MATH 587/589 COURSE NOTES

LOUIGI ADDARIO-BERRY

Abstract. **Course notes for Math 587, Fall 2019 and Math 589, Winter 2020.**
These notes cover all course content contained in the lectures up to and including that of October 29, 2019.

Contents

Date: November 26, 2019.

1. Notation

We write $\mathcal{L}_X$ or μ_X (Make this consistent.) for the distribution of X. Given a σ-finite measure μ on $\mathbb{R}$ and $p > 0$, write $|\mu|_p = \left(\int_{\mathbb{R}} |x|^p d\mu(x)\right)^{1/p}$. If μ is a probability distribution and X has law μ then $|\mu|_p = (\mathbf{E}\left[|X|^p\right])^{1/p}$.

2. Measure theory

Measure theory is the algebraic underpinning of probability theory. It can feel rather abstract; but it is worth setting things up clearly.

2.1. Rings, fields and σ-fields.

Fix a set Ω and a set $\mathcal{A}$ of subsets of Ω with $\emptyset \in \mathcal{A}$. We say $\mathcal{A}$ is a *ring* if the following hold.

Ring

(a) If $E, F \in \mathcal{A}$ then $E \cup F \in \mathcal{A}$.
(b) If $E, F \in \mathcal{A}$ then $F \setminus E \in \mathcal{A}$.

We say $\mathcal{A}$ is a π-system if the following holds.

π-system

(c) If $E, F \in \mathcal{A}$ then $E \cap F \in \mathcal{A}$.

Field

We say $\mathcal{A}$ is a *field* if it is a ring and also the following holds.

(d) If $E \in \mathcal{A}$ then $E^c \in \mathcal{A}$.

We say $\mathcal{A}$ is a *σ-field* if it is a field and also the following holds.

σ-field

(a') For any sequence $(A_n, n \geq 1)$ of elements of $\mathcal{A}$, $\bigcup_{n\geq 1} A_n \in \mathcal{A}$.

In all the above cases, we refer to Ω as the *ground set*. Finally, for an arbitrary set $\mathcal{A}$ of subsets of Ω, the *σ-field generated by $\mathcal{A}$* is

$\sigma(\mathcal{A})$

$$\sigma(\mathcal{A}) := \bigcap_{\{\mathcal{F} \supset \mathcal{A} : \mathcal{F} \text{ a } \sigma-\text{field}\}} \mathcal{F};$$

this is the smallest σ-field containing the set $\mathcal{A}$.

Exercise 2.1. *(i) Show that properties (a) and (d) together imply properties (b) and (c).*
(ii) Show that a field which is closed under countable disjoint *unions is a σ-field.*

Throughout these notes, "countable" means "finite or countably infinite".

Exercise 2.2. *Write $\mathbb{N} := \{1, 2, 3, \ldots\}$. For $n \in \mathbb{N}$ we let $[n] := \{1, 2, \ldots, n\}$. Say that $S \subset \mathbb{N}$* has an asymptotic density *if*

$$\mu(S) := \limsup_{n\to\infty} \frac{|S \cap [n]|}{n} = \liminf_{n\to\infty} \frac{|S \cap \{1, 2, \ldots, n\}|}{n}.$$

Write $\mathcal{A}$ for the set of subsets of $\mathbb{N}$ which have an asymptotic density. Is $\mathcal{A}$ is a π-system? Is it a ring? A field? A σ-field?

Measurable space

2.2. Building measures.

A *measurable space* is a pair $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is a σ-field over Ω. Given such a space, a *measure* μ on $\mathcal{F}$ is a function $\mu : \mathcal{F} \to [0, \infty]$ such that $\mu(\emptyset) = 0$, and for any sequence $(A_n, n \geq 1)$ of disjoint elements of $\mathcal{F}$,

Measure

$$\mu\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mu(A_n).$$

We then call $(\Omega, \mathcal{F}, \mu)$ a *measure space*. You should think of a measure space as a model for a physical system involving randomness. Sometimes this can be quite concrete. For example, one might take $\Omega = [6]$, $\mathcal{F} := 2^\Omega$ is the power set of Ω, and $\mu(S) = |S|/6$, to model the roll of a fair die; here $\mu(S)$ is the probability that the roll yields a value in S. If we took $\Omega = [6]^{[2]} = \{(i, j) : i, j \in [6]\}$ and $\mu(S) = |S|/36$, we could view this as modelling two successive rolls of a fair die.

Measure space

On the other hand, when doing probability it is often useful to leave the details of the measure space rather implicit. There are various tools which justify doing this (change of variables, existence theorems,...), which we'll see later.

Exercise 2.3. *Let μ be a measure on a σ-field $\mathcal{F}$.*

(i) ***[Monotone convergence/Continuity from below.]*** *Show that for any increasing sequence $(E_n, n \geq 1)$ of elements of $\mathcal{F}$, it holds that $\mu(\bigcup_{n\geq 1} E_n) = \lim_{n\to\infty} \mu(E_n)$.*

(ii) ***[Dominated convergence/Continuity from above.]*** *Show that for any decreasing sequence $(E_n, n \geq 1)$ of elements of $\mathcal{F}$ with $\mu(E_1) < \infty$, it holds that $\mu(\bigcap_{n\geq 1} E_n) = \lim_{n\to\infty} \mu(E_n)$.*

(iii) ***[Subadditivity.]*** *Show that for any sequence $(E_n, n \geq 1)$ of elements of $\mathcal{F}$, it holds that $\mu(\bigcup_{n\geq 1} E_n) \leq \sum_{n\geq 1} \mu(E_n)$.*

Exercise 2.4. *Which of the following triples $(\Omega, \mathcal{F}, \mu)$ are measure spaces? Can you think of physical systems which they model?*

(a) $\Omega = \mathbb{N}$, $\mathcal{F}$ *the set of subsets of* $\mathbb{N}$ *which have an asymptotic density,* $\mu(S)$ *the asymptotic density of* S.

(b) $\Omega = \{0,1\}^n$, $\mathcal{F}$ *the power set of* Ω, $\mu(\{\omega\}) = p^{|\{i\in[n]:\omega_i=1\}|}(1-p)^{|\{i\in[n]:\omega_i=0\}|}$, *where* $p \in (0,1)$ *is fixed.*

(c) $\Omega = \{0,1\}^{\mathbb{N}}$, $\mathcal{F}$ *the power set of* Ω, $\mu(\omega) = p^{|\{i\in[n]:\omega_i=1\}|}$, *where* $p \in [0,1]$ *is fixed.*

(d) $\Omega = [0,1]$, $\mathcal{F}$ *the collection of sets* $S \subset [0,1]$ *such that either* S *or* $[0,1] \setminus S$ *is countable, and* $\mu(S) = |S|$.

You have likely seen probability distributions described in terms of *cumulative distribution functions* (CDFs). For example, the standard exponential distribution has CDF $F(x) = (1-e^{-x})\mathbf{1}_{[x\geq 0]}$, corresponding to the fact that for E is a standard exponential random variable, $\mathbf{P}\{E \leq x\} = (1-e^{-x})\mathbf{1}_{[x\geq 0]}$. [1] What $F(x)$ lets us easily compute is probabilities of the form $\mathbf{P}\{E \in (a,b]\} = F(b) - F(a)$, or $\mathbf{P}\{E \in \bigcup_{i=1}^n (a_i, b_i]\} = \sum_{i=1}^n (F(b_i) - F(a_i))$, where $(a_1, b_1], \ldots, (a_n, b_n]$ are disjoint intervals. On the other hand, it's not clear how we would use the above CDF to determine $\mathbf{P}\{E \in \mathbb{Q}\}$, for example, although we know the answer must be zero. If we are going to specify probability distributions in this way, we should really prove that probability measures are uniquely determined by their CDFs; this is a corollary of the coming development.

Indicator of a set

Fix a ring $\mathcal{A}$ over a ground set Ω. A *pre-measure on* $\mathcal{A}$ is a function $\mu : \mathcal{A} \to [0,\infty]$ with $\mu(\emptyset) = 0$ such that for any sequence $(A_n, n \geq 1)$ of disjoint elements of $\mathcal{F}$, *if* $\bigcup_{n\geq 1} A_n \in \mathcal{A}$ then

Pre-measure

$$\mu\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mu(A_n)\,.$$

Pre-measure space

We then say that $(\Omega, \mathcal{A}, \mu)$ is a pre-measure space.

$\mathcal{A}(\mathbb{R})$

Here is a key example of a pre-measure space. We hereafter write

$$\mathcal{A}(\mathbb{R}) = \left\{ \bigcup_{i=1}^n (a_i, b_i] : n \geq 1, -\infty < a_1 \leq b_1 \leq a_2 \leq \ldots \leq a_n \leq b_n < \infty \right\}.$$

Exercise 2.5. *Prove that $\mathcal{A}(\mathbb{R})$ is a ring over $\mathbb{R}$.*

We will see later that if F is a CDF then we can define a function μ on $\mathcal{A}(\mathbb{R})$ by setting $\mu(\bigcup_{i=1}^n (a_i, b_i]) = \sum_{i=1}^n (F(b_i) - F(a_i)$ when $((a_i, b_i], n \geq 1)$ are pairwise disjoint, and that the resulting triple $(\mathbb{R}, \mathcal{A}(\mathbb{R}), \mu)$ is a pre-measure space. The primordial[2] existence theorem for measures is the following.

Theorem 2.1 (Carathéodory extension theorem)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a pre-measure space. Then there exists a σ-field $\mathcal{F}$ containing $\mathcal{A}$ such that μ extends to a measure on $\mathcal{F}$.*

The previous theorem provides existence; the next theorem provides uniqueness.

[1] Here and throughout these notes, for a set S and a subset $T \subset S$, we write $I_T : S \to \{0,1\}$ for the indicator of set T, so $I_T(x) = 1$ for $x \in T$ and $I_T(x) = 0$ otherwise.

[2] Primordial, adj. and n.: That constitutes the origin or starting point from which something else is derived or developed, or on which something else depends; fundamental, basic; elemental. –Oxford English Dictionary

Theorem 2.2 (Dynkin's theorem). *Let $(\Omega, \mathcal{F})$ be a measurable space, and let $\mathcal{P} \subset \mathcal{F}$ be a π-system $\Omega \in \mathcal{P}$ and with $\sigma(\mathcal{P}) = \mathcal{F}$. Fix measures μ_1, μ_2 on $\mathcal{F}$, and suppose that (a) $\mu_1(E) = \mu_2(E)$ for all $E \in \mathcal{P}$ and (b) there exist sets $(\Omega_n, n \geq 1)$ in $\mathcal{P}$ with $\Omega_n \uparrow \Omega$ as $n \to \infty$ and with $\mu_1(\Omega_n) < \infty$. Then $\mu_1 \equiv \mu_2$.*

The proof of the Carathéodory extension theorem consists of two parts. Starting from the pre-measure space $(\Omega, \mathcal{A}, \mu)$ provided by the hypothesis of the theorem, we first use the pre-measure μ provided by to produce an upper bound on any putative[3] extension of μ. Next we show that the upper bound indeed yields a measure on a σ-field extending the ring $\mathcal{A}$.

Proposition 2.3. *Let $(\Omega, \mathcal{A}, \mu)$ be a pre-measure space. For $B \subset \Omega$ let*

$$\mu^*(B) := \inf\left(\sum_{n\geq 1} \mu(A_n) : A_n \in \mathcal{A}, n \geq 1; B \subset \bigcup_{n\geq 1} A_n\right).$$

Then μ^* is an outer measure*:*

Outer measure

(i) $\mu^*(\emptyset) = 0$;
(ii) If $E \subset F$ then $\mu(E) \leq \mu(F)$;
(iii) If $(E_i, i \geq 1)$ are subsets of Ω then $\mu^*(\bigcup_{i\geq 1} E_i) \leq \sum_{i\geq 1} \mu^*(E_i)$.

Note: usually I try to avoid putting definitions within the statements of Theorems, Propositions, Lemmas and so forth; but this is almost the only place where outer measures will be used.

Lemma 2.4 applies to any outer measure, not just μ^*; change notation?

Lemma 2.4 (Carathéodory lemma). *Given an outer measure μ^* over a set Ω, say $A \subset \Omega$ is* μ^*-additive *if for all $B \subset \Omega$,*

$$\mu^*(B) = \mu^*(A \cap B) + \mu^*(A^c \cap B).$$

Let $\mathcal{F} = \{A \subset \Omega : A \text{ is } \mu^*\text{-additive}\}$, and define $\mu^+ : \mathcal{F} \to [0, \infty]$ by $\mu^+(A) := \mu^*(A)$. Then $(\Omega, \mathcal{F}, \mu)$ is a measure space.

I think of μ^*-additive sets as knives; they "sharply cut" any set $B \subset \Omega$ in two without any change of μ^*-measure. I'm not sure how useful this perspective is to others.

Proof of Proposition 2.3. Point (i) is obvious since the empty set is a cover of itself. Point (ii), monotonicity, is also obvious, since if $E \subset F$ then any cover of F is a cover of E, so $\mu^*(E)$ is an infimum over a larger set than $\mu^*(F)$.

Finally, fix subsets $(E_i, i \geq 1)$ of Ω and write $E = \bigcup_{i\geq 1} E_i$. Next fix $\epsilon > 0$, and for each $i \geq 1$, fix a cover $(A_n^i, n \geq 1)$ of E_i with

$$\sum_{n\geq 1} \mu(A_n^i) \leq \mu^*(E_i) + \frac{\epsilon}{2^i};$$

such a cover exists by the definition of $\mu^*(E_i)$. Then $(A_n^i, n, i \geq 1)$ is a cover of E, so

$$\begin{aligned}\mu^*(E) &\leq \sum_{n,i\geq 1} \mu(A_n^i) \\ &\leq \sum_{i\geq 1} \left(\mu^*(E_i) + \frac{\epsilon}{2^i}\right) \\ &= \sum_{i\geq 1} \mu^*(E_i) + \epsilon\,.\end{aligned}$$

Since $\epsilon > 0$ was arbitrary it follows that $\mu^*(E) \leq \sum_{i\geq 1} \mu^*(E_i)$. □

[3]Putative, adj.: That is commonly believed to be such; reputed, supposed; imagined; postulated, hypothetical. –Oxford English Dictionary

Proof of Lemma 2.4. The conclusion of the Carathéodory lemma is that $\mathcal{F}$ is a σ-field over Ω and μ^+ is a measure on $\mathcal{F}$. We prove these in order.

First, for any $B \subset \Omega$ we have $\mu^*(\emptyset \cap B) + \mu^*(\Omega \cap B) = \mu^*(\emptyset) + \mu^*(B) = \mu^*(B)$, so $\emptyset \in \mathcal{F}$. Also, the definition of μ^* additive sets is invariant under complementation, so $A \in \mathcal{F}$ if and only if $A^c \in \mathcal{F}$.

We next show $\mathcal{F}$ is closed under intersections. Fix any sets $A_1, A_2 \in \mathcal{F}$. For any $B \subset \Omega$, we may write B as a disjoint union

$$\begin{aligned} B &= B_0 \cup B_1 \cup B_2 \cup B_{12}, \text{ where} \\ B_0 &= B \cap A_1^c \cap A_2^c, \\ B_1 &= B \cap A_1 \cap A_2^c, \\ B_2 &= B \cap A_1^c \cap A_2, \text{ and} \\ B_{12} &= B \cap A_1 \cap A_2 \,. \end{aligned}$$

This "cuts B into four pieces", according to its intersection with B_1 and B_2. Since A_1 and A_2 are μ^*-additive, we have

$$\begin{aligned} \mu^*(B) &= \mu^*(B \cap A_1) + \mu^*(B \cap A_1^c) \\ &= \mu^*(B_{12}) + \mu^*(B_1) + \mu^*(B_2) + \mu^*(B_0) \end{aligned}$$

If we likewise cut $B \setminus B_{12}$ into four pieces, only the last three pieces will be non-empty, and we obtain

$$\mu^*(B \setminus B_{12}) = \mu^*(B_0) + \mu^*(B_2) + \mu^*(B_1).$$

Together the last two equations give that

$$\mu^*(B) = \mu^*(B_{12}) + \mu^*(B \setminus B_{12}) = \mu^*(B \cap A_1 \cap A_2) + \mu^*(B \cap (A_1 \cap A_2)^c) \,.$$

Thus $A_1 \cap A_2 \in \mathcal{F}$.

At this point we know $\mathcal{F}$ is a field, so to show it is a σ-field it suffices to establish that it is closed under countable disjoint unions. Fix a sequence $(A_i, n \geq 1)$ of disjoint sets in $\mathcal{F}$, and any set $B \subset \Omega$. Writing $A = \bigcup_{i \geq 1} A_i$, we must show that for all $B \subset \Omega$ we have $\mu^*(B) = \mu^*(A \cap B) + \mu^*(A^c \cap B)$. The fact that $\mu^*(B) \leq \mu^*(A \cap B) + \mu^*(A^c \cap B)$ is immediate by subadditivity of outer measure, so we only need to show the reverse inequality.

We will again "cut B into pieces" according to its intersection with the sets A_n. However, since the sets are disjoint, our task is now simpler; we may rewrite

$$\begin{aligned} \mu^*(B) &= \mu^*(B_{12}) + \mu^*(B_1) + \mu^*(B_2) + \mu^*(B_0) \\ &= \mu^*(B_1) + \mu^*(B_2) + \mu^*(B_0) \\ &= \mu^*(B \cap A_1) + \mu^*(B \cap A_2) + \mu^*(B \cap A_1^c \cap A_2^c) \,. \end{aligned}$$

More generally, since $A_n \cap B \cap A_1^c \cap \ldots A_{n-1}^c = A_n \cap B$, by induction we have

$$\mu^*(B) = \mu^*(B \cap A_1^c \cap \ldots \cap A_n^c) + \sum_{i=1}^{n} \mu^*(B \cap A_i)$$

for all n. Now, since $A^c \subset A_1^c \cap \ldots \cap A_n^c$ we have $\mu^*(B \cap A_1^c \cap \ldots \cap A_n^c) \geq \mu^*(B \cap A^c)$ by the monotonicity of outer measure, so

$$\mu^*(B) \geq \mu^*(B \cap A^c) + \sum_{i=1}^{n} \mu^*(B \cap A_i) \,;$$

taking a limit in n now gives

$$\mu^*(B) \geq \mu^*(B \cap A^c) + \sum_{i=1}^{\infty} \mu^*(B \cap A_i) \,.$$

Since $A = \bigcup_{i\geq 1} A_i$, by subadditivity of outer measure we have $\mu^*(B \cap A) \leq \sum_{i=1}^{\infty} \mu^*(B \cap A_i)$, which with the previous bound gives

$$\mu^*(B) \geq \mu^*(B \cap A^c) + \mu^*(B \cap A).$$

Thus $\mu^*(B) = \mu^*(B \cap A^c) + \mu^*(B \cap A)$, so $A \in \mathcal{F}$ and $\mathcal{F}$ is indeed a σ-field.

Finally, note that in proving $\mathcal{F}$ is a σ-field, we also established that the restriction μ^+ of μ^* to $\mathcal{F}$ is countably additive on $\mathcal{F}$. Also, μ^+ is monotone and has $\mu^+(\emptyset) = 0$ by definition; so μ^+ is a measure on $\mathcal{F}$, as required. □

Proof of Theorem 2.1. Let μ^* be as in Proposition 2.3; then μ^* is an outer measure. We first verify that μ^* agrees with μ on $\mathcal{A}$. Fix $A \in \mathcal{A}$ and any sequence $(A_i, i \geq 1)$ of elements of $\mathcal{A}$ which cover A. Writing $B_n = A_n \setminus (A_1 \cup \ldots \cup A_n)$, then $B_n \subset A_n$ and $(A \cap B_n, n \geq 1)$ is another cover of A with elements of $\mathcal{A}$. Since $A = \bigcup_{n\geq 1} A \cap B_n$, by countable additivity for pre-measures we have

$$\begin{aligned}\mu(A) &= \sum_{n\geq 1} \mu(A \cap B_n) \\ &\leq \sum_{n\geq 1} \mu(B_n) \\ &\leq \sum_{n\geq 1} \mu(A_n)\,.\end{aligned}$$

Taking the infimum over covers $(A_n, n \geq 1)$ of A we obtain that $\mu(A) \leq \mu^*(A)$. Also, clearly $\mu(A) \geq \mu^*(A)$ since A covers itself; so $\mu(A) = \mu^*(A)$.

Next, let $\mathcal{F}$ be the collection of μ^*-additive sets, and let μ be the restriction of μ^* to $\mathcal{F}$; we are recycling notation here but this is OK since we already checked that μ and μ^* agree on their common domain of definition. By Lemma 2.4, $(\Omega, \mathcal{F}, \mu^*)$ is a measure space, and by the first paragraph we know that μ^* agrees with μ on $\mathcal{A}$. So, to complete the proof of the theorem it remains to show that $\mathcal{A} \subset \mathcal{F}$, or in other words that the sets in $\mathcal{A}$ are μ^*-additive.

So fix any set $A \in \mathcal{A}$ and any set $B \subset \Omega$. By subadditivity of μ^* we have

$$\mu^*(B) \leq \mu^*(A \cap B) + \mu^*(A^c \cap B)\,;$$

we need to prove the reverse inequality. If $\mu^*(B) = \infty$ then this is obvious, so we suppose $\mu^*(B) < \infty$. Fix $\epsilon > 0$; then we may find a cover $(A_n, n \geq 1)$ of B with elements of $\mathcal{A}$ such that

$$\sum_{n\geq 1} \mu(A_n) \leq \mu^*(B) + \epsilon$$

Finally, $(A \cap A_n, n \geq 1)$ is a cover of $A \cap B$ with elements of $\mathcal{A}$, and $(A^c \cap A_n, n \geq 1)$ is a cover of $A^c \cap B$ with elements of $\mathcal{A}$, so from the definition of μ^* we have

$$\begin{aligned}\mu^*(A \cap B) + \mu^*(A^c \cap B) &\leq \sum_{n\geq 1} \mu(A \cap A_n) + \sum_{n\geq 1} \mu(A^c \cap A_n) \\ &= \sum_{n\geq 1} (\mu(A \cap A_n) + \mu(A^c \cap A_n)) \\ &= \sum_{n\geq 1} \mu(A_n) \\ &\leq \mu(B) + \epsilon\,.\end{aligned}$$

Since $\epsilon > 0$ was arbitrary, it follows that $\mu^*(A \cap B) + \mu^*(A^c \cap B) \leq \mu^*(B)$, as required. □

The proof of Theorem 2.2 relies on one more algebraic/set theoretic closure property, which we now state. We say a collection $\mathcal{A}$ of subsets of a ground set Ω is a *λ-system* if $\Omega \in \mathcal{A}$ and additionally the following both hold. λ-system

(i) For all $E, F \in \mathcal{A}$ with $E \subset F$ we have $F \setminus E \in \mathcal{A}$.

(ii) For any increasing sequence $(A_n, n \geq 1)$ of subsets of $\mathcal{A}$ we have $\bigcup_{n\geq 1} A_n \in \mathcal{A}$.

By *increasing* we mean that $A_n \subset A_{n+1}$ for all $n \geq 1$.

Part (i) is similar to Exercise 2.1.

Exercise 2.6. *(i) If $\mathcal{A}$ is a σ-field then it is a π-system.*

(ii) If $\mathcal{A}$ is both a π-system and a λ-system then it is a σ-field.

(iii) Fix any collection $\{\mathcal{A}_i, i \in I\}$ of λ-systems with a common ground set. Then $\bigcap_{i\in I} \mathcal{A}_i$ is a λ-system.

Lemma 2.5 (Dynkin's π-system lemma)**.** *Let $\mathcal{P}$ be a π-system over a ground set Ω. Then*

$$\bigcap_{\{\mathcal{F}\supset\mathcal{P}:\mathcal{F} \text{ a } \sigma-\text{field}\}} \mathcal{F} = \bigcap_{\{\mathcal{F}\supset\mathcal{P}:\mathcal{F} \text{ a } \lambda-\text{system}\}} \mathcal{F}$$

Proof of Lemma 2.5. The left-hand side is $\sigma(\mathcal{P})$ by definition; temporarily writing $\lambda(\mathcal{P})$ for the right-hand side, we aim to show that $\sigma(\mathcal{P}) = \lambda(\mathcal{P})$.

Since σ-fields are λ-systems we automatically have $\sigma(\mathcal{P}) \supset \lambda(\mathcal{P})$, so to prove the lemma it suffices to show that $\lambda(\mathcal{P})$ is a σ-field. Moreover $\lambda(\mathcal{P})$ is a λ-system by Exercise 2.6 (ii), so by part (iii) of the same exercise, to show it is a σ-field we just have to show it is closed under intersections.

We proceed in two steps. For $E \in \lambda(\mathcal{P})$, say E is *cooperative* if $E \cap F \in \lambda(\mathcal{P})$ for all $F \in \mathcal{P}$, and that E is *helpful* if $E \cap F \in \lambda(\mathcal{P})$ for all $F \in \lambda(\mathcal{P})$.

If $E \in \mathcal{P}$ then $E \cap F \in \mathcal{P}$ for all $F \in \mathcal{P}$ since $\mathcal{P}$ is a π-system; so E is cooperative. Next, if E and E' are both cooperative and $E \subset E'$ then for all $F \in \mathcal{P}$ we have $E \cap F \in \lambda(\mathcal{P})$ and $E' \cap F$ in $\lambda(\mathcal{P})$. Since $\lambda(\mathcal{P})$ is a λ-system, it follows that

$$(E' \setminus E) \cap F = (E' \cap F) \setminus (E \cap F) \in \lambda(\mathcal{P}),$$

so $E' \setminus E$ is cooperative. Third, if $(E_n, n \geq 1)$ is an increasing sequence of cooperative sets then for all $F \in \mathcal{P}$ we have

$$F \cap \bigcup_{n\geq 1} E_n = \bigcup_{n\geq 1} F \cap E_n.$$

Each of the sets $F \cap E_n$ lies in $\lambda(\mathcal{P})$ since the E_n are cooperative. Since $(F \cap E_n, n \geq 1)$ is increasing and $\lambda(\mathcal{P})$ is a λ-system, it follows that $F \cap \bigcup_{n\geq 1} E_n \in \mathcal{P}$, so $\bigcup_{n\geq 1} E_n$ is cooperative.

We've now showed that the cooperative sets in $\lambda(\mathcal{P})$ contain $\mathcal{P}$ and are closed under monotone difference and monotone limits: they are a λ-system; so all sets in $\lambda(\mathcal{P})$ are cooperative.

We now bootstrap this argument. If $E \in \mathcal{P}$ then for any $F \in \lambda(\mathcal{P})$, since F is cooperative we have $E \cap F \in \lambda(\mathcal{P})$; so E is in fact helpful. Next, if E, E' are helpful and $E \subset E'$ then for all $F \in \lambda(\mathcal{P})$, $E' \cap F$ and $E \cap F$ both lie in $\lambda(\mathcal{P})$, so

$$(E' \setminus E) \cap F = (E' \cap F) \setminus (E \cap F) \in \lambda(\mathcal{P}).$$

Finally, if $(E_n, n \geq 1)$ is an increasing sequence of helpful sets then for all $F \in \lambda(\mathcal{P})$ and all $n \in \mathbb{N}$, $F \cap E_n \in \lambda(\mathcal{P})$, so

$$F \cap \bigcup_{n\geq 1} E_n = \bigcup_{n\geq 1} F \cap E_n \in \lambda(\mathcal{P}).$$

Thus $\bigcup_{n\geq 1} E_n$ is helpful. We've just showed that the helpful sets are a λ-system containing $\mathcal{P}$, so all sets in $\lambda(\mathcal{P})$ are helpful. But this means that means that $E \cap F \in \lambda(P)$ for all $E, F \in \lambda(P)$; so $\lambda(P)$ is closed under intersections, as required. □

We now show that Lemma 2.5 easily yields Theorem 2.2.

Proof of Theorem 2.2. Let μ_1,μ_2 be as in the theorem's statement. Fix any set $G \in \mathcal{P}$ with $\mu_1(G) < \infty$, and write $\Lambda = \{E \in \mathcal{F} : \mu_1(E \cap G) = \mu_2(E \cap G)\}$; then Λ contains $\mathcal{P}$ by definition, and in particular $\Omega \in \Lambda$.

Next, if $(E_n, n \geq 1)$ is an increasing sequence of sets in Λ then

$$\mu_1(\bigcup_{n\geq 1} E_n \cap G) = \lim_{n\geq 1} \mu_1(E_n \cap G) = \lim_{n\geq 1} \mu_2(E_n \cap G) = \mu_2(\bigcup_{n\geq 1} E_n \cap G)$$

where we've used countable additivity (as "continuity from below" in the form given in Exercise 2.3 (i)) for the first and third equalities. Thus $\bigcup_{n\geq 1} E_n \in \Lambda$. Also, If $E \subset F$ and $E, F \in \Lambda$ then

$$\mu_1(G \cap (F \setminus E)) = \mu_1(G \cap F) - \mu_1(G \cap E) = \mu_2(G \cap F) - \mu_2(G \cap E) = \mu_2(G \cap (F \setminus E)),$$

where we've used additivity of μ_1 and μ_2, together with the fact that $\mu_1(G) < \infty$, for the first and third equalities. Thus $F \setminus E \in \Lambda$. It follows that Λ is a λ-system containing $\mathcal{P}$, so Λ contains $\mathcal{F} = \sigma(\mathcal{P})$ by Lemma 2.5. If $\mu_1(\Omega) < \infty$ then by taking $G = \Omega$ the result follows.

For the general case, let $(\Omega^n, n \geq 1)$ be elements of $\mathcal{P}$ with $\Omega^n \uparrow \Omega$ and with $\mu_1(\Omega^n) < \infty$ for all n. Then for all $E \in \mathcal{F}$, since μ_1 and μ_2 are measures

$$\mu_1(E) = \lim_{n\to\infty} \mu_1(E \cap \Omega_n) = \lim_{n\to\infty} \mu_2(E \cap \Omega_n) = \mu_2(E),$$

where we have taken $G = \Omega_n$ to deduce that $\mu_1(E \cap \Omega_n) = \mu_2(E \cap \Omega_n)$. □

Remark. We say a measure μ on measurable space $(\Omega, \mathcal{F})$ is *σ-finite* if there exists an increasing sequence $(\Omega_n, n \geq 1)$ of elements of $\mathcal{F}$ with $\bigcup_{n\geq 1} \Omega_n = \Omega$ and with $\mu(\Omega_n) < \infty$ for all $n \geq 1$. Condition (b) in Dynkin's theorem is *stronger* than σ-finiteness, as it requires the approximating sets to in fact lie in $\mathcal{P}$. σ-finite

One might think that the requirement (b) in the statement of Theorem 2.2 could be weakened to simply require σ-finiteness. The result of Exercise 2.9, below, shows that this is not the case. (We must briefly postpone stating the example, until we have defined Borel sets - but they are coming very shortly.)

2.3. **Measures on $\mathbb{R}$.** The above development meant to be in service of defining probability measures in particular. The most fundamental example driving the theory is that of measures on $\mathbb{R}$. We already discussed the specification of probability distributions on $\mathbb{R}$ via their CDFs. Returning to this more formally and slightly more generally, we say $F : \mathbb{R} \to \mathbb{R}$ is a Stieltjes function if F is non-decreasing and right-continuous with left limits. If additionally $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$ then F is called a *cumulative distribution function*. Recall from above that $\mathcal{A}(\mathbb{R})$ is the set of finite unions of intervals of the form $\bigcup_{i=1}^n (a_i, b_i]$. We now define a function $\mu_F : \mathcal{A}(\mathbb{R}) \to [0, \infty]$ starting from the supposition that $\mu_F((a, b]) = F(b) - F(a)$. We are then forced by additivity to set $\mu_F(\bigcup_{i=1}^n (a_i, b_i]) = \sum_{i=1}^n F(b_i) - F(a_i)$ whenever $(a_i, b_i]$ are disjoint intervals. Stieltjes function Cumulative distribution function

Lemma 2.6. *For any Stieltjes function F, μ_F is a pre-measure on $\mathcal{A}(\mathbb{R})$.*

Proof. It is easy to verify that $\mathcal{A}(\mathbb{R})$ is a ring (this is Exercise 2.5). We show that μ_F is a pre-measure in three steps.

The first step is to check that μ_F is well-defined, i.e., that the expression in the definition of μ_F does not depend on how the elements of $\mathcal{A}(\mathbb{R})$ are expressed as finite disjoint unions. To see this, suppose that

$$L := \bigcup_{i=1}^n (a_i, b_i] = \bigcup_{j=1}^m (c_j, d_j]$$

are two ways of expressing the same element of $\mathcal{A}$ as a disjoint union. For each $i \in [n]$ and $j \in [m]$, if $(a_i, b_i] \cap (c_j, d_j]$ is non-empty we denote the intersection by $(\ell_{ij}, r_{ij}]$. Then

$$\begin{aligned}\sum_{i=1}^n F(b_i) - F(a_i) &= \sum_{i=1}^n \sum_{\{j:(a_i,b_i]\cap(c_j,d_j]\neq\emptyset\}} F(r_{ij}) - F(\ell_{ij}) \\ &= \sum_{j=1}^n \sum_{\{i:(a_i,b_i]\cap(c_j,d_j]\neq\emptyset\}} F(r_{ij}) - F(\ell_{ij}) = \sum_{j=1}^m F(d_j) - F(c_j),\end{aligned}$$

so μ_F is indeed well-defined.

We next check that μ_F is additive. This is easy: if $\bigcup_{i=1}^n (a_i, b_i]$ and $\bigcup_{j=1}^m (c_j, d_j]$ are disjoint elements of $\mathcal{A}(\mathbb{R})$ then

$$\mu_F\left(\bigcup_{i=1}^n (a_i, b_i] \cup \bigcup_{j=1}^m (c_j, d_j]\right) = \sum_{i=1}^n (F(b_i) - F(a_i)) + \sum_{j=1}^m (F(d_j) - F(c_j)),$$

which is indeed the sum of the measures of the two elements of $\mathcal{A}(\mathbb{R})$.

Finally, we check that μ_F is a pre-measure. For this, suppose that

$$L := \bigcup_{i=1}^n (a_i, b_i] = \bigcup_{j=1}^\infty (c_j, d_j]$$

where the two unions are over disjoint intervals. Then for all $m \in \mathbb{N}$,

$$\bigcup_{i=1}^n (a_i, b_i] \supset \bigcup_{j=1}^\infty (c_j, d_j],$$

Thus, by monotonicity of μ_F,

$$\mu_F(\bigcup_{i=1}^n (a_i, b_i]) \geq \sup_{m \geq 1} \mu_F(\bigcup_{j=1}^m (c_j, d_j]) = \sum_{j=1}^\infty \mu_F(c_j, d_j]\,;$$

to complete the proof, we must show that in fact equality holds.

Suppose for a contradiction that $\mu_F(L) = \sum_{j=1}^\infty \mu_F(c_j, d_j] + 2\epsilon$, for some $\epsilon > 0$, and write $\Delta_m := L \setminus \bigcup_{i=1}^m (c_i, d_i]$. Note that $\Delta_m \in \mathcal{A}$ — it is a difference of finite unions of intervals — and $\Delta_m \downarrow 0$ as $m \to \infty$. Also, since $L = \Delta_m \cup \bigcup_{i=1}^m (c_i, d_i])$ is a disjoint union, it follows that

$$\mu_F(\Delta_m) = \mu_F(L) - \sum_{i=1}^m \mu_F(c_i, d_i] \geq 2\epsilon$$

for all m.

Choose $D_m \in \mathcal{A}$ with $\overline{D_m} \subset \Delta_m$ and such that $\mu_F(\Delta_m \setminus D_m) \leq \epsilon/2^m$ for all m.[4] Since

$$\Delta_m = \bigcap_{i=1}^m \Delta_i = \bigcap_{i=1}^m D_i \cup (\Delta_i \setminus D_i) \subseteq \bigcap_{i=1}^m D_i \cup \bigcup_{i=1}^m (\Delta_i \setminus D_i),$$

by monotonicity

$$2\epsilon \leq \mu_F(\Delta_m) \leq \mu_F(\bigcap_{i=1}^m D_i) + \sum_{i=1}^m \mu_F(\Delta_i \setminus D_i) \leq \mu_F(\bigcap_{i=1}^m D_i) + \epsilon.$$

Thus $\mu_F(\bigcap_{i=1}^m D_i) \geq \epsilon$ for all m, so $\bigcap_{i=1}^m \overline{D_i} \neq \emptyset$ for all m, so $\bigcap_{i=1}^\infty \Delta_i \supset \bigcap_{i=1}^\infty \overline{D_i} \neq \emptyset$, contradicting the fact that $\Delta_m \downarrow \emptyset$ as $m \to \infty$. □

The σ-field generated by $\mathcal{A}(\mathbb{R})$ is called the *Borel σ-field*, and denoted $\mathcal{B}(\mathbb{R})$; its elements are called *Borel sets* of $\mathbb{R}$. The next exercise asks you to show that $\mathcal{B}(\mathbb{R})$ is the smallest σ-field containing all open sets in $\mathbb{R}$.

$\mathcal{B}(\mathbb{R})$.

Exercise 2.7. *Show that* $\sigma(\mathcal{A}(\mathbb{R})) = \sigma(\{U \subset \mathbb{R} : U \text{ open}\})$.

More generally, given a topological space M, the Borel σ-field over M is defined to be the σ-field generated by the open sets, $\mathcal{B}(M) := \sigma(\{U \subset M : U \text{ open})$.

$\mathcal{B}(M)$.

With Lemma 2.6 under our belt, it now follows easily that Stieltjes functions $\mathbb{R}$ uniquely determine measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Theorem 2.7. *Let F be a Stieltjes function. Then there exists a unique measure μ on $\mathcal{B}(\mathbb{R})$ such that $\mu(a, b] = F(b) - F(a)$ for all $-\infty < a \leq b < \infty$.*

[4] Not hard to see this is possible - add a proof?

Proof. Write $\mathcal{P} = \{(a,b] : -\infty < a \leq b < \infty\}$. By Lemma 2.6, there exists a pre-measure μ on $\mathcal{A}(\mathbb{R})$ such that $\mu(a,b] = F(b) - F(a)$ for all $(a,b) \in \mathcal{P}$. By the Carathéodory Extension Theorem, Theorem 2.1, μ extends to a measure $\mu^+ : \mathcal{F} \to [0,\infty]$ for some σ-field $\mathcal{F}$ containing $\mathcal{A}(\mathbb{R})$. Since $\mathcal{B}(\mathbb{R})$ is the smallest σ-field containing $\mathcal{A}(\mathbb{R})$, the restriction of μ^+ to $\mathcal{B}(\mathbb{R})$ is well-defined, is a measure on $\mathcal{B}(\mathbb{R})$ and has $\mu^+(a,b] = \mu(a,b] = F(b) - F(a)$ for all $(a,b] \in \mathcal{P}$. This proves existence.

Now suppose that μ_1 and μ_2 are measures on $\mathcal{B}(\mathbb{R})$ satisfying the hypotheses of the theorem. Then μ_1 and μ_2 agree on $\mathcal{P}$. But $\mathcal{P}$ is a π-system. Clearly $\sigma(\mathcal{P})$ contains $\mathcal{A}(\mathbb{R})$, so so we must have $\sigma(\mathcal{P}) = \sigma(\mathcal{A}(\mathbb{R})) = \mathcal{B}(\mathbb{R})$. It follows by Dynkin's theorem, Theorem 2.2, that $\mu_1 \equiv \mu_2$. This proves uniqueness. □

The above proof refers to "some σ-field $\mathcal{F}$ containing $\mathcal{A}(\mathbb{R})$". Looking back at the statement of of the Carathéodory lemma reveals that the σ-field $\mathcal{F}$ consists precisely of the μ^*-additive sets.

