_1) = \sigma(X)$ and similarly thus $\sigma(\mathcal{J}_1) = \sigma(Y)$, where $\mathcal{J}_2 = Y^{-1}\mathcal{J}(\mathbb{R})$. Then it follows from previous proposition. □

Similar statements also hold for finite collections for pi-systems.

Exercise: Let $\mathcal{F}_i$ with $i = 1, 2, 3$ be three sigma-algebra on Ω. Assume that each one generate by a pi-system $\mathcal{J}_i$ containing Ω. Then $\mathcal{F}_i$'s are independent iff $\mathbf{P}[I_1 \cap I_2 \cap I_3] = \mathbf{P}[I_1]\mathbf{P}[I_2]\mathbf{P}[I_3]$ for all $I_i \in \mathcal{F}_i$'s.

Sol: (a) If we assume that $\Omega \in \mathcal{J}_i$ for all i, then $\mathbf{P}(I_i \cap I_j) = \mathbf{P}(I_i \cap I_j \cap \Omega) = \mathbf{P}(I_i)\mathbf{P}(I_j)$ for $i_i \in \mathcal{J}_j, I_j \in \mathcal{J}_i, i \neq j$. Then we do as in the proof for pair of sigma-algebras. We will assume that every set has non-zero probability. Define $\tilde{\mathbf{P}}_3(E) = \frac{\mathbf{P}(I_1 \cap I_2 \cap E)}{\mathbf{P}(I_1)\mathbf{P}(I_2)}$ on all of $\mathcal{F}_3$. This is a probability measure because we included Ω. Now this coincides with $\mathbf{P}$ on the pi-system, and hence by Dynkin's identification theorem $\mathbf{P}(E) = \frac{\mathbf{P}(I_1 \cap E \cap E')}{\mathbf{P}(I_1)\mathbf{P}(I_2)}$. Next assume define the measure $\tilde{\mathbf{P}}_2(E) = \frac{\mathbf{P}(I_1 \cap E \cap E')}{\mathbf{P}(I_1)\mathbf{P}(E')}$ for non-zero probability $I_1 \in \mathcal{J}_1, E' \in \mathcal{F}_3$. Again with similar argument and from the first part $\mathbf{P}(E) = \frac{\mathbf{P}(I_1 \cap E \cap E')}{\mathbf{P}(I_1)\mathbf{P}(E')}$. Lastly assume $E_i \in \mathcal{F}_i$ and define measure $\tilde{\mathbf{P}}_2(E_1) = \frac{\mathbf{P}(E_1 \cap E_2 \cap E_3)}{\mathbf{P}(E_2)\mathbf{P}(E_3)}$. Then similar argument and the first two parts we get the claim.

(b) The assumption for the pi-systems to contain Ω was essential to quarantee that the defined measures are actually probability measure. Take for example the set $\Omega = 1, \ldots, 8$ with discrete uniform probability measure $\mathbf{P}$ on it and single element pi-systems $A = \{1, 2, 3, 4\}, B = \{1, 2, 5, 6\}, C = \{2, 5, 7, 8\}$. Then $\mathbf{P}(A \cap B \cap C) = 1/8, \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C) = \frac{1}{2^3}$. But can can notice that $\mathbf{P}(A \cap C) = 1/8, \mathbf{P}(A)\mathbf{P}(C) = 1/2^2$ so the assumption is a necessary.

4.3 Borel-Cantelli

Given a sequence of events $E_1, E_2, \cdots \in \mathcal{F}$ we occasionally care about whether infinitely or finitely many of them occur. Borel-Cantelli lemmas are examples what are called 0-1 laws in probability theory. They state that under some mild conditions that often easy to verify that the probability we are interested in is trivial. For a sequence of events we use the following definitions.

- "E_n occurs infinitely often"

  "E_n i.o. " $= \bigcap_{m \in \mathbb{N}} \bigcup_{n \geq m} E_n = \limsup_n E_n = \{\omega \in \Omega | \omega \in E_n$ for infinitely many indices $n\}$.

- "E_n occurs eventually"

  $"E_n$ i.o. $" = \bigcup_{m \in \mathbb{N}} \bigcap_{n \geq m} E_n = \liminf_n E_n = \{\omega \in \Omega | \omega \in E_n$ for all except finitely many $n\}$.

By de Morgan's laws ("E_n occurs infinitely often"$)^c$ = ("E_n occurs eventually").

Exercise: Consider sequence $E_1, E_2, \dots \subset \Omega$. Show that for all $\omega \in \Omega$ we have

$$\limsup_n \mathbb{I}_{E_n}(\omega) = \mathbb{I}_{\limsup_n E_n}(\omega) \ \ \& \ \ \liminf_n \mathbb{I}_{E_n}(\omega) = \mathbb{I}_{\liminf_n E_n}(\omega).$$

Use this to show that $\liminf_n E_n \subset \limsup_n E_n$.

Sol: Assume $\mathbb{I}_{\limsup_n E_n}(\omega) = 1$ meaning $\omega \in \limsup_n E_n$ meaning for infinitely many n, $\mathbb{I}_{E_n}(\omega) = 1$. Thus $\limsup_n \mathbb{I}_{E_n}(\omega) = \inf_{n \in \mathbb{N}}(\sup_{k \geq n} \mathbb{I}_{E_k}(\omega)) = \inf_{n \in \mathbb{N}}\{1\} = 1$. Conversly if $\limsup_n \mathbb{I}_{E_n}(\omega) = 1$, then $\inf_{n \in \mathbb{N}}(\sup_{k \geq n} \mathbb{I}_{E_k}(\omega)) = 1$. This means that after any n, there must be some index $k \geq n$ for which $\mathbb{I}_{E_k}(\omega) = 1$. Meaning $\omega \in E_k$ for infinitely many indices $k \in \mathbb{N}$.

Assume $\mathbb{I}_{\liminf_n E_n}(\omega) = 1$, then outside a finite set of indices $\mathbb{I}_{E_n}(\omega) = 1$. Write $\liminf_n \mathbb{I}_{E_n}(\omega) = \sup_{n \in \mathbb{N}}(\inf_{k \geq n} \mathbb{I}_{E_k}(\omega))$. Since the vanishing happens only in a finite set we can pick large enough natural number N so $I_{E_k}(\omega) = 1, \forall k > N$. Meaning $\sup_{n \in \mathbb{N}}(\inf_{k \geq n} \mathbb{I}_{E_k}(\omega)) = 1$ since it is eventually the constant sequence $1, 1, 1, \dots$. Conversly assume that $1 = \liminf_n \mathbb{I}_{E_n}(\omega) = \sup_{n \in \mathbb{N}}(\inf_{k \geq n} \mathbb{I}_{E_k}(\omega))$. This means after some amount of indices N, the sequence $I_{E_k}(\omega)$ stabilizes to the constant sequence of 1, so $\omega \in \liminf_n E_n$ and $\mathbb{I}_{\liminf_n E_n}(\omega) = 1$.

To show the inclusion $\liminf_n E_n \subset \limsup_n E_n$ we can see $\mathbb{I}_{\liminf_n E_n} = \liminf_n \mathbb{I}_{E_n} \leq \limsup_n \mathbb{I}_{E_n} = \mathbb{I}_{\limsup_n E_n}$.

The two Borel-Cantelli lemmas: The first Borel-Cantelli lemma says that whenever the probabilities of events E_n decay fast enough, it is almost surely impossible for the events to occur infinitely often.

Lemma 7. *(Convergence lemma) Suppose that $E_1, E_2, \dots \in \mathcal{F}$ are such that $\sum_{n=1}^{\infty} \mathbf{P}[E_n] < +\infty$. Then we have that $\mathbf{P}[$"E_n occurs infinitely often"$] = 0$.*

Proof. Dnote $G_m = \bigcup_{n=m}^{\infty} E_n$, so that the event $E = \limsup_n E_n$ of interest is decreasing limit $G_m \downarrow E$ as $m \to \infty$. By monotone convergence of probability measures $\mathbf{P}[E] = \lim_{m \to \infty} \mathbf{P}[G_m]$. Then

$$0 \leq \mathbf{P}[G_m] = \mathbf{P}[\bigcup_{n=m}^{\infty} E_n] \leq \sum_{n=m}^{\infty} \mathbf{P}[E_n].$$

Since the sum converes, as we crank up m, $\mathbf{P}[G_m] \to 0$. Meaning thus $\mathbf{P}[E] = 0$. □

The second Borel-Cantelli lemma says that probabilities of events E_n not decaying fast enough and if in addition independent, then the events most almost surely occur infinitely often.

Lemma 8. *(divergence) Suppose that $E_1, E_2, \dots \in \mathcal{F}$ are independent events and $\sum_{n=1}^{\infty} \mathbf{P}[E_n] = +\infty$. Then we have that*

$$\mathbf{P}[\text{"}E_n \textit{ occurs infinitely often"}] = 1.$$

Proof. Instead of $E = \bigcap_{m\in\mathbb{N}} \bigcup_{n\geq m} E_n$ consider it's complement $E^c = \bigcup_{m\in\mathbb{N}} \bigcap_{n\geq m} E_n^c$. Denote $F_m = \bigcap_{n\geq m} E_n^c$. In order to use indepencence, we need to restrict to finite intersections $F_{m,l} = \bigcap_{n=m}^{l} E_n^c$ so $F_{m,l} \downarrow F_m$ as l grows large. By independence and the estimate $1 - a \leq e^{-a}$ for $a \in [0,1]$ we have

$$\mathbf{P}[F_{m,l}] = \prod_{n=m}^{l} 1 - \mathbf{P}[E_n] \leq \prod_{n=m}^{l} e^{-\mathbf{P}[E_n]} = \exp\Bigg(- \sum_{n=m}^{l} \mathbf{P}[E_n] \Bigg).$$

So by assumption on the sum diverging, we get that $\mathbf{P}[F_m] = 0$ as $l \to \infty$ by monotone convergence for probability measures. Now $E^c = \bigcup_{m\in\mathbb{N}} F_m$ so $\mathbf{P}[E^c] \leq \sum_{m=1}^{\infty} \mathbf{P}[F_m] = 0$ meaning $\mathbf{P}[E] = 1$. □

Exercise: Why is the independence assumption needed? Show that there is sequence of non independent events whose probabilities sum to $\sum_{n=1}^{\infty} \mathbf{P}[E_n] = +\infty$, but $\mathbf{P}["E_n \text{ occurs infinitely often"}] \neq 1$.

Sol. Take Lebesgue measure on Borel sets on the unit interval. Then the constant sequence $[0, 1/2]$ divergence in summing probabilities. The limit superior of this sequence is obvisouly $[0, 1/2]$ with probability a half. The collection $[0, 1/n]$ actually yields even worse probability, namely zero.

Example: consider an annual sports contest, in which one keeps track of the winner's scores for different years as well as the record score of all past years.

Suppose that the winner's score for nth year is real-valued random variable X_n, and suppose that $X_1, X_2, \ldots$ are independent and identically distributed. Moreover that the cumulative distribution function is continuous (these assumptions are not completely unreasonable).

Define the events $E_n = \{X_n > \max\{X_1, \ldots, X_{n-1}\}\}$ so the event of new record being made. The collection of events $X_1, \ldots, X_n$ has exactly one maximum sinc. Since we assume contious cumulative distribution, the probability that two random variables are the same is zero. Hence the chance of X_n being a new maximum is $1/n$. Also these events are independent. We have that $\mathbf{P}(E_i)(E_j) = \frac{1}{ij}, i \neq j$. Also $\mathbf{P}(E_i \cap E_j) = \mathbf{P}(X_j > \max(X_{j-1}, \ldots, X_1) \cap X_i > \max(X_{i-1}, \ldots, X_1))$. Now this again means that we have probability $1/j$ that X_j is larger than previous scores by continuity assumption and $1/i$ for X_i, thus the intersection of these events must be $\frac{1}{ij}$. Now the sum $\sum_{n=1}^{\infty} \mathbf{P}(E_n)$ is the harmonic series. Borel-Cantelli them implies that E_n i.o. almost surely. So almost surely a new record are made.

Consider the event $F_n = E_n \cap E_{n+1}$ of record being broken two years consecutively. By independence $\mathbf{P}[F_n] = \frac{1}{n(n+1)}$, and thus $\sum_{n=1}^{\infty} \mathbf{P}[F_n] < +\infty$. By Borel-Cantelli F_n i.o. never. So almost surely there is only finite amount of consecutive records.

Exercise: (decimal digits of a uniform random number). Let $\mathbf{P}$ be the uniform probaility measure on the unit intercal $[0, 1]$. Consider the the digits $D_k(\omega), k \in \mathbb{N}$ of the decimal representation of a number $\omega \in [0, 1]$, so $\sum_{k=1}^{\infty} D_k(\omega) 10^{-k}$. For this exercise you can consider it known that the digits $D_1, D_2, \ldots$ are random variables on the sample space $\Omega = [0, 1]$ and that they are independent (think about why).

Let Z_k be the length of zeroes starting from the kth digit:

- $Z_k = 0$ if the kth digit is not zero, $D_k \neq 0$;

- $Z_k = m$ if $D_k = D_{k+1} = \cdots = D_{k+m-1} = 0$ and $D_{k+m} \neq 0$.

(a) Show that $\mathbf{P}[Z_k = m] = \frac{9}{10^{m+1}}$

(b) Fix $m \in \mathbb{Z}_{\geq 0}$. Show that $\mathbf{P}[Z_k = m$ for infinitely many $k] = 1$.

(c) Show that $\mathbf{P}[Z_k = k$ for infinitely many $k] = 0$.

Sol: The reason $D_k \colon [0,1] \to S = \{0, \ldots, 9\}$ is a random variable assuming powerset sigma-algebra on the target space, is that preimage of any singleton $n \in S$ is the union of all the intervals of the form $I_i = [\underbrace{0.\cdots \boxplus n \boxplus i \boxplus \cdots}_{k\text{th digit is } n, i \in S\setminus 9}, \underbrace{0.\cdots \boxplus n \boxplus (i+1) \boxplus \cdots}_{k\text{th digit is } n})$ or $[\underbrace{0.\cdots \boxplus n \boxplus 9 \boxplus \cdots}_{k\text{th digit is } n, n \in S\setminus 9}, \underbrace{0.\cdots \boxplus (n+1) \boxplus 0 \boxplus \cdots}_{k\text{th digit is } n+1})$ or

$$[\underbrace{0.\cdots \boxplus 9 \boxplus 9 \boxplus \cdots}_{k\text{th digit is } 9}, \underbrace{0.\cdots (\text{terminates atleast at the first non nine}) \boxplus 0 \boxplus 0 \boxplus \cdots}_{k\text{th digit is } 0}).$$

Here $\boxplus$ denotes concatenation. One can actually show that the Lebesgue measure of this preimage $\mathcal{L}(D_i^{-1}\{k\}) = 1/10$. One can then check that picking two different i's that this process is independent easily.

(a)

$$\begin{aligned}
\mathbf{P}[Z_k = m] &= \mathbf{P}\left[\left(\bigcap_{n=k}^{k+m-1} (D_n = 0)\right) \cap (D_{k+m} \neq 0)\right] \\
&= \left(\prod_{n=k}^{k+m-1} \mathbf{P}[D_n = 0]\right) \mathbf{P}[D_{k+m} \neq 0] \\
&= (1/10)^m \cdot 9/10 = \frac{9}{10^{m+1}}.
\end{aligned}$$

(b) Since $\sum_{k=1}^{\infty} \mathbf{P}[Z_k = m] = +\infty$ then we apply Borel-Cantelli.

(c) $\sum_{k=1}^{\infty} \mathbf{P}[Z_k = k]$ is a $p-$series for $p > 1$ so it converges, hence apply the other Borel-Cantelli.

5 Events of the infinite horizon

Consider a sequence of random variables. The topic of this lecture is: What information about the sequence is not sensitive to the values of any finite number of individual members of the sequence? Which events can be decided and which random variables are de- termined by such information?

Although it might at first appear surprising, there are in fact many interesting events and random variables which are not affected by any finite number of individual values. We will give a number of examples.

We will moreover return to profound consequences of independence, and ask: Assuming moreover that the sequence of random variables is inde- pendent, what can be said about the probabilities of events that are not sensitive to any finite number of individual values?

Pertaining to the last question, we will prove Kolmogorov's 0-1 law which states that under the independence assumption, any event which is not sensitive to finitely many individual values has probability either zero or one, and any random variable which is not affected by finitely many individual values is almost surely constant. This probabilistic fact underlies some surprising phenomena, in particular related to phase transitions in physical systems.

5.1 Tail sigma-algebras

Given a sequence $X_1, X_2, \dots$ of random variables denote by $\mathcal{X} := \sigma((X_n)_{n\in\mathbb{N}})$. Interpret all the information contained in the sequence. Let $\mathcal{X}_k = \sigma((X_n)_{n\leq k})$ describe then the information contained in the first k terms and $\mathcal{T}_k := \sigma((X_n)_{n>k})$ the information contained without it's k first terms. In other words $\mathcal{T}_k$ is the information in the sequence not affected by the first k terms.

Our main interest lies in the **tail sigma-algebra** $\mathcal{T}_\infty := \cap_{k\in\mathbb{N}}\mathcal{T}_k$. The tail contains less information than any $\mathcal{T}_k$.

Note that $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \cdots \subset \mathcal{X}$ and $\mathcal{T}_\infty \subset \cdots \subset \mathcal{T}_2 \subset \mathcal{T}_1 \subset \mathcal{X}$.

Example: Take real-valued random variables $X_1, X_2, \dots \in m\mathcal{F}$ on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. By definition $X_n \subset m\mathcal{X}_n \subset m\mathcal{X}_l \subset m\mathcal{X}$ for any $n \leq l$. because linear combinations of measurables are measurable, we get that finite avarages $\frac{X_1+\cdots+X_l}{l} \in m\mathcal{X}_l \subset \mathcal{X}$. Because of upper and lower limits of measurable functions are measurable also $\liminf_n X_n, \limsup_n X_n \in m\mathcal{X}$. We will prove this for limit inferior and limit superior is the same kind of method. Write $\liminf_n X_n = \sup_{m\in\mathbb{N}}(\inf_{n\geq m} X_n) = \sup_{m>k}(\inf_{n\geq m} X_n)$ (omitting finite terms does not change the limiting behavior). Now by definition $\inf_{n\geq m} X_n \in m\mathcal{T}_k$. Since this holds for any k, we get that $\inf_{n\geq m} X_n \in m\mathcal{T}_\infty$.

Now provided that the limit $\lim_{n\to\infty} X_n$ exists, it coincides with the lower and upper limits and thus is also tail-measurable. Define the event

$$E = \{\omega \in \Omega | \exists \lim_{n\to\infty} X_n(\omega)\}.$$

A limit does not exist iff the lower and upper limits agree do not agree iff there exists a $q \in \mathbb{Q}$ such that $\liminf_n X_n < q < \limsup_n X_n$ (density of rationals lets us pick a rational). Then we can write the E by complementing this so

$$E = \left(\bigcup_{q\in\mathbb{Q}} (\{\liminf_n X_n < q\}) \cap \{q < \limsup_n X_n\} \right)^c,$$

where the event is only written by sigma-algebra operations of $\mathcal{T}_\infty$ hence E is a tail event.

If the sequence (X_n) was bounded we could've just in slick way defined $D = \limsup_n X_n - \liminf X_n$ where the limit exists iff $D = 0$ since it is linear combination of tail-measurable functions. Then $E = D^{-1}(0) \in \mathcal{T}_\infty$ since $\{0\}$ is Borel.

A slight less obvious is limit of averages $\lim_{l\to\infty} \frac{X_1+\cdots+X_l}{l}$. For a fixed k, obviously $\frac{X_1+\cdots+X_k}{l} \to 0$ as l grows. Thus

$$\begin{aligned}\limsup_{l\to\infty} \frac{X_1+\cdots+X_l}{l} &= \limsup_{l\to\infty} \frac{(X_1+\ldots X_k)+(X_{k+1}+\cdots+X_l)}{l} \\ &= 0+\limsup_{l\to\infty} \frac{X_{k+1}+\cdots+X_l}{l}.\end{aligned}$$

The random variables $(X_n)_{n>k}$ are by construction $\mathcal{T}_k-$measurable, so $\frac{X_{k+1}+\cdots+X_l}{l}$ is $\mathcal{T}_k-$measurable. Thus also $\limsup_{l\to\infty} \frac{X_{k+1}+\cdots+X_l}{l}$ is $\mathcal{T}_k-$measurable. Because this works for every k, the limit of averages is $\mathcal{T}_\infty-$measurable. Just like before now the existance of the limit of an average is a tail event. These examples show that there are some interesting events in the tail sigma-algebra.

Exercise: Let $X_1, X_2, \ldots$ be a sequence of real-valued random variables on $(\Omega, \mathcal{F}, \mathbf{P})$. Investigate which of the following events are tail events.

(a): $\{\omega \in \Omega |$ the series $\sum_{n=1}^{\infty} X_n(\omega)$ converges$\}$

(b): $\{\omega \in \Omega | \sum_{n=1}^{\infty} X_n(\omega) \leq -42\}$

(c): $\{\omega \in \Omega |$ the sequence $X_1, X_2, \ldots$ is bounded$\}$

(d):

$$\begin{aligned}&\{\omega \in \Omega | \forall \ell \in \mathbb{N}\ \exists n \in \mathbb{N} \text{ such that } X_n(\omega) = X_{n+1}(\omega) = \cdots = X_{n+\ell}(\omega)\} \\ &= \{\text{there exists arbitrarily long repetitions in the sequence } X_1, X_2, \ldots\}\end{aligned}$$

Sol: (a) is a limit so a tail event. (b) is not a tail even, take the sequence of constant random variables $-43, 1/2, 1/4, 1/8, \ldots$. The entire sum is -42 so the event is satisfied, but if we remove finite set of invormation, namely -43, then the event is not satisfied. For (c) is a tail event, a finite sequence is always bounded, so removing them doesn't chance the question about boundedness. It can be written as a question about the limsup of the sequence $(|X_n|)_n$. Also (d) is a tail event since removing a finite string of number doesn't change our look for arbituary long repetitions which can be always chosen to be longer than the length of removed terms.

5.2 Kolmogorov's 1-0 law

If we add an independence assumption a sequence we get a remarkable result.

Theorem 5. *(Kolmogorov's 0-1 law) Suppose that $X_1, X_2, \ldots$ is a sequence of independent random variables. Then the following hold*

(a) For any tail event $E \in \mathcal{T}_\infty$ we have $\mathbf{P}[E] = 1$ or 0.

(b) For any $\mathbb{R}-$valued random variable T which is $\mathcal{T}_\infty-$measurable is almost surely a constant. I.e. for some $c \in \mathbb{R}, \mathbf{P}[T = c] = 1$.

The proof strategy is almost as surprising as the statement. We will prove that $\mathcal{T}_\infty \perp\!\!\!\perp \mathcal{T}_\infty$ so independence with itself.

Proof. For (a) take E to be a tail event. If we assume this self-independence, $\mathbf{P}[E] = \mathbf{P}[E \cap E] = \mathbf{P}[E]^2$ which has only $0, 1$ as solutions. For (b) if T is tail-measurable, then $(T \leq x) \in \mathcal{T}_\infty$. By part (a), $\mathbf{P}[T \leq x]$ is either zero or one so the cumulative distribution only takes these values. Define $c := \inf\{x \in \mathbb{R} | \mathbf{P}[T \leq x] = 1\}$. By right-continuity $\mathbf{P}[T \leq c] = 1, \mathbf{P}[T < c] = 0$ so $\mathbf{P}[X = c] = \mathbf{P}[T \leq c] - \mathbf{P}[T < c] = 1$ almost surely. Thus we only have to prove the independence property.

Step 1: First we show that $\mathcal{X}_k \perp\!\!\!\perp \mathcal{T}_k$ using two pi-systems. Let $\mathcal{J}$ denote sets of the form $A = \{\omega \in \Omega | X_1(\omega) \in B_1, \ldots, X_k(\omega) \in B_k\} \subset \Omega$, where $B_1, \ldots, B_k \in \mathcal{B}(\mathbb{R})$. Infact $\sigma(\mathcal{J}) = \mathcal{X}_k$ since A is an intersection of preimages of X_n's of Borel sets for $n \leq k$. Then clearly the sigma-algebra $\sigma(\mathcal{J}) \subset \mathcal{H}_k$. $A \in \mathcal{X}_k$, then it is an intersection, union and complement of preimages of Borel sets of X_n'. So the other inclusion follows as well. It is clear that $\mathcal{J}$ is a pi-system.

