MATH 587 - Advanced Probability I

Pr. Louigi Addario-Berry

Course notes by
Léo Raymond-Belzile
Leo.Raymond-Belzile@mail.mcgill.ca

The current version is that of December 4, 2013
Fall 2013, McGill University

Please signal the author by email if you find any typo.

These notes have not been revised and should be read carefully.

Contents

Chapter 1
Introduction: motivating examples

We start with a combinatorial probability problem, which is due to Polyá, in a paper in 1923.

1.1 Random walks on lattices

Suppose people are walking on an infinite grid $\mathbb{Z}^2$ where the origin is $S_0 = 0$ and S_{n+1} is a neighbor of S_n; where each neighbor is equally likely. The question is whether $(S_n, n \geq 1)$ is **transient** or **recurrent**.

Definition 1.1 (Transient, recurrent)

$$\begin{aligned}\text{Transient: } \mathsf{P}\left(S_n = 0 \text{ for some } n \geq 1\right) < 1\\ \text{Recurrent: } \mathsf{P}\left(S_n = 0\text{for some } n \geq 1\right) = 1.\end{aligned}$$

It turns out that

Theorem 1.2 (Polyá)

The random walk $(S_n, n \geq 1)$ is **recurrent** in $\mathbb{Z}^d$ for $d \leq 2$ and **transient** for $d \geq 3$.

Remark

The lattice here is an infinite system. The dichotomy between the two behaviors is for infinite space. To go back to Polyá question about the two-dimensional random walk for two individuals, we can consider the distance between two individuals as being a vector, and reduce this question to that of the single random walk on the lattice.

In dimension 1 $(d = 1)$, we have $S_0 = 0$ and

$$S_{n+1} = \begin{cases} S_n \quad 1 & \text{with prob. } \frac{1}{2} \\ S_n - 1 & \text{with prob. } \frac{1}{2} \end{cases}$$

we could extend this to having arbitrary probabilities $p, 1 - p$. In this case, one would not expect the process to be recurrent, since we tend to favor one direction, say left. Only in the symmetric case is the event recurrent.

Proposition 1.3

A random walk (RW) is recurrent in $d = 1$.

Proof Let $z_n = \mathsf{P}\left(S_n = 0\right)$ and $f_n = \mathsf{P}\left(S_i \neq 0, i \in \{1, 2, \dots, n-1\}, S_n = 0\right)$. Then, let

$$P(s) = \sum_{i=0}^{\infty} z_n s^n \qquad F(s) = \sum_{n=0}^{\infty} f_n s^n$$

The radius of convergence is definitively one since the series are dominated by a geometric series. We can also view them as a power series, and hope to extract information about the coefficients for the Taylor series.

Claim

$$P(s) = 1 + P(s)F(s)$$

Proof Observe

$$z_n = \sum_{k=1}^{n} f_k z_{n-k};$$

that is the first return at time k, then you need to displacement zero in $n = k$ steps remaining. This is a conditional probability, that is

$$\mathsf{P}\left(S_n = 0 \mid S_k = 0, S_i \neq 0, S_i \neq 0 \text{ for } i \in \{1, 2, \dots, k-1\}\right)$$

that can be obtained using Bayes theorem, to write the product of the independent events. ■

Exercise 1.1

Write a 356 proof of this observation.

Then multiply both sides of the equality by s^n as to get

$$z_n s^n = \sum_{k=1}^{n} (z_{n-k} s^{n-k})(f_k s^k)$$

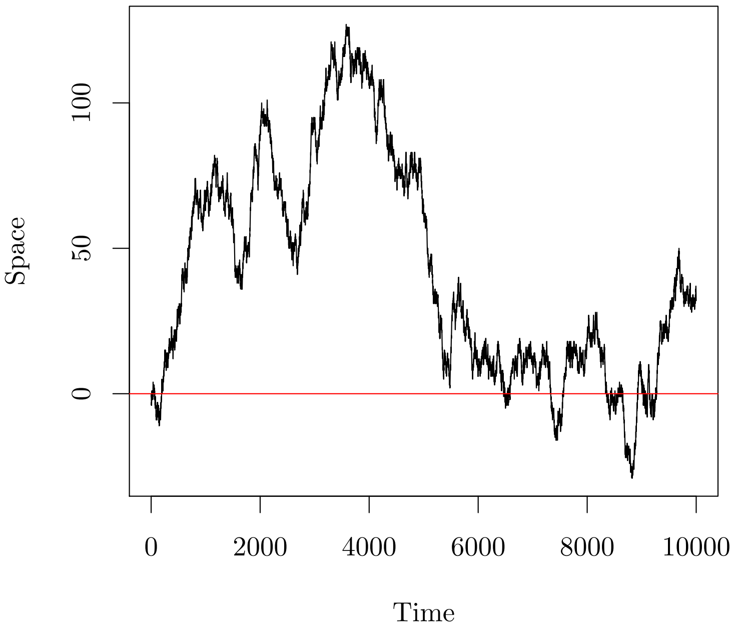


Figure 1: Random walk in space and time

if you hit zero, either the axis was traversed before or it is the first occurence

and sum from $n = 0$ to infinity. For $n = 0$, the left hand side is zero.

$$\begin{aligned} P(s) &= 1 + \sum_{n=1}^{\infty} \sum_{k=1}^{n} (z_{n-k} s^{n-k})(f_k s^k) \\ &= 1 + P(s)F(s) \end{aligned}$$

Observe that the relationship above does not depend on p, but the coefficients of z_n will.

Observation

We can write

$$z_n = \begin{cases} 0 & \text{if } n \text{ is odd} \\ \binom{n}{n/2} p^{\frac{n}{2}} (1-p)^{\frac{n}{2}} & \text{if } n \text{ is even} \end{cases}$$

so that

$$P(s) = \frac{1}{\sqrt{1-4p(1-p)s^2}}$$

and knowledge of $P(s)$ allows us to obtain using the linear recurrence an expression for $F(s)$

$$F(s) = 1 - \sqrt{1-4p(1-p)s^2}.$$

In particular, if $p = \frac{1}{2}$, we get

$$\begin{aligned} P(s) &= \frac{1}{\sqrt{1-s^2}} \\ F(s) &= 1 - \sqrt{1-s^2} \end{aligned}$$

For the remainder of the proof, the probability

$$\mathsf{P}\left(S_n = 0 \text{ for some } n \geq 1\right) = \mathsf{P}\left(\bigcup_{n=1}^{\infty} \{S_n = 0, S_i \neq 0 \; \forall \, i \in \{1, , \ldots, , n-1\}\}\right)$$

since there must be a first return at time n. Since the events are disjoint, this is equal to

$$\sum_{n=1}^{\infty} \mathsf{P}\left(S_n = 0, S_i \neq 0 \; \forall \, i \in \{1, , \ldots, , n-1\}\right) = \sum_{n=1}^{\infty} f_n = \sum_{n=1}^{\infty} f_n 1^n = F(1) = 1$$

What we have not shown in this proof is the countable additivity of the $S_i \neq 0$, and the interchange of the sum and the probability (which is really of limit). We can do this more generally to integrate a function by integrating a sequence of functions as lower bound. This is due to the monotone convergence theorem.

The second thing that needs justification is the rigorous sense of s_n, which is somehow an infinite sequence of numbers. The elementary events are equivalent to infinite sequences taking values in $\{\pm 1\}^{\mathbb{N}}$. Knowledge of the sequence fully determines the behavior of the random walk.

Fact 1.4

It is impossible to define a probability "measure" P (in a non-trivial fashion) so that every subset of $\{\pm 1\}^{\mathbb{N}}$ has a well-defined probability and the basic axioms are satisfied . See Appendix 0 of the book. This is know as the Banach-Tarski paradox. There is a nice paper on this entitled *A non-measurable set from coin tosses.*

If $p \neq \frac{1}{2}$, the random walk will drift. Recurrent means that the expected number of returns to zero is infinite, while for transient (fact) the number of return will be finite. In two-dimension, we can get a similar expression. Using Stirling formula to approximate the factorial, we would get

$$\binom{2m}{m}\left(\frac{1}{2}\right)^{2m} \to \frac{c}{\sqrt{m}}$$

while in two dimension, $\frac{c'}{m}$ and in d, $\frac{c_d}{m^{d/2}}$ and if $d = 1, 2$, the series diverge, but not otherwise.

This also generalize this to an electrical network on a graph, and whether the resistance is finite or infinite is equivalent to transient or recurrent behavior of the random walk. ■

1.2 Galton-Watson trees

Consider a (family) tree, with children and each subtree has descendants or the line goes extinct. In general, one could go down in the line, and line of descent of family names; this is a general model for extinction or survival of the specie. For atoms bumping into each other in a nuclear reactor, this is a good model (chain reaction).

The model then is as follows: each individual has a random number of children B, independently of all others. The questions is survival or extinction of the process. If $\mathsf{E}(B) < 1$, then it is obvious that P (Extinction), say for $\mathsf{E}(B) = \frac{1}{2}$, each tree has an average of 2 individuals per line, and so will go extinct. If $\mathsf{E}(B) > 1$, then P (Extinction) < 1. If $\mathsf{E}(B) = 1$, then extinction happens with probability 1 again, unless B isn't actually random, that is if each line has one child.

If we go back to the symmetric random walk, then with $\mathsf{P}(B = 2) = \frac{1}{2}$ and $\mathsf{P}(B = 0) = \frac{1}{2}$. This turns back to the case of the random walk, either go up by one or not.

The number of children, termed the **branch factor**, is a random variable B taking nonnegative integers as values.

Write Z_n for the number of individuals in generation n. What is the probability that $\mathsf{P}\left(Z_n = 0 \text{ for some } n\right)$, that is extinction.

Theorem 1.5 (Fundamental theorem of Branching Processes)

The probability of extinction is less than 1 if and only if $\mathsf{P}\left(B = 1\right) = 1$ or $\mathsf{E}\left(B\right) = \sum_{k \geq 0} k\mathsf{P}\left(B = k\right) > 1$.

Lemma 1.6

$\mathsf{E}\left(Z_n\right) = \left(\mathsf{E}\left(B\right)\right)^n$

Proof Using conditional expectations and the law of total probability, write

$$\begin{aligned} \mathsf{E}\left(Z_n\right) &= \sum_{k \geq 0} \mathsf{P}\left(B_1 = k\right) \cdot \mathsf{E}\left(Z_n \mid B_1 = k\right) \\ &= \sum_{k \geq 0} \mathsf{P}\left(B_1 = k\right) \cdot k \cdot \mathsf{E}\left(Z_{n-1}\right) \\ &= \mathsf{E}\left(Z_{n-1}\right) \mathsf{E}\left(B\right) \end{aligned}$$

using the fact that $\mathsf{E}\left(Z_{n-1}\right)$ does not depend on k, we can take it outside the sum and use the statement of the theorem. The proof follows by induction. ■

Corollary 1.7

If $\mathsf{E}\left(B\right) < 1$, then $\mathsf{E}\left(\text{total pop. size}\right) = \sum_{n \geq 0} \left(\mathsf{E}\left(B\right)\right)^n = \frac{1}{1 - \mathsf{E}(B)} < \infty$. This tells us that $\mathsf{P}\left(\text{survival}\right) = \mathsf{P}\left(\text{total pop. size} = \infty\right) = 0$.

Define the following:

Definition 1.8

Let $F(z) = \mathsf{E}\left(Z^b\right) = \sum_{k=0}^{\infty} \mathsf{P}\left(B = k\right) z^k$

Proposition 1.9

If $\mathsf{P}\left(B = 1\right) < 1$ then

$$\begin{aligned} \mathsf{P}\left(\text{extinction}\right) \equiv \mathsf{P}\left(\sum_{n \geq 0} Z_n < \infty\right) &\equiv \mathsf{P}\left(\exists n : Z_n = 0\right) \\ &= \mathsf{P}\left(\exists n_0 : \ \forall\, n \geq n_0, Z_n = 0\right) = \min_{x \geq 0} \left\{F(x) = x\right\} \end{aligned}$$

If one was to take the derivative inside the sum, then we can evaluate $F'(1)$; indeed

$$F'(z) = \sum_{k=1}^{\infty} \mathsf{P}\left(B = k\right) k z^{k-1}$$

so that $F'(1) = \mathsf{E}\left(B\right)$. It will thus be below the curve and reach above; by intermediate value theorem, there must be a point $z < 1$ where the two curves intersect.

Lemma 1.10

If $\mathsf{E}\left(B\right) > 1$, then $\min_{x \geq 0}\left\{F(x) = x\right\} < 1$. If $\mathsf{E}\left(B\right) = 1$, but $\mathsf{P}\left(B = 1\right) < 1$, then $\mathsf{P}\left(\text{extinction}\right) = 1$ as $\min_{x \geq 0}\left\{F(x) = x\right\} = 1$.

Proof To show the proposition, want to show that this is a solution, and then all other are bigger. The second part is left as a exercise. Branching processes are the analog of functional composition.

Let $F_1 := F$ and $F_{n+1} = F \circ F_n$

Claim

$\mathsf{P}\left(Z_n = 0\right) = F_n(0)$

Proof We have already established the case $n = 1$. We proceed by induction; using the law of total probability

$$\begin{aligned} \mathsf{P}\left(Z_n = 0\right) &= \sum_{k \geq 0} \mathsf{P}\left(B_1 = k\right) \mathsf{P}\left(Z_n = 0 \mid B_1 = k\right) \\ &= \sum_{k \geq 0} \mathsf{P}\left(B_1 = k\right) \left(\mathsf{P}\left(Z_{n-1} = 0\right)\right)^k && \text{(independence)} \\ &= F(\mathsf{P}\left(Z_{n-1} = 0\right)) && \text{(definition of } F(z)\text{)} \\ &= F(F_{n-1}(0)) && \text{(induction hypothesis)} \end{aligned}$$

since each subtree extinction are independent, the probability of all subtree dying out in stage n (the event is an intersection) is the product of the probabilities of each subtree.

For the remainder of the proof, note that the event are increasing, as if if $Z_k = 0$, then $Z_{k+i} = 0$ for $i \in \mathbb{N}$. We can use monotonicity to exchange the probability and the limits.

$$\begin{aligned} p \equiv \mathsf{P}\left(\bigcup_{n\geq 0}\{Z_n = 0\}\right) &= \lim_{n\to\infty} \mathsf{P}\left(Z_n = 0\right) \\ &= \lim_{n\to\infty} F_n(0) = \lim_{n\to\infty} F(F_n(0)) = F\left(\lim_{n\to\infty} F_n(0)\right) = F(p) \end{aligned}$$

since $\lim x_n = \lim x_{n+1}$ if a sequence converge, we use the fact that F is continuous to again interchange limits. We get thus $p = F(p)$, in other words the extinction probability is the function of this function . ■

■

Some issues that were raised by our calculations above

1. We used an "infinite number of independent copies of B".
2. If instead of **number** Z_n we cared about **weight** W_n, the proof breaks [1]
3. Limits, convergence, countable closure

[1] The conditional expectation on the set of weights, depends on the fact that B took integer values, countable – if the random variable was continuous, we need a new notions of conditional expectation. We need $\mathsf{P}\left(\text{extinction} \mid W_1\right) = \mathsf{P}\left(\text{extinction}\right)^{Z_1}$, conditioning on a random variable that is still random.

Chapter 2
Measures and σ-algebras

2.1 σ-algebras

We recall some facts about sets. Recall that if Ω is a set, then $\mathcal{A} \subset 2^\Omega$ (the power set of Ω) is a **σ-algebra** if

1. $\Omega \in \mathcal{A}$
2. If $E \in \mathcal{A}, \Rightarrow E^\complement \in \mathcal{A}$
3. If $E_i \in \mathcal{A}, i \geq 1 \Rightarrow \bigcup_{i\geq 1} E_i \in \mathcal{A}$

Exercise 2.1

Can replace the third statement by $E_i \in \mathcal{A} \;\forall\; i \geq 1$, then $E_1 \subset E_2 \subset \cdots$, then $\lim_{i\to\infty} E_i = \bigcup_{i\geq 1} E_i \coloneqq E \in \mathcal{A}$ (that is $E_i \uparrow E \in \mathcal{A}$). Hint: if $\{E_i, i \geq 1\}$ are sets, then with $F_i = \bigcup_{j=1}^{i} E_j$, then $F_i \uparrow$ and $\lim_{i\to\infty} F_i = \bigcup_{i\geq 1} E_i$.

Definition 2.1 (Measurable space)

If $\mathcal{F}$ is a σ-algebra over Ω, then $(\Omega, \mathcal{F})$ is called a **measurable space**.

Example 2.1

Rolling ten dice: $\Omega = \{1, 2, 3, 4, 5, 6\}^{10}$, with $\mathcal{F} = 2^\Omega$.[2]

Definition 2.2 (σ-algebra generated by a collection)

If $S \subset 2^\Omega$ then the **σ-algebra generated by** S is denoted $\sigma(S)$ and defined as

$$\sigma(S) \coloneqq \bigcap_{\substack{\mathcal{F}\supset S:\\ \mathcal{F} \text{ is a } \sigma\text{-algebra}}} \mathcal{F}$$

that is the smallest σ-algebra that contains S. This intersection is non-empty since 2^Ω always satisfies the requirements.

Observation

If $\mathcal{F}$ is a σ-algebra and $\mathcal{F} \supset S$, then $\mathcal{F} \supset \sigma(S)$.

Example 2.2

Consider $\Omega = \mathbb{N}$, $S = \{\{1\}, \{2\}, \{3\}, \ldots\}$, then $\sigma(S) = 2^\mathbb{N}$. Indeed, for any set $B \subset \mathbb{N}$, write $B = \{b_i, i \geq i\}$ and take $E_i = \{b_i\}$ and $\bigcup_{i\geq 1} E_i = B$. So if $\mathcal{F}$ is a σ-algebra over $\mathbb{N}$ and $\mathcal{F}$ contains all singleton sets, then $\mathcal{F} = 2^\mathbb{N}$.

[2] Since the set has finite cardinality, there is no distinction between σ-algebras and algebras.

Example 2.3

Take $\Omega = \mathbb{R}$ and $S = \{\text{singleton sets}\}$; we saw last class that the σ-algebra arising from sets which are countable or have countable intersections included. In fact,

$$\sigma(S) = \Big\{ E \in \mathbb{R}, E \text{ countable or } E^{\complement} \text{ countable} \Big\} .$$

Example 2.4

Let $\Omega := \mathbb{M}$ be a metric space and consider S, the open sets in $\mathbb{M}$ (a concrete example of such is when $\mathbb{M} = \mathbb{R}^d$, where in this setting opens sets are countable unions of open balls). $\sigma(S)$ is denoted $\mathbb{B}(\mathbb{M})$, the **Borel sets** of the metric space $\mathbb{M}$.

Remark

In $\mathbb{R}$, we specify the distribution of a random variable X by its CDF

$$F(z) = \mathsf{P}\left(X \leq z\right) = \mathsf{P}\left(X \in (-\infty, z]\right) .$$

In fact, the σ-algebra generated by closed rays (Borel set of the form $(-\infty, z]$) is $\sigma(\{(-\infty, z], z \in \mathbb{R}\}) = \mathbb{B}(\mathbb{R})$.[3]

If we take

$$(-\infty, b) = \bigcup_{b \geq 1} \left(-\infty, b - \frac{1}{n}\right]$$
$$(a, \infty) = (-\infty, a]^{\complement}$$

since we are closed under unions and complements. Since we also are closed under intersections (as $A \cap B = (A^{\complement} \cup B^{\complement})^{\complement}$), and so the intervals (a, b) are in $\mathbb{B}(\mathbb{R})$.

The earlier example of σ-algebra was not rich enough for our purpose.

2.2 Measures

We can get to probability functions, now that we have the sets to which we want to assign probabilities, restricting our attention to sets in the σ-algebra. We will require from them to satisfy the Kolmogorov's axioms.

[3] In French, they are termed *borélien*.

Definition 2.3 (Measures)

Given a measurable space $(\Omega, \mathcal{F})$, a **probability measure** on $(\Omega, \mathcal{F})$ is a function $\mu : \mathcal{F} \to [0,1]$ with

1. $\mu(\Omega) = 1, \mu(\emptyset) = 0$
2. μ is countably additive. If $(E_i, i \geq 1)$ are disjoint elements of $\mathcal{F}$, then

$$\mu\left(\bigcup_{i\geq 1} E_i\right) = \sum_{i\geq 1} \mu(E_i)$$

Remark

μ is a **measure** if the same holds but with $\mu : \mathcal{F} \to [0,\infty]$. μ is a **σ-finite measure** if $\exists \Omega_i \in \mathcal{F}, i \geq 1$ with $\mu(\Omega_i) < \infty \ \forall \ i$ and such that $\Omega_i \uparrow \Omega$.

In this case, we call $(\Omega, \mathcal{F}, \mu)$ a probability space (or measure space, finite measure space, σ-finite measure space).

Example 2.5

$\Omega = S, \mathcal{F} = 2^S$ and ζ is the counting measure, where

$$\zeta(E) = \begin{cases} |E| & \text{if } |E| < \infty \\ \infty & \text{otherwise .} \end{cases}$$

It is σ-finite if Ω is countable. For instance, in this instance $\mathbb{R}$ is not σ-finite.

Example 2.6

Let $\Omega = \mathbb{R}, \mathcal{F} = \mathbb{B}(\mathbb{R})$. For $E \in \mathcal{F}$, $\mu(E) = \int_E \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, \mathrm{d}x$. Then $\mu(\mathbb{R}) = 1$, since the Gaussian density integrates to 1. It is not clear from this that μ is countably additive. If $(E_i, i \geq 1)$ are disjoint Borel sets then

$$\int_{\cup_{i\geq 1} E_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, \mathrm{d}x = \sum_{i\geq 1} \int_{E_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, \mathrm{d}x$$

which is a form of linearity; we need to be clear about which sets we have here. There is a consistency issue here which needs to be resolved.

Example 2.7 (Lebesgue measure)

The last example was giving the measure for any set E_i. Now we again take $\Omega = \mathbb{R}, \mathcal{F} = \mathbb{B}(\mathbb{R})$ and $\mu([a,b]) = b - a$. This is clearly not finite, nor is it a probability measure. It is σ-finite since we can take intervals $[-n,n]$ of measure $\mu([-n,n]) = 2n$ and as $n \to \infty$, converges to $\mathbb{R}$. This is the one-dimensional Lebesgue measure.

2.3 Extension of measures

The question is: how can we show Lebesgue measure exists and is unique? To answer this question, we will do a digression to discuss **multiplicative functions**. These are used in the proof of Carathéodory extension theorem, which we will skip.

A function $\phi : \mathbb{N} \to \mathbb{R}$ is multiplicative if $\phi(mn) = \phi(m)\phi(n)$ whenever $\gcd(m,n) = 1$.

Theorem 2.4 (Extension theorem for multiplicative functions)

Let $P = \{\text{prime powers}\}$, let $\phi_0 : P \to \mathbb{R}$ with $\phi_0(1) = 1$. Then, there exists a unique multiplicative function $\phi : \mathbb{N} \to \mathbb{R}$ with $\phi|_P = \phi_0$.

We have two main results for measures. The first one is the Carathéodory extension theorem. Before this, a few necessary definitions:

Definition 2.5 (Pre-measure)

If $\mathcal{F}_0$ is an algebra over Ω, then μ_0 is a **pre-measure** on $\mathcal{F}_0$ if

(i) $\mu_0(\emptyset) = 0$

(ii) if $A, B \in \mathcal{F}_0$ with $A \cap B = \emptyset$, then $\mu_0 = (A \cup B) = \mu_0(A) + \mu_0(B)$.

(iii) If $E_i \in \mathcal{F}_0, i \geq 1$ are disjoint **and** $\bigcup_{i \geq 1} E_i \in \mathcal{F}_0$, then $\mu_0\left(\bigcup_{i\geq 1} E_i\right) = \sum_{i \geq 1} \mu_0(E_i)$.

Why do we need the last criterion? A problem that arises is that we can break a set such as $(0,1)$ into $\left(0, \frac{1}{2}\right) \cup \left(\frac{1}{2}, \frac{3}{4}\right) \cup \left(\frac{3}{4}, \frac{7}{8}\right) \cup \cdots$. In such case, we are good, but one may not be lucky and there could be two ways to express the measure of a set, as a countable collection or a single element. In such case, we could not extend uniquely.

Exercise 2.2

Define a function that is additive, but not countably additive. Hint: consider $\mathbb{N}$ and take the measure to be $\mu(E) = 0$ if E is finite and $\mu(E) = \infty$ if the set is infinite.

2.4 Lebesgue measure

A crucial measure is the **Lebesgue measure** on $\mathbb{R}$, often denoted λ, where $(\mathbb{R}, \mathbb{B}(\mathbb{R})) = (\Omega, \mathcal{F})$; we want $\mu_0(a,b) = b - a$. We need to check countable, but since $(0,1) \cup (2,3)$ is not an interval, we run into some problems.

[4]

$$\mathcal{F}_0 = \{\text{finite unions } (a_1, b_1] \cup \cdots \cup (a_r, b_r], -\infty \leq a_1 \leq b_1 \leq \cdots \leq a_r \leq b_r \leq \infty\}$$

and we define the pre-measure μ_0 to be equal to $\sum_{i=1}^{r}(b_i - a_i)$ for a set above.

Exercise 2.3

Prove (ii) from the definition of pre-measures, and that the definition of μ_0 doesn't depend on the representation of its argument.

2.5 Carathéodory Extension Theorem

Theorem 2.6 (Carathéodory Extension Theorem)

Let $(\Omega, \mathcal{F})$ be a measurable space and $\mathcal{F}_0 \subset \mathcal{F}$ an **algebra** with $\sigma(\mathcal{F}_0) = \mathcal{F}$. If $\mu_0 : \mathcal{F}_0 \to [0, \infty]$ is a pre-measure on $\mathcal{F}_0$ then $\exists$ a measure $\mu : \mathcal{F} \to [0, \infty]$ on $\mathcal{F}$ that extends μ_0 in the sense that $\mu|_{\mathcal{F}_0} \equiv \mu_0$.

Furthermore, if μ_0 is σ-finite, then μ is the unique (measure) extension of μ_0 to $\mathcal{F}$.

Let us apply Carathéodory Extension Theorem to the Lebesgue measure,

Proposition 2.7

μ_0 is a pre-measure on $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$

Proof μ_0 is well-defined. It is finitely additive; indeed, for two sets of the form $A = (a_1, b_1] \cup \cdots \cup (a_r, b_r]$ and $B = (c_1, d_1] \cup \cdots \cup (c_s, d_s]$ disjoint, their union will be

$$C = (f_1, g_1] \cup (f_2, g_2] \cup \cdots (f_{r+s}, g_{r+s}]$$

since we can use the same intervals; each interval appears exactly once in C, so

$$\begin{aligned}
\mu_0(A \cup B) &= \mu_0(C) \\
&= \sum_{j=1}^{r+s}(g_i - f_i) \\
&= \sum_{i=1}^{r}(b_i - a_i) + \sum_{j=1}^{s}(d_i - c_i) \\
&= \mu_0(A) + \mu_0(B)
\end{aligned}$$

[4] the above notation fiddles to say that it is unbounded to the right or the left, but $a_i, b_i \in \mathbb{R} \; \forall \; i$.

If $A \in \mathcal{F}_0$ and $A = \bigcup_{n\geq 1} A_n$ for A_n disjoint and $A_n \in \mathcal{F}_0$ then $\mu_0(A) = \sum_{n\geq 0} \mu_0(A_n)$.

If A is as previously, then writing $A_{n,i} = A_n \cap (a_i, b_i]$, we see it suffices to consider the case that A is an interval. Replacing each A_n by its component intervals shows we may assume each A_i is an interval.

Exercise 2.4

Show that the above is true when A is unbounded (if $(-\infty, x] = \bigcup_{n=1}^{\infty}(a_n, b_n]$ then $\sum_{n=1}^{\infty}(b_i - a_i) = \infty$).

Now assume A is bounded, say $A = (x, y]$. We replace $A = (x, y]$ by $(0, y - x]$ and $A_n = (x_n, y_n]$ by $(x_n - x, y_n - x]$; thus we can assume that $A = (0, y^*]$ for some y^*. Since the equality

$$\mu_0(A) = \sum_{n\geq 0} \mu_0(A_n)$$

is preserved by dividing by y^*, to replace by $(0, 1]$ and $A_n = (x_n, y_n]$ by $((x_n - x)/y^*, (y_n - x)/y^*]$. It remains to prove that if $A_n, n \geq 1$ are disjoint intervals and $\bigcup_{n\geq 1} A_n = (0, 1]$ then $\sum_{n=1}^{\infty} \mu_0(A_n) = 1$. This sum is a limit of a finite sum. We will replace additivity by monotonicity and take limits of sets.

Let $B_n = \bigcup_{i=1}^{n} A_i$. Then $B_n \uparrow (0, 1]$; so we want to show that $\mu_0(B_n) \uparrow 1$. Let $C_n = (0, 1] \setminus B_n$. Then $C_n \downarrow \emptyset$ and we must show that $\mu_0(C_n) \downarrow 0$. This suffices since $1 = \mu_0((0, 1]) = \mu_0(B_n) + \mu_0(C_n)$ by finite additivity, since each set is in $\mathcal{F}_0$. So $\mu_0(B_n) \to 1 \Leftrightarrow \mu_0(C_n) \downarrow 0$.

Note that if C_n was unbounded and of the form $(-\infty, -n]$, then $\bigcap_{n\geq 1} C_n = \emptyset$ (the limit would be the empty set) while any C_n would be such that $\mu_0(C_n) = \infty$.

We prove the contrapositive: if $C_n \downarrow$ and $\mu_0(C_n) > 2\varepsilon > 0 \; \forall \; n$, then

$$\lim_{n\to\infty} C_n = \bigcap_{n\geq 1} C_n \neq \emptyset.$$

points the

Th idea is to find a sequence $(x_n, n \geq 1)$ for $x_n \in C_n$ such that $x_n \to x \in \bigcap_{n\geq 1} C_n$. For each n, let $D_n \subset C_n \setminus (0, \varepsilon] \equiv C_n'$ with $D_n \in \mathcal{F}_0$, with $\overline{D_n} \subset C_n'$ and $\mu_0(C_n' \setminus D_n) \leq \frac{\varepsilon}{2^{n+1}}$.

Now D_n consisted of open sets, its closure was the union of closed sets, but the intersection is still compact. Finally, let $K_n = \bigcap_{i\le n} \overline{D_i} \subseteq \bigcap_{i \le n} C'_i \subseteq (\varepsilon, 1]$.

Note

$$\begin{aligned}\mu_0\left(\bigcap_{i\le n} D_i\right) &\ge \mu_0(C'_n) - \mu_0\left(\bigcup_{i\le n}(C'_i \setminus D_i)\right)\\ &\ge \mu_0(C'_n) - \sum_{i\le n}\mu_0(C'_i\setminus D_i)\\ &\ge \mu_0(C'_n) - \sum_{i\le n}\frac{\varepsilon}{2^{n+1}}\\ &\ge \mu_0(C'_n) - \frac{\varepsilon}{2}\\ &\ge \frac{\varepsilon}{2}.\end{aligned}$$

Then $K_n \neq 0$ since $K_n \supseteq \bigcap_{i\le n} D_i \neq \emptyset$.

Now pick $x_n \in K_n$ for each $n \in \mathbb{N}$. x_n is a bounded sequence on $(0,1]$; we let $(n_i, i \ge 1)$ be such that x_{n_i} converges as $n_i \to \infty$. Then, for all i, for all $j \ge i$, then $x_{n_j} \in K_{n_j} \subseteq K_{n_i}$ so $(x_{n_j}, j \ge i) \subseteq K_{n_i}$. Since the sequence is in a compact set, the limit belongs to the set. Let $x := \lim_{j\to\infty} x_{n_j} \in K_{n_i}$ since K_{n_i} is compact so $x \in \bigcap_{i\ge n} K_{n_i} = \bigcap_{n\ge 1} K_n \subseteq \bigcap_{n\ge1}\overline{D_n} \subseteq \bigcap_{i\ge 1} C_n$ so $\bigcap_{n\ge1} C_n \neq \emptyset$.

Unraveling the whole procedure, since we have a sequence decreasing to the empty set, its measure must be zero by our contrapositive. Then, shifting and scaling and taking some finite unions allows us to deal with general sets. ■

Proposition 2.8 (π-system lemma)

If $(\Omega, \mathcal{F})$ is a measurable space, $P \in \mathcal{F}$ is a π-system with $\sigma(P) = \mathcal{F}$ and μ_1, μ_2 are σ-finite measures on $(\Omega, \mathcal{F})$ with $\mu_1|_P \equiv \mu_2|_P$, then $\mu_1 \equiv \mu_2$.