The next exercise reveals more information about the collection of μ^*-additive sets, and its relation to the Borel σ-fields. The exercise after that provides an example which shows that condition (b) in Dynkin's theorem can not be replaced by σ-finiteness. The following definition features in first of the two exercises: we say a measure space $(\Omega, \mathcal{F}', \mu')$ *extends* another measure space $(\Omega, \mathcal{F}, \mu)$ if $\mathcal{F} \subseteq \mathcal{F}'$ and $\mu'|_{\mathcal{F}} \equiv \mu$.

Exercise 2.8. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Say $N \in \mathcal{F}$ is a* null set *if $\mu(N) = 0$. Say that $(\Omega, \mathcal{F}, \mu)$ is* complete *if for any null set N, for all $M \subset N$ we have $M \in \mathcal{F}$.*

(a) Write $\overline{\mathcal{F}} := \bigcap_{\{(\Omega,\mathcal{F}',\mu') \text{ extending } (\Omega,\mathcal{F},\mu):(\Omega,\mathcal{F}',\mu') \text{ complete}\}} \mathcal{F}'$. Prove that

$$\overline{\mathcal{F}} = \{E \cup M : E \in \mathcal{F}, M \subset N \text{ for some null set } N \in \mathcal{F}\}.$$

(b) Let $\mu^* : 2^\Omega \to [0,\infty]$ be an outer measure on some ground set Ω, and let $\mathcal{F} = \{A \subset \Omega : A \text{ is } \mu^*\text{-additive}\}$. Show that $(\Omega, \mathcal{F}, \mu^*|_{\mathcal{F}})$ is complete.

(c) Let μ be the Lebesgue pre-measure on $\mathcal{A}(\mathbb{R})$, i.e., with $\mu(a,b] = b - a$ for bounded intervals $(a,b] \subset \mathbb{R}$. Let μ^* be the corresponding outer measure on $\mathbb{R}$, and let $\mathcal{L}(\mathbb{R}) = \{S \subset \mathbb{R} : S \text{ is } \mu^*\text{-additive}\}$. Show that $\mathcal{L}(\mathbb{R})$ is the completion of $\mathcal{B}(\mathbb{R})$.

NB: For (c) you will need the σ-finiteness of Lebesgue measure.

The set $\mathcal{L}(\mathbb{R})$ is known as the *Lebesgue* σ-field (actually, I have only ever seen it called the Lebesgue σ-algebra, but I decided to call them σ-fields, and I'm sticking to it).

Exercise 2.9. *Consider the measures μ_1, μ_2 on $\mathcal{B}(\mathbb{R})$ defined by $\mu_1(B) = |B \cap \mathbb{Q}|$, $\mu_2(B) = 2|B \cap \mathbb{Q}|$.*

(a) Show that μ_1 and μ_2 are σ-finite measures.

(b) Show that $\mu_1(A) = \mu_2(A)$ for all $A \in \mathcal{A}(\mathbb{R})$.

Given a measurable space $(\Omega, \mathcal{F})$ a set S and a function $f : \Omega \to S$, the *push-forward* of $\mathcal{F}$ to S is the set $f^*(\mathcal{F}) = \{B \subset S : f^{-1}(B) \in \mathcal{F}\}$.

Exercise 2.10. *Show that the push-forward $f^*(\mathcal{F})$ is a σ-field.*

2.4. **Independent events.** A *probability space* is a measure space $(\Omega, \mathcal{F}, \mathbf{P})$ with $\mathbf{P}(\Omega) = 1$. Elements of $\mathcal{F}$ are called *events*; elements of Ω are called *elementary events*.[5]

Probability space

We say that events $(E_i, i \in I)$ are *mutually independent* if for all $J \subset I$ finite,

Independent events

$$\mathbf{P}\left\{\bigcap_{j\in J} E_j\right\} = \prod_{j\in J} \mathbf{P}\{E_j\}. \tag{2.1}$$

(Often the word "mutually" is omitted.) For $k \geq 1$, we say the events $(E_i, i \in I)$ are *k-wise independent* if (2.1) holds for all $J \subset I$ with $|J| \leq k$. In particular, they are pairwise independent if $\mathbf{P}\{E_i \cap E_j\} = \mathbf{P}\{E_i\}\mathbf{P}\{E_j\}$ for any distinct $i, j \in I$.

Exercise 2.11. *In this exercise we say that an event E in a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is* non-trivial *if $\mathbf{P}\{E\} \in (0,1)$.*

[5] An unfortunate aspect of this terminology: elementary events need not be events! But it is what it is.

(a) *Let $k \in \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $(E_1, \ldots, E_k)$ be nontrivial, independent events in $\mathcal{F}$. Prove that $|\Omega| \geq 2^k$.*

(b) *Construct a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ with $|\Omega| = 2^{k-1}$ and nontrivial events $(E_1, \ldots, E_k)$, such that for any $1 \leq i \leq n$, the events $(E_j, j \in [k] \setminus \{i\})$ are mutually independent, but $(E_1, \ldots, E_k)$ are not mutually independent.*

The following example is further developed in the homework (and inspired by the use of Rademacher random variables in James Norris's "Probability and measure" lecture notes). It is a hands-on way to model an infinite sequence of independent fair coin tosses. Let $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}([0,1]) = \mathcal{B}(\mathbb{R})|_{[0,1]}$, and let $\mathbf{P}$ be Lebesgue measure on $[0,1]$, which isoften called the *uniform* probability measure in this context; then $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space. For $k \geq 1$ define the event

$$A_k = \bigcup_{\substack{0 \leq i < 2^k \\ i \text{ even}}} \left(\frac{i}{2^k}, \frac{i+1}{2^k} \right]. \tag{2.2}$$

So $A_1 = (0, 1/2]$, $A_2 = (0, 1/4] \cup (1/2, 3/4]$, and so on.

Exercise 2.12. *Show that $(A_k, k \geq 1)$ are mutually independent.*

Note that A_i may be thought of as the set of $x \in (0,1]$ for which the i'th bit in the binary expansion is zero (provided we adopt the convention that we never use infinite strings of zeros in our binary representation).

The *Borel–Cantelli lemmas* are basic and important workhorses of probability theory; stating them will additionally help us away from the language of sets and toward probabilistic terminology. Given a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and events $(E_n, n \geq 1)$, we define

$$\limsup_{n\to\infty} E_n := \bigcap_{n \geq 1} \bigcup_{m \geq n} E_m = \{\omega \in \Omega : \omega \in E_n \text{ for infinitely many } n\}.$$

(In fact, this definition makes sense for any sequence of sets $(E_n, n \geq 1)$ over a common ground set Ω.) Thinking probabilistically, if $\omega \in \limsup_{n\to\infty} E_n$ then infinitely many of the events E_n occur; we therefore introduce $\{E_n \text{ occurs infinitely often}\}$ or simply $\{E_n \text{ i.o.}\}$ as alternative notation for the set $\limsup_{n\to\infty} E_n$.

Similarly, we define

$$\liminf_{n\to\infty} E_n := \bigcup_{n \geq 1} \bigcap_{m \geq n} E_m = \{\omega \in \Omega : \omega \in E_n \text{ for all but finitely many } n\}.$$

Note that $(\limsup_{n\to\infty} E_n)^c = \liminf_{n\to\infty}(E_n^c)$.

As an example, for the events $(A_k, k \geq 1)$ described above, we have

$$\limsup_{n\to\infty} A_n = \{x \in [0,1] : \text{ there are infinitely many zeros in any binary expansion of } x\}, \text{ and}$$
$$\liminf_{n\to\infty} A_n = \{x \in [0,1] : x = k/2^n, \text{ for some integers } n, k \text{ with } n \geq 1, 0 \leq k \leq 2^n\}.$$

Lemma 2.8 (First Borel–Cantelli Lemma)**.** *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $(E_n, n \geq 1)$ be events in $\mathcal{F}$. If $\sum_{n\geq 1} \mathbf{P}\{E_n\} < \infty$ then $\mathbf{P}\{E_n \text{ i.o}\} = 0$.*

Proof. Fix $\epsilon > 0$. Then there exists n_0 such that $\sum_{m \geq n_0} \mathbf{P}\{E_m\} < \epsilon$, so by monotonicity and subadditivity of measures,

$$\mathbf{P}\{E_n \text{ i.o}\} \leq \mathbf{P}\left\{ \bigcap_{n \geq n_0} \bigcup_{m \geq n} E_m \right\} \leq \mathbf{P}\left\{ \bigcup_{m \geq n_0} E_m \right\} \leq \sum_{m \geq n_0} \mathbf{P}\{E_m\} < \epsilon\,.$$

Since $\epsilon > 0$ was arbitrary, the result follows. □

Lemma 2.9 (Second Borel–Cantelli Lemma)**.** *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $(E_n, n \geq 1)$ be mutually independent events in $\mathcal{F}$. If $\sum_{n\geq 1} \mathbf{P}\{E_n\} = \infty$ then $\mathbf{P}\{E_n \text{ i.o}\} = 1$.*

Proof. Note that by definition,

$$\{E_n \text{ i.o}\}^c = \liminf_{n\to\infty}(E_n^c) = \bigcup_{n\geq 1}\bigcap_{m\geq n} E_m^c,$$

so by subadditivity

$$\mathbf{P}\{\{E_n \text{ i.o}\}^c\} \leq \sum_{n\geq 1} \mathbf{P}\left\{\bigcap_{m\geq n} E_m^c\right\}.$$

To prove the lemma we'll show that the summands on the right are all zero.

Writing $p_n = \mathbf{P}\{E_n\}$, for all $1 \leq n \leq N$, by monotonicity and independence we have

$$\mathbf{P}\left\{\bigcap_{m\geq n} E_m^c\right\} \leq \mathbf{P}\left\{\bigcap_{m=n}^{N} E_m^c\right\} = \prod_{m=n}^{N}(1-p_n).$$

Since this holds for all N, and since $1 - p_n \leq e^{-p_n}$, it follows that

$$\mathbf{P}\left\{\bigcap_{m\geq n} E_m^c\right\} \leq \prod_{m=n}^{\infty}(1-p_n) \leq e^{-\sum_{m=n}^{\infty} p_n} = 0,$$

as required. □

The two Borel-Cantelli lemmas together show that if $(E_n, n \geq 1)$ is any sequence of independent events, then $\mathbf{P}\{E_n \text{ i.o.}\} \in \{0, 1\}$. This is a first instance of a *zero-one law*, and a special case of Kolmogorov's zero-one law, which you will meet quite shortly.

We conclude the section by defining independence of σ-fields and establishing a sufficient condition for such independence. Given a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a collection $(\mathcal{G}_i, i \in I)$ of subsets of $\mathcal{F}$, we say that $(\mathcal{G}_i, i \in I)$ are independent if for all $J \subset I$ finite and any events $(E_j, j \in J)$ with each $E_j \in \mathcal{G}_j$, we have

$$\mathbf{P}\left\{\bigcap_{j\in J} E_j\right\} = \prod_{j\in J} \mathbf{P}\{E_j\}.$$

Proposition 2.10. *Fix a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and let $\mathcal{P}, \mathcal{Q}$ be π-systems in $\mathcal{F}$. If $\mathbf{P}\{A \cap B\} = \mathbf{P}\{A\}\mathbf{P}\{B\}$ for all $A \in \mathcal{P}, B \in \mathcal{Q}$, then $\sigma(\mathcal{P})$ and $\sigma(\mathcal{Q})$ are independent.*

Proof. First fix $A \in \mathcal{P}$ and define measures $\mu_A, \mathbf{P}_A$ on $\sigma(\mathcal{Q})$ by

$$\mu_A(B) = \mathbf{P}\{A\}\mathbf{P}\{B\} \text{ and } \mathbf{P}_A(B) = \mathbf{P}\{A \cap B\}.$$

Then $\mu_A(B) = \mathbf{P}_A(B)$ for all $B \in \mathcal{Q}$, and $\mu_A(\Omega) = \mathbf{P}\{A\} = \mathbf{P}_A(\Omega)$, so $\mu_A = \mathbf{P}_A$ by Dynkin's theorem. Thus

$$\mathbf{P}\{A \cap B\} = \mathbf{P}\{A\}\mathbf{P}\{B\}$$

for all $A \in \mathcal{P}$, $B \in \sigma(\mathcal{Q})$.

Next, fix $B \in \sigma(\mathcal{Q})$ and define measures $\nu^B, \mathbf{P}^B$ on $\sigma(\mathcal{P})$ by

$$\nu^B(A) = \mathbf{P}\{A\}\mathbf{P}\{B\} \text{ and } \mathbf{P}^B(A) = \mathbf{P}\{A \cap B\}.$$

Then $\nu^B(A) = \mathbf{P}^B(A)$ for $A \in \mathcal{P}$, and $\nu^B(\Omega) = \mathbf{P}\{B\} = \mathbf{P}^B(\Omega)$, so by Dynkin's theorem we have $\nu^B = \mathbf{P}^B$. Thus $\mathbf{P}\{A \cap B\} = \mathbf{P}\{A\}\mathbf{P}\{B\}$ for all $A \in \sigma(\mathcal{P})$ and $B \in \sigma(\mathcal{Q})$, i.e., $\sigma(\mathcal{P})$ and $\sigma(\mathcal{Q})$ are independent. □

Exercise 2.13. *Fix a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and π-systems $(\mathcal{P}_i, i \in I)$ which are subsets of $\mathcal{F}$. Then the σ-fields $(\sigma(\mathcal{P}_i), i \in I)$ are independent if and only if $\mathbf{P}\left\{\bigcap_{j\in J} E_j\right\} = \prod_{j\in J} \mathbf{P}\{E_j\}$ for all $J \subset I$ finite and any events $E_j \in \mathcal{P}_j$.*

3. **Random variables**

Much of the richness of probability theory arises from the interaction of independence with random variables, but to explore that, we need to define random variables first!

To begin, given measurable spaces $(R, \mathcal{R})$ and $(S, \mathcal{S})$, a $(\mathcal{R}/\mathcal{S})$-*measurable map* is a function $f : R \to S$ such that $f^{-1}(E) \in \mathcal{R}$ for all $E \in \mathcal{S}$.[6] If R and S are topological spaces and $\mathcal{R}, \mathcal{S}$ are the Borel σ-algebras, then f is also called a *Borel* function.

Fix a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. A (real) *random variable* is a $(\mathcal{F}/\mathcal{B}(\mathbb{R}))$-measurable function $X : \Omega \to \mathbb{R}$. In other words, random variables are just measurable maps but where the domain happens to be the ground set of a probability space. Real random variables and extended real random variables are the bread and butter of the course. The laws of large numbers are the jam. Basic measure theory is the plate. Enough with that metaphor. (We write $\mathbb{R}^* = \mathbb{R} \cup \{\pm\infty\}$ for the extended real line; its open sets are generated by those of $\mathbb{R}$ together with sets of the form $(x, \infty]$ and $[-\infty, x)$ for $x \in \mathbb{R}$; an extended real random variable is a $(\mathcal{F}/\mathcal{B}(\mathbb{R}^*))$-measurable map $X : \Omega \to \mathbb{R}^*$.) Oh, and random variables taking values in more general spaces are the *croissants au beurre*. If M is a topological space and $X : \Omega \to M$ is $(\mathcal{F}/\mathcal{B}(M))$-measurable then we call X an M-valued random variable; if $(S, \mathcal{S})$ is a measurable space and $X : \Omega \to S$ is $(\mathcal{F}/\mathcal{S})$-measurable then we call X an S-valued random variable.

For a function $X : \Omega \to \mathbb{R}$ or $X : \Omega \to \mathbb{R}^*$, it's very useful to introduce the notation $\{X \leq r\} := \{\omega \in \Omega : X(\omega) \leq r\}$ and to think of this set as "the event that $X \leq r$". More generally for a function $X : R \to S$ and $U \subset S$ we write $\{X \in U\} := X^{-1}(U)$.

Before diving into the theory, it's worth motivating ourselves (and honing our intuition) by considering an example. We revisit the events A_k defined in (2.2), above, and define $R_k : [0, 1] \to \mathbb{R}$ by $R_k = \mathbf{1}_{[A_k]}$, so $R_k(x) = 1$ if and only if $x \in A_k$.

Exercise 3.1. *Show that R_k is $\mathcal{B}([0, 1])/\mathcal{B}(\mathbb{R})$-measurable.*

Under the uniform probability measure on $[0, 1]$, we have

$$\mathbf{P}\{R_k = 1\} := \mathbf{P}\{\{x : R_k(x) = 1\}\} = \mathbf{P}\{A_k\} = \frac{1}{2}.$$

This agrees with the intuition that for a uniformly random point in $[0, 1]$, each bit of the binary expansion is equally likely to be zero or one. Moreover, intuition suggests that these bits should be independent, and that the asymptotic proportion of ones in the sequence $(R_n, n \geq 1)$ should be $1/2$. More precisely, we expect that

$$\mathbf{P}\left\{\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} R_n = \frac{1}{2}\right\} = 1. \tag{3.1}$$

To make rigorous sense of this assertion, we first need to know that

$$\left\{\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} R_n = \frac{1}{2}\right\} = \left\{x \in [0, 1] : \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} R_n(x) = \frac{1}{2}\right\} \tag{3.2}$$

is a measurable set (otherwise its probability is not defined). Fortunately, this is not hard to see; the closure properties of σ-fields allow us to perform essentially any operations we please with random variables and obtain other random variables, provided we perform at most countably many operations in total. The next theorem provides a useful time-saving device for proving measurability of random variables; the subsequent exercise shows that many of the basic operations of arithmetic and analysis preserve measurability, and in particular implies that the set in (3.2) is measurable.

[6] Notice the similarity to the definition of continuous functions between topological spaces.

Theorem 3.1. *Let $(R, \mathcal{R})$ and $(S, \mathcal{S})$ be measurable spaces and let $f : R \to S$. Suppose that there is $\mathcal{A} \subset \mathcal{S}$ with $\sigma(\mathcal{A}) = \mathcal{S}$ such that $f^{-1}(A) \in \mathcal{R}$ for all $A \in \mathcal{A}$. Then f is $(\mathcal{R}/\mathcal{S})$-measurable.*

Proof. Let $\mathcal{S}_0 = \{E \in \mathcal{S} : f^{-1}(E) \in \mathcal{R}\}$. Then $\mathcal{A} \subset \mathcal{S}_0$ by assumption. Also, if $E \in \mathcal{S}_0$ then

$$f^{-1}(E^c) = \{r \in R : f(r) \in E^c\} = R \setminus \{r \in R : f(r) \in E\} = (f^{-1}(E))^c \in \mathcal{R},$$

so $E^c \in \mathcal{S}_0$. Similarly, if $(E_n, n \geq 1)$ are in $\mathcal{S}_0$ and $E_n \uparrow E_\infty$ then

$$f^{-1}(E_\infty) = \{r \in R : f(r) \in E_\infty\} = \bigcup_{n\geq 1} \{r \in R : f(r) \in E_n\} = \bigcup_{n \geq 1} f^{-1}(E_n) \in \mathcal{R},$$

so $E_\infty \in \mathcal{S}_0$. Thus $\mathcal{S}_0$ is a σ-field, so equals $\mathcal{S}$. □

Here are some examples of how the theorem is useful. Fix a probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

- If $X : \Omega \to \mathbb{R}$ satisfies that $\{X \leq r\} = X^{-1}(-\infty, r] \in \mathcal{F}$ for all $r \in \mathbb{R}$, then X is a real random variable (it is $(\mathcal{F}/\mathcal{B}(\mathbb{R}))$-measurable).
- If $X : \Omega \to \mathbb{R}$ is a real random variable and $f : \mathbb{R} \to \mathbb{R}$ is continuous then $f(X)$ is another random variable. (Since if $U \subset \mathbb{R}$ is open, then $f^{-1}(U)$ is open, so $\{f(X) \in U\} = (f \circ X)^{-1}(U) = X^{-1}(f^{-1}(U)) \in \mathcal{B}(\mathbb{R})$; and the open sets are a π-system generating $\mathcal{B}(\mathbb{R})$.)
- If $\mathrm{X}_J = (X_j, j \in J)$ is a finite collection of random variables then X_J may be viewed as a function from Ω to $\mathbb{R}^J$, sending ω to $(X_j(\omega), j \in J)$. The collection

$$\mathcal{P}_J := \left\{ \prod_{j \in J} (-\infty, b_j] : b_j \in \mathbb{R} \text{ for } j \in J \right\} \tag{3.3}$$

  is a π-system generating the Borel sets $\mathcal{B}(\mathbb{R}^J)$. For any element $\prod_{j\in J}(-\infty, b_j]$ of $\mathcal{P}_J$, we have

$$\mathrm{X}_J^{-1}\Big(\prod_{j\in J} (-\infty, b_j] \Big) = \bigcap_{j \in J} X_j^{-1}(b_j) \in \mathcal{F},$$

  since $\mathcal{F}$ is closed under finite intersections. Thus X is an $\mathbb{R}^J$-valued random variable.
- If R and S are topological spaces, and $h : R \to S$ is such that $h^{-1}(U) \in \mathcal{B}(R)$ for all open $U \subset S$, then h is a Borel function.

The next exercise asks you to check various closure properties of the collection of real-valued measurable maps. Some of these require enlarging the target space from the real numbers to the *extended* real numbers $\mathbb{R}^* := \mathbb{R} \cup \{-\infty, \infty\}$. The open sets of $\mathbb{R}^*$ are generated by $\mathcal{A} = \{(a, b), a, b \in \mathbb{R}\} \cup \{(a, \infty], a \in \mathbb{R}\} \cup \{[-\infty, b), b \in \mathbb{R}\}$, so $\mathcal{A}$ also generates the Borel sets of $\mathbb{R}^*$: that is, $\mathcal{B}(\mathbb{R}^*) = \sigma(\mathcal{A})$.

Exercise 3.2. *Let $(\Omega, \mathcal{F})$ be a measurable space and let X, Y, and $(X_n, n \geq 1)$ be $(\mathcal{F}/\mathcal{B}(\mathbb{R}))$-measurable maps from Ω to $\mathbb{R}$.*

(a) Prove that $\mathbf{1}_{[X\geq 0]}, X + Y, XY, (X/Y)\mathbf{1}_{[Y\neq 0]}$ are all $(\mathcal{F}/\mathcal{B}(\mathbb{R}))$-measurable.
(b) Prove that $\sup_{n\geq 1} X_n, \inf_{n\geq 1} X_n, \limsup_{n\geq 1} X_n$ and $\liminf_{n\geq 1} X_n$ are all $(\mathcal{F}/\mathcal{B}(\mathbb{R}^*))$-measurable.
(c) Prove that if Z is any of the four expressions from part (b), then $Z\mathbf{1}_{[Z\in\mathbb{R}]}$ is $(\mathcal{F}/\mathcal{B}(\mathbb{R}))$-measurable.
(d) Prove that if $f : \mathbb{R}^n \to \mathbb{R}$ is $(\mathcal{B}(\mathbb{R}^n)/\mathcal{B}(\mathbb{R}))$-measurable then $f(X_1, \ldots, X_n)$ is $(\mathcal{F}/\mathcal{B}(\mathbb{R}))$-measurable.

Given a sequence $(a_n, n \geq 1)$ of real numbers, we say that $\lim_{n\to\infty} a_n$ *exists* if either there is $a \in \mathbb{R}$ such that $\lim_{n\to\infty} a_n = a$, or if $\lim_{n\to\infty} a_n = \infty$ or $\lim_{n\to\infty} a_n = -\infty$.

Proposition 3.2. *If $(X_n, n \geq 1)$ is a sequence of random variables on probability space $(\Omega, \mathcal{F}, \mathbf{P})$, then*

$$E := \Big\{ \lim_{n\to\infty} X_n \text{ exists} \Big\} = \Big\{ \omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \text{ exists} \Big\}$$

is an element of $\mathcal{F}$.

Remove **P** since it is not needed in the proposition? But trying to encourage readers to think probabilistically...

Proof. By Exercise 3.2 (b), $\overline{X} := \limsup_{n\geq 1} X_n$ and $\underline{X} := \liminf_{n\geq 1} X_n$ are extended real-valued random variables, so

$$E_\infty := \{\lim_{n\to\infty} X_n = \infty\} = \{\underline{X} = \infty\}$$

is an event, and

$$E_{-\infty} := \{\lim_{n\to\infty} X_n = -\infty\} = \{\overline{X} = -\infty\}$$

is an event. Also,

$$E_{\text{bd}} := \{(X_n, n \geq 1) \text{ is a bounded sequence}\} = \{-\infty < \underline{X}\} \cap \{\overline{X} < \infty\}$$

is an event, so

$$E_{\text{fin}} = \left\{\lim_{n\to\infty} X_n \text{ exists and is finite}\right\} = E_{\text{bd}} \cap \bigcap_{m\in\mathbb{N}} \{\overline{X} - \underline{X} < 1/m\}$$

is an event. Since $E = E_\infty \cup E_{-\infty} \cup E_{\text{fin}}$, this completes the proof. □

Exercise 3.3. *Let $(\Omega, \mathcal{F})$ be a measurable space and let $(X_n, n \geq 1)$ be $(\mathcal{F}/\mathcal{B}(\mathbb{R}))$-measurable maps from Ω to $\mathbb{R}$. Write $E := \{\omega : \lim_{n\to\infty} X_n(\omega) \text{ exists}\}$. Prove that, defining $X_\infty : \Omega \to \mathbb{R}^*$ by*

$$X_\infty(\omega) = \begin{cases} \lim_{n\to\infty} X_n(\omega) & \text{if } \omega \in E \\ 0 & \text{otherwise,} \end{cases}$$

then X_∞ is $(\mathcal{F}/\mathcal{B}(\mathbb{R}^*))$-measurable.

3.1. **Generated σ-fields.** Fix a set R and a measurable space $(S, \mathcal{S})$. Given a collection $(X_i, i \in I)$ of functions from R to S, we define

$$\sigma(X_i, i \in I) := \sigma(\{X_i^{-1}(E) : i \in I, E \in \mathcal{S}\}) = \bigcap_{\substack{\mathcal{F} \text{ a } \sigma-\text{field over } R \\ \forall\, i\in I,\ X_i \text{ is } (\mathcal{F}/\mathcal{S})-\text{measurable}}} \mathcal{F}\,.$$

In words, $\sigma(X_i, i \in I)$ is the smallest σ-field over R to yield measurability of all the maps $(X_i, i \in I)$. If $(R, \mathcal{R})$ is a measurable space and the functions $(X_i, i \in I)$ are all $(\mathcal{R}/\mathcal{S})$-measurable, then $\sigma(X_i, i \in I) \subset \mathcal{R}$ by definition.

The most important example is that of a collection of real random variables $(X_i, i \in I)$ over a common probability space. For $i \in I$ we have $\sigma(X_i) = \{\{X_i \in B\}, B \in \mathcal{B}(\mathbb{R})\} = \sigma(\{X_i \leq b\}, b \in \mathbb{R})$, so it follows that

$$\sigma(X_i, i \in I) = \sigma\left(\bigcup_{i\in I} \{\{X_i \leq b\} : b \in \mathbb{R}\}\right)$$

For any $J \subset I$ finite and $(b_j, j \in J) \in \mathbb{R}^J$, it follows that

$$\{X_j \leq b_j, j \in J\} = \bigcap_{j\in J} \{X_j \leq b_j\} \in \sigma(X_i, i \in I),$$

so we may also write

$$\sigma(X_i, i \in I) = \sigma\left(\{X_j \leq b_j, j \in J\} : J \subset I \text{ finite}, (b_j, j \in J) \in \mathbb{R}^J\right)\,.$$

Exercise 3.4. *Let $(X_i, i \in I)$ be random variables defined on a common probability space. Show that*

$$\sigma(X_i, i \in I) = \bigcup_{J\subset I, J \text{ countable}} \sigma(X_j, j \in J)\,.$$

Exercise 3.5 (Doob-Dynkin Lemma)**.** *Fix a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and let $X, Y : \Omega \to \mathbb{R}$ be random variables. Show that Y is $(\sigma(X)/\mathcal{B}(\mathbb{R}))$-measurable if and only if there exists a Borel function $f : \mathbb{R} \to \mathbb{R}$ such that $Y = f(X)$.*

3.2. **Independence of random variables.** We say a collection of random variables $(X_i, i \in I)$ over a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$ are *mutually independent* if the σ-fields $(\sigma(X_i), i \in I)$ are mutually independent. (We'll often drop the word "mutually".) In other words, $(X_i, i \in I)$ are independent if for any $J \subset I$ finite, and any Borel sets $(B_j, j \in J)$, we have

$$\mathbf{P}\{X_j \in B_j, j \in J\} = \prod_{j \in J} \mathbf{P}\{X_j \in B_j\}.$$

Proposition 3.3. *Real random variables $(X_i, i \in I)$ defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ are mutually independent if and only if for all $J \subset I$ finite, for any real numbers $(b_j, j \in J) \in \mathbb{R}^J$,*

$$\mathbf{P}\{X_j \leq b_j \text{ for all } j \in J\} = \prod_{j \in J} \mathbf{P}\{X_j \leq b_j\}.$$

Proof. By definition, $(X_i, i \in I)$ are mutually independent if and only if the σ-fields $(\sigma(X_i), i \in I)$ are independent. For $i \in I$, set $\mathcal{P}_i = \{\{X_i \leq r\}, r \in \mathbb{R}\}$. Then $\mathcal{P}_i$ is a π-system with $\sigma(\mathcal{P}_i) = \sigma(X_i)$, so by Exercise 2.13, the σ-fields in $(\sigma(X_i), i \in I)$ are independent if and only if for all $J \subset I$ finite, and any choice of events $E_j \in P_j$ for $j \in J$, it holds that $\mathbf{P}\left\{\bigcap_{j \in J} E_j\right\} = \prod_{j \in J} \mathbf{P}\{E_j\}$. This is equivalent to the condition in the proposition. □

Note that this proposition implies that the Rademacher random variables $(R_k, k \geq 1)$ defined earlier are independent, since for any $n \in \mathbb{N}$ and $b_1, \ldots, b_n \in \mathbb{R}$,

$$\mathbf{P}\{R_k \leq b_k \text{ for all } k \in [n]\} = \left(\frac{1}{2}\right)^{\#\{k \in [n] : b_k \in [0,1)\}} = \prod_{k \in [n]} \mathbf{P}\{R_k \leq b_k\}.$$

3.3. **Existence of random variables with given distributions.** You already met cumulative distribution functions of random variables in passing in Section 2.2. Given a real random variable X on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, its cumulative distribution function $F_X : \mathbb{R} \to [0,1]$ is given by $F_X(r) = \mathbf{P}\{X \leq r\}$. Its *distribution* is the measure μ_X on $\mathcal{B}(\mathbb{R})$ given by $\mu_X(B) = \mathbf{P}\{X \in B\}$ for $B \in \mathcal{B}(\mathbb{R})$. In other words, μ_X is the push-forward of the measure $\mathbf{P}$ by X.

μ_X
Note this is a different use of the term "push-forward" from earlier.

It's easy to see that F_X is a Stieltjes function, and that the Borel measure corresponding to F_X — which by Theorem 2.7 is unique — is μ_X. The next proposition says that, in turn, any cumulative distribution function (CDF) is the CDF of some random variable.

Proposition 3.4. *Let F be any CDF. Then there exists a random variable $X : [0,1] \to \mathbb{R}$ on the probability space $([0,1], \mathcal{B}([0,1]), \mathrm{Leb}_{[0,1]})$ such that $F_X = F$.*

Proof. It's both efficient and pedagogically useful to first treat a special case. Suppose F is the Uniform$[0,1]$ CDF; that is,

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x \geq 1. \end{cases}$$

We claim that $U := \sum_{k \geq 1} 2^{-k} R_k$ has $F_U = F$. First, note that $U = \sup_{\ell \geq 1} \sum_{k=1}^{\ell} 2^{-k} R_k$. Each of the terms in the supremum is a finite sum of random variables, so is a random variable; thus U is a random variable by Exercise 3.2.

To see that $F_U = F$, note that for any $n \geq 1$ and $0 \leq m < 2^n$, if we write $m/2^n$ in binary as $m/2^n = 0.b_1 b_2 \ldots b_n$ then

$$\mathbf{P}\left\{U \in \left(\frac{m}{2^n}, \frac{m+1}{2^n}\right]\right\} = \mathbf{P}\{R_1 = b_1, \ldots, R_n = b_n\} = \frac{1}{2^n}.$$

It follows that $\mathbf{P}\{U \leq m/2^n\} = m/2^n$.

Writing $D = \bigcup_{n\geq 1}\{m/2^n, 0 \leq m \leq 2^n\}$ for the dyadic fractions in $[0,1]$, for any $x \in (0,1]$ we may thus find an increasing sequence $(x_k, k \geq 1)$ of elements of D with $x_k \to x$ as $k \to \infty$. For monotone sequences of events we may interchange limit and probability, so

$$\mathbf{P}\{U < x\} = \lim_{k\to\infty} \mathbf{P}\{U \leq x_k\} = \lim_{k\to\infty} x_k = x.$$

We also have $\mathbf{P}\{U \leq x\} \leq \inf\{\mathbf{P}\{U \leq y\} : y \in D, y \geq x\} = x$, so in fact we must have $\mathbf{P}\{U \leq x\} = x$. Thus F is indeed the CDF of U.

For the general case, fix any CDF $F : \mathbb{R} \to [0,1]$, and let $G : [0,1] \to \mathbb{R}^*$ be defined by

$$G(p) := \inf\{x : F(x) \geq p\}.$$

The function G is sometimes called the "right inverse" of F. It is straightforward to check that G is Borel measurable.

Note that for $q \in [0,1]$ and $r \in \mathbb{R}$, if $F(r) \geq q$ then $\{x \in \mathbb{R} : F(x) \geq q\} \subset [r, \infty)$, so

$$G(q) = \inf\{x : F(x) \geq q\} \geq \inf[r, \infty) = r.$$

Conversely, if $F(r) < q$ then, by right-continuity of F, there exists $s > r$ such that $F(s) < q$. For such s we have $\{x \in \mathbb{R} : F(x) \geq q\} \subset (s, \infty)$, so $G(q) \geq s > r$.

The preceding paragraph establishes that $F(r) \geq q$ if and only if $r \geq G(q)$. Now let $X = G(U)$. Then X is a random variable since G is Borel, and for $r \in \mathbb{R}$,

$$\mathbf{P}\{X \leq r\} = \mathbf{P}\{G(U) \leq r\} = \mathbf{P}\{U \leq F(r)\} = F(r).$$

□

There is a simpler way to construct a Uniform$[0,1]$ random variable on the probability space $([0,1], \mathcal{B}([0,1]), \mathrm{Leb}_{[0,1]})$. Simply let $X : [0,1] \to \mathbb{R}$ be the identity function, $X(\omega) = \omega$. Then for $x \in \mathbb{R}$,

$$\mathbf{P}\{X \leq x\} = \mathbf{P}\{\omega \in [0,1] : \omega \leq x\} = \mathrm{Leb}_{[0,1]}\{\omega \in [0,1] : \omega \leq x\} = \begin{cases} 0 & x \leq 0 \\ x & x \in (0,1] \\ 1 & x > 1, \end{cases}$$

so X is Uniform$[0,1]$. Moreover, the function U defined in the course of the proof is essentially just the identity function (expand on this), which may make the proof seem unnecessarily complicated. However, by building a Uniform$[0,1]$ random variable in this way, the argument can be more easily bootstrapped to yield not just a single random variable, but sequences of independent random variables with arbitrary prescribed CDFs.

Theorem 3.5. *Fix any sequence $(F_n, n \geq 1)$ of cumulative distribution functions. Then there exists a sequence of independent random variables $(X_n, n \geq 1)$ such that X_n has CDF F_n.*

Proof. In the previous proof, we constructed a random variable with a given CDF by an appropriate transformation of a uniform random variable. We want to do the same thing but using an independent uniform for each term in the sequence. For this, we begin by splitting the sequence $(R_n, n \geq 1)$ of Rademacher random variables into infinitely many independent groups; there is no canonical way to do this so we just pick one.

List the prime numbers as $(p_i, i \geq 1)$, so $p_1 = 2, p_2 = 3$ and so forth. Then for $j, k \geq 1$ set $Q_{j,k} = R_{p_j^k}$. Then for any $i, j \geq 1$ with $i \neq j$, the sequences $(Q_{i,j}, k \geq 1)$ and $(Q_{i,j}, k \geq 1)$ contain no common terms.

Next, for $i \geq 1$ let $U_i = \sum_{k\geq 1} 2^{-k} Q_{i,k}$. The random variables $(U_i, i \geq 1)$ are each Uniform$[0,1]$ by the same reasoning as in the proof of Proposition 3.4. Moreover, they are independent since $\sigma(U_i) \subseteq \sigma(Q_{i,k}, k \geq 1)$, and the σ-fields $(\sigma(Q_{i,k}, k \geq 1), i \geq 1)$ are independent.

Now, for $n \geq 1$ let $G_n : [0,1] \to \mathbb{R}$ be defined by $G_n(p) = \inf\{x : F_n(x) \geq p\}$, and set $X_n = G_n(U_n)$. Then X_n has CDF F_n by the argument from the proof of Proposition 3.4, and $(X_n, n \geq 1) = (G_n(U_n), n \geq 1)$ are independent since $(U_n, n \geq 1)$ are independent. □

The independence of the random variables $(U_n, n \geq 1)$ constructed in the above proof is a special case of the result of the following exercise.

Exercise 3.6. *If $(Y_i, i \in I)$ are mutually independent random variables, $(I_n, n \geq 1)$ partitions I, and for each n, $g_n : \mathbb{R}^{I_n} \to \mathbb{R}$ is $(\sigma(Y_i, i \in I_n)/\mathcal{B}(\mathbb{R}))$-measurable, then with $X_n = g_n(Y_i, i \in I_n)$, the random variables $(X_n, n \geq 1)$ are independent.*

3.4. **Kolmogorov's zero-one law.** Fix a countable collection $\mathrm{X} = (X_n, n \in N)$ of random variables over a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. For $M \subset N$, write $\mathcal{T}_M = \mathcal{T}_M(\mathrm{X}) := \sigma(X_m, m \in N \setminus M)$. The *tail* σ-field is $\mathcal{T}(\mathrm{X}) := \bigcap_{M \subset N : |M| < \infty} \mathcal{T}_M$. Informally, it contains all information about the sequence $(X_n, n \geq 1)$ that can be obtained while ignoring any given finite set of the random variables. The term "tail" comes from the (standard) setting when $N = \mathbb{N}$, in which case $\mathcal{T}(\mathrm{X}) = \bigcap_{n \geq 1} \sigma(X_m, m > n)$, and from thinking of $\mathbb{N}$ as arranged on the number line.

At first blush, it might seem that if the entries of $(X_n, n \in \mathbb{N})$ are independent then $\mathcal{T}$ ought to be the trivial σ-field $\{\emptyset, \Omega\}$; after all, for any fixed $n \in \mathbb{N}$, it appears not to contain any information about X_n. However, that's not quite the case. For example, the event that $\lim_{n \to \infty} X_n$ exists is a tail event, as is any event of the form

$$\{X_n \in B_n \text{ infinitely often}\} = \bigcap_{n \geq 1} \bigcup_{m \geq n} \{X_m \in B_m\},$$

where $(B_n, n \geq 1)$ are Borel sets in $\mathbb{R}$.

Exercise 3.7. *Prove carefully that the two preceding examples are indeed examples of tail events.*

Kolmogorov's zero-one law says that $\mathcal{T}$ is at least trivial in a somewhat weaker sense.

Theorem 3.6 (Kolmogorov's 0-1 law)**.** *Let $\mathrm{X} = (X_n, n \in N)$ be a countable collection of independent random variables on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Then $\mathbf{P}\{E\} \in \{0, 1\}$ for all $E \in \mathcal{T}(\mathrm{X})$.*

Proof. Fix $E \in \mathcal{T}$. For any $n \in N$ and $F \in \sigma(X_n)$, since $\mathcal{T} \subset \sigma(X_m, m \in N \setminus \{n\})$, the events E and F are independent. Let

$$\mathcal{G} = \sigma(X_n, n \in N) = \sigma\left(\bigcup_{n \in N} \sigma(X_n)\right).$$

Note that E is independent of all events in $\bigcup_{n \in N} \sigma(X_n)$ so is independent of $\mathcal{G}$; that is, for all $F \in \mathcal{G}$, $\mathbf{P}\{E \cap F\} = \mathbf{P}\{E\}\mathbf{P}\{F\}$. However, $\mathcal{T} \subset \mathcal{G}$ so $E \in \mathcal{G}$, so

$$\mathbf{P}\{E \cap E\} = \mathbf{P}\{E\}^2;$$

this is only possible if $\mathbf{P}\{E\} \in \{0, 1\}$. □

One appealing thing about this result is that there are applications that can be described without having developed the theory of integration (expectation). We sketch two.