Secondly define the pi-system $\mathcal{J}'$ of sets of the form $A' = \{\omega \in \Omega | X_{k+1}(\omega) \in B_{k+1}, \ldots, X_{k+r}(\omega) \in B_{k+r}\} \subset \Omega, B_{k+1}, \ldots, B_{k+r} \in \mathcal{B}(\mathbb{R}), r \in \mathbb{N}$. In a similar manner one may show $\sigma(\mathcal{J}') = \mathcal{T}_k$.

Using independence

$$\mathbf{P}[A] = \mathbf{P}[X_1 \in B_1, \ldots, X_k \in B_k] = \prod_{j=1}^{k} \mathbf{P}[X_j \in B_j] \;\; \& \;\; \mathbf{P}[A'] = \prod_{j=k+1}^{k+r} \mathbf{P}[X_j \in B_j].$$

And the intersection

$$\mathbf{P}[A \cap A'] = \prod_{j=1}^{k+r} \mathbf{P}[X_j \in B_j].$$

We previously proved a theorem which said that it is sufficient to look at generating pi-systems to determine independence. Therefore the claim follows.

Step 2: We show $\mathcal{X}_k \perp\!\!\!\perp \mathcal{T}_\infty$. By previous step $\mathcal{X}_k \perp\!\!\!\perp \mathcal{T}_k$ and $\mathcal{T}_\infty \subset \mathcal{T}_k$ so clearly indepence holds since we have less sets.

*Step 3:*We show $\mathcal{X} \perp\!\!\!\perp \mathcal{T}_\infty$. Again we do this by defining suitable pi-systems. Let $\mathcal{U} = \bigcup_{k \in \mathbb{N}} \mathcal{X}_k$ (note that union of sigma-algebras is not a sigma-algebra in general) and claim $\mathcal{U}$ is a pi-system. Taking any two sets $A, A' \in \mathcal{U}$, both belong to $\mathcal{X}_k$ for some k so we can just take their maximum and use the fact that $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \ldots$. Since for some large $A, A' \in \mathcal{X}_k$, since they belong to the same sigma-algebra $A \cap A' \in \mathcal{X}_k \subset \mathcal{U}$.

We then show that $\sigma(\mathcal{U}) = \mathcal{X}$ by double inclusion. For any $n \in \mathbb{N}$ we have that $X_n \in m\mathcal{X}_n$ by definition, abd since $\mathcal{X}_n \subset \bigcup_{k \in \mathbb{N}} \mathcal{X}_k = \mathcal{U} \subset \sigma(\mathcal{U})$, then also $X_n \in m\sigma(\mathcal{U})$. Thus $\sigma(\mathcal{U})$ is a sigma-algebra with respect which each X_n is measurable. By definition $\mathcal{X}$ is the smallest sigma-algebra for which each X_n is measurable, so $\mathcal{X} \subset \sigma(\mathcal{U})$. On the other hand, for each $k \in \mathbb{N}$ we have $\mathcal{X}_k \subset \mathcal{X}$ so also $\mathcal{U} \subset \mathcal{X}$, since $\sigma(\mathcal{U})$ is the smallest with this property.

We now now that $\mathcal{U}$ is a pi-system generating $\mathcal{X}$ and $\mathcal{T}_\infty$ a pi-system generating itself as a sigma-algebra. Take $A \in \mathcal{U}, E \in \mathcal{T}_\infty$ so $A \in \mathcal{X}_k$ for some k. In the previous step we showed that A, E satisfy the independence condition.

Step 4: We showed $\mathcal{X} \perp\!\!\!\perp \mathcal{T}_\infty$, but $\mathcal{X} \supset \mathcal{T}_\infty$ so obviously $\mathcal{T}_\infty \perp\!\!\!\perp \mathcal{T}_\infty$. □

Exercise. Let X be a complete separable metric space. Assume that a random variable $T \colon \Omega \to X$ is $\mathcal{T}_\infty / \mathcal{B}(X)-$measurable, where $\mathcal{T}_\infty$ is tail sigma-algebra of a sequence $X_1, X_2, \dots$ of independent random variables. Show that T is almost surely a constant.

Sol: We begin by revalling small lemma. Every separable metric space can be partioned by Borel sets to arbituary small diameter. Let $S = \{s_1, s_2, \dots\}$ the countable dense subset of X and define $A_{n,1} := \{x \in X | d(x, s_1) < 1/2n\}$ and recursively $A_{n,k+1} := \{x \in X | d(x, s_{k+1}) < 1/2n\} \setminus \cup_{i \leq k} A_{n,i}$. Now $\operatorname{diam}(A_{n,k}) \leq 1/n$ since $d(x,y) \leq d(x, s_k) + d(y, s_k) = 1/n$. The construction is clearly Borel sets, disjoint and covers the entire space. Thus $1 = \mathbf{P}[T \in X] = \sum_{i=1}^{\infty} \mathbf{P}[T \in A_{n,i}]$. Because these are all tail events they are either probability zeros or ones, so exactly one event in the partition must have probability one. Take $\mathbf{P}[T \in A_{n,k}] = 1$ for some $k \in \mathbb{N}$ and for all $n \in \mathbb{N}$. Take the set $A = \bigcap_{n \in \mathbb{N}} A_{n,k}$ as the limit of $\bigcap_{n \leq l} A_{n,k}$ in terms of l. Assuming completeness, by Cantor's intersection theorem for metric space, A is a singleton. Thus by monotone continuity of probability measure $\mathbf{P}[T = c] = \mathbf{P}[T \in A] = \mathbf{P}[\bigcap_{n \in \mathbb{N}} A_{n,k}] = 1$ for $A = \{c\}$.

A few interesting examples: (Random series with independent terms). Let $X_1, X_2, \dots$ be a $\mathbb{R}-$valued indepedent sequence of random variables. Take the random series formed from the sequence $\sum_{n=1}^{\infty} X_n$. Then the event $E = \{\omega \in \Omega | \sum_{n=1}^{\infty} X_n(\omega) \text{ converges}\}$ is a tail event. By Kolmogorov's 0,1 law the probability of E is zero or one. But we do not know which extreme occurs.

(Series with randomly assigned signs) Let $a_1, a_2, \dots \geq 0$ be random sequence of real numbers and $S_1, S_2, \dots$ independent identically distributed $\{\pm 1\}-$valued random variables with $\mathbf{P}[S_n = 1] = 1/2, \mathbf{P}[S_n = -1] = 1/2$. Consider the following series $\sum_{n=1}^{\infty} a_n S_n$. The previous example gives that the series converges or diverges almost surely. It depends on the given sequence $(a_n)_n$. For example the sequence can only converge if it's last terms tend to zero. Since $|s_n S_n| = a_n$, the sequence can oly converge if $a_n \to 0$ as n grows. But for example taking a $a_n = 1$ for all n implies almost surely divergence. On the otherhand if $\sum_{n=1}^{\infty} a_n$ converges, then the series is absolutely convergent, then whatever signs we take, the series converges almost surely.

(Random power series). Let a $Y_1, Y_2, \dots$ be independent sequence of random variables. The power series $F(z) = \sum_{n=0}^{\infty} Y_n z^n$ with the random coefficients defines a random function of real (or complex) variable z. Only difference to the previous examples is that it is a spexial case of a random series $X_n = Y_n z^n$ and indexing starting from zero as it is natural for a power series.

We know that a power series has some radius of convergence $|z| < R$ where by Cauchy-Hadamart $R = \frac{1}{\limsup_n \sqrt[n]{|Y_n|}}$. Note that $x \mapsto \sqrt[n]{|x|}$ is Borel since it is continuous, so $\limsup_n \sqrt[n]{|Y_n|}$ is a random variable. Taking the reciprocal $s \mapsto 1/s$ on $[0, +\infty]$ is continuous, so we see that R is a random variable. Also the radius of convergence R is tail measurable (limsup is insensitive to finite number of chances in the coefficients). Therefore, having assumed the independence of the coefficients, Kolmogorov's 1-0 law tells that R is almost surely a constant. In conclusion, the radius of convergence of this random power series is in fact essentially deterministic (non-random)!

Random walks

Example: (Escape probability of assymetric but simple random walk). Let $\theta \in$

$[0,1]$ be a parameter and $X_1, X_2, \dots$ a sequence of independent and indentically distributed $\{\pm 1\}$–-valued random variables with $\mathbf{P}[X_n = 1] = \theta, \mathbf{P}[X = -1] = 1 - \theta$. Previous examples show that $\sum_{i=1}^{\infty} X_i$ diverges almost surely (in this case we had $\theta = 1/2$ but still works in our case). We think of X_n as the nth step of a random walker $X_n = 1$ interpreted as a step forward and $X_n = -1$ as a step backwards, and the parameter θ gives the probability of a forward step. If the walker starts at $0 \in \mathbb{Z}$, then $W_s := \sum_{n=1}^{s} X_n$ after the first s steps. Let us consider the event $E = \{\omega \in \Omega | \sup_s W_s(\omega) = +\infty\}$. This is a tail event since supremum being infinite is a limiting behavior. By Kolmogorov's 1-0 law $\mathbf{P}[E] = 1$ or 0. So either the walker advances arbituary much almost surely or almost surely not. The answer turns out to depend on the parameter θ : one can show that $\mathbf{P}[E] = 1$ iff $\theta \geq 1/2$. It is natural to refine the question of advancement of the walker slightly. By including also considerations of $\limsup_s W_s$ and $\liminf_s W_s$. One can also show that $\{\lim_{s\to\infty} W_s = +\infty\}$ and $\{\lim_{s\to\infty} W_s = -\infty\}$ are tail events. These denote the escaping towards each extreme of the integer line. Kolmogorov's 1-0 law then says that a walk either almost surely escapes to the extreme or almost surely not. It is the case again that this depends on θ. Infact it goes to $+\infty$ iff $\theta > 1/2$ a.s. and $-\infty$ iff $\theta < 1/2$. What about $\theta = 1/2$? The answer is all over the place where it almost surely goes to $+\infty$ and $-\infty$. It thus also returns to origin infinitely often.

6 Integration against probability measures

One of the main motivators of this chapter is a precise mathematical definition of the expected value, which will discussed more in detail after this section. Namely in the probability space $(\Omega, \mathcal{F}, \mathbf{P})$,

$$\mathbf{E}[X] := \int_{\Omega} X(\omega) d\mathbf{P}(\omega).$$

In the following recall the notations $m\mathcal{F}$ of measurable functions with values $[-\infty, \infty]$, $m\mathcal{F}^+$ of measurabel functions of value $[0, +\infty]$. Similarly for simple functions $s\mathcal{F}, s\mathcal{F}^+$. For these classes denote $\int^{\square}$ as the integral for non-negative simple functions, $\int^{+}$ for non negative measurable functions and $\int$ the general Lebesgue integral.

Recall the following approximating lemma used before. For any $f \in m\mathcal{F}^+$, there exists a sequence from below of non-negative simple functions $s_n(s) \uparrow f(s)$ for all $s \in (S, \mathcal{F})$.

6.1 Integration of non-negative simple functions

Add a measure μ to the measurable space $(S, \mathcal{F})$, then

$$\int^{\square} \mathbb{I}_A d\mu := \mu(A).$$

then expanding linearly for a non-negative simple function $s = \sum_{j=1}^{n} a_j \mathbb{I}_{A_j}$,

$$\int^{\square} \sum_{j=1}^{n} a_j \mathbb{I}_{A_j} := \sum_{i=1}^{n} a_j \mu(A_j).$$

This does not depend on the chosen linear combinations of expressing the the function. One also easily shows $\mathbb{R}-$linearity of the integral.

Lemma 9 (Monotonicity of the integral $\int^{\square}$)**.** *Assume non-negative simple functions $h \leq g$ pointwise, then $\int^{\square} hd\mu \leq \int^{\square} gd\mu$.*

Proof. Assuming disjoint partition, for $h = \sum_{i=1}^{n} a_j \mathbb{I}_{A_j}, g = \sum_{i=1}^{m} a_i \mathbb{I}_{B_i}$, we can write $A_i = h^{-1}(a_i), B_j = g^{-1}(b_j)$. Then the sets $A_i \cap B_j$ are the sets where $a_i \leq b_j$ and also we have a disjoint partition $A_j = \coprod_{j=1}^{m} A_i \cap B_j$ and $B_j = \coprod_{i=1}^{n} A_i \cap B_j$. Thus

$$\int^{\square} hd\mu = \sum_{i,j} a_i \mu[A_i \cap B_j] \leq \sum_{i,j} b_j \mu[A_i \cap B_j] = \int^{\square} gd\mu.$$

□

6.2 Integration of non-negative functions

The following definition is motivated to by preservation of monotonicity. So integral of a function f should be atleast as large as a non-negative simple function $h \leq f$.

Definition 14. For $f \in m\mathcal{F}^{+}$, define

$$\int^{+} fd\mu := \sup_{h \in s\mathcal{F}^{+}, 0 \leq h \leq f} \int^{\square} hd\mu.$$

Note that it is possible even over finite measures that this is infinite. It can be easily checked that this is the $\square-$integral again for non-negative simple functions.

Proposition 5 (Monotonicity of the integral for non-negative functions)**.** *Take $f, g \in m\mathcal{F}^{+}$ and $f \leq g$ poitwise. Then*

$$\int^{+} fd\mu \leq \int^{+} gd\mu.$$

Proof.

$$\sup_{h \in s\mathcal{F}^{+}, 0 \leq h \leq f} \int^{\square} hd\mu \leq \sup_{h \in s\mathcal{F}^{+}, 0 \leq h \leq g} \int^{\square} hd\mu$$

since the supremum is over stricly larger set of non-negative simple functions w.r.t. inclusion. □

Proposition 6. *If $f \in m\mathcal{F}^{+}$ and $\int^{+} fd\mu = 0$, then $\mu[\{s \in S | f(s) > 0\}] = 0$.*

Proof. Write $A_n = \{s \in S | f(s) \geq 1/n\} = f^{-1}[1/n, +\infty] \in \mathcal{F}$. Then one can write $\{s \in S | f(s) > 0\} = \bigcup_n A_n =: A$. Define the non-negative simple function $h = \frac{1}{n}\mathbb{I}_{A_n}$ on A_n. Then $0 = \int^+ f d\mu \geq \int^\Box h d\mu = \mu(A_n)/n$ for every $n \in \mathbb{N}$ so $\mu(A_n) = 0$. By subadditivity $\mu(A) = 0$. □

Let A be as in the above proof and assume $\mu(A) = 0$, then

$$\int^+ f d\mu = \sup_{h \in s\mathcal{F}^+, 0 \leq h \leq f} \int^\Box h d\mu = \sup_{h \in s\mathcal{F}^+, 0 \leq h \leq f} \sum_{i=1}^n a_i \mu[B_i] \leq \sup_{h \in s\mathcal{F}^+, 0 \leq h \leq f} \max\{b_i | 1 \leq i \leq n\} \sum_{i=1}^n \mu[B_i].$$

We may assume that A_i is a disjoint partition not containing sets where h vanishes, so $\sum_{i=1}^n \mu[B_i] \leq \mu[A] = 0$ because. thus $\int^+$ over measure zero sets does vanish.

At this state we can introduce a fundamental way of approximating integrands under pointwise monotone sequences. Namely the **monotone convergesnce theorem (MCT)**.

Theorem 6 (Monotone convergence theorem)**.** *Let $f_1, f_2, \dots \in m\mathcal{F}^+$ and $f_n \uparrow f$ as $n \to \infty$, then we have*

$$\int^+ f_n d\mu \uparrow \int^+ f d\mu$$

as $n \to \infty$.

Proof is very but not that complicated.

6.3 Integral for integrable functions

We call measurable functions **integrable** w.r.t. μ if $\int^+ |f| d\mu < +\infty$. Now we may write any measurable $f \in m\mathcal{F}$ by two non-negative measurables $f_+ = \max(f(s), 0), f_- = \min(-f(s), 9)$. Then $f = f_+ - f_-$ and $|f| = f_+ + f_-$. Set Integrable functions are denoted by $\mathcal{L}^1(\mu)$.

Then the general Lebesgue integral for integrable functions is defined as

$$\int f d\mu = \int^+ f d\mu - \int^+ f_- d\mu$$

We often just omit the measure symbol completely when clear context.

Integrating over a measurable set A is then defined by

$$\int_A f d\mu := \int f \mathbb{I}_A d\mu.$$

In finite measurespaces constants are integrable.

Proposition 7 (Properties of the Lebesgue integral)**.**

1. *$\left|\int f\right| \leq \int |f|$ ($\Delta-$ineq)*

2. *The integral is linear and monotonic.*

6.4 Convergence theorem for integrals

Many concepts in probability theory are formulated by limits and integrals so it is essentialy to know when we can interchange the order of these operations. It is well known from measure theory that this is not the general case that one can do it.

Most important theorems.

- (MCT) Monotone convergence theorem
- (DCT) Dominated convergence theorem (practical for expected values special case of MCT)
- (BCT) Bounded convergence theorem

A result used to get DCT, is Fatou's lemma.

Fatou's lemma is a general convergence result for non-negative integrals, which does not even require the limits to exists! It instead addresses lower limits which always exist. A minor drawback is that the conclusion is merely an inequality in one direction.

Lemma 10 (Fatou's lemma)**.** *For any sequence of non-negative measurables $f_1, f_2, \cdots \in m\mathcal{F}^+$ we have*

$$\int \liminf_n f_n d\mu \leq \liminf_n \int f_n d\mu.$$

Proof. Let $g = \liminf_n f_n$. Define $g_k = \inf_{n\geq k}$ which by construction converges pointwise to g. By MCT

$$\int g_k d\mu \uparrow \int g d\mu \text{ , when } k \to \infty.$$

Since $f_n \geq g_k$ for all $n \geq k$ so by monotonicity $\int f_n \geq \int g_k$ for $n \geq k$. So also $\inf_{n\geq k} \int f_n \geq \int g_k$. Letting $k \to \infty$, the lemma follows. □

A strict inequality for Fatou's lemma is given by $f_n = n\mathbb{I}_{[1,1/N]}$.

Lemma 11 (Reverse Fatou's lemma)**.** *Suppose that $f_1, f_2, \cdots \in m\mathcal{F}^+$ of non-negative measurables functions and that there exists $g \in m\mathcal{F}^+$ which uniformly bounds the sequence, $f_n \leq g$ for all $n \in \mathbb{N}$, and which itself is integrable i.e. $\int g < +\infty$. Then we hvae*

$$\limsup_n \int f_n d\mu \leq \int (\limsup_n f_n) d\mu.$$

Proof. Apply Fatou's lemma to the sequence $g - f_n$ and cancel the finite number $\int g$ from both sides. □

Theorem 7 (Dominated convergence theorem)**.** *Suppose that $f_1, f_2, \cdots \in m\mathcal{F}$ is a sequence of measurable functions such that there exists a non-negative measurable function $g \in m\mathcal{F}^+$ which uniformly bounds the absolute values of the sequence, $|f_n| \leq g$ for all $n \in \mathbb{N}$ and g is integrable i.e. $\int g < +\infty$. Then if the pointwise limit $f = \lim_{n\to\infty} f_n$ exists,*

$$\lim_{n\to\infty} \int |f_n - f| = 0 \quad \& \quad \lim_{n\to\infty} f_n = \int f.$$

Proof. We have that $|f_n - f| \leq 2g$ by the triangle inequality. Apply reverse Fatou's lemma to the sequence $|f_n - f|$ so

$$\limsup_n \int |f_n - f| \leq \int \limsup_n |f_n - f| = \int 0 = 0$$

where the last step follows by pointwise convergence.

Second assertion follows from

$$\left| \int f_n - \int f \right| \leq \int |f_n - f| \xrightarrow{n \to \infty} 0.$$

□

Corollary 6 (Bounded convergence theorem)**.** *Take a finite measure space and a sequence $f_1, f_2, \dots \in m\mathcal{F}$ of bounded measurable functions whose pointwise limit f exists. Then*

$$\lim_{n \to \infty} \int f_n = \int f.$$

Proposition 8. *Change of variables Let $(X_1, \mathcal{F}_1, \mu), (X_2, \mathcal{F}_2, T_*\mu)$ be measure spaces, $T \colon X_1 \to X_2$ measurable, $\mu \colon F_1 \to [0, \infty]$ a measure, then the pushfoward measure $T_*\mu[A] = \mu[T^{-1}[A]]$ satisfies $\int_{X_2} f \, d(T_*\mu) = \int_{X_1} f \circ T \, d\mu$.*

Proof. Basically break down to from downward approximation by simple functions and prove it for simple functions. Use monotone convergence. □

7 Expected value

In a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ for a the **expected value** of a random-variable $X \colon \Omega \to \mathbb{R}$ is just the integral against probability measure

$$\mathbf{E}[X] = \int_\Omega X(\omega) d\mathbf{P}(\omega).$$

Integration results give that

- $X = \mathbb{I}_A, \mathbf{E}[X] = \mathbf{P}[A]$.
- If $X = \sum_{j=1}^n a_j \mathbb{I}_{A_j}$ is simple disjoint partition, then $\mathbf{E}[X] = \sum_{j=1}^n a_j \mathbf{P}[A_j]$.[1]
- If X is non-negative, then

$$\mathbf{E}[X] = \sup_{H \in s\mathcal{F}^+, 0 \leq h \leq x} \mathbf{E}[H].$$

- If X is itegrable, meaning $\mathbf{E}[X_+] < +\infty, \mathbf{E}[X_-] < +\infty$, then we set

$$\mathbf{E}[X] = \mathbf{E}[X_+] - \mathbf{E}[X_-].$$

[1] Since the probability mass in a probability space is finite we don't neeg to assume non-negative to be careful.

- Monotonicity $X \leq Y \Rightarrow \mathbf{E}[X] \leq \mathbf{E}[Y]$.
- $\Delta-$ineq $|\mathbf{E}[X]| \leq \mathbf{E}[|X|]$.

Recall the the "law of a random variable" or "Probability distribution of $X\colon \Omega \to S$ is the pushforward measure $P_X[A] := \mathbf{P}[X \in A]$. Next we show how $g(X)$ for Borel g, can be calculated with respect to the law P_x.

Theorem 8. *Let $X\colon \Omega \to \mathbb{R}$ be a random variable with law P_X and $g\colon \mathbb{R} \to \mathbb{R}$ a Borel function. Then $g(X) \in \mathcal{L}^1(\mathbf{P}) \iff g \in \mathcal{L}^1(P_X)$. If either (then both) of the above hold*

$$\mathbf{E}[g(X)] = \int_{\mathbb{R}} g(x) dP_X(x).$$

Proof. Proof follows a "standard machine": (1) indicators, (2) positive simple functions, (3) non-negative measurable functions and (4) finally all measurable functions.

Step 1: Let $g = \mathbb{I}_B$ for all Borel sets B. Then we can write $\mathbb{I}_B(X(\omega)) = \mathbb{I}_{X^{-1}(B)}(\omega)$. Then

$$\int_B X = \mathbf{E}[\mathbb{I}_B(X)] = \mathbf{E}[\mathbb{I}_{X^{-1}(B)}] = \mathbf{P}[X \in B] = P_X[B] = \int_{\mathbb{R}} \mathbb{I}_B dP_X.$$

Step 2: follows from linearity.

Step 3: Take a non-negative Borel g and monotone increasing approximation g_n non-negative simple functions of g. Then by MCT

$$\mathbf{E}[g(X)] = \lim_{n\to\infty} \mathbf{E}[g_n(X)] = \lim_{n\to\infty} \int_{\mathbb{R}} g_n(x) dP_X(x) = \int_{\mathbb{R}} g dP_X(x).$$

Step 4: take any Borel function. Then $g(X)_+ = g_+ \circ X, g(X)_- = g_- \circ X$ are positive Borel. Thus by previous part $\mathbf{E}[g(X)_\pm] = \int_{\mathbb{R}} g_\pm(x) dP_X(x)$. Then by linearity the claim follows. □

Exercise: (Dicrete random numbers). A random variable X is **discrete** if it's image $A = X(\Omega)$ is countable. The probability mass function of a disrete random variable is define by $p_X(x) = \mathbf{P}[X = x]$. Prove that any discrete real-valued random variable satisfies:

(a) $\mathbf{E}[h(X)] = \sum_{x\in A} h(x) p_X(x)$ for all Borel functions $h\colon \mathbb{R} \to [0, \infty)$.