Before pursuing with the proof, we state the necessary definition.

Definition 2.9 (π-system)

$P \subset 2^\Omega$ is a **π-system** if $\Omega \in P$ and $E, F \in P \Rightarrow E \cap F \in P$

Proof We provide only a sketch. Establish closure properties of the set where μ_1 and μ_2 agree. If one can show that $\{E \in \mathcal{F} : \mu_1(E) = \mu_2 = E\}$ is a σ-algebra, then we are done since the two measures agree on $\mathcal{F}$, since the set is contained in $\mathcal{F}$ and it is smallest. ■

Example 2.8

The set of intervals $\{(-\infty, x], x \in \mathbb{R}\} \cup \{\mathbb{R}\}$ is a π-system generating $\mathbb{B}(\mathbb{R})$.

Exercise 2.5

On the first assignment, you are asked to construct a probability measure on the space $\Omega = \{0,1\}^{\mathbb{N}}$ where $\mathcal{F} = \sigma(\text{cylinder})$; get an infinite square prism; in $\mathbb{R}^3$, you can get constrained shapes like boxes, but in $\{0,1\}^{\mathbb{N}}$, there are many unconstrained parameter. This is equivalent to $\mathcal{F} = \sigma(\{\{\omega : \omega_i = 1\}, i \in \mathbb{N}\}$. Then, any $\omega \in \Omega$ where $\omega = (\omega_k, k \geq 1)$.

Consider $(\mathbb{R}, \mathbb{B}(\mathbb{R}), \lambda)$ where λ denotes the Lebesgue measure. Let $\mathfrak{C}$ be the Cantor set, which can be viewed as $\bigcap C_i$ where C_i is every step of the construction of the Cantor set. Now $\lambda(C_1) = 1, \lambda(C_2) = \frac{2}{3}$, $\lambda(C_i) = \left(\frac{2}{3}\right)^{i-1}$. Thus, $\lambda(\mathfrak{C}) = \lim_{i\to\infty} \lambda(C_i) = 0$. It can also be shown that $\mathfrak{C}$ is uncountable, yet has measure zero.

Now let D be any subset of $\mathfrak{C}$. What is $\lambda(D)$? If $A, B \in \mathbb{B}(\mathbb{R})$ and $A \subset B$, then $\lambda(B) = \lambda(A) + \lambda(B \setminus A) \geq \lambda(A)$. But maybe $D \notin \mathbb{B}(\mathbb{R})$: to fix this if $(\Omega, \mathcal{F}, \mu)$ is a measure space, let $\mathcal{N} = \{D \subset \Omega : \exists E \in \mathcal{F}, \mu(E) = 0, D \subset E\}$. Extend μ to $\mathcal{F} \cup \mathcal{N}$ by setting $\mu(D) = 0 \;\forall\; D \in \mathcal{N}$.[5]

Exercise 2.6

Show that $\sigma(\mathcal{F} \cup \mathcal{N}) = \{E \cup D; E \in \mathcal{F}, D \in \mathcal{N}\}$. For such $E \cup D$, define a measure on $\sigma(\mathcal{F} \cup \mathcal{N}) \coloneqq \mathcal{F}^*$, termed the completion of $\mathcal{F}$.

We will now restrict our attention to probability spaces and discuss the notion of events. The latter is an element of the σ-algebra.

There are many examples, for example coin tosses with $\Omega = \{\pm 1\}^{\mathbb{N}}$, $\mathcal{F} = \sigma(\text{cylinder})$ and returning to the first lecture, the event $E = \{$random walk returns to 0 infinitely

[5] $\mathcal{N}$ stands for null sets

often}={random walk is recurrent}. Then $\{S_n = 0\}$ is the same as $\sum_{i=1}^n \omega_i = 0$, which means that

$$\{S_n = 0\} = \{\#\{1 \leq j \leq n : \omega_j = 1\} = \#\{1 \leq j \leq n : \omega_j = -1\}\} \in \mathcal{F}$$

and even in $\mathcal{F}_0$. [6]

We can thus write the event

$$\{S_n = 0 \text{ i.o.}\} = \{\, \forall\, n, \exists m \geq n : S_m = 0\} = \bigcap_{n\geq 1} \bigcup_{m \geq n} \{S_n = 0\}$$

and because $\mathcal{F}$ is a σ-algebra and is closed under complements and unions. Thus, what we discussed a while ago is an event, and is measurable. Note that we did not need to specify a measure beforehand to determine this property.

This event is called $\limsup_{n\to\infty}\{S_n = 0\}$.

Definition 2.10 (Limit superior)

If $E_n, n \in \mathbb{N}$ are elements of $\mathcal{F}$, then $\limsup_{n\in\mathbb{N}} E_n$ is defined to be $\bigcap_{n\geq 1}\bigcup_{m\geq n} E_m$ and $\liminf \coloneqq \bigcup_{n\geq 1}\bigcap_{m\geq n} E_m$ (for all, but finitely many).

Exercise 2.7

If $(E_n, n \geq 1)$ and $(E'_n, n \geq 1)$ agree on all but finitely many n, then $\limsup_{n\to\infty} E_n = \limsup_{n\to\infty} E'_n$ and $\liminf_{n\to\infty} E_n = \liminf_{n\to\infty} E'_n$.

Note

$\liminf_{n\to\infty} E_n \subseteq \limsup_{n\to\infty} E_n$ and $\limsup_{n\to\infty} E_n^{\complement} \subseteq \liminf_{n\to\infty} E_n^{\complement}$.

Example 2.9

With $(\Omega, \mathcal{F}) = (\{\pm 1\}^{\mathbb{N}}, \mathcal{F})$ and $\mathsf{P}_{\frac{1}{2}}$, the strong law of large numbers states that $\mathsf{P}_{\frac{1}{2}}\left(\frac{S_n}{n} \to 0\right) = 1$. The questions is to whether $\left\{\omega : \frac{S_n(\omega)}{n} \to 0\right\} \in \mathcal{F}$. The answer is yes: for any $\varepsilon > 0$,

let $A_\varepsilon = \{S_n > \varepsilon n \text{ i.o.}\}^{\complement} \in \mathcal{F}$ and take $B_\varepsilon = \{S_n < -\varepsilon n \text{ i.o.}\}^{\complement} \in \mathcal{F}$. The event that $\left\{\frac{S_n}{n} \to 0\right\} = \bigcap_{k\geq 1} A_{\frac{1}{k}} \cap B_{\frac{1}{k}}$.

[6] By doing so, we have defined a mapping $S_n : \Omega \to \mathbb{R}$.

2.6 Probability bounds: reverse Fatou and first Borel-Cantelli lemma

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a given probability space. Then

Proposition 2.11 (Reverse Fatou)

$$\mathsf{P}\left(\limsup_{n\to\infty} E_n\right) \geq \limsup_{n\to\infty} \mathsf{P}\left(E_n\right)$$

Proof Write $\bigcup_{m\geq n} E_m := G_n$ so that $G_n \downarrow \limsup_{n\to\infty} E_n$. Then

$$\begin{aligned}\mathsf{P}\left(\limsup_{n\to\infty} E_n\right) &= \mathsf{P}\left(\bigcap_{n\geq 1} G_n\right) \\ &= \lim_{n\to\infty} \mathsf{P}\left(G_n\right) \\ &\geq \lim_{n\to\infty} \sup_{m\geq n} \mathsf{P}\left(E_m\right) \\ &= \limsup_{n\to\infty} \mathsf{P}\left(E_n\right)\end{aligned}$$

■

Lemma 2.12 (Borel-Cantelli 1)

If $\sum_{n=1}^{\infty} \mathsf{P}\left(E_n\right) < \infty$, then $\mathsf{P}\left(E_n \text{ i.o.}\right) = 0$

Proof We have

$$\mathsf{P}\left(E_n \text{ i.o.}\right) = \lim_{n\to\infty} \mathsf{P}\left(G_n\right) \leq \lim_{n\to\infty} \sum_{m\geq n} \mathsf{P}\left(E_m\right)$$

by countable additivity and using facts about convergent sequences (exercise), if the sum $\sum_{n=1}^{\infty} \mathsf{P}\left(E_n\right) < \infty$, then this in turns imply that $\sum_{m=n}^{\infty} \mathsf{P}\left(E_m\right) \to 0$ as $n \to \infty$. ■

Note that the converse is false in general.

Example 2.10

Take $(\Omega, \mathcal{F}, \mathsf{P}) = ([0,1], \mathbb{B}([0,1]), \lambda)$ and $E_n = \left[0, \frac{1}{n}\right]$. Then $\sum_{n=1}^{\infty} \mathsf{P}\left(E_n\right) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$ and $\mathsf{P}\left(E_n \text{ i.o.}\right) = \mathsf{P}\left(\{0\}\right) = 0$.

In the previous example, we had $S_n((\omega_i, i \geq 1)) = \sum_{i=1}^n \omega_i$ with $\mathsf{P} = \mathsf{P}_{\frac{1}{2}}$. What if we want to model $Z_n, n \geq 1$ independent $\mathcal{N}(0,1)$ random variables, with now $S_n \coloneqq \sum_{i=1}^n Z_i$. Insofar, what we have built are measures $([0,1], \mathbb{B}([0,1]), \lambda)$ and in the assignment, $(\{0,1\}^{\mathbb{N}}, \sigma(\text{cyl.}), \mathsf{P})$ the product measure on the given product space. Neither of these probability space look obviously useful for Gaussian random variables.

Here is nevertheless an idea to model $Z \sim \mathcal{N}(0,1)$: first decide whether $Z \geq 0$ or $Z < 0$. Make then $\left[0, \frac{1}{2}\right]$ corresponds to $Z < 0$ and $\left[\frac{1}{2}, 1\right]$ if $Z \geq 0$. We can further break the first interval into $\left[0, \frac{1}{4}\right]$ with the the 25% quantile of Z, *i.e.* $t : \mathsf{P}\left(\mathcal{N}(0,1) \leq t\right) = \frac{1}{4}$, and similarly for $\left[\frac{1}{4}, \frac{1}{2}\right]$ with $Z \in [t, 0]$, $\left[\frac{1}{2}, \frac{3}{4}\right] \to Z \in [0, -t]$ and finally $\left[\frac{3}{4}, 1\right] \to Z \geq t$.

The above construction is tedious, but could work more generally if we look at quantiles again, with minor twitches for variables that are not absolutely continuous; we will thus look for a more general procedure so as to not have to deal with these details. The philosophy is to assume that $(\Omega, \mathcal{F}, \mathsf{P})$ is "rich enough" to model all needed random variables. One justification is as follows: let $S_i = \{p_i^k : k \geq 1\}$, p_i is the i^{th} prime. Then for $i \neq j$, $S_i \cap S_j = \emptyset$, and $\bigcup_{i \geq 1} S_i \subset \mathbb{N}$, so we can view $\{0,1\}^{\mathbb{N}}$ as

$$\left(\prod_{i \geq 1} \{0,1\}^{S_i} \right) \times \{0,1\}^{\mathbb{N} \setminus \cup_{i \geq 1} S_i}$$

and we can use the countable copies of the original space with same cardinality, and then we can construct with this any countable collection of random variables.

Chapter 3
Random variables

The random variable $S_n : \Omega \to \mathbb{R}$ is a mapping, which we define now. If we say that a random variable is positive, we refer to the sets of ω_i such that their image is greater than zero. We obviously require the sets to be in our σ-algebra and S_n to be measurable.

Definition 3.1 (Random variable)

A **real** random variable is a function $X : \Omega \to \mathbb{R}$ where $(\Omega, \mathcal{F}, \mathsf{P})$ is a probability space, and X is a **measurable map**.

3.1 Measurable maps

Definition 3.2 (Measurable map)

If $(S, \mathcal{F})$ and $(T, \mathcal{G})$ are measurable spaces, a $\mathcal{F}/\mathcal{G}$ **measurable map** from S to T is a function $X : S \to T$ such that

$$\forall\, G \in \mathcal{G}, X^{-1}(G) \coloneqq \{s \in S : X(s) \in G\} \in \mathcal{F}$$

Remark

The measurability depends on the σ-field. For example, if $\mathcal{G} = \{\emptyset, \Omega\}$, then **every** map is measurable. If $\mathcal{G} = 2^T$, then for X to be measurable, we need that $X^{-1}(A) \in \mathcal{F}$ for all $A \in T$. [7]

For (real) random variables, the σ-algebra on $\mathbb{R}$ is $\mathbb{B}(\mathbb{R})$ unless otherwise specified.[8]

Definition 3.3

An $\mathbb{R}^d$-valued random variable is a measurable map $\boldsymbol{X} : \Omega \to \mathbb{R}^d, \mathcal{F} \to \mathbb{B}(\mathbb{R}^d)$.

Generally, if $(\mathbb{M}, d)$ is a metric space, a **random element of** $\mathbb{M}$ is a map $\mathcal{F}/\mathbb{B}(\mathbb{M}, d)$ real valued measurable map $X : \Omega \to \mathbb{M}$ where $(\Omega, \mathcal{F}, \mathsf{P})$ is a probability space.

[7] An analogy with observables and experience is to consider what measurements we want to make with a given sets of tools. The more precise the measurement (*e.g.* the position of a particle in time with momentum in some interval as opposed to the simple observation about whether a particle is in a box or not), the lower the changes that the mapping be measurable: the "experiment" decides what can be measured).

[8] We want to pick the smallest σ-algebra, so it makes sense to consider the Borel sets rather than the Lebesgue sets, since the latter are more complicated and give rise to a larger σ-algebra.

Theorem 3.4

If $(\Omega, \mathcal{F}), (T, \mathcal{G})$ are measurable spaces and $X : \Omega \to T$ and $\exists \mathcal{A} \subset \mathcal{G}$ such that $\sigma(\mathcal{A}) = \mathcal{G}$ such that $X^{-1}(E) \in \mathcal{F}$ for all $E \in \mathcal{A}$, then X is $\mathcal{F}/\mathcal{G}$- measurable.

Example 3.1

- ◦ If $(T, \mathcal{G}) = (\mathbb{R}, \mathbb{B}(\mathbb{R}))$ and $\mathcal{A} = \{(-\infty, x], x \in \mathbb{R}\}$.
- ◦ The same construction extends to $(T, \mathcal{G}) = (\mathbb{R}^d, \mathbb{B}(\mathbb{R}^d))$ with

$$\mathcal{A} = \left\{ \prod_{i=1}^{d} (-\infty, x_i], (x_1, \ldots, x_d) \in \mathbb{R}^d \right\}$$

- ◦ If $(T, \mathcal{G}) = (\mathbb{M}, \mathbb{B}(\mathbb{M}, d))$ with $\mathcal{A} = \{\text{open sets in } \mathbb{M}\}$.

Corollary 3.5

If $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous, then f is measurable.

Definition 3.6

If $X : \Omega \to T$ where $(T, \mathcal{G})$ is a measurable space, then the σ-algebra **generated** by X is

$$\sigma(X) \coloneqq \bigcap_{\substack{\{\mathcal{F}: X \text{ is} \\ \mathcal{F}/\mathcal{G} \text{ meas.}\}}} \mathcal{F} = \left\{ X^{-1}(G) : G \in \mathcal{G} \right\}$$

Note

If A, B in the sets $\mathcal{G}$, then $A = X^{-1}(G)$ and $B = X^{-1}(H)$ so $A \cup B = X^{-1}(G \cup H)$ and if $G, H \in \mathcal{G}$, then $G \cup J \in \mathcal{G}$ so $A \cup B$ are in here.

Suppose Z is a standard normal random variable. Is Z^2 a random variable? What about Z^{-1}? What about the sums? We next look at compositions.

Proposition 3.7 (Composition of random variables)

If $(R, \mathcal{F}), (S, \mathcal{G}), (T, \mathcal{H})$ are measurable spaces and $X : R \to S$, $Y : S \to T$ measurable, then $Y \circ X$ is measurable.

Proof If $E \in \mathcal{H}$, then $Y^{-1}(E) = F \in \mathcal{G}$ so $(Y \circ X)^{-1}(E) = X^{-1}(Y^{-1}(E)) = X^{-1}(F) \in \mathcal{F}$.

■

Corollary 3.8

If $X : \Omega \to \mathbb{R}^n$ is an $\mathbb{R}^n$-valued random variable and $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous, then $f(X)$ is an $\mathbb{R}^m$ valued random variable.

Corollary 3.9

If X is a random variable and $f : \mathbb{R} \to \mathbb{R}$ is **increasing**, then $f(X)$ is a random variable.

Proof

It suffices to show that increasing functions are measurable. If f is increasing, then $f^{-1}((-\infty, x])$ is also an interval so measurable. ■

Theorem 3.10

If $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ are random variables and $f : \mathbb{R}^n \to \mathbb{R}$ is measurable, then $f(X_1, \ldots, X_n) : \Omega \to \mathbb{R}$ is a random variable.

Proof We need to show that the map $(X_1, \ldots X_n) : \Omega \to \mathbb{R}^n, \omega \mapsto (X_1(\omega), \ldots, X_n(\omega))$ is measurable. It suffices to show that the sets formed by the rays of the form

$$\{\omega \in \Omega : X_i(\omega) \leq x_i \text{ for } 1 \leq i \leq n\}$$

are measurable $\forall\ (x_1, \ldots, x_n) \in \mathbb{R}^n$. This is just $\bigcap_{i=1}^n X_i^{-1}\left((-\infty, x_i]\right) \in \mathcal{F}$ since we take finite intersection of elements $X_i^{-1}\left((-\infty, x_i]\right) \in \mathcal{F}$ since X_i is a random variable. ■

We now tackle the issue of limits.

Theorem 3.11

If $(X_n, n \in \mathbb{N})$ random variables $\Omega \to \mathbb{R}$ then $\inf_{n\geq 1} X_n$, $\sup_{n\geq 1} X_n$, $\liminf_{n\geq 1} X_n$ and $\limsup_{n\geq 1} X_n$ are all random variables.

Proof inf and sup are increasing maps, so measurable (infinite sequence here), but

$$\{\inf X_n < a\} = \bigcup_{n\in\mathbb{N}} \{X_n < a\}$$

is measurable since it is a countable union. Next, by the definitions of lim inf and lim sup, we have

$$\limsup_{n\to\infty} X_n = \inf_{n\geq 1} \sup_{m\geq n} X_m$$

is a random variable using the closure properties. Analogously,

$$\liminf_{n\to\infty} X_n = \sup_{n\geq 1} \inf_{m\geq n} X_m$$

is also a random variable. ■

3.2 Limits

We have seen $\limsup X_n$, $\liminf X_n$ were random variables if $\{X_n\}$ were sequences of random variables. For the statement about the

Theorem 3.12 (Strong law of large numbers)

If $(X_n, n \geq 1)$ is a sequence of mutually independent identically distributed random variables with $\mathsf{E}\left(|X_1|\right) < \infty$, then almost surely, [9]

$$\lim_{n\to\infty} \left(\frac{\sum_{i=1}^n X_i}{n}\right) = \mathsf{E}\left(X_1\right)$$

where **almost surely** means

$$\mathsf{P}\left(\lim_{n\to\infty}\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \mathsf{E}\left(X_1\right)\right) = \mathsf{P}\left(\left\{\omega : \lim_{n\to\infty}\left(\frac{\sum_{i=1}^n X_i(\omega)}{n}\right) = \mathsf{E}\left(X_i\right)\right\}\right) = 1$$

Proposition 3.13 (Measurability of limits)

If $\bar{S}_n = \sum_{i=1}^n X_i/n$, since $\limsup_{n\to\infty} \bar{S}_n, \liminf_{n\to\infty} \bar{S}_n$ are random variables, the event

$$\begin{aligned}\{\lim_{n\to\infty} \bar{S}_n \text{ exist}\} = \{&\liminf_{n\to\infty} \bar{S}_n = \infty\} \cup \{\limsup_{n\to\infty} \bar{S}_n = -\infty\} \cup \\ &\left(\left\{\limsup_{n\to\infty} \bar{S}_n - \liminf_{n\to\infty} \bar{S}_n = 0\right\} \cap \left\{\text{either } \limsup_{n\to\infty} \bar{S}_n < \infty, \text{ or } \liminf_{n\to\infty} \bar{S}_n > -\infty\right\}\right)\end{aligned}$$

[9] That is, we have convergence pointwise to a constant function.

Note that the set

$$C = \left(\left(\limsup_{n \to \infty} \bar{S}_n - \liminf_{n \to \infty} \bar{S}_n \right)^{-1} (0) \right) : \Omega \to \mathbb{R} \cup \{-\infty, \infty\}$$

In other words, C has the form $Z^{-1}(0)$ for some random variable Z. Since $X^{-1}(B)$ is measurable for all Borel sets B and all random variables X, it follows that $C = Z^{-1}(0)$ is measurable.

This fiddling around was solely to point out that we have measurable limits even with $\pm\infty$.

Chapter 4

Distribution functions and laws

4.1 Distribution functions

Fix a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ and $X : \Omega \to \mathbb{R}$ a random variable. Then, the **cumulative distribution function** F_X is

$$F_X(x) = \mathsf{P}\left(X \leq x\right).$$

Here are some facts about CDF:

Proposition 4.1

1. F_X is non-decreasing
2. $\lim_{x \to -\infty} F_X(x) = 0, \lim_{x\to\infty} F_x(x) = 1$: use unions/intersections of events to build monotone increasing/decreasing sequences.
3. The CDF is càdlàg: $\forall\, x, \lim_{y \downarrow x} F(y) = F(x)$

Proof We prove the last statement. Write

$$\begin{aligned}\mathsf{P}\left(X \leq x\right) &= \mathsf{P}\left(\bigcap_{n\geq 1} X \leq x + \frac{1}{n}\right)\\ &= \lim_{n\to\infty} \mathsf{P}\left(X \leq x + \frac{1}{n}\right)\\ &= \lim_{n\to\infty} F_X\left(x + \frac{1}{n}\right)\end{aligned}$$

since the limit exist and the function is decreasing, taking any subsequence yield the value of the limit.

Exercise 4.1

The limit does not depend on the sequence: $\forall\, \{y_n, n \geq 1\}$ with $y_n \downarrow x, \lim_{n\to\infty} F_X(y_n) = \lim_{n\to\infty} F_X\left(x + \frac{1}{n}\right)$.

The limit may not take left-value; if $X \equiv 0, F_X(x) = \mathbf{1}_{x\leq 0}$. A function that is right-continuous with left limits is called càdlàg. ■

A function satisfying properties 1.-3. is called a CDF.

Definition 4.2 (Law)

The **law** of X (or the distribution of X) is the function $\mu_X : \mathbb{B}(\mathbb{R}) \to [0,1]$ defined by

$$\mu_X(B) = \mathsf{P}\left(X \in B\right) = \mathsf{P}\left(X^{-1}(B)\right) = \mathsf{P}\left(\omega : X(\omega) \in B\right)$$

Proposition 4.3

μ_X is a probability measure on $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$.

Proof We need to check countable additivity and non-negativity; definitely $\mu_X \geq 0$ and $\mu_X(\mathbb{R}) = \mathsf{P}\left(X \in \mathbb{R}\right) = \mathsf{P}\left(\Omega\right) = 1$. For countable additivity, if we have disjoint collection B_n Borel sets, $n \geq 1$, the pre-image $E_n = X^{-1}(B_n)$ must also be disjoint elements of $\mathcal{F}$. Thus, it follows

$$\begin{aligned} \mu_X\left(\bigcup_{n\geq 1} B_n\right) &= \mathsf{P}\left(X^{-1}\left(\bigcup_{n\geq 1} B_n\right)\right) \\ &= \mathsf{P}\left(\bigcup_{n\geq 1} E_n\right) \\ &= \sum_{n\geq 1} \mathsf{P}\left(E_n\right) \\ &= \sum_{n\geq 1} \mathsf{P}\left(X^{-1}(B_n)\right) \\ &= \sum_{n\geq 1} \mu_X(B_n). \end{aligned}$$

since P is a measure. ■

Now, let $Y : \mathbb{R} \to \mathbb{R}$ be the identity function, $Y(r) = r$. Then, for all $B \in \mathbb{B}(\mathbb{R})$ is

$$\mu_X(Y \in B) = \mu_X(\{r : Y(r) \in B\}) = \mu_X(B) = \mathsf{P}\left(X \in B\right)$$

This is a first example of what is called a **change of variable** . Note that μ_X determines F_X since $F_X(x) = \mathsf{P}\left(X \leq x\right) = \mu_X((-\infty, x])$. Since the half-open intervals generates a π-system, conversely the preceding equality and the π-system lemma together imply that μ_X is uniquely determined by F_X.

4.2 Skorokhod construction

Proposition 4.4 (Skorokhod construction)

Any CDF is the CDF of some random variable.

Proof We use the space $([0,1], \mathbb{B}([0,1]), \lambda)$. We want a function $Z : [0,1] \to \mathbb{R}$ with CDF F. The idea is that if $F(x) = \omega$, then set $Z(\omega) = x$ (we are in a sense flipping the graph and interchanging the domain and the range). Then

$$\mathsf{P}\left(Z \leq z\right) = \mathsf{P}\left(\{\omega : Z(\omega \leq z\}\right) = \mathsf{P}\left([0, F(z)]\right) = \lambda([0, F(z)]) = F(z).$$

this boils down to $F^{-1}((-\infty, z]) = [0, F(z)]]$. The latter holds if F^{-1} exists and is continuous. We could have flat sections: to deal with this, take rationals in these parts (using density of $\mathbb{Q}$ in $[0,1]$) – since we map the flats to rationals, there are at most countably many. Because rationals have measure zero, both have measure zero under corresponding measures.

There is a problem with the above: maybe F is not invertible, *i.e.* not strictly increasing. The solution is to let

$$X^{-1}(\omega) = \sup\{x : F(x) < \omega\}$$
$$X^{+}(\omega) = \sup\{x : F(x) \leq \omega\}.$$

Exercise 4.2

$\mathsf{P}\left(X^+ \neq X^-\right) = \lambda\left(\{\omega : X^+(\omega) \neq X^{-1}(\omega)\}\right) = 0$. Finally, for all $z, \mathsf{P}\left(X \leq z\right) = \lambda([0, F(z)])$ for all z, while

$$\mathsf{P}\left(X^+ \leq x\right) = \lambda([0, F(z)])\mathbf{1}_{F \text{ increasing at } z} + \lambda([0, F(z)))\mathbf{1}_{F(z+\varepsilon)=F(z),\, \varepsilon>0}.$$

Thus, X^+ and X^- both have law F

Exercise 4.3

If $X, Y : \Omega \to \mathbb{R}$ are random variables, $\mathsf{P}\left(X \leq Y\right) = 1$ and X, Y have the same CDF, then $X \overset{\text{a.s.}}{=} Y$, *i.e.* $\mathsf{P}\left(X = Y\right) = 1$.

■

Last class, we show that for any CDF F, there is a random variable X with $F_X \equiv F$.

Another way to see this is to take $(\Omega, \mathcal{F}) = (\mathbb{R}, \mathbb{B}(\mathbb{R}))$, we can take a simple function such as the identity $X : \mathbb{R} \to \mathbb{R}$ and instead work what the measure should be. Reverse-engineering the measure as to have (set) $\mathsf{P}_0 (X \leq x) = \mathsf{P}_0 ((-\infty, x]) = F(x)$; this determines for us what the probability should be on a π-system. We can extend this uniquely to a probability measure on $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$; then X has CDF F_X. This involves checking that P_0 was extended to an additive function and show the latter is a pre-measure, the same argument as in the assignment or as when we constructed the Lebesgue measure.

We will now tackle the monotone class theorem, termed sometimes the standard machine.

4.3 Monotone class theorem

Theorem 4.5 (Monotone class theorem)

Let Ω be a set, $\mathcal{S}$ a π-system on Ω and $\mathcal{H}$ a real vector space of functions $f : \Omega \to \mathbb{R}$.[10] If

i) $\mathbf{1}_A \in \mathcal{H}$ for all $A \in \mathcal{S} \cup \{\Omega\}$, where $\mathbf{1}_A(\omega) = 1$ if $\omega \in A$, and 0 otherwise.
ii) If $f_n \in \mathcal{H}, n \geq 1$ and $f_n \uparrow f$ pointwise, and f bounded, then $f \in \mathcal{H}$.

Then $\mathcal{H}$ contains all $\sigma(\mathcal{S}) - \mathbb{B}(\mathbb{R})$ measurable bounded functions.

We start with a

Definition 4.6

A collection $\mathcal{D}$ of subsets of Ω is a d-system (Dynkin or difference) if

i) $\Omega \in \mathcal{D}$
ii) if $A, B \in \mathcal{D}$ and $A \subset B$, then $B \setminus A \in \mathcal{D}$
iii) if $A_n \in \mathcal{D}, n \geq 1$ are increasing, then $\bigcup_{n \geq 1} A_n \in \mathcal{D}$.

Fact 4.7

- A σ-algebra is both a π-system and a d-system.
- If $\mathcal{D}$ is a π-system and a d-system, then it is a σ-algebra.

[10] We will show that all bounded measurable functions are in some set, measurable with respect to $\sigma(\mathcal{S})$. We assume closure in the preamble, that is multiplication by constant and addition of functions. Starting from indicator function, we can create step functions, then simple functions and construct any continuous function by approximations from below. The hard part of the proof is showing that $\mathbf{1}_A \in \mathcal{S}$ implies that $\mathbf{1}_A$ for all $A \in \sigma(\mathcal{S})$

- If $\mathcal{S}$ is a π-system, $\mathcal{D} \supset \mathcal{S}$ and $\mathcal{D}$ is a difference system, then

$$\mathcal{D} \supset d(\mathcal{S}) \coloneqq \bigcap_{\substack{\mathcal{G},\ d\text{-system} \\ \text{containing } \mathcal{S}}} \mathcal{G}.$$

Theorem 4.8

If $\mathcal{S}$ is a π-system over Ω, then $d(\mathcal{S}) = \sigma(\mathcal{S})$.

Proof First, we show $d(\mathcal{S}) \subseteq \sigma(\mathcal{S})$: since $\sigma(\mathcal{S})$ is a d-system, we have by definition the first part. To show that $\sigma(\mathcal{S}) \subseteq d(\mathcal{S})$, let

$$\mathcal{D}_1 = \{B \in d(\mathcal{S}) : A \cap B \in d(\mathcal{S})\ \forall\ A \in \mathcal{S}\}.$$

Note that $\mathcal{D}_1 \subset d(\mathcal{S})$ by definition. If $\mathcal{D}_1$ is a d-system, then $\mathcal{D}_1 = d(\mathcal{S})$. To check that $\mathcal{D}_1$ is a d-system, we need to check

i) $\Omega \in \mathcal{D}_1$ is obvious

iii) If $B_n \in \mathcal{D}_1, n \geq 1$ and $B_n \uparrow B$, then $\forall\ A \in \mathcal{S}$, $A \cap B_n \in d(\mathcal{S})$. Since $d(\mathcal{S})$ is a d-system, $A \cap B = \lim_{n\to\infty} A \cap B_n \in d(\mathcal{S})$ so $B \in \mathcal{D}_1$.

ii) If $B, C \in \mathcal{D}_1$, $B \subseteq C$, then for all $A \in \mathcal{S}$, $A \cap B \in d(\mathcal{S})$ and $A \cap C \in d(\mathcal{S})$. Since $d(\mathcal{S})$ is closed under differences, $(A \cap C) \setminus (A \cap B) = (A \cap (C \setminus B)) \in d(\mathcal{S})$, so $C \setminus B \in d(\mathcal{S})$ and therefore $C \setminus B \in \mathcal{D}_1$.

Now, let $\mathcal{D}_2 = \{C \in d(\mathcal{S}) : A \cap C \in d(\mathcal{S})$ for all $A \in d(\mathcal{S})\}$. Since $\mathcal{D}_1 = d(\mathcal{S})$, we have $S \subset \mathcal{D}_2$. Thus, **if** $\mathcal{D}_2$ is a d-system, then $d(\mathcal{S}) \subseteq \mathcal{D}_2$ so $d(\mathcal{S}) = \mathcal{D}_2$ so that $d(\mathcal{S})$ is closed under intersection, so $d(\mathcal{S})$ is a π-system, hence a σ-algebra and $d(\mathcal{S}) \supset \sigma(\mathcal{S})$.

■

Exercise 4.4

Mimic the proof that $\mathcal{D}_1$ is a d-system to show that $\mathcal{D}_2$ is a d-system.