First, let $(X_n, n \geq 1)$ be independent random variables, and let $\mathcal{T}$ be their tail σ-field. Write $S_n = \sum_{i=1}^n X_i$ and $M^+ := \limsup_{n \to \infty} S_n/n$, $M^- := \liminf_{n \to \infty} S_n/n$.

For all $x \in \mathbb{R}$, we have $\{M^+ \geq x\} \in \mathcal{T}$, so $\mathbf{P}\{S^+ \geq x\} \in \{0, 1\}$. Letting $x^+ = \sup\{x : \mathbf{P}\{M^+ \geq x\} = 1\}$, then for $y > x$ we have $\mathbf{P}\{M^+ \geq y\} < 1$ so $\mathbf{P}\{M^+ \geq 0\} = 0$. Thus $\mathbf{P}\{M^+ = x^+\} = 1$, and likewise $\mathbf{P}\{M^- = x^-\} = 1$. Moreover, $\mathbf{P}\{\lim_{n \to \infty} S_n/n \text{ exists}\} \in \{0, 1\}$. The strong law of large numbers gives a necessary and sufficient condition for the last probability to equal 1, provided the entries of $(X_n, n \geq 1)$ are identically distributed.

The second example is that of *percolation*, one of the most active areas of modern probability theory. Let $\mathbb{Z}^d$ be the d-dimensional integer lattice; in this context the elements of $\mathbb{Z}^d$ are called *sites*. Fix $p \in [0, 1]$ and let $\mathrm{B} = (B_v, v \in \mathbb{Z}^d)$ be independent Bernoulli(p) random variables on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$. (By Bernoulli($p$) we mean that $\mathbf{P}\{B_v = 1\} = p = 1 - \mathbf{P}\{B_v = 0\}$ for all $v \in \mathbb{Z}^d$.) Let $\mathcal{T}$ be the tail σ-field of B.

We use B to define *site percolation clusters* as follows. Write $\mathbb{Z}^d(\mathrm{B}) = \{v \in \mathbb{Z}^d : B_v = 1\}$. For $x, y \in \mathbb{Z}^d$ say that x *is connected to* y *in* $\mathbb{Z}^d(\mathrm{B})$, and write $x \xrightarrow{\mathrm{B}} y$, if there is a nearest-neighbour path from x to y containing only elements of $\mathbb{Z}^d(B)$. Then for $x \in \mathbb{Z}^d$ define

$$\mathcal{C}(x) := \{y \in \mathbb{Z}^d : x \xrightarrow{\mathrm{B}} y\}.$$

Note that if $y \in \mathcal{C}(x)$ then $\mathcal{C}(x) = \mathcal{C}(y)$.

Now let

$$E = \{\exists x \in \mathbb{Z}^d; |\mathcal{C}(x)| = \infty\} = \{\mathbb{Z}^d(\mathrm{B}) \text{ contains an infinite connected component}\}.$$

An infinite connected component can not be created or destroyed by adding or removing finitely many sites, so E is a tail event; therefore $x(p,d) := \mathbf{P}\{E\} \in \{0,1\}$ by Kolmogorov's 0-1 law. Which of these values is correct depends on the parameter p of the Bernoulli random variables and on the dimension d.

The *critical probability* for site percolation on $\mathbb{Z}^d$ is

$$p_c(\mathbb{Z}^d) := \sup\{p : x(p,d) = 0\}.$$

We necessarily have $x(p,d) = 1$ for all $p > p_c$, but unlike for $\limsup S_n/n$ the first example, this doesn't imply that $x(p_c, d) = 1$. In fact, it is conjectured that $x(p_c, d) = 0$, or in words that there is "no percolation at criticality", in any dimension. This is probably the most famous open question in probability.

The next exercise should take care of any measurability concerns in the definition of percolation. Recall that $2^{\mathbb{Z}^d}$ is the set of all subsets of $\mathbb{Z}^d$. So a set $S \subset 2^{\mathbb{Z}^d}$ is a set of subsets of $\mathbb{Z}^d$; we say such S is a *cylinder set* if

$$S = \{V \subset \mathbb{Z}^d : A \subset V, B \subset V^c\},$$

I haven't checked these exercises carefully, proceed at your own risk

for some finite sets A, B.

Exercise 3.8. *Let* $\mathcal{G} = \sigma(B_v, v \in \mathbb{Z}^d)$ *and let* $\mathcal{G}^* = \mathrm{B}^*(\mathcal{G})$ *be the push-forward of* $\mathcal{G}$ *under the map*

$$\omega \overset{\mathrm{B}}{\mapsto} \{B_v(\omega), v \in \mathbb{Z}^d\} \in \{0,1\}^{\mathbb{Z}^d}.$$

Show that $\mathcal{G}^* = \sigma\left(\text{Cylinder sets in } 2^{\mathbb{Z}^d}\right)$.

Exercise 3.9. *Show carefully that the event* $\{\exists x \in \mathbb{Z}^d; |\mathcal{C}(x)| = \infty\}$ *is in* $\mathcal{T} \subset \mathcal{G}$.

3.5. **Almost sure convergence, convergence in probability and convergence in distribution.** Let $(X_n, 1 \le n \le \infty)$ be a sequence of random variables defined on a common space $(\Omega, \mathcal{F}, \mathbf{P})$. We say X_n *converges almost surely* to X_∞, and write $X_n \overset{\text{a.s.}}{\to} X_\infty$, if

Almost sure convergence

$$\mathbf{P}\left\{\lim_{n\to\infty} X_n = X_\infty\right\} = 1.$$

We say X_n *converges in probability* to X_∞ if for all $\epsilon > 0$,

Convergence in probability

$$\lim_{n\to\infty} \mathbf{P}\{|X_n - X_\infty| > \epsilon\} = 0.$$

Next, given random variables $(X_n, 1 \le n \le \infty)$, with $X_n : \Omega_n \to \mathbb{R}$ for some probability space $(\Omega_n, \mathcal{F}_n, \mathbf{P}_n)$, we say X_n *converges in distribution* to X_∞, and write $X_n \overset{\text{d}}{\to} X_\infty$, if

Convergence in distribution

$$\lim_{n\to\infty} \mathbf{P}_n\{X_n \le x\} = \mathbf{P}_\infty\{X_\infty \le x\}$$

for all x with $\mathbf{P}_\infty\{X_\infty = x\} = 0$. This may seem complicated compared with the previous definitions; the reason for this is that convergence in distribution is really a property of the *distributions* of the random variables (or, equivalently, of their CDFs), and is insensitive to the specific spaces on which they are defined.

Exercise 3.10. *(a) Check that* $X_n \overset{\text{d}}{\to} X_\infty$ *iff* $F_{X_n}(x) \to F_{X_\infty}(x)$ *for all continuity points* x *of* F_{X_∞}.

(b) Show that if $X_n \overset{\text{d}}{\to} X$ *as* $X \to \infty$ *and* $X_n \overset{\text{d}}{\to} Y$ *as* $n \to \infty$ *then* $F_X = F_Y$ *and so* $\mu_X = \mu_Y$.

One warning, which partially explains the restriction to $x \in \mathbb{R}$ with $\mathbf{P}\{X_\infty = x\} = 0$ above, is in order. Write $U_n = \sum_{k=1}^n 2^{-k} R_k$, where $(R_n, n \geq 1)$ are the Rademacher random variables defined earlier, and let $U_\infty = \sum_{k \geq 1} 2^{-k} U_k|$. Then $U_n \to U_\infty$ almost surely, since $|U_n - U_\infty| \leq \sum_{k>n} 2^{-k} = 2^{-n}$. However, $\mathbf{P}\{U_n \in \mathbb{Q}\} = 1$ and $\mathbf{P}\{U_\infty \in Q\} = 0$. This shows that $X_n \overset{\text{a.s.}}{\to} X_\infty$ does *not* in general imply that

$$\mathbf{P}\{X_n \in A\} \to \mathbf{P}\{X_\infty \in A\}$$

for all $A \in \mathcal{B}(\mathbb{R})$; more care is needed.

An easy example also shows that convergence in probability does not imply almost sure convergence. Let $(B_n, n \geq 1)$ be independent with B_n a Bernoulli$(1/n)$ random variable, which is to say $\mathbf{P}\{B_n = 1\} = 1/n = 1 - \mathbf{P}\{B_n = 0\}$. Then for all $\epsilon \in (0,1)$,

$$\mathbf{P}\{|B_n - 0| > \epsilon\} = \mathbf{P}\{B_n = 1\} = \frac{1}{n} \to 0$$

as $n \to \infty$, so $B_n \overset{\text{p}}{\to} 0$. However, $\sum_{n \geq 1} \mathbf{P}\{B_n = 1\} = \sum_{n \geq 1} 1/n = \infty$, so by the second Borel-Cantelli lemma, $\mathbf{P}\{B_n = 1 \text{ i.o.}\} = 1$. It follows that

$$\mathbf{P}\left\{\lim_{n\to\infty} B_n = 0\right\} = \mathbf{P}\{\{B_n = 1 \text{ i.o.}\}^c\} = 1 - \mathbf{P}\{B_n = 1 \text{ i.o.}\} = 0.$$

Thus B_n does *not* converge to 0 almost surely.

We now turn from warning examples to positive results.

Proposition 3.7. *Let $(X_n, n \geq 1)$ be a sequence of random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$. If $X_n \overset{\text{a.s.}}{\to} X_\infty$ then $X_n \overset{\text{p}}{\to} X_\infty$.*

Proof. Fix $\epsilon > 0$. Then we have

$$\mathbf{P}\left\{\lim_{n\to\infty} X_n = X_\infty\right\} \leq \mathbf{P}\left\{\limsup_{n\to\infty} |X_n - X_\infty| \leq \epsilon\right\}. = \mathbf{P}\left\{\exists n \in \mathbb{N} : \sup_{m \geq n} |X_n - X_\infty| \leq \epsilon\right\}.$$

The sequence of events $\{\sup_{m \geq n} |X_m - X_\infty| \leq \epsilon\}$ is increasing in n, and its limit is the event

$$\{\limsup_{n\to\infty} |X_n - X_\infty| \leq \epsilon\},$$

so

$$\mathbf{P}\left\{\limsup_{n\to\infty} |X_n - X_\infty| \leq \epsilon\right\} = \lim_{n\to\infty} \mathbf{P}\left\{\sup_{m \geq n} |X_m - X_\infty| \leq \epsilon\right\} \leq \lim_{n\to\infty} \mathbf{P}\{|X_n - X_\infty| \leq \epsilon\}.$$

It follows that if $\mathbf{P}\{\lim_{n\to\infty} X_n = X_\infty\} = 1$ then $\lim_{n\to\infty} \mathbf{P}\{|X_n - X_\infty| \leq \epsilon\} = 1$. □

Proposition 3.8. *Let $(X_n, n \geq 1)$ be a sequence of random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$. If $X_n \overset{\text{p}}{\to} X_\infty$ then there exists a subsequence $(n_k, k \geq 1)$ such that $X_{n_k} \overset{\text{a.s.}}{\to} X_\infty$ as $k \to \infty$.*

Proof. Suppose that $X_n \overset{\text{p}}{\to} X_\infty$. Then for each $k \in \mathbb{N}$, we may choose $n_k \in \mathbb{N}$ large enough that $\mathbf{P}\{|X_m - X_\infty| > 1/k\} < 1/2^k$ for all $m \geq n_k$. The n_k can clearly be chosen to be increasing, so that $(n_k, k \geq 1)$ is indeed a subsequence of $\mathbb{N}$. Then

$$\sum_{k \geq 1} \mathbf{P}\{|X_{n_k} - X_\infty| > 1/m\} \leq m + \sum_{k \geq m} \frac{1}{2^k} < \infty,$$

so by the first Borel-Cantelli lemma, $\mathbf{P}\{|X_{n_k} - X_\infty| > 1/m \text{ i.o.}\} = 0$. Thus

$$\begin{aligned}
\mathbf{P}\left\{\lim_{k\to\infty} X_{n_k} \neq X_\infty\right\} &= \mathbf{P}\left\{\exists m \in \mathbb{N} : \limsup_{k\to\infty} |X_{n_k} - X_\infty| > 1/m\right\} \\
&\leq \sum_{m \in \mathbb{N}} \mathbf{P}\left\{\limsup_{k\to\infty} |X_{n_k} - X_\infty| > 1/m \text{ i.o.}\right\} \\
&= 0.
\end{aligned}$$

□

Proposition 3.9. *Let $(X_n, n \geq 1)$ be a sequence of random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$. If $X_n \xrightarrow{\text{p}} X_\infty$ then $X_n \xrightarrow{\text{d}} X_\infty$.*

Proof. First, note that for any random variable X, for each $x \in \mathbb{R}$ with $\mathbf{P}\{X = x\} > 0$ the interval

$$(\mathbf{P}\{X < x\}, \mathbf{P}\{X \leq x\})$$

is non-empty, and these intervals are pairwise disjoint for different points $x, y \in \mathbb{R}$. Thus, if for each $x \in \mathbb{R}$ with $\mathbf{P}\{X = x\} > 0$ we choose a point $q(x) \in (\mathbf{P}\{X < x\}, \mathbf{P}\{X = x\}) \cap \mathbb{Q}$, then the values $q(x)$ are distinct rational numbers. We have thus defined an injective map from $\{x \in \mathbb{R} : \mathbf{P}\{X = x\} > 0\}$ to $\mathbb{Q}$, so $\{x \in \mathbb{R} : \mathbf{P}\{X = x\} > 0\}$ is at countable.

Now fix $x \in \mathbb{R}$ with $\mathbf{P}\{X_\infty = x\} = 0$. Then since $\{X_\infty < x\}$ is the increasing limit of the events $\{X_\infty \leq x - \delta\}$ as $\delta \downarrow 0$, we have

$$\mathbf{P}\{X_\infty \leq x\} = \mathbf{P}\{X_\infty < x\} = \lim_{\delta \downarrow 0} \mathbf{P}\{X_\infty \leq x - \delta\}.$$

Also, by continuity from above, $\mathbf{P}\{X_\infty \leq x\} = \lim_{\delta \downarrow 0} \mathbf{P}\{X_\infty \leq x + \delta\}$. Thus, for all $\epsilon > 0$ there is $\delta > 0$ such that

$$\mathbf{P}\{X_\infty \leq x\} - \epsilon < \mathbf{P}\{X_\infty \leq x - \delta\} \leq \mathbf{P}\{X_\infty \leq x + \delta\} < \mathbf{P}\{X_\infty \leq x\} + \epsilon.$$

Now, if $X_n \leq x$ then either $X_\infty \leq x + \delta$ or $|X_n - X_\infty| > \delta$, so

$$\begin{aligned}\limsup_{n\to\infty} \mathbf{P}\{X_n \leq x\} &\leq \limsup_{n\to\infty} (\mathbf{P}\{X_\infty \leq x + \delta\} + \mathbf{P}\{|X_n - X_\infty| > \delta\}) \\ &= \mathbf{P}\{X_\infty \leq x - \delta\} \leq \mathbf{P}\{X_\infty \leq x\} + \epsilon.\end{aligned}$$

Likewise, if $X_\infty \leq x - \delta$ then either $X_n \leq x$ or $|X_n - X_\infty| > \delta$, so

$$\begin{aligned}\liminf_{n\to\infty} \mathbf{P}\{X_n \leq x\} &\geq \liminf_{n\to\infty} (\mathbf{P}\{X_\infty \leq x - \delta\} - \mathbf{P}\{|X_n - X_\infty| > \delta\}) \\ &= \mathbf{P}\{X_\infty \leq x - \delta\} \geq \mathbf{P}\{X_\infty \leq x\} - \epsilon.\end{aligned}$$

Since $\epsilon > 0$ was arbitrary, this completes the proof. □

coupling

For the last, and perhaps most interesting, implication between different modes of convergence, we require an additional definition. Fix a collection of measures $(\mu_i, i \in I)$. A *coupling* of $(\mu_i, i \in I)$ is a collection $(Y_i, i \in I)$ of random variables defined on a *common* probability space $(\Omega, \mathcal{F}, \mathbf{P})$ such that $\mu_{Y_i} = \mu_i$ for all $i \in I$. If $(X_i, i \in I)$ is a collection of random variables, possibly defined on different probability spaces, with $\mu_{X_i} = \mu_i$, we might also refer to $(Y_i, i \in I)$ as a coupling of $(X_i, i \in I)$.

For example, suppose that μ_1 and μ_2 are both the uniform measure on the set $[6] = \{1, 2, 3, 4, 5, 6\}$. Then with $\Omega = [6]$, $\mathcal{F} = 2^{[6]}$ and $\mathbf{P}$ the uniform measure on Ω, setting $Y_1(\omega) = \omega$ and $Y_2(\omega) = 7 - \omega$ gives a coupling of μ_1 and μ_2.[7] Alternately, with $\Omega = [6]^2 = \{(i, j), 1 \leq i, j \leq 6\}$, $\mathcal{F} = 2^\Omega$, and $\mathbf{P}$ the uniform measure on Ω, setting $Y_1(i, j) = i$ and $Y_2(i, j) = j$ gives another coupling of μ_1 and μ_2; this is an "independent coupling" since Y_1 and Y_2 are independent. By Theorem 3.5, if $(\mu_i, i \in I)$ is a countable collection of probability measures then a coupling of $(\mu_i, i \in I)$ always exists.

Theorem 3.10 (Skorohod representation theorem)**.** *Fix random variables $(X_n, 1 \leq n \leq \infty)$, with $X_n : \Omega_n \to \mathbb{R}$ for some probability space $(\Omega_n, \mathcal{F}_n, \mathbf{P}_n)$. If $X_n \xrightarrow{\text{d}} X_\infty$ then there exists a coupling $(Y_n, 1 \leq n \leq \infty)$ of $(X_n, 1 \leq n \leq \infty)$ such that $Y_n \xrightarrow{\text{a.s.}} Y_\infty$.*

Proof. We write $F_n = F_{X_n}$. Our coupling lives on the probability space

$$(\Omega, \mathcal{F}, \mathbf{P}) = ([0, 1], \mathcal{B}([0, 1]), \text{Leb}_{[0,1]}).$$

[7] This is the "glass table" coupling of a die roll: the value that comes up and the value seen by someone lying under the table.

For $1 \leq n \leq \infty$, let $Y_n : \Omega \to \mathbb{R}$ be defined by

$$Y_n(p) = \inf\{x : F_n(x) \geq p\}.$$

Then by the same argument as in the proof of Proposition 3.4, we have $F_{Y_n} = F_n$ for all n, so $(Y_n, 1 \leq n \leq \infty)$ is indeed a coupling of $(X_n, 1 \leq n \leq \infty)$. The bulk of the proof consists in showing that $Y_n \overset{\text{a.s.}}{\to} Y_\infty$.

Note that for all $1 \leq n \leq \infty$, $Y_n(p)$ is increasing in p, so has at most countably many points of discontinuity (reprising the argument from the start of Proposition 3.9 gives an injective map from the discontinuity points into $\mathbb{Q}$). Thus to prove that $Y_n \overset{\text{a.s.}}{\to} Y_\infty$ it is sufficient to prove that $Y_n(p) \to Y_\infty(p)$ whenever Y_∞ is continuous at p.

So fix $p \in [0, 1]$ a continuity point of Y_∞, and write $y = Y_\infty(p) = \inf\{x \in \mathbb{R} : F_\infty(x) \geq p\}$. Then $F_\infty(x) < p$ for $x < y$. Writing $p' = F_\infty(y)$, by right-continuity of F_∞ we must have $p' \geq p$. Moreover, since p is a continuity point of F_∞ we must have $F_\infty(z) > p$ for all $z > y$. (If $F_\infty(z) = p$ for some $z > y$ then for all $q > p$ we have $Y_\infty(q) \geq z$, contradicting that p is a continuity point.)

Now fix $\epsilon > 0$, and choose $x < y < z$ with x, z continuity points of F_∞ and such that

$$y - \epsilon < x < y < z < y + \epsilon.$$

Then $F_\infty(x) < p$ and $F_\infty(z) > p$. Since z is a continuity point of F_∞ and $X_n \overset{\text{d}}{\to} X_\infty$, it follows that

$$F_n(z) \to F_\infty(z) > p$$

so $F_n(z) > p$ for all n sufficiently large. Thus $Y_n(p) \leq z < y + \epsilon$ for n large. Likewise, $F_n(x) \to F_\infty(x) < p$, so $F_n(x) < p$ for n large. Thus for $Y_n(p) \geq x > y - \epsilon$ for n large. Since $\epsilon > 0$ was arbitrary, it follows that $Y_n(p) \to Y_\infty(p)$, as required. □

4. **Integration and expectation**

Let $(\Omega, \mathcal{F}, \mu)$ be a σ-finite measure space. In this section, unless otherwise specified, when we refer to a *measurable function* f, we mean a $\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable function $f : \Omega \to \mathbb{R}$.[8] We say that an event $E \in \mathcal{F}$ occurs *μ-almost everywhere*, or μ-a.e., if $\mu(E^c) = 0$.

Our aim is to define the (definite) integral

$$\int f \mathrm{d}\mu \equiv \int_\Omega f \mathrm{d}\mu \equiv \int_\Omega f(x)\mu(\mathrm{d}x) \equiv \mu(f)$$

for as rich a class of measurable functions as possible. The preceding display lists four different bits of notation for this integral; these notes use at least the first three.

The way the integral is defined is by starting from functions taking only finitely many values, where the correct definition of the integral is obvious, then taking limits. We say a measurable function f is *simple* if it takes only finitely many values. Thus, f is simple if for some $n \in \mathbb{N}$ there are sets $E_1, \ldots, E_n \in \mathcal{F}$ and constants $c_1, \ldots, c_n \in \mathbb{R}$ such that $f = \sum_{i=1}^n c_i \mathbf{1}_{[E_i]}$. Simple function.

Exercise 4.1. *For any simple function $f : \Omega \to \mathbb{R}$, there is a unique choice of pairwise disjoint sets $D_1, \ldots, D_\ell \in \mathcal{F}$ and of distinct constants $b_1, \ldots, b_\ell \in \mathbb{R}$ such that $f = \sum_{i=1}^\ell b_i \mathbf{1}_{[D_i]}$.*

Let $f = \sum_{i=1}^\ell b_i \mathbf{1}_{[D_i]}$ be a simple function from Ω to $\mathbb{R}$, with $(D_1, \ldots, D_\ell)$ pairwise disjoint and $(b_1, \ldots, b_\ell)$ distinct. We say f is *integrable* if $\mu(D_i) < \infty$ for all $1 \leq i \leq \ell$; if this holds then we define

$$\int_\Omega f \mathrm{d}\mu = \sum_{i=1}^n c_i \mu(E_i)\,.$$

If $\mu(\Omega) < \infty$ then every simple function is integrable. The next exercise says that for a simple integrable function, the definition of the integral doesn't depend on the representation of f as a sum of indicators of sets of bounded measure.

[8] Most of what follows also works if $f : \Omega \to \mathbb{R}^*$ is $\mathcal{F}/\mathcal{B}(\mathbb{R}^*)$-measurable, provided one takes appropriate care around situations where $\infty - \infty$ might show up.

Exercise 4.2. *Suppose that* $\sum_{i=1}^n c_i \mathbf{1}_{[E_i]} = \sum_{i=1}^m d_i \mathbf{1}_{[F_i]}$ *define the same function (where* $E_1, \ldots, E_n \in \mathcal{F}$ *and* $F_1, \ldots, F_m \in \mathcal{F}$ *all have finite measure, and* $c_1, \ldots, c_n, d_1, \ldots, d_m \in \mathbb{R}$*. Then* $\sum_{i=1}^n c_i \mu(E_i) = \sum_{i=1}^m d_i \mu(F_i)$*.*

The next proposition states some basic properties of the integral for simple integrable functions.

Proposition 4.1. *Let* $(\Omega, \mathcal{F}, \mu)$ *be a* σ*-finite measure space and let* $f, g : \Omega \to \mathbb{R}$ *be simple integrable functions.*

- *If* $f \geq 0$ μ*-a.e. then* $\int f \mathrm{d}\mu \geq 0$*.*
- *If* $a \in \mathbb{R}$ *then* $\int af + g \mathrm{d}\mu = a \int f \mathrm{d}\mu + \int g \mathrm{d}\mu$*.*
- *If* $f \leq g$ μ*-a.e. then* $\int f \mathrm{d}\mu \leq \int g \mathrm{d}\mu$*.*

Proof. Write $f = \sum_{i=1}^n c_i \mathbf{1}_{[E_i]}$ with $E_1, \ldots, E_n \in \mathcal{F}$ disjoint. If some $c_i < 0$ then since $f \geq 0$ μ-almost everywhere we must have $\mu(E_i) = 0$. Thus

$$\int f \mathrm{d}\mu = \sum_{i : c_i > 0} c_i \mu(E_i) \geq 0 \,,$$

proving (a). Next, write $g = \sum_{j=1}^m d_j \mathbf{1}_{[F_j]}$. Then $af + g = a \sum_{i=1}^n c_i \mathbf{1}_{[E_i]} + \sum_{j=1}^m d_j \mathbf{1}_{[F_j]}$ is simple so by definition

$$\int af + g \mathrm{d}\mu = a \sum_{i=1}^n c_i \mu(E_i) + \sum_{j=1}^m d_j \mu(F_j) = a \int f \mathrm{d}\mu + \int g \mathrm{d}\mu \,,$$

proving (b). Finally, if $f \leq g$ μ-a.e. then $g - f \geq 0$ μ-a.e. so by (a) and (b),

$$0 \leq \int g - f \mathrm{d}\mu = \int g \mathrm{d}\mu - \int f \mathrm{d}\mu \,,$$

proving (c). □

f s.i.: simple integrable

In what follows we'll sometimes write "f s.i." to mean that f is simple and integrable. We extend the definition from simple functions first to non-negative functions, then to general functions. For f a non-negative measurable function, define

$$\int f \mathrm{d}\mu = \sup_{\substack{g \leq f \\ g \text{ s.i.}}} \int g \mathrm{d}\mu \,.$$

Note that if f is itself simple then for $g \leq f$ simple we have $\int g \mathrm{d}\mu \leq \int f \mathrm{d}\mu$ by the previous proposition; it follows that this new definition agrees with the previous definition when f is simple.

One may think of this definition as a "horizontal" definition via lower approximations, whereas the Riemann integral uses a "vertical" approximation. Alternatively, one may say that the Riemann approximation to the integral decomposes the domain, whereas the above definition (one might call it a "Lebesgue approximation") decomposes the range.

Definition of f^+ and f^- for a function f; note that we use this notation differently earlier in the notes.

Finally, for a general measurable function f, write $f^+ = \max(f, 0)$ and $f^- = -\min(f, 0)$. If either $\int f^+ \mathrm{d}\mu < \infty$ or $\int f^- \mathrm{d}\mu < \infty$ then we set

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

and say the integral of f is defined. Note that if $f \geq 0$ then $f = f^+$ and $f^- = 0$, so this definition agrees with the definition for non-negative functions.

Having extended the definition of the integral from simple functions to this more general class, we now need to check again that the basic properties of the integral all hold.

Proposition 4.2. *Let* $(\Omega, \mathcal{F}, \mu)$ *be a* σ*-finite measure space and let* f, g *and* $(f_n, n \geq 1)$ *be measurable functions.*

- **Weak monotonicity.** *If* $f \leq g$ *and* $\int f \mathrm{d}\mu$ *and* $\int g \mathrm{d}\mu$ *are defined then* $\int f \mathrm{d}\mu \leq \int g \mathrm{d}\mu$*.*
- **Weak monotone convergence theorem.** *If* $f_n \geq 0$ *and* $f_n \uparrow f$*, then* $\int f_n \mathrm{d}\mu \uparrow \int f \mathrm{d}\mu$*.*

- **Linearity of expectation.** *If $f, g \geq 0$ and $a \geq 0$ then $\int af + g\mathrm{d}\mu = a\int f\mathrm{d}\mu + \int g\mathrm{d}\mu$.*

To prove linearity of expectation, we need the following lemma.

Lemma 4.3. *Let $f \geq 0$ be measurable. Then there exist non-negative simple functions $(f_n, n \geq 1)$ such that $f_n \uparrow f$ as $n \to \infty$.*

s

Proof. For $0 \leq k < n \cdot 2^n$ let $B_{n,k} = \{k/2^n \leq f < (k+1)/2^n\}$. Then set

$$f_n = \sum_{k=0}^{n2^n-1} \frac{k}{2^n}\mathbf{1}_{[B_{n,k}]}.$$

Then $0 \leq f_n \leq f$, and $f\mathbf{1}_{[f\leq n]} \leq f_n + 1/2^n$ so $\liminf_{n\to\infty} f_n \geq \liminf_{n\to\infty}(f\mathbf{1}_{[f\leq n]} - 2^{-n}) = f$. □

For later use, we remark that the functions f_n constructed in the course of proving the above theorem are all $(\sigma(f)/\mathcal{B}(\mathbb{R}))$-measurable. This means that if $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space and $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are non-negative independent random variables, then there exist $(\sigma(X)/\mathcal{B}(\mathbb{R}))$-measurable random variables $(X_n, n \geq 1)$ and $(\sigma(Y)/\mathcal{B}(\mathbb{R}))$-measurable random variables $(Y_n, n \geq 1)$ such that $X_n \uparrow X$ and $Y_n \uparrow Y$ as $n \to \infty$. The collections $(X_n, n \geq 1)$ and $(Y_n, n \geq 1)$ of random variables then independent due to the independence of X and Y. This extends to more than two variables in an obvious way.

Proof of Proposition 4.2. If $f \geq 0$ then this is obvious because the supremum in the definition of $\int g\mathrm{d}\mu$ is over a larger set than in the definition of $\int f\mathrm{d}\mu$. For general f, since $f \leq g$ we have $f^+ \leq g^+$ and $f^- \geq g^-$ so

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

This proves the first assertion.

Next, suppose $0 \leq f_n \uparrow f$. Then for each n by monotonicity we have $f_n \leq f$ so $\int f_n\mathrm{d}\mu \leq \int f\mathrm{d}\mu$, so

$$\lim_{n\to\infty}\int f_n\mathrm{d}\mu = \sup_{n\in\mathbb{N}}\int f_n\mathrm{d}\mu \leq \int f\mathrm{d}\mu.$$

To prove the reverse inequality, fix any simple function $g = \sum_{i=1}^m c_i\mathbf{1}_{[E_i]}$ with $0 \leq g \leq f$. We may assume that $(E_1, \ldots, E_m)$ are disjoint and that $c_i > 0$ for all $1 \leq i \leq m$.

First suppose $\int f\mathrm{d}\mu = \infty$. For $n \geq 1$ let

$$E_{i,n} = E_i \cap \{f_n > c_i/2\} = \{\omega \in E_i : f_n(\omega) \geq c_i/2\}.$$

Then $E_{i,n} \uparrow E_i$ as $n \to \infty$, so $\mu(E_{i,n}) \uparrow \mu(E_i)$. Since also $f_n \geq \sum_{i=1}^m (c_i/2)\mathbf{1}_{[E_{i,n}]}$, it follows that

$$\begin{aligned}
\liminf_{n\to\infty}\int f_n\mathrm{d}\mu &\geq \liminf_{n\to\infty}\int \sum_{i=1}^m (c_i/2)\mathbf{1}_{[E_{i,n}]}\mathrm{d}\mu \\
&= \liminf_{n\to\infty}\frac{1}{2}\sum_{i=1}^m c_i\mu(E_{i,n}) \\
&= \liminf_{n\to\infty}\frac{1}{2}\sum_{i=1}^m c_i\mu(E_i) \\
&= \frac{1}{2}\int g\mathrm{d}\mu.
\end{aligned}$$

Thus

$$\liminf_{n\to\infty} \int f_n \mathrm{d}\mu \geq \frac{1}{2} \sup_{\substack{g\leq f \\ g \text{ simple}}} \int g \mathrm{d}\mu = \infty.$$

Next suppose $\int f \mathrm{d}\mu < \infty$. Fix $\epsilon > 0$, let $\delta = \epsilon / \int g \mathrm{d}\mu$, and for $n \geq 1$ let

$$E_{i,n} = E_i \cap \{f_n > c_i - \epsilon\}.$$

Then again $E_{i,n} \uparrow E_i$ as $n \to \infty$, so $\mu(E_{i,n}) \uparrow \mu(E_i)$. Moreover,

$$f_n \geq \sum_{i=1}^m (c_i - \delta) \mathbf{1}_{[E_{i,n}]} \geq \sum_{i=1}^m c_i \mathbf{1}_{[E_{i,n}]} - \delta \sum_{i=1}^m \mathbf{1}_{[E_i]},$$

so

$$\begin{aligned}
\liminf_{n\to\infty} \int f_n \mathrm{d}\mu &\geq \liminf_{n\to\infty} \left(\int \sum_{i=1}^m c_i \mathbf{1}_{[E_{i,n}]} \mathrm{d}\mu - \int \delta \sum_{i=1}^m \mathbf{1}_{[E_i]} \mathrm{d}\mu \right) \\
&= \liminf_{n\to\infty} \left(\int \sum_{i=1}^m c_i \mathbf{1}_{[E_{i,n}]} \mathrm{d}\mu \right) - \delta \sum_{i=1}^m c_i \mu(E_i) \\
&= \liminf_{n\to\infty} \sum_{i=1}^m c_i \mu(E_{i,n}) - \epsilon \\
&= \sum_{i=1}^m c_i \mu(E_i) - \epsilon.
\end{aligned}$$

Since $\epsilon > 0$ and $g \leq f$ were arbitrary, it follows that

$$\liminf_{n\to\infty} \int f_n \mathrm{d}\mu \geq \int f \mathrm{d}\mu$$

as before. This proves the second assertion.

Finally, fix non-negative measurable functions f, g and constant $a \geq 0$. Then let $(f_n, n \geq 1)$ and $(g_n, n \geq 1)$ be simple functions with $0 \leq f_n \uparrow f$ and $0 \leq g_n \uparrow g$. Then $af_n + g_n \uparrow af + g$, so

$$\int cf + g \mathrm{d}\mu = \lim_{n\to\infty} \int cf_n + g_n \mathrm{d}\mu = \lim_{n\to\infty} c \int f_n \mathrm{d}\mu + \int g_n \mathrm{d}\mu = c \int f \mathrm{d}\mu + \int g \mathrm{d}\mu,$$

where we have used monotone convergence, plus linearity of integration for simple functions, in the above string of identities. This completes the proof. □

$L_1(\mu)$, μ-integrable

Notice that linearity of integration for non-negative functions implies that $\int |f| \mathrm{d}\mu = \int f^+ \mathrm{d}\mu + \int f^- \mathrm{d}\mu$, since $|f| = f^+ + f^-$. If $\int |f| \mathrm{d}\mu < \infty$ we say that f is *μ-integrable* and write $f \in L_1(\mu)$.

Exercise 4.3. *(a) Show that if f, g are μ-integrable and $a \in \mathbb{R}$ then $\int af + g \mathrm{d}\mu = a \int f \mathrm{d}\mu + \int g \mathrm{d}\mu$.*
(b) Let $f \geq 0$ be measurable. Show that $\int f \mathrm{d}\mu = 0$ if and only if $f = 0$ μ-almost everywhere.

Proposition 4.4 (Monotonicity of integrals)**.** *If $f \leq g$ μ-almost everywhere and both integrals are defined, then $\int f \mathrm{d}\mu \leq \int g \mathrm{d}\mu$.*

Proof. Write $\hat{g} = g + (f - g)\mathbf{1}_{[f>g]}$. Then $\hat{g} \geq f$ so

$$\int \hat{g} \mathrm{d}\mu \geq \int f \mathrm{d}\mu.$$

But $\hat{g} - g = (f - g)\mathbf{1}_{[f>g]}$ is non-negative and μ-a.e. equals zero, so

$$\int g \mathrm{d}\mu = \int \hat{g} \mathrm{d}\mu - \int (\hat{g} - g) \mathrm{d}\mu = \int \hat{g} \mathrm{d}\mu. \qquad \square$$

Note that Proposition 4.4 implies that if $f \overset{\mu-\text{a.e.}}{=} g$ and $\int f \mathrm{d}\mu$ is defined, then $\int g \mathrm{d}\mu$ is defined and $\int f \mathrm{d}\mu = \int g \mathrm{d}\mu$.

We now state and prove the fundamental convergence theorems for sequences of functions. In all of them, $(f_n, n \geq 1)$, f, and g are measurable functions defined on a σ-finite measure space $(\Omega, \mathcal{F}, \mu)$. The first result, the (strong) monotone convergence theorem, is really a corollary of the weak monotone convergence theorem combined with the previous proposition.

Theorem 4.5 (Monotone convergence theorem)**.** *If $(f_n, n \geq 1)$ and f are measurable functions and $0 \leq f_n \uparrow f$ holds μ-almost everywhere then*

$$\int f_n \mathrm{d}\mu \to \int f \mathrm{d}\mu \,,$$

as $n \to \infty$.

Proof. Let

$$E = \{\omega \in \Omega : f_n(\omega) \uparrow f(\omega) \text{ as } n \to \infty\}.$$

Then $\mu(E^c) = 0$ by assumption. Writing $f_n' = f_n \mathbf{1}_{[E]}$ and $f' = f\mathbf{1}_{[E]}$, then $0 \leq f_n' \uparrow f'$ so by the weak monotone convergence theorem $\int f_n' \mathrm{d}\mu \to \int f' \mathrm{d}\mu$. But $f_n' \overset{\mu-\text{a.e.}}{=} f_n$ and $f' \overset{\mu-\text{a.e.}}{=} f$, so Proposition 4.4 we have

$$\int f_n \mathrm{d}\mu = \int f_n' \mathrm{d}\mu \quad \text{and} \quad \int f \mathrm{d}\mu = \int f' \mathrm{d}\mu$$

and the result follows. □

Theorem 4.6 (Fatou's lemma)**.** *If $f_n \geq 0$ for all n then*

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

Proof. Note that $\inf_{k\geq n} f_k$ is increasing in n, and its limit is $\liminf_{n\to\infty} f_n$, so by the monotone convergence theorem

$$\int \liminf_{n\to\infty} f_n \mathrm{d}\mu = \lim_{n\to\infty} \int \inf_{k\geq n} f_k \mathrm{d}\mu \,.$$

But for each $k \geq n$, $\int \inf_{k\geq n} f_k \mathrm{d}\mu \leq \int f_k \mathrm{d}\mu$, so

$$\lim_{n\to\infty} \int \inf_{k\geq n} f_k \mathrm{d}\mu \leq \lim_{n\to\infty} \inf_{k\geq n} \int f_k \mathrm{d}\mu = \liminf_{n\to\infty} \int f_n \mathrm{d}\mu \,.$$ □

Theorem 4.7 (Dominated convergence theorem)**.** *Suppose that $f_n \to f$ μ-almost everywhere. If there exists $g \in L_1(\mu)$ such that $|f_n| \leq g$ μ-almost everywhere then*

$$\int f_n \mathrm{d}\mu \to \int f \mathrm{d}\mu.$$

Proof. We now know that changing a function on a set of measure zero doesn't change its integral, so we can assume that $f_n \to f$ and $|f_n| \leq g$ for all n. It follows that $|f| \leq g$ so $f \in L_1(\mu)$ as well.