(b) $h(X) \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbf{P})$ iff $\sum_{x\in A} |h(x)| p_X(x) < +\infty$

(c) Explain why the formula in (a) is true for all $h \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}, P_X)$.

Sol:

(a) Notice that on the sets $[X = x]$, h is constant.

$$\begin{aligned}
\mathbf{E}[h(X)] &\triangleq \int_\Omega h(X(\omega))d\mathbf{P}(\omega) \\
&= \sum_{x\in A} \int_{[X=x]} h(x)d\mathbf{P}(\omega) \\
&= \sum_{x\in A} \mathbb{I}_{[X=x]} h(x)\mathbf{P}[X = x] \\
&= \sum_{x\in A} h(x)\mathbf{P}[X = x] \\
&\triangleq \sum_{x\in A} h(x)P_X(x).
\end{aligned}$$

(b) Directly by (b) and last theorem.

(c) h is non-negative.

Exercise: Let X, Y be real-valued random variables that are equal almost surely, that is $\mathbf{P}[X = Y] = 1$.

(a) In order to make sure that $\mathbf{P}[X = Y]$ is meaningful probability, explain why the set is measurable inside.

(b) Prove that the laws of X and Y are the same, and conclude that we in particular have $\mathbf{E}[X] = \mathbf{E}[Y]$ (whenever the expected values exist).

Sol:

(a) First of all $[X = Y] = (X - Y)^{-1}(0)$ so a preimage of Borel set.

(b) First of all if the laws are same

$$\mathbf{E}[X] = \int_{\mathbb{R}} xdP_X(x) = \int_R xdP_Y(x) = \mathbf{E}[X].$$

Now $\mathbf{P}[X \in A] = \mathbf{P}[Y \in A]$ because the set where $Y \neq X$ has probability zero.

7.1 Densities of continuous distributions

Let X be a real-valued random variable. If there exists a Borel function

$$f_X : \mathbb{R} \to [0, +\infty)$$

such that $\mathbf{P}_X[B] := \mathbf{P}[X \in B] = \int_B f_X(x)dx$ for all Borel sets of $B \in \mathcal{B}(\mathbb{R})$. Then we say that X has a **continuous distribution** (or a **continuous law**), and we say that f_X is a **density function** X.

Examples: Let $\mathfrak{m} \in \mathbb{R}, \mathfrak{s} > 0$. A random variable X has a **Gaussian distribution** with mean $\mathfrak{m}$ and variance $\mathfrak{s}^2$ if its distribution is continuous and

$$f_X(x) = \frac{1}{\sqrt{2\pi\mathfrak{s}^2}}(x - \mathfrak{m})^2$$

is a density function of X. Gaussian distributions are also called **normal distributions**, and the particular case of zero mean $\mathfrak{m} = 0$, and unit variance $\mathfrak{s}^2 = 1$, it is called the **standard normal distribution**.

X is said to have **exponential distribution** with parameter $\lambda > 0$, if its distribution is continuous and

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

is a density function of X.

Exercise: (Random numbers with continuous distribution). Assume that X has a continuous law with a density function f_X.

(a) Show that X is integrable iff $\int_{\mathbb{R}} |x| f_X(x)dx < +\infty$.

(b) If X is integrable, show that $\mathbf{E}[X] = \int_{\mathbb{R}} x f_X(x)dx$.

(c) if X is integrable with $\mathfrak{m} := E[X]$ show that variance $\mathrm{Var}(X) := \mathbf{E}[(X - \mathbf{E}[X])^2] = \int_{\mathbb{R}}(x - \mathfrak{m})^2 f_X(x)dx$.

(d) Can a random variabel with a continuous law have more than one density function?

Sol

(a) The following chain of equalities gives both sides

$$\begin{aligned} \int_{\mathbb{R}} |x| f_X(x)dx &= \int_{\mathbb{R}} |x| dP_X(x) \\ &= \mathbf{E}[|X|] \\ &= \int_{\Omega} |X| d\mathbf{P}. \end{aligned}$$

(b) Again

$$\mathbf{E}[X] = \int_{\mathbb{R}} x dP_X(x) = \int_{\mathbb{R}} x f_X(x)dx.$$

(c) Let g be the borel function $x \mapsto (x - \mathfrak{m})^2$. So

$$\begin{aligned} \mathbf{E}[(X - \mathbf{E}[X])^2] &= \mathbf{E}[g(X)] \\ &= \int_{\mathbb{R}} (x - \mathfrak{m})^2 f_X(x)dx \end{aligned}$$

(d) Yes just take two functions that differ on a set measure zero.

Exercise: Calculate the expected value and variance of exponentially distributed random variable.

Sol:

$$\begin{aligned}\mathbf{E}[X] &= \int_0^\infty x\lambda e^{-\lambda x}dx \\ &= \Big|_0^\infty (-xe^{-\lambda x}) - \int_0^\infty -e^{-\lambda x}dx \\ &= 0 - \Big|_0^\infty \frac{1}{\lambda}e^{-\lambda x} \\ &= \frac{1}{\lambda}.\end{aligned}$$

For the variance

$$\begin{aligned}\mathrm{Var}(X) &= \mathbf{E}[(X-\frac{1}{\lambda})^2] \\ &= \int_0^\infty (x-\frac{1}{\lambda})^2\lambda e^{-\lambda x}dx \\ &= \int_0^\infty x^2\lambda e^{-\lambda x}dx + \int_0^\infty -\frac{2}{\lambda}x\lambda e^{-\lambda x}dx + \int_0^\infty \frac{1}{\lambda^2}\lambda e^{-\lambda x}dx\end{aligned}$$

The middle integral is $-\frac{2}{\lambda}\mathbf{E}[X] = -\frac{2}{\lambda^2}$. The right most integral is $\frac{1}{\lambda^2}$ times integrating over the pullback measure over the entire space, therefore the integral is one. For the first integral

$$\begin{aligned}\int_0^\infty x^2\lambda e^{-\lambda x}dx &= \Big|_0^\infty -x^2e^{-\lambda x} - \int_0^\infty -2xe^{-\lambda x}dx \\ &= 0 - \left(\Big|_0^\infty -2x\frac{-1}{\lambda}e^{-\lambda x} - \int_0^\infty \frac{-2}{\lambda}e^{-\lambda x}\right) \\ &= 0 - \left(0 + \frac{2}{\lambda}\mathbf{E}[X]\right) = \frac{2}{\lambda^2}.\end{aligned}$$

Therefore $\mathrm{Var}(X) = \frac{1}{\lambda^2}$.

Exercise: Let $Y = |X|$ where X is real-valued random variable. Assume without loss of generaliry that $x \geq 0$, or the event $Y \leq x$ is empty.

(a) Prove that Y is a random variable.

(b) Assume that we know c.d.f. $F_X(x) = \mathbf{P}[X \leq x]$ of X. What is the c.f.d. of Y?

(c) Assume that X has a continuous distribution with a density function $f_X(x)$. Does Y also have a continuous distribution in this case? If yes write down an expression for a density function f_Y of Y in terms of f_X. If not, explain why not.

Sol:

(a) The absolute value is Borel function.

(b) $F_Y(x) = \mathbf{P}[|X| \leq x] = \mathbf{P}[-x \leq X \leq x] = \mathbf{P}[X \leq x] - \mathbf{P}[X < -x] = F_X(x) - \mathbf{P}[X < -x]$.

(c) If X has a continuous distribution, then $\mathbf{P}[X < -x] = \mathbf{P}[X \leq -x]$, so

$$\begin{aligned}
\mathbf{P}[Y \leq x] &= F_Y(x) \\
&= F_X(x) - F_X(-x) \\
&= \int_{-\infty}^{x} f_X(y)dy - \int_{-\infty}^{-x} f_X(y)dy \\
&= \int_{-x}^{x} f_X(y)dy \\
&= \int_{0}^{x} f_X(y)dy + \int_{-x}^{0} f_X(y)dy \\
&= \int_{0}^{x} f_X(y)dy - \int_{0}^{-x} f_X(y)dy \\
&= \int_{0}^{x} f_X(y)dy + \int_{0}^{x} f_X(-y)dy \\
&= \int_{0}^{x} f_X(y) + f_X(y)dy.
\end{aligned}$$

Def $f_Y(y) = f_X(y) + f_X(y)$. Since $\mathbf{P}[X \leq 0] = 0$, we get $\mathbf{P}[Y \leq x] = \int_0^x f_Y(x)dx$.

7.2 Convergence theorems applied for expectation

For non negative variables

Lemma 12. *Let $X \in m\mathcal{F}^+$ such that $\mathbf{E}[X] < +\infty$, then $X < +\infty$ almost surely i.e. $\mathbf{P}[X < +\infty] = 1$.*

Proof. Let $A = \{\omega \in \Omega | X(\omega) = +\infty\}$. Then for any n, $X \geq n\mathbb{I}_A$, so

$$n\mathbf{P}[A] = \mathbf{E}[n\mathbb{I}_A] \leq \mathbf{E}[X].$$

Thus $\mathbf{P}[A] \to 0$ as we $n \to \infty$. □

Next we note that in a random sum with non-negative terms, we are allowed to interchange the order of summation and expected value.

Lemma 13 (Expected value of a series of non-negative random terms)**.** *Let $X_1, X_2, \cdots \in m\mathcal{F}^+$ be a sequence of non-negative random variables. Consider random infinite sequence $\sum_{k=1}^{\infty} X_k$. Then we have $\mathbf{E}[\sum_{k=1}^{\infty} X_k] = \sum_{k=1}^{\infty} \mathbf{E}[X_k]$.*

Proof. By non-negativity, we partial sums are monotonely increasing sequence converging to to entire sequence. Thus

$$\begin{aligned}\mathbf{E}[\sum_{k=1}^{\infty} X_k] &= \mathbf{E}[\lim_{n\to\infty}\sum_{k=1}^{n} X_k] \\ &= \lim_{n\to\infty}\mathbf{E}[\sum_{k=1}^{n} X_k] \text{ (MCT)} \\ &= \lim_{n\to\infty}\sum_{k=1}^{n}\mathbf{E}[X_k] \text{ (Linear)} \\ &= \sum_{k=1}^{\infty}\mathbf{E}[X_k].\end{aligned}$$

□

Proposition 9 (Guaranteeing almost sure convergence of a random series)**.** *Let $X_1, X_2, \dots \in m\mathcal{F}^+$ be a sequence of non-negative random variables. Moreover that we have*

$$\sum_{k=1}^{\infty}\mathbf{E}[X_k] < +\infty.$$

Then almost surely $\sum_{k=1}^{\infty} X_k < +\infty$ and almost surely $X_k \to 0$.

Proof. Notice that series converging is a subevent the series members going to zero. Therefore it is enough to show that

$$\mathbf{P}[\sum_{k=1}^{\infty}\mathbf{E}[X_k] < +\infty] = 1$$

so the other one follows form monotonicity. This is a direct consequence of the previous two lemmas. □

Indicators of events are in particular non-negative random variables. The firts Borel-Cantelli lemma can be seen as a consequece of the above obvervations.

As an application of the dominated convergence theorem, we can verify that expected values and differentiation work relative nicely together.

Exercise: Let $X\colon \Omega \to \mathbb{R}$ be a real-valued random variable. Suppose that $h\colon \mathbb{R}\times(a,b) \to \mathbb{R}$ is continuous, $(x,\lambda) = h(x,\lambda)$. Assume that the partial derivative w.r.t. λ exists and continuous. Assume moreover that for some integrable random variable $Y \in \mathcal{L}^1(\Omega,\mathcal{F},\mathbf{P})$ we have that: for all $\lambda \in (a,b)$ and $\omega \in \Omega$,

$$\left|\frac{\partial h}{\partial \lambda}(X(\omega),\lambda)\right| \le Y(\omega).$$

(a) Show that for any $\lambda, \lambda' \in (a,b)$ with $\lambda \neq \lambda'$ we have

$$\left|\frac{h(X,\lambda') - h(X,\lambda)}{\lambda' - \lambda}\right| \le Y.$$

(b) Show that

$$\frac{d}{d\lambda}\mathbf{E}[h(X,\lambda)] = \mathbf{E}[\frac{\partial h}{\partial \lambda}(X,\lambda)].$$

Sol:

(a)

$$\begin{aligned}|h(X(\omega),\lambda_2) - h(X(\omega),\lambda_2)| &= \left|\int_{\lambda_1}^{\lambda_2} \frac{\partial h(X(\gamma),\lambda)}{\partial \lambda} d\lambda\right| \\ &\leq \int_{\lambda_1}^{\lambda_2} \left|\frac{\partial h(X(\gamma),\lambda)}{\partial \lambda}\right| d\lambda \\ &= \int_{\lambda_1}^{\lambda_2} Y(\omega) \\ &= Y(\omega)(\lambda_2 - \lambda_1).\end{aligned}$$

(b) Assume $\lambda_n \to \lambda$, then using DCT on Y we may pass a limit inside

$$\begin{aligned}\frac{d}{d\lambda}\mathbf{E}[h(X,\lambda)] &= \lim_{n\to\infty} \frac{1}{\lambda_n - \lambda}(\mathbf{E}[h(X,\lambda_n)] - \mathbf{E}[h(X,\lambda)]) \\ &= \lim_{n\to\infty} \mathbf{E}\left[\frac{h(X,\lambda_n) - h(X,\lambda)}{\lambda_n - \lambda}\right] \\ &= \mathbf{E}\left[\lim_{n\to\infty} \frac{h(X,\lambda_n) - h(X,\lambda)}{\lambda_n - \lambda}\right] \\ &= \mathbf{E}\left[\frac{dh}{d\lambda}(X,\lambda).\right]\end{aligned}$$

Exercise: (Differentiating the moment generating function).
Define the moment generating function of $X\colon \Omega \to \mathbb{R}$ by $M_X(\lambda) := \mathbf{E}[e^{\lambda X}]$. Assume that $\mathbb{E}[e^{\varepsilon|X|}] < \infty$ for some $\varepsilon > 0$.

(a) Show that $M_X'(0) = \mathbf{E}[X]$.

(b) Explain, without detailed calculations, why $M_X''(0) = \mathbb{E}[X^2]$. Find also a similar formula for $\mathbf{E}[X^n]$ for all $n \in \mathbb{N}$.

(a) First we show that $|xe^{\lambda x}| \leq Ce^{\varepsilon|x|}$. We will actually prove that $|xe^{|\lambda||x|}| \leq Ce^{\varepsilon|x|}$ that $|x| \leq Ce^{(\varepsilon-\lambda)|x|}$. Let $y = |x|, \delta = \varepsilon - \lambda > 0$. Thus the equation becomes $|x|e^{-(\varepsilon-\lambda)|x|}$. Thus we look at the critical points of $f(y) = ye^{-\delta y}$. By differentiating $f'(y) = e^{\delta y} - y\delta e^{\delta y}$ which has only one critical point at $y = \frac{1}{\delta}$. Clearly $f'(0) > 0$ and $f'(2/\delta) = e^{-2y} - 2e^{-2y} < 0$, so has f a global maximum which we way call C. Now let $h(x,\lambda) = e^{\lambda x}$,

$$\begin{aligned}\left|\frac{d}{d\lambda}h(X(\omega),\lambda)\right| &= \left|X(\omega)e^{\lambda X(\omega)}\right| \\ &\leq Ce^{\varepsilon|x|}\end{aligned}$$

So we have an integrable bound, and therefore we can move the partial derivative inside the integral. Meaning

$$\frac{d}{d\lambda}\mathbf{E}[h(X,\lambda)] = \mathbf{E}[\frac{\partial h}{\partial \lambda}(X,\lambda)] = \mathbf{E}[Xe^{\lambda X}]$$

which evaluated at zero is cleary what we wanted.

(b) Every time we take a derivative we just multiply drop X from the exponent. But this assumes we show that each of the derivatives is bounded by integrable function $|x^n e^{\lambda x}| \leq g$. This results to looking at the critical points of $f(y) = y^n e^{-\delta y}$ for $\delta \geq 0$. It has critical points $y = 0, y = n/\delta$. Since we care about the global behavior and we know that $y = 0$ is a minimum, we need to know if f is downward monotonic to the right of n/δ. Now $f'(y) = y^{n-1}(n - \delta y)e^{-\delta y})$. The non positive term is $(n - \delta)$ which for right of n/δ becomes a negative term. Therefore n/δ is a global maximum so we can pick it as the C.

7.3 $p-$integrability

This is for most analysis students a familiar concept in the realm of L^p-spaces. In probability we call random variable *p-integrable* if $\mathbf{E}[|X|^p] < +\infty$. This concludes if it is meaningful to even talk about the *moment* $\mathbf{E}[X^p]$ or order p. One may remember from analysis that $p-$integrable does not imply $p'-$integrable for for $p' \leq p$ (spaces of infinite mass). But for probability measures this holds.

Lemma 14 (Finiteness of lower order moments.). *Let $0 < r < p$. Suppose that $X \in \mathcal{L}^p(\mathsf{P})$. Then we have $X \in \mathcal{L}^r(\mathsf{P})$, and moreover*

$$\mathsf{E}[|X|^r] \leq 1 + \mathsf{E}[|X|^p].$$

Proof. We have that $|x|^r < |x|^p$ whenever $|x| > 1$ $|x|^r \leq 1$ whenever $|x| \leq 1$. Define the event

$$A = \{\omega \in \Omega | X(\omega)| > 1\}.$$

Then by a pointwise estimate

$|X(\omega)|^r \leq \mathbb{I}_{A^c}(\omega) + \mathbb{I}_A|X(\omega)|^p \leq 1 + |X(\omega)|^p$ which inside expected value means

$$\mathbf{E}[|X|^r] \leq 1 + \mathbf{E}[|X|^p] < +\infty.$$

□

Lemma 15. *The set $\mathcal{L}^p(\Omega, \mathcal{F}, P)$ forms a vector space under addition and scalar multiplication. In addition $\mathbf{E}[|X+Y|^p] \leq 2^p(\mathbf{E}[|X|^p] + \mathbf{E}[|Y|^p])$.*

Proof. The vector space axioms are trivial to check the other claim. By the triangle inequality $|x + y| \leq |x| + |y| \leq 2\max(|x|, |y|)$. Thus

$$|x + y|^p \leq 2^p \max(|x|^p, |y|^p) \leq 2^p(|x|^p + |y|^p).$$

Plug inside expected value and we are done. □

Corollary 7. *Since a constant is $p-$integrbale, $X + c$ is $p-$integrable for any X which is $p-$integrable random variable.*

Note that this relies on finiteness of the measure space.

The bounds on the previous lemmas were maybe not something that special so we want to improve them.

Exercise: (Jensen's ineuqality). Let $I \subset \mathbb{R}$ be an interval and $\phi \colon I \to \mathbb{R}$ convex and X a random-variable with image contained in I. Show that $\phi(\mathbf{E}[X]) \leq \mathbf{E}[\phi(X)]$.

Proof. A proof without any differentiability assumptions. First we show the existance of subgradient. Fix $x \in I$, because the interval might be open, closed, half closed we will look at left and right difference quotients. By convexity for $t > h$

$$\phi(x+h) = \phi\left(\frac{h}{t}(x+t) + \left(1 - \frac{h}{t}\right)x\right) \leq \frac{h}{t}\phi(x+t) + \left(\frac{h}{t} - 1\right)\phi(x)$$

This gives that

$$\frac{\phi(x+h) - \phi(x)}{h} \leq \frac{\phi(x+t) - \phi(x)}{t}.$$

□

Therefore the difference quotient is upwards monotone function of h. This implies that for the fixed x the left and right limits of the different quotient at x exist. Denote these by q_L, q_R for L =left and R =right so

$q_L(x) \leq q_R(x)$. Because we are dealing with any interval it may contain end points, so we will deal with this with $z < x$ and $z > x$ separately.

Case 1: $z < x$ then

$$\frac{\phi(z) - \phi(x)}{x - z} \leq q_L(x)$$

so $\phi(z) - q_L(x)(z - x) \leq \phi(x)$. Let $d = q_L(x) \in \mathbb{R}$ so $\phi(z) - q_L(x)(z - x) = \phi(z) + d(x - z)$.

Case 2: if $z > x$

$$\frac{\phi(z) - \phi(x)}{x - z} \geq q_R(x)$$

so $\phi(x) \geq \phi(z) + q_R(x - z)$ with $d = q_R$.

Hence we can always find a d such that $\phi(x) \geq \phi(z) + (x - z)d$ for all $z \in I$. Since X has image contained.

For next step pick $z = \mathbf{E}[X]$ which is I because the image X is contained in an interval. This is because if $a \leq X \leq b \Rightarrow a \leq \mathbf{E}[X] \leq b$ by monotoninity.

$$\begin{aligned}\mathbf{E}[\phi(X)] &\geq \mathbf{E}[\phi(\mathbf{E}[X]) + (X - \mathbf{E}[X])d] \\ &= \phi(\mathbf{E}[X])\mathbf{E}[1] + d\mathbf{E}[X - \mathbf{E}[X]] \\ &= \phi(\mathbf{E}[X]) + d(\mathbf{E}[X] - \mathbf{E}[X]) = \phi(\mathbf{E}[X]).\end{aligned}$$

Exercise: (The p-norm controls lower norms). Suppose that $X \in \mathcal{L}^p(\mathrm{P})$ for some $p > 0$. Let $0 < r < p$. Using Exercise previous exercise, show that

$$(\mathbf{E}[|X|^r])^{1/r} \leq (\mathbf{E}[|X|^p])^{1/p},$$

Sol: the funcion $(\cdot)^r$ is convex on non-negative reals so

$$\mathbf{E}[|X|^r]^{p/r} \leq \mathbf{E}(|X|^p)$$

Taking pth root yields the claim.

8 Product spaces and Fubini's theorem

The product space of possible outcomes is simply the cartesian product. We need to construct a meaningful sigma-algebra on the product space and measures probability measures aswell. We will then recall Fubini's theorem in the context of probability to make sense of the order of integration for multivariable functions. Key tool here is the monotone class theorem.

For recalling, $m\mathcal{J}, s\mathcal{J}, b\mathcal{J}$ measurable, simple, bounded functions and with $+$ meaning the positive versions.

a collection $\mathcal{H}$ of functions $S \to \mathbb{R}$ is said to be a monotone class if:

(MC-1) The constant function 1 belongs to $\mathcal{H}$.

(MC-$\mathbb{R}$) The class $\mathcal{H}$ is a vector space over $\mathbb{R}$.

(MC-$\uparrow$) If $f_1, f_2, \ldots \in \mathcal{H}$ is an increasing sequence of non-negative functions in $\mathcal{H}$ such that the pointwise limit $f_n \uparrow f$ is a bounded function f, then $f \in \mathcal{H}$.

The statement of the Monotone Class Theorem is the following.

Theorem 9 (Monotone class theorem, Theorem C.2). *Let $\mathcal{H}$ be a monotone class of bounded functions from S to $\mathbb{R}$ and let $\mathcal{J}$ be a π-system on S such that $\sigma(\mathcal{J}) = \mathcal{F}$. Suppose that*

$$\mathbb{I}_A \in \mathcal{H} \quad \textit{for every } A \in \mathcal{J}.$$

Then we have

$$b\mathcal{F} \subset \mathcal{H}.$$

8.1 Product sigma-algebra

Let $(S_1, \mathcal{J}_1), (S_2, \mathcal{J}_2)$ be two sigma-algebras.

Definition 15. The **product sigma-algebra** of $\mathcal{J}_1 \otimes \mathcal{J}_2$ on $S_1 \times S_2$ is the sigma-algebra generated by the projections $\mathrm{pr}_i \colon S_1 \times S_2 \to S_i, i = 1, 2$ where it is the smallest sigma-algebra for which these functions are measurable.