Example 4.1

A set Ω and $A, B \in \Omega$ such that $A \cap B \neq \emptyset$. Then $\mathcal{D} = \{A, B, \Omega, \emptyset, A^{\mathsf{C}}, B^{\mathsf{C}}\}$. This however is not closed under intersection: you can't get $A \cap B$.

Proof The first part is as follows: let $\mathcal{D} = \{E \in \Omega : \mathbf{1}_E \in \mathcal{H}\}$. Then $\mathcal{S} \subset \mathcal{D}$ by i) and $\mathcal{D}$ is a d-system ($\Omega \in \mathcal{H}$ by i), $\mathcal{H}$ closed under increasing limits by ii) and for closure under differences, because if $A \subset B$, then $\mathbf{1}_{B \setminus A} = \mathbf{1}_B - \mathbf{1}_A$ and $\mathcal{H}$ is a vector space).

Thus, $\sigma(\mathcal{S}) = d(\mathcal{S}) \subset \mathcal{D}$. So $\mathcal{H}$ contains $\mathbf{1}_E$ for all $E \in \sigma(\mathcal{S})$. Since $\mathcal{H}$ is a vector space, it follows that $\mathscr{H}$ contains all **simple** $\sigma(\mathcal{S}) - \mathbb{B}(\mathbb{R})$ measurable functions.

Finally, for any bounded $\sigma(\mathcal{S})/\mathbb{B}(\mathbb{R})$ measurable function, f is an increasing limit of simple functions, so $f \in \mathcal{H}$. ■

How to write f as an increasing limit of simple functions? The Lebesgue approximation is to take $0 \cdot \mathbf{1}_{f \in [0, 2^{-n}]}$, slicing horizontally. Similarly, we continue the procedure with $2^{-n}\mathbf{1}_{f \in [2^{-n}, 2 \cdot 2^{-n})}$.

Suppose f is nonnegative and bounded. Let

$$f_n = \sum_{i=0}^{n2^n - 1} \frac{i}{2^n} \mathbf{1}_{f \in \left[\frac{i}{2^n}, \frac{i+1}{2^n}\right)}.$$

Each approximation is such that $f_n \to f$. Either add a constant to make f nonnegative if it is bounded, but not nonnegative. Or, we can break $f = f^+ - f^-$ where $f^+ = \max\{f, 0\}$ and $f^{-1} = -\min\{f, 0\}$ and then $f^+, f^- \in \mathcal{H}$ and so $f^+ - f^- \in \mathcal{H}$.

Read section 3.13 in Chapter 3.

Chapter 5
Independence

5.1 Independence

Definition 5.1 (Independence for events)

A heuristic definition is to say that events A, are independent if $\mathsf{P}(A \cap B) = \mathsf{P}(A)\,\mathsf{P}(B)$. [11] The independence property depends on the measure.

Example 5.1

If $\Omega = \{0,1\}^{\mathbb{N}}$ and $\mathcal{F} = \sigma(\text{cylinders})$. If $A = \{\omega_1 = 1\}, B = \{\omega_2 = 1\}$ then A, B are independent under $\mathsf{P}_{\frac{1}{2}}$: indeed $\mathsf{P}_{\frac{1}{2}}(A \cap B) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathsf{P}_{\frac{1}{2}}(A)\,\mathsf{P}_{\frac{1}{2}}(B)$. But if $\mathbb{Q}$ is a probability measure with $\mathbb{Q}(\{\bar{1}\}) = \frac{1}{2} = \mathbb{Q}(\{\bar{0}\})$, then A and B are **not** $\mathbb{Q}$-independent because

$$\mathbb{Q}(A \cap B) = \mathbb{Q}(\{\bar{1}\}) = \frac{1}{2}, \qquad \mathbb{Q}(A) = \frac{1}{2} = \mathbb{Q}(B).$$

Example 5.2

Suppose $U \overset{d}{=} \mathcal{U}[0,1]$, whose law is the Lebesgue measure on $[0,1]$. Let $A =$\{exactly 5 zeros before first one in binary expansion of U\} and let $B = \{$2 ones before second stretch of zeros$\}$.

Then $\mathsf{P}(A) = \frac{1}{64}$, since the number must lie between $\frac{1}{32}$ and $\frac{1}{64}$. We have for B sequences that are either $0.110\ldots$, or $0.0110\ldots$, which are intervals of length $\frac{1}{8}, \frac{1}{16}, \frac{1}{32}$, since the first one is guaranteed to happen. Thus, adding the terms gives altogether $\mathsf{P}(B) = \frac{1}{4}$. We have $\mathsf{P}(A \cap B) = \frac{1}{256}$,

since we specify 8 bits, that is our binary expansion is of the form $0.00000110\ldots$. Since "5" and "2" may be replaced with any two natural numbers and the same conclusion holds.

Let $X = \#\{$zeros before the first one$\}$, $Y = \#\{$ones before second stretch of zeros$\}$. There are numbers in $[0,1]$ that do not have a unique expansion are dyadic expansions of the form $k/2^n$. Set arbitrarily for such numbers and

[11] This is linked to product spaces, and we will come back to this with Fubini later in the course. We will from now on often be implicit about $(\Omega, \mathcal{F}, \mathsf{P})$

Exercise 5.1

Show that X and Y defined above are measurable.

We saw that $\mathsf{P}(X=5, Y=2) = \mathsf{P}(X=5)\,\mathsf{P}(Y=2)$ and this holds with 5 and 2 replaced by any natural numbers.

Definition 5.2 (Independence)

We say random variables $X, Y : \Omega \to \mathbb{R}$ are independent if $\forall\, E, F \in \mathbb{B}(\mathbb{R})$,

$$\mathsf{P}(X \in E, Y \in F) = \mathsf{P}(X \in E)\,\mathsf{P}(Y \in F)\,.$$

Then X, Y, as define before, are indeed independent.

Remark

Events A, B are independent if and only if $\mathbf{1}_A, \mathbf{1}_B$ are independent, and similarly $A^{\complement}, B$ are independent.

Exercise 5.2

Prove the above remark.

Example 5.3

Let $X \sim \mathcal{B}\left(\frac{1}{2}\right)$, $Y \sim \mathcal{B}\left(\frac{1}{2}\right)$ Bernoulli random variable, where $\sim$ is defined as $\overset{d}{=}$. Here X, Y are independent and we let $Z = (X+Y) \mod 2$. Thus $Z = 0$ if the sum is even, and $Z = 1$ if the sum is odd. Then X, Y are independent,

X, Z are independent and Y, Z are independent. But $\mathsf{P}(X=1, Y=1, Z=1) = 0 \neq \mathsf{P}(Z=1)\,\mathsf{P}(X=1)\,\mathsf{P}(Y=1) = \frac{1}{8}$. Thus, X, Y, Z are not mutually independent.

Definition 5.3 (Mutual independence)

If $(X_i, i \in I)$ are random variables where $X_i : \Omega \to \mathbb{R}$. We say $(X_i, i \in I)$ are **mutually independent** if $\forall\, \{i_1, \ldots, i_n\} \subseteq I$ and all $B_1, \ldots, B_n \in \mathbb{B}(\mathbb{R})$, [12]

$$\mathsf{P}\left(\bigcap_{k=1}^{n} \{X_{i_k}\}\right) = \prod_{k=1}^{n} \mathsf{P}(X_{i_k} \in B_k)$$

[12] From which we can get get any pairs, or smaller for finite number of variables, since we can take $B_i = \mathbb{R}$ so that $\{X_{i_k} \in \mathbb{R}\} = \Omega$, and multiply by 1 on the right.

Definition 5.4 (Independence (σ-algebra))

Given $\mathcal{G}, \mathcal{H}$ sub σ-algebra of $\mathcal{F}$, say $\mathcal{G}, \mathcal{H}$ are independent if $\forall\ A \in \mathcal{G},\ \forall\ B \in \mathcal{H}$, A and B are independent. We define **mutual independence** for σ-algebras accordingly.

Example 5.4

A, B are independent if and only if $\{\emptyset, A, A^{\complement}, \Omega\} = \sigma(\mathbf{1}_A)$ and $\{\emptyset, B, B^{\complement}, \Omega\} = \sigma(\mathbf{1}_B)$ are independent.

Recall that if $X : \Omega \to \mathbb{R}$ is a random variable, then

$$\sigma(X) = \bigcap_{\substack{\mathcal{F}: X \text{ is } \mathcal{F}-\mathbb{B}(\mathbb{R}) \\ \text{measurable}}} \mathcal{F}$$

is the σ-algebra generated by X. More generally, $\sigma((X_i, i \in I))$ is the smallest σ-algebra that makes **all** of the X_i measurable.

For a single event, we have $\sigma(X) = \{X^{-1}(B), B \in \mathbb{B}(\mathbb{R})\}$, [13] and

$$\sigma((X_i, i \in I)) = \sigma\left(\bigcup_{i \in I} \left\{X^{-1}(B) : B \in \mathbb{B}(\mathbb{R})\right\}\right).$$

Exercise 5.3

Check this statement and find σ-algebras $(\mathcal{F}_n, n \geq 1)$ such that their union $\bigcup_{n \geq 1} \mathcal{F}_n$ is not a σ-algebra.

Proposition 5.5

$(X_i, i \in I)$ are independent if and only if $(\sigma(X_i), i \in I)$ are independent.

Proof ($\Rightarrow$) Fix $\{i_1, \ldots, i_n\} \subset I$ and events $E_1, \ldots, E_n$ with $E_k \in \sigma(X_{i_k})$. Then, for $1 \leq k \leq n$, there is $B_k \in \mathbb{B}(\mathbb{R})$ such that $E_k = X_{i_k}^{-1}(B_k)$ so

$$\begin{aligned} \mathsf{P}\left(\bigcap_{k=1}^{n} E_k\right) &= \mathsf{P}\left(\bigcap_{k=1}^{n} X_{i_k}^{-1}(B_k)\right) \\ &= \mathsf{P}\left(\bigcap_{k=1}^{n} \{X_{i_k} \in B_k\}\right) \end{aligned}$$

[13] While intersections of σ-algebras are themselves σ-algebra, the same may not hold for unions

$$
\begin{aligned}
&= \prod_{k=1}^{n} \mathsf{P}\left(X_{i_k} = B_k\right) && \text{(independence)} \\
&= \prod_{k=1}^{n} \mathsf{P}\left(E_k\right)
\end{aligned}
$$

so $(\sigma(X_i), i \in I)$ are independent. The other direction $(\Leftarrow)$ is similar. ■

Lemma 5.6 (π-system)

If $\mathcal{G}, \mathcal{H}$ are sub σ-algebras of $\mathcal{F}$, $\mathcal{I}, \mathcal{J}$ are π-systems where $\mathcal{I} \in \mathcal{G}$ and $\mathcal{J} \in \mathcal{H}$ such that $\sigma(\mathcal{I}) = \mathcal{G}$, $\sigma(\mathcal{J}) = \mathcal{H}$ then if $\mathcal{I}, \mathcal{J}$ are independent, then $\mathcal{G}, \mathcal{H}$ are independent. Here $\mathcal{I}, \mathcal{J}$ are independent, means that $\forall\ A \in \mathcal{I}, B \in \mathcal{J}$, we have $\mathsf{P}\left(A \cap B\right) = \mathsf{P}\left(A\right) \mathsf{P}\left(B\right)$.

Proof Fix some $A \subset \mathcal{I}$ and define measures μ_A, ν_A on $(\Omega, \mathcal{H})$ by setting $\mu_A(B) = \mathsf{P}\left(A \cap B\right)$ and $\nu_A(B) = \mathsf{P}\left(A\right) \mathsf{P}\left(B\right)$. These are measures, yet not probability measure.

Exercise 5.4

The above defined μ_A, ν_A are measures, but are not probability measures.

Note

μ_A, ν_A have the same total measure, *i.e.* $\nu_A(B) = \mu_A(B)\ \forall\ B \in \mathcal{J}$, so $\nu_A \equiv \mu_A$ on $\mathcal{J}$, so $\mu_A = \nu_A$ on $\sigma(\mathcal{J}) = \mathcal{H}$, so $\mathsf{P}\left(A \cap B\right) = \mathsf{P}\left(A\right) \mathsf{P}\left(B\right)$ for all $A \in \mathcal{I}, B \in \mathcal{H}$.

Now fix $B \in \mathcal{H}$, let $\widehat{\mu}_B = \mathsf{P}\left(A \cap B\right)$, $\widehat{\nu}_B(A) = \mathsf{P}\left(A\right) \mathsf{P}\left(B\right)$ for $A \in \mathcal{G}$. Again, these are measures and agree on $\mathcal{I}$ and therefore on $\sigma(\mathcal{I}) = \mathcal{G}$, so $\mathsf{P}\left(A \cap B\right) = \mathsf{P}\left(A\right) \mathsf{P}\left(B\right)$ for all $A \in \mathcal{G}, B \in \mathcal{H}$. This is analogous to the monotone class theorem. This is precisely what it means for $\mathcal{G}, \mathcal{H}$ to be independent.

■

Exercise 5.5

Show that for integer-valued random variables $X : \Omega \to \mathbb{R}$, $Y : \Omega \to \mathbb{R}$,

if $\mathsf{P}\left(X = k, Y = l\right) = \mathsf{P}\left(X = k\right) \mathsf{P}\left(Y = l\right)$ for all $k, l \in \mathbb{Z}$, then X, Y are independent.

Exercise 5.6

If $X, Y : \Omega \to \mathbb{R}$ are random variables, then X, Y are independent (denoted $X \perp\!\!\!\perp Y$) if and only if $\{\{X \leq x\}, x \in \mathbb{R}\} \perp\!\!\!\perp \{\{Y \leq x\}, x \in \mathbb{R}\}$. Hint: use π-system lemma.

Exercise 5.7

If $\{X_{ij}\}_{1\leq i\leq n, 1\leq j\leq m}$ mutually independent and $f_i : \mathbb{R}^m \to \mathbb{R}$ are Borel functions, then each $(f_i(X_{i1}, \ldots X_{i,m}), 1 \leq i \leq n)$ are mutually independent.

The latter could be used for high-dimensional data with dimension reduction, for use in testing independence, say.

Exercise 5.8

If $\mathcal{F}_i$, and $\mathcal{G}_i, i \geq 1$ are σ-algebras over Ω and each $\mathcal{F}_i$ are independent of $\sigma\left(\bigcup_{j\geq 1} \mathcal{G}_j\right)$. Then

$$\sigma\bigg(\bigcup_{j\geq 1} \mathcal{F}_j\bigg) \perp\!\!\!\perp \sigma\bigg(\bigcup_{j\geq 1} \mathcal{G}_j\bigg).$$

You may think of the above as closure properties.

Exercise 5.9

If X_1 is independent of $\sigma(\{X_2, \ldots, X_n\})$, then $X_1 \perp\!\!\!\perp X_2 \cdot X_3 \cdots X_n$.

Remark

For any sequence $(F_i, i \geq 1)$ of CDFs, there exists a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ and random variables $(X_i, i \geq 1)$ on this space such that $(X_i, i \geq 1)$ are mutually independent and F_i is the CDF of X_i. If $F_i \sim \mathcal{U}(0,1)$ CDF for all i, this is from earlier in the term $(\Omega, \mathcal{F}, \mathsf{P}) = ([0,1], \mathbb{B}([0,1]), \lambda)$ and let $X_1(\omega) = 0.\omega_2\omega_{2^2}\omega_{2^3}\omega_{2^4}\ldots$, taking the unique binary expansion. The bits are independent $\mathcal{B}(\frac{1}{2})$ and so $X_1(\omega) \sim \mathcal{U}[0,1]$. For other variables X_i, using powers of a different powers of the i^{th} prime p_i, that is $0.\omega_{p_i}\omega_{p_i^2}\omega_{p_i^3}\ldots$.

Exercise 5.10

Let $B_i : \Omega \to [0,1], B_i(\omega) = \omega_i$. Then $(B_i, i \geq 1)$ are independent Bernoulli $\mathcal{B}(\frac{1}{2})$. Show from this that the $(X_i, i \geq 1)$ are mutually independent $\mathcal{U}[0,1]$.

Let $Y_i = F_i^{-1}(X_i)$, which is a composition of measurable map; then $(Y_i, i \geq 1)$ are independent and Y_i has CDF F_i. This is the Skorokhod construction.

Recall the branching process example from the second lecture. We defined Z_i as the number of individuals in the i^{th} generation. Each individual has a random number of children with CDF F, independently of others. To formally define Z, let (B_i, j)

be independently defined with CDF F. Let $Z_0 = 1$ and $Z_{n+1} = \sum_{k=1}^{Z_i} B_{n+1,k} = \sum_{k=1}^{\infty} B_{n+1,k}\mathbf{1}_{k\leq Z_n}$. By induction, one can see that Z_{n+1} is a random variable, since Z_0 is a deterministic map, measurable with respect to any σ-algebra. Since all random variables are positive, and we have closure under increasing sequence and limits, Z_{n+1} is also a random variable. Another construction is to take $Z_{n+1} = g(Z_n, U_{n+1})$ for suitable $g, (U_i, i \geq 1)$ independent uniforms $\mathcal{U}[0,1]$. For each given step, the next generation is just a finite computation and assign a portion of the interval for this. All probabilities sum to one and then drawing a uniform gives the size of the next generation. It has the advantage of having less information than the previous one. This is the general construction of a Markov chain (discrete Markov chain of a countable state space).

5.2 Second Borel-Cantelli lemma

Recall the first Borel-Cantelli lemma, that stated $\sum_{n=1}^{\infty} \mathsf{P}(B_i) < \infty$, then we had $\mathsf{P}(\limsup_{n\to\infty} B_i) = 0$. We give the second lemma

Lemma 5.7 (Borel-Cantelli II)

Let $(E_n, n \geq 1)$ be mutually independent events. If $\sum_{n\geq 1} \mathsf{P}(E_n) = \infty$, then $\mathsf{P}(E_n \text{ i.o.}) = 1$.

Proof Compute the bound $\mathsf{P}\left((E_n \text{ i.o.})^{\complement}\right)$. This is

$$\begin{aligned}
&\mathsf{P}(\exists n,\ \forall\, m \geq n, E_m \text{ does not occur})\\
&\quad= \mathsf{P}\left(\bigcup_{n\geq 1}\bigcap_{m\geq n} E_m^{\complement}\right)\\
&\quad\leq \sum_{n\geq 1} \mathsf{P}\left(\bigcap_{m\geq n} E_m^{\complement}\right) && \text{(subadditivity)}
\end{aligned}$$

and

$$\begin{aligned}
\mathsf{P}\left(\bigcap_{m\geq n} E_m^{\complement}\right) &= \lim_{M\to\infty} \mathsf{P}\left(\bigcap_{n\leq m\leq M} E_m^{\complement}\right) && \text{(monotonicity)}\\
&= \lim_{M\to\infty} \prod_{m=n}^{M} \mathsf{P}\left(E_m^{\complement}\right)\\
&= \prod_{m=n}^{\infty} \mathsf{P}\left(E_m^{\complement}\right)
\end{aligned}$$

$$\begin{aligned}
&= \prod_{m=n}^{\infty} \left(1 - \mathsf{P}\left(E_m\right)\right) \\
&\leq \prod_{m=n}^{\infty} e^{-\mathsf{P}(E_m)} && (\text{ since } 1 - x \leq e^{-x}) \\
&= \exp\left(-\sum_{m=n}^{\infty} \mathsf{P}\left(E_m\right)\right) \\
&= e^{-\infty} = 0.
\end{aligned}$$

Thus $\mathsf{P}\left(\left(E_n \text{ i.o.}\right)^{\complement}\right) = 0$; it remains to take complements. ■

Sometimes, we wont have mutual independence. We can still get a similar result if we have a statement of the form

$$\mathsf{P}\left(E_{n+1} \mid \sigma(E_1, \ldots, E_n)\right) \geq c_n$$

for some universal lower bound.

Example 5.5

Let $(X_i, i \geq 1) \sim \mathcal{E}(1)$, that is $\mathsf{P}\left(X_i \geq x\right) = e^{-x}, x \geq 0$. Look at $M_n = (\max X_i, 1 \leq i \leq n)$.

Proposition 5.8

$$\limsup_{n\to\infty} \frac{M_n}{\log(n)} \overset{\text{a.s.}}{=} 1.$$

Proof To get to apply Borel-Cantelli 2, we need $\frac{M_n}{\log(n)} \geq 1$ and $\frac{M_n}{\log(n)} \geq 1 + \varepsilon$ at most finitely often.

Lemma 5.9

$$\limsup_{n\to\infty} \frac{X_n}{\log(n)} \overset{\text{a.s.}}{=} 1.$$

Proof Let $E_n = \left\{\frac{X_n}{\log(n)} \geq 1\right\}$. The E_n are independent and

$$\mathsf{P}\left(E_n\right) = \mathsf{P}\left(X_n \geq \log(n)\right) = \frac{1}{n}$$

so $\sum_{n=1}^{\infty} \mathsf{P}(E_n) = \infty$ and $\mathsf{P}(E_n \text{ i.o.}) = 1$. So $\limsup_{n\to\infty} \frac{X_n}{\log(n)} \geq 1$ almost surely. ■

Next, fix $\varepsilon > 0$ and let $F_n = \{X_n \geq (1+\varepsilon)\log(n)\}$. Then $\mathsf{P}(F_n) = \frac{1}{n^{1+\varepsilon}}$ so $\sum_{n\geq 1} \mathsf{P}(F_n) < \infty$. So $\limsup_{n\to\infty} X_n/\log(n) \leq 1+\varepsilon$ almost surely using Borel-Cantelli 1. Since $\varepsilon > 0$ is arbitrary, we can conclude that $\limsup_{n\to\infty} X_n/\log(n) \leq 1$ almost surely.

Many properties depend on the underlying properties of the space. In our example, we care about properties of sequences.

Lemma 5.10

If $\{a_n, n \geq 1\}$ is any sequence of real numbers and $f : \mathbb{N} \to \mathbb{R}$, $f(n) \uparrow \infty$ as $n \to \infty$. Then $a_n \geq f(n)$ for infinitely many n if and only if

$$\max_{1\leq i\leq n} a_i \geq f(n) \text{ for infinitely many } n.$$

Proof The first part is obvious. For the second, consider the graph above.

Fix a sequence $n_k, k \geq 1$ such that $\max_{1\leq i\leq n_k} a_i \geq f(n_k)$ and also $f(n_{k+1}) > \max_{1\leq i\leq n_k} a_i$. Then, for each $k \geq 1$, there is $i_k \in (n_k, n_{k+1})$ such that a_{i_k} is at least $f(n_{k+1})$. Thus $a_{i_k} \geq f(n_{k+1}) \geq f(i_k)$. We have a sequence i_k with the desired property and so $a_n \geq f(n)$ infinitely often. ■

By the previous lemma, we have $a_n/f(n) \geq 1$ if and only if $\max a_n/f(n) \geq 1$. Both limsup must thus be the same, for any sequence. For all ω, we evaluate the function

$$\limsup_{n\to\infty} \frac{M_n(\omega)}{\log(n)} = \limsup_{n\to\infty} \frac{X_n(\omega)}{\log(n)}$$

so

$$\left\{\limsup_{n\to\infty} \frac{M_n(\omega)}{\log(n)} = 1\right\} = \left\{\limsup_{n\to\infty} \frac{X_n(\omega)}{\log(n)}\right\}$$

and by the first lemma, $\mathsf{P}(\limsup_{n\to\infty} M_n/\log(n) = 1) = 1$ and the result is true almost surely. ■

Note

One can show (in almost the same way) that

$$\limsup_{n\to\infty} M_n - \log(n) - \log(\log(n)) = \infty.$$

Since the sum of $\sum_{n=1}^{\infty} \frac{1}{n\log(n)} = \infty$, by Borel-Cantelli 2, we get a divergent series which implies that the event happens infinitely often with probability one. On the other hand

$$\limsup_{n\to\infty} M_n - \log(n) - (1+\varepsilon)\log(\log(n)) = -\infty$$

as $\displaystyle\sum_{n=0}^{\infty} \frac{1}{n\log(n)^{1+\varepsilon}} < \infty.$

We do two last example with Borel-Cantelli lemma 2.

Example 5.6 (St-Petersburg Paradox)

Let

$$Z_i := 2^{\#\text{heads before the first tail}} \quad \text{in a sequence of Bernoulli trials}$$

where $(B_{i,j}, j \geq 1), B_{i,j} \overset{\mathrm{d}}{=} \mathcal{B}\left(\frac{1}{2}\right)$ are independently and identically distributed (iid).

The question is: what is a fair entry fee for a game in which you double your stack every time a head happens, and lose everything at the first tail, starting with 1.

To answer the question, we compute the expected profit:

$$\begin{aligned}\mathsf{E}\left(Z_1\right) &= \sum_{k\geq 0} 2^k \mathsf{P}\left(k \text{ heads before first tail}\right)\\ &= \sum_{k\geq 0} 2^k \frac{1}{2^{k+1}}\\ &= \sum_{k\geq 0} \frac{1}{2} = \infty.\end{aligned}$$

Thus finite random variables can have infinite expectations and yet paradoxically

$$\mathsf{P}\left(Z_i \geq i\log(i) \text{ i.o.}\right) = 1$$

since $(Z_i, i \geq 1)$ are mutually independent and

$$\sum_{i\geq 1} \mathsf{P}\left(Z_i \geq i\log(i)\right) = \sum_{i\geq 1} \frac{1}{2^{k(i)}}$$

where $k(i) = \lceil \log_2(i\log(i)) \rceil$

$$\begin{aligned} &\geq \frac{1}{2}\sum_{i\leq 1} \frac{1}{2^{\log_2(i\log(i))}} \\ &= \frac{1}{2}\sum_{i\geq 1} \frac{1}{i\log(i)} = \infty. \end{aligned}$$

Exercise 5.11

Prove this if it is not obvious to you (use independence).

Example 5.7

Let $X_i, i \geq 1$ be independent such that $\mathsf{P}\left(X_i = k\right) = \frac{3}{\pi^2}\frac{1}{k^2}$, for $k \in \mathbb{Z} \setminus \{0\}$. The expectation of X_i does not exist; the sum of the positive terms goes to infinity and the sum of the negative terms to $-\infty$, as $\sum_{k\geq 1} k\mathsf{P}\left(X_i = k\right) = \frac{3}{\pi^2}\sum_{k\geq 1}\frac{1}{k} = \infty$. [14]

Let $S_n = \sum_{i=1}^n X_i$. Then

$$\mathsf{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \text{ exists and is finite }\right) = 0$$

Proof $\mathsf{P}\left(|X_n| \geq n \text{ i.o.}\right) = 1$ by Borel-Cantelli 2. This is since

$$\begin{aligned} \sum_{n\geq 1} \mathsf{P}\left(|X_n| \geq n\right) &= \sum_{n\geq 1}\left(\sum_{m\geq n} \frac{6}{\pi^2}\frac{1}{m^2}\right) \\ &\geq \sum_{n\geq 1} \frac{6}{\pi^2}\frac{1}{n+1} = \infty \end{aligned}$$

using a lower bound derived from the integral test $\int_n^\infty \frac{\mathrm{d}x}{(x+1)^2} = (n+1)^{-1}$. But for all $c \in \mathbb{R}$ for n sufficiently large and $\varepsilon > 0$ sufficiently small, if $|S_n/n - c| < \varepsilon$ and

[14] In such case, when $\mathsf{E}\left(X^+\right) - \mathsf{E}\left(X^-\right) = \infty$, we say that the expectation is not defined, in the same way a function is not integrable if it is not absolutely integrable.

$|X_{n+1}| \geq n+1$, then

$$\left|\frac{S_{n+1}}{n+1} - c\right| \geq \frac{1}{4}.$$

As we can write $S_{n+1}/n = S_n/n + X_{n+1}/n$ and $S_{n+1}/(n+1) = S_{n+1}/n \cdot \frac{n}{n+1}$ and if $\left|\frac{S_{n+1}}{n} - c\right| > \frac{1}{2}$ and n sufficiently large then $|S_{n+1}/(n+1) - c| > \frac{1}{4}$. More formally, prove that

$$\left\{\exists \lim_{n\to\infty} \frac{S_n}{n} < \infty\right\} \subseteq \left\{\frac{|X_n|}{n} \geq 1 \text{ i.o.}\right\}^{\complement}.$$

This is an exercise in real analysis; fix ω and look at the corresponding sequence. ■

5.3 Kolmogorov's 0-1 law

Definition 5.11 (Tail σ-algebra)

Let $(X_n, n \geq 1)$ be random variables on $(\Omega, \mathcal{F}, \mathsf{P})$ where $X_n : \Omega \to \mathbb{R}$. Let $\mathcal{T}_n = \sigma(X_n, X_{n+1}, \ldots) = \sigma\left(\bigcup_{m\geq n} \sigma(X_n)\right)$ and let $\mathcal{T} = \bigcap_{n\geq 1} \mathcal{T}_n$. $\mathcal{T}$ is called the **tail σ-algebra**.

Exercise 5.12

Give an example where $\mathcal{T} = \mathcal{T}_1$ (so X_1 is $\mathcal{T} - \mathbb{B}(\mathbb{R})$ measurable). Hint: take the first coordinate map of $\{0,1\}^{\mathbb{N}}$. All the information regarding the first coin toss is contained in the first event. Formalize this example.

Note

$\sigma(X_n) \subset \mathcal{F}$ for all n, so $\bigcup_{m\geq n} \sigma(X_m) \subset \mathcal{F}$ for all m and so $\mathcal{T}_n \subset \mathcal{F}$ for all n and $\mathcal{T} \subset \mathcal{F}$ and $\mathcal{F} \supset \mathcal{T}_1 \supset \mathcal{T}_2 \supset \cdots \supset \mathcal{T}$.

Exercise 5.13

The following events are all in $\mathcal{T}$:

1. $\lim_{k\to\infty} X_k$ exists
2. $\sum_k X_k$ converges, i.e. $\lim_k \left(\sum_{j=1}^k X_j\right)$ exists.
3. $\lim_{k\to\infty}(X_1 + X_2 + \cdots + X_k)/k$ exists

Start with sequences, and translate these into probability statements; since there are functions, there is an additional step.

Theorem 5.12 (Kolmogorov's 0-1 law)

If $(X_n, n \geq 1)$ are independent (and $\mathcal{T}$ is their tail σ-algebra), then for all $E \in \mathcal{T}$, either $\mathsf{P}(E) = 0$ or $\mathsf{P}(E) = 1$.

Corollary 5.13

If the limits in the exercise exists, they are almost surely constant.

Proof [of corollary] We prove (iii) only. For all $c \in \mathbb{R}$, $E_c = \left\{\lim_{n\to\infty} \overline{X}_n \text{ exists}, \leq c_n\right\}$. Then, $E_c \in \mathcal{T}$, so $\mathsf{P}(E_c) \in \{0, 1\}$ $\forall\, c \in \mathbb{R}$. These events are increasing (like a CDF): let $c^* = \sup\{c : \mathsf{P}(E_c) = 0\}$. Then $\forall\, c > c^*, \mathsf{P}(E_c) = 1$. So

$$\mathsf{P}\left(\lim_{n\to\infty} \overline{X}_n \text{ exists and is not equal to } c^*\right) = 0.$$

■

Proof We now tackle the proof of Kolmogorov's 0-1 law. The aim is to show that $\forall\, E \in \mathcal{T}$,

E is independent of E, so that $\mathsf{P}(E \cap E) = \mathsf{P}(E)\,\mathsf{P}(E)$ so that $\mathsf{P}(E) = \mathsf{P}(E)^2$ which forces $\mathsf{P}(E) \in \{0, 1\}$

Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_{n-1})$, then (exercise) $\mathcal{F}_n, \mathcal{T}_n$ are independent for all n. Let $\mathcal{P} = \bigcup_n \mathcal{F}_n$; then $\mathcal{P}$ is a π-system: if $A, B \in \mathcal{P}$, then $A, B \in \mathcal{F}_n$ for some n, so the intersection $A \cap B \in \mathcal{F}_n$ and $A \cap B \in \bigcup_{m\geq 1} \mathcal{F}_m = \mathcal{P}$ since the sequence of σ-algebra is nested in one another.

Furthermore, $\mathcal{P}, \mathcal{T}$ are independent (as $\mathcal{F}_n \perp\!\!\!\perp \mathcal{T}\ \forall\, n$). If $E \in \mathcal{P}$ and $F \in \mathcal{T}$, then $E \in \mathcal{F}_n$ for some n and $F \in \mathcal{T}_n, F \in \mathcal{T} \subset \mathcal{T}_n$.