Now apply Fatou's lemma to both $g + f_n$ and $g - f_n$; since $\liminf_{n\to\infty} g + f_n = g + f$ and $\liminf_{n\to\infty} g - f_n = g - f$ we obtain

$$\int g + f \mathrm{d}\mu = \int \liminf_{n\to\infty} (g + f_n) \mathrm{d}\mu \leq \liminf_{n\to\infty} \int (g + f_n) \mathrm{d}\mu = g + \liminf_{n\to\infty} \int f_n \mathrm{d}\mu \,,$$

and

$$\int g - f \mathrm{d}\mu = \int \liminf_{n\to\infty} (g - f_n) \mathrm{d}\mu \leq \liminf_{n\to\infty} \int (g - f_n) \mathrm{d}\mu = g - \limsup_{n\to\infty} \int f_n \mathrm{d}\mu \,.$$

Subtracting $\int g\mathrm{d}\mu$ from both equations, this gives

$$\int f\mathrm{d}\mu \le \liminf_{n\to\infty} \int f_n\mathrm{d}\mu \le \limsup_{n\to\infty} \int f_n\mathrm{d}\mu \le \int f\mathrm{d}\mu\,,$$

so the limit of $\int f_n\mathrm{d}\mu$ must exist and equal $\int f\mathrm{d}\mu$. □

Corollary 4.8. *Suppose $\mu(\Omega) < \infty$. If $f_n \to f$ μ-almost everywhere and there is $M > 0$ such that $|f_n| \le M$ for all $n \ge 1$, then*

$$\int f_n\mathrm{d}\mu \to \int f\mathrm{d}\mu.$$

Proof. In this case the constant function $g \equiv M$ satisfies $\int g\mathrm{d}\mu = M\mu(\Omega) < \infty$, and $|f_n| \le g$ for all $n \ge 1$. □

Exercise 4.4. *(a) Fix $g \in L_1(\mu)$. Suppose that $\sum_{n\ge1} f_n$ converges μ-almost everywhere and that $|\sum_{k=1}^n f_k| \le g$ μ-almost everywhere, for all $n \ge 1$. Show that $f_n \in L_1(\mu)$ for all $n \ge 1$, that $\sum_{n\ge1} f_n \in L_1(\mu)$ and that*

$$\int \sum_{n\ge1} f_n\mathrm{d}\mu = \sum_{n\ge1} \int f_n\mathrm{d}\mu\,.$$

(b) Suppose that $\sum_{n\ge1}\int |f_n|\mathrm{d}\mu < \infty$. Prove that $\sum_{n\ge1} f_n$ is μ-a.e. absolutely convergent and that $\sum_{n\ge1}\int f_n\mathrm{d}\mu = \int \sum_{n\ge1} f_n\mathrm{d}\mu$.

4.1. **Expectation and independence.** All the theorems of the preceding section can be applied to real random variables defined over a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. In this setting, for a random variable $X : \Omega \to \mathbb{R}$ we have one additional way to write the integral: $\mathbf{E}X := \int X\mathrm{d}\mathbf{P}$; in this setting the integral is called the *expected value* of X.

So the theorems of the preceding section imply, for example, that if $0 \le X_n \uparrow X$ almost surely then $\mathbf{E}X_n \to \mathbf{E}X$ as $n \to \infty$; and if $|X_n| \le M$ for all n and $X_n \overset{\text{a.s.}}{\to} X$ then $|X| \le M$ almost surely and $\mathbf{E}X_n \to \mathbf{E}X$.

The main goal of the current section is to exhibit the strong connection between independence of random variables and factorization of expectations into product form.

Theorem 4.9 (Independence means multiply)**.** *Let $(X_i, 1 \le i \le n)$ be random variables defined over a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Then $(X_i, 1 \le i \le n)$ are independent if and only if*

$$\mathbf{E}\left[\prod_{k=1}^n f_k(X_k)\right] = \prod_{k=1}^n \mathbf{E}\left[f_k(X_k)\right] \tag{4.1}$$

for any bounded Borel measurable functions $f_k : \mathbb{R} \to \mathbb{R}$.

This theorem gives us the chance to introduce one of the last "simplifying techniques" of the notes: the *monotone class theorem*.

Theorem 4.10 (Monotone class theorem)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space, and let $\mathcal{P} \subset \mathcal{F}$ be a π-system over Ω with $\Omega \in \mathcal{P}$ and $\sigma(\mathcal{P}) = \mathcal{F}$. Let $\mathcal{S}$ be a collection of functions $f : \Omega \to \mathbb{R}$ with the following properties.*

(a) For all $P \in \mathcal{P}$, $\mathbf{1}_{[P]} \in \mathcal{S}$.
(b) For all $f, g \in \mathcal{S}$ and $c \in \mathbb{R}$, $cf + g \in \mathcal{S}$.
(c) If $(f_n, n \ge 1)$ are elements of $\mathcal{S}$ and $0 \le f_n \uparrow f$ for a bounded function f, then $f \in \mathcal{S}$.

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

Proof. Let $\Lambda = \{F \in \mathcal{F} : \mathbf{1}_{[F]} \in \mathcal{S}\}$. Then $\mathcal{P} \subset \Lambda$ by definition. Moreover, if $E, F \in \Lambda$ and $E \subset F$ then $\mathbf{1}_{[F\setminus E]} = \mathbf{1}_{[F]} - \mathbf{1}_{[E]} \in \mathcal{S}$ by (b) and so $F \setminus E \in \Lambda$. Also, if $F_n \uparrow F$ then $\mathbf{1}_{[F_n]} \uparrow \mathbf{1}_{[F]}$ and $\mathbf{1}_{[F]}$ is bounded so lies in $\mathcal{S}$; thus $F \in \Lambda$. This means that Λ is a λ-system, containing $\mathcal{P}$, so contains $\mathcal{F}$ by Dynkin's π-system lemma (Lemma 2.5).

We now know that $\mathbf{1}_{[F]} \in \mathcal{S}$ for all $F \in \mathcal{F}$. Since by (b), the collection $\mathcal{S}$ is closed under linear combinations, it follows that $\mathcal{S}$ contains all simple functions. Any bounded non-negative function is a monotone limit of simple functions by Lemma 4.3, so by (c) it follows that $\mathcal{S}$ contains all non-negative bounded measurable functions. Finally, for any bounded measurable function f, we may write $f = f^+ - f^-$ as a difference of bounded measurable functions, so another application of (b) shows that $f \in \mathcal{S}$. □

First proof of Theorem 4.9. First suppose that (4.1) holds for any bounded measurable functions $f_1, \ldots, f_n$. Fix any events $E_1, \ldots, E_n \in \mathcal{F}$ with $E_k \in \sigma(X_k)$. Since $\sigma(X_k) = \{X_k^{-1}(B), B \in \mathcal{B}(\mathbb{R})\}$, we may write $E_k = \{X_k \in B_k\}$ for some $B_k \in \mathcal{B}(\mathbb{R})$. It follows that

$$\begin{aligned}\mathbf{P}\left\{\bigcap_{k=1}^n E_k\right\} = \mathbf{P}\left\{\bigcap_{k=1}^n \{X_k \in B_k\}\right\} &= \mathbf{E}\left[\prod_{k=1}^n \mathbf{1}_{[B_k]}(X_k)\right] \\ &= \prod_{k=1}^n \mathbf{E}\left[\mathbf{1}_{[B_k]}(X_k)\right] = \prod_{k=1}^n \mathbf{P}\left\{X_k \in B_k\right\} = \prod_{k=1}^n \mathbf{P}\left\{E_k\right\}.\end{aligned}$$

Thus $X_1, \ldots, X_n$ are independent.

Conversely, suppose $X_1, \ldots, X_n$ are independent. Let

$$\begin{aligned}\mathcal{S}_n := \Big\{ f : \mathbb{R} \to \mathbb{R} \text{ bounded, Borel} : \forall\, B_1, \ldots, B_{n-1} \in \mathcal{B}(\mathbb{R}), \\ \mathbf{E}\Big[f(X_n) \cdot \prod_{k=1}^{n-1} \mathbf{1}_{[X_k \in B_k]}\Big] = \mathbf{E}\left[f(X_n)\right] \cdot \prod_{k=1}^{n-1} \mathbf{P}\left\{X_k \in B_k\right\} \Big\}.\end{aligned}$$

Then by assumption, $\mathcal{S}_n$ contains the indicator functions $\{\mathbf{1}_{[B]} : B \in \mathcal{B}(\mathbb{R})\}$. Moreover, if $f, g \in \mathcal{S}$ and $c \in \mathbb{R}$ then for any $B_1, \ldots, B_{n-1} \in \mathcal{B}(\mathbb{R})$, by linearity of expectation,

$$\begin{aligned}\mathbf{E}\Big[(cf+g)(X_n) \cdot \prod_{k=1}^{n-1} \mathbf{1}_{[X_k \in B_k]}\Big] &= c\mathbf{E}\Big[f(X_n) \cdot \prod_{k=1}^{n-1} \mathbf{1}_{[X_k \in B_k]}\Big] + \mathbf{E}\Big[g(X_n) \cdot \prod_{k=1}^{n-1} \mathbf{1}_{[X_k \in B_k]}\Big] \\ &= c\mathbf{E}\left[f(X_n)\right] \cdot \prod_{k=1}^{n-1} \mathbf{P}\left\{X_k \in B_k\right\} \Big\} + \mathbf{E}\left[g(X_n)\right] \cdot \prod_{k=1}^{n-1} \mathbf{P}\left\{X_k \in B_k\right\} \Big\} \\ &= \mathbf{E}\left[(cf+g)(X_n)\right] \cdot \prod_{k=1}^{n-1} \mathbf{P}\left\{X_k \in B_k\right\} \Big\},\end{aligned}$$

so $cf + g \in \mathcal{S}_n$. Also, if $0 \le f_m \uparrow f$ with f bounded and $f_m \in \mathcal{S}_n$ for all $m \ge 1$, then for any $B_1, \ldots, B_{n-1} \in \mathcal{B}(\mathbb{R})$, by the monotone convergence theorem

$$\begin{aligned}\mathbf{E}\Big[f(X_n) \cdot \prod_{k=1}^{n-1} \mathbf{1}_{[X_k \in B_k]}\Big] &= \lim_{m\to\infty} \mathbf{E}\Big[f_m(X_n) \cdot \prod_{k=1}^{n-1} \mathbf{1}_{[X_k \in B_k]}\Big] \\ &= \lim_{m\to\infty} \mathbf{E}\left[f_m(X_n)\right] \cdot \prod_{k=1}^{n-1} \mathbf{P}\left\{X_k \in B_k\right\} \\ &= \mathbf{E}\left[f(X_n)\right] \cdot \prod_{k=1}^{n-1} \mathbf{P}\left\{X_k \in B_k\right\}.\end{aligned}$$

so $f \in \mathcal{S}_n$. Thus $\mathcal{S}_n$ contains all bounded measurable functions. Next let

$$\begin{aligned}\mathcal{S}_{n-1} := \Big\{ f : \mathbb{R} \to \mathbb{R} \text{ Borel} : \forall\, B_1, \ldots, B_{n-2} \in \mathcal{B}(\mathbb{R}), \forall g : \mathbb{R} \to [0,\infty) \text{ Borel}, \\ \mathbf{E}\Big[f(X_{n-1})g(X_n) \cdot \prod_{k=1}^{n-2} \mathbf{1}_{[X_k \in B_k]}\Big] = \mathbf{E}f(X_{n-1}) \cdot \mathbf{E}g(X_n) \prod_{k=1}^{n-2} \mathbf{P}\left\{X_k \in B_k\right\} \Big\}.\end{aligned}$$

By repeating the same arguments as for $\mathcal{S}_n$, we see that $\mathcal{S}_{n-1}$ contains all bounded measurable functions (the monotone convergence theorem can be used since we took g non-negative). Repeating this argument (i.e. by induction), we obtain that

$$\mathbf{E}\left[\prod_{k=1}^{n} f_k(X_k)\right] = \prod_{k=1}^{n} \mathbf{E}\left[f_k(X_k)\right]$$

for any non-negative bounded Borel functions $f_1, \ldots, f_n$. Using linearity of expectation once more, it follows that this identity indeed holds for any bounded Borel functions. □

Second proof of Theorem 4.9. This proof replaces the use of the monotone class theorem with a direct argument (which has a similar flavour). We refer to (4.1) as "the factorization formula". The proof that if the factorization formula holds then $X_1, \ldots, X_n$ are independent is the same as in the first proof.

Now suppose that $X_1, \ldots, X_n$ are independent. Then for all $B_1, \ldots, B_n \in \mathcal{B}_n$, the events $(\{X_k \in B_k\}, 1 \leq k \leq n)$ are independent, so for any constants $c_1, \ldots, c_n \in \mathbb{R}$, writing $c = \prod_{i=1}^{n} c_i$, we have

$$\mathbf{E}\prod_{k=1}^{n} c_k \mathbf{1}_{[B_k]}(X_k) = c\mathbf{P}\left\{\bigcap_{k=1}^{n}\{X_k \in B_k\}\right\} = c\prod_{k=1}^{n}\mathbf{P}\left\{X_k \in B_k\right\} = \prod_{k=1}^{n}\mathbf{E}c_k\mathbf{1}_{[B_k]}(X_k)\,,$$

proving the factorization formula for (multiples of) indicator functions.

Now let $f_1, \ldots, f_n$ be simple Borel functions. Then we may write $f_k = \sum_{\ell=1}^{m} c_{k,\ell}\mathbf{1}_{[B_{k,\ell}]}$ for some real constants $(c_{k,\ell}, k \in [n], \ell \in [m])$ and Borel sets $(B_{k,\ell}, k \in [n], \ell \in [m])$. (We can always "pad" some of the sums so that they all have the same number of terms.) Then using linearity of expectation and the factorization formula for indicator functions,

$$\begin{aligned}
\mathbf{E}\prod_{k=1}^{n} f_k(X_k) &= \mathbf{E}\prod_{k=1}^{n}\sum_{\ell=1}^{m} c_{k,\ell}\mathbf{1}_{[B_{k,\ell}]}(X_{k,\ell}) \\
&= \sum_{\ell_1,\ldots,\ell_n=1}^{m}\mathbf{E}\prod_{k=1}^{n} c_{k,\ell_k}\mathbf{1}_{[B_{k,\ell_k}]}(X_k) \\
&= \sum_{\ell_1,\ldots,\ell_n=1}^{m}\prod_{k=1}^{n} c_{k,\ell_k}\mathbf{E}\mathbf{1}_{[B_{k,\ell_k}]}(X_k) \\
&= \prod_{k=1}^{n}\mathbf{E}\sum_{\ell=1}^{m} c_{k,\ell}\mathbf{1}_{[B_{k,\ell}]}(X_k) \\
&= \prod_{k=1}^{n}\mathbf{E}f_k(X_k)\,,
\end{aligned}$$

so the factorization formula holds for simple functions.

Now suppose $f_1, \ldots, f_n$ are non-negative Borel functions, and write $Y_k = f_k(X_k)$. Then Y_k is $\sigma(X_k)/\mathcal{B}(\mathbb{R})$-measurable, so $Y_1, \ldots, Y_n$ are independent. By Lemma 4.3, for each $1 \leq k \leq n$ we may find simple functions $(Y_{k,m}, m \geq 1)$ such that $0 \leq Y_{k,m} \uparrow Y_k$ and such that $Y_{k,m}$ is $\sigma(X_k)/\mathcal{B}(\mathbb{R})$-measurable for all $m \in \mathbb{N}$. Then $(Y_{1,m}, \ldots, Y_{n,m})$ are independent for all m, and $\prod_{k=1}^{n} Y_{k,m} \uparrow \prod_{k=1}^{n} Y_k$ as $m \to \infty$, so by the monotone convergence theorem and the factorization formula for simple functions,

$$\mathbf{E}\prod_{k=1}^{n} f_k(X_k) = \mathbf{E}\prod_{k=1}^{n} Y_k = \lim_{m\to\infty}\mathbf{E}\prod_{k=1}^{n} Y_{k,m} = \lim_{m\to\infty}\prod_{k=1}^{n}\mathbf{E}Y_{k,m} = \prod_{k=1}^{n}\mathbf{E}Y_k\,,$$

proving the factorization formula for non-negative functions.

Finally, if $f_1, \dots, f_n$ are bounded and Borel measurable then we can again use linearity of expectation to write $f(X_k) =: Y_k = Y_k^+ - Y_k^-$, and we then have

$$\begin{aligned}\mathbf{E}\prod_{k=1}^n f_k(X_k) &= \mathbf{E}\prod_{k=1}^n (Y_k^+ - Y_k^-)\\ &= \sum_{(\sigma_1,\dots,\sigma_n)\in\{-,+\}^n} (-1)^{\#\{k\in[n]:\sigma_k=-\}}\mathbf{E}\prod_{k=1}^n Y_k^{\sigma_k}\\ &= \sum_{(\sigma_1,\dots,\sigma_n)\in\{-,+\}^n} (-1)^{\#\{k\in[n]:\sigma_k=-\}}\prod_{k=1}^n \mathbf{E}Y_k^{\sigma_k} = \prod_{k=1}^n \mathbf{E}\left[Y_k^+ - Y_k^-\right] = \prod_{k=1}^n \mathbf{E}f_k(X_k)\end{aligned}$$

so the factorization formula holds in general. □

An extremely important corollary of Theorem 4.1 is that the factorization formula holds when the functions $f_1, \dots, f_n$ are simply the identity (this is not an immediate consequence of the theorem as the identity function is unbounded).

Corollary 4.11. *Suppose that $X_1, \dots, X_n$ are independent and either (a) $X_k \geq 0$ for $1 \leq k \leq n$ or (b) $X_k \in L_1(\mathbf{P})$ for $1 \leq k \leq n$. If (b) holds then $\prod_{k=1}^n X_k \in L_1(\mathbf{P})$ and if either (a) or (b) hold then $\mathbf{E}\prod_{k=1}^n X_k = \prod_{k=1}^n \mathbf{E}X_k$.*

In the proof (and later in the notes?), we use the following notation: for a function $f : \Omega \to \mathbb{R}$ and $r > 0$ we write $f_{\leq r} := f\mathbf{1}_{[|f|\leq r]}$. Notation $f_{\leq r}$

Proof. It suffices to prove the corollary when $n = 2$; the result then follows by induction. So suppose X, Y are independent. If X and Y are non-negative then by the monotone convergence theorem and by (4.1),

$$\mathbf{E}[XY] = \lim_{n\to\infty} \mathbf{E}[X_{\leq n}Y_{\leq n}] = \lim_{n\to\infty} \mathbf{E}X_{\leq n}\mathbf{E}Y_{\leq n} = \mathbf{E}X\mathbf{E}Y,$$

so if (a) holds then the factorization formula holds.

Next, if $X, Y \in L_1(\mathbf{P})$ then write $X = X^+ - X^-$ and $Y = Y^+ - Y^-$. Then $|XY| = (X^+ + X^-)(Y^+ + Y^-)$, so by linearity of expectation and the conclusion of the previous paragraph,

$$\begin{aligned}\mathbf{E}|XY| &= \mathbf{E}\left[X^+Y^+\right] + \mathbf{E}\left[X^+Y^-\right] + \mathbf{E}\left[X^-Y^+\right] + \mathbf{E}\left[X^-Y^-\right]\\ &= \mathbf{E}X^+\mathbf{E}Y^+ + \mathbf{E}X^+\mathbf{E}Y^- + \mathbf{E}X^-\mathbf{E}Y^+ + \mathbf{E}X^-\mathbf{E}Y^-\\ &= \mathbf{E}|X|\mathbf{E}|Y| < \infty,\end{aligned}$$

so $XY \in L_1(\mathbf{P})$. We may then again use linearity of expectation to deduce that

$$\begin{aligned}\mathbf{E}XY &= \mathbf{E}\left[X^+Y^+\right] - \mathbf{E}\left[X^+Y^-\right] - \mathbf{E}\left[X^-Y^+\right] + \mathbf{E}\left[X^-Y^-\right]\\ &= \mathbf{E}X^+\mathbf{E}Y^+ - \mathbf{E}X^+\mathbf{E}Y^- - \mathbf{E}X^-\mathbf{E}Y^+ + \mathbf{E}X^-\mathbf{E}Y^-\\ &= \mathbf{E}X\mathbf{E}Y.\end{aligned}$$

□

5. **An interlude: the probabilistic method.**

One of the challenges of teaching a first rigorous probability course is the amount of setup that's required before one gets to "the real stuff". Billingsley's textbook avoids this issue by focussing exclusively on simple functions in the early chapters. This makes the book more engaging at the outset; the cost is that many of the most important random variables (Gaussian, exponential, Gamma, Beta, …) are excluded from consideration.

My approach this course has been to bite the bullet and do the necessary setup, while doing my best to motivate its development. I've also postponed a few things that in most courses would have already been introduced or would be next on the menu: Fubini's theorem, convergence in L_p, Hölder, Minkowski and Cauchy-Schwartz inequalities, to name a few.

Even so, I know that the first third to half of the course can feel like a bit of a slog. To liven things up a bit, I've decided to describe one of the ways in which probability has contributed to other branches of mathematics: the *probabilistic method*. In a nutshell, the idea of the probabilistic method is this. One wishes to show the existence of a mathematical object m with some property P. Rather constructing m directly, one instead constructs a *random* object M and shows that

$$\mathbf{P}\{M \text{ has property } P\} > 0.$$

This immediately shows that there must be at least one object m with property P, proving existence.

Example 1: existence of continuous nowhere differentiable functions. Let $D_n := \{i/2^n, 0 \leq i \leq 2^n\}$, so that $D := \bigcup_{n\geq 0} D_n$ are the dyadic rationals in $[0,1]$. Note that $D_{n-1} \subset D_n$ for each $n \geq 1$. Let $(N_x, x \in D)$ be IID$N(0,1)$ random variables. Define a sequence of random functions B_n from $[0,1]$ to $\mathbb{R}$ as follows.

D_n: n'th level dyadic rationals in $[0,1]$.

For $x \in [0,1]$ let $B_1(x) = xN_1$. Actually, a more wordy but equivalent definition of B_1 presages[9] the subsequent construction more effectively. Let $B_1(0) = 0$ and let $B_1(1) = N_1$; then for $x \in (0,1)$ define $B_1(x)$ by linear interpolation between points of $D_0 = \{0,1\}$.

Inductively, given B_{n-1}, let

$$B_n(x) = \begin{cases} B_{n-1}(x) & \text{if } x \in D_{n-1} \\ B_{n-1}(x) + \frac{N_x}{\sqrt{2^n}} & \text{if } x \in D_n \setminus D_{n-1} \\ pB_n(i/2^n) + (1-p)B_n((i+1)/2^n) & \text{if } x = \frac{pi+(1-p)i+1}{2^n}, p \in (0,1)\,. \end{cases}$$

Then it is possible to show that[10]

$$\mathbf{P}\{(B_n, n \geq 0) \text{ is a uniformly convergent sequence of functions}\} = 1\,,$$

so one may define a random function B_∞ as the almost sure limit of the sequence B_n. The limit B_∞ is *Brownian motion* on the interval $[0,1]$. The fact that B_∞ is a.s. a uniform limit of continuous functions implies that B_∞ is a.s. continuous. However, it turns out that

a.s.: almost surely

$$\mathbf{P}\{B_\infty \text{ is nowhere differentiable}\} = 1.$$

Thus, Brownian motion provides an example of a continuous, nowhere differentiable function. In fact, Brownian motion is (in a sense which can be made precise) a *uniformly random continuous function*, so the above statement can be interpreted as stating that *almost all continuous functions are nowhere differentiable*.

Example 2: small-norm signings of vectors.

This example is a bit more down-to-earth, and we may actually prove all our statements with the machinery we currently have available to us. It is drawn from Alon and Spencer's book "The probabilistic method".

Proposition 5.1. *Let $v_1, \ldots, v_n$ be vectors in $\mathbb{R}^m$ (for some m) with $|v_i| = 1$ for all i. Then there exist $\sigma_1, \ldots, \sigma_n \in \{-1, 1\}$ such that*

$$|\sigma_1 v_1 + \ldots + \sigma_n v_n| \leq \sqrt{n}.$$

Proof. Let $\sigma_1, \ldots, \sigma_n$ be independent and uniform on $\{-1, 1\}$. Set

$$X = |\sigma_1 v_1 + \ldots + \sigma_n v_n|^2 = (\sigma_1 v_1 + \ldots + \sigma_n v_n) \cdot (\sigma_1 v_1 + \ldots + \sigma_n v_n) = \sum_{i,j=1}^n \sigma_i \sigma_j v_i \cdot v_j\,.$$

Then by linearity of expectation,

$$\mathbf{E}X = \mathbf{E}\sum_{i,j=1}^n \sigma_i \sigma_j v_i \cdot v_j = \sum_{i,j=1}^n \mathbf{E}\left[\sigma_i \sigma_j\right] v_i \cdot v_j\,.$$

[9] presage, v.: 1. transitive. a. To constitute a supernatural sign of (a future event); to be an omen of, to portend. b. To be indicative or suggestive of; to be a natural precursor of, to give warning of. –Oxford English Dictionary

[10] discuss measurability issues?

If $i \neq j$ then σ_i and σ_j are independent so $\mathbf{E}[\sigma_i \sigma_j] = \mathbf{E}\sigma_i \mathbf{E}\sigma_j = 0$; so the above identity simplifies to

$$\mathbf{E}X = \sum_{i=1}^{n} \mathbf{E}\left[\sigma_i^2\right] v_i \cdot v_i = \sum_{i=1}^{n} 1 = n$$

since $v_i \cdot v_i = |v_i|^2 = 1$ and $\sigma_i^2 \equiv 1$.

But if $\mathbf{E}X = n$ then $\mathbf{P}\{X \leq n\} > 0$ by monotonicity of expectations; so there must be some choice of $\sigma_1, \ldots, \sigma_n$ which makes $X \leq n$, and for this choice

$$|\sigma_1 v_1 + \ldots + \sigma_n v_n| = X^{1/2} \leq \sqrt{n}\,. \qquad \square$$

6. **Densities and change of variables**

This section is about how to actually do computations with random variables and their expectations.

If X is a random variable taking values in $\mathbb{N}$, then $X = \lim_{n\to\infty} X_{\leq n}$, and this is an increasing limit. For each n, $X_{\leq n}$ is a simple function so by the definition of the integral of simple functions, we have

$$\mathbf{E}X = \lim_{n\to\infty} \mathbf{E}X_{\leq n} = \lim_{n\to\infty} \sum_{k=0}^{n} k\mathbf{P}\{X_{\leq n} = k\} = \sum_{k\geq 0} k\mathbf{P}\{X = k\}\,.$$

More generally, if X is non-negative and takes values in a countable set N, then the same sort of argument gives that $\mathbf{E}X = \sum_{n\in N} n\mathbf{P}\{X = n\}$. This allows us to do computations with discrete random variables. For example, if P is Poisson(λ) then

$$\mathbf{E}P = \sum_{k\geq 0} k \cdot \frac{\lambda^k e^{-\lambda}}{k!} = \lambda \cdot \sum_{k\geq 1} \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!} = \lambda.$$

But it's not yet clear how to tackle computations involving non-discrete random variables. For example, suppose that $(N_i, i \geq 1)$ are independent standard Gaussian random variables, independent of P. How should we compute (or even approximate)

$$\mathbf{P}\left\{\sum_{i=1}^{P} N_i \geq 1\right\} = \mathbf{E}\left[\mathbf{1}_{[\sum_{i=1}^{P} N_i \geq 1]}\right] = \int_\Omega \mathbf{1}_{[\sum_{i=1}^{P} N_i \geq 1]}(\omega)\mathrm{d}\mathbf{P}\,?$$

Using the tools we now develop.

Definition 6.1. *Given a measure space $(\Omega, \mathcal{F}, \mu)$ and $f : \Omega \to \mathbb{R}$ non-negative and $\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable, define a new measure μf on $(\Omega, \mathcal{F})$ by setting*

$$\mu f(A) = \int_A f\mathrm{d}\mu := \int f\mathbf{1}_{[A]}\mathrm{d}\mu\,.$$

If $\nu = \mu f$ then we say ν has density f with respect to μ. Density

Exercise 6.1. *The function μf defined above is a measure on $(\Omega, \mathcal{F})$.*

Exercise 6.2. *If $f' \overset{\mu-\text{a.e.}}{=} f$ then $\mu f = \mu f'$.*

We also say a real random variable X has density f with respect to Lebesgue measure if its distribution μ_X has density f with respect to Lebesgue measure, or in other words if

$$\mu_X(B) = \int_B f(x)\mathrm{d}x$$

for any Borel $B \subset \mathbb{R}$. In this case we also say that f is the *probability density function* of X. These definitions are justified by the following two results.

Proposition 6.2. *Fix a σ-finite measure space $(\Omega, \mathcal{F}, \mu)$ and measurable $f : \Omega \to [0, \infty)$, and suppose ν has density f with respect to μ. Then for measurable $g : \Omega \to \mathbb{R}$,*

$$\int g \mathrm{d}\nu = \int gf \mathrm{d}\mu$$

provided that either $g \geq 0$ or $g \in L_1(\nu)$. Moreover, $g \in L_1(\nu)$ if and only if $gf \in L_1(\mu)$.

Proof. If $g = \mathbf{1}_{[A]}$ for some $A \in \mathcal{F}$ then the equality of the two integrals holds by definition. The theorem then follows straightforwardly using the monotone class theorem and the monotone convergence theorem. □

Proposition 6.3 (Change of variables formula)**.** *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X : \Omega \to \mathbb{R}$ be a random variable. Then for all measurable $g : \mathbb{R} \to \mathbb{R}$ with $\mathbf{E}|g(X)| < \infty$,*

$$\mathbf{E}g(X) = \int_{\mathbb{R}} g(x)\mu_X(\mathrm{d}x)\,. \tag{6.1}$$

Moreover, if X has density f with respect to Lebesgue measure then also $\mathbf{E}g(X) = \int_{\mathbb{R}} g(x)f(x)\mathrm{d}x$.

Proof. Again, the assertions are true if $g = \mathbf{1}_{[B]}$ for Borel $B \subset \mathbb{R}$, and the rest of the proof follows using the monotone class theorem and the monotone convergence theorem. □

More generally, if $(\Omega, \mathcal{F}, \mu)$ is a σ-finite measure space and $(S, \mathcal{S})$ is another measurable space, then for an $(\mathcal{F}/\mathcal{S})$-measurable function $f : \Omega \to S$ we may define

$$\nu(E) = \mu(f^{-1}(E))$$

for $E \in \mathcal{S}$. Then for all non-negative $(\mathcal{S}/\mathcal{B}(\mathbb{R}))$-measurable functions $g : S \to \mathbb{R}$, we have

$$\int g \mathrm{d}\nu = \int g \circ f \mathrm{d}\mu\,.$$

The change of variables formula (6.1) is a special case of this, but this also tells us, for example, that if $\mathrm{X} = (X_1, \ldots, X_n)$ are random variables defined on a common space, then

$$\mathbf{E}g(X_1, \ldots, X_n) = \int_{\mathbb{R}^n} g(\vec{x})\mu_{\mathrm{X}}(\vec{x})\,.$$

Making this a useful computational tool, even for independent random variables, will require Fubini's theorem.

C_1 **Exercise 6.3.** *Prove that if $\varphi : [a, b] \to \mathbb{R}$ is C_1 (continuously differentiable) and strictly increasing then for any Borel function $g : [\phi(a), \phi(b)] \to [0, \infty)$,*

$$\int_{\varphi(a)}^{\varphi(b)} g(y)\mathrm{d}y = \int_a^b g(\phi(y))\varphi'(y)\mathrm{d}y\,.$$

Example: size-biasing the Poisson and folded normal distributions. Let X be a non-negative random variable on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ with $0 < \mathbf{E}X < \infty$. Then $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_X)$ is another probability space. The *size-biased* distribution of X is $\hat{\mu}_X := (\mu_X \cdot X/\mathbf{E}X)$. In other words, for Borel $B \subset \mathbb{R}$,

$$\hat{\mu}_X(A) = (\mu_X \cdot X/\mathbf{E}X)(A) = \mathbf{E}\left[\frac{X}{\mathbf{E}X}\mathbf{1}_{[X\in A]}\right]\,.$$

This is another probability distribution on $\mathbb{R}$, since $\hat{\mu}_X(\mathbb{R}) = \mathbf{E}\left[X/\mathbf{E}X\right] = 1$.

Suppose P is Poisson(λ). Then $\mathbf{E}P = \lambda$ so for Borel $B \subset \mathbb{R}$,

$$\hat{\mu}_P(B) = \left(\mu_P \frac{P}{\lambda}\right)(B) = \mathbf{E}\left[\frac{P}{\lambda}\mathbf{1}_{[P\in B]}\right] = \sum_{k\geq 1, k\in B} \frac{k}{\lambda}\mathbf{P}\{P=k\}$$
$$= \sum_{k\geq 1,k\in B} \frac{k}{\lambda}\frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k\geq 1,k\in B} \frac{\lambda^{k-1}e^{-\lambda}}{(k-1)!} = \mathbf{P}\{P+1\in B\} = \mu_{P+1}(B)\,.$$

In other words, the size-biased distribution of P is just the distribution of $P+1$. Of course, nothing like this need hold in general.

Next suppose N is a standard Gaussian; so N has density $\Phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ with respect to Lebesgue measure. The distribution of $|N|$ is called the *folded normal* distribution; it has density $\psi(x) = 2\Phi(x)\mathbf{1}_{[x\geq 0]} = \sqrt{2/\pi}e^{-x^2/2}\mathbf{1}_{[x\geq 0]}$ with respect to Lebesgue measure.

The size-biasing of the distribution of $|N|$ is $\hat{\mu}_{|N|} = (\mu_{|N|}\cdot |N|/\mathbf{E}|N|)$. To find an explicit formula for this, we first use the change of variables formula to compute

$$\mathbf{E}|N| = \int_{\mathbb{R}} x\mathrm{d}\mu_{|N|} = \int_{[0,\infty} x\cdot\sqrt{\frac{2}{\pi}}e^{-x^2/2}\mathrm{d}x = \left[-\sqrt{\frac{2}{\pi}}e^{-x^2/2}\right]_0^\infty = \sqrt{\frac{2}{\pi}}.$$

It follows that for $B\subset[0,\infty)$ Borel,

$$\hat{\mu}_{|N|}(B) = \int \mathbf{1}_{[B]}\cdot\frac{|N|}{\mathbf{E}|N|}\mathrm{d}\mu = \int_B \frac{x}{\sqrt{2/\pi}}\sqrt{\frac{2}{\pi}}e^{-x^2/2}\mathrm{d}x = \int_B xe^{-x^2/2}\mathrm{d}x\,.$$

Thus, $\hat{\mu}_{|N|}$ has density $xe^{-x^2/2}\mathbf{1}_{[x\geq 0]}$ with respect to Lebesgue measure. This is called the *Rayleigh distribution*.

Rayleigh distribution

Exercise 6.4. *If X, Y are independent standard Gaussians then $\sqrt{X^2+Y^2}$ is Rayleigh distributed.*

6.1. **Product measure and Fubini's theorem.** Let $(\Omega,\mathcal{F},\mathbf{P})$ be a probability space. If $X, Y : \Omega\to\mathbb{R}$ are independent random variables and $f, g : \mathbb{R}\to\mathbb{R}$ are bounded Borel functions then $\mathbf{E}f(X)g(Y) = \mathbf{E}f(X)\mathbf{E}g(Y)$. By Exercise 3.2 (e), the pair (X,Y) is $\Omega/\mathcal{B}(\mathbb{R}^2)$-measurable; in other words, it is an $\mathbb{R}^2$-valued random variable. What is its distribution $\mu_{(X,Y)}$? Note that by the factorization formula, if $A, B\in\mathcal{B}(\mathbb{R})$ then

$$\mu_{(X,Y)}(A\times B) = \mathbf{P}\{(X,Y)\in A\times B\} = \mathbf{P}\{X\in A\}\,\mathbf{P}\{Y\in B\} = \mu_X(A)\mu_Y(B).$$

The collection $\mathcal{B}(\mathbb{R})\times\mathcal{B}(\mathbb{R}) = \{A\times B : A, B\in\mathcal{B}(\mathbb{R})\}$ generates $\mathcal{B}(\mathbb{R}^2)$, so the preceding formula uniquely identifies $\mu_{(X,Y)}$ as the *product measure* of measures μ_X and μ_Y.

This concrete example is generalized as follows. Fix measurable spaces $(M,\mathcal{M})$ and $(N,\mathcal{N})$. Sets $A\times B\in\mathcal{M}\times\mathcal{N}$ are called *rectangles*. Let $\mathcal{M}\boxtimes\mathcal{N}$ be the *field* generated by $\mathcal{M}\times\mathcal{N}$.

Rectangles
$\mathcal{M}\boxtimes\mathcal{N}$

Exercise 6.5. *It holds that*

$$\mathcal{M}\boxtimes\mathcal{N} = \Big\{\bigcup_{i=1}^n A_i\times B_i : n\geq 1; \forall i\in[n], A_i\times B_i\in\mathcal{M}\times\mathcal{N}\Big\}$$
$$= \Big\{\bigcup_{i=1}^n A_i\times B_i : n\geq 1; \forall i\in[n], A_i\times B_i\in\mathcal{M}\times\mathcal{N}; A_1\times B_1,\ldots,A_n\times B_n \textit{ disjoint}\Big\}$$

The product measurable space $(M\times N, \mathcal{M}\otimes\mathcal{N})$ is defined by setting

Product measurable space
$\mathcal{M}\otimes\mathcal{N}$

$$\mathcal{M}\otimes\mathcal{N} := \sigma(\mathcal{M}\times\mathcal{N}) = \sigma(A\times B : A\in\mathcal{M}, B\in\mathcal{N}) = \sigma(\mathcal{M}\boxtimes\mathcal{N}).$$

If μ and ν are σ-finite measures on $(M,\mathcal{M})$ and $(N,\mathcal{N})$ respectively, define a function $\mu\boxtimes\nu$ on $\mathcal{M}\boxtimes\mathcal{N}$ by setting

$$\mu\boxtimes\nu\Big(\bigcap_{i=1}^n A_i\times B_i\Big) := \sum_{i=1}^n \mu(A_i)\nu(B_i)\,,$$

for disjoint rectangles $A_1 \times B_1, \ldots, A_n \times B_n \in \mathcal{M} \times \mathcal{N}$.

Exercise 6.6. *The function $\mu \boxtimes \nu$ is well-defined, in that if $P \in \mathcal{M} \otimes \mathcal{N}$ may be represented as a disjoint union of rectangles in multiple ways,*

$$P = \bigcup_{i=1}^{n} A_i \times B_i = \bigcup_{i=1}^{m} C_i \times D_i$$

then $\sum_{i=1}^{n} \mu(A_i)\nu(B_i) = \sum_{i=1}^{m} \mu(C_i)\nu(D_i)$.

Proposition 6.4. *If $(M, \mathcal{M}, \mu)$ and $(N, \mathcal{N}, \nu)$ are σ-finite measure spaces then $\mu \boxtimes \nu$ is a pre-measure on $\mathcal{M} \boxtimes \mathcal{N}$.*

Assuming the proposition holds, it follows by the Carathéodory Extension Theorem and Dynkin's Uniqueness theorem that $\mu \boxtimes \nu$ extends uniquely to a measure $\mu \otimes \nu$ on $\mathcal{M} \otimes \mathcal{N}$. This extension is called the *product measure* of μ and ν.

Product measure $\mu \otimes \nu$.

Exercise 6.7 (Product measure is commutative)**.** *Suppose $(M, \mathcal{M}, \mu)$ and $(N, \mathcal{N}, \nu)$ are σ-finite measure spaces. Let $\mu \otimes \nu$ be product measure on $\mathcal{M} \otimes \mathcal{N}$ and let $\nu \otimes \mu$ be product measure on $\mathcal{N} \otimes \mathcal{M}$. Prove that for all $B \in \mathcal{M} \otimes \mathcal{N}$, $B^* := \{(b, a) : (a, b) \in B\} \in \mathcal{N} \otimes \mathcal{M}$, and $(\nu \otimes \mu)(B^*) = (\mu \otimes \nu)(B)$.*

We extract two steps of the proof of Proposition 6.4 as separate lemmas. This proof is based on the one given by Norris in his lecture notes on probability and measure.