Lemma 16. *The collection $\mathcal{J} = \{A_1 \times A_2 | A_1 \in \mathcal{J}_1, A_2, \in \mathcal{J}_2\}$ is a pi-system on $S_1 \times S_2$ and $\mathcal{J}_1 \otimes \mathcal{J}_2 = \sigma(\mathcal{J})$.*

Proof. Take two product sets $(A \times A') \cap (B \times B') = (A \cap B) \times (A' \cap B')$ which clearly have both components in the respective sigma-algebras. For the next part take any set $A_i \in S_i$ and the preimage $\mathrm{pr}_i^{-1}(A_i)$ is $S_1 \times A_2$ or $A_1 \times S_2$ for $A_i \in \mathcal{J}_i$. Both of these are clearly measurable sets in $\sigma(\mathcal{J})$, so pr_i must be measurable meaning, $\mathcal{J}_1 \otimes \mathcal{J}_2 \subset \sigma(\mathcal{J})$. Now $\sigma(\mathcal{J})$ is the smallest sigma-algebra containing $\mathcal{J}$, but $\mathcal{J} \subset \mathcal{J}_1 \otimes \mathcal{J}_2$, so the other inclusion follows. One can see this from writing $A_1 \times A_2 = \mathrm{pr}_1^{-1} A_1 \cap \mathrm{pr}_2^{-1} A_2$. □

Note that the product sigma-algebra is a product in the sense of category theory, therefore factors uniquely thought the projections. The product sigma-algebra thought is not always the sigma-algebra that we want on the product. For example for the Lebesgue sigma-algebras, the product does not give Lebesgue sigma-algebra on the product space.

However sometimes the product gives the wanted sigma-algebra. For example take $\mathcal{B}(\mathbb{R})$ and $\mathcal{B}(\mathbb{R}^2)$, we claim that product of two Borel sets on a line yields the Borel sets on a plane.

Sol: The Borel sets on a line are generated by open intervals, and a product of two intervals is a box, so the product is boxes. Therefore $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) \subset \mathcal{B}(\mathbb{R}^2)$. For the other inclusion we one can write any open set of $\mathbb{R}^2$ by countable intersection of open boxes. Take any open U, then for any $p \in U$ there exists an open ball or small enough radius to be contained in U. Inside any ball we can pick an open rectangle small enough to be contained in U. This forms a cover for U, and U being second countable, we have a countable subcover. Any rectangle can be written by products of intervals, so therefore $U \in \mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$.

Now using the monotone class theorem we show the factorization property of the product for real valued measurable functions. Assume $f\colon S_1 \times S_2 \to \mathbb{R}$ is measurable, we show that the projection functions $S_1 \to \mathbb{R}, S_2 \to \mathbb{R}$ are also measurable w.r.t. product sigma-algebra.

Lemma 17 (Freezing a coordinate preserves measurability)**.** *Let $\mathcal{H}$ denote the class of functions $f : S_1 \times S_2 \to \mathbb{R}$ such that $f \in b(\mathcal{J}_1 \otimes \mathcal{J}_2)$ and*

$$\forall s_1 \in S_1 \quad s_2 \mapsto f(s_1, s_2) \quad \textit{is } \mathcal{J}_2\textit{-measurable } S_2 \to \mathbb{R}$$
$$\forall s_2 \in S_2 \quad s_1 \mapsto f(s_1, s_2) \quad \textit{is } \mathcal{J}_1\textit{-measurable } S_1 \to \mathbb{R}.$$

Then we have $\mathcal{H} = b(\mathcal{J}_1 \otimes \mathcal{J}_2)$.

Proof. Clearly $\mathcal{H}$ is a monotone class and by definition $\mathcal{H} \subset b(\mathcal{J}_1 \otimes \mathcal{J}_2)$. We will show that $\mathcal{H}$ has indicator functions. Since $\sigma(\mathcal{J}) = \mathcal{J}_1 \otimes \mathcal{J}_2$, the MCT allows us to deduce $\mathcal{H} \supset b(\mathcal{J}_1 \otimes \mathcal{J}_2)$.

Take a product set $A_1 \times A_2$ and the indicator

$$\mathbb{I}_{A_1 \times A_2} \colon S_1 \times S_2 \to \mathbb{R}.$$

It is clear that

$$\mathbb{I}_{A_1\times A_2}(s_1,s_2)=\mathbb{I}_{A_1}(s_1)\mathbb{I}_{A_2}(s_2).$$

Now case $s_2\mapsto \mathbb{I}_{A_1}(s_1)\mathbb{I}_{A_2}(s_2)$ is clearly $\mathcal{J}_2-$measurable, since it is either zero or the indicator on A_2. Similarly for the other function. Therefore indicators belongs to the class $\mathcal{H}$. □

8.2 Product measure

Take two finite measure spaces $(S_1,\mathcal{F}_1,\mu_1)$, $(S_2,\mathcal{F}_2,\mu_2)$. We will frequently use the fact that in finite measure spaces, constants are integrable.

Suppose that $f\colon S_1\times S_2\to\mathbb{R}$ is bounded $\mathcal{F}_1\otimes\mathcal{F}_2-$measurable. The functions $s_2\mapsto f(s_1,s_2), s_1\mapsto f(s_1,s_2)$ are also thus bounded and measurable by the previous lemma, and thus integrable. Thus define

$$\mathfrak{I}_1^f(s_1):=\int_{S_2} f(s_1,s_2)\,d\mu_2(s_2)$$

$$\mathfrak{I}_2^f(s_2):=\int_{S_1} f(s_1,s_2)\,d\mu_1(s_1).$$

Theorem 10 (Fubini's theorem for bounded functions on finite measure spaces)**.** *With the above notions we have the following:*

(i) $s_1\mapsto\mathfrak{I}_1^f(s_1)$ *is bounded* $\mathcal{F}_1-$*measurable* $S_1\to\mathbb{R}$.

(ii) $s_2\mapsto\mathfrak{I}_2^f(s_2)$ *is bounded* $\mathcal{F}_1-$*measurable* $S_2\to\mathbb{R}$.

(iii)

$$\int_{S_1}\mathfrak{I}_1^f(s_1)d\mu_1(s_1)=\int_{S_2}\mathfrak{I}_2^f(s_2)d\mu_2(s_2).$$

Proof. Let $\mathcal{H}$ be the collection of $f\in b(\mathcal{F}_1\otimes\mathcal{F}_2)$ for which (i),(ii) and (iii) hold. Clearly $b(\mathcal{F}_1\otimes\mathcal{F}_2)\supset\mathcal{H}$, so we want to so the other inclusion which follows from MCT. We thus need to show the indicator property to apply it. Firstly

$$\mathfrak{I}^{\mathbb{I}_{A_1\times A_2}}(s_1)=\mu_2[A_2]\mathbb{I}_{A_1}(s_1).$$

This is clearly $\mathcal{F}_1-$measurable and bounded. The other side is symmetric. Hence the assumption for MCT is cleared. Thus it remains to prove that $\mathcal{H}$ is a monotone class. Conditions MC-1 and MC-$\mathbb{R}$ are trivial so we show the increasing condition. Suppose that $f_n(s_1,s_2)\uparrow f(s_1,s_2)$ of non-negative functions for all (s_1,s_2) where f is bounded. Then for fixed s_1 we get that $s_2\mapsto f_n(s_1,s_2)\uparrow s_2\mapsto f(s_1,s_2)$ pointwise. Thus by monotone convergence $\mathfrak{I}_1^{f_n}(s_1)\uparrow\mathfrak{I}_1^f(s_1)$. Since f is bounded and μ_2 finite, $s_1\mapsto\mathfrak{I}_1^f(s_1)$ is bounded. Thus property (i) holds for f. Similarly one gets property (ii). The last step follows from applying monotone convergence twice for the integral, so $f\in\mathcal{H}$. This finishes the proof. □

The above shows that two formulas below are euqla and the following definition is unabbiguous.

Definition 16. The product measure $\mu_1 \otimes \mu_2$ on $S_1 \times S_2$ is defined by

$$(\mu_1 \otimes \mu_2)[B] := \int_{S_1} \left(\int_{S_2} \mathbb{I}_B(s_1, s_2) d\mu_2(s_2) \right) d\mu_1(s_1) = \int_{S_2} \left(\int_{S_1} \mathbb{I}_B(s_1, s_2) d\mu_1(s_1) \right) d\mu_2(s_2)$$

for any $B \in \mathcal{F}_1 \otimes \mathcal{F}_2$.

This is a measure since $\emptyset$ evauluates to zero clearly. Take an increasing sequence of disjoint measurable sets $B_1, B_2, \ldots, \in \mathcal{F}_1 \otimes \mathcal{F}_2$ and let $B = \bigcup_{n\in\mathbb{N}} B_n$. Thus it can be also written as the upward limit of $U_n = B_1 \cup \cdots \cup B_n$. By disjointness $\mathbb{I}_{U_n} = \sum_{i \leq n} \mathbb{I}_{B_i}$. By linearity of the integral

$$(\mu_1 \otimes \mu_2)[U_n] = \sum_{i \leq n} (\mu_1 \otimes \mu_2)[B_n].$$

On the other hand the indicators form an increasing sequence $\mathbb{I}_{U_n} \uparrow \mathbb{I}_B$, so applying monotone convergence twice for both μ_1, μ_2 one gets

$$(\mu_1 \otimes \mu_2)[B] = \lim_{n \to \infty} (\mu_1 \otimes \mu_2)[U_n].$$

Combining these two, one get that

$$(\mu_1 \otimes \mu_2)[\bigcup_{i \in \mathbb{N}} B_i] = \sum_{i=1}^{\infty} (\mu_1 \otimes \mu_2)[B_i].$$

This measure satisfies universal property of a product.

Lemma 18. *The measure $\mu_1 \otimes \mu_2$ is the unique measure ν on $(S_1 \times S_1, \mathcal{F}_1 \otimes \mathcal{F}_2)$ that satisfies $\nu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ for all $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$.*

Proof. An easy calculation shows that $\mu_1 \otimes \mu_2$ satisfies the above condition. Since finite measures can be normalized by the total mass to make them into probability measures. Since these measures agree on the pi-system $\{A_1 \times A_2 | A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}$, by Dynkin's identification they must agree. □

Theorem 11 (Fubini's theorem)**.** *For a function $f \colon S_1 \times S_2 \to [-\infty, +\infty]$, consider the following integrals*

$$\int_{S_1 \times S_2} f d(\mu_1 \otimes \mu_2)$$

$$\int_{S_1} \left(\int_{S_2} f(s_1, s_2) d\mu_2(s_2) \right) d\mu_1(s_1)$$

$$\int_{S_2} \left(\int_{S_1} f(s_1, s_2) d\mu_1(s_1) \right) d\mu_2(s_2).$$

We have:

(a) If f is non-negative and measurable, $f \in m(\mathcal{F}_1 \otimes \mathcal{F}_2)^+$, then the integrals evaluate all in $[0, +\infty]$, and they are all equal.

(b) *If f is integrable, $f \in \mathcal{L}^1(\mu_1 \otimes \mu_2)$, then the integral are all in $\mathbb{R}$ and they are all equal.*

Proof. First for part (a) we claim that it holds for $f \in b(\mathcal{F}_1 \otimes \mathcal{F}_2)$, the above integrals are equal as real numbers. The equality of the last two was infact shown in the bounded Fubuni's theorem already. Let $\mathcal{H}$ be the set of functions $f \in b(\mathcal{F}_1 \otimes \mathcal{F}_2)$ for which we have (i) and (ii) from the bounded Fubini theorem and the above three integrals equal. With minor modifications to the proof of Fubini's theorem for bounded functions one can show that $\mathcal{H}$ is a monotone class and contains indicators, so $\mathcal{H} = b(\mathcal{F}_1 \otimes \mathcal{F}_2)$. In particular all non-negative simple functions are bounded and satisfy the integrals above. Take $f \in m(\mathcal{F}_1 \otimes \mathcal{F}_2)^+$ and non-negative pointwise approximation $f_n \uparrow f$. Hence for any f_n the above properties hold. Using monotone convergence theorem three times for the measures $\mu_1 \otimes \mu_2, \mu_2, \mu_1$ all the equalities follow.

Part (b) take $f \in \mathcal{L}^1(\mu_1 \otimes \mu_2)$, meaning

$$\int_{S_1 \times S_2} |f| d(\mu_1 \otimes \mu_2) < +\infty.$$

Write $f = f_+ - f_-$ so $f_+, f_- \in m(\mathcal{F}_1 \otimes \mathcal{F}_2)^+$. Apply part (a).

□

Remark: In part (b) we assumed the integrability w.r.t. the product measure. By part(a) however, this integrability follows if f is measurable and either of the integrals

$$\int_{S_1} \left(\int_{S_2} |f(s_1, s_2)| d\mu_2(s_2) \right) d\mu_1(s_1) \text{ or } \int_{S_2} \left(\int_{S_1} |f(s_1, s_2)| d\mu_1(s_1) \right) d\mu_2(s_2))$$

are finite

Product of two sigma-finite measures: In general one cannot change the order of integration. Take two measure spaces $([0,1], \mathcal{B}([0,1]), \mathcal{L})$ (Lebesgue measure on borel sets) and $([0,1], \mathcal{B}([0,1]), \mu_{\#})$ (counting measure on Borel sets). Take the functions $f\colon [0,1]^2 \to \mathbb{R}$ by $f(x,y) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x \neq y. \end{cases}$ This function is product measurable, but one can see that the product measure integral order changes the value.

Definition 17. A measure space X is **sigma-finite** if it can be covered by countably measurable sets that have finite measure.

Note that it is always possible to choose a disjoint such cover. Take for example Euclidean space with the Lebesgue measure where we can approximate stuff with countably many balls. On the other hand the counting measure is sigma-finite only when the space is finite.

If one takes two measure spaces $(S_1, \mu_1), (S_2, \mu_2)$ with disjoint covering of finite measure sets $S_1 = \bigcup_{i \in \mathbb{N}} A_i^1, S_2 = \bigcup_{i \in \mathbb{N}} A_i^2$. Recall the truncation measure $A \mapsto \mu[A \cap B]$. Let μ_1^n be the truncation of μ_1 to A_n^1 and similarly for μ_2^n for μ_2 are two finite measure. Then take the product measures $\mu_1^n \otimes \mu_2^m$, then the product of two sigma-finite measures can be defined as the countable sum $\sum_{(n,m)\in\mathbb{N}^2} \mu_1^n \otimes \mu_2^m$. One can verify that the chosen disjoint covering does not effect the construction. Fubini's theorem continuous to hold for the sigma-finite product splitting to the countably many pieces $A_n^1 \times A_m^2 \subset S_1 \times S_2$ of finite measure.

9 Probability on product spaces

In this section the previous construction will be applied to probability.

9.1 Joint laws and densities

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probaility space. We defined the law of a random variable $X : \Omega \to S$ by the pushfoward measure $P_X[B] = \mathbf{P}[X \in B]$. In the case $S = \mathbb{R}$, the law is a probability measure on the Borel sets of $\mathbb{R}$. Recall that we had for Borel functions $h : \mathbb{R} \to \mathbb{R}$ that

$$\mathbf{E}[h(X)] = \int_{\mathbb{R}} h(x) dP_X(x)$$

, when ever $h(X) \in \mathcal{L}^1(\mathbf{P})$ or equivalently $h \in \mathcal{L}^1(P_X)$.

Take two real-valued random variables X, Y on Ω

Definition 18. For X, Y the **joint law (or joint distribution)** of X and Y is the probability measure on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ giveb by

$$P_{X,Y}[A] := \mathbf{P}[(X,Y) \in A].$$

Remark the pair (X, Y) is interpreted as a random vector in $\mathbb{R}^2$ or more precisely $Z : \Omega \to \mathbb{R}^2, \omega \mapsto Z(\omega) = (X(\omega), Y(\omega))$. The reason this is measurable is looking at the preimages of the pi-system $\{B_1 \times B_2 | B_1, B_2 \in \mathcal{B}(\mathbb{R})\}$ which generates $\mathcal{B}(\mathbb{R}^2) = \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$. Analogously to the Borel function theorem w.r.t. the expected values, one can write value of $h(X, Y)$ in terms of the law $P_{X,Y}$ as

$$\mathbf{E}[h(X,Y)] = \int_{\mathbb{R}^2} h(x,y) dP_{X,Y}(x,y),$$

provided that $h(X, Y) \in \mathcal{L}^1(\mathbf{P})$ or equivalently $h \in \mathcal{L}^1(P_{X,Y})$. Idea for the proof is similar for this case, using the Monotone class theorem may be more convenient.

Lemma 19. *The law $P_{X,Y}$ is uniquely characteriz by the property that*

$$P_{X,Y}[B_1 \times B_2] = \mathbf{P}[(X,Y) \in B_1 \times B_2]$$

for all $B_1, B_2 \in \mathcal{B}(\mathbb{R})$.

Proof. Use the above generating pi-system and Dynkin's identification. □

Exercise: (Transition probability kernels) Let K be a transition probability kernel on $(S, \mathcal{F})$, i.e., a mapping $S \times \mathcal{F} \to [0, +\infty)$ denoted by $(s, A) \mapsto K_s[A]$.

- For any $A \in \mathcal{F}$, the map $s \mapsto K_A[A]$ is $\mathcal{F}-$measurable $S \to [0, +\infty)$.
- For any $s \in S$, the mapping $A \mapsto K_s[A]$ is a probability measure on $(S, \mathcal{F})$.

Let μ be another probability measure on $S, \mathcal{F}$

(a) Define μK by

$$(\mu K)[A] = \int_S K_s[A] d\mu(s), \text{ for } A \in \mathcal{F}.$$

Show that μK is a probability measure on $(S, \mathcal{F})$.

(b) Define for $\mathcal{F} \otimes \mathcal{F}-$ measurable subsets $B \subset S \times S$

$$\nu[B] = \int_S \left(\int_S \mathbb{I}_B(s_1, s_2) dK_{s_1}(s_2) \right) d\mu(s_1).$$

Show that ν is a probability measure on $(S \times S, \mathcal{F} \otimes \mathcal{F})$.

(c) Let $X = (X_1, X_2)$ be a random "vector" in $S \times S$ with distribution $P_X = \nu$ given by (b). Show that the distributions P_{X_1}, P_{X_2} of its components X_1, X_2 are μ and μK respectively.

Sol:

1. Since K_s is a probability measure $K_s(\emptyset) = 0$, so $\mu K(\emptyset) = 0$. Also $\mu K(S) = 1$ follow by this fact and μ being a probability measure. Take disjoint sequence $A_1, A_2, \ldots$ of measurables. Then by linearity and monotone convergence

$$\begin{aligned}(\mu K)[A] &= \int_S K_s[\bigcup_n A_n] d\mu(s) \\ &= \sum_n \int_S K_s[A_n] d\mu(s) \\ &= \sum_n (\mu K)[A_n].\end{aligned}$$

2. Again taking $\emptyset$ inside the definition of ν, the indicator terminates everything to zero. Taking the entire space $S \times S$, $\nu(S \times S) = \int_S K_{s_1}(S) d\mu(s_1) = \mu(S) = 1$. Taking a disjoint sequence of sets one just pulls out the sum $\sum_n \mathbb{I}_{A_n} = \mathbb{I}_B$ by linearity and monotone convergence.

3. First we reason why $P_{X_1} = \nu(- \times S)$ and $P_{X_2} = \nu(S \times -)$. The set $\mathbf{P}[X_1 \in A] = \mathbf{P}[b \in S | X_1(b) \in A] = \mathbf{P}[b \in S | X(b) \in A \times S] = \mathbf{P}[X \in A \times S] = \nu[A \times S]$. Similar reasoning works for X_2.

Take set of the form $B \times S$ so

$$\begin{aligned}\nu[B \times S] &= \int_S \left(\int_S \mathbb{I}_{B\times S}(s_1, s_2) dK_{s_1}(s_2) \right) d\mu(s_1) \\ &= \int_S \left(\int_S \mathbb{I}_B(s_1)\mathbb{I}_S(s_2) dK_{s_1}(s_2) \right) d\mu(s_1) \\ &= \int_S \left(\int_S \mathbb{I}_B(s_1) dK_{s_1}(s_2) \right) d\mu(s_1) \\ &= \int_S \mathbb{I}_B(s_1) d\mu(s_1) = \mu[B]\end{aligned}$$

and for sets of the form $S \times B$

$$\begin{aligned}\nu[S \times B] &= \int_S \left(\int_S \mathbb{I}_{S\times B}(s_1, s_2) dK_{s_1}(s_2) \right) d\mu(s_1) \\ &= \int_S \left(\int_S \mathbb{I}_B(s_2) dK_{s_1}(s_2) \right) d\mu(s_1) \\ &= \int_S K_{s_1}[B] d\mu(s_1) = (\mu K)[B].\end{aligned}$$

The notion of joint law of n real valued random variables is a straightforward gener alization and it is nothing but the law of the n-dimensional random vector whose components are the real valued random variables.

Recall from previously that a random variables $X : \Omega \to \mathbb{R}$ is said to have a **continuous distribution** if there is a non-negative Borel function $f_X : \mathbb{R} \to [0, +\infty]$ such that $P_X[B] = \mathbf{P}[X \in B] = \int_B f_X d\Lambda$ for all $B \in \mathcal{B}$. Then f_X is called the **probability density** of (the law of) X. Now on Borel sets the 2-dimensional Lebesgue measure Λ^2 on $\mathbb{R}^2$ is the product measure $\Lambda^2 = \Lambda \otimes \Lambda$. On Borel sets it is just the Area measure. The **joint density** of two real-valued random variables X, Y is a measurable function $f_{X,Y} : \mathbb{R}^2 \to [0, +\infty]$ for which $\mathbf{P}_{X,Y}[A] = \mathbf{P}[(X, Y) \in A] = \int_A f_{X,Y} d\Lambda^2$ for all Borel sets in $\mathbb{R}^2$.

For more than two variables it is again just straight forward generalization.

Proposition 10 (Existance of marginal densities from a joint density)**.** *If X and Y have a joint density $f_{X,Y}$, then X has a density $f_X : \mathbb{R} \to [0, \infty]$ given by currying*

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y) d\Lambda(y)$$

and similarly for f_Y.

Proof. This proof relies on Fubini's theorem. Write

$$P_X[B] = \mathbf{P}[X \in B] = \mathbf{P}[X \in B, Y \in \mathbb{R}] = \mathbf{P}[(X, Y) \in B \times \mathbb{R}].$$

Then by the existance of joint density

$$P_X[B] = \int_{B\times\mathbb{R}} f_{X,Y} d\Lambda^2 = \int_{\mathbb{R}^2} \mathbb{I}_{B\times\mathbb{R}} f_{X,Y} d\Lambda^2.$$

Using the fact that $\mathbb{I}_{B\times\mathbb{R}}(x,y) = \mathbb{I}_B(x)$ and Λ^2 viewed as product measure,

$$\begin{aligned} P_X[B] &= \int_{\mathbb{R}^2} \mathbb{I}_B(x) f_{X,Y}(x,y) d(\Lambda\otimes\Lambda)(x,y) \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathbb{I}_B(x) f_{X,Y}(x,y) d\Lambda(x) \right) d\Lambda(x) \\ &= \int_{\mathbb{R}} \mathbb{I}_B(x) f_X(x) d\Lambda(x). \end{aligned}$$

The other density is similar. □

The converse does not hold thought. Meaning existance of marginal densities does not imply existance of a joint density.

Example: (Marginal densities do not guarantee existence of joint density). Take real valued X with continuous distribution f_X and vector $Z = (X, X)$. The joint law $P_{X,X}$ has support $\{x = y\} \subset \mathbb{R}^2$ which has Lebesgue measure zero.

Also not all probability distributions have densities.

Exercise: Show that any Borel function $\mathbb{R}^d \to \mathbb{R}$ is integrable with respect to the the Dirac measure at a point $a \in \mathbb{R}^d$ defined by $\delta_a[A] = \begin{cases} 1 & \text{, if } a \in A \\ 0 & \text{, else} \end{cases}$. Does the measure δ_a have a probability density function?

