MA 361: Probability Theory

Naman Mishra

August 2024

Contents

Lectures

The course

Lecture 1.
Thursday
August 1

Grading

- Homework: 20%
- Two midterms: 15% each
- Final: 50%

Chapter I

Review of discrete probability

Definition I.1 (Discrete probability space)**.** A discrete probability space is a pair (Ω, p) where Ω is a finite or countable set called *sample space* and $p : \Omega \to [0, 1]$ is a function giving the *elementary probabilities* of each $\omega \in \Omega$ such that

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

Examples.

- "Toss a fair n times" is modeled as

  $$\Omega = \{0, 1\}^n$$

  with

  $$p(\omega) \equiv \frac{1}{2^n}.$$

- "Throw r balls randomly into m bins" is modeled as

  $$\Omega = [m]^r$$

  with p given by the multinomial distribution (assuming uniformity).

- "A box has N coupons, draw one of them."

  $$\Omega = [N]$$
  $$p = \omega \mapsto \frac{1}{N}.$$

- "Toss a fair coin countably many times." The set of outcomes is clear: $\Omega = \{0, 1\}^{\mathbb{N}}$. What about the elementary probabilities?

  Probabilities of some events are also fairly intuitive. For example, the event

  $$A = \{\underline{\omega} \in \Omega \mid \omega_1 = 1, \omega_2 = 1, \omega_3 = 0\}$$

  has probability $1/8$. Similarly $B = \{\underline{\omega} \in \Omega \mid \omega_1 = 1, \omega_2 = 0\}$ has probability $1/4$. Where does this come from?

What about this event:

$$C = \{\underline{\omega} \in \Omega \;\Big|\; \frac{1}{n}\sum_{i=1}^{n} \omega_i \to 0.6\}$$

What about:

$$D = \{\underline{\omega} \in \Omega \;\Big|\; \sum_{i=1}^{n} \omega_i = \frac{n}{2} \text{ for infinitely many } n\}^{1}$$

- "Draw a number uniformly at random from $[0,1]$." Ω is obviously $[0,1]$. Again some events have obvious probabilities.

  $$A = [0.1, 0.3] \implies \mathbf{P}(A) = 0.2$$

  Similarly

  $$B = [0.1, 0.2] \cup (0.7, 1) \implies \mathbf{P}(B) = 0.4$$

  What about $C = \mathbb{Q} \cap [0,1]$? What about D, the $\frac{1}{3}$-Cantor set?

  The $\frac{1}{3}$-Cantor set is given by the limit of the following sequence of sets.

  $$\begin{aligned} K_0 &= [0,1] \\ K_1 &= [0,1/3] \cup [2/3,1] \\ K_2 &= [0,1/9] \cup [2/9,1/3] \cup [2/3,7/9] \cup [8/9,1] \\ &\vdots \end{aligned}$$

  where each K_{n+1} is obtained by removing the middle third of each interval in K_n.[2]

The resolution for the above examples is achieved by taking the 'obvious' cases as definitions.

What we agree on:

(∗) $\mathbf{P}([a,b]) = b - a$ for all $0 \le a \le b \le 1$.

What we wish for:

(#1) If $A \cap B = \varnothing$, then $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$.

(#2) If $A_n \downarrow A$, then $\mathbf{P}(A_n) \downarrow \mathbf{P}(A)$.

Question: Does there exist a $\mathbf{P}\colon 2^{[0,1]} \to [0,1]$ that satisfies (∗), (#1) and (#2)? **No.**

Question: Does there exist a $\mathbf{P}\colon 2^{[0,1]} \to [0,1]$ that satisfies (∗), (#1) and even *translational invariance*? **Yes!**
However, it is not unique.

[1] $\mathbf{P}(C) = 0$ and $\mathbf{P}(D) = 1$.
[2] $\mathbf{P}(C) = \mathbf{P}(D) = 0$.

What about the same for a probability measure on $[0,1]^2$ that is translation and rotation invariant?
What about $[0,1]^3$?[3]

Lack of uniqueness is a disturbing issue. The way out is the following: restrict the class of sets on which $\mathbf{P}$ is defined to a σ-algebra.

[3]The Banach-Tarski paradox gives a "no" for the 3D case.

Chapter II

Measure-theoretic probability

II.1 σ-algebras

Definition II.1 (σ-algebra)**.** Given a set Ω, a collection $\mathcal{F} \subseteq 2^\Omega$ is called a σ-algebra if

(ς1) $\varnothing \in \mathcal{F}$.

(ς2) $A \in \mathcal{F} \implies A^c \in \mathcal{F}$.

(ς3) If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{n=1}^\infty A_n \in \mathcal{F}$.

This gives us a modified question.

Question: Does there exist *any* σ-algebra $\mathcal{F}$ on $[0,1]$ and a function $\mathbf{P}\colon \mathcal{F} \to [0,1]$ that satisfies (∗), (#1) and (#2)?

Answer: Yes, and it is sort-of unique.

Exercise II.2. *Suppose (∗) and (#1) hold. Prove that (#2) is equivalent to the following: if $(B_n)_\mathbb{N}$ are pairwise disjoint, then*

$$\mathbf{P}\Big(\bigcup B_n\Big) = \sum \mathbf{P}(B_n). \tag{II.1}$$

Solution. If $A_1 \supseteq A_2 \supseteq \cdots \supseteq A$, then $A_1^c \subseteq A_2^c \subseteq \cdots \subseteq A^c$. Let $B_n = A_n^c \setminus A_{n-1}^c$, with $B_1 = A_1^c$. First note that (∗) and (#1) imply the following:

(1) $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$, since $\mathbf{P}(A) + \mathbf{P}(A^c) = \mathbf{P}[0,1] = 1$.

(2) If $A \subseteq B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$, since $\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(B \setminus A)$. Specifically, $\mathbf{P}(A_1) \geq \mathbf{P}(A_2) \geq \cdots \geq \mathbf{P}(A)$.

Thus $\mathbf{P}(A_n) \downarrow \lim \mathbf{P}(A_n) \geq \mathbf{P}(A)$.

Then

$$\sum_{n=1}^\infty \mathbf{P}(B_n) = \lim_{n\to\infty} \mathbf{P}(A_n^c) \quad \text{and} \quad \mathbf{P}(A^c) = \mathbf{P}\Big(\bigcup B_n\Big).$$

If $\mathbf{P}(A_n) \downarrow \mathbf{P}(A)$, then $\mathbf{P}(A_n^c) \uparrow \mathbf{P}(A^c)$ and so $\sum \mathbf{P}(B_n) = \mathbf{P}(\bigcup B_n)$.
If $\sum \mathbf{P}(B_n) = \mathbf{P}(\bigcup B_n)$, then $\lim \mathbf{P}(A_n^c) = \mathbf{P}(A^c)$ and so $\mathbf{P}(A_n) \downarrow \mathbf{P}(A)$. ■

A σ-algebra that works for our case is the *smallest* one that contains all intervals.

Exercise II.3. *If $\{\mathcal{F}_i\}_{i\in I}$ are σ-algebras on Ω, then $\bigcap_{i\in I} \mathcal{F}_i$ is also a σ-algebra.*

Proof. $\varnothing$ is in each $\mathcal{F}_i$ and hence in the intersection. If A is in each $\mathcal{F}_i$, then so is A^c. If $A_1, A_2, \ldots$ are in each $\mathcal{F}_i$, then so is $\bigcup_{n=1}^{\infty} A_n$. ■

This allows us to make sense of the word 'smallest' above.

Definition II.4. Let $\mathcal{S} \subseteq 2^\Omega$. The *smallest* σ-algebra containing $\mathcal{S}$ is given by the intersection of all σ-algebras on Ω that contain $\mathcal{S}$. We denote this by $\sigma(\mathcal{S})$.

This will contain $\mathcal{S}$ since 2^Ω itself is a σ-algebra.

Example (Borel σ-algebra). The *Borel σ-algebra* on $[0,1]$ is the smallest σ-algebra containing all intervals in $[0,1]$. It is denoted by $\mathcal{B}_{[0,1]}$.

II.2 Probability spaces

Lecture 2. Tuesday August 6

Definition II.5 (probability space)**.** A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbf{P})$, where Ω is a non-empty set called the *sample space*, $\mathcal{F}$ is a σ-algebra on Ω, and $\mathbf{P}$ is a *probability measure* on $\mathcal{F}$.

A *probability measure* on a σ-algebra $\mathcal{F}$ is a function $\mathbf{P}\colon \mathcal{F} \to [0,1]$ such that $\mathbf{P}(\Omega) = 1$ and

$$\mathbf{P}\left(\bigsqcup_n A_n\right) = \sum_n \mathbf{P}(A_n)$$

for any sequence of pairwise disjoint sets $A_n \in \mathcal{F}$ (countable additivity).

Countable additivity is a stronger condition than finite additivity.

Exercise II.6. *Prove that countable additivity is equivalent to the following two conditions taken together:*

(1) ***finite additivity:*** *if $A \cap B = \varnothing$, then $\mathbf{P}(A \sqcup B) = \mathbf{P}(A) + \mathbf{P}(B)$*

(2) If $A_n \uparrow A$, then $\mathbf{P}(A_n) \uparrow \mathbf{P}(A)$.

Solution. Identical to exercise II.2. ■

Where do Ω, $\mathcal{F}$, and $\mathbf{P}$ come from?

Ω is simply the set of all possible outcomes.

II.2.1 The σ-algebra

$\mathcal{F} = 2^\Omega$ and $\mathcal{F} = \{\varnothing, \Omega\}$ are bullshit choices. In reality, $\mathcal{F}$ is always chosen to be the smallest σ-algebra containing some specified sets of interest. That is, for some $\mathcal{S} \subseteq 2^\Omega$, $\mathcal{F} = \sigma(\mathcal{S})$.

This is sometimes called the σ-algebra "generated by" $\mathcal{S}$. However, this can create a misconception. Recall the similar notion of the *span* of a set of vectors. We can define the span of a set $S \subseteq V$ of vectors in two ways:

- (external) the smallest subspace containing S.
- (internal) the set of all linear combinations of vectors in S.

For $\sigma(\mathcal{S})$, there is no "internal" definition. $\sigma(\mathcal{S})$ cannot be generated by unions, intersections, etc. of sets in $\mathcal{S}$.

A frequent choice for $\mathcal{S}$ is the following.

Definition II.7 (Borel σ-algebra)**.** Let (X, d) be a metric space. The *Borel σ-algebra* on X is the smallest σ-algebra containing all open sets in X, and is denoted $\mathcal{B}(X)$.

Exercise II.8 (self)**.** *Show that* $\sigma\{(a, b) \mid a, b \in \mathbb{R}\} = \mathcal{B}(\mathbb{R})$.

Solution. Let $\Sigma = \sigma\{(a, b) \mid a, b \in \mathbb{R}\}$. It is obvious that $\Sigma \subseteq \mathcal{B}(\mathbb{R})$, since the set of intervals is a subset of the set of all open sets.