Lemma 6.5. *Under the hypotheses of Proposition 6.4, if $f : M \otimes N \to \mathbb{R}$ is $(\mathcal{M} \otimes \mathcal{N})/\mathcal{B}(\mathbb{R})$-measurable then for all $a \in \mathcal{M}$, the function $f_a : \mathcal{N} \to \mathbb{R}$ given by $f_a(b) := f(a, b)$ is $(\mathcal{N}/\mathcal{B}(\mathbb{R}))$-measurable.*

Proof. Write

$$\mathcal{S} := \{f : M \times N \to \mathbb{R} : \forall a \in \mathcal{M}, f_a \text{ is } (\mathcal{N}/\mathcal{B}(\mathbb{R})) - \text{measurable}\}.$$

We aim to show $\mathcal{S}$ contains all $(\mathcal{M} \otimes \mathcal{N})/\mathcal{B}(\mathbb{R})$-measurable functions.

First, if $f = \mathbf{1}_{[A \times B]}$ for $A \times B \in \mathcal{M} \times \mathcal{N}$ then for $a \in A$, $f_a \equiv \mathbf{1}_{[B]}$, and for $a \notin A$, $f_a \equiv 0$. In both cases f_a is measurable so$f \in \mathcal{S}$; thus $\mathcal{S}$ contains indicators of rectangles.

Next, if $f, g \in \mathcal{S}$ and $c \in \mathbb{R}$ then for all $a \in \mathcal{M}$,

$$(cf + g)_a(b) = (cf + g)(a, b) = cf(a, b) + g(a, b) = cf_a(b) + g_a(b) = (cf_a + g_a)(b)\,,$$

so $(cf + g)_a$ is a linear combination of $(\mathcal{N}/\mathcal{B}(\mathbb{R}))$-measurable functions and so is $(\mathcal{N}/\mathcal{B}(\mathbb{R}))$-measurable. Therefore $\mathcal{S}$ is closed under linear combinations.

Moreover, if $(f^{(n)}, n \geq 1)$ is a sequence of elements of $\mathcal{S}$ and $0 \leq f^{(n)} \uparrow f$ for some bounded function f, then for all $a \in \mathcal{M}$, $f_a^{(n)} \uparrow f_a$. As a monotone limit of measurable functions, f_a is $(\mathcal{N}/\mathcal{B}(\mathbb{R}))$-measurable; thus $f \in \mathcal{S}$.

Since rectangles form a π-system generating $\mathcal{M} \otimes \mathcal{N}$, by the monotone class theorem it follows that $\mathcal{S}$ contains all bounded $(\mathcal{M} \otimes \mathcal{N})/\mathcal{B}(\mathbb{R})$-measurable functions.

Finally, if $f : M \times N \to \mathbb{R}$ is any $(\mathcal{M} \otimes \mathcal{N})/\mathcal{B}(\mathbb{R})$-measurable we may write f as a limit of bounded measurable functions $f = \lim_{n\to\infty} f^{(n)}$ by taking $f^{(n)} = f\mathbf{1}_{[|f| \leq n]}$. For all $a \in \mathcal{M}$ we then have $f_a = \lim_{n\to\infty} f_a^{(n)}$ so f_a is a limit of $\mathcal{N}/\mathcal{B}(\mathbb{R})$-measurable functions and so is $\mathcal{N}/\mathcal{B}(\mathbb{R})$-measurable. Thus $f \in \mathcal{S}$. □

Lemma 6.6. *Under the hypotheses of Proposition 6.4, Let $f : M \times N \to \mathbb{R}$ be $\mathcal{M} \otimes \mathcal{N}/\mathcal{B}(\mathbb{R})$-measurable and either bounded or non-negative. Define $f_M : M \to \mathbb{R} \cup \{+\infty\}$ by*

$$f_M(a) := \int_N f(a, b)\nu(\mathrm{d}b).$$

If $\nu(N) < \infty$ and f is bounded then f_M is bounded and $\mathcal{M}/\mathcal{B}(\mathbb{R})$-measurable. Also, if f is non-negative then $f_M : M \to [0, \infty]$ is $\mathcal{M}/\mathcal{B}(\mathbb{R}^*)$-measurable.

Proof. Note that by the definition of f_a we have

$$f_M(a) := \int_N f_a(b)\nu(\mathrm{d}b),$$

so the integral in the lemma statement at least makes sense by Lemma 6.5.

Suppose $\nu(N) < \infty$. We wish to show that for any bounded $\mathcal{M}\otimes\mathcal{N}/\mathcal{B}(\mathbb{R})$-measurable function f, the function f_M is $\mathcal{M}/\mathcal{B}(\mathbb{R})$-measurable.

First, if $f = \mathbf{1}_{[A\times B]}$ for $A \times B \in \mathcal{M} \times \mathcal{N}$, then for $a \in A$,

$$f_M(a) = \int_N \mathbf{1}_{[B]}(b)\nu(\mathrm{d}b) = \nu(B) < \infty,$$

and for $a \notin A$, $f_M(a) = \int_N 0\nu(\mathrm{d}b) = 0$. Thus $f_M \equiv \nu(B)\mathbf{1}_{[A]}$ is bounded and $\mathcal{M}/\mathcal{B}(\mathbb{R})$-measurable. Next, if f, g are bounded functions such that f_M and g_M are $\mathcal{M}/\mathcal{B}(\mathbb{R})$-measurable and $c \in \mathbb{R}$ then $(cf + g)_M = cf_M + g_M$ by linearity of integration, so $(cf + g)_M$ is bounded and $\mathcal{M}/\mathcal{B}(\mathbb{R})$-measurable. Finally, if $0 \le f^{(n)} \uparrow f$ then by the monotone convergence theorem,

$$f_M(a) = \int_N f_a(b)\nu(\mathrm{d}b) = \lim_{n\to\infty} \int_N f_a^{(n)}(b)\nu(\mathrm{d}b) = \lim_{n\to\infty} f_M^{(n)}(a),$$

so f_M is an increasing limit of measurable functions and thus measurable. The first assertion of the lemma follows by the monotone class theorem.

The second assertion of the lemma follows by a similar argument. □

In the course of the preceding proof, we derived that if $f = \mathbf{1}_{[A\times B]}$ for a rectangle $A \times B$, then $f_M = \nu(B)\mathbf{1}_{[A]}$, which implies that

$$\int_M \int_N f(a, b)\nu(\mathrm{d}b)\mu(\mathrm{d}a) = \int_M f_M(a)\mu(\mathrm{d}a) = \int_M \nu(B)\mathbf{1}_{[A]}(a)\mu(\mathrm{d}a) = \nu(B)\mu(A)\,; \quad (6.2)$$

we will use this in the next proof.

Proof of Proposition 6.4. The function $\mu\boxtimes\nu$ is obviously non-negative, and it is additive by definition. To prove $\mu \boxtimes \nu$ is a pre-measure, it suffices to show that it is countably additive.

So suppose that $\bigcup_{i=1}^k A_i \times B_i \in \mathcal{M} \boxtimes \mathcal{N}$ is a finite disjoint union of rectangles which may also be represented as an infinite disjoint union of rectangles,

$$\bigcup_{i=1}^k A_i \times B_i = \bigcup_{i\ge 1} C_i \times D_i.$$

Using (6.2), we have

$$\mu \otimes \nu\Big(\bigcup_{i=1}^k A_i \times B_i\Big) = \sum_{i=1}^k \mu(A_i)\nu(B_i) = \sum_{i=1}^k \int_M \int_N \mathbf{1}_{[A_i\times B_i]}(a, b)\mu(\mathrm{d}a)\nu(\mathrm{d}b)\,.$$

We may use linearity of integration twice to bring the sum inside the two integrals in the final term. Since $\sum_{i=1}^k \mathbf{1}_{[A_i\times B_i]} = \mathbf{1}_{[\bigcup_{i=1}^k A_i\times B_i]}$, it follows that

$$\mu \otimes \nu\Big(\bigcup_{i=1}^k A_i \times B_i\Big) = \int_M \int_N \mathbf{1}_{[\bigcup_{i=1}^k A_i\times B_i]}\mathrm{d}\mu\mathrm{d}\nu = \int_M \int_N f\mathrm{d}\mu\mathrm{d}\nu\,,$$

where we have taken $f := \mathbf{1}_{[\bigcup_{i=1}^k A_i\times B_i]}$.

Now write $f^{(n)} = \mathbf{1}_{[\bigcup_{i=1}^n C_i\times D_i]}$. Repeating the above logic gives

$$\mu \otimes \nu\Big(\bigcup_{i=1}^n C_i \times D_i\Big) = \int_M \int_N \mathbf{1}_{[\bigcup_{i=1}^n C_i\times D_i]}\mathrm{d}\mu\mathrm{d}\nu = \int_M \int_N f^{(n)}\mathrm{d}\mu\mathrm{d}\nu\,,$$

Also, $f^{(n)} \uparrow f$ since $\bigcup_{i=1}^{\infty} C_i \times D_i = \bigcup_{i=1}^{k} A_i \times B_i$, so for all $a \in M$, $f_a^{(n)} \uparrow f_a$, so by the monotone convergence theorem,

$$f_M^{(n)}(a) = \int_N f_a^{(n)}(b)\nu(\mathrm{d}b) \nearrow \int_N f_a(b)\nu(\mathrm{d}b) = f_M(a)\,.$$

Since this convergence is monotone, another application of the monotone convergence theorem gives that

$$\int_M \int_N f^{(n)} \mathrm{d}\nu \mathrm{d}\mu = \int_M f_M^{(n)} \mathrm{d}\mu \to \int_M f \mathrm{d}\mu = \int_M \int_N f \mathrm{d}\nu \mathrm{d}\mu = \mu \otimes \nu\Big(\bigcup_{i\geq 1} C_i \times D_i\Big)\,.$$

But also

$$\int_M \int_N f^{(n)} \mathrm{d}\nu \mathrm{d}\mu = \mu \otimes \nu\Big(\bigcup_{i=1}^{n} C_i \times D_i\Big) = \sum_{i=1}^{n} \mu(C_i)\nu(D_i) \to \sum_{i=1}^{\infty} \mu(C_i)\nu(D_i)\,.$$

Comparing the two preceding displays, we see that

$$\mu \otimes \nu\Big(\bigcup_{i\geq 1} C_i \times D_i\Big) = \sum_{i=1}^{\infty} \mu(C_i)\nu(D_i) = \sum_{i\geq 1} \mu \otimes \nu(C_i \times D_i)\,;$$

thus $\mu \otimes \nu$ is indeed a pre-measure. □

Theorem 6.7 (Fubini's theorem)**.** *Let $(M, \mathcal{M}, \mu)$ and $(N, \mathcal{N}, \nu)$ be σ-finite measure spaces, and let $f : M \times N \to \mathbb{R}$ be $(\mathcal{M} \otimes \mathcal{N})/\mathcal{B}(\mathbb{R})$-measurable.*

(a) If $f \geq 0$ then

$$\int f \mathrm{d}(\mu \otimes \nu) = \int_M \int_N f \mathrm{d}\nu \mathrm{d}\mu\,. \tag{6.3}$$

(b) If $f \in L_1(\mu \otimes \nu)$ then with $F := \{a \in M : \int_N |f(a,b)|\nu(\mathrm{d}b) < \infty\}$, it holds that $\mu(M \setminus F) = 0$. Moreover, setting

$$\hat{f}_M(a) = \begin{cases} \int_N f(a,b)\nu(\mathrm{d}b) & \text{if } a \in F \\ 0 & \text{if } a \notin F \end{cases}$$

then $\hat{f}_M \in L_1(\mu)$, and

$$\int \hat{f}_M \mathrm{d}\mu = \int f \mathrm{d}(\mu \otimes \nu).$$

Part (b) of the theorem implies the following. Set $Z = F \times N$. Then $(\mu \otimes \nu)(Z^c) = \mu(M \setminus F)\nu(N) = 0 \cdot \nu(N) = 0$, so $f\mathbf{1}_{[Z]} : M \times N \to \mathbb{R}$ is $(\mu \otimes \nu)$-a.e. equal to f, and

$$\int f \mathrm{d}(\mu \otimes \nu) = \int f\mathbf{1}_{[Z]} \mathrm{d}(\mu \otimes \nu) = \int_M \int_N f(a,b)\mathbf{1}_{[Z]}(a,b)\nu(\mathrm{d}b)\mu(\mathrm{d}a)\,. \tag{6.4}$$

The only thing preventing us from removing the indicator from the double integral is that otherwise there can exist points $a \in M$ where the inner integral is not defined.

Proof. We first assume both measure spaces are finite. First, if $f = \mathbf{1}_{[A\times B]}$ for $A \times B \in \mathcal{M} \times \mathcal{N}$ then the identity holds by (6.2). Write

$$\mathcal{S} = \Big\{ f : M \times N \to \mathbb{R} : \int f \mathrm{d}(\mu \otimes \nu) = \int_M \int_N f \mathrm{d}\nu \mathrm{d}\mu \Big\}\,.$$

Using linearity of integration and the monotone convergence theorem, it is not hard to check the conditions to see that $\mathcal{S}$ satisfies the conditions of the monotone class theorem. It then follows that (6.2) holds for all bounded $(\mathcal{M} \otimes \mathcal{N})/\mathcal{B}(\mathbb{R})$-measurable functions $f : M \times N \to \mathbb{R}$.

Next, suppose f is non-negative, and for $n \geq 1$ write $f^{(n)} = \min(f, n)$. Then $f^{(n)} \uparrow f$ so by the monotone convergence theorem

$$\int f^{(n)} \mathrm{d}\mu \otimes \nu \uparrow \int f \mathrm{d}\mu \otimes \nu.$$

Writing $f_a^{(n)}(b) = f^{(n)}(a, b)$, for all $a \in M$, we also have $f_a^{(n)} \uparrow f_a$, so

$$f_M^{(n)}(a) = \int_N f_a^{(n)}(b)\nu(\mathrm{d}b) \nearrow \int_N f_a(b)\nu(\mathrm{d}b)\,,$$

so

$$\int_M \int_N f^{(n)} \mathrm{d}\nu \mathrm{d}\mu \to \int_M \int_N f^{(n)} \mathrm{d}\nu \mathrm{d}\mu\,.$$

Since $f^{(n)}$ is bounded, we have

$$\int f^{(n)} \mathrm{d}(\mu \otimes \nu) = \int_M \int_N f^{(n)} \mathrm{d}\nu \mathrm{d}\mu$$

for all n, so it follows that $\int f \mathrm{d}(\mu \otimes \nu) = \int_M \int_N f \mathrm{d}\nu \mathrm{d}\mu$, proving (a).

Next suppose $f \in L_1(\mu \otimes \nu)$ and let

$$|f|_M(a) := \int_N |f(a, b)|\nu(\mathrm{d}b),\ f_M^{(+)}(a) := \int_N f^+(a, b)\nu(\mathrm{d}b), \text{ and } f_M^{(-)}(a) = \int_N f^-(a, b)\nu(\mathrm{d}b)\,.$$

Note that all three functions are $(\mathcal{M} \otimes \mathcal{N})/\mathcal{B}(\mathbb{R})$-measurable by Lemma 6.6; the lemma only guarantees this with $\mathcal{B}(\mathbb{R})$ replaced by $\mathcal{B}(\mathbb{R}^*)$, but the condition that $f \in L_1(\mu \otimes \nu)$ ensures that everything stays finite. Since $|f| \geq 0$, we may apply part (a) of the theorem to deduce that

$$\int_M |f|_M \mathrm{d}\mu = \int_M \int_N |f| \mathrm{d}\nu \mathrm{d}\mu = \int |f| \mathrm{d}(\mu \otimes \nu) < \infty.$$

Thus $|f|_M$ is μ-almost everywhere finite; i.e., $\mu(M \setminus F) = 0$.

Finally, $\hat{f}_M = (f_M^+ - f_M^-)\mathbf{1}_{[F]}$, at least if we are willing to accept the convention that $(\infty - \infty) \cdot 0 = 0$, and so

$$\begin{aligned}
\int f \mathrm{d}(\mu \otimes \nu) &= \int f^+ \mathrm{d}(\mu \otimes \nu) - \int f^- \mathrm{d}(\mu \otimes \nu) && \text{linearity of integration} \\
&= \int f_M^{(+)} \mathrm{d}\mu - \int f_M^{(-)} \mathrm{d}\mu && \text{by part (a)} \\
&= \int (f_M^{(+)} - f_M^{(-)}) \mathrm{d}\mu && \text{linearity of integration} \\
&= \int (f_M^{(+)} - f_M^{(-)}) \mathbf{1}_{[F]} \mathrm{d}\mu && \text{since } \mathbf{1}_{[F]} \overset{\mu\text{-a.e.}}{=} 1 \\
&= \int \hat{f}_M \mathrm{d}\mu\,,
\end{aligned}$$

proving (b).

The extension to the case that $(M, \mathcal{M}, \mu)$ and $(N, \mathcal{N}, \nu)$ are σ-finite follows by letting $(M_k, k \geq 1)$ be measurable sets in $\mathcal{M}$ with $\mu(M_k) < \infty$ and $M_k \uparrow M$, and $(N_k, k \geq 1)$ be measurable sets in $\mathcal{N}$ with $\mu(N_k) < \infty$ and $N_k \uparrow N$. The finite measure case of Fubini's theorem can be applied to the restriction $(M_k \times N_k, \mathcal{M}_k \otimes \mathcal{N}_k, \mu_k \otimes \nu_k)$, where $\mathcal{M}_k = \mathcal{M}|_{M_k}$ and $\mu_k = \mu|_{M_k}$, and N_k and ν_k are defined accordingly. The conclusions of Fubini's theorem in the σ-finite case can then be deduced by letting $k \to \infty$ and applying the monotone convergence theorem and linearity of integration. □

By an exactly parallel development to the above, we may obtain an analogue of Fubini's theorem for the product measure $\nu \otimes \mu$, where the iterated integral has $\int_M$ as the inner integral. By Exercise 6.7, it follows that (6.3) extends to the identity

$$\int f \mathrm{d}(\mu \otimes \nu) = \int_M \int_N f \mathrm{d}\nu \mathrm{d}\mu = \int \int_N \int_M f \mathrm{d}\mu \mathrm{d}\nu, .$$

Proposition 6.8. *Under the conditions of Fubini's theorem, if $f \in L_1(\mu \otimes \nu)$ then there exists $E \in \mathcal{M} \otimes \mathcal{N}$ with $\mu \times \nu(Z^c) = 0$ such that*

$$\int f \mathrm{d}(\mu \otimes \nu) = \int_M \int_N f(a,b) \mathbf{1}_{[E]}(a,b) \nu(\mathrm{d}b) \mu(\mathrm{d}a) = \int_N \int_M f(a,b) \mathbf{1}_{[E]}(a,b) \nu(\mathrm{d}a) \mu(\mathrm{d}b) .$$

Proof. Applying Fubini's theorem, we may obtain sets Z_M and Z_N as in (6.4), i.e., so that $\mu \otimes \nu(Z_M^c) = \mu \otimes \nu(Z_N^c) = 0$ and so that

$$\int f \mathrm{d}(\mu \otimes \nu) = \int_M \int_N f(a,b) \mathbf{1}_{[Z_M]}(a,b) \nu(\mathrm{d}b) \mu(\mathrm{d}a)$$

and

$$\int f \mathrm{d}(\mu \otimes \nu) = \int_N \int_M f(a,b) \mathbf{1}_{[Z_N]}(a,b) \nu(\mathrm{d}b) \mu(\mathrm{d}a) .$$

Taking $E = Z_M \cap Z_N$, the result follows. □

Corollaries added Oct 22

Corollary 6.9. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ and $(M, \mathcal{M})$, $(N, \mathcal{N})$ be measurable spaces, and let $X : \Omega \to M$ and $Y : \Omega \to N$ be independent random variables (M-valued and N-valued, respectively), with distributions μ and ν. If $h : M \times N \to \mathbb{R}$ is $(\mathcal{M} \otimes \mathcal{N}/\mathcal{B}(\mathbb{R}))$-measurable and either $h \geq 0$ or $\mathbf{E}|h(X,Y)| < \infty$, then*

$$\mathbf{E}h(X,Y) = \int_M \int_N h(x,y) \nu(\mathrm{d}y) \mu(\mathrm{d}x) .$$

Proof. For all $A \in \mathcal{M}$ and $B \in \mathcal{N}$, by independence,

$$\mathbf{P}\{(X,Y) \in A \times B\} = \mathbf{P}\{X \in A\} \mathbf{P}\{Y \in B\} = \mu(A)\nu(B).$$

Since $\mathcal{M} \times \mathcal{N}$ is a π-system generating $\mathcal{M} \otimes \mathcal{N}$, it follows that the distribution of (X,Y) is $\mu \otimes \nu$. If either $h \geq 0$ or $\mathbf{E}|h(X,Y)| < \infty$, it then follows by the change of variables formula and Fubini's theorem that

$$\mathbf{E}h(X,Y) = \int_{M \times N} h(x,y) \mathrm{d}(\mu \otimes \nu) = \int_M \int_N h(x,y) \nu(\mathrm{d}y) \mu(\mathrm{d}x) . \qquad \square$$

Corollary 6.10. *In the setting of Corollary 6.9, for any $E \in \mathcal{M} \otimes \mathcal{N}$,*

$$\mathbf{P}\{(X,Y) \in E\} = \int_M \mathbf{P}\{(x,Y) \in E\} \mu(\mathrm{d}x) .$$

Proof. Apply Corollary 6.9 to the non-negative function $h(x,y) = \mathbf{1}_{[(x,y) \in E]}$ to get

$$\mathbf{P}\{(X,Y) \in E\} = \mathbf{E}h(X,Y) = \int_M \int_N \mathbf{1}_{[(x,y) \in E]} \nu(\mathrm{d}y) \mu(\mathrm{d}x) = \int_M \mathbf{P}\{(x,y) \in E\} \mu(\mathrm{d}x) ,$$

the last equality holding by change of variables. □

Exercise 6.8. *If X and Y are independent real random variables, and X and Y have respective densities f and g with respect to Lebesgue measure on $\mathbb{R}$, then (X,Y) has density $h(x,y) = f(x)g(y)$ with respect to Lebesgue measure on $\mathbb{R}^2$.*

The final exercise of the section describes an important instance of the "independence means multiply" heuristic, and provides a natural segue to the following section, which is about sums of independent random variables. Given a random variable X with distribution $\mu_X = \mu$, the *moment generating function* of X is

Moment generating function

$$G_X(s) := \mathbf{E}\left[e^{-sX}\right] = \int_{\mathbb{R}} e^{-sx} \mu(\mathrm{d}x) \in (0, \infty] .$$

Exercise 6.9. *If X, Y are independent random variables then $G_{X+Y} = G_X G_Y$.*

7. Sums of independent random variables

Everything in this section works for random variables taking values in a separable Banach space, but we restrict to $\mathbb{R}$ for concreteness.

7.1. **Convolutions.** If μ, ν are Borel measures on $\mathbb{R}$ then the *convolution* $\mu * \nu$ is the Borel measure on $\mathbb{R}$ given by

$$\mu * \nu(B) = \int_{\mathbb{R}} \nu(B - x)\mu(\mathrm{d}x)\,,$$

for Borel B, where $B - x := \{b - x : b \in B\}$. (Exercise: to check that this definition makes sense, verify that $x \mapsto \nu(B - x)$ is a Borel function.)

Proposition 7.1. *If X, Y are independent random variables with respective laws μ and ν then $X + Y$ has law $\mu * \nu$.*

Proof. For any Borel $A \subset \mathbb{R}$, by Fubini's theorem,

$$\mathbf{P}\{X + Y \in A\} = \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbf{1}_{[x+y\in A]}\nu(\mathrm{d}y)\mu(\mathrm{d}x) = \int_{\mathbb{R}} \nu(A - x)\mu(\mathrm{d}x)\,. \qquad \square$$

If $f, g : \mathbb{R} \to [0, \infty)$ are Borel functions then we likewise define the convolution of f and g as

$$f * g(x) = \int_{\mathbb{R}} f(x - y)g(y)\mathrm{d}y.$$

The next exercise states that the connection between convolution and sums of independent random variables also holds at the level of densities.

Exercise 7.1. *If X and Y are independent real random variables, and X and Y have respective densities f and g with respect to Lebesgue measure on $\mathbb{R}$, then $X + Y$ has density $f * g$ with respect to Lebesgue measure on $\mathbb{R}$.*

Exercise 7.2. *Let μ, ν be Borel measures on $\mathbb{R}$ and let $f, g : \mathbb{R} \to [0, \infty)$ be Borel functions. Prove that $\mu * \nu = \nu * \mu$ and that $f * g = g * f$.*

It's worth seeing an example. For $\alpha, \gamma > 0$, the *Gamma*(α, λ) *density* is

$$\gamma(x) = \gamma_{\alpha,\lambda}(x) := \frac{\lambda^\alpha x^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)}\mathbf{1}_{[x\geq 0]}.$$

Here $\Gamma(\alpha) := \int_{[0,\infty]} x^{\alpha-1}e^{-x}\mathrm{d}x$ is the Gamma function. A real random variable X is *Gamma*(α, λ)*-distributed* if it has density $\gamma_{\alpha,\lambda}$ with respect to Lebesgue measure. The next exercise describes a scaling property of Gamma random variables in the second coordinate.

Exercise 7.3. *If X is Gamma(α, λ)-distributed then λX is Gamma$(\alpha, 1)$-distributed.*

Suppose X and Y are independent, X is Gamma(α, λ)-distributed and Y is Gamma(β, λ)-distributed. We claim that $Z = X + Y$ is Gamma$(\alpha + \beta, \lambda)$-distributed.

To see this, first note that by Exercise 7.3 we may assume $\lambda = 1$. (I.e. it suffices to show that $\lambda X + \lambda Y$ is Gamma$(\alpha+\beta, 1)$-distributed.) We restrict to this case, and then note that by the above exercise, the density of Z with respect to Lebesgue measure is

$$\begin{aligned} f_Z(x) &= \int_{[0,x]} \gamma_{\alpha,1}(y)\gamma_{\beta,1}(x - y)\mathrm{d}y \\ &= \int_{[0,x]} \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)}\frac{(x - y)^{\beta-1}e^{-(x-y)}}{\Gamma(\beta)}\mathrm{d}y \\ &= \frac{e^{-x}}{\Gamma(\alpha)\Gamma(\beta)}\int_0^x y^{\alpha-1}(x - y)^{\beta-1}\mathrm{d}y\,. \end{aligned}$$

Making the change of variables $u = y/x$, this becomes

$$f_Z(x) = \frac{x^{\alpha+\beta-1}e^{-x}}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 u^{\alpha-1}(1 - u)^{\beta-1}\mathrm{d}u.$$

Since $\int_{[0,\infty]} f_Z(x)\mathrm{d}x = \mathbf{P}\{Z \geq 0\} = 1$ and, by definition, $\int_{[0,\infty]} x^{\alpha+\beta-1}e^{-x} = \Gamma(\alpha+\beta)$, it follows that

$$1 = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\mathrm{d}u,$$

which combined with the preceding display gives that

$$f_Z(x) = \frac{x^{\alpha+\beta-1}e^{-x}}{\Gamma(\alpha+\beta)},$$

so Z is indeed Gamma$(\alpha+\beta, 1)$-distributed.

Another important example is introduce din the next exercise. For $\alpha \in \mathbb{R}$ and $\sigma > 0$, the $N(\alpha, \sigma^2)$ density is given by

$$\varphi_{\alpha,\sigma^2}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\alpha)^2/(2\sigma^2)}.$$

Exercise 7.4. *(a) Use change of variables and Fubini's theorem to prove that* $(\int_{\mathbb{R}} e^{-x^2}\mathrm{d}x)^2 = \pi$. *(You've perhaps seen this before and know how the proof goes. If not: look for an integral over* $\mathbb{R}^2$*, and consider a switch to polar coordinates.)*

(b) Show that if X *and* Y *are independent normals with densities* $\varphi_{\alpha,\sigma^2}(x)$ *and* φ_{β,τ^2} *respectively, then* $X+Y$ *has density* $\varphi_{\alpha+\beta,\sigma^2+\tau^2}$*; in particular* $X+Y$ *is again a normal random variable.*

8. **Laws of large numbers**

In the previous section we saw that summing independent random variables corresponds to convolution of their distributions. What happens if there are large number of summands? If $X_1, \ldots, X_n$ are independent random variables with a common distribution μ, then by Proposition 7.1, their sum $S_n := X_1 + \ldots + X_n$ has distribution μ^{*n}, the n-fold convolution of μ with itself. *Laws of large numbers* describe the first-order asymptotic behaviour of S_n (or equivalently of μ^{*n}) when $n \to \infty$.

Rather than jumping straight to the most general results, we start with a result that has an easy proof, and has the advantage of introducing one of the most important techniques for controlling the behaviour of random variables, namely *moment methods*. These are essentially all variants of the following simple inequality

Proposition 8.1 (Markov's inequality)**.** *If* X *is a non-negative random variable then* $\mathbf{P}\{X \geq t\} \leq \mathbf{E}X/t$ *for all* $t > 0$*.*

Proof. Since $X \geq X\mathbf{1}_{[X\geq t]}$, by monotonicity and by linearity of expectation,

$$\mathbf{E}X \geq \mathbf{E}\left[X\mathbf{1}_{[X\geq t]}\right] \geq \mathbf{E}\left[t\mathbf{1}_{[X\geq t]}\right] = t\mathbf{P}\{X \geq t\}. \qquad \square$$

Here are some important special cases. For a random variable X with $\mathbf{E}|X| < \infty$, we write $\mathbf{Var}(X) := \mathbf{E}\left[(X - \mathbf{E}X)^2\right] \in [0, \infty]$; the quantity $\mathbf{Var}(X)$ is called the *variance* of X.

Variance of a random variable

Corollary 8.2 (Chebyshev's inequality)**.** *For any random variable* X *with* $\mathbf{E}|X| < \infty$*, for all* $t > 0$*,*

$$\mathbf{P}\{|X - \mathbf{E}X| \geq t\} \leq \frac{\mathbf{Var}(X)}{t^2}.$$

Proof. Note that $|X - \mathbf{E}X| \geq t$ if and only if $(X - \mathbf{E}X)^2 \geq t^2$; then apply Markov's inequality. $\square$

Corollary 8.3 (Chernoff bound)**.** *For any random variable* X*, for all* $t \in \mathbb{R}$ *,*

$$\mathbf{P}\{X \geq t\} \leq \inf_{a>0} \frac{\mathbf{E}\left[e^{aX}\right]}{e^{at}}.$$

Proof. Fix $c > 0$. Then by Markov's inequality,

$$\mathbf{P}\{X \geq t\} = \mathbf{P}\left\{e^{aX} \geq e^{at}\right\} \leq \frac{\mathbf{E}\left[e^{aX}\right]}{e^{at}}.$$

Since this bound holds for each $a > 0$, the result follows. □

In general, if X is a random variable taking values in a (possibly unbounded) interval $I \subseteq \mathbb{R}$ and $\phi : I \to [0, \infty$ is strictly increasing, then for any we may use Markov's inequality to obtain that for any $t \in I$,

$$\mathbf{P}\{X \geq t\} = \mathbf{P}\{\phi(X) \geq \phi(t)\} \leq \frac{\mathbf{E}\phi(X)}{\phi(t)};$$

both Chebyshev's inequality and the Chernoff bound are special cases of this general bound.

We next use Chebyshev's inequality and the Chernoff bound to control the deviations of sums of independent random variables from their expected values. Before giving the details, we make a few simple observations.

Let X be a random variable with and let $0 \leq q \leq p$. Then

$$\mathbf{E}\left[|X|^q\right] \leq \mathbf{E}\left[\max(1, |X|^q)\right] \leq \mathbf{E}\left[\max(1, |X|^p)\right] \leq \mathbf{E}\left[1 + |X|^p\right], \tag{8.1}$$

so if $\mathbf{E}\left[|X|^p\right] < \infty$ then $\mathbf{E}\left[|X|^q\right] < \infty$. In particular, if $\mathbf{E}\left[X^2\right] < \infty$ then $X \in L_1(\mathbf{P})$ and so by linearity of expectation,

$$\mathbf{Var}(X) = \mathbf{E}\left[(X - \mathbf{E}X)^2\right] = \mathbf{E}\left[X^2 - 2X\mathbf{E}X + (\mathbf{E}X)^2\right] = \mathbf{E}\left[X^2\right] - (\mathbf{E}X)^2 \leq \mathbf{E}\left[X^2\right]. \tag{8.2}$$

Also, if a random variable X almost surely satisfies $a \leq X \leq b$ then we always have $|X - \mathbf{E}X| \leq b - a$, and so

$$\mathbf{Var}(X) = \mathbf{E}\left[(X - \mathbf{E}X)^2\right] \leq |b - a|^2.$$

Exercise 8.1. *Strengthen the above bound to* $|b - a|^2/4$.

Example: Gaussian tails for sums of bounded random variables. Fix $C > 1$ and let $(X_i, i \geq 1)$ be independent random variables with $|X_i| \leq C$ for all i. As before, write $x_i = \mathbf{E}X_i$, let $S_n := X_1 + \ldots + X_n$ and let $s_n = \mathbf{E}S_n = \sum_{i=1}^n x_i$. Then by the Chernoff bound,

$$\begin{aligned}
\mathbf{P}\{|S_n - s_n| \geq t\} &\leq \inf_{a>0} e^{-at}\mathbf{E}\left[e^{a(S_n - s_n)}\right] \\
&= \inf_{a>0} e^{-at}\mathbf{E}\left[\prod_{i=1}^n e^{a(X_i - x_i)}\right] \\
&= \inf_{a>0} e^{-at}\prod_{i=1}^n \mathbf{E}\left[e^{a(X_i - x_i)}\right],
\end{aligned}$$

where we have used the factorization formula in the last step. We now use that if $|y| \leq 1$ then $|e^y - 1 - y| \leq y^2$. Since $|X_i| \leq C$, necessarily $|X_i - x_i| \leq 2C$, so if $a \leq 1/(2C)$ then

$$e^{a(X_i - x_i)} \leq 1 + a(X_i - x_i) + a^2(X_i - x_i)^2.$$

Taking $t = xn^{1/2}$ and $a = x/(2C^2n^{1/2})$, we then obtain

$$\mathbf{P}\left\{|S_n - s_n| \geq xn^{1/2}\right\} \leq e^{-x^2/2C^2}\prod_{i=1}^n \left(\mathbf{E}\left[1 + a(X_i - x_i) + a^2(X_i - x_i)^2\right]\right)$$

For each $i \in [n]$, by linearity of expectation and since $\mathbf{E}\left[X_i - x_i\right] = 0$ and $\mathbf{Var}(X_i) \leq C^2$,

$$\mathbf{E}\left[1 + a(X_i - x_i) + a^2(X_i - x_i)^2\right] = 1 + a^2\mathbf{E}\left[(X_i - x_i)^2\right] \leq 1 + a^2C^2 = 1 + \frac{x^2}{4C^2n}.$$

Combining this with the preceding bound gives

$$\begin{aligned}\mathbf{P}\left\{|S_n - s_n| \geq xn^{1/2}\right\} &\leq e^{-x^2/2C^2}\left(1 + \frac{x^2}{4C^2 n}\right)^n \\ &\leq e^{-x^2/2C^2} e^{x^2/4C^2} \\ &= e^{-x^2/4C^2},\end{aligned}$$

where in the second inequality we used that $1 + x \leq e^x$.

The next example introduces the notation of *covariance* of random variables. If X, Y are random variables with $\mathbf{E}|X|, \mathbf{E}|Y|, \mathbf{E}|XY| < \infty$, the covariance of X and Y is defined to be $\mathrm{Cov}(X, Y) = \mathbf{E}\left[(X - \mathbf{E}X)(Y - \mathbf{E}Y)\right]$. If $\mathrm{Cov}(X, Y)$ is defined and equals zero, then X and Y are said to be *uncorrelated*. Chebyshev's inequality gives clean bounds for sums of uncorrelated random variables, that are useful frequently enough to be stated as a separate corollary.

Covariance, uncorrelated random variables

Corollary 8.4 (Chebyshev's inequality for sums)**.** *Let $(X_i, i \geq 1)$ be uncorrelated random variables with $\mathbf{E}|X_i| < \infty$ for all $i \geq 1$. Let $S_n := X_1 + \ldots + X_n$ and let $s_n = \mathbf{E}S_n$. Then for all $t > 0$,*

$$\mathbf{P}\left\{|S_n - s_n| \geq t\right\} \leq \frac{\sum_{i=1}^n \mathbf{Var}\left(X_i\right)}{t^2}.$$

Proof. Write $x_i = \mathbf{E}X_i$, so $s_n = x_1 + \ldots + x_n$. Then

$$\begin{aligned}\mathbf{Var}\left(S_n\right) &= \mathbf{E}\left[(S_n - s_n)^2\right] \\ &= \mathbf{E}\left[\Big((X_1 - x_1) + \ldots + (X_n - x_n)\Big)^2\right] \\ &= \sum_{i=1}^n \mathbf{E}\left[(X_i - x_i)^2\right] + \sum_{1 \leq i \neq j \leq n} \mathbf{E}\left[(X_i - x_i)(X_j - x_j)\right].\end{aligned}$$

By independence, for $i \neq j$ we have $\mathbf{E}\left[(X_i - x_i)(X_j - x_j)\right] = \mathbf{E}\left[X_i - x_i\right]\mathbf{E}\left[X_j - x_j\right] = 0$, so the second sum vanishes. The first sum is simply $\sum_{i=1}^n \mathbf{Var}\left(X_i\right)$, so the result follows by Chebyshev's inequality. □

Example: weak law of large numbers for uncorrelated random variables. Using Chebyshev's inequality for sums, we can easily prove a first law of large numbers.

Theorem 8.5 (Weak law of large numbers for sums of uncorrelated random variables with bounded variance)**.** *Let $(X_i, i \geq 1)$ be independent random variables with $\sup_{i \geq 1} \mathbf{E}\left[X_i^2\right] = C < \infty$. Write $x_i = \mathbf{E}X_i$, let $S_n := X_1 + \ldots + X_n$ and let $s_n = \mathbf{E}S_n$. Then*

$$\frac{S_n - s_n}{n} \to 0 \tag{8.3}$$

in probability.

Proof. By Chebyshev's inequality for sums we have

$$\mathbf{P}\left\{|S_n - s_n| > t\right\} \leq \frac{1}{t^2}\sum_{i=1}^n \mathbf{Var}\left(X_i\right) \leq \frac{Cn}{t^2}.$$

In the last line we have used that $\mathbf{Var}\left(X_i\right) \leq \mathbf{E}\left[X_i^2\right] \leq C$ for each $1 \leq i \leq n$. For any $\epsilon > 0$, taking $t = \epsilon n$ above gives

$$\mathbf{P}\left\{\left|\frac{S_n - s_n}{n}\right| > \epsilon\right\} = \mathbf{P}\left\{|S_n - s_n| > \epsilon n\right\} \leq \frac{\mathbf{Var}\left(S_n\right)}{\epsilon^2 n^2} \leq \frac{C}{\epsilon^2 n} \to 0$$

as $n \to \infty$; thus $(S_n - s_n)/n \to 0$ in probability as claimed. □

Remarks.

- We have just proved a weak law of large numbers for independent random variables with bounded second moments. We'll next see how to combine this with *truncation* and Markov's inequality to prove that $S_n/s_n \to 1$ under only a first-moment assumption, but additionally assuming that the random variables are identically distributed.
- If the random variables $(X_i, i \geq 1)$ are also identically distributed, then $s_n = n\mathbf{E}X_1$, in which case (8.3) asserts that $S_n/n \to \mathbf{E}X_1$ in probability; this is a more classical way to state a law of large numbers.

Exercise 8.2. *Modify the above proof to show that, under the same assumptions, if $f : \mathbb{N} \to [0, \infty)$ and $f(n) \to \infty$ as $n \to \infty$ then*

$$\frac{S_n - s_n}{f(n)n^{1/2}} \to 0$$

in probability, as $n \to \infty$.