Sol: Since $\mathbb{R}^d \setminus \{a\}$ has measure zero, we only care about the integral $\int_{\{a\}} |f| d\delta_a$. A function supported on a singleton, is constant on that singleton, so $\int_{\{a\}} |f| d\delta_a = |f(a)|\delta_a[\{a\}] = |f(a)| < +\infty$. Now do we have a probability density functions, meaning that $\delta_a[A] = \int_{\mathbb{R}} \mathbb{I}_A(x) g(x) dx$ for all $A \in \mathcal{B}$. In the case $a \notin A$, we can just pick constant zero function as the density. If $a \in A$, then $\int_{\mathbb{R}} \mathbb{I}_A(x) g(x) dx = 1$ for all such A. But this cannot happen. Take $a = 1, A = [0,1], B[0,2]$. Then

$$1 = \delta_1[B] = \int_{\mathbb{R}} \mathbb{I}_B g = \int_{\mathbb{R}} \mathbb{I}_A g + \int_{\mathbb{R}} \mathbb{I}_{B\setminus A} g = \delta_1[A] + \delta_1[B \setminus A] = 1 + 1.$$

9.2 Variance and covariance of square integrable random variables

Variances are important statistics of distributions of real valued random variables, and covariances are important statistics of the joint distributions of pairs of random variables. In order for these to be well-defined, we need the second moments of the random variables to exist. This is that $\mathbf{E}[X^2] < +\infty$. In the case of probability measures this implied $\mathcal{L}^1-$integebility. Next we list some well known analysis results without proofs.

Theorem 12 (Cauchy-Schwartz). *For $X, Y \in \mathcal{L}^2(\mathbf{P})$, $|\mathbf{E}[XY]| \leq \sqrt{\mathbf{E}[X^2]\mathbf{E}[Y^2]}$ so therefore XY is integrable.*

Corollary 8. *For $X \in \mathcal{L}^2(\mathbf{P})$, $\mathbf{E}[X]^2 \leq \mathbf{E}[X^2]$.*

Definition 19 (Variance and covariance)**.** The **varience** of $X, Y \in \mathcal{L}^2(\mathbf{P})$ is defined by

$$\operatorname{Var}(X) = \mathbf{E}[(X - \mathbf{E}[X])^2]$$

and the **covariance** by

$$\operatorname{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])].$$

Infact variance is a special case of covarience with $X = Y$. Variance measures the spread from the mean of a random variable. Covariance on the otherhand measures how var from being independent two random variables are.

Proposition 11. *We have that* $\operatorname{Var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$ *and* $\operatorname{Cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$.

9.3 Independence and products

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and X, Y two real-valued random varibles.

Theorem 13 (Independence equivalent condition of random numbers)**.** *Let X, Y be as above, then the following conditions are equivalent.*

(i) $X \perp\!\!\!\perp Y$

(ii) $P_{X,Y} = P_X \otimes P_Y$

(iii) for all $x, y \in \mathbb{R}$ we have that $\mathbf{P}[X \leq x, Y \leq y] = \mathbf{P}[X \leq x]\mathbf{P}[Y \leq y]$

If moreover X, Y have a joint density, then the following statement if equivalent to all above: $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ for Λ^2-almost all (x, y).

Proof. (i) $\iff$ (iii) was already proven in the chapter for independence. Thus we prove (i) $\Rightarrow$ (ii) and (ii) $\Rightarrow$ (iii) to get the equivalence of first three. Then under the extra hypothesis show (ii) $\Rightarrow$ (iv) and (iv) $\Rightarrow$ (iii).

(i) $\Rightarrow$ (ii): ASsume that $X \perp\!\!\!\perp Y$ and take $B_1, B_2 \in \mathcal{B}$ whose preimages $X^{-1}(B_1), Y^{-1}B_2$ are independent events. Thus

$$\begin{aligned} P_{X,Y}[B_1 \times B_2] &= \mathbf{P}[X^{-1}(B_1) \cap Y^{-1}B_2] \\ &= \mathbf{P}[X \in B_1]\mathbf{P}[Y \in B_2] \\ &= P_X[B_1]P_Y[B_2]. \end{aligned}$$

But this property uniquely determines the product measure.

(ii) $\Rightarrow$ (iii): Suppose $P_{X,Y} = P_X \otimes P_Y$, then

$$\begin{aligned} \mathbf{P}[X \leq x, Y \leq y] &= P_{X,Y}[(-\infty, x]] \times (-\infty, y]] \\ &= P_X[(-\infty, x]]P_Y[(-\infty, y]] \\ &= \mathbf{P}[X \leq x]\mathbf{P}[Y \leq y]. \end{aligned}$$

(ii) ⇒ (iv) whilst assuming joint density: Suppose a joint density $f_{X,Y}$ exists and $P_{X,Y} = P_X \otimes P_Y$. By Fubini's theorem, this means that

$$\int_{A\times B} f_{X,Y} d\Lambda^2 = P_{X,Y}[A\times B] = P_X \otimes P_Y[A\times B] = P_X[A]P_Y[B] = \int_A f_X d\Lambda \int_B f_Y d\Lambda$$
$$= \int_A \int_B f_X f_Y d\Lambda d\Lambda = \int_{A\times B} f_X F_Y d\Lambda^2.$$

These are equal hence equal almost everywhere since they are both non-negative functions.

(iv) ⇒ (iii) assuming existance of joint density: If $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for Λ^2-almost all (x,y), then their integrals have to agree. Thus again using Fubini, definitions, set theory and given assumptions

$$\begin{aligned}
\mathbf{P}[X \leq x, Y \leq y] &= \mathbf{P}[X^{-1}(-\infty,x] \cap Y^{-1}(-\infty,y]] \\
&= \mathbf{P}[(X,Y)^{-1}((-\infty,x] \times (-\infty,y])] \\
&= \mathbf{P}[(X,Y) \in (-\infty,x] \times (-\infty,y]] \\
&= P_{X,Y}[(-\infty,x] \times (-\infty,y]] \\
&= \int_{(-\infty,x]\times(-\infty,y]} f_{X,Y} d\Lambda^2 \\
&= \int_{(-\infty,x]\times(-\infty,y]} f_X f_Y d\Lambda^2 \\
&= \int_{(-\infty,x]} f_X d\Lambda \int_{(-\infty,y]} f_Y d\Lambda \\
&= P_X[(-\infty,x]]P_Y[(-\infty,y]] \\
&= \mathbf{P}[X \leq x]\mathbf{P}[Y \leq y]
\end{aligned}$$

□

Exercise: Let $X \perp\!\!\!\perp Y$ be real-valued random variables with laws P_X, P_Y. Let $h\colon \mathbb{R}^2 \to \mathbb{R}$ be Borel, prove that

(a) $\mathbf{E}[|h(X,Y)|] = \int_{\mathbb{R}} \mathbf{E}[|h(x,Y)|]dP_X(x) = \int_{\mathbb{R}} \mathbf{E}[|h(X,y)|]dP_Y(y)$.

(b) If $\mathbf{E}[|h(X,Y)|] < +\infty$,

$$\mathbf{E}[h(X,Y)] = \int_{\mathbb{R}} \mathbf{E}[h(x,Y)]dP_X(x) = \int_{\mathbb{R}} \mathbf{E}[h(X,y)]dP_Y(y).$$

Sol: (a) Using Fubini and the product measure

$$\begin{aligned}
\mathbf{E}[|h(X,Y)|] &= \int_\Omega |h(X,Y)| d\mathbf{P} \\
&= \int_{\mathbb{R}^2} |h(x,y)| dP_{X,Y} \text{ (pushforward measure change of variables)} \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} |h(x,y)| dP_X dP_y \text{ (Fubini and product measure from the independence)}
\end{aligned}$$

Now one can clearly see by changing the order of integration that the equalities hold. For (b) the same thing apply fubini, since we got the integrability assumption.

Exercise: (Product of two uniformly distributed numbers) Let U_1, U_2 two independent uniformly distributed random variables on $[0,1]$, so that both have $\mathbb{I}_{[0,1]}$ as their density function. Define $X = U_1U_2$.

(a) Calculate the c.d.f $F_X(x) = \mathbf{P}[X \leq x]$ of X.

(b) What is the distribution $P_X[B] = \mathbf{P}[X \in B]$ of X?

(c) Does X have a probability density function?

Sol: For (a). Notice that U_1, U_2 are almost surely non-negative given by the density functions

$$\mathbf{P}[U_1 \leq x] = \int_{-\infty}^{x} \mathbb{I}_{[0,1]} d\Lambda = \mathbf{P}[U_2 \leq x].$$

First we show that uniformly distributed real-valued random variables have a joint distribution. For independent variables the joint distribution should be $f_{U_1}f_{U_1}$ so we should verify it that this is a joint distribution. Take any open Borel set, which may be assumed to be a box $B = I \times J = [a,b] \times [c,d]$ since any open set can be written as a countable union of boxes. Then any Borel set we can just integrate over the interior.

$$\begin{aligned}
\int_B f_{U_1}f_{U_1} d\lambda^2 &= \int_{\mathbb{R}^2} \mathbb{I}_{[0,1]^2} d\lambda^2 \\
&= \int_J \mathbb{I}_{[0,1]} \int_I \mathbb{I}_{[0,1]} dv du \text{ (Fubini)} \\
&= \mathbf{P}[U_1 \in J]\mathbf{P}[U_2 \in I] \\
&= \mathbf{P}[U_1 \in J, U_2 \in I] \text{ (indepencence).} \\
&= \mathbf{P}[(U_1, U_2) \in B]
\end{aligned}$$

Thus (U_1, U_2) has the density function $f_{U_1}f_{U_1}$. Rewrite the set $U_1U_2 \leq x$ as a two dimensional event. $A = \{(a,b) | ab \leq x\}$, then

$$\begin{aligned}
\mathbf{P}[U_1U_2 \leq x] &= \mathbf{P}[(U_1, U_2) \in A] \\
&= \int_A \mathbb{I}_{[0,1]^2} d\lambda^2 \\
&= \int_{A \cap [0,1]^2} d\lambda^2 \\
&= |A \cap [0,1]^2|
\end{aligned}$$

We can get one of the middle integrals to deal with

$$\begin{aligned}\int_A \mathbb{I}_{[0,1]^2} d\lambda^2 &= \int_{[0,1]^2} \mathbb{I}_{A\cap[0,1]^2} d\lambda^2 \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{I}_{A\cap[0,1]^2}(u,v)dudv \\ &= \int_0^1\int_0^1 \mathbb{I}_A(u,v)dudv \\ &= \int_0^1\int_0^{x/v} dudv \\ &= \int_0^1 \min(1,\frac{x}{v})dv \\ &= \int_0^x 1dv + \int_x^1 \frac{x}{v}dv \\ &= x - x\log x.\end{aligned}$$

This gets us (a). For (b) we calculate the density function of X so lets do (c) first.

Notice that $\mathbf{P}[X \in B]$, we can think of $B \subset (0,1)$ as a countable union of open intervals. Therefore lets look at the situation $\mathbf{P}[a \leq X \leq B] = F_X(b) - F_X(a) = \int_a^b \frac{d}{dx}F_X dx$. On our interval function is differentiable, so $f_X(x) = \frac{d}{dx}F_X(x) = 1 - \log x - 1 = \log x$.

Lastly for (b) we get that $P_X(B) = \mathbf{P}[X \in B] = \int_B -\mathbb{I}_{[0,1]} \log x dx$ since in the calculation we assumed $B \subset (0,1)$.

9.4 Independence and expected value

Theorem 14 (Expected value of independent real-valued random variables)**.** *Suppose that X, Y are real-valued integrable random variables that are independent and that their product is integrable. Then*

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y].$$

Proof. We prove the statement using the "standard machine", i.e., successively for (1): indicators, (2): simple random variables, (3): non-negative random variables, and (4): all integrable random variables.

Step 1: Taking two indicators of independent events $\mathbb{I}_A, \mathbb{I}_B$, then $\mathbf{E}[\mathbb{I}_A\mathbb{I}_B] = \mathbf{E}[\mathbb{I}_{A\cap B}] = \mathbf{P}[A\cap B] = \mathbf{P}[A]\mathbf{P}[B] = \mathbf{E}[\mathbb{I}_A]\mathbf{E}[\mathbb{I}_B]$.

Step 2: Two simples $X = \sum_i a_i\mathbb{I}_{A_i}$, $Y = \sum_j b_j\mathbb{I}_{B_j}$ assuming disjoint partition. Then each set $A_i = X^{-1}(a_i), B_j = Y^{-1}(b_j)$ implies for all i, j, $A_i \perp\!\!\!\perp B_j$. Then one expands linearly and uses the first step.

Step 3: For non-negatives X, Y and take the staircase approximation of $s_n \uparrow \mathrm{id}_{\mathbb{R}}$ and define $X_n = s_n \circ X$ and $Y_n = s_n \circ Y$, so by construction pointwise $X_n \uparrow X, Y_n \uparrow Y, X_nY_n \uparrow XY$. Then just apply monotone convergence for the expected value.

Step 4: For integrable random variables split into positive and negative parts. □

Proposition 12 (Variance is additive for independent random variables)**.** *If X, Y are square integrable and independent, then* $\mathrm{Cov}(X, Y) = 0$ *and* $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$*.*

Proof. Since $\mathrm{Cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$, the claim follows from previous theorem. For the variance part use one of the defining equalities for variance so,

$$\begin{aligned}\mathsf{Var}(X+Y) &= \mathsf{E}[(X+Y)^2] - \mathsf{E}[X+Y]^2 \\ &= \mathsf{E}[X^2 + 2XY + Y^2] - (\mathsf{E}[X] + \mathsf{E}[Y])^2 \\ &= \mathsf{E}[X^2] + 2\mathsf{E}[XY] + \mathsf{E}[Y^2] - \mathsf{E}[X]^2 - 2\mathsf{E}[X]\mathsf{E}[Y] - \mathsf{E}[Y]^2.\end{aligned}$$

Using previous theorem, two terms vanish which yields the result. □

The additivity generalizes for n independent variables. Note that the independence assumption is crucial.

9.5 A formula for expected value of non-negative random variable and it's moments

Proposition 13 (Expected value in terms of c.d.f.)**.** *For a non-negative real-valued random variable X, we have that*

$$\mathbf{E}[X] = \int_0^\infty (1 - F_X(x))dx.$$

Proof. Consider the product spcae $\Omega \times [0, +\infty)$ equipped with product sigma-algebra $\mathcal{F} \otimes \mathcal{B}([0, +\infty))$ and the product measure $\mathbf{P} \otimes \Lambda$. The subset $A = \{(\omega, t) | t < X(\omega)\} \subset \Omega \times [0, +\infty)$ is measurable. It can be written as $(X \circ \mathrm{pr}_1 - \mathrm{pr}_2)^{-1}(0, +\infty)$. Therefore it's indicator is non-negative measurable

$$\mathbb{I}_A(\omega, t) = \begin{cases} 1, & \text{if } t < X(\omega) \\ 0, & \text{if } t \geq X(\omega). \end{cases}$$

By Fubini

$$\int_0^\infty \left(\int_\Omega \mathbb{I}_A(\omega, t) d\mathbf{P}(\omega) \right) dt = \int_\Omega \left(\int_0^\infty \mathbb{I}_A(\omega, t) dt \right) d\mathbf{P}(\omega).$$

The LHS evalueates to $1 - \mathbf{P}[t \geq X] = 1 - F_X(t)$. The RHS becomes the expected value. □

Exercise: Show that for $p-$integrable X non-negative

$$\mathbf{E}[X^p] = p \int_0^\infty t^{p-1}(1 - F_X(t))dt$$

Write $x^p = \int_0^x \frac{t^{p-1}}{p} dt = \int_0^\infty \mathbb{I}_{(0,x)} p t^{p-1} dt$. Then

$$\begin{aligned}
\mathbf{E}[X^p] &= \int_\Omega \left(\int_0^\infty \mathbb{I}_{(0,X(\omega))} p t^{p-1} dt \right) d\mathbf{P}(\omega) \\
&= \int_0^\infty p t^{p-1} \left(\int_\Omega \mathbb{I}_{(0,X(\omega))} d\mathbf{P}(\omega) \right) dt \\
&= \int_0^\infty p t^{p-1} \mathbf{P}[0 \le t \le X] dt \\
&= \int_0^\infty p t^{p-1} \mathbf{P}[t \le X] dt \text{ non-negativity of } t \\
&= \int_0^\infty p t^{p-1} (1 - F_X(t)) dt.
\end{aligned}$$

Exercise: Given $p-$integrable non-negative X, X is $p-$integrable iff $|X|^p$ is integrable. Assume that for $\alpha > 0$ the c.f.t. F_X of X satisfies $\lim_{x \to \infty}(x^\alpha(1 - F_X(x))) = c > 0$. Prove that $X \in \mathcal{L}^p$ iff $p < \alpha$.

Sol: If $X \in \mathcal{L}^p$, then

$$+\infty > E[X^p] = p \int_0^\infty t^{p-1}(1 - F_X(t)) dt$$

which means that $t^{p-1}(1 - F_X(t)) \to 0$ as $t \to \infty$. Thus clearly $p - 1 < \alpha$. Conversly the limit does not go to zero, the integral will not converge.

10 Notions of convegence in probability theory

Notions of convergence in probability theory Pointwise convergence of a sequence of random variables is often too much to ask.

Definition 20. We say that X_n tends to X *almost surely* as $n \to \infty$, if

$$\mathsf{P}\left[\lim_{n \to \infty} X_n = X\right] = 1.$$

In this case we denote $X_n \xrightarrow{\text{a.s.}} X$.

This probabilistic notion of a limit should be intuitively easy to understand — we are giving up pointwise convergence only on an exceptional event E^c which has probability zero. Although we have thus relaxed the extremely stringent requirement of pointwise convergence, this is still a very strong notion of convergence.

Occasionally almost sure convergence is sitll too much to hope for. We have even less restrictive notion.

Definition 21. We say that X_n tends to X *in probability* as $n \to \infty$, if for all $\varepsilon > 0$ we have

$$\lim_{n \to \infty} \mathsf{P}\left[|X_n - X| < \varepsilon\right] = 1. \tag{XI.3}$$

In this case we denote $X_n \xrightarrow{\mathsf{P}} X$.

Often we instead of look equivalently when $\lim\limits_{n\to\infty} \mathbf{P}[|X_n - X| \geq \varepsilon] = 0$ because there exists many techniques to upper bound probabilities.

One may check that reverse of these implications does not hold. There is some form of reverse implications thought. Convergence in probability implies almost sure convergence along a subsequence.

Proof: Assume $\lim_{n\to\infty} \mathbf{P}[|X_n - X| < \varepsilon] = 1$. Take sequences $1/n$ and $1/2^m$, since we know that $\lim\limits_{n\to\infty} \mathbf{P}[|X_n - X| \geq \varepsilon]$ we can always pick some n_k for which $\mathbf{P}[|X_{n_k} - X| \geq \frac{1}{k}] \leq \frac{1}{2^k}$. Let $E_k = \{|X_{n_k} - X| \geq \frac{1}{k}\}$, so $\sum_{k\in\mathbb{N}} \mathbf{P}[E_k] \leq \sum_{k\in\mathbb{N}} \frac{1}{2^k} < +\infty$. Then by Borel-Cantelli $\mathbf{P}[E_k \text{ i.o.}] = 0$ which is by definition

$$\mathbf{P}[\limsup_{k\to\infty} |X_{n_k} - X| \geq \frac{1}{k}] = 0.$$

Why this implies almost sure convergence is that $\mathbf{P}[\limsup_{k\to\infty} |X_{n_k} - X| \neq 0]$ is upper bounded by $\mathbf{P}[\limsup_{k\to\infty} |X_{n_k} - X| \geq \frac{1}{k}]$.

Exercise: (Continuous transformations and convergence in probability) Let $X, X_1, X_2, \ldots$ sequence of real-valued random variables and suppose $X_n \xrightarrow{\mathbf{P}} X$.

(a) For uniformly continuous $h\colon \mathbb{R} \to \mathbb{R}$, show that $h(X_n) \xrightarrow{\mathbf{P}} h(X)$. The first case also holds for just continuous h. Prove it.

(b) For any bounded uniformly continuous $h\colon \mathbb{R} \to \mathbb{R}$, show we have

$$\mathbf{E}[|h(X_n) - h(X)|] \to 0 \quad \& \quad \mathbf{E}[h(X_n)] \to \mathbf{E}[h(X)].$$

The conclusions are not valid without the assumption of boundedness. Can you give a counter example in that case?

Sol:

(a) By uniform continuity for all $\varepsilon > 0$ there exists $\delta > 0$ such that for every pair of points x, y, $|x - y| < \delta \Rightarrow |h(x) - h(y)| < \varepsilon$. Thus

$$\mathbf{P}[|h(X_n) - h(X)| < \varepsilon] \geq \mathbf{P}[|X_n - X| < \delta] \xrightarrow{n\to\infty} 1.$$

How would be argue this for for an arbituary continuous function? We show that $\mathbf{P}[|h(X_n) - h(X)| \geq \varepsilon] < k$ for any positive K. We can pick some M large enough that $\mathbf{P}[|X| \geq M] < k/2$ and on the interval $[-(M+1), M+1]$ we can use uniform convergence to find $\delta_M < 1$ relying on M to get that $\mathbf{P}[|X_n - X| \geq \delta_M] < k/2$ for large n. These assumptions yield that if $|X| \leq M$,

$$|X_n| \leq |X| + |X_n - X| \leq M + 1.$$

So X_n must be on the interval $[-(M+1), M+1]$. Therefore we get two different cases what may happen, meaning upper bound

$$\mathbf{P}[|h(X_n) - h(X)| \geq \varepsilon] \leq \mathbf{P}[|X_n - X| \geq \delta_M] + \mathbf{P}[|X| \geq M] \leq k/2 + k/2 = k.$$

(b) Let δ be the uniform constant for ϵ and M the uniform bound $|h| \leq M$. Then one gets

$$\begin{aligned}\mathbf{E}[|h(X_n) - h(X)|] &= \int_\Omega |h(X_n) - h(X)| d\mathbf{P} \\ &= \int_{\{|h(X_n)-h(X)|<\varepsilon\}} |h(X_n) - h(X)| d\mathbf{P} + \int_{\{|h(X_n)-h(X)|\geq\varepsilon\}} |h(X_n) - h(X)| d\mathbf{P} \\ &\leq \int_{\{|h(X_n)-h(X)|<\varepsilon\}} \varepsilon d\mathbf{P} + \int_{\{|h(X_n)-h(X)|\geq\varepsilon\}} 2M d\mathbf{P} \\ &= \varepsilon\mathbf{P}[|h(X_n) - h(X)| < \varepsilon\}] + 2M\mathbf{P}[\{|h(X_n) - h(X)| \geq \varepsilon\}] \to \varepsilon.\end{aligned}$$

Choice of epsilon was arbituary so to convergence holds. For the other part we can apply dominated convergence w.r.t. $|h(X_n)| \leq M$,

$$\begin{aligned}\lim_{n\to\infty} \mathbf{E}[|h(X_n)|] &= \lim_{n\to\infty} \int_\Omega |h(X_n)| d\mathbf{P} \\ & \int_\Omega |h(X)| d\mathbf{P} \\ &= \mathbf{E}[|h(X)|].\end{aligned}$$

Boundedness critirion is atleast must for the use of dominated convergence in general. Define a sequence $X_n = \begin{cases}\mathbf{P}[X_n = n] = 1/n \\ \mathbf{P}[X_n = 0] = 1 - 1/n\end{cases}$ and take $h = \mathrm{id}_\mathbb{R}$. Then $X_n \xrightarrow{\mathbf{P}} 0$, but $\mathbf{E}[h(X_n)] = n\frac{1}{n} = 1$.

10.1 Weak and strong laws of large numbers

Let $X_1, X_2, \ldots$ be a sequence on real-valued random variables with $\mathbf{E}[X_j] = \mathfrak{m}$ for all $j \in \mathbb{N}$. Denote also any uniform upper bound constant for kth moment by $\mathbf{E}[X_j^k] \leq K_k$.