We will show that each open set can be written as a countable union of open intervals. Then $\{U \subseteq \mathbb{R} \mid U \text{ is open}\}$ would be necessarily contained in Σ by (ς3), and so $\mathcal{B}(\mathbb{R}) \subseteq \Sigma$.

Let $U \in \mathbb{R}$ be open. For each $x \in U$, there exists a bounded open interval $I_x = (a_x, b_x) \subseteq U$ containing x. Let $(\alpha_n)_{n\in\mathbb{N}}$ be an enumeration of the rationals, and define

$$I_n = \bigcup_{I_x \ni \alpha_n} I_x.$$

Observe that $I_n = (\inf a_x, \sup b_x)$, where the inf and sup are taken over $I_x \ni \alpha_n$.

But each I_x contains a rational number, so $U = \bigcup_n I_n$ is a countable union of open intervals. ■

Homework 1, problem 8 presents a neater argument.

II.2.2 The probability measure

There is some collection $\mathcal{S} \subseteq \Omega$ for which we know what the probabilities "should" be, $\mathbf{P}\colon \mathcal{S} \to [0,1]$.

Question II.9. *Does* $\mathbf{P}$ *extend to a probability measure on* $\sigma(\mathcal{S})$*? If so, is it unique?*

Uniqueness does not hold.

Example. Let $\Omega = \{1,2,3,4\}$ and $\mathcal{S} = \{\{1,2\},\{2,3\},\{3,4\}\}$. $\mathcal{F} = \sigma(\mathcal{S}) = 2^\Omega$.

Then the probability measures given by

$$\underline{p} = (.25,.25,.25,.25)$$
$$\underline{q} = (.5,0,.5,0)$$

agree on $\mathcal{S}$ but differ on $\mathcal{F}$.

When does uniqueness hold?

Uniqueness

Definition II.10 (π-system)**.** A collection $\mathcal{S} \subseteq 2^\Omega$ is a *π-system* if it is closed under finite intersections. That is, for any $A, B \in \mathcal{S}$, $A \cap B \in \mathcal{S}$.

Definition II.11 (λ-system)**.** A collection $\mathcal{C} \subseteq 2^\Omega$ is a *λ-system* if it contains Ω and is closed under

- proper differences: if $A, B \in \mathcal{C}$ and $B \subseteq A$, then $A \setminus B \in \mathcal{C}$.
- increasing limits: if $A_n \in \mathcal{C}$ and $A_n \uparrow A$, then $A \in \mathcal{C}$.

Theorem II.12. *If* $\mathcal{F} = \sigma(\mathcal{S})$ *where* $\mathcal{S}$ *is a* π*-system and* P, Q *are probability measures on* $\mathcal{F}$ *that agree on* $\mathcal{S}$*, then* $P = Q$*.*

Proof. Consider $\mathcal{G} = \{A \in \mathcal{F} \mid P(A) = Q(A)\}$. Then $\mathcal{G} \supseteq \mathcal{S}$. Further, if $A \in \mathcal{G}$, then $A^c \in \mathcal{G}$ since $P(A^c) = 1 - P(A) = 1 - Q(A) = Q(A^c)$.

If $A, B \in \mathcal{G}$ are disjoint, then

$$P(A \sqcup B) = P(A) + P(B) = Q(A) + Q(B) = Q(A \sqcup B).$$

But how do we deal with A, B not disjoint? We would need to show that $A, B \in \mathcal{G} \implies A \cap B \in \mathcal{G}$.

Resolution: Show that $\mathcal{G}$ is a λ-system, and then apply the π-λ theorem. Suppose $A, B \in \mathcal{G}$ with $B \subseteq A$. Then $P(A \setminus B) = P(A) - P(B) = Q(A) - Q(B) = Q(A \setminus B)$. Thus $\mathcal{G}$ is closed under proper differences.

If $A_n \uparrow A$ are in $\mathcal{G}$, then $P(A_n) \uparrow P(A)$ and $Q(A_n) \uparrow Q(A)$. But $P(A_n) = Q(A_n)$ for all n, so $P(A) = Q(A)$. Thus $\mathcal{G}$ is closed under increasing limits.

$\mathcal{G}$ contains Ω since $P(\Omega) = Q(\Omega) = 1$.

Thus by the π-λ theorem, $\mathcal{G} \supseteq \mathcal{F}$. ■

Theorem II.13 (π-λ theorem)**.** *Let S be a π-system and $\mathcal{C}$ be a λ-system. If $\mathcal{C} \supseteq S$, then $\mathcal{C} \supseteq \sigma(S)$.*

This is due to Sierpiński and Dynkin.

Proof. It suffices to show that $\sigma(S) = \lambda(S)$. Call this simply λ. $\lambda \supseteq S$ already contains Ω, and is closed under complements and increasing limits by virtue of being a λ-system. All that's required is closure under finite unions. Manju shows closure under finite intersections, which is slightly easier. I will show closure under finite unions.

For any $A \in S$, let $\Lambda_A = \{B \in \lambda \mid A \cup B \in \lambda\}$.

- Clearly $\Omega \in \Lambda_A$.
- If $B_n \uparrow B$ are from Λ_A, then $A \cup B_n \uparrow A \cup B$, so $A \cup B \in \lambda$ and hence $B \in \Lambda_A$.
- If $B \in \Lambda_A$, then $A \cup B \in \lambda$, so $B \setminus A = (A \cup B) \setminus A \in \lambda$. Then
$$\Omega \setminus (B \setminus A) = \Omega \cap (B \cap A^c)^c = B^c \cup A \in \lambda,$$
so $B^c \in \Lambda_A$.
- Let $B \in S$. HOW?

Thus Λ_A is a λ-system containing S. This proves that $\Lambda_A = \lambda$. In particular, λ is closed under union with elements of S.

Now let $A \in \lambda$ and define Λ_A as before. By the above, $S \subseteq \Lambda_A$. Now again

- $\Omega \in \Lambda_A$.
- If $B_n \uparrow B$ are from Λ_A, then $A \cup B_n \uparrow A \cup B$, so $A \cup B \in \lambda$ and hence $B \in \Lambda_A$.
- If $B \in \Lambda_A$, then $A \cup B \in \lambda$, so $B \setminus A = (A \cup B) \setminus A \in \lambda$. Then
$$\Omega \setminus (B \setminus A) = \Omega \cap (B \cap A^c)^c = B^c \cup A \in \lambda,$$
so $B^c \in \Lambda_A$.

Thus $\Lambda_A = \lambda$. That is, λ is closed under union with elements of λ. ■

What about existence?

Existence

In the general case, obviously not. Consider $\Omega = [0,1]$ with

$$\mathcal{S} = \{(0, \frac{1}{2}), (0, \frac{1}{4}), (\frac{1}{4}, \frac{1}{2})\}$$
$$\mathbf{P}(a,b) = (b-a)^2.$$

Then the sum of $\mathbf{P}(0, \frac{1}{4})$ and $\mathbf{P}(\frac{1}{4}, \frac{1}{2})$ is less that $\mathbf{P}(0, \frac{1}{2})$.

Let us impose some necessary conditions.

Definition II.14 (algebra)**.** A collection $\mathcal{A} \subseteq 2^\Omega$ is an *algebra* if it $\varnothing \in \mathcal{A}$ and it is closed under complements and finite unions.

Definition II.15 (monotone class)**.** A collection $\mathcal{M} \subseteq 2^\Omega$ is a *monotone class* if it is closed under increasing and decreasing limits.

Proposition II.16 (monotone algebra theorem)**.** *Let $\mathcal{A}$ be an algebra and $\mathcal{M}$ be a monotone class. If $\mathcal{M} \supseteq \mathcal{A}$, then $\mathcal{M} \supseteq \sigma(\mathcal{A})$.*

This is very similar to the π-λ theorem.

Proof. It again suffices to show that $\sigma(\mathcal{A}) = \mathcal{M}(\mathcal{A})$. Call this simply $\mathcal{M}$. $\mathcal{M} \supseteq \mathcal{A}$ already contains $\varnothing$ and is closed under increasing and decreasing limits. It remains to show closure under complements and finite unions.

Let $\mathcal{G} = \{A \in \mathcal{M} \mid A^c \in \mathcal{M}\}$. Since $\mathcal{A}$ is closed under complements, $\mathcal{A} \subseteq \mathcal{G}$. Now if $A_n \uparrow A$ (resp. $A_n \downarrow A$) are from $\mathcal{G}$, then $A \in \mathcal{M}$. Moreover, $A_n^c \downarrow A^c$ (resp. $A_n^c \uparrow A^c$) are in $\mathcal{G}$, so $A^c \in \mathcal{M}$. Thus $A \in \mathcal{G}$. This shows that $\mathcal{G} \subseteq \mathcal{M}$ is a monotone class containing $\mathcal{A}$, so $\mathcal{G} = \mathcal{M}$. Thus $\mathcal{M}$ is closed under complements.

Now let $A \in \mathcal{A}$ and set $\mathcal{M}_A = \{B \in \mathcal{M} \mid A \cup B \in \mathcal{M}\}$. Again $\mathcal{A} \subseteq \mathcal{M}_A$. If $B_n \uparrow B$ (resp. $B_n \downarrow B$) are from $\mathcal{M}_A$, then $A \cup B_n \uparrow A \cup B$ (resp. $A \cup B_n \downarrow A \cup B$). So $A \cup B \in \mathcal{M}$ and hence $B \in \mathcal{M}_A$. Thus $\mathcal{M}_A \subseteq \mathcal{M}$ is a monotone class containing $\mathcal{A}$, so $\mathcal{M}_A = \mathcal{M}$. This shows that $\mathcal{M}$ is closed under union with elements of $\mathcal{A}$.

Now let $A \in \mathcal{M}$ and define $\mathcal{M}_A$ as before. By the above, $\mathcal{A} \subseteq \mathcal{M}_A$. Closure under limits is the same as before. Thus $\mathcal{M}$ is closed under finite unions. ∎

Theorem II.17 (Carathéodory's extension theorem)**.** *Let $\mathcal{S}$ be an algebra. Assume that $P\colon \mathcal{S} \to [0,1]$ is countably additive. Then there exists an extension of P to a probability measure $\mathbf{P}$ on $\mathcal{F} = \sigma(\mathcal{S})$.*

Corollary II.18. *The above extension is unique.*

Proof. An algebra is a π-system. Theorem II.12 applies. ∎

	$\varnothing, \Omega$	A^c	$\bigcap_{i=1}^{n}$	$\bigcup_{i=1}^{n}$	$\bigcap_{i=1}^{\infty}$	$\bigcup_{i=1}^{\infty}$	$A \setminus B$ $(B \subseteq A)$	$A_n \uparrow A$
π-system			✓					
λ-system	✓	✓					✓	✓
algebra	✓	✓	✓	✓			✓	
σ-algebra	✓	✓	✓	✓	✓	✓	✓	✓

Table II.1: Various systems of sets

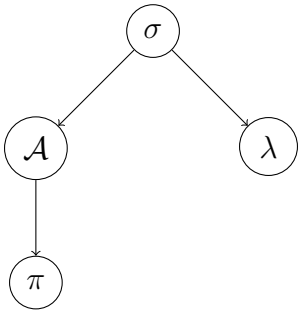


Figure II.1: Heirarchy of systems of sets under inclusion

II.3 Existence of Lebesgue measure

Theorem II.19. *There is a unique probability measure λ on $[0,1]$ with the Borel σ-algebra such that*

$$\lambda[a,b] = b - a \qquad \forall\, 0 \le a \le b \le 1.$$

Proof. Let $\Omega = [0,1)$.