Thus, $\sigma(\mathcal{P})$ and $\mathcal{T}$ are independent. But

$$\begin{aligned}\sigma(\mathcal{P}) &= \sigma\left(\bigcup_{n\geq 1} \sigma(X_1, \ldots, X_{n-1})\right)\\ &\supseteq \sigma\left(\bigcup_{n\geq 1} \sigma(X_{n-1})\right)\end{aligned}$$

$$= \sigma(X_n, n \geq 1).$$

and $\mathcal{F}_n \subseteq \sigma(X_n, n \geq 1)$ so $\mathcal{P} = \bigcup_{n\geq 1} \mathcal{F}_n \subset \sigma(X_n, n \geq 1)$ and thus $\sigma(\mathcal{P}) \subseteq \sigma(X_n, n \geq 1)$. So $\sigma(\mathcal{P}) = \sigma(X_n, n \geq 1)$ and $\sigma(X_n, n \geq 1)$ and $\mathcal{T}$ are independent. But $\mathcal{T}_n \subseteq \sigma(X_n, n \geq 1)$ for all n and $\mathcal{T} = \bigcap_{n\geq 1} \mathcal{T}_n \subset \sigma(X_n, n \geq 1)$. So $\mathcal{T}$ is independent of $\mathcal{T}$, i.e. for all $E, F \in \mathcal{T}$, $\mathsf{P}(E \cap F) = \mathsf{P}(E)\,\mathsf{P}(F)$. Now take $E = F$. ■

Chapter 6

Integration and expectation

6.1 Defining integrals

We consider a function f on $[0,1]$ and its area under the curve, that is $\int_0^1 f(x)\,\mathrm{d}x$, which corresponds to E (height of a random point under the curve). The region under the curve is the set $\{(x,y) : 0 \leq y \leq f(x)\}$. [15] We define integrals in four steps, starting with

1. indicator functions
2. simple functions: $f = \sum_{i=1}^m c_i \mathbf{1}_{E_i}$
3. nonnegative functions (as increasing limits of simple functions)
4. general functions f (by decomposing $f = f^+ - f^-$).

For the rest of the development, fix a measure space $(\Omega, \mathcal{F}, \mu)$ which we will assume is σ-finite (that is $\Omega_n \uparrow \Omega, \Omega_n \in \mathcal{F}, \mu(\Omega_n) < \infty$).

Step 1: If $f = \mathbf{1}_E$ for $E \in \mathcal{F}$, let $\int f\,\mathrm{d}\mu = \mu(E)$.[16] One can check easily the linearity of expectation for simple functions.

Step 2: If f is simple, $f = \sum_{i=1}^m c_i \mathbf{1}_{E_i}$, let $\int f\,\mathrm{d}\mu = \sum_{i=1}^m c_i \mu(E_i)$.

Proposition 6.1

The integral of simple functions is well-defined.

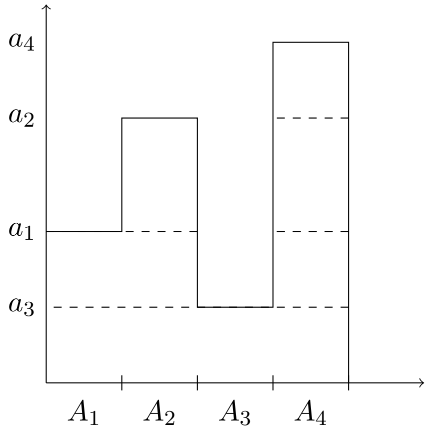
Proof Let f be simple and list the distinct values of f as $a_1, \ldots, a_l$. Let $A_i = f^{-1}(\{a_i\})$ and suppose

$$f = \sum_{i=1}^m c_i \mathbf{1}_{E_i} = \sum_{j=1}^l a_j \mathbf{1}_{A_j}.$$

and then split $E_1, \ldots E_m$ into smaller sets $G_{i,j}, 1 \leq j \leq l$ where $G_{i,j} = E_i \cap A_j$.

[15] Not a point on the line of f, rather a point under the curve.

[16] The book sometimes use μf instead as opposed to $f\,\mathrm{d}\mu$.



Then, we must have

$$\sum_{i=1}^{m} c_i \mathbf{1}_{E_i} = \sum_{i=1}^{m} c_i \left(\sum_{j=1}^{l} \mathbf{1}_{G_{i,j}} \right)$$

So

$$\begin{aligned} \mu(E_i) &= \int \mathbf{1}_{E_i} \, \mathrm{d}\mu \\ &= \sum_{j=1}^{l} \int \mathbf{1}_{G_{i,j}} \, \mathrm{d}\mu \\ &= \sum_{j=1}^{l} \mu(G_{i,j}) \end{aligned}$$

and thus

$$\begin{aligned} \sum_{i=1}^{m} c_i \mu(E_i) &= \sum_{i=1}^{m} c_i \left(\sum_{j=1}^{l} \mu(G_{i,j}) \right) \\ &= \sum_{j=1}^{l} \left(\sum_{i=1}^{m} c_i \mu(G_{i,j}) \right) \end{aligned}$$

where i is varying and j is fixed. ■

Proposition 6.2 (Basic properties of integrals for simple functions)

(i) If f, g are simple, then $\int (cf + g) \, \mathrm{d}\mu = c \int f \, \mathrm{d}\mu + \int g \, \mathrm{d}\mu$.

(ii) If f, g are simple and $f \leq g$, then $\int f \, \mathrm{d}\mu \leq \int g \, \mathrm{d}\mu$.

(iii) If f, g are simple, then $f \vee g$ (the maximum) and $f \wedge g$ (the minimum) are simple.

(iv) If f, g are simple and if $\mu(\{f \neq g\}) = 0$, then $\int f \,\mathrm{d}\mu = \int g \,\mathrm{d}\mu$. We say $f = g$ almost everywhere (a.e.).

Example 6.1

More generally, for extended real-valued functions, if $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and $g(x) = f(x)\mathbf{1}_{x \in \mathbb{R}\setminus\mathbb{Q}} + \infty \mathbf{1}_{x\in\mathbb{Q}}$, then $\int_A f(x)\,\mathrm{d}x = \int_A g(x)\,\mathrm{d}x$ for all Borel A and so g is "a" density for the standard normal distribution. We use the convention $\infty \cdot 0 = 0$.

Proof

(ii) $g - f$ is nonnegative and simple, so $g - f = \sum_{i=1}^n d_i \mathbf{1}_{F_i}$ and thus using (i),

$$\int g \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu + \int (g-f)\,\mathrm{d}\mu = \int f \,\mathrm{d}\mu + \sum_{i=1}^n \mathrm{d}_i \mu(F_i) \geq \int f \,\mathrm{d}\mu + 0.$$

(iv) write $f = g - \sum_{i=1}^n v_i \mathbf{1}_{\{f = g - v_i\}}$. Then

$$\begin{aligned} \int f \,\mathrm{d}\mu &= \int g \,\mathrm{d}\mu - \sum_{i=1}^n v_i \mu(\{f = g - v_i\}) \\ &= \int g \;\mu - \sum_{i=1}^n v_i 0 \\ &= \int g \,\mathrm{d}\mu. \end{aligned}$$

since simple functions take only finitely many values (as they are piecewise linear).

■

Example 6.2

In the St-Petersburg paradox (Example 5.6), we had $Z_i = 2^k$ with probability 2^{-k-1} for all $k \geq 0$. We could stop playing eventually (this corresponds to a truncation of the process) as to have $Z_i^{(N)} = Z_i \mathbf{1}_{Z_i \leq 2^N} + 2^N \mathbf{1}_{Z_i > 2^N}$. Since this now only takes finitely many values, we can define

$$\int Z_I^{(N)} \,\mathsf{dP} = \sum_{k=0}^N 2^k \mathsf{P}\left(\{Z_i = 2^k\}\right) = \mathsf{E}\left(Z_i^{(N)}\right) = \frac{N+1}{2}.$$

Note

$Z_i^{(N)} \leq Z_i^{(N+1)} \leq \cdots \leq Z_i$. So by Proposition 6.2 (ii),

$$\int Z_i^{(N)} \, \mathrm{dP} \leq \int Z_i^{(N+1)} \, \mathrm{dP}$$

and in fact

$$\mathsf{E}\left(Z_i^{(N)}\right) \uparrow \infty \qquad \text{as } n \to \infty$$

so we should have $\mathsf{E}\left(Z_i\right) = \infty$.[17]

Step 3: For f measurable nonnegative function, define

$$\int f \, \mathrm{d}\mu \coloneqq \sup \left\{ \int g \, \mathrm{d}\mu : g \text{ simple}, g \leq f \right\}.$$

Exercise 6.1

If f itself is simple, then this makes sense and agrees with the previous definition.

6.2 Monotone convergence theorem

Theorem 6.3 (Monotone Convergence Theorem)

If $f_n, n \geq 1$ are non-negative measurable functions and $f_n \uparrow f$, then $\int f_n \uparrow \int f$.[18] Since $\lim_{n\to\infty} f_n = f$ pointwise, then

$$\int f \, \mathrm{d}\mu = \int \lim_{n\to\infty} f_n \, \mathrm{d}\mu = \lim_{n\to\infty} \int f_n \, \mathrm{d}\mu.$$

Lemma 6.4

If $f \geq 0, \int f \, \mathrm{d}\mu = 0$, then $f = 0$ a.e.

Proof Let $E_n = \{f \geq 1/n\}$. Then $E_n \uparrow \{f > 0\}$ so

$$\int \mathbf{1}_{E_n} \, \mathrm{d}\mu \uparrow \int \mathbf{1}_{\{f \geq 0\}} \, \mathrm{d}\mu$$

[17] The book has $\mu(F_n) \uparrow \mu(f)$ if $f_n \uparrow f$.

[18] Note this is a theorem about interchange of limits and integral.

but $\int \mathbf{1}_{E_n} \, \mathrm{d}\mu = \mu(E_n)$ and

$$\int \mathbf{1}_{E_n} \, \mathrm{d}\mu = n \int \frac{1}{n} \mathbf{1}_{E_n} \, \mathrm{d}\mu \leq n \int f \, \mathrm{d}\mu$$

as $\frac{1}{n}\mathbf{1}_{E_n} \leq f$. Thus if $\mu(E) > 0$, then $\exists n$ such that $\mu(E_n) > 0$, so

$$0 \leq \mu(E_n) = \int \mathbf{1}_{E_n} \, \mathrm{d}\mu \leq n \int f \, \mathrm{d}\mu$$

so $\int f \, \mathrm{d}\mu > 0$. ■

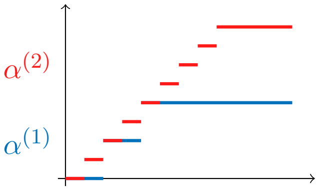
Note that countable additivity is equivalent to this limiting closure property. Recall that we defined for $f : \Omega \to \mathbb{R}$ non-negative measurable functions, we defined $\int f \, \mathrm{d}\mu = \sup\{\int g \, \mathrm{d}\mu : g \text{ simple}, g \leq f\}$. We will check this consistency property.

Definition 6.5

For $k \geq 1$, let $\alpha^{(k)} : [0, \infty) \to [0, \infty)$ be given by

$$\alpha^{(k)}(x) = \begin{cases} 0 & \text{if } x = 0 \\ \frac{i}{2^k} & \text{if } x \in \left(\frac{i}{2^k}, \frac{i+1}{2^k}\right], 0 \leq i \leq k2^k \\ k & \text{if } x > k \end{cases}$$

so for example $\alpha^{(1)}$ and $\alpha^{(2)}$ look like



For any measurable $f : \Omega \to [0, \infty)$, $\alpha^{(k)}(f)$ is simple, if $f_n \uparrow f$, then $\alpha^{(k)}(f_n) \uparrow \alpha^{(k)}(f)$ and furthermore, for any $g : \Omega \to [0, \infty)$, then $\alpha^{(k)} \uparrow g$ as $k \to \infty$.

The strategy for proving the MCT: take a **double** limit as $k, n \to \infty$, show that the limit doesn't depend on the order of those limits.

Lemma 6.6

If $(y_n^{(k)})$, $n, k \geq 1$ is increasing in n and k, then

$$\lim_{n\to\infty} \lim_{k\to\infty} y_n^{(k)} = \lim_{k\to\infty} \lim_{n\to\infty} y_n^{(k)}$$

$y_1^{(1)}$	$y_2^{(1)}$	$y_3^{(1)}$	$\cdots$	$y_\infty^{(1)}$
$y_1^{(2)}$	$y_2^{(2)}$	$y_3^{(2)}$	$\ddots$	$y_\infty^{(2)}$
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$
$y_1^{(\infty)}$	$y_2^{(\infty)}$	$\cdots$	$\cdots$	$y_\infty^{(\infty)}$

Lemma 6.7 (f_n simple, f indicator)

If f_n is simple and $f_n \uparrow \mathbf{1}_E$, then $\int f_n \, \mathrm{d}\mu \uparrow \int \mathbf{1}_E \, \mathrm{d}\mu$.

Proof We prove the upper bound first, if $f_n, \mathbf{1}_E$ is simple, then $f_n \leq \mathbf{1}_E$ so $\int f_n \, \mathrm{d}\mu \leq \int \mathbf{1}_E \, \mathrm{d}\mu$ and thus $\lim_{n\to\infty} \int f_n \, \mathrm{d}\mu \leq \int \mathbf{1}_E \, \mathrm{d}\mu$. For the lower bound, fix $\varepsilon > 0$ and let

$$F_n = \{f_n \geq (1-\varepsilon)\mathbf{1}_E\} = \{\omega \in \Omega : f_n(\omega) \geq (1-\varepsilon)\mathbf{1}_E(\omega)\}.$$

Then, since $f_n \uparrow \mathbf{1}_E$, $F_n \uparrow E$ so

$$\begin{aligned} \int \mathbf{1}_E \, \mathrm{d}\mu = \mu(E) &= \lim_{n\to\infty} \mu(F_n) \\ &= \lim_{n\to\infty} \int \mathbf{1}_{F_n} \, \mathrm{d}\mu \\ &= \lim_{n\to\infty} \frac{1}{1-\varepsilon} \int (1-\varepsilon)\mathbf{1}_{F_n} \, \mathrm{d}\mu \leq \frac{1}{1-\varepsilon} \lim_{n\to\infty} \int f_n \, \mathrm{d}\mu \end{aligned}$$

using the fact that on E, $(1-\varepsilon)\mathbf{1}_{F_n} \leq f_n$. The bound follows since ε was arbitrary. ■

Lemma 6.8 (f_n simple, f simple)

If $0 \leq f_n \uparrow f$ and f_n, f simple, then $\int f_n \, \mathrm{d}\mu \uparrow \int f \, \mathrm{d}\mu$.

Proof Write $f = \sum_{i=1}^m c_i \mathbf{1}_{E_i}$, $E_1, \dots E_m$ using the canonical decomposition. Since

$f_n \uparrow f$, we have $f_n \mathbf{1}_{E_i} \uparrow f \mathbf{1}_{E_i} = c_i \mathbf{1}_{E_i}$ so

$$\int f_n \mathbf{1}_{E_i} \, \mathrm{d}\mu = c_i \int \frac{1}{c_i} f_n \mathbf{1}_{E_i} \, \mathrm{d}\mu \uparrow c_i \int \mathbf{1}_{E_i} \, \mathrm{d}\mu = \int c_i \int_{E_i}$$

Let $E_{m+1} = \Omega \setminus (\bigsqcup_{i=1}^m E_i)$, $c_{m+1} = 0$. Then,

$$\begin{aligned} \int f \, \mathrm{d}\mu &= \sum_{i=1}^{m+1} c_i \int \mathbf{1}_{E_i} \, \mathrm{d}\mu \\ &= \lim_{n \to \infty} \sum_{i=1}^{m+1} \int f_n \mathbf{1}_{E_i} \, \mathrm{d}\mu \\ &= \lim_{n \to \infty} \int f_n \left(\sum_{i=1}^{m+1} \mathbf{1}_{E_i} \, \mathrm{d}\mu \right) \\ &= \lim_{n \to \infty} \int f_n \, \mathrm{d}\mu \end{aligned}$$

■

Lemma 6.9 (Consistency for simple sequences)

If $f_n, g_n \geq 0$ are simple, $f_n \uparrow f, g_n \uparrow f$ then $\lim_{n \to \infty} \int f_n \, \mathrm{d}\mu = \lim_{k \to \infty} \int g_k \, \mathrm{d}\mu$.

Proof Let $h_n^{(k)} = \min\{f_n, g_k\}$; then $(h_n^{(k)})$ are simple and $(h_n^{(k)})_{n,k \geq 1}$ is increasing in n, k. Then

$$\lim_{n \to \infty} \int f_n \, \mathrm{d}\mu = \lim_{n \to \infty} \lim_{k \to \infty} \int h_n^{(k)} \, \mathrm{d}\mu$$

By definition, we have $\int f_n \, \mathrm{d}\mu = \lim_{k \to \infty} \int h_n^{(k)} \, \mathrm{d}\mu$, since for all $\omega \in \Omega$, $h_n^{(k)}(\omega) = \min\{f_n(\omega), g_k(\omega)\} \uparrow f_n(\omega)$ as $k \nearrow$. Thus, the claimed convergence follows from Lemma 6.8. Likewise, by Lemma 6.8,

$$\lim_{k \to \infty} \int g_k \, \mathrm{d}\mu = \lim_{k \to \infty} \lim_{n \to \infty} \int h_n^{(k)} \, \mathrm{d}\mu.$$

Since $h_n^{(k)}(\omega)$ is an array of number, by Lemma 6.6,

$$\lim_{n \to \infty} \lim_{k \to \infty} \int h_n^{(k)} \, \mathrm{d}\mu = \lim_{k \to \infty} \lim_{n \to \infty} \int h_n^{(k)} \, \mathrm{d}\mu.$$

■

Lemma 6.10 (MCT for simple approximant)

If $0 \leq f_n \uparrow f$ and f_n is simple then $\lim_{n\to\infty} \int f_n \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu$. [19]

Proof ($\leq$) We can appeal to monotonicity of $\int$ for non-negative functions or more simply use f_n as a simple lower bound for f in the definition, so $\int f_n \,\mathrm{d}\mu$ appears in the supremum that defines $\int f \,\mathrm{d}\mu$.

($\geq$) Since $\int f \,\mathrm{d}\mu = \sup\{\int g \,\mathrm{d}\mu, g \text{ simple}, g \leq f\}$. We can find a sequence g_k with $k \geq 1$ of simple lower bounds for f such that $\int g_k \,\mathrm{d}\mu \uparrow \int f \,\mathrm{d}\mu$. Let $h_n = \max(f_1, \dots f_n, g_1, \dots g_n)$ is simple. Then $h_n \uparrow f$, so by Lemma 6.9, we know

$$\lim_{n\to\infty} \int f_n \,\mathrm{d}\mu = \lim_{n\to\infty} \int h_n \,\mathrm{d}\mu \geq \lim_{n\to\infty} \int g_n \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu$$

by monotonicity for simple functions, as $h_n \geq g_n$. ■

Proof [of MCT] Suppose $0 \leq f_n \uparrow f$. Then $\alpha^{(k)}(f_n) \uparrow \alpha^{(k)}(f)$ as $n \to \infty$ and $\alpha^{(k)}(f_n) \uparrow f_n$ as $k \to \infty$ because $\alpha^{(k)}$ increases to the identity. The first is simple, but the second one isn't. Then, by Lemma 6.10

$$\begin{aligned}
\lim_{n\to\infty} \int f_n \,\mathrm{d}\mu &= \lim_{n\to\infty} \lim_{k\to\infty} \int \alpha^{(k)}(f_n) \,\mathrm{d}\mu && \text{(Lemma 6.6)} \\
&= \lim_{k\to\infty} \lim_{n\to\infty} \int \alpha^{(k)}(f_n) \,\mathrm{d}\mu && \text{(Lemma 6.8, 6.10)} \\
&= \lim_{k\to\infty} \int \alpha^{(k)}(f) \,\mathrm{d}\mu \\
&= \int f \,\mathrm{d}\mu. && \text{(Lemma 6.10)}
\end{aligned}$$

since we can interchange limits by Lemma 6.6 and by 6.10, $\alpha^{(k)}$ is a simple approximation of f . ■

Exercise 6.2

If $0 \leq f \leq g$ and $\int f \,\mathrm{d}\mu \leq \int g \,\mathrm{d}\mu$, then

[19] Note that we need positivity, otherwise we could take $-\frac{1}{n}$ on $\mathbb{R}$. The integral is $-\infty$, but the limit is the zero function which has integral 0.

- linearity of integration for non-negative functions.
- If $f \leq g$ and $\int f \, \mathrm{d}\mu = \int g \, \mathrm{d}\mu$, then $\mu(\{\omega f(\omega) \neq g(\omega)\}) = 0$
- If $f \overset{\text{a.e.}}{=} g$, then $\int f \, \mathrm{d}\mu = \int g \, \mathrm{d}\mu$.

The last step is to get this for negative functions as well.

6.3 Expectation

Let $f : \Omega \to [0, \infty)$, then $\int f \, \mathrm{d}\mu$ is defined. If $f : \Omega \to \mathbb{R}$ is measurable, say $f \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathsf{P})$ if $\int f^+ \, \mathrm{d}\mu < \infty, \int f^- \, \mathrm{d}\mu < \infty$, that is if $\int |f| \, \mathrm{d}\mu < \infty$. For $f \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu)$, let $\int f \, \mathrm{d}\mu = \int f^+ \, \mathrm{d}\mu - \int f^- \, \mathrm{d}\mu$.[20]

Exercise 6.3

Check basic properties (e.g. linearity)

Now let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space. In this case, for $X : \Omega \to \mathbb{R}$ measurable, $X \geq 0$ or $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathsf{P})$, and define

$$\mathsf{E}\left(X\right) = \int X \, \mathrm{d}\mathsf{P},$$

to be the expectation of X

Exercise 6.4

If X is **discrete**, ($\exists$ countable S such that $\mathsf{P}\left(X = S\right) = 1$), then

$$\mathsf{E}\left(X\right) = \sum_{x \in S} x \mathsf{P}\left(X = x\right).$$

Proof Monotone convergence theorem for use with the simple function approximation. ■

[20] where as before $f^+ := \max(f, 0)$ and $f^- = \min(-f, 0)$.

6.4 Change of variables

Definition 6.11 (Change of variable)

Given a measurable function $h :\to \mathbb{R}$, $h \geq 0$, define a measure on $(\Omega, \mathcal{F})$ by

$$\mu_h(A) = \int h \mathbf{1}_A \, \mathrm{d}\mathsf{P} = \mathsf{E}\left(h \mathbf{1}_A\right).$$

Proposition 6.12

μ_h defined above is a measure.

Proof $\mu_h(A) \geq 0 \; \forall \; A \in \mathcal{F}$. Since $h, \mathbf{1}_A$ both are measurable, so is their product. We need thus to check that if $A_i, i \geq 1$ are disjoint, then

$$\begin{aligned}
\mu_h\left(\bigcup_{i\geq 1} A_i\right) &= \mathsf{E}\left(h \mathbf{1}_{\cup_{i\geq 1} A_i}\right) \\
&= \mathsf{E}\left(h \sum_{i\geq 1} \mathbf{1}_{A_i}\right) \\
&= \sum_{i\geq 1} \mathsf{E}\left(h \mathbf{1}_{A_i}\right) && \text{(linearity)} \\
&= \sum_{i\geq 1} \mu_H(A_i)
\end{aligned}$$

as

$$\begin{aligned}
\mathsf{E}\left(h \sum_{i\geq 1} \mathbf{1}_{A_i}\right) &= \mathsf{E}\left(\sum_{i\geq 1} h \mathbf{1}_{A_i}\right) \\
&= \mathsf{E}\left(\lim_{n\to\infty} \sum_{i=1}^{n} h \mathbf{1}_{A_i}\right) \\
&= \lim_{n\to\infty} \mathsf{E}\left(\sum_{i=1}^{n} h \mathbf{1}_{A_i}\right) && \text{(MCT)} \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \mathsf{E}\left(h \mathbf{1}_{A_i}\right) \\
&= \sum_{i=1}^{\infty} \mathsf{E}\left(h \mathbf{1}_{A}\right)
\end{aligned}$$

One has to be careful about linearity since it holds for finite, not necessarily countable. ■

Theorem 6.13

If $f : \Omega \to \mathbb{R}$ measurable and $f \geq 0$ or $f \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu_h)$, then

$$\int f \, \mathrm{d}\mu_h = \int fh \, \mathrm{d}\mathsf{P} = \mathsf{E}\left(fh\right).$$

Note

Change of variables works starting from any σ-finite measurable space, not just a probability space.

Proof We use the standard machine

1. Let $f = \mathbf{1}_E, E \in \mathcal{F}$. Then

$$\int f \, \mathrm{d}\mu_h = \mu_h(E) = \mathsf{E}\left(h\mathbf{1}_E\right) = \mathsf{E}\left(hf\right)$$

2. If f is simple, and $f = \sum_{i=1}^n c_i \mathbf{1}_{E_i}$, then by linearity

$$\begin{aligned}\int f \, \mathrm{d}\mu_h &= \int \sum_{i=1}^n c_i \mathbf{1}_{E_i} \, \mathrm{d}\mu_h \\ &= \sum_{i=1}^n c_i \int \mathbf{1}_{E_i} \, \mathrm{d}\mu_h \\ &= \sum_{i=1}^n c_i \mathsf{E}\left(h\mathbf{1}_{E_i}\right) \\ &= \mathsf{E}\left(\sum_{i=1}^n c_i h \mathbf{1}_{E_i}\right) \\ &= \mathsf{E}\left(hf\right).\end{aligned}$$

3. Let $f \geq 0$; we use MCT twice and (2) in the middle. Then let f_n simple and $0 \leq f_n \uparrow f$. Then

$$\begin{aligned}\int f \, \mathrm{d}\mu_h &= \lim_{n\to\infty} \int f_n \, \mathrm{d}\mu_h && \text{(MCT)} \\ &= \lim_{n\to\infty} \mathsf{E}\left(hf_n\right) && (2)\end{aligned}$$

$$\begin{aligned} &= \mathsf{E}\left(hf\right) && (hf_n \uparrow hf, \mathrm{MCT}) \\ &= \mathsf{E}\left(hf\right) \end{aligned}$$

4. Let $f \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu_h)$ and write $f = f^+ - f^-$ and use linearity (3).

■

Definition 6.14 (Density)

A random variable X has a **density** $f : \mathbb{R} \to [0, \infty)$ if $\forall\, B \in \mathbb{B}(\mathbb{R})$,

$$\mathsf{P}\left(X \in B\right) = \int f \mathbf{1}_B \,\mathrm{d}\lambda = \int_B f(x) \,\mathrm{d}x$$

where λ is the Lebesgue measure.

Recall we had ν law of X if

$$\nu(B) = \mathsf{P}\left(X \in B\right) = \int \mathbf{1}_B \,\mathrm{d}\nu.$$

Saying that X has a density is saying that the law of X may be obtained from Lebesgue measure by a change of variables.

Proposition 6.15

If X has density f and law ν, then for all $h : \mathbb{R} \to \mathbb{R}$, $h(X) \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathsf{P})$ if and only if $h \in \mathcal{L}^1(\mathbb{R}, \mathbb{B}(\mathbb{R}), \nu) \Leftrightarrow hf \in \mathcal{L}^1(\mathbb{R}, \mathbb{B}(\mathbb{R}), \lambda)$. In this case,

$$\mathsf{E}\left(h(X)\right) = \int h \,\mathrm{d}\nu = \int hf \,\mathrm{d}\lambda = \int h(x) f(x) \,\mathrm{d}x$$

In particular,

Example 6.3

If $X \sim \mathcal{N}(0,1)$, then $\mathsf{E}\left(h(X)\right) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} h(x) \,\mathrm{d}x$. If $X \sim \mathcal{P}(\lambda), \mathsf{E}\left(h(X)\right) = \sum_{i=0}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} h(i)$. The Poisson does not have a density, we can still use the change of variable formula to it.

Note

Densities are **not** unique, since if $f \overset{\text{a.e.}\lambda}{=} g$, then $\int f \mathbf{1}_B \,\mathrm{d}\lambda = \int g \mathbf{1}_B \,\mathrm{d}\lambda$ and $\lambda(\{f \neq g\}) = 0$, but the functions f, g are arbitrary on the null set.

6.5 Restriction of measures

Definition 6.16 (Restriction of measures)

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space. If $A \in \mathcal{F}$, write $\mathsf{P}(\cdot; A)$ for the function from $\Omega : [0,1]$ with $\mathsf{P}(B; A) = \mathsf{P}(B \cap A)$ and $\mathsf{E}(\cdot; A)$ for the function with $\mathsf{E}(X; A) = \mathsf{E}(X\mathbf{1}_A)$.

Fact 6.17

$\mathsf{P}(\cdot; A)$ is a measure called the **restriction** of P to A. For a general measure space, the restriction is denoted μ_A.

Proposition 6.18

If $(\Omega, \mathcal{F}, \mu)$ is a measurable space and $f : \Omega \to \mathbb{R}$ with $f \geq 0$ or $f \in \mathcal{L}^1$, then $\forall\, A \in \mathcal{F}$,

$$\int_A f \, \mathrm{d}\mu := \int f\mathbf{1}_A \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu_A.$$

The proof of this fact follows from the standard machine. Fix A, and build up the collection of functions for which this hold.

Differentiation under the integral sign is a useful concept, but will be left as reading.

Example 6.4

Let

$$X_n = \begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

Show that $\mathsf{P}(S_n \geq 0 \text{ i.o.}) = 1$.

Proof First, suppose that we have the tail event $\{S_n \geq 0 \text{ i.o.}\} \in \{0, 1\}$ (which is not true). It thus suffices to show that $\mathsf{P}(S_n \geq 0 \text{ i.o.}) \geq 0$. Either $\{S_n \geq 0 \text{ i.o.}\}$ or $\{S_n \leq 0 \text{ i.o.}\}$ or both so $\Omega = \{S_n \geq 0 \text{ i.o.}\} \cup \{S_n \leq 0 \text{ i.o.}\}$ so $\mathsf{P}(S_n \geq 0 \text{ i.o.}) + \mathsf{P}(S_n \leq 0 \text{ i.o.}) \geq 1$. But these probabilities are equal by symmetry so each is $\geq \frac{1}{2}$.

Remark

$\{S_n \geq 0 \text{ i.o.}\}$ is not a tail event. Replace the latter by $\{\limsup_{n\to\infty} S_n = \infty\}$. But now

$$\Omega \neq \{\limsup_{n\to\infty} S_n = \infty\} \cup \{\limsup_{n\to\infty} (-S_n) = \infty\} \cup \{\exists M : |S_n| < M \;\forall\, n\}$$

Showing that $\{\exists M : |S_n| \leq M \;\forall\; n\}$ is a zero event does the job. For a simple symmetric random walk, look at the $2M$ steps in the same direction guarantees the walk drifts and leaves the interval. Let

$$E_i = \{X_{(i-1)2M+1} = X_{(i-1)2M+2} = X_{i2M} = 1\}.$$

Then

$$\{|S_n| < M \;\forall\, n\} \subseteq \bigcap_{i\geq 1} E_i.$$

Thus

$$\mathsf{P}\left(|S_n| \leq M \;\forall\, n\right) \leq \mathsf{P}\left(\bigcap_{i\geq 1} E_i^{\complement}\right) = \left(1 - \frac{1}{2^{2m}}\right)^{\infty} = 0.$$

By subadditivity,

$$\begin{aligned}\mathsf{P}\left(\exists M \in \mathbb{N} : |S_n| < M \;\forall\, n\right) &= \mathsf{P}\left(\bigcup_{M\geq 1} \{|S_n| \leq M \;\forall\, n\}\right) \\ &\leq \sum_{M\geq 1} \mathsf{P}\left(\{|S_n| \leq M \;\forall\, n\}\right) = 0\end{aligned}$$

■

6.6 Integral Inequalities

Let $f_n, n \geq 1$ be a sequence of measurable function. We state without proof the following results:

Lemma 6.19 (Fatou)

If $f_n \geq 0$, then

$$\int \liminf_{n\to\infty} f_n \,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int f_n \,\mathrm{d}\mu$$

Lemma 6.20 (Reverse Fatou)

If $0 \leq f_n \leq g$ for all n, then

$$\int \limsup_{n\to\infty} f_n \,\mathrm{d}\mu \geq \limsup_{n\to\infty} \int f_n \,\mathrm{d}\mu$$

Proof Apply Fatou to the sequence $h_n = g - f_n$. ■

Exercise 6.5

Check necessity of the conditions for both lemmas.