Theorem 15 (Weak law of large numbers with bounded second moments)**.** *If* $\mathbf{E}[X_j^2] \leq K_2 < +\infty$, *then*

$$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{\mathbf{P}} \mathfrak{m}, \ \text{as } n \to \infty$$

Theorem 16 (Strong law of large numbers with bounded second moments)**.** *If* $\mathbf{E}[X_j^4] \leq K_4 < +\infty$, *then*

$$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{a.s.} \mathfrak{m}, \ \text{as } n \to \infty.$$

We will prove these later but first.

Exercise: Let $X_3, X_4, \ldots$ be a sequence of independent random variables, with $\mathbf{P}[X_k = 0] = 1 - \frac{1}{k\log k}$ and $\mathbf{P}[X_k = +k] = \frac{1}{2k\log k} = \mathbf{P}[X_k = -k]$.

(a) Calculate $\mathbf{E}[X_k]$ and $\mathrm{Var}(X_k)$

(b) Show that $\sum_{j=3}^{\infty} \frac{1}{j \log(j)} = \infty$ ja $\frac{1}{n^2} \sum_{j=3}^{n} \frac{j}{\log j} \to 0$, as $n \to \infty$.

(c) For $n \geq 3$ define the average $A_n = \frac{1}{n-2} \sum_{k=3}^{n} X_k$. Does the sequence of averages above converge almost surely?

(d) Does the sequence of averages converge in probability?

Sol:

(a) For the expectation

$$\begin{aligned}\mathbf{E}[X_k] &= \int_{\mathbb{R}} x dP_{X_k}(x) \\ &= 0 \cdot \mathbf{P}[X_k = 0] + \sum_{j=3}^{\infty} -kP_{X_k}(-k) + kP_{X_k}(k) \\ &= 0.\end{aligned}$$

We get by definition that

$$\begin{aligned}\mathrm{Var}(X_k) &= \mathbf{E}[X_k^2] \\ &\int_{\mathbb{R}} x^2 dP_{X_k}(x) \\ &= 0^2 \cdot \mathbf{P}[X_k = 0] + \sum_{j=3}^{\infty} (-k)^2 P_{X_k}(-k) + k^2 P_{X_k}(k) \\ &= \frac{1}{k \log k}.\end{aligned}$$

We have a lower bound since the terms of the sum can be taught lower sums of a decreasing function.

(b)

$$\sum_{j=3}^{\infty} \frac{1}{j \log(j)} \geq \int_{3}^{\infty} \frac{1}{x \log x} dx = \lim_{n \to \infty} \log \log n - \log \log 3 = \infty.$$

Now for the other part, we have quotient of two sequences a_n / b_n where $b_n = 1/n^2$ and a_n the partial sums of the above sum. By Stolz–Cesàro theorem

$$\lim_{n \to \infty} \frac{a_n}{b_n} = \lim_{n \to \infty} \frac{a_n - a_{n-1}}{b_n - b_{n-1}}.$$

This quotient is

$$\frac{a_n - a_{n-1}}{b_n - b_{n-1}} = \frac{\left(\frac{1}{n \log n}\right)}{n^2 - (n^2 - 1)^2} = \frac{2n + 1}{n \log n} \to 0, \text{ as } n \to \infty.$$

(c) For A_n to converge we need that $X_n/n \to 0$ since

$$X_n/(n-2) = A_n - A_{n-1}$$

and after taking limits from both sides we get that $\lim_{n\to\infty} X_n(n-1) = L - L = 0$. Thus we look at the event $|X_n| \geq n, n \in \mathbb{N}$, which happens only at vlaues $\pm n$ so

$$\mathbf{P}[|X_n| \geq m] = 2\frac{1}{2k\log k}.$$

Then by previous exercise $\sum_{n=3}^{\infty} \mathbf{P}[|X_n| \geq n] = +\infty$, where each event is independent. Then $|X_n| \geq n$ infinitely often by Borel-Cantelli. This means that the ratio $|X_n|/n \geq 1$ infinitley often, so after any large natural number the sequence will still have 1, so it cannot converge to zero in almost surely.

(d) One easily sees by linearity that $\mathbf{E}[A_n] = 0$ for all n. Now thus $\mathrm{Var}(A_n) = \mathbf{E}[A_n^2]$, but for independent variables

$$\mathrm{Var}(A_n) = \sum_{k=3}^{n} \mathrm{Var}\left(\frac{X_k}{n-2}\right) = \sum_{k=3}^{n} \frac{1}{(n-2)^2}\mathbf{E}(X_k^2) = \sum_{k=3}^{n} \frac{1}{(n-2)^2 k \log k}.$$

We will apply Chebyshev's inequality proven below this, so

$$\mathbf{P}[|A_n| \geq \varepsilon] \leq \frac{\mathrm{Var}(A_n)}{\varepsilon^2} \to 0, \quad \text{as } n \to \infty.$$

meaning $A_n \xrightarrow{\mathbf{P}} 0$.

10.2 Proof of the weak law

Following lemmas will be extremely useful.

Lemma 20 (Markov's inequality)**.** *If $X\colon \Omega \to \mathbb{R}$ is a random variables, then for any $a > 0$ we hvae*

$$\mathbf{P}[|X| \geq a] \leq \frac{1}{a}\mathbf{E}[|X|].$$

Proof. Define the event $E = \{\omega \in \Omega | |X(\omega)| \geq a\} = X^{-1}((-a,a)^c)$. For all $\omega \in \Omega$, we have that $|X(\omega)| \geq a\mathbb{I}_E(\omega)$, so by monotonicity of expected value $\mathbf{E}[|X|] \geq \mathbf{E}[a\mathbb{I}_E] = a\mathbf{P}[E]$. Dividing by a grants the inequality. □

Lemma 21 (Chebyshev's inequality)**.** *Suppose that $X \in \mathcal{L}^2(\mathbf{P})$, denote $\mathfrak{m} = \mathbf{E}[X]$ and $\mathfrak{s}^2 := \mathrm{Var}(X) = \mathbf{E}[(X-\mathfrak{m})^2]$. The for any $c > 0$, we hvae*

$$\mathbf{P}[|X - \mathfrak{m}| \geq c] \leq \frac{\mathfrak{s}^2}{c^2}.$$

Proof. Let $Y = (X-\mathfrak{m})^2$, so $\mathbf{E}[|Y|] = \mathfrak{s}^2$. Apply Markov's inequality to the event $|Y| \geq c^2$ since $\{|X - \mathfrak{m}| \geq c\} = \{|Y| \geq c^2\}$ as events. □

Now we can start to prove the weak law of large numbers.

Let $S_n = \sum_{j=1}^n X_j$ and $Y_n = \frac{1}{n} S_n$. Goal is to show that $Y_n \xrightarrow{\mathbf{P}} \mathfrak{m}$ assuming $\mathbf{E}[X_j^2] \leq K_2$ for all j. By linearity we get $\mathbf{E}[S_n] = n\mathfrak{m}$ and $\mathbf{E}[Y_n] = \mathfrak{m}$. By the second moment bound assumptipon $\operatorname{Var}(X_j) = \mathbf{E}[X_j^2] - \mathbf{E}[X_j]^2 \leq \mathbf{E}[X_j^2] \leq K_2$, by independence $\operatorname{Var}(S_n) = \sum_{j=1}^n \operatorname{Var}(X_j) \leq nK_2$ and $\operatorname{Var}(Y_n) = \frac{1}{n^2}\operatorname{Var}(S_n) \leq \frac{K_2}{n}$. By Chebyshev

$$\mathbf{P}[|Y - \mathfrak{m}| \geq \varepsilon] \leq \frac{\operatorname{Var}(Y_n)}{\varepsilon^2} \leq \frac{K_2}{n\varepsilon^2} \to 0 \text{ as } n \to \infty.$$

□

The weak law of large numbers can be applied to prove the famous Weierstrass' approximation theorem. Meaning polynomials are dense in continuous functions with sup-norm. The proof will be as follows: Let $U_1, U_2, \dots$ be indepedent random vairbales with uniform distribution on $[0,1]$. For $p \in [0,1]$, consider the events $E_j^{(p)} := \{U_j \leq p\}$ for $j \in \mathbb{N}$.

(a) Define $S_n^{(p)} = \sum_{j=1}^n \mathbb{I}_{E_j^{(p)}}$, calculate the expected value of $\mathbf{E}[S_n^{(p)}]$, and show $\operatorname{Var}(S_n^{(p)}) \leq \frac{n}{4}$.

(b) Show that for any $\delta > 0$, we have $\mathbf{P}[|S_n^{(p)}/n - p| \geq \delta] \leq \frac{1}{4n\delta^2}$.

(c) Let $f \in C[0,1]$, show that $B_n(p) := \mathbf{E}[f(S_n^{(p)}/n)]$ is polynomial in p.

(d) Show that $|B_n(p) - f(p)| \leq \mathbf{E}[|f(S_n^{(p)}/n) - f(p)|]$

(e) Since f is uniformly continuous by the compact domain, choose the corresponding $\delta > 0$ for $\epsilon > 0$ such that $|p - q| < \delta \Rightarrow |f(p) - f(q)| < \varepsilon$. Also let K a uniform bound for f. Consider the event $A_n^{(p)} = \{|S_n^{(p)}/n - p| \geq \delta\}$. Show that we have

$$|f(S_n^{(p)}/n) - f(p)| \leq 2K\mathbb{I}_{A_n^{(p)}} + \varepsilon\mathbb{I}_{\Omega\setminus A_n^{(p)}}.$$

(f) Prove the theorem: For any $\varepsilon > 0$, there exists a polynomial B_n such that $|f(p) - B_n(p)| < 2\varepsilon$ for all $p \in [0,1]$.

The details are as follows:

(a) From linearity, and the c.d.f of uniform distirbution it easily follows that $\mathbf{E}[S_n^{(p)}] = np$. Now variance is additive on independent events so $\operatorname{Var}(S_n^{(p)}) = \sum_{j=1}^n \operatorname{Var}(\mathbb{I}_{E_j^{(p)}})$. $\operatorname{Var}(I_{E_j^{(p)}}) = \mathbf{E}[I^2_{E_j^{(p)}}] - \mathbf{E}[I_{E_j^{(p)}}]^2 = \mathbf{E}[I_{E_j^{(p)}}] - \mathbf{E}[I_{E_j^{(p)}}]^2 = p - p^2$, so summing over the total variance is $np(1-p)$. Then we calculate the maximum of this function of p, by analyzing it's critical points $n(1-2p) = 0 \iff p = 1/2$. Both of the end points are zero, and evaluated at $p = 1/2$ is $\frac{n}{4}$.

(b) Apply Chebyshev's to $p = \mathbf{E}[S_n^{(p)}/n]$ so

$$\mathbf{P}[|S_n^{(p)}/n - p| \geq \delta] = \frac{\operatorname{Var}(S_n^{(p)}/n)}{\delta^2} = \frac{\frac{1}{n^2}\operatorname{Var}(S_n^{(p)})}{\delta^2} \leq \frac{\left(\frac{1}{4n}\right)}{\delta^2} = \frac{1}{4n\delta^2}.$$

(c) We need to compute the law $P_{S_n^{(p)}}$, first we notice that $S_n^{(p)}$ is discrete with values in $\{0, 1, \dots, p\}$. Therefore

$$\mathbf{E}\left[f\left(\frac{S_n^{(p)}}{n}\right)\right] = \int_{\mathbb{R}} f(x/n)dP_{S_n^{(p)}}(x) = \sum_{i=0}^{p} f(i/n)\mathbf{P}[S_n^{(p)} = i] = (*).$$

We need to calculate $\mathbf{P}[S_n^{(p)} = i]$ which is the number possible weighted probability p coin flip sequences such that i heads happen which is given by the binomial formula $\mathbf{P}[S_n^{(p)} = i] = \binom{n}{i}p^i(1-p)^{n-i}$, so

$$(*) = \sum_{i=0}^{p} f(i/n)\binom{n}{i}p^i(1-p)^{n-i}.$$

This is a polynomial of p.

(d)

$$\begin{aligned}|B_n(p) - f(p)| &= |\mathbf{E}[f(S_n^{(p)}) - \mathbf{E}[f(p)]]\\ &= |\mathbf{E}[f(S_n^{(p)}) - f(p)]|\\ &\leq \mathbf{E}[|f(S_n^{(p)}) - f(p)|].\end{aligned}$$

(e) One notices that from the uniform continuity for $\omega \in \Omega \setminus A_n^{(p)}$, $|f(S_n^{(p)}(\omega)) - f(p)| < \varepsilon$ and otherwise we can just use the uniform bound K twice by triangle inequality.

(f) By Chebyshev $\mathbf{P}[A_n^{(p)}] \leq \frac{1}{4n\delta^2}$ so

$$\begin{aligned}|f(p) - B_n(p)| &\leq \mathbf{E}[|f(S_n^{(p)}) - f(p)|]\\ &\leq \mathbf{E}[2K\mathbb{I}_{A_n^{(p)}} + \varepsilon\mathbb{I}_{\Omega\setminus A_n^{(p)}}]\\ &= \frac{2K}{4n\delta^2} + \varepsilon(1 - \frac{1}{4n\delta^2})\\ &= \frac{2K-\varepsilon}{4n\delta^2} + \varepsilon\end{aligned}$$

where the n cna be pumped as large as we want, since none of the constants here depend on n.

10.3 Proof of the strong law of large numbers

First prove the statement for $\mathfrak{m} = 0$ and after that reduce the general case to this particular case by centering appropriately.

Case $\mathfrak{m} = 0$: Assume $\mathfrak{m} = 0$ and denote

$$S_n = \sum_{j=1}^{n} X_j.$$

The assertion of the strong law of large numbers concerns the event

$$E := \left\{ \omega \in \Omega | \lim_{n\to\infty} \frac{S_n(\omega)}{n} \to 0 \right\}$$

occuring almost surely. However we will consider another event

$$E' := \left\{ \omega \in \Omega | \sum_{n=1}^{\infty} \left(\frac{S_n(\omega)}{n} \right)^4 < +\infty \right\},$$

and show that this is almost sure, which will in turn imply E almost surely. Note that

$$\sum_{n=1}^{\infty} \left(\frac{S_n(\omega)}{n} \right)^4 < \infty, \forall \omega \in \mathbf{E}'$$

implies that

$$\left(\frac{S_n(\omega)}{n} \right)^4 \to 0, n \to \infty, \forall \omega \in \Omega.$$

By continuity of $t \mapsto t^{1/4}$ assets

$$\frac{S_n(\omega)}{n} \to 0, n \to \infty, \forall \omega \in \Omega$$

so $E' \subset E$.

To prove E' occuts almost surely, the main body of work consists of showing

$$\sum_{n=1}^{\infty} \mathbf{E}\left[\left(\frac{S_n}{n} \right)^4 \right] < +\infty.$$

Once shown we can use the convergence results for non-negative random variables in expected values imply $\sum_{n=1}^{\infty} (\frac{S_n}{n})^4 < +\infty$ almost surely, meaning $1 = \mathbf{P}[E'] \leq \mathbf{P}[E]$.

Note first of all that since $X_j \in \mathcal{L}^4(\mathrm{P})$ for all j by assumption, Lemma VIII.10 shows that also $S_n = X_1 + \cdots + X_n \in \mathcal{L}^4(\mathrm{P})$. We will compute the fourth moment of S_n by expanding the multinomial

$$\begin{aligned} S_n^4 &= (X_1 + \cdots + X_n)^4 \\ &= \sum_{1 \leq i \leq n} X_i^4 + \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} \frac{4!}{3!\,1!} X_i^3 X_j + \frac{1}{2!} \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} \frac{4!}{2!^2} X_i^2 X_j^2 \\ &\quad + \frac{1}{2!} \sum_{\substack{1 \leq i,j,k \leq n \\ i,j,k \text{ different}}} \frac{4!}{2!\,1!^2} X_i^2 X_j X_k + \frac{1}{4!} \sum_{\substack{1 \leq i,j,k,\ell \leq n \\ i,j,k,\ell \text{ different}}} \frac{4!}{1!^4} X_i X_j X_k X_\ell. \end{aligned}$$

Because $\mathbf{E}[X_j] = 0$, and independence makes expectation multiplicative, the second, forth and fifth terms above vanish. Meaning

$$\begin{aligned}
\mathbf{E}[S_n^4] &= \sum_{i\leq i\leq n} \mathbf{E}[X_i^4] + \frac{1}{2!} \sum_{1\leq i,j\leq n, i\neq j} \frac{4!}{2!^2}\mathbf{E}[X_i^2 X_j^2] \\
&\overset{C-S}{\leq} \sum_{i\leq i\leq n} K_4 + \frac{1}{2!} \sum_{1\leq i,j\leq n, i\neq j} \frac{4!}{2!^2} K_4 \\
&= nK_4 + n(n-1)\frac{1}{2!}\frac{4!}{2!^2}K_4 \\
&\leq 3n^2 K_4.
\end{aligned}$$

This gives that

$$\sum_{n=1}^{\infty} \mathbf{E}\left[\left(\frac{S_n}{n}\right)^4\right] \leq \sum_{n=1}^{\infty} \frac{1}{n^4}\mathbf{E}[S_n^4] \leq \sum_{n=1}^{\infty} \frac{3n^2K_4}{n^4} \leq 3K_4 \sum_{n=1}^{\infty} \frac{1}{n^2} < +\infty.$$

Case $\mathfrak{m} \neq 0$: If $\mathbf{E}[X_j] = \mathfrak{m}$, then $\tilde{X}_j := X_j - \mathfrak{m}$ has expected value 0. By the first case $\frac{1}{n}\sum_{j=1}^n \tilde{X}_j \xrightarrow{a.s.} 0$ which is equivalent to $\frac{1}{n}\sum_{j=1}^n X_j \xrightarrow{a.s.} \mathfrak{m}$. We just need to check that the forth moments assumption and independence holds for $\tilde{X}_j$'s. Independence follows from Borel-measurability of $x \mapsto x + \mathfrak{m}$. For forth moments

$$\mathbf{E}[|\tilde{X}_j|^4] = \mathbf{E}[|X_j - \mathfrak{m}|^4] \leq 2^4(\mathbf{E}[|X_j^4| + \mathbf{E}[\mathfrak{m}^4]]) \leq 16(K_4 + \mathfrak{m}^4).$$

10.4 Kolmogorov's strong law of large numbers

The two versions of law of large numbers required the notion of bounded second or forth moments. But the existance of expected value only needs the condition $X_1, X_2, \dots \in \mathcal{L}^1(\mathbf{P})$.

Definition 22. Sequence $X_1, X_2, \dots \in \mathcal{L}^1(\mathbf{P})$ converges **in** $\mathcal{L}^1$ if

$$\lim_{n\to\infty} \mathbf{E}[|X_n - X|] = 0$$

denote $X_n \xrightarrow{\mathcal{L}^1} X$.

This actually is a stronger notion than convergence in probability. By Markov's inequality assume that there exists some $0 < M = \lim_{n\to\infty} \mathbf{E}[|X_n - X|]$, then

$$\mathbf{P}[|X_n - X| \geq \varepsilon] \leq \frac{1}{\varepsilon}\mathbf{E}[|X_n - X|] \to 0, n \to \infty.$$

Theorem 17 (Kolmogorov's strong law of large numbers)**.** *Let $X_1, X_2, \dots \in \mathcal{L}^1(\mathbf{P})$ be independent and indentically distributed random variables with $\mathbf{E}[X_j] = \mathfrak{m}$ for all $j \in \mathbb{N}$. Then*

$$\frac{1}{n}(X_1 + \dots + X_n) \xrightarrow{a.s.} \mathfrak{m}$$

and

$$\frac{1}{n}(X_1 + \dots + X_n) \xrightarrow{\mathcal{L}^1} \mathfrak{m}$$

as $n \to \infty$.

Note, however, that while the above result is thus relaxing the moment assumptions, it is not strictly speaking a generalization of the strong law of Theorem XI.5, because it assumes that the sequence consists of identically distributed random variables. Both formulations are useful.

11 Convergence in distirbution and the central limit theorem

Let $X_1, X_2, \ldots$ be independent identivally distributed random numbers, and form the sums

$$S_n = X_1 + \cdots + X_n.$$

We are interested in the behavrior of S_n for large n. If $X_j \in \mathcal{L}^1(\mathbf{P}), \mathbf{E}[X_j] = \mathfrak{m}$ for, then we have $\mathbf{E}[S_n] = n\mathfrak{m}$. The law of (Kolmogorov's strong) large numbers then says that the sum concentrates around the value $n\mathfrak{m}$, more precicely

$$\frac{S_n - n\mathfrak{m}}{n} \to 0$$

(with respect to the notion of convergence a.s., in probability, $\mathcal{L}^1$). If the random variables are square integrable $X_j \in \mathcal{L}^2(\mathbf{P})$, with expectation $\mathbf{E}[X_j] = \mathfrak{m}$ and variance $\mathrm{Var}(X_j) = \mathfrak{s}^2$, then by the independence of the the terms, $\mathrm{Var}(S_n) = n\mathfrak{s}^2$. Chebyshev's inequality then says that the fluctuations of the sum S_n around the value $n\mathfrak{m}$ do note exceed a scale propotional to $\sqrt{n}$, more precicely

$$\mathbf{P}\left[\frac{|S_n - n\mathfrak{m}|}{\sqrt{n}} \geq c\right] \leq \frac{n\mathfrak{s}^2}{(c\sqrt{n})^2} = \frac{\mathfrak{s}^2}{c^2}, \text{ for any } c > 0.$$

To understand the behavior of sum S_n for large n, in detail, it is therefore meaningful to look at $S_n - n\mathfrak{n}$ on a scale proportional to $\sqrt{n}$. One can show that

$$\frac{S_n - n\mathfrak{m}}{\sqrt{n}}$$

does not converge almost surely or in probability. But however we can look at it's behavior in **distribution**, known as **convergence in distribution**. The central limit theorem asserts that the above quotient approaches the Gaussian distribution, which has density function

$$\frac{1}{\sqrt{2\pi\mathfrak{s}^2}} \exp\left(-\frac{1}{2\mathfrak{s}^2}(x - \mathfrak{m})^2\right).$$

An elementary interpretation of this is that the cumulative distribution functions have the following limit

$$\mathbf{P}\left[\frac{S_n - n\mathfrak{m}}{\mathfrak{s}\sqrt{n}} \leq x\right] \xrightarrow{n\to\infty} \Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

The right hand side above is the cumulative distribution of the standard normal distribution (Gaussian distribution with mean zero and unit variance).

11.1 Characteristic functions

An important tool in the proof of the central limit theorem and in various other places in probability theory is characteristic functions. A characteristic functions of a real valued random variable is a sinhe function which encodes it's distribution. It is essentially the Fourier transform of the law. More generally, characteristic functions of vector valued random variables could be defined, and have properties parallel with what we show in the setup of real valued random variables.

The characteristic functions are complex valued, so we begin with a few remarks about complex valued random variables. The plane $\mathbb{C}$ has Borel sets of $\mathbb{R}^2$, a compelx valued random variable has the form $Z = X+iY$ where X, Y are real-valued random variables viewed as random vectors $Z = (X, Y)$. We have that Z is integrable if $X, Y \in \mathcal{L}^1(\mathbf{P})$, $\mathbf{E}[Z] := \mathbf{E}[X] + i\mathbf{E}[Y]$. Note that if Z is integrable $\mathbf{E}[|Z|] \leq \mathbf{E}[|X| + |Y|] \leq \mathbf{E}[|X|] + \mathbf{E}[|Y|] < +\infty$.

Proposition 14.
- *The complex expected value is $\mathbb{C}-$linear.*
- *Triangle inequality $\mathbf{E}[|Z|] \leq |\mathbf{E}[Z]|$ holds.*
- *Dominate convergence holds: Suppose $Z_1, Z_2, \dots$ is a sequence of $\mathbb{C}-$valued rancom variables and $X \in \mathcal{L}^1(\mathbf{P})$ integrable which dominates $|Z_n|$ for all $n \in \mathbb{N}$. Then if the pointwise limit exists, $\mathbf{E}[\lim_{n\to\infty} Z_n] = \lim_{n\to\infty} \mathbf{E}[Z_n]$.*
- *For independent, integrable Z_1, Z_2, $\mathbf{E}[Z_1 Z_2] = \mathbf{E}[Z_1]\mathbf{E}[Z_2]$.*

Definition 23 (Characteristic function)**.** Let X be a real-valued random variable, then the **characterisic function** $\varphi_X : \mathbb{R} \to \mathbb{C}$ is defined to be $\varphi_X(\theta) = \mathbf{E}[e^{i\theta X}] = \mathbf{E}[\sin(\theta X)] + i\mathbf{E}[\cos(\theta X)]$.