Let $\mathcal{S}_0 = \{[a,b) \mid 0 \le a \le b \le 1\}$. Half-open intervals are nice because they are closed under complements $[a,b)^c = [0,a) \sqcup [b,1)$ and intersections $[a,b) \cap [c,d) = [a \vee c, b \wedge d)$.

Let

$$\mathcal{S} = \{I_1 \sqcup \cdots \sqcup I_k \mid k \ge 1, I_j \in \mathcal{S}_0 \text{ disjoint}\}$$

be the collection of all finite disjoint unions of half-open intervals. It is obvious that $\mathcal{S}$ is an algebra. Define

$$\lambda_{\mathcal{S}}(I_1 \sqcup \cdots \sqcup I_k) = \sum_{j=1}^{k} (\sup I_j - \inf I_j).$$

We need to show that this is countably additive, in order that Carathéodory's extension theorem applies. We will proceed via exercise II.6.

- Finite additivity is obvious.

- Let $A_n, A \in \mathcal{S}$ with $A_n \uparrow A$. Then it should be sort of obvious that $\lambda_{\mathcal{S}}(A_n) \uparrow \lambda_{\mathcal{S}}(A)$. Yes?

By Carathéodory's extension theorem, there exists a unique probability measure λ on $\mathcal{F} = \sigma(\mathcal{S})$ that extends $\lambda_{\mathcal{S}}$. ■

Lecture 3. Thursday August 8

II.4 New measures from old

Definition II.20. Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be two sets with σ-algebras. A function $T\colon \Omega \to \Omega'$ is *measurable* if

$$T^{-1}(B) \in \mathcal{F} \quad \text{for all } B \in \mathcal{F}'.$$

II.4.1 Push forward

Lemma II.21. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(\Omega', \mathcal{F}')$ be a set with a σ-algebra. Let $T\colon \Omega \to \Omega'$ be measurable. Then $Q := P \circ T^{-1}$ is a probability measure on $\mathcal{F}'$.*

Proof. We need to show that $Q(\Omega') = 1$ and Q is countably additive. The first is immediate as $Q(\Omega') = P(T^{-1}(\Omega')) = P(\Omega) = 1$.

Notice that if B_1 and B_2 are disjoint, so are $T^{-1}(B_1)$ and $T^{-1}(B_2)$. Let $(B_n)_{\mathbb{N}}$ be a sequence of pairwise disjoint sets in $\mathcal{F}'$. Then $(T^{-1}(B_n))_{\mathbb{N}}$ are pairwise disjoint in $\mathcal{F}$. Thus

$$\begin{aligned} Q\Big(\bigsqcup B_n\Big) &= P\Big(T^{-1}\Big(\bigsqcup B_n\Big)\Big) \\ &= P\Big(\bigsqcup T^{-1}(B_n)\Big) \\ &= \sum P\Big(T^{-1}(B_n)\Big) \\ &= \sum Q(B_n). \end{aligned}$$

■

Definition II.22 (cumulative distributive function)**.** A *cumulative distributive function* (CDF) is a function $F\colon \mathbb{R} \to [0,1]$ such that

(1) (increasing) $x \le y \implies F(x) \le F(y)$

(2) (right-continuous) $\lim_{h \searrow 0} F(x+h) = F(x)$

(3) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$

Let $P(\mathbb{R})$ be the set of all probability measures on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. If $\mu \in P(\mathbb{R})$, then $F_\mu(x) := \mu(-\infty, x]$ is a CDF (increasing, right-continuous with $F(-\infty) = 0$, $F(\infty) = 1$)).

Lecture 4. Tuesday August 13

Theorem II.23. *Given a CDF $F\colon \mathbb{R} \to [0,1]$, there exists a unique probability measure $\mu \in P(\mathbb{R})$ such that $\mu(-\infty, x] = F(x)$ for all $x \in \mathbb{R}$.*

Proof. Consider $((0,1), \mathcal{B}, \lambda)$ and define

$$\begin{aligned} T\colon (0,1) &\to \mathbb{R} \\ u &\mapsto \inf\{x \in \mathbb{R} : F(x) \geq u\} \end{aligned}$$

The set is non-empty since $F(x) \to 1$ as $x \to \infty$. Moreover, T is increasing since

$$\{x \in \mathbb{R} : F(x) \geq u\} \subseteq \{x \in \mathbb{R} : F(x) \geq v\}$$

whenever $u \leq v$. T is left-continuous.

Finally, $T(u) \leq x \iff F(x) \geq u$. (This is reminiscent of the inverse property: $T(u) = x \iff F(x) = u$.) If $F(x) \geq u$, then $x \in F^{-1}[u,1)$, so $T(u) \leq x$. If $T(u) \leq x$, then $x + \frac{1}{n} \in F^{-1}[u,1)$ for all $n \in \mathbb{N}$. By right-continuity, $F(x) \geq u$.

Now T is Borel-measurable, so

$$\mu \coloneqq \lambda \circ T^{-1}$$

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

Further, $\mu(-\infty, x] = \lambda(T^{-1}(-\infty, x]) = \lambda(0, F(x)] = F(x)$.

Uniqueness if by the π-system thingy. ∎

Examples.

- Take $f\colon \mathbb{R} \to [0,\infty)$ measurable whose total integral is 1. Then $F = x \mapsto \int_{-\infty}^{x} f(u)\,\mathrm{d}u$ is a CDF.
- (Cantor measure) Consider the $\frac{1}{3}$-Cantor set $K = K_1 \cap K_2 \cap \dots$ where

$$\begin{aligned} K_1 &= \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right] \\ K_2 &= \left[0, \frac{1}{9}\right] \cup \left[\frac{2}{9}, \frac{1}{3}\right] \cup \left[\frac{2}{3}, \frac{7}{9}\right] \cup \left[\frac{8}{9}, 1\right] \\ &\vdots \end{aligned}$$

  Notice that

$$K = \{x \in [0,1] : x = \sum_{n=1}^{\infty} \frac{x_n}{3^n}, x_n = 0 \text{ or } 2\}.$$

  We can construct the measurable function

$$\begin{aligned} T\colon [0,1] &\to \mathbb{R} \\ \sum_{n=1}^{\infty} \frac{x_n}{2^n} &\mapsto \sum_{n=1}^{\infty} \frac{2x_n}{3^n} \end{aligned}$$

where we are considering the non-terminating binary expansion of x on the left. It is obvious that T maps only to K. Since $T^{-1}(K) = [0,1]$, we have that $\mu(K) = 1$. However, $\lambda(K) = 0$. Thus the CDF cannot arise from a density. However, the CDF is continuous!

- (just for fun) Fix a $\theta > 2$ and define

$$\begin{aligned} T_\theta &: [0,1] \to [0,1] \\ \sum_{n=1}^{\infty} \frac{x_n}{2^n} &\mapsto \sum_{n=1}^{\infty} \frac{x_n}{\theta^n} \end{aligned}$$

define $\mu_\theta = \lambda \circ T_\theta^{-1}$. $\mu_2 = \lambda$. It is known that for $\theta > 2$, μ_θ has no density. What about $1 < \theta < 2$? This is an open problem. "Bernoulli convolution problem".

II.4.2 Structure of $P(\Omega, \mathcal{F})$

What is the structure of $P(\Omega, \mathcal{F})$? Is it a vector space? A group?

One thing to note is that $P(\Omega, \mathcal{F})$ is convex. That is, given any $\mu, \nu \in P(\Omega, \mathcal{F})$ and $0 \leq t \leq 1$, $(1-t)\mu + t\nu \in P(\Omega, \mathcal{F})$. This is called a *mixture* of μ and ν.

We would like to study *closeness* of probability measures. Consider a computer generating a random number between 0 and 1, by generating a sequence of 8 random bits. The computer is actually sampling from the uniform distribution

$$\mu_{2^8} = \text{Unif}\left\{\frac{0}{2^8}, \frac{1}{2^8}, \ldots, \frac{2^8 - 1}{2^8}\right\}.$$

However, we do accept μ as an approximation of λ. We will thus attempt to define a *metric* on $P(\mathbb{R})$.

Attempt 1. (total variation distance) Define

$$d(\mu, \nu) = \sup_{A \in \mathcal{B}_\mathbb{R}} |\mu(A) - \nu(A)|.$$

This does not work for out for our use case, as

$$d(\mu_{2^8}, \lambda) = 1.$$

Attempt 2. (Kolmogorov-Smirnov metric) Choose a suitable $\mathcal{C} \in \mathcal{B}_\mathbb{R}$ and define

$$d(\mu, \nu) = \sup_{A \in \mathcal{C}} |\mu(A) - \nu(A)|.$$

$\mathcal{C}$ should be "measure-determining".

Attempt 3. (Lévy metric)

$$d(\mu,\nu) = \inf\{\varepsilon > 0 : F_\mu(x+\varepsilon) + \varepsilon \geq F_\nu(x) \text{ and } F_\nu(x+\varepsilon) + \varepsilon \geq F_\mu(x) \text{ for all } x \in \mathbb{R}\}.$$

This is symmetric by sheer obviousness. For $\triangle$, consider three measures μ, ν, ρ.

$$\begin{aligned} t > d(\mu,\nu) &\implies F_\mu(x+t) + t \geq F_\nu(x) \\ s > d(\nu,\rho) &\implies F_\nu(x+s) + s \geq F_\rho(x) \end{aligned}$$

Thus

$$F_\mu(x+t+s) + t + s \geq F_\nu(x+s) + t \geq F_\rho(x)$$

Thus $t + s \geq d(\mu,\rho)$. $\triangle$ holds.

Finally, suppose $d(\mu,\nu) = 0$. Let $\varepsilon_n \downarrow 0$ be a sequence such that $F_\mu(x+\varepsilon_n) + \varepsilon_n \geq F_\nu(x)$ for all x for all n. Taking limits, we have $F_\mu(x) \geq F_\nu(x)$ by right-continuity. By symmetry, $F_\mu(x) = F_\nu(x)$.