Theorem 6.21 (Dominated convergence)

If $f_n \xrightarrow{\text{a.e.}} f$ and $|f_n| \leq g$ a.e. with $\int g \,\mathrm{d}\mu < \infty$. Then

$$\int f \,\mathrm{d}\mu = \lim_{n\to\infty} f_n \,\mathrm{d}\mu = \lim_{n\to\infty} \int f_n \,\mathrm{d}\mu.$$

The proof follows from Fatou's lemma.

Corollary 6.22 (Bounded convergence)

In a probability space, if $X_n \xrightarrow{\text{a.s.}} X$ and $|X_n| \leq M \;\forall\, n$ with $M \geq 0$ fixed, then

$$\mathsf{E}\left(X\right) = \lim_{n\to\infty} \mathsf{E}\left(X_n\right)$$

Proof $\mathsf{E}\left(M\right) = M < \infty$ since unlike general abstract space, the integral of constant functions is bounded and not infinite. ■

Note

Boundedness is **not** sufficient if the space has infinite total measure.

Example 6.5

Take $f_n = \frac{1}{n}\mathbf{1}_{|x|\leq n}$. Then $f_n \to 0$ pointwise and $|f_n| \leq 1$, but $\int f_n \,\mathrm{d}\lambda = 2 \neq 0 = \int 0 \,\mathrm{d}\lambda$.

6.7 Convergence, $\mathcal{L}^p$ spaces and $\mathcal{L}^p$ convergence

$\mathcal{L}^p(\Omega, \mathcal{F}, \mu)$ is the space of measurable functions $f : \Omega \to \mathbb{R}$ with $\int |f|^p \,\mathrm{d}\mu < \infty$. We say $f_n \xrightarrow{\mathcal{L}^p} f$ if $f_n \in \mathcal{L}^p \;\forall\, n, f \in \mathcal{L}^p$ and

$$\int |f_n - f|^p \,\mathrm{d}\mu \to 0.$$

Why not require that the difference between the integral goes to zero? In fact, since the latter would be asking that they both have same expectation, which tells nothing about the functions being close together. One could consider a Gaussian random variable with mean 1, an Exponential random variable or even a Poisson; these have little in common.

Note

One can view $\mathcal{L}^p$ as a vector space. Indeed,

$$\int |cf|^p \,\mathrm{d}\mu = |c|^p \int |f|^p \,\mathrm{d}\mu$$

and

$$\begin{aligned}\int |f+g|^p \,\mathrm{d}\mu &\le \int (2\max(|f|, |g|))^p \,\mathrm{d}\mu. \\ &= 2^p \int \max(|f|, |g|)^p \,\mathrm{d}\mu \\ &= 2^p \int \max(|f|^p, |g|^p) \,\mathrm{d}\mu \\ &\le 2^p \int (|f|^p + |g|^p) \,\mathrm{d}\mu \\ &= 2^p \left(\int |f|^p \,\mathrm{d}\mu + \int |g|^p\right)\end{aligned}$$

Thus $\mathcal{L}^p(\Omega, \mathcal{F}, \mu)$ is a real vector space. In fact, it has a **norm**:

$$\|f\|_p = \left(\int |f|^p \,\mathrm{d}\mu\right)^{\frac{1}{p}}$$

which follows from Minkowski inequality: thus $\|\cdot\|_p$ satisfies the triangle inequality.

$\mathcal{L}^p$ convergence and other types of convergence

Note

If $f_n \xrightarrow{\mathcal{L}^p} f$ then for any $\varepsilon > 0$,

$$\int |f_n - f|^p \,\mathrm{d}\mu \geq \int |f_n - f|^p \,\mathbf{1}_{|f_n - f| > \varepsilon} \,\mathrm{d}\mu \geq \int \varepsilon^p \mathbf{1}_{|f_n - f| > \varepsilon} \,\mathrm{d}\mu = \varepsilon^p \mu(\{|f_n - f| \geq \varepsilon\})$$

but the leftmost term converges to zero, so $\mu(\{|f_n - f| \geq \varepsilon\}) \to 0$ as well. In other words, $\mathcal{L}^p$ convergence implies convergence in measure.

Definition 6.23 (Convergence in measure)

$f_n \xrightarrow{\mu} f$ (in measure) if $\forall\ \varepsilon > 0$, $\mu(|f_n - f| \geq \varepsilon) \to 0$. For random variables, $\mathsf{P}\left(|X_n - X| \geq \varepsilon\right) \to 0$.

Example 6.6

If M_n is the maximum win in the first n St-Petersburg games, then $M_n/(2^n \log(n)) \xrightarrow{\mathsf{p}} 0$ because $\mathsf{P}\left(|M_n/(n \log(\log(n)))| > \varepsilon\right) \to 0$ but $\limsup_{n\to\infty} M_n/(n\log(\log(n))) \xrightarrow{\text{a.s.}} \infty$.

This means that convergence in probability is stronger than convergence in measure.

Note

$\mathcal{L}^p$ convergence does not implies a.e. convergence, neither is it implied by.

Example 6.7 (Moving bumps)

Consider a function indexed by integers: for $n \geq 1, 0 \leq k \leq 2^n$, write

$$f_{2^n+k} = n\mathbf{1}_{|x| \in \left[\frac{k}{2^n}, \frac{k+1}{2^n}\right]}$$

Then $f_n : [0,1] \to \mathbb{R}$, $f_n \xrightarrow{\mathcal{L}^p} 0 \;\forall\; p$ as

$$\int |f_n - 0|^p \,\mathrm{d}\lambda = \left(\frac{n}{2^n}\right)^p \to 0$$

It is in fact true that $f_n(x)$ not converging pointwise at any point x.

Example 6.8

To show that a.e. convergence does not imply $\mathcal{L}^p$-convergence, take a box $n\mathbf{1}_{|x|<n}$ which is a box of area $2 \;\forall\; n$ while it converges a.e. to the zero function.

Proposition 6.24

For $1 \leq p < q < \infty$, if $f_n \xrightarrow{\mathcal{L}^q} f$, then $f_n \xrightarrow{\mathcal{L}^p} f$. This is subtle, using Jensen's inequality.

Chapter 7

Tightness and inequalities

7.1 Tightness

This is an extract from Louigi's blog.

When teaching a first rigorous course on probability, I think it is valuable to emphasize that real-valued random variables should be thought of as random real numbers. What I mean by this is that one should investigate the properties of random variables by analogy with the behaviour of real numbers. In this post, I take this approach, and ask: what is the probabilistic analogue of the Bolzano-Weirstrass theorem?

Theorem 7.1 (Bolzano-Weierstrass theorem)

Any bounded sequence of real numbers has a convergent subsequence.

Proof Fix a bounded sequence $(x_n, n \geq 1)$ of real numbers; for convenience assume that $x_n \in [0,1]$ for all n. Then either infinitely many of the x_n are at least $1/2$, or infinitely many are at most $1/2$ (or both), so we may find an increasing sequence $(n_i^{(1)}, i \geq 1)$ such that the subsequence $(x_{n_i^{(1)}}, i \geq 1)$ of $(x_n, n \geq 1)$ lies either in $[0, 1/2]$ or in $[1/2, 1]$. Repeating this argument, we may find a sub-subsequence $(x_{n_i^{(2)}}, i \geq 1)$ that lies in an interval of width $1/4$, and a further subsequence $(x_{n_i^{(3)}}, i \geq 1)$ lying in an interval of width $1/8$, and so on.

Then use a diagonal argument; in other words, consider the subsequence $\boldsymbol{x} = (x_{n_i^{(i)}}, i \geq 1)$, by construction, for all $k \geq 1$, $(x_{n_i^{(i)}}, i \geq k)$ lies in an interval of width 2^{-k}. Thus $\boldsymbol{x}$ is Cauchy; since $\mathbb{R}$ is complete, it follows that $\boldsymbol{x}$ is convergent. ■

Note the essential use of completeness in the above argument. The argument generalizes to compact subsets of complete metric spaces, by choosing appropriate replacements for the "dyadic covers" of the previous proof.

Exercise 7.1

Show that if $(\mathbb{M}, d)$ is a complete metric space and $K \subset \mathbb{M}$ is compact, and $(x_n, n \geq 1)$ is a sequence of points in K, then $(x_n, n \geq 1)$ has a convergent subsequence.

What is the analogous result for sequences of random variables? To start, consider a sequence $(X_n, n \geq 1)$ of real random variables. How might one show that $(X_n, n \geq 1)$

has a convergent subsequence? It is tempting to argue as in the Bolzano-Weierstrass theorem and begin by saying that either infinitely many of the X_n take their value in $[1/2, 1]$ or in $[0, 1/2]$. This doesn't work, however; the X_n are functions, not numbers. We could, however, consider the sequence $(x_n, n \geq 1)$ with $x_n = \mathsf{P}\left(X_n \leq 0\right)$. This is a bounded sequence – each x_n is between 0 and 1 – so has a convergent subsequence $(x_{n_i^{(1)}}, i \geq 1)$.

We then have at least some information about the sequence $(X_{n_i^{(1)}}, i \geq 1)$, which is that $\mathsf{P}\left(X_{n_i^{(1)}} \leq 0\right)$ converges as $n \to \infty$. But we may then repeat this argument to obtain a subsubsequence along which $\mathsf{P}\left(X_{n_i^{(2)}} \leq 10\right)$, say, converges. Enumerate the rationals as $(q_i, i \geq 1)$ and apply this argument at each rational, then diagonalize as in Bolzano-Weierstrass; the result is a sequence $(X_{n_i^{(i)}}, i \geq 1)$ along which $\mathsf{P}\big(X_{n_i^{(i)}} \leq q\big)$ converges to some number $f_q \in [0, 1]$ for every rational number q.

Does this mean the sequence $(X_{n_i^{(i)}}, i \geq 1)$ converges? Not necessarily: perhaps $f_q = 0$ for every q (so $X_{n_i^{(i)}}$ "tends to infinity"), or instead is 1 for every q. To rule this out in the case of real numbers, we insisted the initial sequence should be bounded. In this case, we require the random variables to be stochastically bounded (or tight). This means that we may uniformly control the probability that the random variables take large values. More precisely, say that $(X_n, n \geq 1)$ is tight if for all $\varepsilon > 0$, there is $K(\varepsilon)$ such that $\mathsf{P}\left(|X_n| \geq K\right) < \varepsilon$ for all n. Writing μ_n for the law of X_n, we equivalently have that $(X_n, n \geq 1)$ is tight if and only if

$$\lim_{N\to\infty} \sup_{n\in\mathbb{N}} \mu_n(\mathbb{R} \setminus [-N, N]) = 0.$$

The latter condition defines tightness for a family of measures on $\mathbb{R}$, and more generally applicable as it does not require the measures to be probability measures.

If $(X_n, n \geq 1)$ is indeed tight then $f_{-K(\varepsilon)} < \varepsilon$ and $f_{K(\varepsilon)} > 1 - \varepsilon$. Thus, the sequence $(f_q, q \in \mathbb{Q})$ satisfies $f_q \to 0$ as $q \to -\infty$ and $f_q \to 1$ as $q \to \infty$. It is easily verified that f_q is non-decreasing in q. In order to define a cumulative distribution function, let

$$F(r) = \inf_{q>r, q\in\mathbb{Q}} f_q = \lim_{q\downarrow r, q\in\mathbb{Q}} f_q.$$

Exercise 7.2

Suppose that the random variables $(X_n, n \geq 1)$ form a tight family. Show that in this case the function F defined above is a cumulative distribution function (cdf). Let X be a random variable with cdf F. Show that $X_{n_i^{(i)}}$ converges to X in distribution as $i \to \infty$, i.e., that $\mathsf{P}(X_{n_i^{(i)}} \leq x) \to \mathsf{P}\left(X \leq x\right)$ for all x points of continuity of F.

Exercise 7.3

More generally, if $(X_n, n \geq 1)$ are random variables taking value in a metric space $(\mathbb{M}, d)$, we say that $(X_n, n \geq 1)$ is tight if for all $\varepsilon > 0$ there is a compact $K \subset \mathbb{M}$ such that $\mathsf{P}\left(X_n \notin K\right) < \varepsilon$ for all n. Show that in this case $(X_n, n \geq 1)$ has a subsequence $(X_{n_i^{(i)}}, i \geq 1)$ such that for all bounded, continuous functions $g : \mathbb{M} \to \mathbb{R}$, $g(X_{n_i^{(i)}})$ converges in distribution. (Hint: first prove this for a countable collection of indicator functions, then use a monotone class argument.)

Exercise 7.4 (Portmanteau theorem for real random variables)

Let $(X_n, n \geq 1)$ be a tight sequence of real random variables. Show that X_n converges in distribution if and only if for all bounded continuous functions $g : \mathbb{R} \to \mathbb{R}$, $\mathsf{E}\left(g(X_n)\right)$ converges.

We conclude with a word on uniform integrability. Say that the random variables $(X_n, n \geq 1)$ are uniformly integrable if

$$\lim_{N \to \infty} \sup_{n \in \mathbb{N}} \mathsf{E}\left(|X_n|; |X_n| \geq N\right) = 0.$$

Let μ_n be the law of X_n, and let $\widehat{\mu}_n$ be the linear tilt of μ_n, obtained by setting

$$\widehat{\mu}_n(A) = \mathsf{E}\left(X_n; A\right) .$$

Exercise 7.5 (Uniform integrability and tightness)

Show that $(X_n, n \geq 1)$ are uniformly integrable if and only if the family of measures $(\widehat{\mu}_n, n \geq 1)$ are tight.

7.2 Inequalities

Theorem 7.2 (Jensen inequality)

Let X be a random variable, $f : \mathbb{R} \to \mathbb{R}$ convex and $X, f(X) \in \mathcal{L}^1$, then

$$f(\mathsf{E}\,(X)) \leq \mathsf{E}\,(f(X))$$

Proof The idea is to look at the graph. Another proof is as follow: let $|\text{supp}(X) = 2$. Then $\exists \{a, b\} : \mathsf{P}\,(X \in (a, b)) = 1$. Prove for finite support by induction, then approximate by simple functions. ■

Corollary 7.3

The $\| \cdot \|_p$ norms are increasing, that is $\|X\|_p \leq \|X\|_r$ whenever $1 \leq p \leq r$. Recall that

$$\|X\|_p = \mathsf{E}\,(|X|^p)^{\frac{1}{p}}$$

Proof $|X|^r = (|X|^p)^{\frac{r}{p}}$ so take $f(x) = |x|^{\frac{r}{p}}$, a convex function. Then, by Jensen,

$$\begin{aligned}
\mathsf{E}\,(|X|^r) &= \mathsf{E}\left((|X|^p)^{\frac{r}{p}}\right) \\
&= \mathsf{E}\,(f\,(|x|^p)) \\
&\geq f\,(\mathsf{E}\,(|X|^p)) \\
&= (\mathsf{E}\,(X_n^p))^{\frac{r}{p}}
\end{aligned}$$

and now take r^{th} roots. But we didn't check that $|X|^p$ or $f(|X|^p)$ are in $\mathcal{L}^1$.

The solution is to use monotone convergence theorem, with

$$\begin{aligned}
\mathsf{E}\,(|X|^p) &= \lim_{n \to \infty} \mathsf{E}\,(\min(|X|, n)^p) \\
\mathsf{E}\,(|X|^r) &= \lim_{n \to \infty} \mathsf{E}\,(\min(|X|, n)^r)
\end{aligned}$$

and as $r \to \infty$, $(\mathsf{E}\,(|X|^r))^{\frac{1}{r}} \to \text{esssup}(X)$ if $\|X\|_\infty < \infty$ and where

$$\text{esssup}(X) = \sup\{x : \mathsf{P}\,(X > x) > 0\}$$

■

Lemma 7.4 (Scheffé)

Fix $p \geq 1$ and suppose that $X_n \xrightarrow{\text{a.s.}} X$. Then

$$\mathsf{E}\left(|X_n|^p\right) \to \mathsf{E}\left(|X|^p\right) < \infty \quad \Leftrightarrow \quad \mathsf{E}\left(|X_n - X|^p\right) \to 0$$

Proof We start with sufficiency ($\Leftarrow$). If $\mathsf{E}\left(|X_n - X|^p\right) \to 0$, then $\|X_n - X\|_p \to 0$. Next,

$$\begin{aligned} \|X_n\|_p &= \|X + (X_n - X)\|_p \\ &\leq \|X\|_p + \|X_n - X\|_p \end{aligned}$$

using the triangle inequality, so

$$\limsup_{n\to\infty} \|X_n\|_p \leq \limsup_{n\to\infty} \left(\|X\|_p + \|X_n - X\|_p\right) = \|X\|_p$$

To conclude, replace X_n, X by $-X_n, X$ and repeat.

For necessity ($\Rightarrow$), Let $Y = -X$; then

$$\begin{aligned} |X_n - X|^p &= |X_n + Y|^p \\ &< 2^p \left(|X_n|^p + |Y|^p\right) \\ &= 2^p \left(|X_n|^p + |X|^p\right) \end{aligned}$$

Now, if we let $Z_n = 2^p \left(|X_n|^p + |X|^p\right) - |X_n - X|^p \geq 0$ so by Fatou's lemma,

$$\mathsf{E}\left(\liminf_{n\to\infty} Z_n\right) \leq \liminf_{n\to\infty} \mathsf{E}\left(Z_n\right)$$

but the left hand side is $2^{p+1}\mathsf{E}\left(|X|^p\right)$ using the fact that $X_n \xrightarrow{\text{a.s.}} X$ and the right hand side is by linearity (provided that the random variables are in $\mathcal{L}^1$, which they are if the limit is finite)

$$\liminf_{n\to\infty} \mathsf{E}\left(Z_n\right) = 2^p\mathsf{E}\left(|X|^p\right) + \liminf_{n\to\infty} 2^p\mathsf{E}\left(|X_n|^p\right) - \liminf_{n\to\infty} \mathsf{E}\left(|X_n - X|^p\right)$$

$$=2^p\mathsf{E}\left(|X|^p\right)+\liminf_{n\to\infty}2^p\mathsf{E}\left(|X|^p\right)-\liminf_{n\to\infty}\mathsf{E}\left(|X_n-X|^p\right)$$

Taking differences, we get

$$0\leq-\liminf_{n\to\infty}\mathsf{E}\left(|X_n-X|^p\right)$$

so the latter implies $\limsup_{n\to\infty}\mathsf{E}\left(|X_n-X|^p\right)=0$. It is always true that convergence in $\mathcal{L}^p$ implies convergence in norms. ■

We now give the proof of Hölder inequality. One may view it as a concentration inequality. For X,Y centered random variable, with $\mathsf{E}\left(X\right)=0,\mathsf{E}\left(Y\right)=0$, we say X,Y are **negatively correlated**. For $X,Y,XY\in\mathcal{L}^1$, X,Y are negatively correlated if

$$\mathsf{E}\left((X-\mathsf{E}\left(X\right))(Y-\mathsf{E}\left(Y\right))\right)\leq 0$$

or in other words

$$\mathsf{E}\left(XY\right)\leq\mathsf{E}\left(X\right)\mathsf{E}\left(Y\right).$$

Theorem 7.5 (Hölder inequality)

For $p,q\geq$ conjugate exponents, *i.e.* $\frac{1}{p}+\frac{1}{q}=1$, if $X\in\mathcal{L}^p,Y\in\mathcal{L}^q$, then $XY\in\mathcal{L}^1$ and

$$\begin{aligned}\mathsf{E}\left(|XY|\right)&\leq\mathsf{E}\left(|X|^p\right)^{\frac{1}{p}}\mathsf{E}\left(|Y|^q\right)^{\frac{1}{q}}\\\|XY\|_1&\leq\|X\|_p\|Y\|_q\end{aligned}$$

This can be used to optimize. A particular case is when $X,Y\in\mathcal{L}^2$, with $p=q=\frac{1}{2}$. We get then $\mathsf{E}\left(|XY|\right)\leq\sqrt{\mathsf{E}\left(X^2\right)\mathsf{E}\left(Y^2\right)}$, the so-called **Cauchy-Schwartz inequality.**

Note

We can make the proof and our lives easier by rescaling, without affecting the inequality. In fact, we may look at the equivalent statement

$$\mathsf{E}\left(\frac{|X|}{\mathsf{E}\left(|X|^p\right)^{\frac{1}{p}}}\frac{|Y|}{\mathsf{E}\left(|Y|^q\right)^{\frac{1}{q}}}\right)\leq\mathsf{E}\left(\frac{|X|^p}{\mathsf{E}\left(|X|^p\right)}\right)^{\frac{1}{p}}\mathsf{E}\left(\frac{|Y|^q}{\mathsf{E}\left(|Y|^q\right)}\right)^{\frac{1}{q}}=1$$

or

$$\mathsf{E}\left(\frac{|X|}{\|X\|_p}\frac{|Y|}{\|Y\|_q}\right) \leq \frac{(\mathsf{E}\,(|X|^p))^{\frac{1}{p}}}{\|X\|_p}\frac{(\mathsf{E}\,(|Y|^q))^{\frac{1}{q}}}{\|Y\|_q} = 1$$

If we relabel $\frac{|X|}{\mathsf{E}(|X|^p)^{\frac{1}{p}}} = X'$ and similarly for Y', we get $\|X'\|_p = \|Y'\|_q = 1$. So it suffices to show that if X, Y are nonnegative and $\|X\|_p = \|Y\|_q = 1$, then $\mathsf{E}\,(XY) \leq 1$.

The proof will follow easily from

Lemma 7.6 (Young's inequality)
For all $a, b > 0$,

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

Proof Take logarithms, then

$$\begin{aligned}\log(ab) &= \log(a) + \log(b) \\ &= \frac{1}{p}\log(a^p) + \frac{1}{q}\log(b^q) \\ &\leq \log\left(\frac{1}{p}a^p + \frac{1}{q}b^q\right).\end{aligned}$$

since logarithms are concave, so the weighted average is below the value of the value at the weighted average. (Recall that for concave functions f, $f(\theta\alpha + (1-\theta)\beta) \geq \theta f(\alpha) + (1-\theta)f(\beta)$ ■

Proof We just learned pointwise that

$$XY \leq \frac{X^p}{p} + \frac{Y^q}{q}$$

Taking expectations, the statement still holds true

$$\mathsf{E}\,(XY) \leq \frac{1}{p}\mathsf{E}\,(|X|)^p + \frac{1}{q}\mathsf{E}\,(|Y|^q)$$

$$
\begin{aligned}
&= \frac{1}{p}\|X\|_p^p + \frac{1}{q}\|Y\|_q^q \\
&= \frac{1}{p} + \frac{1}{q} = 1.
\end{aligned}
$$

Remark that we could get a bound that is less sharp than Hölder inequality, if p, q were not conjugate exponents. ■

Theorem 7.7 (Markov's inequality)

If $X \geq 0$ and $f : \mathbb{R} \to [0, \infty)$ increasing, then for all $t \geq 0$

$$\mathsf{P}\left(X \geq t\right) \leq \frac{\mathsf{E}\left(f(X)\right)}{f(t)}$$

Proof

$$\mathsf{E}\left(f(X)\right) \geq \mathsf{E}\left(f(X)\mathbf{1}_{X\geq t}\right) \geq \mathsf{E}\left(f(t)\mathbf{1}_{X\geq t}\right) = f(t)\mathsf{E}\left(\mathbf{1}_{X\geq t}\right) = f(t)\mathsf{P}\left(X \geq t\right).$$

■

Some important cases are when

- $f(x) = |x|, \mathsf{P}\left(|X| \geq t\right) \leq \mathsf{E}\left(|X|\right)/t$ for $t > 0$
- $f(x) = x^2, \mathsf{P}\left(|X| \geq t\right) \leq \mathsf{E}\left(X^2\right)/t^2$
- If $X \in \mathcal{L}^1$, then for $Y = X - \mathsf{E}\left(X\right)$,

$$\mathsf{P}\left(|X - \mathsf{E}\left(X\right)| \geq t\right) \leq \frac{\mathsf{E}\left(Y^2\right)}{t^2} = \frac{\mathsf{Var}\left(X\right)}{t^2}$$

- **Chebyshev's inequality**

$$\mathsf{Var}\left(X\right) = \mathsf{E}\left(\left(X - \mathsf{E}\left(X\right)\right)^2\right)$$

which is equal to $\mathsf{E}\left(X^2\right) - \left(\mathsf{E}\left(X\right)\right)^2$ if $X \in \mathcal{L}^1$, using linearity of expectation.

- Chebyshev inequality for sums: if $S = \sum_{i=1}^k X_i$ and $X_1, \ldots, X_k$ are pairwise independent, then $\mathsf{Var}\left(S\right) = \sum_{i=1}^k \mathsf{Var}\left(X_i\right)$. We prove for the mean zero case.

Proof Wlog, say $\mathsf{E}\left(X_i\right) = 0$, we then have

$$\begin{aligned}
\mathsf{E}\left(S^2\right) &= \mathsf{E}\left(\left(\sum_{i=1}^{k} X_i\right)^2\right) \\
&= \mathsf{E}\left(\sum_{i=1}^{k} X_i^2 + 2\sum_{1\le i<j\le k} X_i X_j\right) \\
&= \sum_{i=1}^{k} \mathsf{E}\left(X_i^2\right) + 2\sum_{1\le i<j\le k} \mathsf{E}\left(X_i X_j\right) \\
&= \sum_{i=1}^{k} \mathsf{Var}\left(X_i\right)
\end{aligned}$$

where we use the fact (which will prove shortly) that for X, Y independent, $\mathsf{E}\left(XY\right) = \mathsf{E}\left(X\right)\mathsf{E}\left(Y\right)$ so the cross term vanish. ■

Thus

$$\mathsf{P}\left(\left|S - \mathsf{E}\left(S\right)\right| > t\right) \le \frac{\sum_{i=1}^{k} \mathsf{Var}\left(X_i\right)}{t^2}.$$

Note that this bound is useful when $t^2 > k \sup_{1\le i\le k} \mathsf{Var}\left(X_i\right)$.

Proposition 7.8 (Chernoff bound)

If $(X_i, 1 \le i \le k)$ are iid with mean 0, then with $f(x) = e^{cx}$, using Markov inequality, we have for $S = \sum_{i=1}^{k} X_i$ for all $c > 0$,

$$\begin{aligned}
\mathsf{P}\left(S \ge t\right) &\le \frac{\mathsf{E}\left(e^{cS}\right)}{e^{ct}} && \text{(Markov)} \\
&= e^{-ct}\mathsf{E}\left(\prod_{i=1}^{k} e^{cX_i}\right) \\
& e^{-ct}\prod_{i=1}^{k} \mathsf{E}\left(e^{cX_i}\right) && \text{(independence)} \\
&= e^{-ct}\left(\mathsf{E}\left(e^{cX_i}\right)\right)^k && \text{(identically distributed)}
\end{aligned}$$

noting that $S \ge t \Leftrightarrow e^{cS} \ge e^{ct}$. Write $t = \alpha k$, then write

$$\mathsf{E}\left(e^{cX_1}\right) = \exp\left(\log\left(\mathsf{E}\left(e^{cX_1}\right)\right)\right)$$

to obtain

$$\mathsf{P}\left(S \geq \alpha k\right) \leq e^{-c\alpha k} \exp\left(k \log\left(\mathsf{E}\left(e^{cX_1}\right)\right)\right) = \exp\left(k \log\left(\mathsf{E}\left(e^{cX_1}\right) - \alpha c\right)\right)$$

If the expectation is finite and we do not run into paradoxs such as the St-Petersburg's one, Chernoff bound is useful. For given α, we can optimize for c and the bound is very good, even in comparison with Sterling formula, exponentially small is $\sqrt(n)$ error, which is not so bad. The "fonction Legendre" is what appear in the exponent, describing entropy or energy trade-off by choosing c.

We will now use Chebyshev inequality to prove a weak law of large number. We then prove Weirstrass approximation theorem, to show polynomials are dense in the bounded functions. We will start strengthening the weak law, with only finite mean, or existing variance. The book assumes finite fourth moment for the proof.

Proposition 7.9 (Independence means multiply)
If $X, Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathsf{P})$ for $X \perp\!\!\!\perp Y$, then $XY \in \mathcal{L}^1$ and $\mathsf{E}\left(XY\right) = \mathsf{E}\left(X\right)\mathsf{E}\left(Y\right)$.

We did not needed to use the former fact since the variance were finite and thus the statement could be deduced from Hölder.

Proof We use the standard machine. For indicators, if $\mathbf{1}_A \perp\!\!\!\perp \mathbf{1}_B \Leftrightarrow A \perp\!\!\!\perp B$, then

$$\mathsf{E}\left(\mathbf{1}_A \mathbf{1}_B\right) = \mathsf{E}\left(\mathbf{1}_{A\cap B}\right) = \mathsf{P}\left(A \cap B\right) = \mathsf{P}\left(A\right)\mathsf{P}\left(B\right) = \mathsf{E}\left(\mathbf{1}_A\right)\mathsf{E}\left(\mathbf{1}_B\right)$$

and the rest follows, we only need to prove for non-negative functions since checking that the functions are in $\mathcal{L}^1$ is equivalent to this. . ■

If $X_1, \ldots, X_k \in \mathcal{L}^1$ are mutually independent, then $\prod_{i=1}^k X_i \in \mathcal{L}^1$,

$$\mathsf{E}\left(\prod_{i=1}^k X_i\right) = \prod_{i=1}^k \mathsf{E}\left(X_i\right)$$

Chapter 8

Laws of large numbers

Theorem 8.1 (Weak law of large numbers)

If $(X_i, i \geq 1)$ are pairwise independent with mean zero and $\sup_{i\geq 1} \mathsf{Var}\left(X_i\right) < \infty$, then setting $S_n = \sum_{i=1}^n X_i$, we have $S_n/n \xrightarrow{d} 0$, that is $\forall\, \varepsilon > 0$,

$$\mathsf{P}\left(\left|\frac{S_n}{n}\right| > \varepsilon\right) \to 0 \text{ as } n \to \infty$$

Proof Using Chebyshev. Write $K = \sup_{i\geq 1} \mathsf{Var}\left(X_i\right)$, then

$$\begin{aligned}\mathsf{P}\left(\left|\frac{S_n}{n}\right| > \varepsilon\right) &= \mathsf{P}\left(|S_n| > \varepsilon n\right)\\ &\leq \frac{\sum_{i=1}^n \mathsf{Var}\left(X_i\right)}{\varepsilon^2 n^2} \leq \frac{nK}{\varepsilon^2 n^2} = \frac{K}{\varepsilon^2} \cdot \frac{1}{n}.\end{aligned}$$

■

Remark

If we had a bound that was stronger than $1/n$ and did converge, we could use Borel-Cantelli lemma to prove the strong law of large numbers. Assuming fourwise independence and using the fourth moment (assuming $\sup_{i\geq 1} \mathsf{E}\left(X_i^4\right) < \infty$), we get a tighter bound replacing Chebyshev inequality with Markov for $f(x) = x^4$. This is what is done in the book.

Remark

We can replace pairwise independence with pairwise uncorrelatedness[21]. We can also weaken the rate of convergence of the variance relative to n. The variancegrowing sublinearly, say $\log(n)$, is a sufficient condition. Changing the mean zero by some stabilization also works, since we are bounding deviations from the mean of the sum.

Proposition 8.2 (Weak law of large numbers in $\mathcal{L}^1$)

If $(X_i, i \geq 1)$ are identically distributed, pairwise independent with mean zero, and measure μ, then $S_n/n \xrightarrow{\mathcal{L}} 0$

[21] Complete relaxation would not work, if for example we take a coin toss as X_1 and all other random variables replication of this experiment. Then S_n/n is 1 $\forall\, n$

Remark

If $a \leq X \leq b$, then $|X - \mathsf{E}(X)| \leq b - a$ so that $\mathsf{Var}(X) \leq (b-a)^2$.

Exercise 8.1

The worst case scenario happens when $\mathsf{P}(X = b) = 1/2$ and $\mathsf{P}(X = a) = 1/2$. Improve this bound to $(b-a)^2/4$.

Proof For $i \geq 1, N \geq 0$, write $X_i^{\leq N} = X_i \mathbf{1}_{|X_i| \leq N}$ and $X_i^{>N} = X_i - X_i^{\leq N} = X_i \mathbf{1}_{|X_i| > N}$. The first elements have finite variance, so we can readily apply WLLN. The second have small mean, so do not contribute much.