We now use the same idea for random variables with possibly infinite variance (but additionally assuming the random variables are identically distributed). We obviously can't directly use the same proof in this case; we will instead argue by *truncation*.

Theorem 8.6. *Let $(X_n, n \geq 1)$ be independent identically distributed random variables with $\mathbf{E}|X_n| < \infty$, and write $S_n = X_1 + \ldots + X_n$. Then for all $\epsilon > 0$,*

$$\mathbf{P}\left\{\left|\frac{S_n}{n} - \mathbf{E}X_1\right| \geq \epsilon\right\} \to 0\,,$$

as $n \to \infty$.

Proof. For fixed $N > 0$, we define $X_k^{\leq N}$ and $X_k^{>N}$ as follows: $X_k^{\leq N} = X_k \mathbf{1}_{|X_k| \leq N}$ and $X_k^{>N} = X_k - X_k^{\leq N}$.

We then have that $|X_1^{\leq N}|$ increases to $|X_1|$ as $N \to \infty$, so by monotone convergence

$$\mathbf{E}|X_1^{\leq N}| \to \mathbf{E}|X_1|,$$

again as $N \to \infty$. Since $|X_1| = |X_1^{\leq N}| + |X_1^{>N}|$ (check if it isn't obvious to you), it follows that as $N \to \infty$ we also have

$$\mathbf{E}|X_1^{>N}| = \mathbf{E}|X_1| - \mathbf{E}|X_1^{\leq N}| \to 0.$$

Now fix $\epsilon > 0$, and let N be large enough that $\mathbf{E}|X_1^{>N}| < \epsilon^2/8$. By Chebyshev's inequality for sums, we then have

$$\mathbf{P}\left\{|\bar{S}_n^{\leq N} - \mathbf{E}\bar{S}_n^{\leq N}| > \epsilon/2\right\} \leq \frac{1}{(\epsilon/2)^2 n}\mathbf{Var}(X_1^{\leq N}) \leq \frac{4N^2}{\epsilon^2 n},$$

the last inequality since $-N \leq X_1^{\leq N} \leq N$ so $\mathbf{Var}\{X_1^{\leq N}\} \leq (2N)^2/4 = N^2$. The last expression is less than $\epsilon/2$ for $n > 8N^2/\epsilon^3$. We then have

$$\begin{aligned}
\mathbf{P}\left\{|\bar{S}_n^{>N} - \mathbf{E}\bar{S}_n^{>N}| > \epsilon/2\right\} &\leq \frac{\mathbf{E}\left[|\bar{S}_n^{>N} - \mathbf{E}\bar{S}_n^{>N}|\right]}{(\epsilon/2)} && \text{(Markov's inequality)}\\
&\leq \frac{\mathbf{E}\left[|\bar{S}_n^{>N}|\right] + |\mathbf{E}\bar{S}_n^{>N}|}{(\epsilon/2)} && \text{(Triangle inequality)}\\
&\leq \frac{4\mathbf{E}\left[|\bar{S}_n^{>N}|\right]}{\epsilon} && \text{(Move absolute value inside expectation)}\\
&\leq \frac{\epsilon}{2} && \text{(Since } \mathbf{E}|\bar{S}_n^{>N}| \leq \mathbf{E}|X_1^{>N}| < \epsilon^2/8\text{).}
\end{aligned}$$

It follows that for $n > 8N^2/\epsilon^3$,

$$\mathbf{P}\left\{|\bar{S}_n - \mathbf{E}\left[X_1\right]| > \epsilon\right\} \leq \mathbf{P}\left\{|\bar{S}_n^{\leq N} - \mathbf{E}\bar{S}_n^{\leq N}| > \epsilon/2\right\} + \mathbf{P}\left\{|\bar{S}_n^{>N} - \mathbf{E}\bar{S}_n^{>N}| > \epsilon/2\right\} < 2\epsilon.$$

Since $\epsilon > 0$ was arbitrary this shows convergence in probability. □

This argument was straightforward enough that it's worth seeing if we can squeeze a little more out of it. Our goal is to end up proving a *strong* law of large numbers; we want to prove that

$$\frac{S_n}{n} \stackrel{\text{a.s.}}{\to} \mathbf{E}X_1,$$

strengthening the convergence in probability shown above. How might we naturally proceed?

Well, we did see one way to deduce almost sure convergence from convergence in probability: Proposition 3.8, which states that if $(Z_n, 1 \le n \le \infty)$ are a sequence of random variables with $Z_n \stackrel{\text{p}}{\to} Z_\infty$, then there exists a subsequence $(n_k, k \ge 1)$ such that $Z_{n_k} \stackrel{\text{a.s.}}{\to} Z_\infty$ as $k \to \infty$. It is reasonable to ask how "dense" a subsequence we can pick, without working too hard, and obtain a subsequential strong law of large numbers using nothing more than the bounds we proved in the course of proving the weak law.

We say a sequence $(n_k, k \ge 1)$ is *lacunary* if it is increasing and there exists $c > 1$ such that $n_{k+1} > cn_k$ for all k sufficiently large. We will prove the following theorem.

Theorem 8.7 (Lacunary Strong Law of Large Numbers)**.** *Let $(X_n, n \ge 1)$ be independent identically distributed random variables with $\mathbf{E}|X_n| < \infty$, and write $S_n = X_1 + \ldots + X_n$. Then for any lacunary sequence of positive integers $(n_k, k \ge 1)$,*

$$\mathbb{P}\left(\lim_{k\to\infty} \frac{S_{n_k}}{n_k} = \mathbf{E}X_1\right) = 1.$$

Proof. Let $(n_k, k \ge 1)$ be a lacunary sequence, and let $c > 1$ be such that $n_{k+1} \ge cn_k$ for $k \ge k_0$. As before, for any $\epsilon > 0$, $N > 0$ and $n \ge 1$ we have

$$\mathbf{P}\left\{|\bar{S}_n^{\le N} - \mathbf{E}\bar{S}_n^{\le N}| > \epsilon/2\right\} \le \frac{1}{(\epsilon/2)^2 n}\mathbf{E}\left[(X_1^{\le N})^2\right] \le \frac{4N^2}{\epsilon^2 n}.$$

We could have used $\mathbf{Var}\left(X_1^{\le N}\right)$ rather than $\mathbf{E}\left[(X_1^{\le N})^2\right]$ above, and obtained a better upper bound; but using $\mathbf{E}\left[(X_1^{\le N})^2\right]$ will make things a little cleaner later. From the preceding bound it follows that

$$\begin{aligned}
\sum_{k\ge 1} \mathbf{P}\left\{|\bar{S}_{n_k}^{\le N} - \mathbf{E}\bar{S}_{n_k}^{\le N}| > \epsilon/2\right\} &\le k_0 + \sum_{k>k_0} \mathbf{P}\left\{|\bar{S}_{n_k}^{\le N} - \mathbf{E}\bar{S}_{n_k}^{\le N}| > \epsilon/2\right\} \\
&\le k_0 + \sum_{k>k_0} \frac{\mathbf{E}\left[(X_1^{\le N})^2\right]}{(\epsilon/2)^2 n_k} \\
&\le k_0 + \sum_{k>k_0} \frac{4N^2}{\epsilon^2 c^{k-k_0} n_{k_0}} \\
&< \infty, \qquad (8.4)
\end{aligned}$$

Say something about the fact that IIDness is key here?

the last bound holding since the summands of the final sum are geometrically decreasing. It follows by the first Borel-Cantelli lemma that with probability 1, for all k sufficiently large, $|\bar{S}_n^{\le N} - \mathbf{E}\bar{S}_n^{\le N}| \le \epsilon/2$.

That worked out well. But the situation isn't so good when we turn to the unbounded summands. There we have the bound

$$\mathbf{P}\left\{|\bar{S}_n^{>N} - \mathbf{E}\bar{S}_n^{>N}| > \epsilon/2\right\} \le \frac{4\mathbf{E}|\bar{S}_n^{>N}|}{\epsilon}.$$

We can make $\mathbf{E}\left[|\bar{S}_n^{>N}|\right]$ as small as we like by choosing N large, but there are infinitely many summands, so a fixed positive bound on $\mathbf{E}\left[|\bar{S}_n^{>N}|\right]$ isn't good enough to let us apply the Borel-Cantelli lemma. Can we find a more explicit/more useful bound? Well, the triangle inequality is a reasonable attack:

$$|\bar{S}_n^{>N}| = \frac{1}{n}|X_1^{>N} + \ldots + X_n^{>N}| \le \frac{1}{n}|X_1^{>N}| + \ldots + |X_n^{>N}|,$$

and the summands on the right are IID, so this gives

$$\mathbf{P}\left\{|\bar{S}_n^{>N} - \mathbf{E}\bar{S}_n^{>N}| > \epsilon/2\right\} \leq \frac{4\mathbf{E}|X_1^{>N}|}{\epsilon}\,. \tag{8.5}$$

We can make $\mathbf{E}|X_1^{>N}|$ small by taking N large. But we can't make it zero, so this doesn't give a finite bound on summation.

Are we stuck? Well, if a fixed positive bound isn't good enough, maybe we can let N vary. Let's look back at the control we achieved for the bounded summands. If instead of picking a single value N we choose a sequence of different values $(N_k, k \geq 1)$, then we obtain

$$\sum_{k\geq 1}\mathbf{P}\left\{|\bar{S}_{n_k}^{\leq N_k} - \mathbf{E}\bar{S}_{n_k}^{\leq N_k}| > \epsilon/2\right\} \leq k_0 + \sum_{k>k_0}\frac{\mathbf{E}\left[(X_1^{\leq N_k})^2\right]}{(\epsilon/2)^2 n_k}\,.$$

If we can show that the last sum is finite, then by the first Borel-Cantelli lemma we again obtain that almost surely $|\bar{S}_{n_k}^{\leq N_k} - \mathbf{E}\bar{S}_{n_k}^{\leq N_k}| \leq \epsilon/2$ except for finitely many values of k. We'd like to make this argument work for a sequence $(N_k, k \geq 1)$ growing as quickly as possible, since the larger the values N_k the easier our work will be when we turn to controlling the unbounded summands.

At this point the first natural thing to do is to use the bound $\mathbf{E}\left[(X_1^{\leq N_k})^2\right] \leq N_k^2$. If we do that then we will end up with a finite bound provided we choose N_k such that (N_k^2/n_k) is summable. For example, taking $N_k = n_k^{1/4}$ would yield the bound

$$k_0 + \sum_{k>k_0}\frac{1}{(\epsilon^2/2)^2 n_k^{1/2}} \leq k_0 + \sum_{k>k_0}\frac{4}{\epsilon^2 c^{(k-k_0)/2} n_{k_0}^{1/2}} < \infty\,.$$

This already looks promising. But we can squeeze out a slightly stronger result, and simultaneously simplify our notation, by explicitly considering which values of k have $X_1^{\leq N_k} \neq 0$. That is, let $J = J(\omega) = \min\{k : N_k \geq |X_1(\omega)|\}$. Then $X_1^{\leq N_k} = 0$ for $k < J$, so

$$\sum_{k=k_0}^{\infty}\frac{\mathbf{E}\left[(X_1^{\leq N_k})^2\right]}{n_k} = \mathbf{E}\left[\sum_{k=k_0}^{\infty}\frac{X_1^2\mathbf{1}_{|X_1|\leq N_k}}{n_k}\right] = \mathbf{E}\left[\sum_{k=\max(k_0,J)}^{\infty}\frac{X_1^2}{n_k}\right] \leq \frac{c}{c-1}\mathbf{E}\left[\frac{X_1^2}{n_{\max(k_0,J)}}\right]\,.$$

In the last bound we used that $\frac{n_{k+1}}{n_k} \geq c$ for $k > k_0$, so $\sum_{k=\max(k_0,J)}^{\infty} n_k^{-1} \leq n_{\max(k_0,J)}\cdot\sum_{i\geq 0} c^{-1}$.

How can we make this bound finite? Well, if $N_k \leq n_k$ then by the definition of J we have $n_{\max(k_0,J)} \geq N_{\max(k_0,J)} \geq |X_1|$, so $\mathbf{E}\left[X_1^2/n_{\max(k_0,J)}\right] \leq \mathbf{E}|X_1| < \infty$. We want to choose N_k to be as large as possible, since this should make our lives easier when it comes to controlling the unbounded summands; so let's take $N_k = n_k$ henceforth.[11] Summarizing the story to date, we now have that

$$\sum_{k\geq 1}\mathbf{P}\left\{|\bar{S}_{n_k}^{\leq n_k} - \mathbf{E}\bar{S}_{n_k}^{\leq n_k}| > \epsilon/2\right\} \leq k_0 + \sum_{k>k_0}\frac{\mathbf{E}\left[(X_1^{\leq n_k})^2\right]}{(\epsilon/2)^2 n_k} \leq k_0 + \frac{4c}{\epsilon^2(c-1)}\mathbf{E}|X_1| < \infty\,,$$

so by the first Borel-Cantelli lemma,

$$\mathbf{P}\left\{|\bar{S}_{n_k}^{\leq n_k} - \mathbf{E}\bar{S}_{n_k}^{\leq n_k}| > \epsilon/2 \text{ i.o.}\right\} = 0\,. \tag{8.6}$$

We are in good shape for the sums of the bounded parts. For the unbounded summands, from (8.5) we have

$$\mathbf{P}\left\{|\bar{S}_{n_k}^{>n_k} - \mathbf{E}\bar{S}_{n_k}^{>n_k}| > \epsilon/2\right\} \leq \frac{4\mathbf{E}\left[|X_1^{>n_k}|\right]}{\epsilon}\,.$$

What happens if we sum the right-hand side over $k \geq k_0$?

[11] To make N_k even bigger we could take $N_k = An_k$ for some constant $A > 1$, but given that our proof has to work for all lacunary sequences, it's not hard to see that such a change would not make any difference to the success or failure of our argument.

Well, bad news, friends: the sum may be infinite. For example, it could be that $n_k = 2^k$ (in which case $k_0 = 1$). By linearity of expectation, we would then have

$$\sum_{k \geq k_0} \mathbf{E}\left[|X_1^{>n_k}|\right] = \mathbf{E}\left[\sum_{k \geq 1} |X_1| \mathbf{1}_{[|X_1| > 2^k]}\right] = \mathbf{E}\left[|X_1| \lfloor \log_2 \max(1, |X_1| - 1) \rfloor\right]$$
$$\leq \mathbf{E}\left[|X_1| \log_2(X_1 + 1)\right] .$$

We assumed $\mathbf{E}|X_1| < \infty$, but $\mathbf{E}\left[|X_1| \log(|X_1| + 1)\right]$ need not be, so this bound may be useless. On the other hand, it's worth recording now that if n_k is any lacunary sequence then similar logic would yield the bound

$$\sum_{k \geq k_0} \mathbf{E}\left[|X_1^{>n_k}|\right] = O(\mathbf{E}\left[|X_1| \log(|X_1| + 1)\right]) ,$$

so if $\mathbf{E}\left[|X_1| \log(|X_1| + 1)\right] < \infty$ then we can actually finish the proof along these lines.

At this point the situation may seem bleak. We are stuck trying to bound

$$\mathbf{P}\left\{|\bar{S}_{n_k}^{>n_k} - \mathbf{E}\bar{S}_{n_k}^{>n_k}| > \epsilon/2\right\} ,$$

and our tricks have all failed. But our sleeves are not yet empty. We will go back to the very basics and try to exploit subadditivity of probabilities. By this I mean the following. Since

$$\bar{S}_{n_k}^{>n_k} = \frac{1}{n_k} S_{n_k}^{>n_k} = \frac{1}{n} \sum_{i=1}^{n_k} X_{n_k}^{>n_k} ,$$

by the triangle inequality,

$$|\bar{S}_{n_k}^{>n_k} - \mathbf{E}\bar{S}_{n_k}^{>n_k}| \leq \frac{1}{n_k} \sum_{i=1}^{n} |X_{n_k}^{>n_k} - \mathbf{E}X_{n_k}^{>n_k}| ,$$

so if $|\bar{S}_{n_k}^{>n_k} - \mathbf{E}\bar{S}_{n_k}^{>n_k}| > \epsilon/2$ then there must be $1 \leq i \leq n_k$ such that $|X_{n_k}^{>n_k} - \mathbf{E}X_{n_k}^{>n_k}| > \epsilon/2$. We thus have

$$\mathbf{P}\left\{|\bar{S}_{n_k}^{>n_k} - \mathbf{E}\bar{S}_{n_k}^{>n_k}| > \epsilon/2\right\} \leq \mathbf{P}\left\{\exists i \in [n_k] : |X_{n_k}^{>n_k} - \mathbf{E}X_{n_k}^{>n_k}| > \epsilon/2\right\}$$
$$\leq n_k \mathbf{P}\left\{|X_1^{>n_k} - \mathbf{E}X_1^{>n_k}| > \epsilon/2\right\} .$$

But $n_k \to \infty$ as $k \to \infty$, so $\mathbf{E}X_1^{>n_k} \to 0$, and so there is $k_1 \geq k_0$ such that $|\mathbf{E}X_1^{>n_k}| < \epsilon/2$ for $k \geq k_1$. For such k, if $|X_1^{>n_k} - \mathbf{E}X_1^{>n_k}| > \epsilon/2$ then in particular $X_1^{>n_k} \neq 0$, in which case necessarily $|X_1| > n_k$. Using this observation to bound the final probability above, we obtain

$$\mathbf{P}\left\{|\bar{S}_{n_k}^{>n_k} - \mathbf{E}\bar{S}_{n_k}^{>n_k}| > \epsilon/2\right\} \leq \mathbf{P}\left\{\exists i \in [n_k] : |X_{n_k}^{>n_k} - \mathbf{E}X_{n_k}^{>n_k}| > \epsilon/2\right\}$$
$$\leq n_k \mathbf{P}\left\{|X_1| > n_k\right\} .$$

This is excellent news. Because $(n_k, k \geq 1)$ is a lacunary sequence, this expectation is actually finite! To see this, recall that $J = \min\{k : n_k \geq |X_1|\}$; then $|X_1| > n_k$ only for $k < J$, so

$$\begin{aligned}
\sum_{k=k_1}^{\infty} n_k \mathbf{P}\left\{|X_1| > n_k\right\} &= \mathbf{E}\left[\sum_{k=k_1}^{J-1} n_k \mathbf{1}_{X_1 > n_k}\right] \\
&= \mathbf{E}\left[\sum_{k=k_1}^{J-1} n_k\right] && (\mathbf{1}_{X_1 > n_k} = 0 \text{ for } i \geq J) \\
&\leq \mathbf{E}\left[\sum_{k=k_1}^{J-1} c^{-(J-1-k)} X_1\right] && \text{(lacunarity)} \\
&\leq \mathbf{E}\left[\sum_{k=0}^{\infty} c^{-k} X_1\right] \\
&= \frac{c}{c-1}\mathbf{E}\left[X_1\right] < \infty.
\end{aligned}$$

Thus $\sum_{k\geq 1} \mathbf{P}\left\{|\bar{S}^{>n_k}_{n_k} - \mathbf{E}\bar{S}^{>n_k}_{n_k}| > \epsilon/2\right\} < \infty$, so again by the first Borel-Cantelli lemma,

$$\mathbf{P}\left\{|\bar{S}^{>n_k}_{n_k} - \mathbf{E}\bar{S}^{>n_k}_{n_k}| > \epsilon/2 \text{ i.o.}\right\} = 0.$$

But if $|\bar{S}_{n_k} - \mathbf{E}\bar{S}_{n_k}| > \epsilon$ infinitely often then either the bounded or unbounded partial sums must differ from their mean by at least $\epsilon/2$ infinitely often; so the theorem follows from the preceding equality and (8.6) □

Let's summarize the situation. We proved the weak law of large numbers, stating conditions which guarantee that $S_n/n \to \mathbf{E}\left[X_1\right]$ in probability. Under these conditions, Proposition 3.8 then guarantees the existence of subsequences $(n_k, k \geq 1)$ along which $S_n/n \overset{\text{a.s.}}{\to} \mathbf{E}\left[X_1\right]$. The lacunary law of large numbers gave a quantitative strengthening of Proposition 3.8 in this setting, by showing that $(n_k, k \geq 1)$ can be taken to be any lacunary sequence.

This quantitative bound was not trivial to prove, but it was worth the effort, as the general strong law of large numbers ends up being a quite straightforward consequence. Its proof will proceed by first reducing to the case that the summands are non-negative, then using the monotonicity of the partial sums to relate convergence along lacunary subsequences to convergence of the whole sequence. For the second step, the key analytic fact is described in the following exercise.

Exercise 8.3. *Let $(s_n, n \geq 0)$ be a non-decreasing sequence with $s_0 = 0$. Fix $\mu > 0$, $\epsilon \in (0, 1/3)$, and define a sequence by $n_k = \lceil (1+\epsilon)^k \rceil$.*

- *(a) Show that for all n sufficiently large (i.e. $n \geq n_0(\epsilon)$), if $s_n \geq \mu n(1+3\epsilon)$ then letting k be such that $n_{k-1} < n \leq n_k$, we have $s_{n_k} \geq \mu n_k(1+\epsilon)$.*
- *(b) Show that for all n sufficiently large, if $s_n \leq \mu n(1-3\epsilon)$ then letting k be such that $n_{k-1} < n \leq n_k$, we have $s_{n_{k-1}} \leq \mu n_{k-1}(1-\epsilon)$.*
- *(c) Conclude that if $\limsup_n |s_n - \mu n|/n > 3\epsilon\mu$ then $\limsup_k |s_{n_k} - \mu n_k|/n_k > \epsilon\mu$.*

Theorem 8.8. *Let $(X_n, n \geq 1)$ be independent identically distributed random variables with $\mathbf{E}|X_1| < \infty$. Write $S_n = X_1 + \ldots + X_n$ for $n \geq 1$. Then*

$$\frac{S_n}{n} \overset{\text{a.s.}}{\to} \mathbf{E}X_1\,.$$

Proof. Write $X_n = X_n^+ - X_n^-$ and $S_n^+ = X_1^+ + \ldots + X_n^+$ and $S_n^- = X_1^- + \ldots + X_n^-$. If $\omega \in \Omega$ is such that $S_n^+(\omega)/n \to \mathbf{E}X_1^+$ and $S_n^-(\omega)/n \to \mathbf{E}X_1^-$ then

$$\frac{S_n(\omega)}{n} = \frac{S_n^+(\omega) + S_n^-(\omega)}{n} \to \mathbf{E}X_1^+ + \mathbf{E}X_1^- = \mathbf{E}X_1.$$

So we see that to prove $S_n/n \to \mathbf{E}(X_1)$ almost surely, it suffices to prove that

$$\frac{S_n^+}{n} \xrightarrow{\text{a.s.}} \mathbf{E}(X_1^+),$$

and that $S_n^-/n \overset{\text{a.s.}}{\to} \mathbf{E}\left[X_1^-\right]$. The point of this reduction is that summands $(X_n^+, n \geq 1)$ are all non-negative, and likewise for $(X_n^-, n \geq 1)$.

So we may now assume (by replacing $(X_i, i \geq 1)$ by either $(X_i^+, i \geq 1)$ or $(X_i^-, i \geq 1)$) that $\mathbf{P}\{X_1 \geq 0\} = 1$; in this case $(S_n, n \geq 1)$ is almost surely non-decreasing. Fix $\epsilon \in (0, 1/3)$ and for $k \geq 1$ let $n_k = n_k(\epsilon) := \lceil (1+\epsilon)^k \rceil$. Then by Exercise 8.3,

$$\left\{\omega : \limsup_{n\to\infty} \left|\frac{S_n(\omega)}{n} - \mathbf{E}X_1\right| > 3\epsilon\right\} \subseteq \left\{\omega : \limsup_{k\to\infty} \left|\frac{S_{n_k}(\omega)}{n_k} - \mathbf{E}X_1\right| > \epsilon\right\}.$$

By Theorem 8.7, we have $\mathbf{P}\{\limsup_{k\to\infty} |S_{n_k}/n_k - \mathbf{E}X_1| > \epsilon\} = 0$; it follows that

$$\limsup_{n\to\infty} \mathbf{P}\left\{\left|frac S_n n - \mathbf{E}X_1\right| > 3\epsilon\right\} = 0.$$

Since this holds for all $\epsilon > 0$, it follows that

$$\frac{S_n}{n} \overset{\text{a.s.}}{\to} \mathbf{E}X_1\,. \qquad \square.$$

9. **Convexity, inequalities, and L_p spaces**

We begin with convexity. A function $f : \mathbb{R} \to \mathbb{R}$ is convex if $f(px + (1-p)y) \leq pf(x) + (1-p)f(y)$ for all $x, y \in \mathbb{R}$ and $p \in [0,1]$.

Exercise 9.1. *If $f : \mathbb{R} \to \mathbb{R}$ is convex then it is continuous, so Borel measurable.*

Theorem 9.1 (Jensen's inequality)**.** *If X is a real random variable with $\mathbf{E}|X| < \infty$, and $f : \mathbb{R} \to \mathbb{R}$ is convex and $\mathbf{E}f(X)$ is defined, then $f(\mathbf{E}X) \leq \mathbf{E}f(X)$.*

Proof. Fix $h > 0$ and $0 < p < 1$. For any $x \in \mathbb{R}$ we have $x + hp = (1-p)x + p(x+h)$ so

$$f(x + hp) \leq (1-p)f(x) + pf(x+h),$$

which after rearrangement gives

$$\frac{f(x+hp) - f(x)}{hp} \leq \frac{f(x+h) - f(x)}{h},$$

In other words, $(f(x+h) - f(x))/h$ is increasing in h for all $x \in \mathbb{R}$; we define

$$f'_+(x) := \lim_{h \downarrow 0} \frac{f(x+h) - f(x)}{h}\,.$$

Likewise, $(f(x) - f(x-h))/h$ is decreasing in h, so the limit

$$f'_-(x) := \lim_{h \downarrow 0} \frac{f(x) - f(x-h)}{h}\,.$$

Convexity also gives

$$f(x) - f(x-h) \leq f(x+h) - f(x),$$

from which it follows that $f'_-(x) \leq f'_+(x)$.

Now let $c := \mathbf{E}f(X)$, and fix $a \in \mathbb{R}$ with $f'_-(c) \leq a \leq f'_+(c)$. Then the line ℓ given by $\ell(x) = f(c) + a(x-c)$ has $\ell \leq f$ and $\ell(c) = f(c)$. By linearity of expectation and monotonicity,

it follows that

$$\begin{aligned} f(\mathbf{E}(X)) &= f(c) \\ &= \ell(c) \\ &= f(c) + a(\mathbf{E}X - c) \\ &= \mathbf{E}\left[f(c) + a(X-c)\right] \\ &= \mathbf{E}\ell(X) \\ &\leq \mathbf{E}f(X)\,, \end{aligned}$$

as required. □

For X a random variable and $p \geq 0$ we write $\|X\|_p := (\mathbf{E}\left[X^p\right])^{1/p}$, and call $\|X\|_p$ the L_p-norm of X. Jensen's inequality immediately yields monotonicity of the L_p norms: if $0 \leq p \leq q$ then using the convexity of the function $x \mapsto x^{p/q}$,

$$\begin{aligned} \|X\|_p^p = \mathbf{E}\left[|X|^p\right] = \mathbf{E}\left[(|X|^q)^{p/q}\right] &= \lim_{n\to\infty} \mathbf{E}\left[(|X^{\leq n}|^q)^{p/q}\right] \\ &\geq \lim_{n\to\infty} (\mathbf{E}\left[|X^{\leq n}|^q\right])^{p/q} = \mathbf{E}\left[|X|^q\right]^{p/q} = \|X\|_q^p,; \end{aligned}$$

which in a sense strengthens (8.1). (We had to use the monotone convergence theorem because it's possible that $\mathbf{E}\left[|X|^q\right] = \infty$, in which case Jensen's inequality doesn't apply directly.)

Given random variables $(X_n, 1 \leq n \leq \infty)$ defined over a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$, we say that $X_n \to X_\infty$ *in* $L_p(\mathbf{P})$, and write $X_n \xrightarrow{L_p} X_\infty$, if $X_n \in L_p(\mathbf{P})$ for all $1 \leq n \leq \infty$ and $\|X_n - X_\infty\|_p \to 0$ as $n \to \infty$.

Exercise 9.2. *For any $p > 0$ and any random variables $(X_n, n \geq 1), X, Y \in L_p(\mathbf{P})$, if $X_n \to X$ in $L_p(\mathbf{P})$ and $X_n \to Y$ in $L_p(\mathbf{P})$ then $X \overset{\text{a.s.}}{=} Y$.*

The monotonicity of the L_p norms immediately implies that for $0 < q \leq p$, if $X_n \xrightarrow{L_p} X_\infty$ then $X_n \xrightarrow{L_q} X_\infty$. The next proposition states that convergence in L_p is at least as strong as convergence in probability.

Proposition 9.2. *Let $(X_n, 1 \leq n \leq \infty)$ be real random variables defined on a common space. For any $p > 0$, if $X_n \xrightarrow{L_p} X_\infty$ then $X_n \xrightarrow{\text{p}} X_\infty$.*

Proof. If $X_n \xrightarrow{L_p} X_\infty$ then for any $\epsilon > 0$,

$$\mathbf{P}\left\{|X_n - X_\infty| \geq \epsilon\right\} = \mathbf{P}\left\{|X_n - X_\infty|^p \geq \epsilon^p\right\} \leq \frac{\mathbf{E}\left[|X_n - X_\infty|^p\right]}{\epsilon^p} \to 0\,,$$

as $n \to \infty$. □

The next exercise asks you to analyze examples which show that convergence in probability does not imply convergence in $L_p(\mathbf{P})$, which in turn does not imply almost sure convergence.

Exercise 9.3. *Let $(B_n, n \geq 1)$ be independent random variables with B_n Bernoulli$(1/n)$-distributed. Fix $p > 0$ and for $n \geq 1$ let $X_n = n^{1/p} B_n$. Show that $B_n \xrightarrow{L_p} 0$ but that B_n does not converge to 0 almost surely. Show further that $X_n \xrightarrow{\text{p}} 0$ but that X_n does not converge to 0 in L_p.*

Jensen's inequality also allows us to prove Hölder's inequality, which provides a tool for showing that a product of random variables is integrable.

Theorem 9.3 (Hölder's inequality)**.** *Fix $p, q \geq 1$ with $1 \leq p, q \leq \infty$. If $1/p + 1/q = 1$ then for any random variables X, Y defined on a common probability space,*

$$\|XY\|_1 \leq \|X\|_p \|Y\|_q\,.$$

Proof. We may assume that $\|X\|_1 > 0$ and that $\|Y\|_1 > 0$ or else the left-hand side is zero. Similarly, we may assume that $\|X\|_p < \infty$ and that $\|Y\|_q < \infty$ or else the right-hand side is infinite. Finally, we may assume that $X \geq 0$ and $Y \geq 0$ since the values of both the left- and right-hand sides are unchanged if we replace X by $|X|$ and Y by $|Y|$.

We now write

$$\begin{aligned}\mathbf{E}|XY| &= \mathbf{E}\left[e^{\log(XY)}\right] \\ &= \mathbf{E}\left[e^{\log X + \log Y}\right] \\ &= \mathbf{E}\left[e^{\frac{1}{p}\log(X^p)+\frac{1}{q}\log(Y^q)}\right]\end{aligned}$$

Since $u \mapsto \log u$ is concave, $\frac{1}{p}\log(X^p) + \frac{1}{q}\log(Y^q) \leq \log(\frac{1}{p}X^p + \frac{1}{q}Y^q)$, so □

The Cauchy-Schwarz inequality is the case $p = q = 2$ of Hölder's inequality.

Corollary 9.4 (Cauchy-Schwarz inequality for random variables)**.** *For any random variables X, Y defined on a common probability space, $\|XY\|_1 \leq \|X\|_2\|Y\|_2$.*

The next exercise asks you to verify the "$p = 1, q = \infty$" case of Hölder's inequality. Given a random variable X we write $\operatorname{ess\,sup} X := \sup\{c \in \mathbb{R} : \mathbf{P}\{X > c\} > 0\}$ and call $\operatorname{ess\,sup} X$ the *essential supremum* of X. We write $\|X\|_\infty := \operatorname{ess\,sup}|X|$, and let

$$L_\infty(\Omega, \mathcal{F}, \mathbf{P}) := \{X : \Omega \to \mathbb{R} : X \text{ is } (\mathcal{F}/\mathcal{B}(\mathbb{R}))\text{-measurable and } \|X\|_\infty < \infty\}.$$

$L_\infty(\Omega, \mathcal{F}, \mathbf{P})$.

Exercise 9.4. *Let X, Y be random variables defined on a common space. Show that* $\operatorname{ess\,sup}|X| = \lim_{p\to\infty}\|X\|_p$. *Then show that*

$$\|XY\|_1 \leq \|X\|_\infty\|Y\|_1.$$

A clever application of Hölder's inequality yields *Minkowski's inequality*, which is the triangle inequality for L_p spaces.

Theorem 9.5 (Minkowski's inequality)**.** *Let X, Y be random variables in $L_1(\mathbf{P})$. Then for all $p \geq 1$, $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$.*

Proof. When $p = 1$ this follows from monotonicity of expectation, since $|X + Y| \leq |X| + |Y|$ by the triangle inequality. For $p > 1$ we also use the triangle inequality, as follows:

$$\|X+Y\|_p^p = \mathbf{E}\left[|X + Y|^p\right] = \mathbf{E}\left[|X + Y||X + Y|^{p-1}\right] \leq \mathbf{E}\left[|X||X + Y|^{p-1}\right] + \mathbf{E}\left[|Y||X + Y|^{p-1}\right].$$

Applying Hölder's inequality to each of the above expectations gives

$$\mathbf{E}\left[|X||X + Y|^{p-1}\right] \leq (\mathbf{E}\left[|X|^p\right])^{1/p}(\mathbf{E}\left[|X + Y|^p\right])^{(p-1)/p} = \|X\|_p\|X + Y\|_p^{(p-1)/p}$$

and

$$\mathbf{E}\left[|Y||X + Y|^{p-1}\right] \leq (\mathbf{E}\left[|Y|^p\right])^{1/p}(\mathbf{E}\left[|X + Y|^p\right])^{(p-1)/p} = \|Y\|_p\|X + Y\|_p^{(p-1)/p},$$

so

$$\|X + Y\|_p^p \leq (\|X\|_p + \|Y\|_p)\|X + Y\|_p^{(p-1)/p}.$$

Dividing by $\|X + Y\|_p^{(p-1)/p}$ completes the proof. □

We would like to think of $L_p(\Omega, \mathcal{F}, \mathbf{P})$ as a metric space, but this isn't quite right because a metric space (M, d) is supposed to satisfy that $d(x, y) = 0$ if and only if $x = y$. But $\|X - Y\|_p = 0$ provided that $X \overset{\text{a.s.}}{=} Y$, and almost sure equality is not the same as equality.

There are two ways to deal with this. The first approach, which is the most standard in probability, is to simply accept that instead of a metric space we only have a pseudo-metric space. (A pseudo-metric space is a pair (M, d) where $d : M \times M \to [0, \infty)$ is a symmetric function satisfying the triangle inequality. In a pseudo-metric space it is possible to have $d(x, y) = 0$ for distinct points

x, y.) The other is to quotient by almost sure equality. In other words, for $X \in L_p(\Omega, \mathcal{F}, \mathbf{P})$ we may write $[X] = \{Y \in L_p(\Omega, \mathcal{F}, \mathbf{P}) : X \stackrel{\text{a.s.}}{=} Y\}$ and $[L_p(\Omega, \mathcal{F}, \mathbf{P})] = \{[X] : X \in L_p(\Omega, \mathcal{F}, \mathbf{P})\}$.

Exercise 9.5.
- *Check that almost sure equality is an equivalence relation.*
- *Check that if $X, Y \in L_p(\Omega, \mathcal{F}, \mathbf{P})$ and $X' \in [X]$ and $Y' \in [Y]$, then $\|X' - Y'\|_p = \|X - Y\|_p$. That is, L_p distance is a class function for the "almost sure equality" equivalence relation.*

The next theorem implies that $[L_p(\Omega, \mathcal{F}, \mathbf{P})]$ is a complete metric space for any probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

Theorem 9.6. *Fix a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and $p \geq 1$, and let $(X_n, n \geq 1)$ be a Cauchy sequence in $L_p(\mathbf{P})$. Then there is $X \in L_p(\mathbf{P})$ such that $X_n \xrightarrow{L_p} X$. Moreover, for any other random variable $X' : \Omega \to \mathbb{R}$, if $\|X_n - X'\|_p \to 0$ then $X' \stackrel{\text{a.s.}}{=} X$.*

Proof. Since $(X_n, n \geq 1)$ is Cauchy, we can find an increasing sequence of integers $(n(k), k \geq 1)$ such that for all $m, n \in \mathbb{N}$, if $m, n \geq n(k)$ then $\|X_m - X_n\|_p \leq 2^{-k}$.

Then write $Y_k = X_{n(k)}$. By our choice of the sequence $(n(k), k \geq 1)$ we have $\|Y_{k+1} - Y_k\|_p \leq 2^{-k}$, so

$$\mathbf{E}\left[\sum_{k\geq 1} |Y_{k+1} - Y_k|\right] = \sum_{k\geq 1} \mathbf{E}|Y_{k+1} - Y_k| = \sum_{k\geq 1} \|Y_{k+1} - Y_k\|_1 \leq \sum_{k\geq 1} \|Y_{k+1} - Y_k\|_p \leq 1\,.$$

It follows that $\mathbf{P}\left\{\sum_{k\geq 1} |Y_{k+1} - Y_k| < \infty\right\} = 1$, or in other words that $(Y_{k+1} - Y_k, k \geq 1)$ is almost surely absolutely convergent. Letting $Y := \limsup_{k\to\infty} Y_k$, it follows that $\mathbf{P}\{\lim_{k\to\infty} Y_k = Y\} = 1$.

Now, for $n \geq n(k)$, note that

$$X_{n(k)} + \sum_{\ell \geq k} (Y_{\ell+1} - Y_\ell) = Y_k + \sum_{\ell \geq k} (Y_{\ell+1} - Y_\ell) \stackrel{\text{a.s.}}{=} Y,$$

so for $n \geq n(k)$,

$$\begin{aligned}
\|X_n - Y\|_p &= \|X_n - X_{n(k)} - \sum_{\ell \geq k} (Y_{\ell+1} - Y_\ell)\|_p \\
&\leq \|X_n - X_{n(k)}\|_p + \sum_{\ell \geq k} \|Y_{\ell+1} - Y_\ell\|_p \\
&\leq \frac{1}{2^k} + \sum_{\ell \geq k} \frac{1}{2^\ell} = \frac{1}{2^{k-2}}\,.
\end{aligned}$$

Thus $\|X_n - Y\|_p \to 0$ as $p \to \infty$. Since $\|Y\|_p = \|X_n - Y - X_n\|_p \leq \|X_n - Y\|_p + \|X_n\|_p$, it follows that $\|Y\|_p < \infty$, so $X_n \xrightarrow{L_p} Y$. Finally, the almost sure uniqueness of the limit is Exercise 9.2. □

9.1. **The geometric structure of L_2.** The space $L_2(\mathbf{P})$ is special because it can be endowed with a natural inner product structure, which allows us to harness the power of geometry. For $X, Y \in L^2(\mathbf{P})$, let $\langle X, Y\rangle := \mathbf{E}[XY]$; that the right-hand side is defined follows from the Cauchy-Schwarz inequality. You should check that $\langle \cdot, \cdot \rangle : L_2(\mathbf{P}) \times L_2(\mathbf{P})$ satisfies the axioms of an inner product (up to almost sure equivalence): it is symmetric and bilinear, and $\langle X, X\rangle = 0$ if and only if $X \stackrel{\text{a.s.}}{=} 0$. (The "true" inner product space is $[L_2(\Omega, \mathcal{F}, \mathbf{P})]$, but we will continue working with random variables, at the cost of occasionally having to use the phrase "almost sure".)