Lecture 5. Tuesday August 20

Definition II.24. If $\mu_n, \mu \in P(\mathbb{R})$ and $d(\mu_n,\mu) \to 0$ then we say that μ_n *converges in distribution* to μ and write $\mu_n \xrightarrow{\mathrm{d}} \mu$.

Remark. This is also called *weak convergence* and hence sometimes written $\mu_n \xrightarrow{\mathrm{w}} \mu$. Yet others write $\mu_n \Rightarrow \mu$.

We now prove an extremely powerful result for showing convergence of probability measures.

Proposition II.25. *Let $\mu_n, \mu \in P(\mathbb{R})$. Then*

$$\mu_n \xrightarrow{d} \mu \iff F_{\mu_n}(x) \to F_\mu(x) \text{ for all } x \text{ where } F_\mu \text{ is continuous.}$$

Examples.

- $\delta_{\frac{1}{n}} \xrightarrow{\mathrm{d}} \delta_0$ because

$$\lim_{n\to\infty} F_{\delta_{1/n}}(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 & \text{if } x > 0. \end{cases}$$

  So $F_{\delta_{1/n}}(x) \to F_{\delta_0}(x)$ for all $x \neq 0$.

- $\delta_{-\frac{1}{n}}(x) \to \delta_0(x)$ everywhere.

Proof. Write $F_\mu = F$ and $F_{\mu_n} = F_n$.

Suppose $\mu_n \xrightarrow{\mathrm{d}} \mu$ and let F be continuous at $x \in \mathbb{R}$. Let $\varepsilon > 0$. Then

$$\begin{aligned} F(x+\varepsilon) + \varepsilon &\geq F_n(x) \\ F_n(x) + \varepsilon &\geq F(x-\varepsilon) \end{aligned}$$

for all large n. Thus we have

$$\begin{aligned} \limsup_{n\to\infty} F_n(x) &\leq F(x+\varepsilon) + \varepsilon \\ \liminf_{n\to\infty} F_n(x) &\geq F(x-\varepsilon) - \varepsilon. \end{aligned}$$

But this holds for all $\varepsilon > 0$. Letting $\varepsilon \downarrow 0$ gives

$$\begin{aligned} \limsup_{n\to\infty} F_n(x) &\leq F(x) \\ \liminf_{n\to\infty} F_n(x) &\geq F(x). \end{aligned}$$

Thus $\lim_{n\to\infty} F_n(x) = F(x)$.

Now suppose $F_n(x) \to F(x)$ for all x where F is continuous. Fix $\varepsilon > 0$ and pick $x_1 < \cdots < x_p$ such that

- each x_j is a continuity point of F,
- $x_{j+1} - x_j < \varepsilon$ for all j,
- $F(x_1) \leq \varepsilon$ and $F(x_p) \geq 1 - \varepsilon$.

Then $\exists N \in \mathbb{N}$ such that $\forall n \geq N$ we have

$$|F_n(x_j) - F(x_j)| < \varepsilon \text{ for all } j. \tag{II.2}$$

Let $x \in \mathbb{R}$ and $n \geq N$. We have three cases.

($x_j \leq x \leq x_{j+1}$) Then

$$F_n(x+\varepsilon) + \varepsilon \geq F_n(x_{j+1}) + \varepsilon \geq F(x_{j+1}) \geq F(x).$$

The first and last inequalities are by the increasing nature of CDFs. The middle inequality is by equation (II.2). Similarly

$$F(x+\varepsilon) + \varepsilon \geq F(x_{j+1}) + \varepsilon \geq F_n(x_{j+1}) \geq F_n(x).$$

($x < x_1$) Then

$$F_n(x+\varepsilon) + \varepsilon \geq \varepsilon \geq F(x_1) \geq F(x).$$

The other direction requires a bigger jump.

$$F(x+2\varepsilon) + 2\varepsilon \geq 2\varepsilon \geq F(x_1) + \varepsilon \geq F_n(x_1) \geq F_n(x).$$

($x > x_p$)

Thus $d(\mu_n, \mu) \to 0$. ∎

Remarks.

- We will now frequently show $F_{\mu_n} \to F_\mu$ at all continuity points of F_μ, to show that $\mu_n \xrightarrow{\text{d}} \mu$. In fact, many authors use this proposition as the *definition* of convergence, without even mentioning the Lévy metric.
- Notice that the converse did not use the continuity of F at all. All that was required is that the points of continuity of F are dense. Thus we have the following proposition immediately.

Proposition II.26. *Let $\mu_n, \mu \in P(\mathbb{R})$ and let D be a dense subset of $\mathbb{R}$. Then*

$$F_{\mu_n}(x) \to F_\mu(x) \text{ for all } x \in D \implies \mu_n \xrightarrow{d} \mu.$$

$(P(\mathbb{R}), d_{\text{Lévy}})$ is a metric space. It is interesting to ask what the *compact* subsets of this space are, so that we can exploit convergence of subsequences.

Definition II.27. A subset $\mathcal{A} \subseteq P(\mathbb{R})$ is *tight* if for all $\varepsilon > 0$ there exists a compact set K_ε such that

$$\mu(K_\varepsilon^c) \leq \varepsilon \text{ for all } \mu \in \mathcal{A}.$$

For $\mathbb{R}$, it only makes sense to consider $K_\varepsilon = [-M_\varepsilon, M_\varepsilon]$. Such an M_ε exists for each $\mu \in \mathcal{A}$ individually, but not neccessarily for all $\mu \in \mathcal{A}$ simultaneously.

Examples.

- $\mathcal{A} = \{\delta_n\}_{n\in\mathbb{Z}}$ is *not* tight. No matter what M is chosen, $\delta_{\lceil M+1 \rceil}$ will have all of its mass outside of $[-M, M]$.
- Similarly, $\{N(\mu, 1)\}_{\mu\in\mathbb{R}}$ is not tight, but $\{N(\mu, 1)\}_{-16\leq\mu\leq 32768}$ is tight.
- $\{N(\mu, \sigma^2)\}_{-10\leq\mu\leq 10}$ is not tight, but $\{N(\mu, \sigma^2)\}_{\substack{-10\leq\mu\leq 10 \\ 0<\sigma<10}}$ is.
- $\{\delta_{\frac{1}{n}}\}_{n\in\mathbb{N}}$ is tight.

Definition II.28. A set $E \subseteq (X, d)$ is *pre-compact* if its closure $\overline{E}$ is compact.

Theorem II.29. *Any $\mathcal{A} \subseteq P(\mathbb{R})$ is pre-compact iff it is tight.*

We will cover two prerequisites before we prove this theorem.

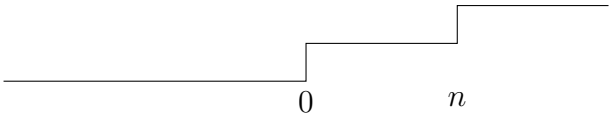
Theorem II.30 (Helly's selection principle)**.** *Let $\mu_n \in P(\mathbb{R})$. Then there is a subsequence $n_1 < n_2 < \dots$ and an increasing, right continuous function $F\colon \mathbb{R} \to [0, 1]$ such that*

$$F_{\mu_{n_k}}(x) \to F(x) \text{ for all } x \text{ where } F \text{ is continuous.}$$

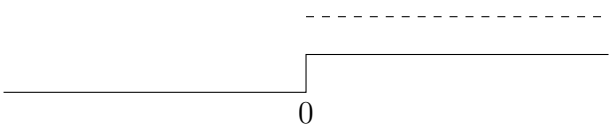
F may be a "defective CDF". It need not go to 0 to the left, nor 1 to the right.

Examples.

- Let $\mu_n = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_n$. $F_{\mu_n}(x)$ looks like



  The pointwise limit of any subsequence is



  This is *not* a CDF.

- The limit for $\mu_n = N(0, n)$ is the constant function $F(x) = \frac{1}{2}$.

Proof. Fix a dense countable set $D = \{x_1, x_2, \ldots\} \subseteq \mathbb{R}$. By compactness of $[0, 1]$, there exists a subsequence $(n_k)_{k\in\mathbb{N}}$ such that $F_{n_k}(x_1)$ converges, say to c_1.

Choose a further subsequence $(n_{k_l})_{l\in\mathbb{N}}$ such that $F_{n_{k_l}}(x_2)$ converges, say to c_2.

Choose a further subsequence $\left(n_{k_{l_m}}\right)_{m\in\mathbb{N}}$ such that $F_{n_{k_{l_m}}}(x_3)$ converges, say to c_3.

The limit of doing this infinitely many times may give an empty subsequence. The key is *diagonalization*.

Let us relabel these subsequences as $(n_{1,k})_{k\in\mathbb{N}}$, $(n_{2,k})_{k\in\mathbb{N}}$, $(n_{3,k})_{k\in\mathbb{N}}$, ...

$$\begin{array}{c|cccc} n_1 & n_{1,1} & n_{1,2} & n_{1,3} & \cdots \\ n_2 & n_{2,1} & n_{2,2} & n_{2,3} & \cdots \\ n_3 & n_{3,1} & n_{3,2} & n_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

Walk the diagonal. $F_{n_{j,j}}(x_i) \to c_i$ for each i.

Thus we have constructed a subsequence, which we will finally label $(n_k)_{k\in\mathbb{N}}$ such that

$$F_{n_k}(x_i) \to c_i \text{ for all } i.$$

All that remains is to extend this preserving right-contiunity. Define

$$F(x) \coloneqq \inf\{c_i \mid i \in \mathbb{N} \text{ such that } x < x_i\}.$$

All that remains is to check that

- F is increasing and right-continuous,
- $F_{n_k}(x) \to F(x)$ if F is continuous at x.

Suppose $x_1 \leq x_2$. Then $F(x_1) = \inf\{c_i \mid x_i > x_1\} \leq \inf\{c_i \mid x_i > x_2\} = F(x_2)$ since the second set is a subset of the first.

Now let $x \in \mathbb{R}$ and $\varepsilon > 0$. Then $F(x) \geq c_i - \varepsilon$ for some i such that $x_i \geq x$. Let $y \in (x, x_i)$. Then $F(y) \leq c_i$ by definition of F (c_i is a witness for y). Thus $F(x) \leq F(y) \leq F(x) + \varepsilon$. ■

Lecture 6. Thursday August 22

When does Helly's selection give a defective CDF? Whenever some mass escapes out to $\pm\infty$. For example, in $\mu_n = \frac{1}{4}\delta_{-n} + \frac{1}{2}\delta_0 + \frac{1}{4}\delta_n$, whose limit is the constant $x \mapsto \frac{1}{2}$. If the mass does not escape, we should get a proper CDF. This is where tightness comes in (theorem II.29).