Note

$|X_i^{\leq N}| \uparrow |X_i|$ so by the monotone convergence theorem,

$$\mathsf{E}\left(\left|X_I^{\leq N}\right|\right) \uparrow \mathsf{E}(|X_i|)$$

Only one of these is only non-zero, both are different one can apply the triangle inequality. Indeed, $X_i^{\leq N} + X_i^{<N} = X_i$, which is also true in absolute value,

so $\mathsf{E}\left(|X_I^{>N}\right) \downarrow 0$.

Fix $\varepsilon > 0$ and let $N = N(\varepsilon)$ large enough that

$$\mathsf{E}\left(\left|X_1^{>N}\right|\right) \leq \frac{\varepsilon^2}{8}.$$

Then, writing

$$S_n^{\leq N} = \sum_{i=1}^{n} X_i^{\leq N}$$

and so the probability for $X_i^{\leq N}$ can be bounded as

$$\mathsf{P}\left(\left|\frac{S_n^{\leq N}}{n} - \mathsf{E}\left(X_1^{\leq N}\right)\right| > \frac{\varepsilon}{2}\right) \leq \frac{\mathsf{Var}\left(X_i^{\leq N}\right)}{\varepsilon/2^2 n} \leq \frac{4N^2}{\varepsilon^2}\frac{1}{n} < \frac{\varepsilon}{2}$$

for n large. Second,

$$
\begin{aligned}
&\mathsf{P}\left(\left|\frac{S_n^{>N}}{n}-\mathsf{E}\left(X_1^{>N}\right)\right|>\frac{\varepsilon}{2}\right)\\
&\quad\leq\frac{2}{\varepsilon}\mathsf{E}\left(\left|\frac{S_n^{>N}}{n}-\mathsf{E}\left(X_1^{>N}\right)\right|\right)\\
&\quad=\frac{2}{n\varepsilon}\mathsf{E}\left(\left|\sum_{i=1}^{n}X_i^{>N}-\mathsf{E}\left(X_1^{>N}\right)\right|\right)\\
&\quad\leq\frac{2}{n\varepsilon}n\mathsf{E}\left(\left|X_i^{>N}-\mathsf{E}\left(X_1^{>N}\right)\right|\right) && (\triangle\text{ineq.})\\
&\quad\leq\frac{2}{\varepsilon}\left(\mathsf{E}\left(\left|X_1^{>N}\right|\right)+\left|\mathsf{E}\left(X_1^{>N}\right)\right|\right) && (\triangle\text{ineq.})\\
&\quad\leq\frac{4}{\varepsilon}2\mathsf{E}\left(\left|X_1^{>N}\right|\right) && (|\mathsf{E}\left(\cdot\right)|\leq\mathsf{E}\left(|\cdot|\right))\\
&\quad\leq\frac{4}{\varepsilon}\frac{\varepsilon^2}{8}\leq\frac{\varepsilon}{2}.
\end{aligned}
$$

Combining these bounds with the $\triangle$ inequality yields

$$\mathsf{P}\left(|S_n/n|>\varepsilon\right)\leq\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon$$

for n large enough. This is enough to yield convergence in probability. ■

Exercise 8.2

Draw the picture of the different events. Convince yourself that when ε decreases, the bound gets better and we get convergence in probability.

8.1 Strong law of large numbers

We now target the strong law of large numbers, proving a weaker statement and building up to a general result.

Theorem 8.3 (Strong law of large numbers)

Let $(X_n, n\geq 1)$ be iid and $X_i\in\mathcal{L}^1$. If $S_n=\sum_{i=1}^{n}X_i$, then $S_n/n\xrightarrow{\text{a.s.}}\mathsf{E}\left(X_i\right)$, that is

$$\mathsf{P}\left(\lim_{n\to\infty}\frac{S_n}{n}\text{ exists, }\lim_{n\to\infty}\frac{S_n}{n}=\mathsf{E}\left(X_i\right)\right)=1$$

Exercise 8.3

Prove the WLLN from the SLLN

We need first a definition:

Definition 8.4 (Lacunary)

A sequence $(n_i, i \geq 1)$ is **lacunary** if $\exists c > 1$ such that for all i sufficiently large, $n_{i+1}/n_i > c$.

Theorem 8.5 (Nonnegative lacunary strong law of large numbers)

In the setup of SLLN, if $X_i \overset{\text{a.s.}}{\geq} 0$, then for any lacunary sequences $(n_k)_{k\geq 1}$, then

$$\mathsf{P}\left(\lim_{k\to\infty}\frac{S_{n_k}}{n_k} \text{ exists}, \lim_{k\to\infty}\frac{S_{n_k}}{n_k}\right) = \mathsf{E}\left(X_i\right)$$

Notation

We will write $\bar{S}_n$ for S_n/n. We write $S_n^{\geq N} = \sum_{i=1}^n X_i^{\geq N}/n$

Proof It suffices to show that $\forall\, \varepsilon > 0$,

$$\sum_{k=1}^{\infty}\mathsf{P}\left(\left|\bar{S}_{n_k} - \mathsf{E}\left(X_i\right)\right| > \varepsilon\right) < \varepsilon \tag{8.1}$$

Since then

$$\begin{aligned}\mathsf{P}\left(\left(\lim_{k\to\infty}\frac{S_{n_k}}{n_k} = \mathsf{E}\left(X_i\right)\right)^{\mathsf{C}}\right) &= \mathsf{P}\left(\bigcup_{m\geq 1}\left\{\left|\bar{S}_{n_k} - \mathsf{E}\left(X_i\right)\right| > \frac{1}{m} \text{ for } \infty \text{ many } k\right\}\right)\\ &\leq \sum_{m\geq 1}\mathsf{P}\left(\left|\bar{S}_{n_k} - \mathsf{E}\left(X_i\right)\right| > \frac{1}{m} \text{ for } \infty \text{ many } k\right) = 0\end{aligned}$$

using subadditivity, (8.1) and the first Borel-Cantelli lemma 1.

So for $\varepsilon > 0$, let i_0 be such that $\mathsf{E}\left(X^{\leq n_{i_0}}\right) \geq \mathsf{E}\left(X_i\right) - \varepsilon$ as $\mathsf{E}\left(X_i^{(>n_{i_0}}\right) < \varepsilon$.[22] Then $\bar{S}_{n_k} = \bar{S}_{n_k}^{\leq n_k} + \bar{S}_{n_k}^{>n_k}$. So for $k \geq i_0$, [23]

[22] For convergence, only the tail of the sequence matters, so we can only look at values past i_0

[23] Taking $A = \{\left|\bar{S}_{n_k} - \mathsf{E}\left(X_i\right)\right| > 2\varepsilon\}$, $B = \{\left|\bar{S}_{n_k}^{\leq n_k} - \mathsf{E}\left(X_i^{\leq n_k}\right)\right| > \varepsilon\}$ and $C = \{\bar{S}_{n_k}^{>n_k} \neq 0\}$ and breaking A into $B \cap C \sqcup B \cap C^{\mathsf{C}} \subset B \cap C^{\mathsf{C}} \sqcup C$.

$$\begin{aligned}
&\mathsf{P}\left(\left|\bar{S}_{n_k} - \mathsf{E}\left(X_i\right)\right| > 2\varepsilon\right) \\
&\quad = \mathsf{P}\left(\bar{S}_{n_k}^{>n_k} \neq 0\right) + \mathsf{P}\left(\left|\bar{S}_{n_k}^{\leq n_k} - \mathsf{E}\left(X_i\right)\right| > 2\varepsilon\right) \\
&\quad \leq \mathsf{P}\left(\bar{S}_{n_k}^{>n_k} \neq 0\right) + \mathsf{P}\left(\left|\bar{S}_{n_k}^{\leq n_k} - \mathsf{E}\left(X_i^{\leq n_k}\right)\right| > \varepsilon\right)
\end{aligned}$$

Next

$$\begin{aligned}
\mathsf{P}\left(\bar{S}_{n_k}^{>n_k} \neq 0\right) &= \mathsf{P}\left(\sum_{i=1}^{n_k} X_i^{>n_k} \neq 0\right) \\
&= \mathsf{P}\left(\bigcup_{i=1}^{n_k} \{X_i^{>n_k} \neq 0\}\right) && \text{(non-negativity)} \\
&\leq \sum_{i=1}^{n_k} \mathsf{P}\left(\{X_i^{>n_k} \neq 0\}\right) && \text{(subadditivity)} \\
&= n_k \mathsf{P}\left(\{X_i^{>n_k} \neq 0\}\right) && \text{(iid)}
\end{aligned}$$

and

$$\begin{aligned}
\mathsf{P}\left(\left|\bar{S}_{n_k}^{\leq n_k} - \mathsf{E}\left(X_i^{\leq n_k}\right)\right| > \varepsilon\right) &\leq \frac{\mathsf{Var}\left(X_i^{\leq n_k}\right)}{\varepsilon^2 n_k} \\
&\leq \frac{\mathsf{E}\left(\left(X_i^{\leq n_k}\right)^2\right)}{\varepsilon^2 n_k} \overset{?}{\leq} \infty
\end{aligned}$$

using Chebyshev inequality.

Let $J = \min\{k : n_k \geq X_i\}$. Then

$$\begin{aligned}
\sum_{k \geq i_0} n_k \mathsf{P}\left(X_i^{>n_k} \neq 0\right) &= \sum_{k \geq i_0} \mathsf{E}\left(n_k \mathbf{1}_{X_i > n_k}\right) \\
&= \mathsf{E}\left(\sum_{k=i_0}^{\infty} n_k \mathbf{1}_{X_i > n_k}\right) \\
&= \mathsf{E}\left(\sum_{k=i_0}^{J-1} n_k \mathbf{1}_{X_i > n_k}\right) \qquad (8.2)
\end{aligned}$$

because $\mathbf{1}_{X_i>n_k}$ is zero for $k \geq J$. Because n_k is lacunary, $n_{J-1} < X_i$, $n_{J-2} < \frac{n_{J-1}}{/}c < X_i/c$, $n_{J-3} \leq X_i/c^2$. So

$$\sum_{k=i_0}^{J-1} n_k \mathbf{1}_{X_i>n_k} < \sum_{k=i_0}^{J-1} \frac{X_i}{c^{J-1-k}} \leq X_i \sum_{i\geq 0} c^{-i} = \frac{X_i}{c-1}$$

So (8.2) $\leq \mathsf{E}\left(X_1/(c-1)\right) < \infty$. Likewise,

$$\begin{aligned}\sum_{k=i_0}^{\infty} \frac{\left(X_i^{<n_k}\right)^2}{n_k} &= \sum_{k=\max(J,i_0)}^{\infty} \frac{\left(X_i^{<n_k}\right)^2}{n_k} \\ &\leq \sum_{k=\max(J,i_0)}^{\infty} \frac{X_i n_J}{n_k} \\ &\leq \sum_{k=\max(J,i_0)}^{\infty} \frac{1}{c^{k-J}} X_i \\ &\leq \frac{X_i}{c-1}\end{aligned}$$

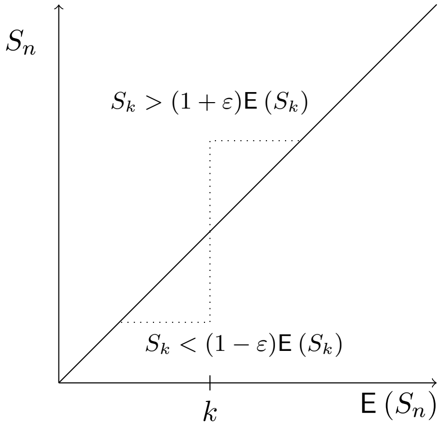
using the bounds $X_i = X_i^{\leq n_k}$ ad $X_i < n_J$ and lacunary in the subsequent step. So

$$\mathsf{E}\left(\sum_{k=i_0}^{\infty} \frac{\left(X_i^{<n_k}\right)^2}{n_k}\right) \leq \mathsf{E}\left(\frac{X_i}{c-1}\right) < \infty.$$

We used Borel-Cantelli to get a finite bound rather than prove the terms went to zero. We used subadditivity and took gaps to bound the sum using the lacunary assumption to replace the sum by summable terms. The variance of the truncated term is finite, so we used Chebychev to replace (ε^2 was a constant). The bounds behave like numbers, we proved statements pointwise for every fixed ω. ■

Proposition 8.6 (Nonnegative strong law of large numbers)
In the setup of the SLLN, if $X - i \overset{\text{a.s.}}{\geq} 0$, then $\bar{S}_n \xrightarrow{\text{a.s.}} \mathsf{E}\left(X_i\right)$

Proof Assume for simplicity that $\mathsf{E}\left(X_1\right) = 1$. A plot of the expected value of the sum $\mathsf{E}\left(S_n\right)$ (the target) against $S_n = n$ is the line $y = x$. Say $S_k > (1+\varepsilon)\mathsf{E}\left(S_k\right)$ for some k, an upper bound for some interval. At $x = k(1+\frac{\varepsilon}{2})$, for every $l \in [k, k(1+\varepsilon/2)$. If we take a lacunary sequence $c = 1+\varepsilon/3$. Fix $\varepsilon > 0$, let $n_k = \lceil(1+\varepsilon)^k\rceil$. By


S_n
$S_k > (1+\varepsilon)\mathsf{E}\left(S_k\right)$
$S_k < (1-\varepsilon)\mathsf{E}\left(S_k\right)$
k
$\mathsf{E}\left(S_n\right)$

the nonnegative lacunary SLLN, $\bar{S}_{n_k} \xrightarrow{\text{a.s.}} 1$ as $k \to \infty$. We show if $|\bar{S}_n - 1| > 2\varepsilon$ for infinitely many n, then $|\bar{S}_{n_k} - 1| > \frac{\varepsilon}{3}$ for infinitely many k.

Assuming this, the theorem follows easily. To see this, fix n large and let k be minimal such that $n_k \geq n$. If $\bar{S}_n \geq 1 + 2\varepsilon$, then $S_n \geq n(1+2\varepsilon)$ so

$$S_{n_k} \geq n(1+2\varepsilon) \geq \frac{n_k - 1}{1+\varepsilon}(1+2\varepsilon) \geq n_k\left(1 + \frac{\varepsilon}{3}\right)$$

for n large enough and ε small. ∎

Exercise 8.4

Prove the corresponding lower bound.

For an increasing series macroscopic fluctuations are witnessed by geometric subsequences.

Proof For the strong law, we can use $X = X^+ - X^-$ where $S_n^+ = \sum_{i=1}^n X_i^+$ and $S_n^- = \sum_{i=1}^n X_n^-$; then $S_n = S_n^+ - S_n^-$ and by the nonnegative SLLN, $\bar{S_n}^+ \xrightarrow{\text{a.s.}} \mathsf{E}\left(X_i^+\right)$,

$\bar{S_n}^- \xrightarrow{\text{a.s.}} \mathsf{E}\left(X_i^-\right) < 0$ and $S_n \xrightarrow{\text{a.s.}} \mathsf{E}\left(X_i^+\right) - \mathsf{E}\left(X_i^-\right) = \mathsf{E}\left(X_i^+ - X_i^-\right) = \mathsf{E}\left(X_i\right)$ using decomposition into positive and negative functions. We have linearity of expectation. ∎

Exercise 8.5

Read and understand the Weierstrass approximation theorem.

8.2 Weirstrass approximation

Theorem 8.7 (Weirstrass approximation)

If f is a continuous function on $[0,1]$ and $\varepsilon > 0$, then there exists a polynomial B such that $\sup_{[0,1]} |B(x) - f(x)| \leq \varepsilon$.

Proof Let X_i be iid binary random variables with expectation p and define $S_n = \sum_{i=1}^n X_i$. Then, since X_i are Bernoulli random variates, it follow that S_n is a Binomial random variable with parameters (n,p) and hence

$$B_n(p) \coloneqq \mathsf{E}\left(fn^{-1}S_n\right) = \sum_{k=0}^{n} f(n^{-1}k)\binom{n}{k}p^k(1-p)^{n-k}$$

B refers to the Bernstein polynomials. f is bounded on the interval $[0,1]$ and thus attains a maximum value so $|f(y)| \leq K \; \forall \; y \in [0,1]$. It is also uniformly continuous, which means $|x-y| \leq \delta$ implies $|f(x) - f(y)| < \frac{1}{2}\varepsilon$. Let $p \in [0,1]$, then

$$|B_n(p) - f(p)| = \left|\mathsf{E}\left(f(n^{-1}S_n)\right) - f(p)\right|.$$

Write $Y_n \coloneqq |f(n^{-1}S_n) - f(p)|$ and $Z_n \coloneqq |n^{-1}S_n - p|$. Then $Z_n < \delta$ implies that $Y_n < \frac{1}{2}\varepsilon$ and thus

$$\begin{aligned}
|B_n(p) - f(p)| &\leq \mathsf{E}\left(Y_n\right) \\
&= \mathsf{E}\left(Y_n; Z_n \leq \delta\right) + \mathsf{E}\left(Y_n; Z_n > \delta\right) \\
&\leq \frac{1}{2}\varepsilon \mathsf{P}\left(Z_n \leq \delta\right) + 2K\mathsf{P}\left(Z_n > \delta\right) \\
&\leq \frac{1}{2}\varepsilon + \frac{2K}{4n\delta^2}
\end{aligned}$$

since from Chebyshev's inequality

$$\mathsf{P}\left(|n^{-1}S_n - p| > \delta\right) \leq \frac{1}{4n\delta^2}$$

Choosing n such that $K/(2n\delta^2) < \frac{1}{2}\varepsilon$ gives us $|B_n(p) - f(p)| < \varepsilon$ for all $p \in [0,1]$. ■

Chapter 9

Product spaces, multiple integrals and Fubini theorem

9.1 Product space

We will work in this section in finite measurable spaces, but everything goes through to σ-finite spaces.

Definition 9.1 (Product space)

If $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ are measurable spaces, we define the product space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$ by

$$\begin{aligned} \Omega_1 \times \Omega_2 &= \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\} \\ \mathcal{F}_1 \times \mathcal{F}_2 &= \sigma(\{E_1 \times E_2 : E_1 \in \mathcal{F}_1, E_2 \in \mathcal{F}_2\}) \end{aligned}$$

and $\{E_1 \times E_2 : E_1 \in \mathcal{F}_1, E_2 \in \mathcal{F}_2\}$ is a π-system.[24]

We have seen an example of product space before when working with random walks with step ±1. Rather than working with rectangles, we were using the cylinder map. In general, for uncountable products, we want to work with the latter. This motivates the following

Note

$$\mathcal{F}_1 \times \mathcal{F}_2 = \sigma(\text{cylinders}) = \sigma(\{E_1 \times \Omega_2 : E_1 \in \mathcal{F}_1\} \cup \{\Omega_1 \times E_2 : E_2 \in \mathcal{F}_2\})$$

More generally, for $\{(\Omega_i, \mathcal{F}_i), i \in I\}$, where $\mathcal{F}_i$ are σ-algebras, define the product σ-algebra by

$$\prod_{i \in I} \mathcal{F}_i = \sigma\left(\bigcup_{i \in I} \left\{\left\{\overline{\omega} \in \prod_{j \in I} \Omega_j : \omega_i \in E_i\right\}, E_i \in \mathcal{F}_i\right\}\right)$$

and denote by $\Omega \coloneqq \prod_{j \in I} \Omega_j$. For I countable, this is the **smallest** σ-algebra such that given functions $f_i : \Omega_i \to \mathbb{R}$, if each f_i is $\mathcal{F}_i - \mathbb{B}(\mathbb{R})$ measurable, then the

[24] $\mathcal{F}_1 \times \mathcal{F}_2$ is not in general a σ-algebra, so we mean by this the closure of this set. This abuse of notation is common.

function $f : \Omega \to \mathbb{R}$, $f((\omega_j, j \in I)) = f_i(\omega_i)$ is measurable for all $i \in I$.

Example 9.1

Consider $\mathbb{R}^n$ with $\mathbb{B}(\mathbb{R}^n) = \sigma(\text{open sets in } \mathbb{R}^n)$ versus $\mathbb{B}^n(\mathbb{R}) = \mathbb{B}(\mathbb{R}) \times \mathbb{B}(\mathbb{R}) \times \cdots \times \mathbb{B}(\mathbb{R})$ which is equal to $\sigma(\text{rectangles in } \mathbb{R}^n)$. For any open set $A \subset \mathbb{R}^n$, write

$$A = \bigcup_{q \in \mathbb{Q}^n \cap A} \bigcup_{\substack{r \in \mathbb{Q}, r > 0 \\ B_q(r) \in A}} B_q(r)$$

using separability of $\mathbb{R}^n$ (it has a countable dense subset, namely $\mathbb{Q}^n$ is dense in $\mathbb{R}^n$) so $\mathbb{B}(\mathbb{R}^n) \subset \mathbb{B}(\mathbb{R})^n$.

Exercise 9.1

Show that $\mathbb{B}(\mathbb{R})^n \subset \mathbb{B}(\mathbb{R}^n)$. More generally, this works on finite products of separable metric spaces. In $\mathcal{L}^\infty$ metric, the circle is a square; we can use the fact that all $\mathcal{L}^p$ distances give rise to the same topology, so this is clear from that point of view.

Proposition 9.2

If $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$, f is $\mathcal{F}_1 \times \mathcal{F}_2$ measurable, then for all $\omega_1 \in \Omega_1$, $f(\omega_1, \bullet) : \Omega_2 \to \mathbb{R}$ is $\mathcal{F}_2$-measurable and for all $\omega_2 \in \Omega_2$, then $f(\bullet, \omega_2) : \Omega_1 \to \mathbb{R}$ is $\mathcal{F}_1$-measurable. If f satisfies this property, say f is "good".

To prove the above, we will use the Monotone class theorem.

Suppose $(\Omega, \mathcal{F})$ is a measure space and let $\mathcal{H}$ be a collection of functions $f : \Omega \to \mathcal{F}$ and $\mathcal{P}$ is a π-system with $\sigma(\mathcal{P}) \in \mathcal{F}$. If

- $\mathbf{1}_f \in \mathcal{H}$ for all $f \in \mathcal{P}$
- $\mathcal{H}$ is a real vector space, $f, g \in \mathcal{H}$ implies $f + g \in \mathcal{H}$
- $\mathcal{H}$ closed under bounded monotone limits, then

$$\forall \text{ bounded } f \in \Omega \to \mathbb{R}, f \text{ is } \mathcal{F}\text{-measurable}, f \in \mathcal{H}$$

Proof Let $\mathcal{H} = \{f : \Omega_1 \times \Omega_2 \to \mathbb{R} : f\text{“good”}\}$. Let

$$\mathcal{P} = \{E_1 \times E_2 : E_1 \in \mathcal{F}_1, E_2 \in \mathcal{F}_2\}.$$

Then for $f = \mathbf{1}_{E_1 \times E_2}$ and $\omega_1 \in \Omega_1$. There are two cases for $f(\omega)1, \bullet) = \mathbf{1}_{(\omega_1, \bullet) \in E_1 \times E_2}$. If $\omega_1 \in E_1$, then $f = \mathbf{1}_{E_2}$, while if $\omega_1 \notin E_1$, then $f \equiv 0$. Applying a symmetric argument to $f(\bullet, \omega_w)$ shows f is "good". So $\mathbf{1}_E \in \mathcal{H}$ for all $E \in \mathcal{P}$.

Next, $\mathcal{H}$ is a vector space because if $h = cf + g$, for $f, g : \Omega_1 \times \Omega_2 \to \mathbb{R}$, then $h(\omega_1, \bullet) = cf(\omega_1, \bullet) + g(\omega_1, \bullet)$, linearity combinations are also measurable in the one-dimensional setting using closure properties established before, so $f(\omega_1, \bullet)$ and $g(\omega_1, \bullet)$ are $\mathcal{F}_2$-measurable implies that $h(\omega_1, \bullet)$ is as well and any slice will be. For any slice, if the surface is increasing, then the picture is clear: this step is left to the reader.

Exercise 9.2

Show "goodness" is preserved under bounded monotone limits.

By the monotone class theorem, it follows that all bounded $\mathcal{F}_1 \times \mathcal{F}_2$-measurable functions $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ are "good".

For $f \geq 0$, write $f = \lim_{n \to \infty} \min(f, n)$ and redo the monotonicity argument. For the general case, write $f = f^+ - f^-$ and use vector space property (linearity). ■

9.2 Multiple integrals

Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be measurable spaces, write $\Omega, \mathcal{F}) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$ and let $f : \Omega \to \mathbb{R}$ be a bounded $\mathcal{F}$-measurable function.

We want to define

$$\iint f \,\mathrm{d}\mu_1 \,\mathrm{d}\mu_2$$

we need to decide which to fix one coordinate and integrate over the second. We need to know that $\int f(x, \omega_2) \,\mathrm{d}\mu_1(x)$ is $\mathcal{F}_2$-measurable. Similarly, if we define $\iint f \,\mathrm{d}\mu_2 \,\mathrm{d}\mu_1$, we want $\int f(\omega_1, y) \,\mathrm{d}\mu_2(y)$ is $\mathcal{F}_1$-measurable. Call a function satisfying both of these "satisfying".

Proposition 9.3

Bounded $\mathcal{F}$-measurable functions are "satisfying".

Proof We use again the monotone class theorem. If $f = \mathbf{1}_{E_1 \times E_2}$, then

$$f_2(\mu_2) = \int \mathbf{1}_{E_1 \times E_2}(x, \omega_2) \,\mathrm{d}\mu_1(x) = \mu_1(E_1) \mathbf{1}_{E_2}(\omega_2)$$

which is $\mathcal{F}_2$-measurable. Likewise, the function

$$f_1(\omega_1) = \int \mathbf{1}_{E_1 \times E_2}(\omega_1, y)\,\mathrm{d}\mu_2(y)$$

is $\mathcal{F}_1$-measurable. Next, if $h = c_f + g$ and f, g are "satisfying", then

$$\begin{aligned}\int h(x,\omega)2\,\mathrm{d}\mu_1(x) &= \int \left[cf(x,\omega_2) + g(x,\omega_2)\right]\,\mathrm{d}\mu_1(x) \\ &= c\int f(x,\omega_2)\,\mathrm{d}\mu_1(x) + \int g(x,\omega_2)\,\mathrm{d}\mu_1(x)\end{aligned}$$

Since f, g are "satisfying", this is a linear combination of measurable functions so measurable. A symmetric argument shows $\int h(\omega_1, y)\,\mathrm{d}\mu_2(y)$ is $\mathcal{F}_2$-measurable, so h is satisfying. Finally, if $h_n \uparrow h$ bounded, and h_n are "satisfying", we can assume h_n is positive by adding the lower bound. Then by the MCT in the first coordinate pointwise for ω_2

$$\int h(x,\omega_2)\,\mathrm{d}\mu_1(x) = \lim_{n\to\infty} \int h_n(x,\omega_2)\,\mathrm{d}\mu_1(x)$$

for all $\omega_2 \in \Omega_2$. So the LHS is an increasing limit of $\mathcal{F}_2$ measurable functions so is itself $\mathcal{F}_2$-measurable. The proposition follows. ■

Insofar, we have constructed $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$, saw that $\mathbb{B}^n(\mathbb{R}) = \mathbb{B}(\mathbb{R}^n)$ and defined iterated integrals $\int\int f\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2$ for $f : \Omega_1 \times \Omega_2$ bounded and $\mathcal{F}_1 \times \mathcal{F}_2$ measurable.

Definition 9.4 (Product measure)

Given measurable spaces $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$, define the product measure $\mu = \mu_1 \times \mu_2$ as follows: for a **rectangle** $E = E_1 \times E_2 \in \mathcal{F}_1 \times \mathcal{F}_2$, the σ(rectangle), let $\mu(E) = \mu_1(E_1) \cdot \mu_2(E_2)$.

Claim

μ is a pre-measure on the algebra of finite disjoint unions of rectangles. If this holds, then μ extends uniquely to a measure on $\mathcal{F}_1 \times \mathcal{F}_2$ (we still assume μ_1, μ_2 are finite measures).

Note

$$\mu\left(\bigcup_{i=1}^{n} E_{i,1}\times E_{i,2}\right)=\sum_{i=1}^{n}\mu\left(E_{i,1}\times E_{i,2}\right)$$

Proof By monotonicity,

$$E=\bigcup_{i=1}^{n} E_{i,1}\times E_{i,2}\subset\bigcup_{j=1}^{m} F_{j,1}\times F_{j,2}=F$$

One way to prove this uses the fact that $\bigcup_{j=1}^{m} F_{j,1}\times F_{j,2}\setminus\left(\bigcup_{i=1}^{n} E_{i,1}\times E_{i,2}\right)$ is a finite disjoint union of rectangles. We proceed however with integrals.

Using monotonicity, then $\mathbf{1}_E\leq\mathbf{1}_F$ and $\mathbf{1}_E,\mathbf{1}_F$ are $\mathcal{F}_1\times\mathcal{F}_2$ measurable, so

$$\int\left(\int\mathbf{1}_E\,\mathrm{d}\mu_1\right)\mathrm{d}\mu_2\leq\int\left(\int\mathbf{1}_F\,\mathrm{d}\mu_1\right)\mathrm{d}\mu_2$$

since $\mathbf{1}_E<\mathbf{1}_F$ for all ω_2, thus $\int\mathbf{1}_E(\bullet,\omega_2)\,\mathrm{d}\mu_1$ and now integrate over Ω_2.

but $\mathbf{1}_E=\sum_{i=1}^{n}\mathbf{1}_{E_{i,1}\times E_{i,2}}$ so for every $\omega_2\in\Omega_2$,

$$\begin{aligned}\int\mathbf{1}_E(\bullet,\omega_2)\,\mathrm{d}\mu_1&=\int\sum_{i=1}^{n}\mathbf{1}_{E_{i,1}\times E_{i,2}}(\bullet,\omega_2)\,\mathrm{d}\mu_1\\&=\sum_{i=1}^{n}\int\mathbf{1}_{E_{i,1}\times E_{i,2}}(\bullet,\omega_2)\,\mathrm{d}\mu_1\end{aligned}$$

Now integrate with respect to ω_2 :

$$\begin{aligned}\iint\mathbf{1}_E\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2&=\int\left(\sum_{i=1}^{n}\int\mathbf{1}_{E_{i,1}\times E_{i,2}}(\bullet,\omega_2)\,\mathrm{d}\mu_1\right)\mathrm{d}\mu_2\\&=\sum_{i=1}^{n}\left(\int\mathbf{1}_{E_{i,1}\times E_{i,2}}(\bullet,\omega_2)\,\mathrm{d}\mu_1\right)\mathrm{d}\mu_2\\&=\sum_{i=1}^{n}\int\mu_1(E_{i,1})\mathbf{1}_{E_{i,2}}\,\mathrm{d}\mu_2\\&=\sum_{i=1}^{n}\mu_1(E_{i,1})\cdot\mu_2(E_{i,2})\end{aligned}$$

$$= \mu(E)$$

using linearity and additivity proven last class.

The same logic shows $\iint \mathbf{1}_F \, \mathrm{d}\mu_1 \, \mathrm{d}\mu_2 = \mu(F)$ so $\mu(E) \leq \mu(F)$. ■

We thus have the following:

- ○ finite additivity: exercise (apply the definition)
- ○ countable consistency: if E is a (finite union) of rectangle(s) and also $E = \bigcup_{i\geq 1} A_i \times B_i$ then $\mu(E) = \sum_{i\geq 1} \mu(A_i \times B_i)$ disjoint. To prove this, note that $\mu(E) = \iint \mathbf{1}_E \, \mathrm{d}\mu_1 \, \mathrm{d}\mu_2$ and also, for each n,

$$\mu\left(\bigcup_{i=1}^n A_i \times B_i\right) = \iint \mathbf{1}_{\bigcup_{i=1}^n A_i \times B_i} \, \mathrm{d}\mu_1 \, \mathrm{d}\mu_2.$$

But $\mathbf{1}_{\bigcup_{i=1}^n A_i \times B_i} \uparrow \mathbf{1}_E$ as $n \to \infty$. So by the monotone convergence theorem, we have pointwise convergence for each ω_2

$$\int \mathbf{1}_{\bigcup_{i=1}^n A_i \times B_i}(\bullet, \omega_2) \, \mathrm{d}\mu_1 \uparrow \int \mathbf{1}_E(\bullet, \omega_2) \, \mathrm{d}\mu_1$$

Now, applying MCT a second time gives

$$\lim_{n\to\infty} \iint \mathbf{1}_{\bigcup_{i=1}^n A_i \times B_i} \, \mathrm{d}\mu_1 \, \mathrm{d}\mu_2 = \iint \mathbf{1}_E \, \mathrm{d}\mu_1 \, \mathrm{d}\mu_2$$

so

$$\begin{aligned}\mu(E) &= \sum_{i=1}^n \lim_{n\to\infty} \iint \mathbf{1}_{\bigcup_{i=1}^n A_i \times B_i} \, \mathrm{d}\mu_1 \, \mathrm{d}\mu_2 \\ &= \lim_{n\to\infty} \sum_{i=1}^n \lim_{n\to\infty} \iint \mathbf{1}_{A_i \times B_i} \, \mathrm{d}\mu_1 \, \mathrm{d}\mu_2 \\ &= \lim_{n\to\infty} \sum_{i=1}^n \mu(A_i \times B_i) \\ &= \sum_{i\geq 1} \mu(A_i \times B_i).\end{aligned}$$

Thus μ is a pre-measure, so extends uniquely to a measure on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$ since $\mathcal{F}_1 \times \mathcal{F}_2$ is generated by finite disjoint unions of rectangles.