Note that $x \mapsto e^{i\theta x}$ is Borel, so $\mathbf{E}[e^{i\theta X}] = \int_{\mathbb{R}} e^{i\theta x} dP_X(x)$. Therefore the characteristic function only depends on the distribution of X. Soon we will show that infact, φ_X completely determines P_X.

Examples: Let $X \sim \text{Exp}(\lambda)$, so it has density function $f_X(x) = \lambda e^{-\lambda x}\mathbb{I}_{[0,+\infty)}(x)$, so

$$\begin{aligned}\varphi_X(\theta) &= \mathbf{E}[e^{i\theta X}]\\ &= \int_{\mathbb{R}} e^{i\theta x} f_X(x) dx\\ \lambda &\int_0^\infty e^{x(-\lambda - i\theta)} dx\\ &= \frac{1}{1 + i\theta/\lambda}.\end{aligned}$$

If $X \sim \text{Poisson}(\lambda)$ has $P_X(n) = \mathbf{P}[X = n] = e^{-\lambda}\frac{\lambda^n}{n!}, n \in \mathbb{Z}_{\geq 0}$. Then

$$\begin{aligned}\varphi_X(\theta) &= \mathbf{E}[e^{i\theta X}]\\ &= \sum_{n=0}^{\infty} p_X(n)e^{i\theta n}\\ &= e^{-\lambda}\sum_{n=0}^{\infty}\frac{1}{n!}(\lambda e^{i\theta})^n = \exp\bigl(\lambda(e^{i\theta}-1)\bigr).\end{aligned}$$

Exercise:
If $X \sim N(0,1)$, then $f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}, x \in \mathbb{R}$.

(a) Let $t \in \mathbb{R}$. Showw $\mathbf{E}[e^{tX}] = e^{t^2/2}$.

(b) For $x, t \in \mathbb{R}$, show that $e^{|tx|} \le e^{tx} + e^{-tx}$. Using this, prove that for any $t \in \mathbb{R}$ we have

$$\mathrm{E}\left[\sum_{n=0}^{\infty}\frac{1}{n!}|tX|^n\right] < +\infty.$$

(c) Prove that for any $t \in \mathbb{R}$ we have

$$\mathrm{E}[e^{tX}] = \sum_{n=0}^{\infty}\frac{1}{n!}t^n\mathrm{E}[X^n].$$

(d) By comparing (a) with (c), deduce that for $n \in \mathbb{N}$ we have

$$\mathrm{E}[X^n] = \begin{cases}\prod_{j=1}^{n/2}(2j-1) & \text{if } n \text{ is even}\\ 0 & \text{if } n \text{ is odd.}\end{cases}$$

(e) Prove that

$$\varphi_X(\theta) = e^{-\frac{1}{2}\theta^2} \qquad \text{for } \theta \in \mathbb{R}.$$

Sol:

(a) Let $x' = x - t$, then

$$\begin{aligned}\mathbf{E}[e^{tX}] &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{tx-\frac{x^2}{2}}dx\\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{t(x'+t)-\frac{(x'+t)^2}{2}}dx' = (*)\end{aligned}$$

Now simplying

$$t(x'+t) - \frac{(x'+t)^2}{2} = tx' + t^2 - \frac{x'^2}{2} - xt' - t^2 = -\frac{x'^2}{2}$$

yields

$$(*) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{x'^2}{2}}dx' = 1.$$

(b) If $tx \geq 0$ then $e^{|tx|} = e^{tx} \leq e^{tx} + e^{-tx}$. If $tx < 0$, then $e^{|tx|} = e^{-tx} \leq e^{tx} + e^{-tx}$. Now for the second part

$$\begin{aligned}\mathbf{E}\left[\sum_{n=0}^{\infty}\frac{1}{n!}|tX|^n\right] &= \mathbf{E}\left[\exp\left(|tX|\right)\right] \\ &\leq \mathbf{E}[\exp\left(tX\right)] + \mathbf{E}[\exp\left(-tX\right)] \\ &= 2e^{t^2/2} < +\infty.\end{aligned}$$

(c)

$$\begin{aligned}\mathbf{E}[e^{tX}] &= \mathbf{E}\left[\sum_{n=0}^{\infty}\frac{1}{n!}t^nX^n\right] \\ &= \sum_{n=0}^{\infty}\frac{1}{n!}t^n\mathbf{E}\left[X^n\right] \text{ (by DCT applied to previous part's function)}\end{aligned}$$

(d)

$$\sum_{n=0}^{\infty}\frac{1}{n!}t^n\mathbf{E}\left[X^n\right] = e^{t^2/2} = \sum_{n=0}^{\infty}\frac{1}{n!}(t^2/2)^n.$$

RHS has only powers of t that are even, meaning for odd n by the uniqueness of power series (that has radius of convergence greater than 0), $\mathbf{E}[X^n] = 0$ thinking in terms of t being the variable. If n is even, lets say $n = 2k$, then the LHS term is for t^{2k} is $\frac{t^{2k}}{(2k)!}\mathbf{E}[X^{2k}]$ and on RHS $\frac{t^{2k}}{k!2^k}$. Meaning

$$\mathbf{E}[X^{2k}] = \frac{t^{2k}(2k)!}{k!2^kt^{2k}} = \frac{(2k)!}{k!2^k} = \frac{2^k(k!)\cdot\text{product of odd nums}}{2^kk!} = \text{product of odd nums}.$$

Now product of odd nums $= (2k-1)(2k-3)\cdots 5\cdot 3\cdot 1$ which by substitutin $n = 2k$, becomes $(n-1)$

(e) Now

$$\begin{aligned}\varphi_X(\theta) &= \mathbf{E}[e^{i\theta X}] \\ &= \sum_{n=0}^{\infty}\frac{1}{n!}(\theta i)^n\mathbf{E}\left[X^n\right] \text{ (like in (c))} \\ &= \sum_{n=0}^{\infty}\frac{1}{(2n)!}(\theta i)^{2n}\mathbf{E}\left[X^{2n}\right] \\ &= \sum_{n=0}^{\infty}\frac{1}{2^nn!\prod_{n=1}^{n}(2j-1)}(-\theta^2)^n\prod_{n=1}^{n}(2j-1) \\ &= \sum_{n=0}^{\infty}\frac{1}{2^nn!}(-\theta^2)^n \\ &= \sum_{n=0}^{\infty}\frac{(-\theta^2/2)^n}{n!} \\ &= e^{\theta^2/2}.\end{aligned}$$

Proposition 15 (Basic properties of characteristic functions)**.** *Characteristic functions have the following properties:*

(a) We have $\varphi_X(0) = 1$.

(b) We have $|\varphi_X(\theta)| \leq 1$ for all $\theta \in \mathbb{R}$.

(c) The function $\varphi_X : \mathbb{R} \to \mathbb{C}$ is continuous.

(d) For any $a, b \in \mathbb{R}$ we have $\varphi_{aX+b}(\theta) = e^{ib\theta}\varphi_X(a\theta)$ for all $\theta \in \mathbb{R}$.

(e) We have $\varphi_{-X}(\theta) = \overline{\varphi_X(\theta)}$ for all $\theta \in \mathbb{R}$.

(f) We have $\varphi_X(-\theta) = \varphi_X(\theta)$ for all $\theta \in \mathbb{R}$.

Proof. (a) $\varphi_X(0) = \mathbf{E}[1] = 1$.

(b) Triangle inequality $\varphi_X(\theta) \leq \mathbf{E}[1] = 1$.

(c) Let $\theta_n \to \theta$ be any sequence of reals. By continuity $e^{i\theta_n X(\omega)} \to e^{i\theta X(\omega)}$. But it is not only continuous, but bounded. By bounded convergence the limit thus exists inside the integral, to

$$\varphi_X(\theta_n) = \mathbf{E}[e^{i\theta_n X(\omega)}] \to \mathbf{E}[e^{i\theta X(\omega)}] = \varphi_X(\theta).$$

(d) Follows from linearity of expectation.

(e) Complex conjugation.

(f) Complex conjugation.

□

We can apply this to get the distribution of $X \sim N(\mathfrak{m}, \mathfrak{s}^2)$ where $\tilde{X} := (X - \mathfrak{m})/\mathfrak{s} \sim N(0,1)$. Thus

$$\varphi_X(\theta/s) = e^{-i\frac{-\mathfrak{m}}{\mathfrak{s}}\theta}\varphi_{\tilde{X}}(\theta) = e^{i\frac{\mathfrak{m}}{\mathfrak{s}}\theta-\theta^2/2}.$$

This means that

$$\varphi_X(\theta) = e^{i\mathfrak{m}\theta-(\mathfrak{s}^2\theta^2)/2}.$$

Another fundamental property of characteristic functions is that the characteristic function of a sum of independent terms is the pointwise product of the characteristic functions.

In fact $\varphi_{X+Y}(\theta) = \varphi_X(\theta)\varphi_Y(\theta)$.

This follows from

$$\varphi_{X+Y}(\theta) = \mathbf{E}[e^{i\theta X}e^{i\theta Y}] = \mathbf{E}[e^{i\theta X}]\mathbf{E}[e^{i\theta Y}]$$

if $e^{i\theta X} \perp\!\!\!\perp e^{i\theta Y}$. But the independence follows from the function $x \mapsto e^{i\theta x}$ being Borel.

A fundamental property of the characteristic function of a random variable is that it contains all the information about the distribution of the random variable. This fact is made explicit by Lévy's inversion theorem, below.

Theorem 18 (Lévy's inversion theorem). *Let $X \in m\mathcal{F}$ be a real-valued random variable, P_X its distribution (a Borel probability measure on $\mathbb{R}$), and $\varphi_X : \mathbb{R} \to \mathbb{C}$ its characteristic function. Then for any $a, b \in \mathbb{R}$, $a < b$, we have*

$$\begin{aligned}&\lim_{T\to+\infty} \frac{1}{2\pi} \int_{-T}^{+T} \frac{e^{-\mathrm{i}\theta a} - e^{-\mathrm{i}\theta b}}{\mathrm{i}\theta} \varphi_X(\theta)\,\mathrm{d}\theta \\ &= P_X[(a,b)] + \frac{1}{2}P_X[\{a\}] + \frac{1}{2}P_X[\{b\}].\end{aligned}$$

In particular, φ_X uniquely determines P_X.
Moreover, if $\int_{\mathbb{R}} |\varphi_X(\theta)|\,\mathrm{d}\theta < +\infty$, then X has a continuous probability density function f_X given by

$$f_X(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-\mathrm{i}\theta x} \varphi_X(\theta)\,\mathrm{d}\theta.$$

We will use short hand notation

$$\nu[B] = P_X[B] = \mathbf{P}[X \in B],$$

$$\varphi(\theta) = \varphi_X(\theta) = \mathbf{E}[e^{i\theta x}] = \int_{\mathbb{R}} e^{i\theta x} d\nu(x),$$

$$F(x) = F_X(x) = \nu[(-\infty, x]].$$

The proof will be split in some lemmas. Then this turns into

$$\lim_{T\to+\infty} \frac{1}{2\pi} \int_{-T}^{+T} \frac{e^{-\mathrm{i}\theta a} - e^{-\mathrm{i}\theta b}}{\mathrm{i}\theta} \varphi(\theta)\,\mathrm{d}\theta = \nu[(a,b)] + \frac{1}{2}\nu[\{a\}] + \frac{1}{2}\nu[\{b\}]$$

and

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-\mathrm{i}\theta x} \varphi(\theta)\,\mathrm{d}\theta.$$

The first formula is occasionally written in terms of the c.f.t as well. Revall that c.f.t F is increasing and right-continuoud: If $x_n \downarrow x$, then $F(x_n) \downarrow F(x)$. Since it is increasing bounded, left limits also exists: They are defined by

$$F(x^-) = \lim_{x' \uparrow x} F(x').$$

By monotone convergence of probability measures, the left lmits can be expressed in terms of the distribution ν as following measures of open semi-infinite intervals

$$\begin{aligned}F(x^-) &= \lim_{n\to\infty} F(x - 1/n) \\ &= \lim_{n\to\infty} \nu[(-\infty, x - 1/n]] \\ &= \nu\left[\bigcup_{n=1}^{\infty} (-\infty, x - 1/n]\right] \\ &= \nu[(-\infty, x)].\end{aligned}$$

In particular, if there is a jump of discontinuity at a point x, the size of the jump $F(x) - F(x^-)$ is the probability mass located at the single point x,

$$F(x) - F(x') = \nu[(-\infty, x]] - \nu[(-\infty, x)] = \nu[\{x\}]$$

(at a continuity point there is no probability mass).

For any $a < b$, the increment from a to b of the c.f.t is

$$F(b) - F(a) = \nu[(-\infty, b]] - \nu[(-\infty, a]] = \nu[(a, b]].$$

If we replace $F(x)$ by the average $\frac{1}{2}(F(x) - F(x'))$ of the left and right limits, then the corresponding increment becomes

$$\frac{1}{2}(F(b) - F(b^-)) - \frac{1}{2}(F(a) + F(a^-)) = \nu[(a, b)] + \frac{1}{2}\nu[\{a\}] + \frac{1}{2}\nu[\{b\}].$$

Lemma 22 (An auxiliary integral)**.** *For $r \in \mathbb{R}$ define*

$$S(r) := \int_0^r \frac{\sin\theta}{\theta} d\theta.$$

Then the limits $r \uparrow +\infty$ and $r \downarrow -\infty$ of this integral are

$$\lim_{r\uparrow+\infty} S(r) = \frac{\pi}{2} \quad \& \quad \lim_{r\downarrow-\infty} S(r) = -\frac{\pi}{2}.$$

Moreover for any $c \in \mathbb{R}$, we have

$$\int_0^r \frac{\sin c\theta}{\theta} d\theta = S(cr).$$

Proof. The last part is direct one step change of variables. In particular $S(-r) = S(r)$. Now we have that

$$\int_{-\infty}^{\infty} \frac{\sin\theta}{\theta} d\theta = \operatorname{Im}\left(\int_{-\infty}^{\infty} \frac{e^{i\theta}}{\theta} d\theta\right) = \operatorname{Im}(\pi i \operatorname{Res}|_0) = \operatorname{Im}(\pi i).$$

Result then follows from symmetry. □

The contour here is chosen that it is the Cauchy PV-integral (which equls to the improper integral iff the improper integral exists. Note that the sinc integral is not Lebesgue integrable for example over positive reals). *Proof of the theorem:* For any $v, w \in \mathbb{R}$, we have $|e^{iv} - e^{iw}| \leq |\int_v^w ie^{iz}dz| \leq |v - w|$. Hence

$$\int_{\mathbb{R}} \left(\int_{-T}^{+T} \frac{e^{i\theta(x-a)} - e^{i\theta(x-b)}}{i\theta}\right) d\nu(x) \leq 2T|b - a| < +\infty.$$

This will allow use to use Fubini

$$\begin{aligned}\int_{-T}^{+T}\frac{e^{-i\theta a}-e^{-i\theta b}}{i\theta}\varphi(\theta)d\theta &= \int_{-T}^{+T}\frac{e^{-i\theta a}-e^{-i\theta b}}{i\theta}\left(\int_{\mathbb{R}} e^{i\theta x}d\nu(x)\right)d\theta \\ &= \int_{\mathbb{R}}\int_{-T}^{+T}\frac{e^{i\theta(x-a)}-e^{i\theta(x-b)}}{i\theta}d\theta d\nu(x)\end{aligned}$$

Then we just evaluate the real part

$$\begin{aligned}\int_{-T}^{+T}\frac{e^{i\theta(x-a)}-e^{i\theta(x-b)}}{i\theta}d\theta &= \int_{-T}^{T}\frac{\sin(\theta(x-a))-\sin(\theta(x-b))}{\theta}d\theta \\ &= \int_{-T}^{T}\frac{\sin(\theta(x-a))}{\theta}d\theta - \int_{-T}^{T}\frac{\sin(\theta(x-b))}{\theta}d\theta \\ &= 2\int_{0}^{T}\frac{\sin(\theta(x-a))}{\theta}d\theta - 2\int_{0}^{T}\frac{\sin(\theta(x-b))}{\theta}d\theta \\ &= 2S((x-a)T)-2S((x-b)T)\end{aligned}$$

As T grows the limit is

$$\begin{cases}2\pi, & a<x<b\\ \pi, & x=a \text{ or } x=b\\ 0, & x<a \text{ or } x>b.\end{cases}$$

By bdd convergence theorem we can put limit inside integral

$$\lim_{T\to\infty}\int_{\mathbb{R}}\left(\int_{-T}^{+T}\frac{e^{i\theta(x-a)}-e^{i\theta(x-b)}}{i\theta}\right)d\nu(x)\to 2\pi\nu[(a,b)]+\pi\frac{1}{2}\nu[\{a\}]+\pi\frac{1}{2}\nu[\{b\}].$$

Proof of the density part: Suppose now that $\int_{\mathbb{R}}|\varphi(x)|dx<+\infty$. Note that the function

$$f_X(x)=\frac{1}{2\pi}\int_{\mathbb{R}}e^{-i\theta x}\varphi(\theta)d\theta$$

is continuous. Let $x_n\to x$, then the funtion inside is continuous. Moreover the function inside is dominated by $|\varphi(\theta)|$ so we can pass any limit inside by dominated convergence, hence $f_X(x_n)\to f_X(x)$. It remains now to show that this is the probability density for the measure ν. Next note that

$$\left|\frac{e^{i\theta(x-a)}-e^{i\theta(x-b)}}{i\theta}\varphi(x)\right|\le|a-b||\varphi(x)|.$$

We can apply dominated convergence on constant multiple of $|\varphi(x)|$ we can write the LHS of the first part as

$$\frac{1}{2\pi}\int_{\mathbb{R}}\frac{e^{-i\theta a}-e^{-ib}}{i\theta}d\theta.$$

Also by dominated convergence with the same domanation argument for $a_n \to a, b_n \to b$

$$\frac{1}{2\pi}\int_{\mathbb{R}} \frac{e^{-i\theta a_n} - e^{-ib_n}}{i\theta} d\theta \to \frac{1}{2\pi}\int_{\mathbb{R}} \frac{e^{-i\theta a} - e^{-ib}}{i\theta} d\theta.$$

Using the result of the firts part and rewriting $2\pi\nu[(a,b)] + \pi\frac{1}{2}\nu[\{a\}] + \pi\frac{1}{2}\nu[\{b\}]$ by the left-right limit averages, we get

$$\frac{F(b_n) + F(b_n^-)}{2} - \frac{F(a_n) + F(a_n^-)}{2} \to \frac{F(b) + F(b^-)}{2} - \frac{F(a) + F(a^-)}{2}.$$

This shows that that the c.f.t is continuous. Again

$$\left|\frac{e^{-i\theta a} - e^{-i\theta b}}{i\theta(a-b)}\varphi(x)\right| \leq |\varphi(x)|,$$

we may calculate the derivative of F by dominated convergence as follows:

$$F'(a) = \lim_{b\to a}\frac{F(b) - F(a)}{b-a} = \lim_{b\to a}\frac{1}{2\pi}\int_{\mathbb{R}} \frac{e^{-i\theta a} - e^{-i\theta b}}{i\theta(b-a)}\varphi(\theta)d\theta = \frac{1}{2\pi}\int_{\mathbb{R}} e^{i\theta a}\varphi(\theta)d\theta.$$

Thus the derivative is precicely $F'(x) = f_X(x)$, meaning

$$\nu[(a,b)] = F(b) - F(a) = \int_a^b F'(x)dx = \int_a^b f_X(x)dx.$$

Since open intervals form a pi-system which uniquely determine ν, we are done. □

This allows us to prove a lemma which allows us to control the amount of probability mass outside a large interval in terms of the bahavior of the characteristic function $\varphi(\theta)$ near $\theta = 0$.

Lemma 23. *Let ν be a probability measure on $(\mathbb{R}, \mathcal{B})$, and let $\varphi(\theta) = \int_{\mathbb{R}} e^{i\theta x}\,d\nu(s)$ be its characteristic function. Then for any $r > 0$ we have*

$$\nu\left[\mathbb{R} \setminus [-r, +r]\right] \leq \frac{r}{2}\int_{-2/r}^{2/r} (1 - \varphi(\theta))\,d\theta.$$

Proof. The RHS integral on the inequlity evaluates to $2u(1 - \sin(ux)/ux)$ with $u = r/2$ for convinience. Then integrating both sides over $\mathbb{R}$ w.r.t ν, we get by Fubini

$$\int_{\mathbb{R}}\left(\int_{-u}^{u}(1 - e^{i\theta x}d\theta\right)d\nu(x) = \int_{-u}^{u}\left(\int_{\mathbb{R}}(1 - e^{i\theta x}d\nu(x))\right)d\theta = \int_{-u}^{u}(1 - \varphi(\theta))d\theta.$$

Dividing by u,

$$\frac{1}{u}\int_{-u}^{u}(1 - \varphi(\theta))d\theta = 2\int_{\mathbb{R}} 1 - sin(ux)/uxd\nu(x).$$

Since $\sin(\zeta)/\zeta \leq 1$, we see that the integrand on the RHS is non-negative, therefore omitting from the integral $[-r, r]$, we get

$$\frac{1}{u}\int_{-u}^{u}(1-\varphi(\theta))d\theta \geq 2\int_{\mathbb{R}\setminus[-r,r]} 1 - sin(ux)/uxd\nu(x).$$

For $|x| > r = 2/u$, we have $\left|\frac{\sin(ux)}{ux}\right| \leq \frac{1}{u|x|} \leq 1/2$ so

$$\frac{1}{u}\int_{-u}^{u}(1-\varphi(\theta))d\theta \geq 2\int_{\mathbb{R}\setminus[-r,r]} \frac{1}{2}\nu(x) = \nu[\mathbb{R}\setminus[-r,r]].$$

□

Example: If $X_1 \perp\!\!\!\perp X_2, X_1 \sim N(\mathfrak{m}_1, \mathfrak{s}_1^2)$ and $X_2 \sim N(\mathfrak{m}_2, \mathfrak{s}_2^2)$, then we show that $X_1 + X_2 \sim N(\mathfrak{m}_1 + \mathfrak{m}_2, \mathfrak{s}_1^2 + \mathfrak{s}_2^2)$.

Sol: By properties of characteristic functions $\varphi_{X_1+X_2}(\theta) = e^{i(\mathfrak{m}_1+\mathfrak{m}_2)\theta - \frac{1}{2}(\mathfrak{s}_1^2+\mathfrak{s}_2^2)\theta^2}$. Now one can see that taking the absolute value gets an integrable function, which is a Gaussian integral. Therefore by Lévy's inversion

$$f_{X_1+X_2}(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\theta x}e^{i(\mathfrak{m}_1+\mathfrak{m}_2)\theta - \frac{1}{2}(\mathfrak{s}_1^2+\mathfrak{s}_2^2)\theta^2}d\theta = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\theta x}e^{i(\mathfrak{m}_1+\mathfrak{m}_2)\theta - \frac{1}{2}(\mathfrak{s}_1^2+\mathfrak{s}_2^2)\theta^2}d\theta.$$

Write $s^2 = \mathfrak{s}_2^2 + \mathfrak{s}_1^2, m = \mathfrak{m}_1 + \mathfrak{m}_2$.

Then the integral becomes

$$\frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\theta x}e^{im\theta - \frac{1}{2}s^2\theta^2}d\theta$$

which is the density function of distirbution $N(m, s^2)$ by Lévy's inversion.

Exercise:

(a) Calculate the characteristic function $\varphi_B(\theta) = \mathbf{E}[e^{i\theta B}]$ of a random variables B such that $\mathbf{P}[B = 1] = p$ and $\mathbf{P}[B = 0] = p - 1$ (denoted $B \sim$ Bernoulli(p)).

(b) Let $p \in [0, 1], n \in \mathbb{N}$. Calculate the characteristic function $\varphi_X(\theta) = \mathbf{E}[e^{i\theta X}]$ of a random variables X s.t. $\mathbf{P}[X = k] = \binom{n}{k}p^k(1-p)^{n-k}$ for all $k \in \{0, 1, 2, \ldots, n\}$ (we denote $X \sim$ Bin(n, p)).