Lemma II.31. *Suppose $\mu_n \in P(\mathbb{R})$ and F is a possibly defective CDF. Suppose $F_{\mu_n} \to F$ at all continuity points of F. Then $F = F_\mu$ for some $\mu \in P(\mathbb{R})$ iff $\{\mu_n\}$ is tight.*

Proof. ($\Longrightarrow$) Suppose $F = F_\mu$. Let $\varepsilon > 0$ be given. Let M_1, M_2 be such that $F(M_1) < \varepsilon$ and $F(M_2) > 1 - \varepsilon$. We can choose M_1, M_2 to be continuity points of F, since it is continuous at all but countably many points.

Since $F_{\mu_n} \to F$ at all continuity points of F, $F_{\mu_n}(M_1) \to F(M_1) < \varepsilon$ and $F_{\mu_n}(M_2) \to F(M_2) > 1 - \varepsilon$. Thus there is some N such that for all $n \geq N$, $F_{\mu_n}(M_1) < \varepsilon$ and $F_{\mu_n}(M_2) > 1 - \varepsilon$, that is,

$$\mu_n[M_1, M_2] > 1 - 2\varepsilon \text{ for all } n \geq N.$$

We need to show this for all n. Simply pick $M_1' < M_1$ and $M_2' > M_2$ such that $\mu_n[M_1', M_2'] > 1 - 2\varepsilon$ for all $n < N$, which are only finitely many. Thus $\{\mu_n\}$ is tight.

($\Longleftarrow$) Now suppose $\{\mu_n\}$ is tight. Let $\varepsilon > 0$. Pick $M_1 < M_2$ such that $\mu_n[M_1, M_2] > 1 - \varepsilon$ for all n, ensuring again that F is continuous at M_1, M_2. Then

$$F(M_1) = \lim F_{\mu_n}(M_1) \leq \varepsilon \quad \text{and} \quad F(M_2) = \lim F_{\mu_n}(M_2) \geq 1 - \varepsilon.$$

Thus F is not defective. ■

We can now prove theorem II.29.

Proof of theorem II.29. ($\Longrightarrow$) Suppose $\mathcal{A}$ is not tight. That is, there is some $\varepsilon > 0$ such that for all $M > 0$, there is some $\mu \in \mathcal{A}$ for which $\mu[-M, M]^c > \varepsilon$. Thus we have a sequence $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{A}$ such that $\mu_n[-n, n]^c > \varepsilon$ for all n.

Note that no subsequence of (μ_n) is tight. Thus the previous lemma gives that no subsequence of (μ_n) can converge to a proper CDF, and hence $\mathcal{A}$ is not pre-compact.

($\Longleftarrow$) Suppose $\mathcal{A}$ is tight. Let $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{A}$. By Helly's selection principle, there exists a subsequence $(\mu_{n_k})_{k\in\mathbb{N}}$ and a possibly defective CDF F such that $F_{\mu_{n_k}} \to F$ at all continuity points of F. But (μ_{n_k}) is tight, so by the previous lemma, F is a proper CDF. ■

Recap: We have covered the following so far.

- Probability spaces $(\Omega, \mathcal{F}, \mathbf{P})$ in section II.2.
- Where $\mathcal{F}$ and $\mathbf{P}$ come from.
- Construction of probability measures:
  - Lebesgue measure
  - Coin-tossing measure
  - Every measure on $\mathbb{R}$.
- Lévy metric and convergence in distribution in terms of CDFs.
- Tightness and Helly's selection.

Chapter III

The Lebesgue integral

Fix a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. For this chapter, we will let RV denote the collection of all (real-valued) random variables, and RV_+ denote the collection of all non-negative random variables.

That is, all functions $X\colon \Omega \to \overline{\mathbb{R}}$ such that for each $B \in \mathcal{B}(\overline{R})$, $X^{-1}(B) \in \mathcal{F}$.

Notice that the codomain of X is $\overline{\mathbb{R}}$, the extended real numbers. This is because it is often convenient to allow random variables to take infinite values. In fact, whenever we say "real-valued", we will mean "extended real-valued".

We will need to define the Borel σ-algebra on $\overline{\mathbb{R}}$. For this we define the following metric.

Definition III.1 (Metric on $\overline{\mathbb{R}}$)**.** For $x, y \in \overline{\mathbb{R}}$, we define the metric

$$d_{\overline{\mathbb{R}}}(x, y) = \left| \frac{x}{1+|x|} - \frac{y}{1+|y|} \right|.$$

Exercise III.2. *Check that any function* $X\colon (\Omega, \mathcal{F}) \to \left(\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}})\right)$ *is measurable iff*

$$\{X \le t\} \coloneqq \{\omega \in \Omega \mid X(\omega) \le t\} \in \mathcal{F} \quad \textit{for all } t \in \mathbb{R}.$$

Solution. The forward direction is by definition. Assume $\{X \le t\} \in \mathcal{F}$ for all $t \in \mathbb{R}$. Let $\mathcal{G} = \{A \in \mathcal{B}(\overline{\mathbb{R}}) \mid X^{-1}(A) \in \mathcal{F}\}$. By the assumption, $\mathcal{G}$ contains all $(-\infty, t]$. $\mathcal{G}$ contains $\varnothing$, and if $A \subseteq B$ are in $\mathcal{G}$, then $X^{-1}(B \backslash A) = X^{-1}(B) \setminus X^{-1}(A) \in \mathcal{F}$. ■

Theorem III.3 (existence and uniqueness of expectation)**.** *There is a unique function* $E\colon \mathrm{RV}_+ \to [0,\infty]$ *called the* expectation *such that*

(E1) (pseudo-linearity) $E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y]$ *for every* $X, Y \in \mathrm{RV}_+$ *and* $\alpha, \beta \geq 0$*;*

(E2) (positivity) $E[X] \geq 0$ *with equality iff* $X = 0$ *almost surely;*

(E3) (indicator) $E[\mathbf{1}_A] = \mathbf{P}(A)$ *for all* $A \in \mathcal{F}$*;*

(E4) (monotone convergence) If $X_n \uparrow X$ *almost surely (that is,* $\mathbf{P}\{\omega \in \Omega \mid X_n(\omega) \uparrow X(\omega)\} = 1$*), then* $E[X_n] \uparrow E[X]$*.*

Exercise III.4. *Let* $X_n \in \mathrm{RV}$*. Show that the following are measurable sets.*

(1) $\{\omega \mid \lim X_n = 0\}$

(2) $\{\omega \mid \lim X_n$ *exists*$\}$

Definition III.5 (Expectation)**.** For $X \in \mathrm{RV}$, let $X_+ = X \vee 0$ and $X_- = (-X) \vee 0$. Then $X_+, X_- \in \mathrm{RV}_+$ and $X = X_+ - X_-$, $|X| = X_+ + X_-$. If $E|X| < \infty$, we say X is *integrable* or that X *has expectation* and define $\mathbf{E}[X] \coloneqq E[X_+] - E[X_-]$.

Proposition III.6.

(1) (linearity) If $X, Y \in \mathrm{RV}$ *are integrable and* $\alpha, \beta \in \mathbb{R}$*, then* $\alpha X + \beta Y$ *is integrable and* $\mathbf{E}[\alpha X + \beta Y] = \alpha\, \mathbf{E}[X] + \beta\, \mathbf{E}[Y]$*.*

(2) (positivity) If $X \in \mathrm{RV}_+$ *then* $\mathbf{E}[X] \geq 0$*, with equality only if* $X = 0$ *almost surely.*

(3) (indicator) $\mathbf{E}[\mathbf{1}_A] = \mathbf{P}(A)$ *for all* $A \in \mathcal{F}$*.*

III.1 Lebesgue as super-Riemann

The expectation is a generalization of the Riemann integral.

Proposition III.7. *Fix* $(\Omega, \mathcal{F}, \mathbf{P}) = ([0,1], \mathcal{B}([0,1]), \lambda)$*. Let* $f\colon [0,1] \to \mathbb{R}$ *be continuous. Then* $f \in \mathrm{RV}$ *and* $\mathbf{E}[f] = \int_0^1 f(x)\,\mathrm{d}x$*.*

Proof. f is measurable since the pre-image of each open set is open. f is bounded by the extreme value theorem.

Let $M = \sup|f(x)|$. Then $\mathbf{E}|f| \leq M\, \mathbf{E}[\mathbf{1}_{[0,1]}] = M$ is well-defined.

Let $(f_n)_n$ be a sequence of step functions bounded above by f that converges pointwise to f. Then $\mathbf{E}[f_n] = \int_0^1 f_n(x)\,\mathrm{d}x$ by indicators and

linearity. By the monotone convergence theorem, $\mathbf{E}[f_n] \uparrow \mathbf{E}[f]$. Thus $\mathbf{E}[f] = \int_0^1 f(x)\,\mathrm{d}x$. ■

Lecture 7.
Tuesday
August 27

Proof of theorem III.3 (uniqueness). Let $X \in \mathrm{RV}_+$. Define

$$X_n(\omega) = \sum_{k=0}^{n2^n-1} \frac{k}{2^n} \left[X(\omega) \in \left[\frac{k}{2^n}, \frac{k+1}{2^n} \right) \right].$$

Observe that $X_n(\omega) \le X_{n+1}(\omega)$ for all n and ω. As the partition becomes finer, X_n converges to X pointwise. Thus, by the monotone convergence theorem, $\mathbf{E}\,X_n \uparrow \mathbf{E}\,X$. But we can find $\mathbf{E}\,X_n$ explicitly:

$$\mathbf{E}\,X_n = \sum_{k=0}^{n2^n-1} \frac{k}{2^n}\,\mathbf{P}\left(X \in \left[\frac{k}{2^n}, \frac{k+1}{2^n} \right) \right)$$

The limit exists axiomatically, so

$$\mathbf{E}\,X = \lim_{n\to\infty} \sum_{k=0}^{n2^n-1} \frac{k}{2^n}\,\mathbf{P}\left(X \in \left[\frac{k}{2^n}, \frac{k+1}{2^n} \right) \right)$$

is uniquely determined. ■

Once we have expectation, various interesting quantitites can be defined.