If $X > 0$ and $Y > 0$ are random variables in $L_2(\mathbf{P})$ then we may use the inner product to define an angle $\theta_{XY} \in [0, \pi]$ by the formula

$$\cos\theta_{XY} = \frac{\langle X, Y\rangle}{\|X\|_2 \|Y\|_2}\,.$$

Note that $\cos\theta_{XX} = \mathbf{E}\left[X^2\right]/\|X\|_2^2 = 1$, so $\theta_{XX} = 0$. This geometric structure is closely related to the covariance of the random variables: recall that for $X, Y \in L_2(\mathbf{P})$,

$$\mathrm{Cov}(X,Y) = \mathbf{E}\left[(X - \mathbf{E}X)(Y - \mathbf{E}Y)\right] = \langle X, Y\rangle - \mathbf{E}X\mathbf{E}Y\,.$$

In the case that $\mathbf{E}X = 0$ or $\mathbf{E}Y = 0$, it follows that X and Y are uncorrelated if and only if $\langle X, Y\rangle = 0$ or equivalently if and only if $\theta_{X,Y} = \pi/2$. We also have that

picture?

$$\|X+Y\|_2^2 = \mathbf{E}\left[(X+Y)^2\right] = \mathbf{E}\left[X^2\right] + 2\mathbf{E}\left[XY\right] + \mathbf{E}\left[Y^2\right] = \|X\|_2^2 + \langle X, Y\rangle + \|Y\|_2^2\,,$$

so $\|X+Y\|_2^2 = \|X\|_2^2 + \|Y\|_2^2$ if and only if $\langle X, Y\rangle = 0$.

Exercise 9.6 (Parallelogram Law)**.** *For $U, V \in L_2(\mathbf{P})$ we have*

$$\|U+V\|_2^2 + \|U-V\|_2^2 = 2\|U\|_2^2 + 2\|V\|_2^2.$$

Covariance has a very direct relation to the geometric structure of $L_2(\Omega, \mathcal{F}, \mathbf{P})$. Another feature of the geometry which we will exploit is the ability to perform *projections* onto subspaces. Consider a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a sub-σ-field $\mathcal{G}$ of $\mathcal{F}$. If $X : \Omega \to \mathbb{R}$ is $(\mathcal{G}/\mathcal{B}(\mathbb{R}))$-measurable then it is $(\mathcal{F}/\mathcal{B}(\mathbb{R}))$-measurable, so for any $p \geq 1$, if $Z \in L_p(\Omega, \mathcal{G}, \mathbf{P})$ then $Z \in L_p(\Omega, \mathcal{F}, \mathbf{P})$. In other words, "up to almost sure equality" the space $L_p(\Omega, \mathcal{G}, \mathbf{P})$ is a complete subspace of $L_p(\Omega, \mathcal{F}, \mathbf{P})$. In the case $p = 2$, the existence of a notion of orthogonality then allows us to consider projections onto $L_p(\Omega, \mathcal{G}, \mathbf{P})$.

We are abusing notation by writing $(\Omega, \mathcal{G}, \mathbf{P})$ rather than $(\Omega, \mathcal{G}, \mathbf{P}|_{\mathcal{G}})$, but this shouldn't cause confusion.

Theorem 9.7. *Fix a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a sub-σ-field $\mathcal{G}$ of $\mathcal{F}$. Fix $X \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ and let $\Delta = \inf\{\|X - Y\|_2 : Y \in L_2(\Omega, \mathcal{G}, \mathbf{P})\}$. Then there is an almost surely unique $\in L_2(\Omega, \mathcal{G}, \mathbf{P})$ such that $\|X - Z\|_2 = \Delta$.*

Proof. Let $(Y_n, n \geq 1)$ be elements of $L_2(\Omega, \mathcal{G}, \mathbf{P})$ with $\|X - Y_n\|_2 \leq \Delta + 1/n$. For $m, n \geq 1$, we apply the parallelogram law with $U + V = X - Y_n$ and $U - V = X - Y_m$. This means $2U = 2X - Y_n + Y_m$ and $2V = Y_m - Y_n$, so we obtain

$$\|X - Y_n\|_2^2 + \|X - Y_m\|_2^2 = 2\|X - (Y_n - Y_m)/2\|_2^2 = \frac{1}{2}\|Y_m - Y_n\|_2^2.$$

The left-hand side is at most $2\Delta^2 + 1/m + 1/n$ by our choice of Y_m and Y_n. Also, $(Y_n - Y_m)/2 \in L_2(\Omega, \mathcal{G}, \mathbf{P})$, so by the definition of Δ we have $\|X - (Y_n - Y_m)/2\|_2^2 \geq \Delta^2$. From the above equality combined with these two bounds we obtain

$$\frac{1}{2}\|Y_m - Y_n\|_2^2 \leq \frac{1}{m} + \frac{1}{n}\,,$$

so $(Y_n, n \geq 1)$ is a Cauchy sequence. By Theorem 9.6, it follows that there is $Y_\in L_2(\Omega, \mathcal{G}, \mathbf{P})$ such that $\|Y_n - Y\|_2 \to 0$. For any $n \geq 1$, by the triangle inequality, we then have

$$\|X - Y\|_2 \leq \|X - Y_n\| + \|Y_n - Y\|_2 \leq \Delta + \frac{1}{n} + \|Y_n - Y\|_2\,,$$

and taking $n \to \infty$ shows that $\|X - Y\| \leq \Delta$; by the definition of Δ we must then have $\|X - Y\|_2 = \Delta$.

Finally, suppose Z is another random variable with $\|X - Z\|_2 = \Delta$. Then apply the parallelogram law with $U + V = X - Y$ and $U - V = X - Z$. We then obtain

$$2\Delta^2 = \|X - Y\|_2^2 + \|X - Z\|_2^2 = 2\|X - (Y+Z)/2\|_2 + \frac{1}{2}\|Y - Z\|_2^2 \geq 2\Delta^2 + \frac{1}{2}\|Y - Z\|_2^2\,,$$

so it must be that $\|Y - Z\| = 0$. □

Let's call the (a.s. unique) minimizer Z in the above theorem the *closest $\mathcal{G}$-measurable random variable to X*. (This is cumbersome - we'll introduce a shorter name shortly.)

Corollary 9.8. *With the setup of Theorem 9.7, for $Z \in L_2(\Omega, \mathcal{G}, \mathbf{P})$ we have $\|X - Z\|_2 = \Delta$ if and only if $\langle Y, X - Z\rangle = 0$ for all $Y \in L_2(\Omega, \mathcal{G}, \mathbf{P})$.*

Proof. First, suppose that $Z \in L_2(\Omega, \mathcal{G}, \mathbf{P})$ is such that $\langle Y, X - Z\rangle = 0$ for all $Y \in L_2(\Omega, \mathcal{G}, \mathbf{P})$. Then for any $Z' \in L_2(\Omega, \mathcal{G}, \mathbf{P})$ we have

$$\begin{aligned}\mathbf{E}\left[(X - Z')^2\right] &= \mathbf{E}\left[(X - Z - (Z - Z'))^2\right] \\ &= \mathbf{E}\left[(X - Z)^2\right] - 2\langle X - Z, Z - Z'\rangle\rangle + \mathbf{E}\left[(Z - Z')^2\right] \\ &= \mathbf{E}\left[(X - Z)^2\right] + \mathbf{E}\left[(Z - Z')^2\right] \\ &\geq \mathbf{E}\left[(X - Z)^2\right] .\end{aligned}$$

Conversely, suppose that $\|X - Z\|_2 = \Delta$, and fix any $Y \in L_2(\Omega, \mathcal{G}, \mathbf{P})$. Then for any $t \in \mathbb{R}$ we have $Z + tY \in L_2(\Omega, \mathcal{G}, \mathbf{P})$ so

$$\begin{aligned}\Delta^2 &\leq \mathbf{E}\left[(X - Z - tY)^2\right] \\ &\mathbf{E}\left[(X - Z)^2\right] - 2t\mathbf{E}\left[Y(X - Z)\right] + t^2\mathbf{E}\left[Y^2\right] \\ &= \Delta^2 - 2t\langle Y, X - Z\rangle + t^2\mathbf{E}\left[Y^2\right] .\end{aligned}$$

The only way this can hold for small t is if $\langle Y, X - Z\rangle = 0$. □

The condition that $\langle Y, X - Z\rangle = 0$ for all $Y \in L_2(\Omega, \mathcal{G}, \mathbf{P})$ may be rewritten as saying that

$$\mathbf{E}\left[Y(X - Z)\right] = 0$$

for all $Y \in L_2(\Omega, \mathcal{G}, \mathbf{P})$. This is easy enough to verify that we can start doing examples.

Examples. The first several examples will relate to a sequence $(X_i, i \geq 1)$ of independent random variables in $L_2(\Omega, \mathcal{F}, \mathbf{P})$.

1. Suppose that $\mathbf{E}[X_i] = 0$ for all i. Fix $n \geq 1$, let $X = \sum_{i=1}^n X_i$, and let $\mathcal{G} = \sigma(X_1, \ldots, X_{n-1})$. We claim that $Z = X_1 + \ldots + X_{n-1}$ is the closest $\mathcal{G}$-measurable random variable to X. To see this, we fix any $Y \in L_2(\Omega, \mathcal{G}, \mathbf{P})$ and compute

$$\mathbf{E}\left[Y(X - Z)\right] = \mathbf{E}\left[YX_n\right] = \mathbf{E}\left[Y\right]\mathbf{E}\left[X_n\right] = 0.$$

The second equality holds since $\mathcal{G} = \sigma(X_1, \ldots, X_{n-1})$ and $\sigma(X_n)$ are independent, so Y and X_n are independent.

2. Again take $X = \sum_{i=1}^n X_i$, but this time don't assume the random variables have zero mean. Write $\mathbf{E}[X_i] = c_i$, fix some set $S \subset [n]$ and let $\mathcal{G} = \sigma(X_i, i \in S)$. If $Z_0 = \sum_{i\in S} X_i$ then for $Y \in L_2(\Omega, \mathcal{G}, \mathbf{P})$ we have

$$\mathbf{E}\left[Y(X - Z_0)\right] = \mathbf{E}\left[Y(\sum_{i\notin S} X_i\right] = \mathbf{E}\left[Y\right]\mathbf{E}\left[\sum_{i\notin S} X_i\right] = \mathbf{E}\left[Y\right]\cdot\sum_{i\notin S} c_i.$$

This need not be zero - we should *recenter* Z_0 to take account of what direction the remaining summands are heading in. Taking $Z = Z_0 + \sum_{i\notin S} c_i$, we then get

$$\mathbf{E}\left[Y(X - Z)\right] = \mathbf{E}\left[Y(X - Z_0)\right] - \mathbf{E}\left[Y\cdot\sum_{i\notin S} c_i\right] = 0\,,$$

so the closest $\mathcal{G}$-measurable random variable to X is $\sum_{i\in S} X_i + \sum_{i\notin S} c_i$.

3. Let $X = \prod_{i=1}^n X_i$ and take $\mathcal{G} = \sigma(X_1, \ldots, X_{n-1})$. Then with $c = \mathbf{E}X_n$, the closest $\mathcal{G}$-measurable random variable to X is $Z = c\cdot\prod_{i=1}^{n-1} X_i$. To see this, choose $Y \in L_2(\Omega, \mathcal{G}, \mathbf{P})$. Since the random variables $X_1, \ldots, X_n$ are independent, both X and Z are in $L_2(\Omega, \mathcal{F}, \mathbf{P})$ (exercise!). It follows by Cauchy-Schwarz that $YZ \in L_1(\Omega, \mathcal{F}, \mathbf{P})$; since X_n is independent of YZ and $YX = YZX_n/c$, we thus have

$$\mathbf{E}\left[YX\right] = \mathbf{E}\left[YZX_n/c\right] = \mathbf{E}\left[YZ\right]\mathbf{E}\left[X_n\right]/c = \mathbf{E}\left[YZ\right].$$

4. Fix $c \in \mathbb{R}$ and suppose that $\mathbf{E}X_i = c$ and (to avoid integrability issues) that $X_i \geq 0$ for all $i \geq 1$. Let N be a positive integer random variable independent of $(X_i, i \geq 1)$, and take $X = \sum_{i=1}^N X_i$

and $\mathcal{G} = \sigma(N)$. We claim that $Z = cN$. To see this, we transform the random sum into a deterministic sum by writing

$$X = \sum_{i=1}^{N} X_i = \sum_{n=1}^{\infty} (\mathbf{1}_{[N=n]} \cdot \sum_{i=1}^{n} X_i) .$$

For $Y \in L_2(\Omega, \mathcal{G}, \mathbf{P})$ we then have

$$\begin{aligned} \mathbf{E}[YX] &= \mathbf{E}\left[Y \cdot \sum_{n=1}^{\infty} (\mathbf{1}_{[N=n]} \cdot \sum_{i=1}^{n} X_i) \right] \\ &= \sum_{n=1}^{\infty} \mathbf{E}\left[Y \mathbf{1}_{[N=n]} \cdot \sum_{i=1}^{n} X_i \right] . \end{aligned}$$

Now, $\mathcal{G} = \sigma(N)$ and $\sigma(X_i, i \geq 1)$ are independent, so the random variables $Y\mathbf{1}_{[N=n]}$ and $\sum_{i=1}^{n} X_i$ are independent. Applying the factorization formula to the right-hand side above then gives

$$\begin{aligned} \mathbf{E}[YX] &= \sum_{n=1}^{\infty} \mathbf{E}\left[Y \mathbf{1}_{[N=n]} \right] \cdot \mathbf{E}\left[\sum_{i=1}^{n} X_i \right] \\ &= \sum_{n=1}^{\infty} \mathbf{E}\left[Y \mathbf{1}_{[N=n]} \right] \cdot cn \\ &= \mathbf{E}\left[\sum_{n=1}^{\infty} Y \mathbf{1}_{[N=n]} \cdot cn \right] \\ &= \mathbf{E}[Y \cdot cN] . \end{aligned}$$

5. This example is chattier. The idea behind it is a bit different from the others, and is quite important. Let Ω be the set of all individuals who filed an income tax return in Canada in 2018, and let $\mathbf{P}$ be the uniform measure on $(\Omega, 2^{\Omega})$. Define a random variable $X : \Omega \to \mathbb{R}$ by taking $X(\omega)$ to be the amount of income tax paid by individual ω.

Define another random variable $R : \Omega \to \{1, \ldots, 13\}$ by taking $R(\omega)$ to be the province or territory of residence of individual ω in 2018[12], and let $\mathcal{G} = \sigma(R)$. This means that (for example) $\Omega_1 := R^{-1}(1)$ is the set of taxpayers in Alberta, and $\Omega_{13} := R^{-1}(13)$ is the set of taxpayers in the Yukon.

Note that $\mathcal{G}$ is generated by the sets $\Omega_1, \ldots, \Omega_{13}$. That means a random variable $U : \Omega \to \mathbb{R}$ is $(\mathcal{G}/\mathcal{B}(\mathbf{R}))$-measurable if and only if for any $B \in \mathcal{B}(\mathbf{R})$, the set $U^{-1}(B)$ is a union of some or all of the sets $\Omega_1, \ldots, \Omega_{13}$. In other words, whether $U(\omega) \in B$ must only depend on the value of $R(\omega)$, so $U(\omega)$ must be the same for every taxpayer in a given province.

What is the closest $\mathcal{G}$-measurable random variable to X? We seek a random variable Z which assigns the same value to every taxpayer in a province, and satisfies

$$\mathbf{E}[XY] = \mathbf{E}[ZY]$$

for any other random variable Y which also assigns the same value to every taxpayer in a province. Suppose Y has that property, so we may represent Y as $Y(i) = \sum_{i=1}^{13} y_i \mathbf{1}_{[\omega \in \Omega_i]}$. Then

$$\mathbf{E}[XY] = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} X(\omega) Y(\omega) = \frac{1}{|\Omega|} \sum_{i=1}^{13} y_i \cdot \sum_{\omega \in \Omega_i} X(\omega) .$$

[12]Order the provinces and territories in some way...

If $Z = \sum_{i=1}^{13} z_i \mathbf{1}_{[\omega \in \Omega_i]}$ then

$$\mathbf{E}[ZY] = \frac{1}{|\Omega|}\sum_{\omega\in\Omega} Z(\omega)Y(\omega) = \frac{1}{|\Omega|}\sum_{i=1}^{13} |\Omega_i| y_i z_i \,.$$

To make the last two expressions equal, we see that we should take $z_i = |\Omega_i|^{-1}\sum_{\omega\in\Omega_i} X_i(\omega)$. This last value is just the average tax paid by taxpayers in province/territory i! Calling that value μ_i, we then have

$$Z(\omega) = \sum_{i=1}^{13} \mu_i \mathbf{1}_{[\Omega_i]}(\omega)\,.$$

It's worth comparing this example to the previous ones. In examples 1, 2 and 3, the closest $\mathcal{G}$-measurable random variable to X ended up being obtained by essentially "replacing the part not lying in the subspace by its expected value". In example 4, the "expectation of the independent part" also came into the picture, but in a more involved way, as X did not separate as cleanly as in the first three cases. In example 5, the random variable Z is a sort of "coarsening" of X, obtained by taking expectations over subsets whenever $\mathcal{G}$ gives us no information about the behaviour of X within those subsets. If you think of X as a lookup table where the first column lists taxpayers and the second lists the amount they paid, then Z is a table which only lists the average (expected) amount paid per province or territory.

This motivates the name we will use for such random variables for the rest of the class; rather than calling Z the closest $\mathcal{G}$-measurable random variable to X, we call it the *conditional expectation of X given $\mathcal{G}$*, and denote it $\mathbf{E}[X \mid \mathcal{G}]$.

This notation takes some time to get used to. The conditional expectation $\mathbf{E}[X \mid \mathcal{G}]$ is not a *number*: it is a random variable, which "tries to be like X" but is forced to be simpler than X by the constraint that it must be $(\mathcal{G}/\mathcal{B}(\mathbb{R}))$-measurable. The next section is devoted to conditional expectation and its properties.

10. **Conditional expectation**

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra. For a random variable $X \in L_1(\Omega, \mathcal{F}, \mathbf{P})$, we say that $Z : \Omega \to \mathbb{R}$ is a version of $\mathbf{E}[X \mid \mathcal{G}]$ if

(a) $Z \in L_1(\Omega, \mathcal{G}, \mathbf{P})$, and
(b) For all $E \in \mathcal{G}$, $\mathbf{E}\left[X\mathbf{1}_{[E]}\right] = \mathbf{E}\left[Z\mathbf{1}_{[E]}\right]$.

We will momentarily show existence and (almost sure) uniqueness of conditional expectations of L_1 random variables. First, however, we establish a monotonicity property of conditional expectations. We use the result of the following easy exercise.

Did I already state this exercise or a close variant earlier?

Exercise 10.1. *Suppose that random variables U and V have $\mathbf{P}\{U \geq V\} = 1$ and $\mathbf{E}U \leq \mathbf{E}V$. Then $U \overset{\text{a.s.}}{=} V$.*

Proposition 10.1 (Monotonicity of conditional expectation)**.** *Suppose that $X, X' \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ satisfy $\mathbf{P}\{X \leq X'\} = 1$, and that Z, Z' are versions of $\mathbf{E}[X \mid \mathcal{G}]$ and $\mathbf{E}[X' \mid \mathcal{G}]$, respectively. Then $\mathbf{P}\{Z \leq Z'\} = 1$.*

Proof. Since $Z, Z \in L_1(\Omega, \mathcal{G}, \mathbf{P})$ we have $Z - Z' \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ so $\{Z \geq Z'\} = \{Z - Z' \geq 0\} \in \mathcal{G}$. Thus, by the defining property (b) of conditional expectation and monotonicity of expectation,

$$\mathbf{E}\left[Z\mathbf{1}_{[Z\geq Z']}\right] = \mathbf{E}\left[X\mathbf{1}_{[Z\geq Z']}\right] \leq \mathbf{E}\left[X'\mathbf{1}_{[Z\geq Z']}\right] = \mathbf{E}\left[Z'\mathbf{1}_{[Z\geq Z']}\right]\,.$$

But $Z\mathbf{1}_{[Z\geq Z']} \geq Z'\mathbf{1}_{[Z\geq Z']}$, so it follows by the above exercise that $\mathbf{P}\left\{Z\mathbf{1}_{[Z\geq Z']} = Z'\mathbf{1}_{[Z\geq Z']}\right\} = 1$, which is equivalent to the assertion that $\mathbf{P}\{Z \leq Z'\} = 1$. □

Theorem 10.2 (Existence of conditional expectation)**.** *For any random variable $X \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ and any sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$, there exists a version of $\mathbf{E}[X \mid \mathcal{G}]$. Moreover, if Y and Y' are two versions of $\mathbf{E}[X \mid \mathcal{G}]$ then $Y \overset{\text{a.s.}}{=} Y'$.*

Proof. We first prove the uniqueness claim. Suppose that Z, Z' are two versions of $\mathbf{E}[X \mid \mathcal{G}]$. Applying Proposition 10.1 with $X' = X$ we have $\mathbf{P}\{Z \leq Z'\} = 1$; by symmetry we then have $\mathbf{P}\{Z = Z'\} = 1$, establishing the uniqueness claimed in the theorem statement.

To prove existence, first suppose $X \in L_2(\Omega, \mathcal{F}, \mathbf{P})$. Then by Corollary 9.8, there is $Z \in L_2(\Omega, \mathcal{G}, \mathbf{P})$ such that

$$\mathbf{E}[XY] = \mathbf{E}[ZY]$$

for all $Y \in L_2(\Omega, \mathcal{G}, \mathbf{P})$. In particular this holds when $Y = \mathbf{1}_{[E]}$ for $E \in \mathcal{G}$, so Z is a version of $\mathbf{E}[X \mid \mathcal{G}]$.

Now suppose $X \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ is non-negative. Then for for each $n \geq 1$, since $X^{\leq n}$ is bounded it is in $L_2(\Omega, \mathcal{F}, \mathbf{P})$, so we may find a version Z_n of $\mathbf{E}\left[X^{\leq n} \mid \mathcal{G}\right]$. By monotonicity of conditional expectation, the random variables $(Z_n, n \geq 1)$ are almost surely increasing. Set $Z = \limsup_{n\to\infty} Z_n$, so that Z_n almost surely increases to Z. Then for any event $E \in \mathcal{G}$, by two applications of the monotone convergence theorem we then have

$$\mathbf{E}\left[X\mathbf{1}_{[E]}\right] = \lim_{n\to\infty} \mathbf{E}\left[X^{\leq n}\mathbf{1}_{[E]}\right] = \lim_{n\to\infty} \mathbf{E}\left[Z_n\mathbf{1}_{[E]}\right] = \mathbf{E}\left[Z\mathbf{1}_{[E]}\right],$$

so Z is a version of $\mathbf{E}[X \mid \mathcal{G}]$.

Finally, for arbitrary $X \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ we may write $X = X^+ - X^-$ and let Z_+ and Z_- be versions of $\mathbf{E}[X^+ \mid \mathcal{G}]$ and $\mathbf{E}[X^- \mid \mathcal{G}]$, respectively. Then for $E \in \mathcal{G}$, using linearity of expectation we have

$$\mathbf{E}\left[X\mathbf{1}_{[E]}\right] = \mathbf{E}\left[X^+\mathbf{1}_{[E]}\right] - \mathbf{E}\left[X^-\mathbf{1}_{[E]}\right] = \mathbf{E}\left[Z_+\mathbf{1}_{[E]}\right] - \mathbf{E}\left[Z_-\mathbf{1}_{[E]}\right] = \mathbf{E}\left[(Z_+ - Z_-)\mathbf{1}_{[E]}\right],$$

so $Z_+ - Z_-$ is a version of $\mathbf{E}[X \mid \mathcal{G}]$. □

It is immediate from the definition that in the five examples with which we concluded the preceding section, the "closest $\mathcal{G}$-measurable random variables to X" were in fact versions of $\mathbf{E}[X \mid \mathcal{G}]$.

Exercise 10.2. *Use the monotone class theorem to show that if Z is a version of $\mathbf{E}[X \mid \mathcal{G}]$ then for any $Y \in L_\infty(\Omega, \mathcal{G}, \mathbf{P})$, $\mathbf{E}[XY] = \mathbf{E}[ZY]$.*

We'll spend some time on further examples of conditional expectations, but first discuss notation a little bit more. We'll sometimes start writing $\mathbf{E}[X \mid \mathcal{G}]$ rather than referring to versions of $\mathbf{E}[X \mid \mathcal{G}]$. Also, if $\mathcal{G} = \sigma(V)$ for some random variable V, it's standard to write $\mathbf{E}[X \mid V]$ rather than $\mathbf{E}[X \mid \mathcal{G}]$ or $\mathbf{E}[X \mid \sigma(V)]$.

More examples.

1. Our first example generalizes the last example from the previous section. Suppose $X \in L_1(\Omega, \mathcal{F}, \mathbf{P})$. Let $(\Omega_n, n \geq 1)$ be a partition of Ω with all parts in $\mathcal{F}$, and let $\mathcal{G} = \sigma(\{\Omega_n, n \geq 1\})$. Write $z_n = \mathbf{E}\left[X\mathbf{1}_{[\Omega_n]}\right] / \mathbf{P}\{\Omega_n\}$. Then the random variable $Z = \sum_{n\geq 1} z_n \mathbf{1}_{[\Omega_n]}$ is a version of $\mathbf{E}[X \mid \mathcal{G}]$. To see this is easy: any event E in $\mathcal{G}$ may be written as

$$E = \sum_{i\in S} \Omega_i$$

for some $S \subset \mathbb{N}$, and then

$$\mathbf{E}\left[X\mathbf{1}_{[E]}\right] = \sum_{n\in S} \mathbf{E}\left[X\mathbf{1}_{[\Omega_n]}\right] = \sum_{i\in S} z_n \mathbf{P}\{\Omega_n\} = \sum_{n\in S} z_n \mathbf{E}\left[\mathbf{1}_{[\Omega_n]}\right] = \mathbf{E}\left[\sum_{n\in S} Z\mathbf{1}_{[\Omega_n]}\right] = \mathbf{E}\left[Z\mathbf{1}_{[E]}\right]$$

2. Say that $X, Y \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ have *joint density* f if $f : \mathbb{R}^2 \to [0, \infty)$ is a Borel function which is a density for the $\mathbb{R}^2$-valued random variable (X, Y); that is, for any $B \in \mathcal{B}(\mathbb{R}^2)$,

$$\mathbf{P}\{(X, Y) \in B\} = \int_B f(x, y)\mathrm{d}x \otimes \mathrm{d}y\,,$$

where we use $\mathrm{d}x \otimes \mathrm{d}y$ to denote Lebesgue measure on $\mathbb{R}^2$. Suppose X and Y have joint density f.

The "natural formula" for $\mathbf{E}[X \mid Y = y]$ would be

$$\mathbf{E}[X \mid Y = y] = \frac{\int_{\mathbb{R}} xf(x,y)\mathrm{d}x}{\int_{\mathbb{R}} f(x,y)\mathrm{d}x}.$$

The top is just the integral along the slice, and the bottom is a normalization factor. If we

Now, cast your mind back to the development of product measure and Fubini's theorem. Lemma 6.5 tells us that $f(x,y)$ is a Borel function of x; Lemma 6.6 tells us that

$$\int_{\mathbb{R}} f(x,y)\mathrm{d}x$$

is a Borel function of y, and Fubini's theorem tells us that

$$\int_{\mathbb{R}} |f(x,y)|\mathrm{d}x < \infty$$

almost everywhere. We can thus define

$$\phi(y) = \begin{cases} \int_{\mathbb{R}} f(x,y)\mathrm{d}x & \text{if } \int_{\mathbb{R}} |f(x,y)|\mathrm{d}x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

and by Fubini's theorem, for $A \in \mathcal{B}(\mathbb{R})$ we have

$$\mathbf{P}\{Y \in A\} = \mathbf{P}\{(X,Y) \in \mathbb{R} \times A\} = \int_{\mathbb{R}\times A} f(x,y)\mathrm{d}x \otimes \mathrm{d}y = \int_A \phi(y)\mathrm{d}y.$$

In other words, ϕ is a density for Y. Now let

$$f(x|y) = \begin{cases} \frac{f(x,y)}{\phi(y)} & \text{if } \phi(y) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

This is a Borel function from $\mathbb{R}^2$ to $\mathbb{R}$ (exercise!), and so is a Borel function in either coordinate (when the other is held fixed).

We now want to define $\mathbf{E}[X \mid Y] = \int_{\mathbb{R}} xf(x|Y)\mathrm{d}x$. We really need to work on the set that the integral is defined, so let $Z = \int_{\mathbb{R}} xf(x|Y)\mathrm{d}x \cdot \mathbf{1}_{[Y\in F]}$. Then Z is a composition of Y with a Borel map, so is $\mathcal{G}/\mathcal{B}(\mathbb{R})$ measurable, and using the change of variables formula and monotonicity of probability (or Jensen's inequality),

$$\begin{aligned} \mathbf{E}|Z| &= \mathbf{E}\left[\left|\int_{\mathbb{R}} xf(x|Y)\mathrm{d}x\right| \mathbf{1}_{[Y\in F]}\right] \\ &= \int_{\mathbb{R}} \left|\int_F xf(x|y)\mathrm{d}x\right| \phi(y)\mathrm{d}y \\ &= \int_{\mathbb{R}} \left|\int_F xf(x,y)\mathrm{d}x\right| \mathrm{d}y \\ &\le \int_{\mathbb{R}}\int_{\mathbb{R}} xf(x,y)\mathrm{d}x\mathrm{d}y \\ &= \mathbf{E}|X| < \infty. \end{aligned}$$

Thus $Z \in L_1(\Omega, \mathcal{G}, \mathbf{P})$.

Finally, for any $E \in \sigma(Y)$ we can write $E = \{Y \in A\}$ for some $A \in \mathcal{B}(\mathbb{R})$, so by two applications of the change of variables formula,

$$\begin{aligned}\mathbf{E}\left[X\mathbf{1}_{[Y\in A]}\right] &= \int_{\mathbb{R}\times A} xf(x,y)\mathrm{d}x\otimes\mathrm{d}y\\ &= \int_A\int_{\mathbb{R}} xf(x,y)\mathrm{d}x\mathrm{d}y\\ &= \int_A\int_{\mathbb{R}} xf(x|y)\phi(y)\mathrm{d}x\mathrm{d}y\\ &= \int_A\int_{\mathbb{R}} xf(x|y)\mathrm{d}x\phi(y)\mathrm{d}y\\ &= \mathbf{E}\left[\int_{\mathbb{R}} xf(x|Y)\mathbf{1}_{[Y\in A]}\right].\end{aligned}$$

Thus $\int_{\mathbb{R}} xf(x|Y)\mathrm{d}x$ is indeed a version of $\mathbf{E}[X \mid Y]$.

3. Suppose that X and Y are independent and that $\phi : \mathbb{R}^2 \to \mathbb{R}$ is a Borel function with $\mathbf{E}|\phi(X,Y)| < \infty$. Let $F = \{y \in \mathbb{R} : \mathbf{E}|\phi(X,y)| < \infty\}$ and set $g(y) := (\mathbf{E}\phi(X,y))\mathbf{1}_{[y\in F]}$. We claim that $g(Y) \overset{\text{a.s.}}{=} \mathbf{E}\phi(X,Y) \mid Y$.

To see this, first note that by the change of variables formula and monotonicity of integration,

$$\begin{aligned}\mathbf{E}|g(Y)| &= \int_F |\mathbf{E}\phi(X,y)|\mathrm{d}\mu_Y\\ &= \int_F \left|\int_{\mathbb{R}} \phi(x,y)\mathrm{d}\mu_X\right|\mathrm{d}\mu_Y\\ &\le \int_F\int_{\mathbb{R}} |\phi(x,y)|\mathrm{d}\mu_X\mathrm{d}\mu_Y\,.\end{aligned}$$

Since X and Y are independent, the pair (X,Y) has joint law $\mathrm{d}_X \otimes \mathrm{d}_Y$, so by Fubini's theorem and another application of the change of variables formula,

$$\begin{aligned}\int_F\int_{\mathbb{R}} |\phi(x,y)|\mathrm{d}\mu_X\mathrm{d}\mu_Y &= \int_{\mathbb{R}^2} |\phi(x,y)|\mathrm{d}(\mu_X\otimes\mu_Y)\\ &= \mathbf{E}|\phi(X,Y)| < \infty\,,\end{aligned}$$

so $g(Y) \in L_1(\Omega, \sigma(Y), \mathbf{P})$. Next, for any $A \in \mathcal{B}(\mathbb{R})$, again using change of variables and Fubini we have

$$\begin{aligned}\mathbf{E}\left[g(Y)\mathbf{1}_{[Y\in A]}\right] &= \int_F \mathbf{E}\phi(X,y)\mathbf{1}_{[A]}(y)\mathrm{d}\mu_Y\\ &= \int_F\int_{\mathbb{R}} \phi(x,y)\mathbf{1}_{[A]}(y)\mathrm{d}\mu_X\mathrm{d}\mu_Y\\ &= \int_{A\times\mathbb{R}} \phi(x,y)\mathrm{d}(\mu_X\otimes\mu_Y)\\ &= \mathbf{E}\left[\phi(X,Y)\mathbf{1}_{[Y\in A]}\right]\,,\end{aligned}$$

as required.

4. This example is a straightforward generalization of the previous one to more than two random variables, and a detailed justification is omitted (only the construction is given). Suppose $(X_1,\ldots,X_n)$ are independent random variables on a probability space$(\Omega,\mathcal{F},\mathbf{P})$. Fix any Borel function $\phi : \mathbb{R}^n \to \mathbb{R}$ such that $\phi(X_1,\ldots,X_n) \in L_1(\Omega,\mathcal{F},\mathbf{P})$. Fix $1 \le i \le n$ and let $\mathcal{G} = \sigma(X_1,\ldots,X_i)$.

Let

$$F = \{(x_1,\ldots,x_i) \in \mathbb{R}^i : \mathbf{E}|\phi(x_1,\ldots,x_i,X_{i+1},\ldots,X_n)| < \infty\}.$$

Define $g : \mathbb{R}^i \to \mathbb{R}$ by

$$g(x_1, \ldots, x_i) = \mathbf{E}\left[\phi(x_1, \ldots, x_i, X_{i+1}, \ldots, X_n)\right] \mathbf{1}_{[(x_1,\ldots,x_i)\in F]}.$$

Then $g(X_1, \ldots, X_i)$ is a version of $\mathbf{E}\left[\phi(X_1, \ldots, X_n) \mid \mathcal{G}\right]$. This construction subsumes[13] examples 1 through 4 from Section 9.1.

10.1. **Properties of conditional expectation.** In this section we record a litany[14] of basic properties satisfied by conditional expectation. We always assume $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space, that $X, Y \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ and that $\mathcal{G}$ is a sub-σ-field of $\mathcal{F}$.

(i) $\mathbf{E}\left[\mathbf{E}\left\{X \mid \mathcal{G}\right\}\right] = \mathbf{E}X$

(ii) If $\sigma(X) \subset \mathcal{G}$ then $\mathbf{E}\left\{X \mid \mathcal{G}\right\} \overset{\text{a.s.}}{=} X$.

Proving the first two properties is an exercise in understanding and applying the definition of conditional expectation.

(iv) If $\sigma(X)$ and $\mathcal{G}$ are independent then $\mathbf{E}\left\{X \mid \mathcal{G}\right\} \overset{\text{a.s.}}{=} \mathbf{E}X$.
Proof: For all $A \in \mathcal{G}$, by the independence assumption,

$$\mathbf{E}\left[X\mathbf{1}_{[A]}\right] = \mathbf{E}X \cdot \mathbf{E}\left[\mathbf{1}_{[A]}\right] = \mathbf{E}\left[(\mathbf{E}X) \cdot \mathbf{1}_{[A]}\right]. \qquad \square$$

(v) **Linearity of conditional expectation.** For all $a \in \mathbb{R}$, $\mathbf{E}\left\{aX + Y \mid \mathcal{G}\right\} \overset{\text{a.s.}}{=} a\mathbf{E}\left\{X \mid \mathcal{G}\right\} + \mathbf{E}\left\{Y \mid \mathcal{G}\right\}$.

(vi) **Montonicity.** If $X \le Y$ almost surely then $\mathbf{E}\left\{X \mid \mathcal{G}\right\} \le \mathbf{E}\left\{Y \mid \mathcal{G}\right\}$ almost surely.

The last fact is just a restatement of Proposition 10.1.

For the next three properties, we additionally require a sequence $(X_n, n \ge 1)$ of random variables over $(\Omega, \mathcal{F}, \mathbf{P})$. The first is left as an exercise.

(vi) **Conditional Monotone Convergence Theorem.** If $0 \le X_n \uparrow X$ almost surely then $\mathbf{E}\left\{X_n \mid \mathcal{G}\right\} \uparrow \mathbf{E}\left\{X \mid \mathcal{G}\right\}$ almost surely.

(vii) **Conditional Fatou's Lemma.** If $X_n \ge 0$ for all n then $\mathbf{E}\left\{\liminf_{n\to\infty} X_n \mid \mathcal{G}\right\} \le \liminf_{n\to\infty} \mathbf{E}\left\{X_n \mid \mathcal{G}\right\}$.
Proof: For any $n \ge 1$, for all $n' \ge n$ we have $X_{n'} \ge \inf_{m\ge n} X_m$, and it follows by monotonicity of conditional expectation that

$$\inf_{m\ge n} \mathbf{E}\left\{X_m \mid \mathcal{G}\right\} \overset{\text{a.s.}}{\ge} \mathbf{E}\left\{ \inf_{m\ge n} X_m \;\middle|\; \mathcal{G}\right\}.$$

Taking $n \to \infty$ on both sides gives

$$\liminf_{n\to\infty} \mathbf{E}\left\{X_n \mid \mathcal{G}\right\} \overset{\text{a.s.}}{\ge} \lim_{n\to\infty} \mathbf{E}\left\{ \inf_{m\ge n} \;\middle|\; \mathcal{G}\right\} \overset{\text{a.s.}}{=} \mathbf{E}\left\{ \liminf_{n\to\infty} X_n \;\middle|\; \mathcal{G}\right\},$$

where the almost sure equality follows from the conditional monotone convergence theorem. $\square$

(viii) **Conditional Dominated Convergence Theorem.** If $X_n \overset{\text{a.s.}}{\to} X$ almost surely and $|X_n| \le Y$ almost surely for all n, then $\mathbf{E}\left\{X_n \mid \mathcal{G}\right\} \overset{\text{a.s.}}{\to} \mathbf{E}\left\{X \mid \mathcal{G}\right\}$.

The conditional dominated convergence theorem follows from the conditional Fatou's lemma in essentially the same way as the dominated convergence theorem follows from Fatou's lemma.

We next turn to inequalities related to convexity.