(c) Let $B_1, \ldots, B_n$ be independent indentically distributed with $\mathbf{P}[B_j = 1] = p$ and $\mathbf{P}[B_j = 0] = 1 - p$, for all j. Compute the characteristic function of $S = B_1 + \cdots + B_n$ using part independence formula for characteristic funtion.

Sol:

(a)

$$\varphi_B(\theta) = \int_{\mathbb{R}} e^{i\theta x}dP_B(x) = 1 - p + e^{i\theta}p.$$

(b)

$$\varphi_X(\theta) = \int_{\mathbb{R}} e^{i\theta x}dP_X(x) = \sum_{k=0}^{n} e^{i\theta k}\binom{n}{k}p^k(1-p)^{n-k}.$$

(c) If we hvae n independent Bernoulli trials then

$$\varphi_{B_1+\cdots+B_n}(\theta) = \varphi_{B_1}(\theta)\cdots\varphi_{B_n}(\theta) = (1-p+e^{i\theta}p)^n.$$

By the Binomial formula the RHS is equal to $\sum_{k=0}^n \binom{n}{k}(1-p)^{n-k}(e^{i\theta}p)^k = \varphi_X(\theta)$.

11.2 Taylor expansion of a characteristic function

By Lévy's inversion theorem, the characteristic function φ_X of a random variable X contains all the information about the distribution P_X of X. In particular it should contain information about the expected value, variance etc. To see why this is atleast formalyl true write the power series expansion

$$e^{i\theta X(\omega)} = \sum \frac{1}{n!}(i\theta X(\omega))^n = 1 + i\theta X(\omega) - \frac{1}{2}i\theta^2 X(\omega)^2 + \cdots \quad ,\forall \omega \in \Omega.$$

If the expected value can be taken, then we would get

$$\varphi_X(\theta) = \mathbf{E}[e^{i\theta X}] \overset{?}{=} 1 + 1\theta\mathbf{E}[X] - \frac{1}{2}\theta^2\mathbf{E}[X^2] + \cdots$$

Formally, therefore the expected value seems to be encoded in the firts order term in the taylor expansion of $\varphi_X(\theta)$ around the point $\theta = 0$. And variance in encoded up in terms up to order two, and more. This is of course only meaningful assuming the random variables belongs in the right moment class $X \in \mathcal{L}^p(\mathbf{P})$. Following lemma mames precise sense for square integrable random variables.

Proposition 16. *Let $X \in \mathcal{L}^2(\mathsf{P})$ be a square integrable random variable and let $\varphi_X : \mathbb{R} \to \mathbb{C}$ be its characteristic function. Then we have*

$$\varphi_X(\theta) = 1 + \mathrm{i}\theta\mathsf{E}[X] - \frac{1}{2}\theta^2\mathsf{E}[X^2] + \epsilon(\theta),$$

where the function $\epsilon : \mathbb{R} \to \mathbb{C}$ is an error term of smaller order than θ^2 in the sense that

$$\frac{|\epsilon(\theta)|}{|\theta|^2} \to 0 \qquad \text{as } \theta \to 0.$$

Proof. The idea is to Taylor expand $e^{i\theta X}$ up to order two, with controlled error term. Firstly $\frac{d}{du}e^{i\theta u} = i\theta e^{i\theta u}$, so for any $x \in \mathbb{R}$

$$\int_0^x i\theta e^{i\theta u} du = e^{i\theta x} - 1.$$

Solving for $e^{i\theta u}$, we get

$$\begin{aligned} e^{i\theta u} &= 1 + i\theta \int_0^x e^{i\theta u} du \\ &= 1 + i\theta \int_0^x 1 + (i\theta \int_0^u e^{i\theta v} dv) du \\ &= 1 + i\theta x - \theta^2 \int_0^x \int_0^u e^{i\theta v} dv du \\ &= 1 + i\theta x - \theta^2 \int_0^x \int_0^u (1-1) + e^{i\theta v} dv du \\ &= 1 + i\theta x - \theta^2 \frac{x^2}{2} - \theta^2 \int_0^x \int_0^u (e^{i\theta v} - 1) dv du \end{aligned}$$

Let $R(\theta, x) := \int_0^x \int_0^u (e^{i\theta v} - 1) dv du$. We investigate it's magnitude.

$$|R(\theta, x)| \leq \int_0^{|x|} \int_0^v |e^{i\theta v} - 1| dv du \leq 2 \int_0^{|x|} \int_0^v |\sin(\theta v/2)| dv du.$$

If we estimate sine term by ≤ 1, we get $|R(\theta, x)| \leq |x|^2$. If we estimate $|\sin(\theta v/2)| \leq |\theta||v|/2$, then $|R(\theta, x)| \leq |\theta||x|^3/6$. In particular this shows $|R(\theta, x)| \to 0$ as $\theta \to 0$. By above observations pointwise

$$e^{i\theta X(\omega)} 1 + i\theta X(\omega) - \theta^2 \frac{X(\omega)^2}{2} - \theta^2 R(\theta, X(\omega)).$$

Taking the expectation, we get that

$$\epsilon(\theta) := \varphi_X(\theta) - (1 + \mathrm{i}\theta \mathsf{E}[X] - \frac{1}{2}\theta^2 \mathsf{E}[X^2]) = \mathbf{E}[e^{i\theta X} - 1 - \mathrm{i}\theta X + \frac{1}{2}\theta^2 X^2]$$

which is precicely $-\theta^2 \mathbf{E}[R(\theta, X)]$. Thus $|\epsilon(\theta)| \leq |\theta|^2 \mathbf{E}[|R(\theta, X)|]$ which by above estimates

$$\frac{|\epsilon(\theta)|}{|\theta|^2} \leq \mathbf{E}[|R(\theta, X)|] \to 0$$

as $\theta \to 0$. □

11.3 Convergence in distribution

The convergence in distribution is a different notion of convergence. Not stircly a convergence of random variables, but measures on the real line.

Theorem 19 (Equivalent notions of convergence in distribu)**.** *Let $X_1, X_2, \ldots$ and X be real-valued random variables. Let also $F_1, F_2, \ldots$ and F be their cumulative distribution functions, and let $\varphi_1, \varphi_2, \ldots$ and φ be their characteristic functions, respectively. Then the following conditions are equivalent:*

(i) For all bounded continuous functions $f : \mathbb{R} \to \mathbb{R}$ we have $\mathsf{E}[f(X_n)] \to \mathsf{E}[f(X)]$ as $n \to \infty$.

(ii) We have $F_n(x) \to F(x)$ as $n \to \infty$ for all points $x \in \mathbb{R}$ such that F is continuous at x.

(iii) We have $\varphi_n(\theta) \to \varphi(\theta)$ as $n \to \infty$ for all $\theta \in \mathbb{R}$.

We skip this proof since it is very long.

Definition 24 (Convergence in distribuion)**.** Let $X_1, X_2, \dots$ and X be real-valued random variables. We say that X_n *tend to X in distribution (or in law)* as $n \to \infty$ and denote $X_n \xrightarrow{\text{law}} X$, if any (then all) of the equivalent conditions of above theorem hold.

The definition in form of distributions: Let $f \colon \mathbb{R} \to \mathbb{R}$ be bdd continuous function, so $\mathbf{E}[f(X_n)]$ and $\mathbf{E}[f(X)]$ can be written using the distributions P_{X_n} and P_X. Therefore this condition reads

$$\int_{\mathbb{R}} f dP_{X_n} \to \int_{\mathbb{R}} f dP_X$$

as $n \to \infty$ for all such f.

11.4 Central limit theorem

Theorem 20 (Central limit theorem)**.** *Let $X_1, X_2, \dots \in \mathcal{L}^2(\mathbf{P})$ independent identically distributed random variables. Denote*

$$\mathfrak{m} := \mathbf{E}[X_j] \;\;\&\;\; \mathfrak{s}; = \sqrt{\mathrm{Var}(X_j)} > 0.$$

(variance is zero only for a.s. constants which are not interesting) Let $S_n = X_1 + \dots + X_n$, then

$$\frac{S_n - n\mathfrak{m}}{\mathfrak{s}\sqrt{n}} \xrightarrow{law} Z$$

where $Z \sim N(0,1)$.

Proof. We may always assume $\mathfrak{m} = 0$ by renormalization argument. Likewise we may assume that $\mathfrak{s} = 1$. The goal is then to show that $\frac{S_n}{\sqrt{n}} \xrightarrow{\text{law}} Z$. By assumptions all the characteristic functions of X_j's are equal, so just denote it by $\varphi(\theta)$. By the series expansion and $\mathfrak{m} = 0, \mathfrak{s} = 1$, we $\varphi(\theta) = 1 - \frac{1}{2}\theta^2 + \epsilon(\theta)$ where $\epsilon(\theta)/|\theta|^2 \to 0$ as $\theta \to 0$. Hence

$$\varphi_{S_n}(\theta) = \mathbf{E}[e^{i\theta \sum_{j=1}^n X_n}] = \mathbf{E}[\prod_{j=1}^n e^{i\theta X_j}] \overset{\perp\!\!\!\perp}{=} \prod_{j=1}^n \mathbf{E}[e^{i\theta X_j}] = \varphi(\theta)^n.$$

Thus $\varphi_{S_n/\sqrt{n}}(\theta) = \varphi(\theta/\sqrt{n})^n$ Also $\varphi(\theta/\sqrt{n}) = 1 - \theta^2/2n + r_n$ where $r_n/(1/n) \to 0$ as $n \to \infty$, so

$$\varphi_{S_n/\sqrt{n}}(\theta) = (1 - \theta^2/2n + r_n)^n.$$

We will show that this limit approaches $e^{-\frac{1}{2}\theta^2}$. Substitute $u = -\theta^2/2$ so we look at the limit $\lim_{n\to\infty}(1 + \frac{u}{n} + r^n)^n = e^u$ and since the characteric function determine the distribution, we are done. Taking principle log on both sides makes we want to show $\lim_{n\to\infty} n \log\bigl(1 + \frac{u}{n} + r_n\bigr) = u$. Since $r_n \to 0, u/n \to 0$ we can use the Taylor expansion of log for large n so it is in the radius of convergence to write

$$n \log\left(1 + \frac{u}{n} + r_n\right) = n \left(\sum_{m=1}^{\infty} (-1)^m \frac{(\frac{u}{n} + r_n)^m}{m} \right).$$

We will look at the tail of this when n grows even larger

$$n \left(\sum_{m=1}^{\infty} (-1)^m \frac{(\frac{u}{n} + r_n)^m}{m} \right) = u + nr_n + n\text{Error}.$$

The Error term behaves asymptotically for the symbol $z_n := \frac{u}{n} + r_n$ like $O(z_n^2)$ when $z_n \to 0$. So nError $\approx O(z_n)$, therefore the entire thing approaches u. □

12 Martingales

12.1 Conditional expectancy

Given a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, integrable random variables Y and sub-sigma-algebra $\mathcal{G} \subset \mathcal{F}$, we will define and study the **conditional expeced value** $\mathbf{E}[Y|\mathcal{G}]$ of Y given $\mathcal{G}$. This should be interpreted as the best estimate of Y that can be based on information given by $\mathcal{G}$. First we will assume Y is square-integrable.

If $\mathcal{G}$ is as above, then $\mathcal{L}^2(\mathbf{P}) \cap m\mathcal{G}$ is a closed (any convergeing sequence in the 'norm' is also $\mathcal{L}^2(\mathbf{P})$) subspace (w.r.t. the $\mathcal{L}^2-$norm) of $\mathcal{L}^2(\mathbf{P})$. Note that we do not quotient by a.s., so the 'norm' is not really a norm and 'closed' is not really topologically closed. The 'inner-product' will be just $\langle X, Y\rangle := \mathbf{E}[XY]$.

Not our task is to estimate Y by some $\hat{Y}$ based on information $\mathcal{G}$, where $Y - \hat{Y}$ is the difference between the actual value and estimate. The 'norm' $\left\|Y - \hat{Y}\right\|$ is the magnitude of error. In a sense then the best estimation would be $\hat{Y} \in \mathcal{L}^2(\mathbf{P}) \cap m\mathcal{G}$ that minimizes $\left\|Y - \hat{Y}\right\|$. This is given by the orthogonal projection.

Proposition 17 (Orthogonal projection in $\mathcal{L}^2(\mathbf{P})$)**.** *Let $\mathcal{V} \subset \mathcal{L}^2(\mathrm{P})$ be a closed subspace of square integrable random variables and let $Y \in \mathcal{L}^2(\mathrm{P})$ be a square integrable random variable. Define the distance of Y to the subspace $\mathcal{V}$ by*

$$\Delta := \inf_{X \in \mathcal{V}} \|Y - X\|. \tag{E.6}$$

Then for a random variable $Z \in \mathcal{V}$ the following conditions are equivalent:

$$(i) : \|Y - Z\| = \Delta \qquad\qquad (ii) : Y - Z \perp \mathcal{V}.$$

Furthermore, there exists a random variable $Z \in \mathcal{V}$ with these properties, and if $\widetilde{Z} \in \mathcal{V}$ is another such random variable, then we have $Z = \widetilde{Z}$ almost surely.

Lemma 24. *Let $\mathcal{G} \subset \mathcal{F}$ be a sub-sigma-algebra and $Y \in \mathcal{L}^2(\mathbf{P})$, $\hat{Y}$ orthogonal projection of Y to $\mathcal{L}^2(\mathbf{P}) \cap m\mathcal{G}$. Then for any $G \in \mathcal{G}$ we have*

$$\mathbf{E}[\mathbb{I}_G \hat{Y}] = \mathbf{E}[\mathbb{I}_G Y]].$$

Proof. $0 = \langle Y - \hat{Y}, \mathbb{I}_G\rangle = \langle Y, \mathbb{I}_G\rangle - \langle \hat{Y}, \mathbb{I}_G\rangle = \mathbf{E}[\mathbb{I}_G Y] - \mathbf{E}[\mathbb{I}_G \hat{Y}]].$ □

This motivates a general definition let $Y \in \mathcal{L}^1(\mathbf{P})$ and $\mathcal{G} \subset \mathcal{F}$ be a sub-sigma-algebra.

Definition 25. A random variable is said the be (a version of) the **conditional expected value** (denoted $\mathbf{E}[Y|\mathcal{G}]]$ of Y given $\mathcal{G}$, if $\hat{Y} \in m\mathcal{G}$ and

$$\mathbf{E}[\mathbb{I}_G \hat{Y}] = \mathbf{E}[\mathbb{I}_G Y]] \text{ for all } G \in \mathcal{G}.$$

In particular, in the $\mathcal{L}^2-$case (a version of) is given by orthogonal projection. We still have to show the existance of such conditional expected values.

Lemma 25. *If $\hat{Y}, \hat{Y}'$ are conditional expected values of Y given $\mathcal{G}$. Then $\hat{Y} \overset{a.s.}{=} \hat{Y}'$.*

Proof. Define

$$G_n = \{\omega \in \Omega | \hat{Y}(\omega) - \hat{Y}'(\omega) \geq 1/n\}.$$

Then $G_n \in \mathcal{G}$, since both $\hat{Y}, \hat{Y}'$ are $\mathcal{G}-$measurable. By Markov's inequality

$$\frac{1}{n}\mathbf{P}[G_n] \leq \mathbf{E}[\mathbb{I}_{G_n}(\hat{Y} - \hat{Y}')] = \mathbf{E}[\mathbb{I}_{G_n}\hat{Y}] - \mathbf{E}[\mathbb{I}_{G_n}\hat{Y}'] = \mathbf{E}[\mathbb{I}_{G_n}Y] - \mathbf{E}[\mathbb{I}_{G_n}Y] = 0.$$

Os $\mathbf{P}[G_n] = 0$. This means that

$$\mathbf{P}[\hat{Y} > \hat{Y}'] = \mathbf{P}[\cup_{n\in\mathbb{N}} G_n] \leq \sum_{n\in\mathbb{N}} \mathbf{P}[G_n] = 0.$$

By symmetry $\mathbf{P}[\hat{Y}' > \hat{Y}] = 0$, so $\mathbf{P}[\hat{Y} = \hat{Y}'] = 1$. □

In general we will denote expected values of Y givne $\mathcal{G}$ by $\mathbf{E}[Y|\mathcal{G}] \in \mathcal{L}^1(\mathbf{P}) \cap m\mathcal{G}$ not worrying about the different choices that could be made. Assuming the conditional expected values exists one can show that $Y \geq 0$ a.s. $\Rightarrow \mathbf{E}[Y|\mathcal{G}] \geq 0$ a.s.

Proof. Let $\hat{Y}$ be a conditional expectancy for Y given $\mathcal{G}$. The set $G = \{\hat{Y} < 0\} \in \mathcal{G}$, so

$$0 \leq \mathbf{E}[\mathbb{I}_G Y] = \mathbf{E}[\mathbb{I}_G \hat{Y}].$$

But we have that $(\mathbb{I}_G \hat{Y})(\omega) < 0$ for all $\omega \in G$ and else zero. This implies that G must have measure zero for $\mathbf{E}[\mathbb{I}_G \hat{Y}]$ to be non-negative. □

To show that generally $\mathcal{L}^1-$integrbale conditional expectation exists, we first condier non-negatives and then generla case. Define the truncation at levle n of non-negative random variable Y by $Y \wedge n(\omega) = \min(Y(\omega), n)$, so $Y \wedge n \uparrow Y$. We use this approach on square integrable's to approximate Y.

Lemma 26. *Let Y be non-negative random variables, Z_n the orthogonal projection of $Y \wedge n$ to $\mathcal{L}^2(\mathbf{P}) \cap m\mathcal{G}$. Then there exists $m\mathcal{G}-$measurable integrable randon variable Z, such that $Z_n \uparrow Z$ almost surely, and we have*

$$\mathbf{E}[\mathbb{I}_G Z] = \mathbf{E}[\mathbb{I}_G Y] \textit{ for all } G \in \mathcal{G}.$$

Proof. Since $Y \wedge (n+1) \leq Y \wedge n$, then by linearity of projection $Z_n \leq Z_{n+1}$ almost surely. This means that Z_n converges to some $Z \in m\mathcal{G}$ from below almost surely (the pointwise supremum). Applying monotone convergence twice

$$\begin{aligned} \mathbf{E}[\mathbb{I}_G Z] &= \lim_{n\to\infty} \mathbf{E}[\mathbb{I}_G Z_n] \\ &= \lim_{n\to\infty} \mathbf{E}[\mathbb{I}_G (Y \wedge n)] \\ &= \mathbf{E}[\mathbb{I}_G Y]. \end{aligned}$$

Since Y is integrable, letting $\Omega = G$ one gets $\mathbf{E}[Z] = \mathbf{E}[Y] < +\infty$ to get integrability. □

Proposition 18. *The conditional expectation* $\mathbf{E}[Y|\mathcal{G}]$ *of* $Y \in \mathcal{L}^1(\mathbf{P})$ *exists.*

Proof. Split into $Y = Y_+ - Y_-$ and define $\mathbf{E}[Y|\mathcal{G}] = \hat{Y}_+ - \hat{Y}_-$ which are given by the non-negative lemma above. □

Theorem 21 (Properties of conditional expectation)**.** *Conditional expected values satisfy the following properties (interpreted in the almost sure sense), when* Y *and* $Y_1, Y_2, \ldots$ *are integrable random variables.*

(i) If $Y \in \mathrm{m}\mathcal{G}$*, then we have* $\mathsf{E}[Y|\mathcal{G}] = Y$*.*

(ii) We have $\mathsf{E}[\mathsf{E}[Y|\mathcal{G}]] = \mathsf{E}[Y]$*.*

(iii) For $c_1, c_2 \in \mathbb{R}$*, we have* $\mathsf{E}[c_1Y_1 + c_2Y_2 \mid \mathcal{G}] = c_1\mathsf{E}[Y_1 \mid \mathcal{G}] + c_2\mathsf{E}[Y_2 \mid \mathcal{G}]$*.*

(iv) If $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$ *are* σ*-algebras, then we have* $\mathsf{E}[\mathsf{E}[Y|\mathcal{G}] \mid \mathcal{H}] = \mathsf{E}[Y|\mathcal{H}]$*.*

(v) If $Z \in \mathrm{m}\mathcal{G}$ *and* $ZY \in \mathcal{L}^1(\mathsf{P})$*, then we have* $\mathsf{E}[ZY|\mathcal{G}] = Z\,\mathsf{E}[Y|\mathcal{G}]$*.*

(vi) If $\mathcal{G} \perp\!\!\!\perp \sigma(Y)$*, then we have* $\mathsf{E}[Y|\mathcal{G}] = \mathsf{E}[Y]$*.*

(vii) Known quantities can be treated like constants when forming best estimates.

(viii) Any information that is independent of Y *can not be used to estimate* Y *any better than the expected value* $\mathbf{E}[Y]$ *of* Y*.*

Remark: (Interpretations of the properties above)

(i) Best estimate of a known quantity $Y \in m\mathcal{G}$ is the quantity itself.

(ii) The best estimate of a quantity Y is unbiased, in the sense that it has the same expected values as the quantity Y itself.

(iii) The best estimate of a linear combination of quantities is the corresponding linear combination of the best estimates.

(iv) Suppose that a person H possesses less information than a person G. If H tries to form an estimate about the best estimate that G makes about some quantity Y, then the best she can do is to use her own best estimate of the quantity Y

Also for example Monotone convergence theorem, Dominated convergence theorem, Fatou's lemma, Jensen's inequality, etc. hold in the appropriate form for conditional expected values, and their proofs are straightforward modifications of the corresponding ones for usual expected values.

12.2 Martingales

Definition 26. A **filtered space** $(\sigma, \mathcal{F}, \{F_n\}, \mathbf{P})$ consists of the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a **filtration** $\{\mathcal{F}_n | n \geq 0\}$ that is an increasing family of sub-sigma-algebras of $\mathcal{F}$:

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}.$$

We define $\mathcal{F}_\infty = \sigma(\cup_{n \geq 0} \mathcal{F}_n) \subset \mathcal{F}$.

The intuitive idea: The information about $\omega \in \Omega$ available to us at (or, if you prefer 'just after') time n consists precisely of the values of $Z(\omega)$ for all $\mathcal{F}_n-$measurable functions Z. Usually, $\{\mathcal{F}_n\}$ is the **natural filtration**

$$\mathcal{F}_n = \sigma(W_0, \ldots, W_n)$$

of some (stochastic) process $W = (W_n | n \in \mathbb{Z}_{\geq 0})$, and then the information about ω which we have at time n consists of the values

$$W_0(\omega), \ldots, W_n(\omega).$$

Definition 27. A process $X = (X_n | n \in \mathbb{Z}_{\geq 0})$ is called **adapted** (to the filtration $\{\mathcal{F}_n\}$ if for each n, X_n is $\mathcal{F}_n-$measurable.

Intuitive idea: If X is apapted, the value $X_n(\omega)$ is known to us at time n. Usually $\mathcal{F}_n = \sigma(W_0, \ldots, W_0)$ and $X_n = f_n(W_0, \ldots, W_n)$ for borel $f\colon \mathbb{R}^{n+1} \to \mathbb{R}$.

Definition 28. A process X as above is called a **martingale** (relative to $\{\mathcal{F}_n, \mathbf{P}\}$ if

- X is adapted
- $\mathbf{E}[|X_n|] < +\infty, \forall n$
- $\mathbf{E}[X_n | \mathcal{F}_{n-1}] = X_{n-1}$ a.s. $(n \geq 1)$.

Supermartingale if $\mathbf{E}[X_n|\mathcal{F}_{n-1}] \leq X_{n-1}$ and **submartingale** if $\mathbf{E}[X_n|\mathcal{F}_{n-1}] \geq X_{n-1}$. Supermatringales 'decrease on average' and submartingales 'increase on average'. [Supermartingale corresponds to superharmonic: a function $f\colon \mathbb{R}^n \to \mathbb{R}$ is superharmonic iff for a Brownian motion B on $\mathbb{R}^n$, $f(B)$ is a local supermartingale relative to the natural filtration of B.]

13 Markov chains