- **Moments:** if $p \in \mathbb{N}$ and X^p is integrable, then $\mathbf{E}[X^p]$ is called the p-th moment of X. More generally, if $|X|^p$ is integrable, we say that the p-th moment of X exists.
- **Variance:** if the second moment exists, we define

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

  By linearity,

  $$\begin{aligned} \operatorname{Var} X &= \mathbf{E}[X^2 - 2X(\mathbf{E}\,X) + (\mathbf{E}\,X)^2] \\ &= \mathbf{E}\,X^2 - (2\,\mathbf{E}\,X)\,\mathbf{E}\,X + (\mathbf{E}\,X)^2\,\mathbf{E}[\mathbf{1}] \\ &= \mathbf{E}\,X^2 - (\mathbf{E}\,X)^2. \end{aligned}$$

  This exists, since $|X| \le X^2 + 1$.
- **Moment generating function:** If $\mathbf{E}[e^{\theta x}]$ exists for all $\theta \in I = (-a, b)$, we define

  $$\begin{aligned} \phi\colon I &\to \mathbb{R} \\ \theta &\mapsto \mathbf{E}[e^{\theta X}]. \end{aligned}$$
- **Characteristic function:** We define

  $$\begin{aligned} \psi\colon \mathbb{R} &\to \mathbb{C} \\ \theta &\mapsto \mathbf{E}[e^{i\theta X}] = \mathbf{E}[\cos(\theta X)] + i\,\mathbf{E}[\sin(\theta X)]. \end{aligned}$$

Exercise III.8. *If* $\mathbf{E}[e^{\theta X}]$ *exists for all* $\theta \in (-\delta, \delta)$ *for some* $\delta > 0$*, show that* X *has all moments.*

Theorem III.9 (inequalities)**.** *Consider a probability space* $(\Omega, \mathcal{F}, \mathbf{P})$*. Let* X, Y *be random variables on* Ω*.*

(1) If $\mathbf{E}\, X^2 < \infty$ *and* $\mathbf{E}\, Y^2 < \infty$*, then* XY *is integrable and*

$$(\mathbf{E}[XY])^2 \le \mathbf{E}[X^2]\, \mathbf{E}[Y^2].$$

(2) $(\mathbf{E}\, X)^2 \le \mathbf{E}[X^2]$*.*

(3) Let $1 < p, q < \infty$ *with* $\frac{1}{p} + \frac{1}{q} = 1$*. Let* $X, Y \in \mathrm{RV}_+$ *and* $\mathbf{E}\, X^p$*,* $\mathbf{E}\, Y^q$ *exist. Then*

$$\mathbf{E}[XY] \le \mathbf{E}[X^p]^{\frac{1}{p}}\, \mathbf{E}[Y^q]^{\frac{1}{q}}.$$

(4) Let $1 \le p < \infty$ *and* $\mathbf{E}|X|^p, \mathbf{E}|Y|^p < \infty$*. Then*

$$\mathbf{E}[|X+Y|^p]^{\frac{1}{p}} \le \mathbf{E}[|X|^p]^{\frac{1}{p}} + \mathbf{E}[|Y|^p]^{\frac{1}{p}}.$$

Proof. Consider the set

$$\mathcal{V} = \{X \in \mathrm{RV} \mid \mathbf{E}\, X^2 < \infty\}$$

be the space of square-integrable random variables. Then for any $X, Y \in \mathcal{V}$, we have

$$|XY| \le \frac{X^2 + Y^2}{2}$$

is integrable. Thus

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

is a pseudo-inner product on $\mathcal{V}$. Cauchy-Schwarz follows.

More directly, let $X, Y \in \mathcal{V}$. Then

$$\begin{aligned} 0 &\le \mathbf{E}[(X - \lambda Y)^2] \\ &= \mathbf{E}\, X^2 - 2\lambda\, \mathbf{E}[XY] + \lambda^2\, \mathbf{E}\, Y^2 \end{aligned}$$

for all $\lambda \in \mathbb{R}$. Thus the discriminant is nonpositive, so

$$(\mathbf{E}[XY])^2 \le \mathbf{E}\, X^2\, \mathbf{E}\, Y^2.$$

The equality holds iff there is some λ such that $X = \lambda Y$ a.s.

(2) follows from Cauchy-Schwarz with $X = Y$. Alternatively, follows from $\mathrm{Var}(X) \ge 0$.

For Hölder's inequality, define

$$A = \frac{X}{\mathbf{E}\, X^p} \quad \text{and} \quad B = \frac{Y}{\mathbf{E}\, Y^q}.$$

From Hölder's inequality for real numbers, we have

$$\frac{XY}{(\mathbf{E}\,X^p)^{\frac{1}{p}}(\mathbf{E}\,Y^q)^{\frac{1}{q}}} \le \frac{1}{p}\frac{X^p}{\mathbf{E}\,X^p} + \frac{1}{q}\frac{Y^q}{\mathbf{E}\,Y^q}.$$

The expectation is thus bounded by

$$\frac{1}{p}\frac{\mathbf{E}\,X^p}{\mathbf{E}\,X^p} + \frac{1}{q}\frac{\mathbf{E}\,Y^q}{\mathbf{E}\,Y^q} = 1.$$

This gives

$$\mathbf{E}[XY] \le \mathbf{E}[X^p]\,\mathbf{E}[Y^q].$$

Finally, we come to Minkowski's inequality. $p = 1$ is obvious, so consider $p > 1$, and let $q = \frac{p}{p-1}$.

$$\begin{aligned}
\mathbf{E}|X+Y|^p &= \mathbf{E}|X+Y|^{p-1}|X+Y| \\
&\le \mathbf{E}|X+Y|^{p-1}|X| + \mathbf{E}|X+Y|^{p-1}|Y| \\
&\le (\mathbf{E}|X|^p)^{\frac{1}{p}}\Big(\mathbf{E}|X+Y|^{(p-1)q}\Big)^{\frac{1}{q}} + (\mathbf{E}|Y|^p)^{\frac{1}{p}}\Big(\mathbf{E}|X+Y|^{(p-1)q}\Big)^{\frac{1}{q}} \\
&= (\mathbf{E}|X|^p)^{\frac{1}{p}} + (\mathbf{E}|Y|^p)^{\frac{1}{p}}(\mathbf{E}|X+Y|^p)^{\frac{1}{q}}.
\end{aligned}$$

■

Theorem III.10 (Jensen's inequality)**.** *Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a convex function and let X be an integrable random variable. Then*

$$\mathbf{E}[\phi(X)] \ge \phi(\mathbf{E}\,X)$$

Proof. We will use that for any $x_0 \in \mathbb{R}$, there is a line $y = \phi(x_0) + (x - x_0)m$ that lies below the raph of ϕ. Let $x_0 = \mathbf{E}\,X$ and take expectations.

$$\phi(x_0) = \mathbf{E}[\phi(x_0)] \le \mathbf{E}[\phi(X)].$$

■

Lecture 8. Thursday August 29

III.2 Change of variables

Tutorial 2. Tuesday August 27

Proof of theorem III.3 (existence). Fix a probability space $(\Omega, \mathcal{F}, P)$.

We will only consider non-negative random variables. We define a *simple function* on Ω to be a random variable whose range is finite. For a simple function X taking values $x_1, \dots, x_n$ on sets $A_1, \dots, A_n \in \mathcal{F}$, we define the expectation of X to be

$$\mathbf{E}[X] = \sum_{i=1}^{n} x_i P(A_i).$$

For a general random variable X, we define the expectation of X to be

$$\mathbf{E}[X] = \sup\{\mathbf{E}[\phi] : 0 \le \phi \le X, \phi \text{ simple}\}.$$

We have to show

- $\mathbf{E}[X]$ is well-defined and agrees with the first definition when X is simple.
- $\mathbf{E}[\mathbf{1}_A] = P(A)$ for any $A \in \mathcal{F}$.
- $\mathbf{E}[X]$ is linear.
- $\mathbf{E}[X] \leq \mathbf{E}[Y]$ if $X \leq Y$.
- $$\mathbf{E}\Big[\phi \mathbf{1}_{\bigsqcup_{i=1}^{\infty} A_i}\Big] = \sum_{i=1}^{\infty} \mathbf{E}[\phi \mathbf{1}_{A_i}] \tag{III.1}$$

Let $\varepsilon > 0$ be arbitrary and

$$E_n = \{\omega \in \Omega : X_n(\omega) \geq (1-\varepsilon)\phi(\omega)\}.$$

Note that $E_n \subseteq E_{n+1}$ and $\bigcup_{n=1}^{\infty} E_n = \Omega$. That is, $E_n \uparrow \Omega$. Now

$$\begin{aligned}\mathbf{E}[X_n] &\geq \mathbf{E}[X_n \mathbf{1}_{E_n}] \\ &\geq \mathbf{E}[(1-\varepsilon)\phi \mathbf{1}_{E_n}] \\ &= (1-\varepsilon)\,\mathbf{E}[\phi \mathbf{1}_{E_n}]\end{aligned}$$

Since $E_n \uparrow \Omega$, we have

$$\lim_{n\to\infty} \mathbf{E}[X_n] \geq (1-\varepsilon)\,\mathbf{E}[\phi]$$

by equation (III.1). ■

Proposition III.11 (simple function approximation)**.** *Let $X\colon (\Omega, \mathcal{F}, \mathbf{P}) \to \mathbb{R}$ be a random variable, and let $f\colon \mathbb{R} \to \mathbb{R}$. Then*

$$\mathbf{E}[f(X)] = \int f(X)\,\mathrm{d}\,\mathbf{P} = \int f(x)\,\mathrm{d}\mu,$$

where μ is the push-forward measure

$$\begin{aligned}\mu\colon \mathbb{R} &\to \mathbb{R} \\ A &\mapsto \mathbf{P}(X^{-1}(A))\end{aligned}$$

Lecture 10. Thursday September 5

Definition III.12. Let $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ be measurable spaces and $X\colon \Omega \to \Lambda$ be measurable. We define

$$\sigma(X) = \{X^{-1}(A) \mid A \in \mathcal{G}\}.$$

This is the smallest σ-algebra on Ω with respect to which X is measurable.

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

Definition III.13 (Independence)**.** $\mathcal{G}_1, \dots, \mathcal{G}_n$ sub σ-algebras of $\mathcal{F}$ are *independent* if

$$\mathbf{P}(A_1 \cap \dots \cap A_n) = \mathbf{P}(A_1) \dots \mathbf{P}(A_n) \quad \text{for all } A_k \in \mathcal{G}_k \tag{III.2}$$

Random variables $X_1, \dots, X_n$ are independent if $\sigma(X_1), \dots, \sigma(X_n)$ are.

Equivalently, X_i's are independent if

$$\mathbf{P}\{X_1 \in A_1, \dots, X_n \in A_n\} = \mathbf{P}\{X_1 \in A_1\} \dots \mathbf{P}\{X_n \in A_n\}$$

for all $A_1, \dots, A_n \in \mathcal{B}(\mathbb{R})$.