This is the unique measure on $\mathcal{F}_1 \times \mathcal{F}_2$ for which the integral $\int \mathbf{1}_{A\times B}\,\mathrm{d}\mu = \mu_1(A) \cdot \mu_2(B) = \iint \mathbf{1}_{A\times B}\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2$. This agrees with our rectangle area of base times height if $A \times B$ is a rectangle. We have also checked iterated integrals for indicators, the first step in Fubini's theorem which we tackle next.

Theorem 9.5 (Fubini)

Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be σ-finite measure spaces,

and $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ are $\mathcal{F}_1 \times \mathcal{F}_2$ measurable functions. If $f \in \mathcal{L}^1(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, \mu_1 \times \mu_2)$ or $f \geq 0$, then

$$\iint f\,\mathrm{d}\mu_2\,\mathrm{d}\mu_1 = \int f\,\mathrm{d}\mu = \iint f\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2.$$

Proof Assume μ_1, μ_2 are finite.[25] Apply the monotone class theorem to

$$\mathcal{H} = \left\{ f \text{ bounded, product measurable, } \iint f\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2 = \int f\,\mathrm{d}\mu = \iint f\,\mathrm{d}\mu_2\,\mathrm{d}\mu_1 \right\}$$

Then

- $\mathbf{1}_{A\times B} \in \mathcal{H}$ proven already.
- $f, g \in \mathcal{H} \Rightarrow cf + g \in \mathcal{H}$ by linearity of integration.
- if $f_n \geq 0, f_n \in \mathcal{H}, f_n \uparrow f$ bounded, then

$$\int f\,\mathrm{d}\mu = \lim_{n\to\infty} \int f_n\,\mathrm{d}\mu = \lim_{n\to\infty} \iint f_n\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2 = \iint \lim_{n\to\infty} f_n\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2 = \iint f\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2$$

using monotone convergence theorem three times. Likewise, $\int f\,\mathrm{d}\mu = \iint f\,\mathrm{d}\mu_2\,\mathrm{d}\mu_1$

Thus, $\mathcal{H}$ contains all bounded measurable functions, $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$.

This is not general enough. For Tonelli's theorem, if $f \geq 0$, then we use MCT three more times. If $f \geq 0$, then $f = \lim_{n\to\infty} f_n$ where $f_n = \min(f, n)$ so $\int f\,\mathrm{d}\mu = \lim_{n\to\infty} \int f_n\,\mathrm{d}\mu = \lim_{n\to\infty} \iint f_n\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2$ since $f_n \in \mathcal{H}$. Apply MCT twice more to write this as $\iint f\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2$ and likewise, $\int f\,\mathrm{d}\mu = \iint f\,\mathrm{d}\mu_2\,\mathrm{d}\mu_1$.

For $f \in \mathcal{L}^1$, then $\int f^+\,\mathrm{d}\mu, \int f^-\,\mathrm{d}\mu < \infty$ and so

$$\int f\,\mathrm{d}\mu = \int f^+\,\mathrm{d}\mu - \int f^-\,\mathrm{d}\mu$$

[25] For linearity, say with $\int (f-1)\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu - \int \mathrm{d}\mu$ and we could have to subtract ∞ from ∞.

$$
\begin{aligned}
&= \iint f^+ \,\mathrm{d}\mu_1 \,\mathrm{d}\mu_2 - \iint f^- \,\mathrm{d}\mu_1 \,\mathrm{d}\mu_2 \\
&= \iint f \,\mathrm{d}\mu_1 \,\mathrm{d}\mu_2
\end{aligned}
$$

with two applications of linearity, to recombine the inner integral. ■

Corollary 9.6

For a non-negative random variable X,

$$\mathsf{E}\left(X\right) = \int_0^\infty \mathsf{P}\left(X > x\right) \,\mathrm{d}x$$

A special case: if X is an integer valued random variable, then

$$
\begin{aligned}
\mathsf{E}\left(X\right) &= \sum_{i=0}^\infty \mathsf{P}\left(X > i\right) \\
&= \sum_{i\geq 0} \sum_{j>i} \mathsf{P}\left(X = j\right) \\
&= \sum_{k\geq 0} k\mathsf{P}\left(X = k\right)
\end{aligned}
$$

Proof The idea is that

$$
\begin{aligned}
\mathsf{P}\left(X > x\right) = \mathsf{E}\left(\mathbf{1}_{X>x}\right) &= \int \mathbf{1}_{X\in(x,\infty)} \,\mathrm{d}\mathsf{P} \\
&= \int \mathbf{1}_{(x,\infty)} \,\mathrm{d}\mu_X
\end{aligned}
$$

By Fubini,

$$
\begin{aligned}
\iint \mathbf{1}_{(x,\infty)}(y) \,\mathrm{d}\mu_X(y) \,\mathrm{d}x &= \iint_0^\infty \mathbf{1}_{(x,\infty)}(y) \,\mathrm{d}x \,\mathrm{d}\mu_X \\
&= \int y \,\mathrm{d}\mu_X(y) = \int X \,\mathrm{d}\mathsf{P} = \mathsf{E}\left(X\right).
\end{aligned}
$$

Recall change of variables, we had

$$\mathsf{E}\left(f(Z)\right) = \int f(z) \,\mathrm{d}\mathsf{P} \,(=) \int f(u) \,\mathrm{d}\mu_Z(u)$$

where in our case $f(u) = u$. To justify this, we use the product space $[0,\infty) \times [0,\infty)$,

$\mathbb{B}(\mathbb{R}) \times \mathbb{B}(\mathbb{R}), \lambda \times \mu_X$. ■

Corollary 9.7

If $\iint |f| \,\mathrm{d}\mu_1 \,\mathrm{d}\mu_2 < \infty$, then $f \in \mathcal{L}^1$ and so Fubini holds.

Proof The function $|f| \geq 0$ so by Fubini, $\int |f| \,\mathrm{d}\mu = \iint |f| \,\mathrm{d}\mu_1 \,\mathrm{d}\mu_2 < \infty$, so $f \in \mathcal{L}^1$. ■

Some warnings regarding the use of Fubini's theorem:

Example 9.2

Consider the spaces $(\Omega_1, \mathcal{F}_1, \mu_1) = ([0,1], \mathbb{B}([0,1]), \lambda)$ where λ is the Lebegues measure and $(\Omega_2, \mathcal{F}_2, \mu_2) = ([0,1], \mathbb{B}([0,1]), \zeta)$; the latter measure $\zeta(S) = \#$ elements in S (the counting measure, which is not σ-finite). Let $D = \{(x,x) : x \in [0,1]\}$ and let $f = \mathbf{1}_D \geq 0$. We compute the iterated integral:

$$\iint f \,\mathrm{d}\mu_1 \,\mathrm{d}\mu_2 = \int \left(\int \mathbf{1}_D(\bullet, y) \,\mathrm{d}\lambda(x) \right) \mathrm{d}\zeta(y) = \int 0 \,\mathrm{d}\zeta(y) = 0.$$

while

$$\iint f \,\mathrm{d}\mu_2 \,\mathrm{d}\mu_1 = \int \left(\int \mathbf{1}_D(x, \bullet) \,\mathrm{d}\zeta(y) \right) \mathrm{d}\lambda(x) = \int 1 \,\mathrm{d}\lambda(y) = 1.$$

Example 9.3

Let $(\Omega_1, \mathcal{F}_1, \mu_1) = (\Omega_2, \mathcal{F}_2, \mu_2) = (\mathbb{N}, 2^{\mathbb{N}}, \zeta)$ and let

$$fm,n) = \begin{cases} 1 & \text{if } n = m \\ -1 & \text{if } n = m+1 \\ 0 & \text{otherwise} \end{cases}$$

This can be represented by a matrix. First note that

$$\iint |f| \,\mathrm{d}\mu_1 \,\mathrm{d}\mu_2 = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} |f(m,n)| = \infty$$

so $f \notin \mathcal{L}^1$. If we compute the iterated integrals, we obtain summing over columns,

$$\begin{array}{ccccccc}
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
0 & 0 & 0 & 0 & 1 & -1 & \cdots \\
0 & 0 & 0 & 1 & -1 & 0 & \cdots \\
0 & 0 & 1 & -1 & 0 & 0 & \cdots \\
0 & 1 & -1 & 0 & 0 & 0 & \cdots
\end{array}$$

then rows

$$\begin{aligned}
\iint f \,\mathrm{d}\mu_1 \,\mathrm{d}\mu_2 &= \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} f(m,n) \\
&= 1 + \sum_{m=2}^{\infty}\sum_{n=1}^{\infty} f(m,n) \\
&= 1
\end{aligned}$$

while if we sum first over row, then over columns

$$\begin{aligned}
\iint f \,\mathrm{d}\mu_2 \,\mathrm{d}\mu_1 &= \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} f(m,n) \\
&= 0
\end{aligned}$$

9.3 Joint laws and joint densities

Definition 9.8 (Joint law)

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ be random variables. The **joint law** of X, Y, denoted $\mu_{X,Y}$ is a probability measure on $(\mathbb{R}^2, \mathbb{B}(\mathbb{R}^2))$ with

$$\mu_{X,Y}(E) = \mathsf{P}\left((X,Y) \in E\right)$$

Example 9.4

Let $X = Y \sim \mathcal{N}(0,1)$. Then $\mu_{X,Y}(E) = \mu_{\mathcal{N}(0,1)}(\pi_1(E \cap D))$, since

$$\mathsf{P}\left((X,Y) \in E\right) = \mathsf{P}\left((X,Y) \in E \cap D\right) = \mathsf{P}\left(X \in \{x : (x,x) \in E \cap D\}\right)$$

where π stands for projection.

Example 9.5

Let X, Y be independent $\mathcal{N}(0,1)$ random variables. Then what is $\mu_{X,Y}(E)$? Suppose first that $E = A \times B$ is a rectangle. Then

$$\begin{aligned}\mu_{X,Y}(E) &= \mathsf{P}\left((X,Y) \in A \times B\right) \\ &= \mathsf{P}\left(X \in A, Y \in B\right) \\ &= \mathsf{P}\left(X \in A\right)\mathsf{P}\left(Y \in B\right) && \text{(independence)} \\ &= \mu_X(A)\mu_Y(B) \\ &= (\mu_X \times \mu_Y)(A \times B).\end{aligned}$$

So $\mu_{X,Y}$ is a probability measure agreeing with $\mu_X \times \mu_Y$ on rectangles, so by the π-system lemma, $\mu_{X,Y} = \mu_X \times \mu_Y$ so for $E \in \mathbb{B}(\mathbb{R}^2)$, then $\mu_{X,Y}(E) = \iint \mathbf{1}_E(x,y)\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2$. Since X has a density, we can write as

$$\begin{aligned}\mu_{X,Y}(E) &= \iint \mathbf{1}_E(x,y)\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x\,\mathrm{d}\mu_Y \\ &= \int\left(\int \mathbf{1}_E(x,y)\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x\right)\frac{e^{-y^2/2}}{\sqrt{2\pi}}\,\mathrm{d}y \\ &= \iint_E \frac{e^{-(x^2+y^2)/2}}{2\pi}\,\mathrm{d}x\,\mathrm{d}y\end{aligned}$$

This works in more generality if the two random variables with a given law.

Proposition 9.9

Let $(\Omega, \mathcal{F}, \mathsf{P})$ and $X, Y : \Omega \to \mathbb{R}$. Then $\mu_{X,Y} = \mu_X \times \mu_Y \Leftrightarrow X \perp\!\!\!\perp Y$.

Proof The direction ($\Leftarrow$) is the same as in the example. For the other case, ($\Rightarrow$), if $\mu_{X,Y} = (\mu_X \times \mu_Y)$, then for any Borel $A, B \in \mathbb{B}(\mathbb{R})$, then

$$\begin{aligned}\mathsf{P}\left(X \in A, Y \in B\right) &= \mathsf{P}\left((X,Y) \in A \times B\right) = \mu_{X,Y}(A \times B) \\ &= (\mu_X \times \mu_Y)(A \times B) = \mu_X(A) \cdot \mu_Y(B) = \mathsf{P}\left(X \in A\right)\mathsf{P}\left(Y \in B\right).\end{aligned}$$

■

Definition 9.10

A non-negative measurable function $f : \mathbb{R}^2 \to [0, \infty)$ is a **joint density** for (X, Y)

if for all $E \in \mathbb{B}(\mathbb{R}^2)$, we have

$$\mathsf{P}\left((X,Y) \in E\right) = \iint_E f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Proposition 9.11

If $X \perp\!\!\!\perp Y$ are independent random variables with densities f_X, f_Y, then $f(x,y) = f_X(x)f_Y(y)$ is the joint density for X, Y. Conversely, if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ is a joint density for (X,Y), then X and Y are independent.

Exercise 9.3

If $f_{X,Y}$ is a joint density for (X,Y), then setting

$$f_X(x) = \int f_{X,Y}(x,y)\,\mathrm{d}y, \qquad f_Y(y) = \int f_{X,Y}(x,y)\,\mathrm{d}x$$

then f_X and f_Y are densities for X, Y. Integrate $B \times \mathbb{R}$

9.4 n-fold and ∞ products

If $(\Omega_i, \mathcal{F}_i, \mu_i), i \geq 1$ are probability spaces, then we can define $\prod_{i=1}^n(\Omega_i, \mathcal{F}, \mu_i)$ either by setting $\mu(A_1 \times \cdots \times A_n) = \prod_{i=1}^n \mu_i(A_i)$ or by repeatedly forming 2-dimensional product spaces; the result is the same.

Again, if $X_1, \ldots, X_n$ are random variables, their joint law $\mu_{X_1,\ldots,X_n} = \mu_1 \times \mu_2 \times \cdots \times \mu_{X_n}$ if and only if $X_1, \ldots, X_n$ are mutually independent and Fubini holds under the same conditions and any of the $n!$ iterated integrals are equal.

To construct the infinite product space, measure, we let $\Omega \prod_{i \geq 1} \Omega_i$, $\mathcal{F} = \sigma(\text{cylinders})$, these are sets of the form $A_1 \times \cdots \times A_n \times \Omega_{n+1} \times \cdots$ for $n \geq 1$ and $A_i \in \mathcal{F}_i$. Here, μ is determined by the rule

$$\mu(A_1 \times \cdots \times A_n \times \Omega_{n+1} \times \cdots) = \prod_{i=1}^n \mu_i(A_i)$$

You need to check that μ is a premeasure on Ω with the algebra $\mathcal{A}$ given by finite unions of cylinders.[26]

[26] To prove these technicalities, you need the property that product of compact spaces are compact, but Tychonoff requires the axiom of choice. Here, we need something weaker.

A key fact is that if $(\Omega_i, \mathcal{F}_i) = (\mathbb{R}, \mathbb{B}(\mathbb{R}))$ for all i, then setting $X_i(\omega) = \omega_i$ to be the i^{th} coordinate map, this gives us an infinite sequence of random variables $(X_i, i \geq 1)$ which are independent with laws $(\mu_i, i \geq 1)$ since for any $n \geq 1$ and any Borel set $A_1, \dots, A_n \in \mathbb{B}(\mathbb{R})$, the product measure

$$\begin{aligned}
\mu(X_1 \in A_1, \dots, X_n \in A_n) &= \mu(\{\omega_1 \in A_1\}, \cdots, \{\omega_n \in A_n\}) \\
&= \mu(A_1 \times \cdots A_n \times \mathbb{R} \times \cdots) \\
&= \prod_{i=1}^{n} \mu_i(A_i) \\
&= \prod_{i=1}^{n} \mu(X_i \in A_i).
\end{aligned}$$

If we have $(\mathbb{R}^{\mathbb{N}}, \mathbb{B}(\mathbb{R}^{\mathbb{N}}))$, [27] this space is generated by

$$\bigcup_{n \geq 1} \{A_1 \times \cdots \times A_n \times \mathbb{R} \times \mathbb{R} \times \cdots, \text{ where } A_1, \dots A_n \in \mathbb{B}(\mathbb{R})\};$$

In summary, if $(\mathbb{R}, \mathbb{B}(\mathbb{R}), \mu_i)$ are probability spaces, $(\Omega, \mathcal{F}, \mathsf{P}) = \prod_{i \geq 1}(\mathbb{R}, \mathbb{B}(\mathbb{R}), \mu_i)$ and $X_i(\omega_i) = \omega_i$, then the X_i are independent with laws μ_i

For next week material, Chapter 4 from Durrett, *Probability theory and examples*.

[27] For general $\mathbb{R}^{\infty}$,one would need to define a topology for $\mathbb{R}^{\infty}$. It is not clear how to do this, since defining open sets in an infinite product spaces.

Chapter 10
Random walks

Let $\{X_n; n \geq 1\}$ are iid real valued random variables and consider $S_n = \sum_{i=1}^{n} X_i$. One may ask what are $\mathsf{P}\left(S_n \in B \text{ i.o.}\right)$ equal to for some $B \in \mathbb{B}(\mathbb{R})$ and what is $\mathsf{P}\left(\limsup_{n\to\infty} S_n/c_n \geq 1\right)$ equal to for c_n sequence of constants. We start with a formal construction of the underlying probability space $(\Omega, \mathcal{F}, \mathsf{P}) = \bigotimes_{n=1}^{\infty}(\mathbb{R}, \mathbb{B}(\mathbb{R}), \mu)$ where μ denotes the aw of X_i. Under this construction, $X_n(\omega) = \omega_n$ for some $\omega = (\omega_1, \omega_2, \dots) \in \Omega$.

Definition 10.1 (Exchangeable)

An event $A \in \mathcal{F} = \sigma(X_n; n \geq 1)$ is **exchangeable** if for all **finite** permutation π

$$[(X_1, X_2, \dots) \in B] = [X_{\pi_1}, X_{\pi_2}, \dots) \in B]$$

for $A = [(X_1, X_2, \dots) \in B], B \in \mathbb{B}(\mathbb{R}^{\mathbb{N}})$.

Remark

$[S_n \in B \text{ i.o.}]$ and $\limsup_{n\to\infty} S_n/c_n \geq 1]$ are exchangeable. This is because, for example,

$S_n = X_1 + X_2 + \cdots + X_n = X_2 + X_1 + \cdots + X_n.$

Exercise 10.1

Verify that $\mathcal{E} = \{A \in \mathcal{F}; A \text{ is exchangeable}\}$ is a σ-algebra.

Theorem 10.2 (Hewitt-Savage 0-1 law)

For all $A \in \mathcal{E}$, $\mathsf{P}\left(A\right) \in \{0, 1\}$.

We begin with a heuristic argument. Suppose we have an event $A \in \mathcal{E}$ and assume without loss of generality that this event depends on only finitely many coordinates, that is

$$A = [(X_1, \dots, X_n) \in B_n] \quad \text{for some }_n \in \mathbb{B}(\mathbb{R}^n). \tag{10.3}$$

We consider the permutation

$$\pi(j) = \begin{cases} j+n & \text{if } 1 \leq j \leq n \\ j-n & \text{if } n+1 \leq k \leq 2n \\ j & \text{if } j \geq 2n+1. \end{cases}$$

So using exchangeability,

$$A = [X_{\pi_1}, X_{\pi_2}, \ldots) \in B] = [X_{n+1}, X_{n+2}, \ldots) \in B] \tag{10.4}$$

So we have

$$A = [(X_1, \ldots, X_n) \in B_n] = [X_{n+1}, X_{n+2}, \ldots) \in B]$$

by (10.3) and (10.4), we have $\mathsf{P}(A) \in \{0, 1\}$ by independence. The only possibility for such A is $A = \Omega$, by exchangeability ...We will replace the exact equality $A = [(X_1, \ldots, X_n) \in B_n]$ by an approximating equality. The key step for this approximation is the following

Lemma 10.3 (Approximation)
Suppose that $\mathcal{G} \in \mathcal{F}$ is an algebra and that $\sigma(\mathcal{G}) = \mathcal{F}$. Then

$$\mathcal{F} = \{A \in \mathcal{F};\ \forall\, \varepsilon > 0, \exists B \in \mathcal{G} \text{ such that } \mathsf{P}(A \triangle B) < \varepsilon\}$$

Proof Take $\mathcal{H} = \{A \in \mathcal{F}; A \text{ is "good"}\}$, where "good" means the approximation is valid. It suffices to check that $\mathcal{H}$ is a σ-algebra because $\mathcal{G} \subset \mathcal{H} \subset \mathcal{F}$, which entail $\mathcal{F} = \sigma(\mathcal{G}) \subseteq \mathcal{H} \subseteq \mathcal{F}$ and so therefore $\mathcal{H} = \mathcal{F}$.

i) $\Omega \in \mathcal{H}$: therefore $\Omega \in \mathcal{G}$.

ii) $A \in \mathcal{H} \Rightarrow A^{\complement} \in \mathcal{H}$: by definition, since A is "good", $\forall\, \varepsilon > 0, \exists B \in \mathcal{G}, \mathsf{P}(A \triangle B) < \varepsilon$. But

$$\begin{aligned} A \triangle B &= (A \setminus B) \sqcup (B \setminus A) \\ &= (A \cap B^{\complement}) \sqcup (B \cap A^{\complement}) \\ &= A^{\complement} \triangle B^{\complement} \end{aligned}$$

and thus $\mathsf{P}\left(A^{\complement}\cap B^{\complement}\right)<\varepsilon$. Also, since $\mathcal{G}$ is an algebra, $B^{\complement}\in\mathcal{G}$, $B^{\complement}$ is an approximation of $A^{\complement}$ and $A^{\complement}$ is "good" as well.

iii) If $A_n\in\mathcal{H}$, then $\bigcup_n A_n\in\mathcal{H}$: $\forall\,\varepsilon>0$, pick an $\varepsilon/2^n$ approximation $B_n\in\mathcal{G}$ of A_n : that is $\mathsf{P}\left(A_n\triangle B_n\right)\leq\varepsilon/2^n$. So for $A=\bigcup_n A_n$, we take $B=\bigcup_{n=1}^{N}B_n$ for N to be determined. Note that $B\in\mathcal{G}$. We have

$$\begin{aligned}\mathsf{P}\left(A\triangle B\right)&=\mathsf{P}\left(\bigcup_{n=1}^{\infty}A_n\triangle\bigcup_{n=1}^{N}B_n\right)\\&=\leq\mathsf{P}\left(\bigcup_{n=1}^{N}A_n\triangle\bigcup_{n=1}^{N}B_n\right)+\mathsf{P}\left(\bigcup_{n=N+1}^{\infty}A_n\right)\\&\text{wrong: assume }\mathcal{G}\text{ is a }\sigma\text{-algebra instead}\\&\leq\mathsf{P}\left(\bigcup_{n=1}^{\infty}(A_n\triangle B_n)\right)\\&\leq\sum_{n=1}^{\infty}\mathsf{P}\left(A_n\triangle B_n\right)\\&\leq\sum_{n=1}^{\infty}\frac{\varepsilon}{2^n}=\varepsilon\end{aligned}$$

■

Proof If $A\in\mathcal{E}$, then $\mathsf{P}(A)=0$ or $\mathsf{P}(A)=1$. We use the approximation: $\forall\,\varepsilon>0,\exists n\geq1$ and $B\in\mathbb{B}(\mathbb{R}^n)$ such that

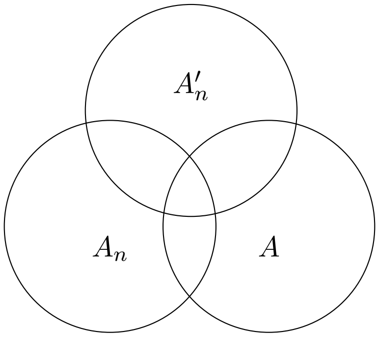
$$\mathsf{P}\left(A\triangle[(X_1,\ldots,X_n)\in B]\right)<\varepsilon.$$

This means that $A\simeq[(X_1,\ldots,X_n)\in B]$. If $A=[(X_1,X_2,\ldots)\in C]$ for $C\in\mathbb{B}(\mathbb{R}^n)$, then $A=[(X_{\pi(1)},X_{\pi(2)},\ldots,)\in C]$ for any finite permutation π. On the other hand, we apply a finite permutation π to

$$A\cap[(X_1,\ldots,X_n)\in B]=[(X_1,\ldots,X_n)\in C]\cap[(X_1,\ldots,X_n)\in B]$$

This gives

$$[(X_{\pi(1)},X_{\pi(2)},\ldots,)\in C]\cap[(X_{\pi(1)},X_{\pi(2)},\ldots,)\in B]$$



for the particular permutation $\pi(j)$ where

$$\pi(j) = \begin{cases} j+n & \text{if } 1 \leq j \leq n \\ j-n & \text{if } n+1 \leq k \leq 2n \\ j & \text{if } j \geq 2n+1. \end{cases}$$

This implies

$$\begin{aligned} &\mathsf{P}\left([(X_{\pi(1)}, X_{\pi(2)}, \ldots,) \in C] \cap [(X_{\pi(1)}, \ldots, X_{\pi(n)}, \ldots,) \in B]\right) \\ &\quad = \mathsf{P}\left([(X_1, X_2, \ldots,) \in C] \cap [(X_1, \ldots, X_n) \in B]\right) < \varepsilon \end{aligned}$$

Denote $(X_{\pi(1)}, \ldots, X_{\pi(n)}) = A'_n$ and $(X_1, \ldots, X_n) = A_n$. This means $A \simeq [(X_{\pi(1)}, \ldots X_{\pi(n)}) \in B]$ too. On the other hand, $A \simeq [(X_1, \ldots, X_n) \in B]$ so

$$[(X_1, \ldots X_n) \in B] \simeq [(X_{\pi(1)}, \ldots X_{\pi(n)}) \in B]$$

Finally, we consider

$$\begin{aligned} \mathsf{P}\left(A_n \triangle A'_n\right) &\leq \mathsf{P}\left((A_n \triangle A) \cup (A'_n \triangle A)\right) \\ &\leq \mathsf{P}\left(A_n \triangle A\right) + \mathsf{P}\left(A'_n \triangle A\right) \leq 2\varepsilon \end{aligned}$$

If we pick $\varepsilon = 1/k$ for $k \in \mathbb{N}$ and write $A_{n(k)}, A'_{n(k)}$, then $\mathsf{P}\left(A_{n(k)} \triangle A'_{n(k)}\right) \to 0$ which implies

$$\begin{aligned}\mathsf{P}\left(A\right)^2 &= \lim_{k\to\infty} \mathsf{P}\left(A_{n(k)}\right)\mathsf{P}\left(A'_{n(k)}\right)\\ &= \lim_{k\to\infty} \mathsf{P}\left(A_{n(k)} \cap A'_{n(k)}\right)\\ &= \mathsf{P}\left(A\right)\end{aligned}$$

■

Back to $S_n = \sum_{j=1}^n X_j$. We have the following corollary:

Corollary 10.4

For a random walk on $\mathbb{R}$, there are only four possibilities, one of which has probability 1.

1. $S_n = 0 \; \forall \; n$
2. $\lim_{n\to\infty} S_n = \infty$
3. $\lim_{n\to\infty} S_n = -\infty$
4. $-\infty = \liminf_{n\to\infty} S_n < \limsup_{n\to\infty} S_n = \infty$.

Proof $\limsup_{n\to\infty} S_n$ is a constant $c \in [-\infty, \infty]$ almost surely. If $c \in (-\infty, \infty)$, then we consider $S'_n = S_{n+1} - X_1 = \sum_{j=2}^{n+1} X_j$, which implies $c = \limsup_{n\to\infty} S'_n = \liminf S_{n+1} - X_1 = c - X_1$ and so $X_=0$ a.s. Otherwise, $\lim_{n\to\infty} S_n \in \{-\infty, \infty\}$ almost surely and similarly $\liminf_{n\to\infty} S_n \in \{-\infty, \infty\}$. So we can only have one of the above cases. ■

Corollary 10.5

If $(X_i; i \geq 1)$ are iid random variables with mean zero and $X_i \overset{\mathrm{d}}{=} -X_i$ but nondegenerate (*i.e.*, $\mathsf{P}\left(X_i = 0\right) < 1$), then we have

$$-\infty = \liminf_{n\to\infty} S_n < \limsup_{n\to\infty} S_n = \infty.$$

Proof If $\lim_{n\to\infty} S_n = \infty$, then $\lim_{n\to\infty} -S_n = \infty$ almost surely since S_n and $-S_n$ have the same law. Contradiction. ■

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space and $\mathcal{G}$ and algebra with $\mathcal{G} \subseteq \mathcal{F}$ and $\sigma(\mathcal{G}) = \mathcal{F}$, then $\forall\, A \in \mathcal{F}$, $\forall\, \varepsilon > 0$, $\exists B \in \mathcal{G}$ such that $\mathsf{P}\left(A \triangle B\right) < \varepsilon$

Proof Let $\mathcal{H} = \{A \in \mathcal{F}; A \text{ satisfies the approximation}\}$ we want to show for $(A_n) \in \mathcal{H}$, then $\bigcup_n A_n \in \mathcal{H}$. For all $\varepsilon > 0$, choose $B_n \in \mathcal{G}$ such that $\mathsf{P}\left(A_n \triangle B_n\right) < \frac{\varepsilon}{2^n}$. Let $B \in \mathcal{G}, B = \bigcup_{n=1}^{N} B_n$ for $N \in \mathbb{N}$ large enough. Then by the triangle inequality

$$\begin{aligned}\mathsf{P}\left(A \triangle B\right) &= \mathsf{E}\left(|\mathbf{1}_A - \mathbf{1}_B|\right) \\ &\leq \mathsf{E}\left(\left|\mathbf{1}_A - \mathbf{1}_{\bigcup_{n=1}^{N} A_n}\right|\right) + \mathsf{E}\left(\left|\mathbf{1}_{\bigcup_{n=1}^{N} A_n} - \mathbf{1}_B\right|\right) \\ &\leq \mathsf{P}\left(\bigcup_{n=N+1}^{\infty} A_n\right) + \mathsf{P}\left(\bigcup_{n=1}^{N} A_n \triangle B_n\right)\end{aligned}$$

As an exercise, verify that $\bigcup_{n=1}^{N} A_n \triangle B = \bigcup_{n=1}^{N} A_n \triangle \bigcup_{n=1}^{N} B_n \subseteq \bigcup_{n=1}^{N} (A_n \triangle B_n)$

$$\begin{aligned} &\leq \mathsf{P}\left(\bigcup_{n=N+1}^{\infty} A_n\right) + \sum_{n=1}^{N} \mathsf{P}\left(A_n \triangle B_n\right) \\ &= \mathsf{P}\left(\bigcup_{n=N+1}^{\infty} A_n\right) + \varepsilon \leq 2\varepsilon\end{aligned}$$

if N is large enough. ■

10.1 Stopping time of random walk

The question here in a simplify setting concerns (X_n) iid real-valued random variables with $\mathsf{E}\left(|X_i|\right) < \infty$. Clearly, $\forall\, n \in \mathbb{N}$, $\mathsf{E}\left(S_n\right) = n\mathsf{E}\left(X_i\right)$; we can generalize this by $\mathsf{E}\left(S_N\right) = \mathsf{E}\left(N\right)\mathsf{E}\left(X_i\right)$ for a "suitable random variable N taking values in $\mathbb{N} \cup \{\infty\}$. For any n, $Cf_n = \sigma(S_1, \ldots, S_n)$ captures the information of the random walk up to time n. $(\mathcal{F}_n)_{n \geq 1}$ captures the "information flow" of the random walk and is termed a **filtration** .