---

[13] Subsume, v.: 6. transitive. a. To take up or absorb (a concept, thing, person, etc.) into another, esp. one which is larger or higher; to include in. b. To bring (an idea, principle, etc.) under another; to instance or include (a case, term, etc.) under a rule, category, etc. –Oxford English Dictionary

[14] Litany, n.: 2. *transferred.* A form of supplication (e.g. in non-Christian worship) resembling a litany; also, a continuous repetition or long enumeration resembling those of litanies. –Oxford English Dictionary

(ix) **Conditional Jensen's inequality.** If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex and $\varphi(X) \in L_1(\Omega, \mathcal{F}, \mathbf{P})$, then $\varphi(\mathbf{E}\{X \mid \mathcal{G}\}) \leq \mathbf{E}\{\varphi(X) \mid \mathcal{G}\}$.
Proof: We may fix a sequence of linear functions $\ell_n(x) = a_n x + b_n$ such that for all $x \in \mathbb{R}$, $\varphi(x) = \sup_{n\geq 1}(a_n x + b_n)$. We then have $\varphi(X) \geq a_n X + b_n$ for all n, so by monotonicity and linearity of conditional expectation,

$$\mathbf{E}\{\varphi(X) \mid \mathcal{G}\} \overset{\text{a.s.}}{\geq} \mathbf{E}[a_n X + b_n]\,\mathcal{G} \overset{\text{a.s.}}{=} a_n \mathbf{E}\{X \mid \mathcal{G}\} + b_n.$$

Taking a supremum over $n \geq 1$ gives

$$\mathbf{E}\{\varphi(X) \mid \mathcal{G}\} \overset{\text{a.s.}}{\geq} \sup_{n\geq 1}(a_n \mathbf{E}\{X \mid \mathcal{G}\} + b_n) = \varphi(\mathbf{E}\{X \mid \mathcal{G}\}). \qquad \square$$

(x) For all $p \geq 1$, $\|\mathbf{E}\{X \mid \mathcal{G}\}\|_p \leq \|X\|_p$.
Proof: This is obvious if $\|X\|_p = \infty$. Otherwise, by the conditional Jensen's inequality applied to the function $\phi(x) = |x|$ with the random variable $X^p \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ we have $|\mathbf{E}\{X \mid \mathcal{G}\}|^p \leq \mathbf{E}\{|X|^p \mid \mathcal{G}\}$. It follows by monotonicity and the definition of conditional expectation that

$$\begin{aligned}\|\mathbf{E}\{X \mid \mathcal{G}\}\|_p^p &= \mathbf{E}\left[|\mathbf{E}\{X \mid \mathcal{G}\}|^p\right] \\ &\overset{\text{a.s.}}{\leq} \mathbf{E}\left[\mathbf{E}\{|X|^p \mid \mathcal{G}\}\right] \\ &= \mathbf{E}\left[\mathbf{E}\{|X|^p \mid \mathcal{G}\}\,\mathbf{1}_{[\Omega]}\right] = \mathbf{E}\left[|X|^p \mathbf{1}_{[\Omega]}\right] = \|X\|_p^p. \qquad \square\end{aligned}$$

(xi) **Conditional Hölder's inequality.** For $p, q \geq 1$ with $1/p + 1/q = 1$. If $X \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ and $Y \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ then $XY \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ and

$$\mathbf{E}\{|XY| \mid \mathcal{G}\} \leq (\mathbf{E}\{|X|^p \mid \mathcal{G}\})^{1/p} (\mathbf{E}\{|X|^q \mid \mathcal{G}\})^{1/q}.$$

We briefly delay the proof of Hölder's inequality as it uses a property of conditional expectation we have not yet seen.

The next three properties are perhaps less "intuitive", as they are not simply conditional versions of facts you have already seen. The first is related to the fact that the projection operation is idempotent. The second says that if $\sigma(Y) \subset \mathcal{G}$ then Y "acts like a constant" with respect to conditional expectations given $\mathcal{G}$. The third says (informally) that conditioning a conditional expectation on another independent σ-field doesn't change anything.

(xii) **The tower property.** If $\mathcal{H} \subset \mathcal{G}$ is another σ-field, then

$$\mathbf{E}\{\mathbf{E}\{X \mid \mathcal{G}\} \mid \mathcal{H}\} \overset{\text{a.s.}}{=} \mathbf{E}\{X \mid \mathcal{H}\} \overset{\text{a.s.}}{=} \mathbf{E}\{\mathbf{E}\{X \mid \mathcal{H}\} \mid \mathcal{G}\}.$$

Proof: First, $\mathbf{E}\{X \mid \mathcal{H}\}$ is $\mathcal{H}/\mathcal{B}(\mathbb{R})$-measurable, and $\mathcal{H} \subset \mathcal{G}$, so by property (ii),

$$\mathbf{E}\{\mathbf{E}\{X \mid \mathcal{H}\} \mid \mathcal{G}\} \overset{\text{a.s.}}{=} \mathbf{E}\{X \mid \mathcal{H}\}.$$

Next, let Z be a version of $\mathbf{E}\{\mathbf{E}\{X \mid \mathcal{G}\} \mid \mathcal{H}\}$. Then by definition, $Z \in L_1(\Omega, \mathcal{H}, \mathbf{P})$ and for all $A \in \mathcal{H}$,

$$\mathbf{E}\left[Z\mathbf{1}_{[A]}\right] = \mathbf{E}\left[\mathbf{E}\{X \mid \mathcal{G}\}\,\mathbf{1}_{[A]}\right].$$

By the definition of $\mathbf{E}\{X \mid \mathcal{G}\}$, we also have $\mathbf{E}\left[\mathbf{E}\{X \mid \mathcal{G}\}\,\mathbf{1}_{[A]}\right] = \mathbf{E}\left[X\mathbf{1}_{[A]}\right]$. It follows that $\mathbf{E}\left[Z\mathbf{1}_{[A]}\right] = \mathbf{E}\left[X\mathbf{1}_{[A]}\right]$, so Z is a version of $\mathbf{E}\{X \mid \mathcal{H}\}$. $\square$

(xiii) **Moving variables out of conditional expectations.** For any random variable $Z \in L_\infty(\Omega, \mathcal{G}, \mathbf{P})$ it holds that $\mathbf{E}\{XZ \mid \mathcal{G}\} \overset{\text{a.s.}}{=} \mathbf{E}\{X \mid \mathcal{G}\}\,Z$.
Proof: Let

$$\mathcal{S} = \left\{Z \in L_\infty(\Omega, \mathcal{G}, \mathbf{P}) : \mathbf{E}\{XZ \mid \mathcal{G}\} \overset{\text{a.s.}}{=} \mathbf{E}\{X \mid \mathcal{G}\}\,Z\right\}.$$

We aim to show that $\mathcal{S} = L_\infty(\Omega, \mathcal{G}, \mathbf{P})$. First suppose $Z = \mathbf{1}_{[B]}$ for some $B \in \mathcal{G}$. Then for all $A \in \mathcal{G}$,

$$\mathbf{E}\left[\mathbf{E}\{X \mid \mathcal{G}\}\,Z\mathbf{1}_{[A]}\right] = \mathbf{E}\left[\mathbf{E}\{X \mid \mathcal{G}\}\,\mathbf{1}_{[A\cap B]}\right] = \mathbf{E}\left[X\mathbf{1}_{[A\cap B]}\right] = \mathbf{E}\left[XZ\mathbf{1}_{[A]}\right],$$

so by the definition of conditional expectation, $\mathbf{E}\{X \mid \mathcal{G}\} Z$ is a version of $\mathbf{E}\{XZ \mid \mathcal{G}\}$ and therefore $Z \in \mathcal{S}$.

Next, if $Z, Z' \in \mathcal{S}$ and $a \in \mathcal{R}$ then by linearity of conditional expectation,

$$\begin{aligned}\mathbf{E}\{X \mid \mathcal{G}\}(aZ + Z') &= a\mathbf{E}\{X \mid \mathcal{G}\} Z + \mathbf{E}\{X \mid \mathcal{G}\} Z' \\ &\stackrel{\text{a.s.}}{=} \mathbf{E}\left\{aXZ + XZ' \;\middle|\; \mathcal{G}\right\} = \mathbf{E}\left\{X(aZ + Z') \;\middle|\; \mathcal{G}\right\},\end{aligned}$$

so $aZ + Z' \in \mathcal{S}$.

Next, if $0 \le Z_n \in S$ for $n \ge 1$ and $Z_n \uparrow Z \in \mathcal{S}$ as $n \to \infty$, then $X^+ Z_n \uparrow X^+ Z$ as $n \to \infty$, so by the conditional monotone convergence theorem,

$$\mathbf{E}\left\{X^+ \;\middle|\; \mathcal{G}\right\} Z = \lim_{n\to\infty} \mathbf{E}\left\{X^+ \;\middle|\; \mathcal{G}\right\} Z_n \stackrel{\text{a.s.}}{=} \lim_{n\to\infty} \mathbf{E}\left\{X^- Z_n \;\middle|\; \mathcal{G}\right\} \stackrel{\text{a.s.}}{=} \mathbf{E}\left\{X^- Z \;\middle|\; \mathcal{G}\right\}.$$

Likewise $\mathbf{E}\{X^- \mid \mathcal{G}\} Z \stackrel{\text{a.s.}}{=} \mathbf{E}\{X^- Z \mid \mathcal{G}\}$, so by linearity of conditional expectation,

$$\mathbf{E}\{X \mid \mathcal{G}\} Z \stackrel{\text{a.s.}}{=} (\mathbf{E}\left\{X^+ \;\middle|\; \mathcal{G}\right\} + \mathbf{E}\left\{X^- \;\middle|\; \mathcal{G}\right\}) Z \stackrel{\text{a.s.}}{=} \mathbf{E}\left\{X^+ Z + X^- Z \;\middle|\; \mathcal{G}\right\} = \mathbf{E}\{XZ \mid \mathcal{G}\}.$$

Thus $Z \in \mathcal{S}$. It follows by the monotone class theorem that $\mathcal{S} = L_\infty(\Omega, \mathcal{G}, \mathbf{P})$. □

(xiv) **Adding an independent conditioning changes nothing.** If $\mathcal{H} \subset \mathcal{F}$ is another σ-algebra and $\sigma(X, \mathcal{G}) := \sigma(\sigma(X) \cup \mathcal{G})$ is independent of $\mathcal{H}$ then $\mathbf{E}\{X \mid \sigma(\mathcal{G}, \mathcal{H})\} = \mathbf{E}\{X \mid \mathcal{G}\}$.

The last property is left as an exercise. We also state the following strengthening of (xiii) as an exercise.

Exercise 10.3. *Prove that if $Z : \Omega \to \mathbb{R}$ is $\mathcal{G}/\mathcal{B}(\mathbb{R})$-measurable and $X, XZ \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ then* $\mathbf{E}\{XZ \mid \mathcal{G}\} \stackrel{\text{a.s.}}{=} \mathbf{E}\{X \mid \mathcal{G}\} Z$.

Proof of Holder's inequality. To avoid issues of integrability and dividing by zero, for $\epsilon \ge 0$ write $U_\epsilon = (\mathbf{E}\{|X|^p \mid \mathcal{G}\} + \epsilon)^{1/p}$ and $V_\epsilon = (\mathbf{E}\{|Y|^q \mid \mathcal{G}\} + \epsilon)^{1/q}$. Then let $X_\epsilon = \frac{|X|}{U_\epsilon}$ and $Y_\epsilon = \frac{Y}{U_\epsilon}$.

For $\epsilon > 0$ we then have

$$X_\epsilon Y_\epsilon = \exp\left(\frac{1}{p}\log(X_\epsilon^p) + \frac{1}{q}\log(Y_\epsilon^q)\right) \le \exp\left(\log\frac{X_\epsilon^p}{p} + \frac{Y_\epsilon^q}{q}\right) = \frac{X_\epsilon^p}{p} + \frac{Y_\epsilon^q}{q},$$

so

$$\begin{aligned}\mathbf{E}\{X_\epsilon Y_\epsilon \mid \mathcal{G}\} &\le \frac{1}{p}\mathbf{E}\{X_\epsilon^p \mid G\} + \frac{1}{p}\mathbf{E}\{Y_\epsilon^q \mid G\} \\ &= \frac{1}{p}\mathbf{E}\left\{|X|^p U_\epsilon^{-p} \;\middle|\; \mathcal{G}\right\} \frac{1}{q}\mathbf{E}\left\{|Y|^q V_\epsilon^{-q} \;\middle|\; \mathcal{G}\right\}\end{aligned}$$

The terms U_ϵ^{-p} and V_ϵ^{-q} are in $L_\infty(\Omega, \mathcal{G}, \mathbf{R})$, so by (xiii) we may move them outside the conditional expectations. The previous bound then beomes

$$\mathbf{E}\{X_\epsilon Y_\epsilon \mid \mathcal{G}\} \stackrel{\text{a.s.}}{\le} \frac{1}{p}\frac{\mathbf{E}\{|X|^p \mid \mathcal{G}\}}{U_\epsilon^p} + \frac{1}{q}\frac{\mathbf{E}\{|Y|^q \mid \mathcal{G}\}}{V_\epsilon^q} = \frac{1}{p}\frac{U_0^p}{U_\epsilon^p} + \frac{1}{q}\frac{V_0^q}{V_\epsilon^q} \le 1.$$

Again using that U_ϵ^{-p} and V_ϵ^{-q} are in $L_\infty(\Omega, \mathcal{G}, \mathbf{R})$, we also have

$$\mathbf{E}\{X_\epsilon Y_\epsilon \mid \mathcal{G}\} \stackrel{\text{a.s.}}{=} \frac{\mathbf{E}\{|XY| \mid \mathcal{G}\}}{U_\epsilon V_\epsilon},$$

which combined with the previous inequality gives

$$\mathbf{E}\{|XY| \mid \mathcal{G}\} \stackrel{\text{a.s.}}{\le} U_\epsilon V_\epsilon.$$

Taking $\epsilon \downarrow 0$, the result follows. □

10.2. **Conditional expectations, tightness and uniform integrability.** Fix a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a collection $\mathrm{X} = (X_i, i \in I)$ of random variables in $L_1(\Omega, \mathcal{F}, \mathbf{P})$.

Write μ_i for the distribution of X_i. The family $(\mu_i, i \in I)$ of probability measures is *tight* if for all $\epsilon > 0$ there is a compact set $K \subset \mathbb{R}$ tight

$$\sup_{i \geq 1} \mu_i(\mathbb{R} \setminus K) < \epsilon.$$

uniformly integrable The collection X is *uniformly integrable* with respect to $\mathbf{P}$ if for all $\epsilon > 0$ there is a compact set $K \subset \mathbb{R}$ such that

$$\sup_{i \in I} \mathbf{E}\left[|X_i| \mathbf{1}_{[|X_i| \notin K]}\right] < \epsilon.$$

The two conditions are syntactically similar. They are connected by using the *size-biasing* operation introduced earlier. Write $\hat{\mu}_i$ for the size-biasing of μ_i, so

$$\hat{\mu}_i(B) = \left(\mu_i \cdot \frac{|X_i|}{\mathbf{E}|X_i|}\right)(B) = \mathbf{E}\left[|X_i| \mathbf{1}_{[X_i \in B]}\right]$$

Then X is uniformly integrable if and only if $(\hat{\mu}_i, i \in I)$ is tight (exercise).

Exercise 10.4. *Let $(\mu_n, n \geq 1)$ be a tight family of probability measures. Then there exists a subsequence $(n_k, k \geq 1)$ such that μ_{n_k} converges in distribution. (I.e. such that if X_{n_k} has distribution μ_{n_k} then X_{n_k} converges in distribution.)*

The next proposition connects uniform integrability and conditional expectations, and is the first step toward martingales and martingale convergence theorems.

Proposition 10.3. *Fix a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a random variable $X \in L_1(\Omega, \mathcal{F}, \mathbf{P})$. Then $\{\mathbf{E}\{X \mid \mathcal{G}\} : \mathcal{G} \subset \mathcal{F}$ a sub-σ-field$\}$ is a uniformly integrable collection of random variables.*

Lemma 10.4. *If $X \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ then for all $\epsilon > 0$ there is $\delta > 0$ such that for all $A \in \mathcal{F}$, if $\mathbf{P}\{A\} \leq \delta$ then $\mathbf{E}\left[|X| \mathbf{1}_{[A]}\right] < \epsilon$.*

Proof. Suppose that the assertion of the lemma is false. Then we may find $\epsilon > 0$ and events $(A_n, n \geq 1)$ in $\mathcal{F}$ with $\mathbf{P}\{A_n\} \leq 2^{-n}$ such that $\mathbf{E}\left[|X| \mathbf{1}_{[A_n]}\right] \geq \epsilon$ for all n.

We now show this implies that $\mathbf{E}\left[|X| \mathbf{1}_{[A_n \text{ i.o.}]}\right] \geq \epsilon$. By definition, $\{A_n \text{ i.o.}\} = \bigcap_{n \geq 1} \bigcup_{m \geq n} A_m$, so $\mathbf{1}_{[A_n \text{ i.o.}]} = \mathbf{1}_{[\bigcap_{n \geq 1} \bigcup_{m \geq n} A_m]} = \lim_{n \to \infty} \mathbf{1}_{[\bigcup_{m \geq n} A_m]}$

For any event $E \in \mathcal{F}$ we have $|X| \mathbf{1}_{[E]} \leq |X|$, so by the dominated convergence theorem,

$$\mathbf{E}\left[X \mathbf{1}_{[A_n \text{ i.o.}]}\right] = \mathbf{E}\left[\lim_{n \to \infty} |X| \mathbf{1}_{[\bigcup_{m \geq n} A_m]}\right] = \lim_{n \to \infty} \mathbf{E}\left[|X| \mathbf{1}_{[\bigcup_{m \geq n} A_m]}\right] \geq \epsilon\,.$$

On the other hand, $\sum_{n \geq 1} \mathbf{P}\{A_n\} = 1 < \infty$, so by the first Borel-Cantelli lemma, $\mathbf{P}\{A_n \text{ i.o.}\} = 0$ and thus $\mathbf{E}\left[X \mathbf{1}_{[A_n \text{ i.o.}]}\right] = 0$, a contradiction. □

Proof of Proposition 10.3. Fix $\epsilon > 0$ and let $\delta > 0$ be such that $\mathbf{E}\left[|X| \mathbf{1}_{[A]}\right] < \epsilon$ whenever $\mathbf{P}\{A\} \leq \delta$; such δ exists by the lemma. Then for any sub-σ-field $\mathcal{G} \subset \mathcal{F}$, by the conditional Jensen's inequality

$$|\mathbf{E}\{X \mid \mathcal{G}\}| \leq \mathbf{E}\{|X| \mid \mathcal{G}\},$$

so

$$\mathbf{E}\left[|\mathbf{E}\{X \mid \mathcal{G}\}|\right] \leq \mathbf{E}\left[\mathbf{E}\{|X| \mid \mathcal{G}\}\right] = \mathbf{E}|X|,$$

Taking $K = [-\mathbf{E}|X|/\delta, \mathbf{E}|X|/\delta]$, it follows that

$$\begin{aligned} \mathbf{P}\{|\mathbf{E}\{X \mid \mathcal{G}\}| \notin K\} &= \mathbf{P}\{|\mathbf{E}\{X \mid \mathcal{G}\}| > \mathbf{E}|X|/\delta\} \\ &\leq \mathbf{E}\left[|\mathbf{E}\{X \mid \mathcal{G}\}|\right] \mathbf{E}|X|/\delta \\ &\leq \delta. \end{aligned}$$

It follows that

$$\begin{aligned}\mathbf{E}\left[|\mathbf{E}\{X \mid \mathcal{G}\}|\mathbf{1}_{[|\mathbf{E}\{X \mid \mathcal{G}\}|\notin K]}\right] &\leq \mathbf{E}\left[\mathbf{E}\{|X| \mid \mathcal{G}\}\mathbf{1}_{[|\mathbf{E}\{X \mid \mathcal{G}\}|\notin K]}\right] \\ &= \mathbf{E}\left[\mathbf{E}\left\{|X|\mathbf{1}_{[|\mathbf{E}\{X \mid \mathcal{G}\}|\notin K]} \;\middle|\; \mathcal{G}\right\}\right] \\ &= \mathbf{E}\left[|X|\mathbf{1}_{[|\mathbf{E}\{X \mid \mathcal{G}\}|\notin K]}\right] \\ &\leq \epsilon\,,\end{aligned}$$

the last bound holding since $\mathbf{P}\{|\mathbf{E}\{X \mid \mathcal{G}\}| \notin K\} \leq \delta$. □

11. **Martingales**

A stochastic process is simply a family of random variables $(X_i, i \in I)$ defined over a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Martingales are stochastic processes which model "fair games", or random systems which evolve in time without a bias in any particular direction. They are one of the most important general classes of stochastic processes; the next part of these notes is devoted to defining martingales and understanding their properties.

A *filtration* is an increasing sequence of σ-algebras $(\mathcal{F}_n)_{n\geq 0}$ over a common ground set. A *filtered probability space* is a tuple $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbf{P})$, where $(\mathcal{F}_n)_{n\geq 0}$ is a filtration over Ω and $\mathcal{F}_n \subset \mathcal{F}$ for all $n \geq 0$.

A sequence $\mathrm{X} = (X_n)_{n\geq 0}$ of random variables is $(\mathcal{F}_n)$-*adapted* if X_n is $\mathcal{F}_n/\mathcal{B}(\mathbb{R})$-measurable for all $n \geq 0$. It is *integrable* if $X_n \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ for all $n \geq 1$. Finally, it is an $(\mathcal{F}_n)$-*martingale* (or just a martingale for short) if it is integrable and adapted and satisfies the *martingale property*: for all $n > 0$,

$$\mathbf{E}\{X_n \mid \mathcal{F}_{n-1}\} = X_{n-1}.$$

If you think of $(X_n)_{n\geq 1}$ as a stock value (for example), then the martingale property states that the best prediction for the stock's future value given its past performance is simply its present value.[15]

Example: Simple random walk. Let $(Z_i, i \geq 1)$ be iid random variables in $L_1(\Omega, \mathcal{F}, \mathbf{P})$ with $\mathbf{E}Z_1 = 0$ and let $X_n = Z_1 + \ldots + Z_n$. Then with

$$\mathcal{F}_n = \sigma(Z_1, \ldots, Z_n) = \sigma(X_1, \ldots, X_n),$$

the sequence $\mathrm{X} = (X_n, n \geq 0)$ is an $(\mathcal{F}_n)$-martingale.

Example: branching processes. This example is developed to some extent in another set of notes, available on the website. I will expand on the martingale connection here when I have the chance.

Exercise 11.1. *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbf{P})$ be a filtered probability space and let $X \in L_1(\Omega, \mathcal{F}, \mathbf{P})$. Write $X_n \overset{\text{a.s.}}{=} \mathbf{E}\{X \mid \mathcal{F}_n\}$. Show that $(X_n, n \geq 0)$ is a martingale relative to the filtration $(\mathcal{F}_n, n \geq 0)$.*

The main goal of the section is to find conditions which guarantee that a martingale $(X_n)_{n\geq 0}$ converges to some limit X in some sense. However, the convergence theory is not the only point and, in fact, the theory will be easier to understand and will appear better motivated if we first approach the subject from a more applied point of view.

Continuing the analogy with stock prices, suppose that $\mathrm{X} = (X_n)_{n\geq 0}$ is an $(\mathcal{F}_n)$-*adapted* process, and think of it as tracking a stock price over time. At time n you can choose to invest some amount money C_{n+1} for the next unit of time, based on your observation of the stock's behaviour to date. In the next unit of time your profit/loss will then be $C_{n+1}(X_{n+1} - X_n)$.

An integrable stochastic process $(C_n)_{n\geq 1}$ is $(\mathcal{F}_n)$-*previsible* if $C_{n+1} \in L_1(\Omega, \mathcal{F}_n, \mathbf{P})$ for all $n \geq 0$. In the stock market analogy, saying that that C_{n+1} should be chosen based on past observations precisely means that means that $(C_n)_{n\geq 1}$ should be $(\mathcal{F}_n)$-previsible. If this is the case, then by the

[15] The efficient markets hypothesis in economics is essentially a statement that stocks behave like martingales.

properties of conditional expectation (and assuming the random variables C_n are bounded), the profit/loss in the time unit from n to $n+1$ is

$$\begin{aligned}\mathbf{E}\left[C_{n+1}(X_{n+1}-X_n)\right] &= \mathbf{E}\left[\mathbf{E}\left\{C_{n+1}(X_{n+1}-X_n) \mid \mathcal{F}_n\right\}\right] \\ &= \mathbf{E}\left[\mathbf{E}\left\{C_{n+1}X_{n+1} \mid \mathcal{F}_n\right\}\right] - \mathbf{E}\left[\mathbf{E}\left\{C_{n+1}X_n \mid \mathcal{F}_n\right\}\right] \\ &= \mathbf{E}\left[C_{n+1}\mathbf{E}\left\{X_{n+1} \mid \mathcal{F}_n\right\}\right] - \mathbf{E}\left[C_{n+1}X_n\right] .\end{aligned}$$

We've used that C_{n+1} and X_n are $(\mathcal{F}_n/\mathcal{B}(\mathbb{R}))$-measurable to extract them from the conditional expectation. If X is an $(\mathcal{F}_n)$-martingale then by the martingale property, the last line equals zero, which means that gambling on this stock yields no expected profit or loss.

In the above setup, the total profit/loss by time n is

$$\sum_{i=1}^{n} C_i(X_i - X_{i-1})$$

which is our first glimpse at stochastic integration; it looks like a discrete analogue of an integral $\int_0^n X_i \mathrm{d}C_i$. This perspective has been fruitfully developed into an entire academic discipline.

The theory of martingales is, among other things, a computational tool. Basic facts about martingales allow some expected values to be identified by appeal to general theory rather than via ad hoc calculations. For example, imagine that $(R_n)_{n\geq 0}$ tracks the dollar value of your current bankroll[16] in a gambling game. You may choose to stop gambling at the first time T that either $R_n \geq 1000$ or $R_n = 0$. You will then return home with R_T dollars, and may care to know the expected value $\mathbf{E}[R_T]$. The *optional stopping theorem* says that, if you were playing a fair game, then $\mathbf{E}[R_T] = R_0$; you expect to walk out with whatever you brought in. Of course, most casinos don't offer fair games. (If you are inclined to split hairs[17], there are other issues with this as a model for gambling play; what is the meaning of R_n for $n > T$, for example?)

To state the optional stopping theorem we first need to define stopping times, and take the opportunity to state some elementary facts about them. Given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbf{P})$

stopping time

a random variable $T : \Omega \to \mathbb{N} \cup \{+\infty\}$ is an $(\mathcal{F}_n)$-stopping time (or just "stopping time") if for all $n \in \mathbb{N}$, the event $\{T \leq n\} \in \mathcal{F}_n$. The idea to have in mind is, if $\mathcal{F}_n$ is the information available to you at time n, then saying T is a stopping time means that, if you are trying to stop at time T, then enough information is available to you that you will know when to stop (gambling, riding the bus, owning a stock,...).

In the gambling example from two paragraphs ago, we could take $T^* = \inf\{n : R_n \in \mathbb{R} \setminus (0, 1000)\}$, which could also be written as $T^* = \min(T_0, T_1)$, where $T_0 = \inf\{n : R_n \leq 0\}$ and $T_1 = \inf\{n : R_n \geq 1000\}$. All three of T_0, T_1 and T^* are stopping times. An example of a non-stopping time would be this: "I'll play for 100 rounds and stop whenever my bankroll is largest". This corresponds to the random variable $T_2 = \arg\max(R_n, 0 \leq n \leq 100)$; [18] but to stop at time T_2 would require foreknowledge of $(R_n, T_2 \leq n \leq 100)$. Laws against insider trading are in a sense legislating that decisions about when to buy and sell stocks must be stopping times.

Exercise 11.2. *Show that T_0, T_1 and T^* defined above are all stopping times with respect to the filtration $\mathcal{F}_n = \sigma(R_m, 0 \leq m \leq n)$.*

Given an $(\mathcal{F}_n)$-stopping time T, we define the *stopped σ-field* as follows: let $\mathcal{F}_\infty = \sigma(\bigcup_{n\geq 0} \mathcal{F}_n)$, and let

$$\mathcal{F}_T := \{A \in \mathcal{F}_\infty : \forall n \geq 0, A \cap \{T \leq n\} \in \mathcal{F}_n\}.$$

For example, the event A that (R_n) first exceeds 1000 before first reaching 0 is in $\mathcal{F}_{T^*}$, since (informally) at time T^* we know which of 0 and 1000 was first reached by (R_n). The next exercise is

[16] Bankroll, n. originally and chiefly U.S. A roll of banknotes; (in extended use) the money a person possesses; funds, financial resources; (Gambling) the amount of money a person sets aside for a given session or period of gambling. Frequently with possessive. –Oxford English Dictionary

[17] Couper les cheveux en quatre (fr)/Fendre les cheveux en quatre (qc)/S'enfarger dans les fleurs du tapis (qc)

[18] Given a finite collection $(x_i, i \in I)$ of real numbers, $\arg\max(x_i, i \in I)$ returns the value of i for which x_i is largest.

a special case of a fact from the subsequent proposition, but is perhaps worth doing separately to make sure you're comfortable with these basic ideas.

Exercise 11.3. *Let $A = \{T_1 \leq T_0\}$. Show that A is in $\mathcal{F}_{T_0}, \mathcal{F}_{T_1}$, and $\mathcal{F}_{T^*}$.*

Proposition 11.1 (Basic facts about stopping times)**.** *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbf{P})$ be a filtered probability space, let $(X_n)_{n\geq 0}$ be an $(\mathcal{F}_n)$-adapted process, and let S, T be two $(\mathcal{F}_n)$-stopping times. Then the following all hold.*

(1) $\min(S, T)$ *is a stopping time.*
(2) $\mathcal{F}_T$ *is a σ-field.*
(3) If $S \leq T$ then $\mathcal{F}_S \subseteq \mathcal{F}_T$.
(4) $X_T \mathbf{1}_{[T<\infty]}$ *is $\mathcal{F}_T/\mathcal{B}(\mathbb{R})$-measurable.*
(5) The process $(X_{\min(T,n)}, n \geq 0)$ is $(\mathcal{F}_n)$-adapted.[19]
(6) If $(X_n)_{n\geq 0}$ is integrable then $(X_{\min(T,n)})_{n\geq 0}$ is integrable.

The proofs of the facts stated in the proposition are left as exercises.

An $(\mathcal{F}_n)$-adapted integrable process $(X_n)_{n\geq 0}$ is a *supermartingale* if $\mathbf{E}\{X_n \mid \mathcal{F}_m\} \overset{\text{a.s.}}{\leq} X_m$ for all $0 \leq m \leq n$. It is a *submartingale* if $\mathbf{E}\{X_n \mid \mathcal{F}_m\} \overset{\text{a.s.}}{\geq} X_m$ for all $0 \leq m \leq n$. You might expect the inequalities to go the other way; in its current form it is more in line with the definitions of sub/superharmonic functions, but this is hard to explain rigorously without a large digression. So for the time being you'll just have to find your own way to remember.

Submartingale, supermartingale

Theorem 11.2 (Optional stopping theorem)**.** *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbf{P})$ be a filtered probability space and let $(X_n)_{n\geq 0}$ be an $\mathcal{F}_n$-supermartingale. Then for any bounded stopping times $0 \leq S \leq T$, it holds that $\mathbf{E}X_T \leq \mathbf{E}X_S$.*

Before proving the theorem, we note its immediate corollary for martingales.

Corollary 11.3. *Suppose $(X_n)_{n\geq 0}$ is in fact an $\mathcal{F}_n$-martingale. Then for any bounded stopping times $0 \leq S \leq T$, it holds that $\mathbf{E}X_T = \mathbf{E}X_S$.*

To prove the corollary, note that if $(X_n)_{n\geq 0}$ is a martingale then both $(X_n)_{n\geq 0}$ and $(-X_n)_{n\geq 0}$ are supermartingales; then apply Theorem 11.2. The optional stopping theorem is a consequence of the following theorem, which lists three necessary and sufficient conditions for an adapted integrable process to be a supermartingale.

Theorem 11.4. *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbf{P})$ be a filtered probability space and let $(X_n)_{n\geq 0}$ be an $(\mathcal{F}_n)$-adapted integrable process. Then the following are equivalent.*

(a) $(X_n)_{n\geq 0}$ *is an $(\mathcal{F}_n)$-supermartingale.*
(b) For any bounded $(\mathcal{F}_n)$-stopping time T and any $(\mathcal{F}_n)$-stopping time S, $\mathbf{E}\{X_T \mid \mathcal{F}_S\} \overset{\text{a.s.}}{\leq} X_{\min(S,T)}$.
(c) For any $(\mathcal{F}_n)$-stopping time T, the process $(X_{\min(T,n)})_{n\geq 0}$ is an $(\mathcal{F}_n)$-supermartingale.
(d) For any bounded $(\mathcal{F}_n)$-stopping times S and T with $S \overset{\text{a.s.}}{\leq} T$, $\mathbf{E}X_T \leq \mathbf{E}X_S$.

Proof. **[(c) ⇒ (a)].** Let T be the constant function which is identically equal to n. Then $X_{\min(T,n)} = X_n$ and $X_{\min(T,n-1)} = n-1$, so by the assumption in (c),

$$\mathbf{E}\{X_n \mid \mathcal{F}_{n-1}\} = \mathbf{E}\left\{X_{\min(T,n)} \;\middle|\; \mathcal{F}_{n-1}\right\} \overset{\text{a.s.}}{\leq} X_{\min(T,n-1)} = X_{n-1},$$

so X is an $(\mathcal{F}_n)$-supermartingale.
[(b) ⇒ (c)]. Let T be a stopping time, fix $n \geq 1$ and take $S \equiv n-1$. Then $\mathcal{F}_S = \mathcal{F}_{n-1}$ (exercise), and $\min(T, n)$ is a bounded stopping time, so by (b),

$$\mathbf{E}\left\{X_{\min(T,n)} \;\middle|\; \mathcal{F}_{n-1}\right\} = \mathbf{E}\left\{X_{\min(T,n)} \;\middle|\; \mathcal{F}_S\right\} \overset{\text{a.s.}}{\leq} X_{\min(\min(T,n),S)} = X_{\min(T,n-1)}, \overset{\text{a.s.}}{\leq}$$

so $(X_{\min(T,n)})_{n\geq 0}$ is an $(\mathcal{F}_n)$-supermartingale.

[19] Hairs unsplit.

[(b) $\Rightarrow$ (d)]. If S and T are both bounded stopping times with $S \overset{\text{a.s.}}{\leq} T$ then (b) gives us that

$$\mathbf{E}\{X_T \mid \mathcal{F}_S\} \overset{\text{a.s.}}{\leq} X_{\min(S,T)} \overset{\text{a.s.}}{=} X_S.$$

Taking expectations on both sides, it follows that $\mathbf{E}X_T \leq \mathbf{E}X_S$.

[(a) $\Rightarrow$ (b)]. Suppose $(X_n)_{n\geq 0}$ is a supermartingale, let S be a stopping time and T be a bounded stopping time, and choose $n \in \mathbb{N}$ such that $\mathbf{P}\{T \leq n\} = 1$. Then we can write

$$X_T = X_{\min(S,T)} + \sum_{k=0}^{n} (X_{k+1} - X_k)\mathbf{1}_{[S\leq k<T]}.$$

For any event $A \in \mathcal{F}_S$ and $k \in \mathbb{N}$, by definition we have $A \cap \{S \leq k\} \in \mathcal{F}_k$. Also, $\{T \leq k\} \in \mathcal{F}_k$ so $\{T > k\} \in \mathcal{F}_k$. Using this measurability together with the tower law, and then using supermartingale property, it follows that

$$\begin{aligned}
\mathbf{E}\left[(X_{k+1} - X_k)\mathbf{1}_{[S\leq k<T]}\mathbf{1}_{[A]}\right] &= \mathbf{E}\left[\mathbf{E}\left\{(X_{k+1} - X_k)\mathbf{1}_{[T>k]}\mathbf{1}_{[A\cap\{S\leq k\}]} \;\middle|\; \mathcal{F}_k\right\}\right] \\
&= \mathbf{E}\left[\mathbf{E}\{X_{k+1} - X_k \mid \mathcal{F}_k\}\mathbf{1}_{[T>k]}\mathbf{1}_{[A\cap\{S\leq k\}]}\right] \\
&\leq \mathbf{E}\left[0 \cdot \mathbf{1}_{[T>k]}\mathbf{1}_{[A\cap\{S\leq k\}]}\right] \\
&= 0\,.
\end{aligned}$$

Combined with the previous identity for X_T, it follows that for any event $A \in \mathcal{F}_S$,

$$\mathbf{E}\left[X_T\mathbf{1}_{[A]}\right] \leq \mathbf{E}\left[X_{\min(S,T)}\mathbf{1}_{[A]}\right]\,.$$

From this and the definition of conditional expectation, it follows that

$$\mathbf{E}\left[\mathbf{E}\{X_T \mid \mathcal{F}_S\}\mathbf{1}_{[A]}\right] = \mathbf{E}\left[X_T\mathbf{1}_{[A]}\right] \leq \mathbf{E}\left[X_{\min(S,T)}\mathbf{1}_{[A]}\right]\,.$$

Since both $\mathbf{E}\{X_T \mid \mathcal{F}_S\}$ and $X_{\min(S,T)}$ are $\mathcal{F}_S/\mathcal{B}(\mathbb{R})$-measurable, it follows that $\mathbf{E}\{X_T \mid \mathcal{F}_S\} \overset{\text{a.s.}}{\leq} X_{\min(S,T)}$, so (b) holds.

[(d) $\Rightarrow$ (a)]. We must show that for all $n \geq 1$

$$\mathbf{E}\{X_n \mid \mathcal{F}_{n-1}\} \overset{\text{a.s.}}{\leq} X_{n-1}$$

To establish this inequality, it suffices to show that for any event $A \in \mathcal{F}_{n-1}$,

$$\mathbf{E}\left[\mathbf{E}\{X_n \mid \mathcal{F}_{n-1}\}\mathbf{1}_{[A]}\right] \leq \mathbf{E}\left[X_{n-1}\mathbf{1}_{[A]}\right]\,.$$

So fix $n \geq 1$ and $A \in \mathcal{F}_{n-1}$. Let $T \equiv n$ and let $S = (n-1)\mathbf{1}_{[A]} + n\mathbf{1}_{[A^c]}$. It is not hard to see that S is a stopping time exercise. Since $S \leq T$, it follows from (d) that

$$\mathbf{E}X_n = \mathbf{E}X_T \leq \mathbf{E}X_S.$$

But $X_S = X_{n-1}\mathbf{1}_{[A]} + X_n\mathbf{1}_{[A^c]}$, so this gives

$$\mathbf{E}X_n \leq \mathbf{E}\left[X_{n-1}\mathbf{1}_{[A]}\right] + \mathbf{E}\left[X_n\mathbf{1}_{[A^c]}\right];$$

rearranging gives $\mathbf{E}\left[X_n\mathbf{1}_{[A]}\right] \leq \mathbf{E}\left[X_{n-1}\mathbf{1}_{[A]}\right]$. But by definition, $\mathbf{E}\left[X_n\mathbf{1}_{[A]}\right] = \mathbf{E}\left[\mathbf{E}\{X_n \mid \mathcal{F}_{n-1}\}\mathbf{1}_{[A]}\right]$, so the required inequality follows. □

List of notation and terminology

λ-system	A set $\mathcal{A} \subset 2^\Omega$ with $\Omega \in \mathcal{A}$, closed under monotone limits and relative complements. 6
$\mathcal{L}_X$	The distribution of random variable X. 2
Measure	A countably additive function $\mu : \mathcal{F} \to [0, \infty]$ on a σ-field, with $\mu(\emptyset) = 0$. 2
G_X	The moment generating function of X, $G_X(s) = \mathbf{E}\left[e^{-sX}\right]$. 39
μ_X	The distribution of random variable X. 16
Outer measure	A monotone, countably subadditive function $\mu : 2^\Omega \to [0, \infty]$ with domain the power set of some set Ω, such that $\mu(\emptyset) = 0$. 4
π-system	A collection of subsets of a ground set, closed under finite intersection. 2
Pre-measure	A σ-additive function $\mu : \mathcal{A} \to [0, \infty]$ on a ring $\mathcal{A}$ with $\mu(\emptyset) = 0$. 3
Pre-measure space	A triple $(\Omega, \mathcal{A}, \mu)$ where $\mathcal{A}$ is a ring over Ω and μ is a pre-measure on $\mathcal{A}$. . . 3
Probability space	A measure space $(\Omega, \mathcal{F}, \mathbf{P})$ with $\mathbf{P}(\Omega) = 1$. 10
Ring	A collection of subsets of a ground set, closed under finite union and relative complement. 2
$\sigma(\mathcal{A})$	The smallest σ-field containing $\mathcal{A}$. 2
σ-field	A collection of subsets of a ground set, closed under countable union and complement. 2
Stieltjes Function	A non-decreasing function $F : \mathbb{R} \to \mathbb{R}$ which is right continuous with left limits. 8

Department of Mathematics and Statistics, McGill University, Montréal, Canada.
Email address: louigi.addario@mcgill.ca
URL: http://problab.ca/louigi/