Lemma III.14. *Let $\mathcal{G}_k = \sigma(\mathcal{S}_k)$ where $\mathcal{S}_k$ is a π-system containing Ω. If equation (III.2) folds for $A_k \in \mathcal{S}_k$, then $\mathcal{G}_k$ are independent.*

Example. If

Exercise III.15. *If $(\mathcal{H}_i)_{i \in I}$ and $(\mathcal{G}_i)_{i \in I}$ are such that $\mathcal{H}_i \subseteq \mathcal{G}_i$ for all $i \in I$, then independence of $(\mathcal{G}_i)_{i \in I}$ implies independence of $(\mathcal{H}_i)_{i \in I}$.*

Lemma III.16. *Suppose $(\mathcal{G}_i)_{i \in I}$ are independent. Let $I = \bigsqcup_{r \in R} I_r$ be a partition of I. Define $\widetilde{\mathcal{G}}_r = \sigma(\bigcup_{i \in I_r} \mathcal{G}_i)$. Then $(\widetilde{\mathcal{G}}_r)_{r \in R}$ are independent.*

Proof. For each $r \in R$, let $\mathcal{S}_r$ be the collection of all finite intersections of elements in $\bigcup_{i \in I_r} \mathcal{G}_i$. Then $\mathcal{S}_r$ is a π-system generating $\widetilde{\mathcal{G}}_r$. Furthermore, equation (III.2) holds since TODO ∎

Example. If $X_1, \dots, X_{10}$ are independent and

$$\begin{aligned} Y_1 &= f_1(X_1, X_2, X_3), \\ Y_2 &= f_2(X_4, X_5), \\ Y_3 &= f_3(X_6, X_7), \\ Y_4 &= f_4(X_8, X_9, X_{10}), \end{aligned}$$

where f_i's are measurable. Then Y_1, Y_2, Y_3, Y_4 are independent, since $\sigma(Y_1) \subseteq \sigma(X_1 \cup X_2 \cup X_3)$, etc.

Proposition III.17. *$\mathcal{G}_1, \dots, \mathcal{G}_n$ are independent if and only if*

$$\mathbf{E}[X_1 X_2 \dots X_n] = \mathbf{E}[X_1] \, \mathbf{E}[X_2] \dots \mathbf{E}[X_n]$$

for all random variables X_k measurable with respect to $\mathcal{G}_k$.

Proof. "if" is trivial, setting $X_k = \mathbf{1}_{A_k}$.

For "only if", we start with simple random variables. Since both sides are multilinear in X_k's and equality holds for indicators, it holds for simple functions.

Now let X_k's be positive random variables and choose simple random variables $X_{k,m} \uparrow X_k$. Then $X_{1,m}X_{2,m}\dots X_{n,m} \uparrow X_1X_2\dots X_n$ and by monotone convergence,

$$\begin{aligned}&\mathbf{E}[X_{1,m}X_{2,m}\dots X_{n,m}] = \mathbf{E}[X_{1,m}]\,\mathbf{E}[X_{2,m}]\dots\mathbf{E}[X_{n,m}]\\ \xRightarrow{\lim}\ &\mathbf{E}[X_1X_2\dots X_n] = \mathbf{E}[X_1]\,\mathbf{E}[X_2]\dots\mathbf{E}[X_n].\end{aligned}$$

■

Corollary III.18. *$X_1,\dots,X_n$ are independent if and only if*

$$\mathbf{E}[f_1(X_1)f_2(X_2)\dots f_n(X_n)] = \mathbf{E}[f_1(X_1)]\,\mathbf{E}[f_2(X_2)]\dots\mathbf{E}[f_n(X_n)]$$

for all $f_k\colon \mathbb{R}\to\mathbb{R}$ bounded and Borel measurable.

Theorem III.19 (Daniell-Kolmogorov)**.** *Given $\mu_1,\mu_2,\dots$ in $P(\mathbb{R})$, there exists a probability space $(\Omega,\mathcal{F},\mathbf{P})$ and random variables random variables $X_1,X_2,\dots\colon\Omega\to\mathbb{R}$ such that*

(1) $X_n\sim\mu_n$, and

(2) $X_1,X_2,\dots$ are independent.

Proof.

Case 1: $\mu_n = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$. Then $([0,1],\mathcal{B},\lambda)$ and

$$X_n(\omega) = n^{\text{th}} \text{ digit in the binary expansion of } \omega$$

works. TODO

Case 2: $\mu_n = [0,1]$.

Claim. *Suppose $(\Omega,\mathcal{F},\mathbf{P})$ and $\varepsilon_1,\varepsilon_2,\dots$ are independent* $\mathrm{Ber}(\frac{1}{2})$ *random variables. Then $Y_n = \sum_{k=1}^n \frac{\varepsilon_k}{2^k}\sim[0,1]$.*

Choose $([0,1],\mathcal{B},\lambda)$ and X_k's as before. Let

$$M = \begin{pmatrix}1 & 3 & 5 & 7 & \dots\\ 2 & 6 & 10 & 14 & \dots\\ 4 & 12 & 20 & 28 & \dots\\ 8 & 24 & 40 & 56 & \dots\\ \vdots & \vdots & \vdots & \vdots & \ddots\end{pmatrix}.$$

By the claim,

$$Y_n = \sum_{i=1}^{\infty}\frac{m_{ni}}{2^i}$$

are $[0,1]$. By lemma III.16, they are independent.

Case 3: $\mu_n \in P(\mathbb{R})$. Let G_n be the generalized inverse of the CDF of μ_n. Set

$$Z_n = G_n(Y_n).$$

Then $Z_n \sim \mu_n$ and $Z_1, Z_2, \ldots$ are independent.

■

Let $\mu_1, \mu_2, \mu_3 \in P(\mathbb{R}^2)$. Do there exist random variables X_1, X_2, X_3 on a common $(\Omega, \mathcal{F}, \mathbf{P})$ such that

$$(X, Y) \sim \mu_1,$$
$$(Y, Z) \sim \mu_2,$$
$$(Z, X) \sim \mu_3?$$

No, since $\mu_1(A \times \mathbb{R}) = \mu_3(\mathbb{R} \times A)$ for all $A \in \mathcal{B}(\mathbb{R})$ is necessary.

Fact III.20 (Kolmogorov's consistency theorem)**.** *For each n, let $\mu_n \in P(\mathbb{R}^n)$. Suppose that $\{\mu_n\}_{n\geq 1}$ are consistent in the sense that*

$$\mu_n \circ (\pi_1, \ldots, \pi_{n-1})^{-1} = \mu_{n-1} \quad \text{for all } n \geq 2.$$

Then there exists a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and random variables $X_1, X_2, \ldots$ such that

$$(X_1, \ldots, X_n) \sim \mu_n$$

for all n.

Example. Given an infinite symmetric and positive semi-definite matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \ldots \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \ldots \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

does there exist an $(\Omega, \mathcal{F}, \mathbf{P})$ and random variables $X_1, X_2, \ldots$ such that

$$(X_1, \ldots, X_n) \sim N_n(0, \Sigma_n)$$

for all n? (Where Σ_n is the top-left $n \times n$ submatrix of Σ.) Kolmogorov's consistency theorem says yes. Alternatively, we can construct them using iid standard normal random variables, which would not require this theorem.

Let $Z_1, Z_2, \ldots$ be iid standard normal random variables. Write $\Sigma = LL^\top$ where L is lower triangular. Define $X_1, X_2, \ldots$ by

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \end{pmatrix} = L \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ \vdots \end{pmatrix}$$

Lecture 13.
Tuesday
September 24

III.3 First and second moment methods

The first moment method is simply Markov's inequality. If $X \geq 0$, then

$$\mathbf{P}\{X \geq t\} \leq \frac{\mathbf{E}[X]}{t} \quad \text{for all } t > 0.$$

In particular,

$$\mathbf{P}\{X \geq k\,\mathbf{E}[X]\} \leq \frac{1}{k} \quad \text{for all } k > 0.$$

"X is within a few multiples of $\mathbf{E}[X]$ with high probability."

The other direction does not hold.

$$\mathbf{P}\{X \leq \frac{1}{2}\,\mathbf{E}[X]\} \leq \varepsilon$$

cannot be guaranteed for any ε. Consider

$$X_n \sim \Big(1 - \frac{1}{n}\Big)\delta_0 + \frac{1}{n}\delta_{n^2}. \tag{III.3}$$

Then

$$\mathbf{E}[X_n] = n, \text{ but}$$
$$\mathbf{P}\{X = 0\} = 1 - \frac{1}{n} \to 1.$$

If the second moment exists, we can use Chebyshev's inequality to get a bound.

$$\mathbf{P}\Big\{X \leq \frac{1}{2}\mu\Big\} \leq \Pr\Big\{|X - \mu| \geq \frac{1}{2}\mu\Big\} \leq \frac{\mathrm{Var}(X)}{\mu^2/4}.$$

In general,

$$\mathbf{P}\{X \leq \delta\mu\} \leq \mathbf{P}\{|X - \mu| \geq (1-\delta)\mu\} \leq \frac{\mathrm{Var}(X)}{(1-\delta)^2\mu^2}.$$

This only gives non-trivial bounds if $\mathrm{Var}(X) < (1-\delta)^2\mu^2$. If $\mathrm{Var}(X) \geq \mu^2$, no δ gives a non-trivial bound!

The second moment method refers to the Paley-Zygmund inequality.

Proposition III.21 (Paley-Zygmund inequality)**.** *For a random variable $X \geq 0$,*

$$\mathbf{P}\{X \geq \delta\,\mathbf{E}[X]\} \geq (1-\delta)^2\frac{\mathbf{E}[X]^2}{\mathbf{E}[X^2]} \quad \textit{for all } 0 \leq \delta \leq 1.$$

When does this give a non-trivial lower bound?

$$\frac{\mathbf{E}[X]^2}{\mathbf{E}[X^2]} = \frac{\mu^2}{\mu^2 + \sigma^2} = \frac{1}{1 + \sigma^2/\mu^2}.$$

This gives non-trivial bounds no matter how small σ^2/μ^2 is.

Proof. We first prove it for $\delta = 0$. Write $X = X\mathbf{1}_{X>0}$. Then from Cauchy-Schwarz,

$$\begin{aligned}\mathbf{E}[X]^2 &\le \mathbf{E}[X^2]\,\mathbf{E}[\mathbf{1}_{X>0}^2] \\ &= \mathbf{E}[X^2]\,\mathbf{P}\{X > 0\}.\end{aligned}$$

Now for any δ, let $Y = (X - \delta\mu)_+ \ge 0$. Then

$$\mathbf{P}\{Y > 0\} \ge \frac{\mathbf{E}[Y]^2}{\mathbf{E}[Y^2]}. \tag{III.4}$$

The LHS is $\mathbf{P}\{X \ge \delta\mu\}$. We need to relate moments of X and Y.

$$\begin{aligned}\mu = \mathbf{E}[X] &= \mathbf{E}[X\mathbf{1}_{X>\delta\mu}] + \mathbf{E}[X\mathbf{1}_{X\le\delta\mu}] \\ &\le \mathbf{E}[Y] + \mathbf{E}[\delta\mu\mathbf{1}_{X>\delta\mu}] + \mathbf{E}[\delta\mu\mathbf{1}_{X\le\delta\mu}] \\ &= \mathbf{E}[Y] + \delta\mu.\end{aligned}$$

Then

$$\mathbf{E}[Y] \ge (1-\delta)\mu.$$

Since $|Y| \le |X|$, we have

$$\mathbf{E}[Y^2] \le \mathbf{E}[X^2].$$

Thus equation (III.4) gives

$$\mathbf{P}\{X \ge \delta\mu\} \ge \frac{(1-\delta)^2\mu^2}{\mathbf{E}[X^2]}.$$ ■

III.3.1 Coupon collector

A box contains N coupons labelled $1, 2, \dots, N$. Draw repeatedly uniformly at random with replacement until all coupons are collected. That is, let $X_n \sim ([N])$ be iid, and define

$$T_N = \min\{t \in \mathbb{N} \mid \{X_1, \dots, X_t\} = [N]\}.$$

We wish to study T.