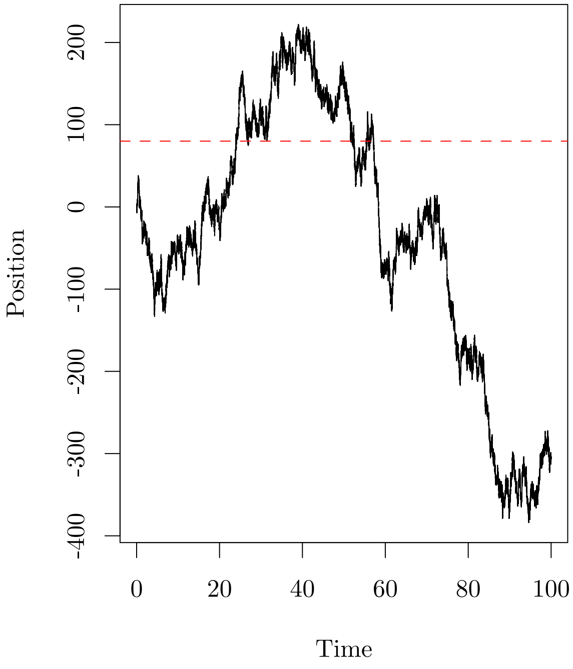
Definition 10.6 (Stopping time)

A random variable N taking values in $N \cup \{\infty\}$ is called a **stopping time** if $\forall\, n \in \mathbb{N}$, the event $[T \leq n] \in \mathcal{F}_n$

Exercise 10.2

N is a stopping time if and only if $[N \leq n] \in \mathcal{F}_n$, $\forall\ n \in \mathbb{N}$. So informally, N captures the information of the sequence $(S_1, \dots, S_N)$.

Hitting times for $n = 80$ for a binary random walk



Example 10.1 (Hitting time)

We have $N_A = \inf\{n; S_n \in A\}$ with the convention that $\inf\{\emptyset\} = \infty$. Check

$$[N_A = n] = S_1 \notin A, \dots S_{n-1} \notin A, S_n \in A] \in \mathcal{F}_n$$

In general, we can define the n^{th} hitting times of A inductively by $N_A^{(1)} = N_A$ and $N_A^{(n+1})= \inf\{n < N_A^{(n)}; S_n \in A\}$.

Exercise 10.3

Each $N_A^{(n)}$ is a stopping time.

In general, if N is random, we have $\mathcal{F}_N = \sigma(S_{N\wedge n}, n \geq 1)$. This is not a universally used definition.

Definition 10.7 (σ-algebra for stopping times)

For any stopping time N,

$$\mathcal{F}_N = \{A \in \mathcal{F}, A \cap [T \leq n] \in \mathcal{F}_n,\ \forall\, n \in \mathbb{N}\}$$

Exercise 10.4

Show that statement generates the same σ-algebra

$$\mathcal{F}_N = \{A \in \mathcal{F}, A \cap [T = n] \in \mathcal{F}_n,\ \forall\, n \in \mathbb{N}\}$$

Exercise 10.5

If S and T are stopping time and $S \leq T$, then $\mathcal{F}_S \subseteq \mathcal{F}_T$.

Exercise 10.6

If M and N are stopping time, $N \leq M$ and $A \in \mathcal{F}$, then

$$T = \begin{cases} N & \text{on } A \\ M & \text{on } A^{\complement} \end{cases}$$

is again a stopping time.

Lemma 10.8

If N is a stopping time and $S_n \neq S_{n+1}$, $\forall\, n$, then $\mathcal{F}_N = \sigma(S_{N\wedge n}; n \geq 1)$.

Proof Starting with inclusion $\supseteq$, write $S^N = (S_{N\wedge 1}, S_{N\wedge 2}, \ldots)$. An event is of the form $[S^N \in B]$ for some $B \in \mathcal{B}(\mathbb{R}^{\mathbb{N}})$. We check that $[S^N \in] \cap [N = n] \in \mathcal{F}_n$ for all $n \in \mathbb{N}$, then

$$[S^N \in B] \cap [N = n] = [(S_1, \ldots, S_n, S_n \ldots) \in B] \cap [N = n]$$

and both events are in $\mathcal{F}_n$ so $[S^N \in B] \in \mathcal{F}_N$.

As for the inclusion $\subseteq$, take $A \in \mathcal{F}_N$ (we want $A = [S^N \in B]$): by definition, $A \cap [N = n\ \in \mathcal{F}_n$ implies $A \cap [N = n] = [(X_1, \ldots, X_n) \in B_n]$ for some $B_n \in \mathbb{B}(\mathbb{R}^n)$, with the additional property that $\forall\, (x_1, \ldots x_n) \in B_n$, $x_j \neq x_{j+1}\ \forall\, j$, $(S_j \neq S_{j+1}\ \forall\, j)$. So we take

$$B = \bigcup_{n=1}^{\infty} (B_n \times \mathbb{R}^{\mathbb{N}})$$

It suffices to show $\forall\, n \in \mathbb{N} \cup \{\infty\},\ A \cap [N = n] = [S^N \in B] \cap [N = n]$. Now

$$\begin{aligned}
[S^N \in B] \cap [N = n] &= \bigcup_{m=1}^{\infty} [S^N \in B_m \times \mathbb{R}^{\mathbb{N}}] \cap [N = n] \\
&\overset{?}{=} [S^N \in B_m \times \mathbb{R}^{\mathbb{N}}] \cap [N = n] \\
&= (S_1, \ldots, S_n, S_n, \ldots) \in B_m \times \mathbb{R}^{\mathbb{N}}] \cap [N = n] \\
&= [(S_1, \ldots, S_n) \in B_m] \cap [N = n] \\
&= [(S_1, \ldots, S_n) \in B_m] = A \cap [N = n]
\end{aligned}$$

To check the claim on the second line, write for $m < n$,

$$\begin{aligned}
[S^N \in B_M \times \mathbb{R}^{\mathbb{N}}] \cap [N = n] &= [(S_1, \ldots S_n, S_n, \ldots) \in B_m \times \mathbb{R}^{\mathbb{N}} \cap [N = n] \\
&= [(S_1, \ldots S_m) \in B_m \times \mathbb{R}^{\mathbb{N}} \cap [N = n] \subseteq [N = m] \cap [N = n] = \emptyset
\end{aligned}$$

while for $m > n$

$$\begin{aligned}
[S^N \in B_M \times \mathbb{R}^{\mathbb{N}}] \cap [N = n] &= [(S_1, \ldots S_n, S_n, \ldots) \in B_m \times \mathbb{R}^{\mathbb{N}} \cap [N = n] \\
&= \emptyset \cap [N = n] = \emptyset
\end{aligned}$$

■

Theorem 10.9 (Wald's identity)

Let (X_n) be iid real valued with $\mathsf{E}\left(|X_i|\right) < \infty$. If N is a stopping time, then $\mathsf{E}\left(S_N\right) = \mathsf{E}\left(N\right) \mathsf{E}\left(X_i\right)$.

Proof

$$\begin{aligned}
\mathsf{E}\left(S_N\right) &= \sum_{n=1}^{\infty} \mathsf{E}\left(S_n \mathbf{1}_{N=n}\right) \\
&= \sum_{n=1}^{\infty} \sum_{m=1}^{n} \mathsf{E}\left(X_m \mathbf{1}_{N=n}\right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{?}{=} \sum_{m=1}^{\infty}\sum_{n=m}^{\infty} \mathsf{E}\left(X_m \mathbf{1}_{N=n}\right) \\
&= \sum_{m=1}^{\infty} \mathsf{E}\left(X_m \mathbf{1}_{N\geq m}\right)
\end{aligned}
$$

But $[N \geq m] = [N \leq m-1]^{\complement} \in \mathcal{F}_{m-1} = \sigma(X_1, \dots X_{m-1}) = \sigma(S_1, \dots, S_{m-1})$

$$
\begin{aligned}
&= \sum_{m=1}^{\infty} \mathsf{E}\left(X_i\right) \mathsf{P}\left(N \geq m\right) \\
&= \mathsf{E}\left(X_i\right) \sum_{m=1}^{\infty} \mathsf{P}\left(N \geq m\right) \\
&= \mathsf{E}\left(X_i\right) \mathsf{E}\left(N\right)
\end{aligned}
$$

To use Fubini, check

$$
\sum_{n=1}^{\infty}\sum_{m=1}^{n} \left|\mathsf{E}\left(X_m \mathbf{1}_{N=n}\right)\right| < \infty
$$

but

$$
\begin{aligned}
\sum_{n=1}^{\infty}\sum_{m=1}^{n} \mathsf{E}\left(|X_m| \mathbf{1}_{N=n}\right) &= \sum_{m=1}^{\infty}\sum_{n=m}^{\infty} \mathsf{E}\left(|X_m| \mathbf{1}_{N=n}\right) \\
&= \sum_{m=1}^{\infty} \mathsf{E}\left(|X_m \mathbf{1}_{N\geq m}\right) = \mathsf{E}\left(|X_i|\right) \mathsf{E}\left(N\right) < \infty.
\end{aligned}
$$

■

Example 10.2

Consider a simple random walk: let (X_n) be iid with $\mathsf{P}\left(X_n = \pm 1\right) = \frac{1}{2}$ and $T_x = \inf\{n \geq 1; S_n = x\}$ for $x \neq 0, x \in \mathbb{Z}$. Starting with $\mathsf{E}\left(T_a \wedge T_b\right) < \infty$ for $a < 0 < b$. We have

$$
\mathsf{P}\left(X + S_{b-a} \notin (a,b)\right) \geq \frac{1}{2^{b-a}} \;\forall\, x \in (a,b)
$$

which implies

$$\mathsf{P}\left(T_a \wedge T_b > n(b-a)\right) \leq \left(1 - \frac{1}{2^{b-a}}\right)^n$$

because

$$\mathsf{P}\left(T_a \wedge T_b > n(b-a)\right) \leq \mathsf{P}\left(S_{b-a} \in (a,b), S_{2(b-a)} \in (a,b), \ldots, S_{n(b-a)} \in (a,b)\right)$$

so the expectation is finite and we can use Wald's identity to answer our questions.

1. $\mathsf{P}\left(T_a < T_b\right) = \frac{b}{b-a}$ and $\mathsf{P}\left(T_b < T_a\right) = \frac{-a}{b-a}$
Proof Using Wald's identity, $\mathsf{E}\left(X_i\right)\mathsf{E}\left(T_a \cap T_b\right) = \mathsf{E}\left(S_{T_a \cap T_b}\right)$. The term $\mathsf{E}\left(X_i\right) = 0$ by symmetry, while the right end side is

$$a\mathsf{P}\left(S_{T_a \wedge T_b} = a\right) + b\mathsf{P}\left(S_{T_a \wedge T_b} = b\right) = a\mathsf{P}\left(T_a < T_b\right) + b\mathsf{P}\left(T_b < T_a\right)$$

We have a system of linear equations

$$\begin{aligned} 0 &= a\mathsf{P}\left(T_a < T_b\right) + b\mathsf{P}\left(T_b < T_a\right) \\ 1 &= \mathsf{P}\left(T_a < T_b\right) + \mathsf{P}\left(T_b < T_a\right) \end{aligned}$$

■

2. $\mathsf{P}\left(T_x < \infty\right) = 1, \ \forall\, x \neq 0.$
Proof Write $\mathsf{P}\left(T_x < \infty\right) \geq \mathsf{P}\left(T_x < T_M\right) = \frac{M}{M-x} \to 1$. Because $(-S_n)$ has the same distribution as S_n, $\mathsf{P}\left(T_x < \infty\right) = 1 \ \forall\, x \in \mathbb{N}$. ■

3. $\mathsf{E}\left(T_x\right) = \infty$ for $x \neq 0$
Proof Using Wald's identity, if $\mathsf{E}\left(T_x\right) < \infty$, then $0 = \mathsf{E}\left(X_i\right)\mathsf{E}\left(T_x\right) = \mathsf{E}\left(S_{T_x}\right) = x$. ■

Chapter 11
Conditional probability and conditional expectation

We remark that $\mathsf{P}(E) = \mathsf{E}(\mathbf{1}_E)$. Similarly, $\mathsf{P}(E \mid F) = \mathsf{E}(\mathbf{1}_E \mid F)$. Expectations are "more general" so we define conditional expectation, for which conditional probability follows as a consequence. The idea is

$$\mathsf{E}(X \mid Y) \quad \text{should be} \quad \mathsf{E}(X, \text{ pretending you know } Y),$$

so we average over what is known.

Definition 11.1 (Conditional expectation)

Let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathsf{P})$, let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra of $\mathcal{F}$. We say that $C \in \mathcal{L}^1(\Omega, \mathcal{G}, \mathsf{P})$ is a version of $\mathsf{E}(X \mid \mathcal{G})$ if for all $E \in \mathcal{G}$,

$$\mathsf{E}(C\mathbf{1}_E) = \mathsf{E}(X\mathbf{1}_E)$$

C is almost surely unique.

We write $\mathsf{E}(X \mid Y)$ for $\mathsf{E}(X \mid \sigma(Y))$ and $\mathsf{E}(X \mid Y_i, i \in I)$ for $\mathsf{E}(X \mid \sigma(Y_i, i \in I))$.

Example 11.1

If $X, Y \sim \mathcal{U}([0,1])$ are independent, then

$$\mathsf{E}\left(e^{XY} \mid Y\right) = \mathsf{E}\left(e^{X} \mid Y\right) = \int_0^1 e^{xY}\,\mathrm{d}x = \frac{e^Y - 1}{Y} = C$$

using our knowledge of conditional probability (calculate $\mathsf{E}(\exp(X) \mid Y = y)$) but keeping in mind that Y stays random.

Note

Fix $E \in \sigma(Y)$. Then, we can write $E = \{Y \in B\}$ for some $B \in \mathbb{B}(\mathbb{R})$. So $\mathsf{E}(C\mathbf{1}_E) = \mathsf{E}\left(\frac{e^Y-1}{Y}\mathbf{1}_{Y\in B}\right)$ and

$$\mathsf{E}\left(e^{XY}\mathbf{1}_{Y\in B}\right) = \int e^{xy}\mathbf{1}_{y\in B}\,\mathrm{d}(\mu_X \times \mu_Y)$$

μ_X is the law of X, μ_Y the law of Y. By Fubini, the join law factors and we may write the product measure as $\mathrm{d}x\,\mathrm{d}y$ since both have Lebesgue measure on $[0,1]$.

$$\begin{aligned}\int_0^1\int_0^1 e^{xy}\mathbf{1}_{y\in B}\,\mathrm{d}x\,\mathrm{d}y &= \int_0^1 \left(\mathbf{1}_{y\in B}\int_0^1 e^{xy}\,\mathrm{d}x\right)\,\mathrm{d}y \\ &= \int_0^1 \mathbf{1}_{y\in B}\frac{e^y-1}{y}\,\mathrm{d}y = \mathsf{E}\left(\frac{e^Y-1}{Y}\mathbf{1}_{Y\in B}\right)\end{aligned}$$

Example 11.2 (Random walk)

Let $(X_i, i\geq 1)$ be iid random variable in $\mathcal{L}^1(\Omega,\mathcal{F},\mathsf{P})$ where $\mathsf{E}\left(X_i\right)=\mu, S_n = X_1+\cdots+X_n$. Then

$$\begin{aligned}\mathsf{E}\left(S_{n+1}\mid S_n\right) &= \mathsf{E}\left(S_n+X_{n+1}\mid S_n\right) \\ &= \mathsf{E}\left(S_n\mid S_n\right)+\mathsf{E}\left(X_{n+1}\mid S_n\right) \\ &= S_n+\mu\end{aligned}$$

by linearity of conditional expectation, and independence between S_n and X_{n+1}, so we fully average. If $X\in\mathcal{G}$, then $\mathsf{E}\left(X\mid\mathcal{G}\right)=X$ (this is common abuse of notation).

Example 11.3 (Conditional expectation for independent random variables)

If $X \perp\!\!\!\perp Y$, then $\mathsf{E}\left(X\mid Y\right)=\mathsf{E}\left(X\right)$ and the proof follows by similar calculations using the defining property of conditional expectations. This is because if $\mathsf{E}\left(X\right)=C$, then

$$\mathsf{E}\left(C\mathbf{1}_{Y\in B}\right)=\mathsf{E}\left(X\right)\mathsf{E}\left(\mathbf{1}_{Y\in B}\right)=\mathsf{E}\left(X\right)\mathsf{P}\left(Y\in B\right)$$

while

$$\mathsf{E}\left(X\mathbf{1}_{Y\in B}\right)=\int_{\mathbb{R}}\int_{\mathbb{R}} x\mathbf{1}_{y\in B}\,\mathrm{d}\mu_X\,\mathrm{d}\mu_Y=\int\mathbf{1}_{y\in B}\mathsf{E}\left(X\right)\,\mathrm{d}\mu_Y=\mathsf{E}\left(X\right)\mathsf{P}\left(Y\in B\right)$$

Example 11.4 (Conditional expectation for random variables with joint density)

If X,Y have joint density, $f_{X,Y}(x,y)$, what should the conditional density of X given $Y=y$ be?

$$f_{X|Y}(x\mid y)=\frac{f_{X,Y}(x,y)}{\int f_{X,Y}(x,y)\,\mathrm{d}x}=\frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Then we should have $\mathsf{E}\left(X \mid Y = y\right) = \int x f_{X|Y}(x \mid y)\,\mathrm{d}x$ and so a good guess for

$$\begin{aligned}\mathsf{E}\left(X \mid Y\right) &= \int x f_{X|Y}(x \mid Y)\,\mathrm{d}x \\ &= \int x f_{X,Y}(x, Y)\,\mathrm{d}x \\ &= \frac{\int x f_{X,Y}(x, Y)\,\mathrm{d}x}{\int f_{X,Y}(x, Y)\,\mathrm{d}x}\end{aligned}$$

and the denominator is $\sigma(Y)$-measurable

Example 11.5 (Discrete conditional expectation)

If Ω is discrete, $\mathcal{F} = 2^{\Omega}$. Let $(\Omega_i, i \geq 1)$ be a partition of Ω. We say $\mathcal{F} = \sigma(\Omega_i, i \geq 1)$, that is $\mathcal{G} = \sigma(Y)$ where $Y(\omega) = i\mathbf{1}_{\omega \in \Omega_i}$. If X is $\mathcal{F}$-measurable, then

$$\mathsf{E}\left(X \mid Y\right)(\omega) = C(\omega) = \frac{\mathsf{E}\left(X\mathbf{1}_{\Omega_i}\right)}{\mathsf{P}\left(\Omega_i\right)}$$

whenever $\omega \in \Omega_i$. This function is constant over partition classes.

Example 11.6

If $\varphi : \mathbb{R}^2 \to \mathbb{R}$ measurable and $X, Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathsf{P})$, for $X \perp\!\!\!\perp Y$ with $\mathsf{E}\left(|\varphi(X,Y)|\right) < \infty$. What is $\mathsf{E}\left(\varphi(X,Y) \mid X\right)$? Let $g(x) = \int_{\mathbb{R}} \varphi(x,y)\,\mathrm{d}\mu_Y$. We claim

$$\mathsf{E}\left(\varphi(X,Y) \mid X\right) = g(X) = \int_{\mathbb{R}} \varphi(X,y)\,\mathrm{d}\mu_Y$$

We verify once again the defining property: if $E \in \sigma(X)$, write $E = \{X \in B\}$, so

$$\begin{aligned}\mathsf{E}\left(\varphi(X,Y)\mathbf{1}_E\right) &= \int \varphi(x,y)\mathbf{1}_{x \in B}\,\mathrm{d}(\mu_X \times \mu_Y) \\ &= \int \mathbf{1}_{x \in B} \int \varphi(x,y)\,\mathrm{d}\mu_Y\,\mathrm{d}\mu_X \\ &= \int \mathbf{1}_{x \in B} g(x)\,\mathrm{d}\mu_X \\ &= \mathsf{E}\left(g(X)\mathbf{1}_E\right)\end{aligned}$$

11.1 Existence and uniqueness of conditional expectation

We now tackle the proof of uniqueness

Proposition 11.2 (Uniqueness of conditional expectation)

If C and D are versions of $\mathsf{E}\left(X \mid \mathcal{G}\right)$, then $C \overset{\text{a.s.}}{=} D$.

Proof By assumption, $C, D \in \mathcal{G}$, so $\{C > D\} \in \mathcal{G}$. If $C \overset{\text{a.s.}}{\neq} D$, then wlog $\mathsf{P}\left(C > D\right) > 0$. So $\mathsf{P}\left(C > D + \frac{1}{n}\right) > 0$ for some n. Let $E = C > D + \frac{1}{n}$; then $\mathsf{E}\left(C\mathbf{1}_E\right) = \mathsf{E}\left(X\mathbf{1}_E\right) = ED\mathbf{1}_E$. But on the other hand,

$$\mathsf{E}\left(C\mathbf{1}_E\right) \geq \mathsf{E}\left(\left(D + \frac{1}{n}\right)\mathbf{1}_E\right) = \mathsf{E}\left(D\mathbf{1}_E\right) + \frac{1}{n}\mathsf{E}\left(\mathbf{1}_E\right) > \mathsf{E}\left(D\mathbf{1}_E\right)$$

using monotonicity, linearity and $\mathsf{P}\left(E\right) > 0$. ■

Proposition 11.3 (Existence of conditional expectation)

For all $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathsf{P})$ for all $G \in \mathcal{F}$ sub-σ-algebras, there exists a random variable Y that is a version of $\mathsf{E}\left(X \mid \mathcal{G}\right)$.

$\mathcal{L}^2$ has a Hilbert space structure with an inner product.[28]

Suppose the proposition is true. Then for $X \geq 0$, $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathsf{P})$. Let $X_n = \min(X, n)$; let Y_n be a version of $\mathsf{E}\left(X_n \mid \mathcal{G}\right)$. Then $Y_n \leq Y_{n+1}$ (otherwise there exists k such that $\mathsf{P}\left(\{Y_{n+1} < Y_n - \frac{1}{k}\}\right) > 0$ which (taking expectations and using defining property of conditional expectation) contradicts that $X_n \leq X_{n+1}$ almost surely.) Then let $Y = \limsup_{n\to\infty} Y_n$. We then have $\forall\, E \in \mathcal{G}$,

$$\mathsf{E}\left(Y\mathbf{1}_E\right) = \lim_{n\to\infty} \mathsf{E}\left(Y_n\mathbf{1}_E\right) = \lim_{n\to\infty} \mathsf{E}\left(X_n\mathbf{1}_E\right) = \mathsf{E}\left(X\mathbf{1}_E\right)$$

by the defining property so Y is a version of $\mathsf{E}\left(X \mid \mathcal{G}\right)$. $Y \in \mathcal{L}^1(\Omega, \mathcal{G}, \mathsf{P}), Y = \lim Y_n$, is $\mathcal{G}$-measurable because is it an increasing limit and each $Y_n \in \mathcal{L}^1(\Omega, \mathcal{G}, \mathsf{P})$ and $\mathsf{E}\left(Y\right) = \mathsf{E}\left(X\right) < \infty$. This is the usual closure properties.

Next, for $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathsf{P})$, let $X = X^+ - X^-$ and let $Y^+ - Y^-$ be versions of $\mathsf{E}\left(X^\pm \mid \mathcal{G}\right)$. Then

$$\mathsf{E}\left(Y^+ - Y^-\right) = \mathsf{E}\left(X^+\right) + \mathsf{E}\left(X^-\right) \qquad \text{(defining property)}$$

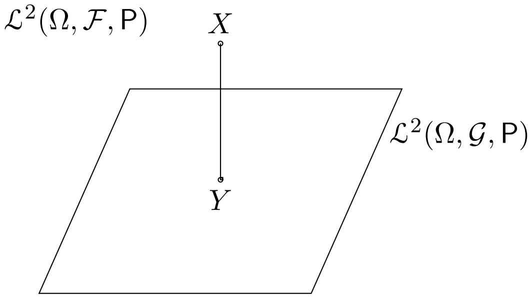
[28] To identify random variables that are almost equivalent, we (would) need to work with equivalence classes for random variables with distance zero. To go from $X \in \mathcal{L}^2$ to $\mathcal{L}^1$, we will use the last steps of the standard machine.

$$= \mathsf{E}\left(|X|\right) < \infty.$$

so $Y^+ - Y^- \in \mathcal{L}^1(\Omega, \mathcal{G}, \mathsf{P})$ and for $E \in \mathcal{G}$,

$$\begin{aligned}
\mathsf{E}\left((Y^+ - Y^-)\mathbf{1}_E\right) &= \mathsf{E}\left(Y^+\mathbf{1}_E\right) - \mathsf{E}\left(Y^-\mathbf{1}_E\right) && \text{(linearity)} \\
&= \mathsf{E}\left(X^+\mathbf{1}_E\right) - \mathsf{E}\left(X^-\mathbf{1}_E\right) && \text{(defining property)} \\
&= \mathsf{E}\left(X\mathbf{1}_E\right) && \text{(linearity)}
\end{aligned}$$

so $Y^+ - Y^-$ is a version of $\mathsf{E}\left(X \mid \mathcal{G}\right)$.


$\mathcal{L}^2(\Omega, \mathcal{F}, \mathsf{P})$
X
$\mathcal{L}^2(\Omega, \mathcal{G}, \mathsf{P})$
Y

11.2 Conditional expectation as least-square predictor

Lemma 11.4 (Completeness of $\mathcal{L}^2$)

$\mathcal{L}^2(\Omega, \mathcal{F}, \mathsf{P})$ is complete, with norm $\|X\|_p = \mathsf{E}\left(X^p\right)^{\frac{1}{p}}$.

Proof Let $X_n, n \geq 1$ be Cauchy in $\mathcal{L}^p$. We need to show X_n has an almost sure limit. Choose a subsequence $(k_n, n \in \mathbb{N})$, $k_n \uparrow \infty$ such that for all $r, s \geq k_n$, $\|X_r - X_s\|_p \leq 2^{-n}$. Then

$$\mathsf{E}\left(|X_{k_{n+1}} - X_{k_n}|\right) \leq \|X_{k_{n+1}} - X_{k_n}\|_p \leq 2^{-n}$$

so

$$\mathsf{E}\left(\sum_{n \geq 1} |X_{k_{n+1}} - X_{k_n}|\right) < \infty$$

so

$$\sum_{n \geq 1} (X_{k_{n+1}} - X_{k_n}) \text{ converges almost surely}$$

which implies X_{n_k} converges almost surely as $k \to \infty$.. We thus let $X = \lim_{k_n \to \infty} X_{k_n}$. To check $X \in \mathcal{L}^p$ and $X_r \xrightarrow{\text{a.s.}} X$, note that for $n \in \mathbb{N}$, $r \geq k_n$, and all $t \geq n$

$$\mathsf{E}\left(|X_r - X_{k_t}|^p\right) \leq 2^{-np}$$

By Fatou's lemma,

$$\mathsf{E}\left(|X_r - X|^p\right) = \mathsf{E}\left(\lim_{t\to\infty} |X_r - X_{k_t}|^p\right) \leq \liminf_{t\to\infty} \mathsf{E}\left(|X_r - X_{k_t}|^p\right) \leq \frac{1}{2^{np}}.$$

Finally, $\|X\|_p \leq \|X - X_r + X_r\| \leq \|X - X_r\|_p + \|X_r\|_p \leq 2^{-n} + \|X_r\|_p < \infty$ using the triangle inequality and since $X_r \in \mathcal{L}^p$. For $r \geq k_n$, $\|X_r - X\|_p \leq 2^{-n}$ and so $X_r \to X$ in $\mathcal{L}^p$. ■

Theorem 11.5 (Orthogonal projection)

Let $\mathcal{L}^2(\Omega, \mathcal{G}, \mathsf{P})$ be a subspace of $\mathcal{K} \coloneqq \mathcal{L}^2(\mathcal{G}) \coloneqq \mathcal{L}^2(\Omega, \mathcal{G}, \mathsf{P})$ (where $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$) which is complete in that whenever (V_n) is a sequence in $\mathcal{K}$ which is Cauchy, that is

$$\sup_{r,s\geq k} \|V_r - V_s\|_2 \to 0 \quad \text{as } k \to \infty$$

then there exists a V in $\mathcal{L}^2(\mathcal{G})$ such that $\|V_n - V\|_2 \to 0$ as $n \to \infty$. Then given $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathsf{P})$, there exists $Y \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathsf{P})$ such that

$$\|X - Y\|_2 = \Delta \coloneqq \inf\{\|X - W\| : W \in \mathcal{L}^2(\mathcal{G})\},$$

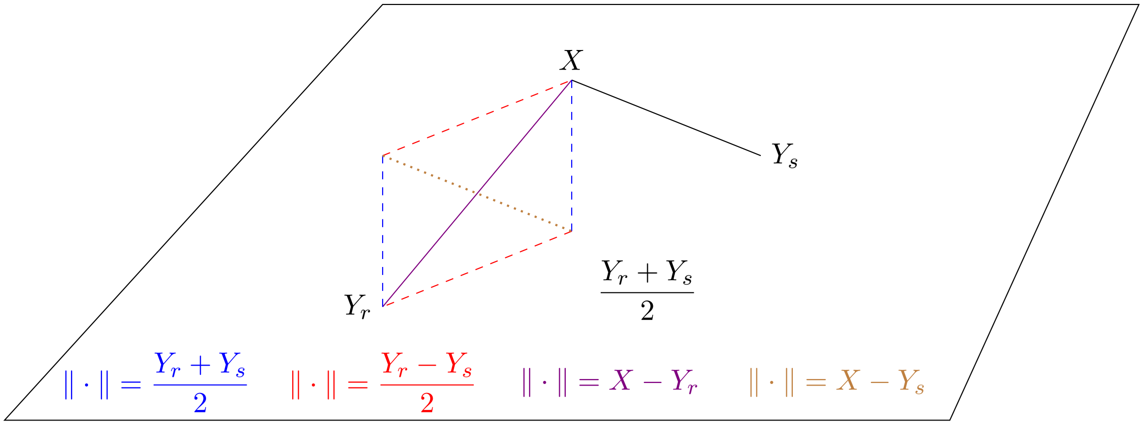
and

$$\langle X - Y, Z\rangle = 0, \quad \forall\ Z \text{ in } \mathcal{L}^2(\mathcal{G})$$

No w if $G \in \mathcal{G}$, then taking $Z \coloneqq \mathbf{1}_G \in \mathcal{L}^2(\mathcal{G})$ gives $\mathsf{E}\left(Y\mathbf{1}_G\right) = \mathsf{E}\left(X\mathbf{1}_G\right)$ and hence Y is a version of $\mathsf{E}\left(X \mid \mathcal{G}\right)$, as required.

For the proof, recall the parallelogram law: $(a+b)^2 + (a-b)^2 = 2(a^2+b^2)$

Proof Let $\Delta = \inf\{\|X - Z\| : Z \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathsf{P})\}$. Let $(Y_n, n \geq 1) \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathsf{P})$ be such that $\|Y_n - X\| \to \Delta$. If Y_n is Cauchy, then it has an almost sure limit Y.



Then

$$\|X-Y_r\|_2^2+\|X-Y_s\|_2^2 = 2\|X-(Y_r+Y_s)/2\|_2^2+2\|(Y_s-Y_r)/2\|_2^2$$

so

$$\|Y_r-Y_s\|_2^2 = \underbrace{2\|X-Y_r\|_2^2+2\|X-Y_s\|_2^2}_{\leq 4\Delta^2+\varepsilon} - \underbrace{4\|X-(Y_r+Y_s)/2\|_2^2}_{\geq 4\Delta^2}$$

when r, s are large. So for r, s sufficiently large, $\|X_r-X_s\|_2^2 \leq \varepsilon$ and thus $(X_n, n \geq 1)$ is Cauchy. To show uniqueness of the minimizer, we show it is a version of the conditional expected value.

Claim

Y is a version of $\mathsf{E}(X \mid \mathcal{G})$

Proof It suffices to show $\mathsf{E}((X-Y)\mathbf{1}_E) = 0$ for all $E \in \mathcal{G}$. We prove the stronger fact: $\mathsf{E}((X-Y)Z) = 0$ for all $Z \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathsf{P})$. So for $Z \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathsf{P})$, for all $t \in \mathbb{R}$, $Y + tZ \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathsf{P})$. So

$$\|X-(Y+tZ)\|_2^2 \geq \|X-Y\|_2^2$$

that is

$$\mathsf{E}\left(((X-Y)-tZ)^2\right) \geq \mathsf{E}\left((X-Y)^2\right)$$

so

$$-2t\mathsf{E}(Z(X-Y)) + t^2\mathsf{E}\left(Z^2\right) \geq 0$$

This hold true for all t, but quadratic terms tend to zero faster than linear (for any $a, b \neq 0, at^2 - 2bt < 0$ for some t), so for this to hold true, we need $\mathsf{E}\left(Z(X-Y)\right) = 0$.

■

■

Index