To analyze this, define $U_t = U_t^{(N)}$ as

$$U_t = N - \#\{X_1, \dots, X_t\}.$$

We can further write it as

$$U_t = \sum_{k=1}^{N} U_{t,k}, \quad U_{t,k} = \prod_{i=1}^{t} \mathbf{1}_{X_i \ne k}.$$

We can write $T > t \iff U_t \ge 1$, so

$$T_N = \min\{t \in \mathbb{N} \mid U_t = 0\}.$$

We shall use the first and second moment methods on U_t.

$$\begin{aligned}
\mathbf{E}[U_{t,k}] &= \prod_{i=1}^{t} \mathbf{P}\{X_i \neq k\} \\
&= \left(1 - \frac{1}{N}\right)^t \\
\implies \mathbf{E}[U_t] &= N\left(1 - \frac{1}{N}\right)^t \\
\mathbf{E}[U_t^2] &= \mathbf{E}\left[\sum_{k=1}^{N} U_{t,k} + \sum_{k \neq \ell} U_{t,k} U_{t,\ell}\right] \\
&= N\left(1 - \frac{1}{N}\right)^t + N(N-1)\left(1 - \frac{2}{N}\right)^t
\end{aligned}$$

Proposition III.22 (elementary inequalities)**.** *For all x,*

$$1 - x \leq e^{-x}.$$

For $|x| < \frac{1}{2}$,

$$e^{-x-x^2} \leq 1 - x \leq e^{-x}.$$

Thus we have

$$\mathbf{E}[U_t] \leq N e^{-t/N} = e^{-t/N + \log N}.$$

If $t = N(\log N + h_N)$ where $h_N \to \infty$ as $N \to \infty$, we have

$$\mathbf{E}[U_t] \leq e^{-h_N} \to 0.$$

Define this t to be t_N^+. Then

$$\mathbf{P}\{T_N > t_N^+\} = \mathbf{P}\{U_{t_N^+} \geq 1\} \leq \mathbf{E}[U_{t_N^+}] \to 0.$$

What if $t = N(\log N - h_N)$? Call this t_N^-.

$$\begin{aligned}
\mathbf{P}\{T_N > t_N^-\} &= \mathbf{P}\{U_{t_N^-} \geq 1\} \\
&\geq \frac{\mathbf{E}[U_t]^2}{\mathbf{E}[U_t^2]} \\
&\geq \frac{N^2 e^{-\frac{2t}{N} - \frac{2t}{N^2}}}{N e^{-\frac{t}{N}} + N(N-1) e^{-\frac{2t}{N}}} \\
&\geq \frac{e^{-\frac{2t}{N^2}}}{\frac{1}{N} e^{\frac{t}{N}} + \left(1 - \frac{1}{N}\right)}.
\end{aligned}$$

We have concluded that for any $h_N \to \infty$,

$$\boxed{\mathbf{P}\{N(\log N - h_N) \leq T_N \leq N(\log N + h_N)\} \to 1.}$$

III.4 Random graphs

Definition III.23 (Erdös-Rényi graph)**.** The *Erdös-Rényi random graph* $\mathcal{G}_{n,p}$ is the random graph on n vertices where each edge is present with probability p independently.

That is, we have random variables $\{X_{i,j} \mid 1 \le i < j \le n\}$ iid $\mathrm{Ber}(p)$.

Theorem III.24. *For a $\delta > 0$ and let $p_n^+ = (1+\delta)\frac{\log n}{n}$ and $p_n^- = (1-\delta)\frac{\log n}{n}$. Then*

$$\mathbf{P}\{\mathcal{G}_{n,p_n^+} \text{ is connected}\} \to 1$$
$$\mathbf{P}\{\mathcal{G}_{n,p_n^-} \text{ is connected}\} \to 0.$$

Define

$$\mathcal{C}_n = \text{number of connected components of size} \le \frac{n}{2}.$$
$$\mathcal{I}_n = \text{number of isolated vertices.}$$

Of course $\mathcal{I}_n \le \mathcal{C}_n$, and

$$\mathcal{G}_{n,p} \text{ is disconnected} \iff \mathcal{C}_n \ge 1 \impliedby \mathcal{I}_n \ge 1.$$

Lecture 15.
Tuesday
October 1

Definition III.25 (tail)**.** Fix a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a collection $(\mathcal{G}_n)_n$ of sub σ-algebras of $\mathcal{F}$. The *tail* of $(\mathcal{G}_1, \mathcal{G}_2, \dots)$ is the σ-algebra

$$\tau = \bigcap_{N=1}^{\infty} \sigma(\mathcal{G}_N, \mathcal{G}_{N+1}, \dots).$$

More usefully, if $\mathcal{G}_n = \sigma(Y_n)$ for some random variables $(Y_n)_n$, then

$$\tau = \bigcap_{N=1}^{\infty} \sigma(Y_N, Y_{N+1}, \dots).$$

A random variable X is *tail-measurable* iff

$$X = f_N(Y_N, Y_{N+1}, \dots) \text{ for some } f_N \text{ for all } N.$$

Examples.

- $X = \limsup_{n\to\infty} Y_n$ is tail-measurable.
- $X = \lim_{n\to\infty}(Y_1 + Y_2 + \cdots + Y_n)$ is *not* tail-measurable.
- $X = \lim_{n\to\infty} \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)$ is tail-measurable.

Theorem III.26 (Kolmogorov's 0–1 law)**.** *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $(\mathcal{G}_n)_n$ be* independent *sub σ-algebras of $\mathcal{F}$. Then the tail τ of $(\mathcal{G}_n)_n$ is trivial, i.e., $\mathbf{P}(A) \in \{0, 1\}$ for all $A \in \tau$.*

Corollary III.27. *If $(Y_n)_n$ are independent random variables, then any tail random variable is almost surely constant.*

Proof. If $\mathbf{P}(A) \in \{0,1\}$ for all $A \in \tau$, then for any tail-measurable X, the event $\{X \le t\}$ is in τ and hence has probability 0 or 1. Thus the CDF of X is either 0 or 1 for all t.

$$F_X(t) = \begin{cases} 0 & t < t_0 \\ 1 & t \ge t_0 \end{cases} \text{ for some } t_0 \in \overline{\mathbb{R}}.$$

Thus $X = t_0$ almost surely. ■

Proof of Kolmogorov's 0–1 law. Since

$$\mathcal{G}_1, \dots, \mathcal{G}_N, \mathcal{G}_{N+1}, \dots \text{ are independent,}$$

so are

$$\mathcal{G}_1, \dots, \mathcal{G}_N, \sigma(\mathcal{G}_{N+1}, \mathcal{G}_{N+2}, \dots).$$

But $\tau \subseteq \sigma(\mathcal{G}_{N+1}, \mathcal{G}_{N+2}, \dots)$, so

$$\mathcal{G}_1, \dots, \mathcal{G}_N, \tau \text{ are independent.}$$

This holds for all N.

An infinite set of σ-algebras is independent iff any finite subset is independent. Thus

$$\tau, \mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \dots \text{ are independent.}$$

Chunking them up again, we have that

$$\tau, \sigma(\mathcal{G}_1 \cup \mathcal{G}_2 \cup \dots) \text{ are independent.}$$

Again τ is a subset of the second, so finally

$$\tau, \tau \text{ are independent.}$$

That is, $\mathbf{P}(A \cap B) = \mathbf{P}(A)\,\mathbf{P}(B)$ for all $A, B \in \tau$. Thus $\mathbf{P}(A) \in \{0,1\}$ for all $A \in \tau$. ■

III.4.1 Percolation

Fix a $p \in [0,1]$. For each edge e of $\mathbb{Z}^d$, let $X_e \sim \mathrm{Ber}(p)$ independently. Let G be the random graph with vertex set $\mathbb{Z}^d$ and edge set $\{e \mid X_e = 1\}$. Does G have an infinite cluster (connected component)?

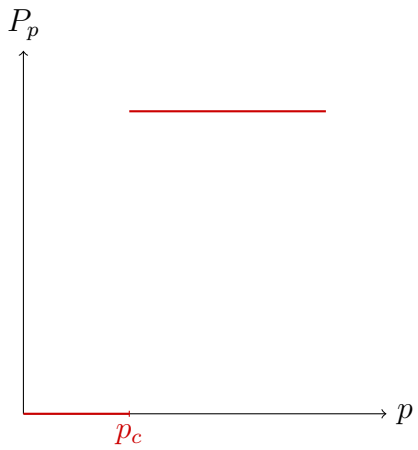
This has probability 0 or 1. Enumerate the edges as $e_1, e_2, \dots$ and let the corresponding random variables be $Y_n = X_{e_n}$. Finitely many edges do not affect the existence of an infinite cluster, so the existence of an infinite cluster is a tail event with respect to $(Y_n)_n$. Since Y_n are independent, the existence of an infinite cluster is a 0–1 event.

Remark. This tries to model a fluid percolating through a porous medium. If the fluid is allowed to flow only through the edges of the graph, then the question is whether the fluid can flow from top to bottom.

Let P_p be the probability that there is an infinite cluster in G for a given $p \in [0,1]$. Observe that P_p is increasing in p (coupling). Thus there is a critical value $p_c \in [0,1]$ such that

$$P_p = \begin{cases} 0 & \text{if } p < p_c, \\ 1 & \text{if } p > p_c, \end{cases}$$

and $P_{p_c} \in \{0,1\}$.



This shows a discontinuous behaviour, where none was present in the underlying model. Close links to phase transitions.

III.4.2 Random independent series

Let $(X_n)_n$ be a sequence of independent random variables. Consider the event that $\sum_{n=1}^{\infty} X_n$ converges. This is a tail event, so it has probability 0 or 1 (independence).

Examples.

- $X_n = c^n \xi_n$, where $c \in \mathbb{R}$ and $\xi_n \sim \mathrm{Ber}(\frac{1}{2})$ are independent.
  - If $|c| < 1$, then $\sum_{n=1}^{\infty} X_n$ converges.
  - If $c = 1$, then $\sum_{n=1}^{\infty} X_n$ diverges almost surely.

Theorem III.28 (Khinchine)**.** *Let $(X_n)_n$ be independent random variables with finite variances and zero means. If $\sum_{n=1}^{\infty} \mathrm{Var}(X_n) < \infty$, then $\sum_{n=1}^{\infty} X_n$ converges almost surely.*

Examples.

- Let $\xi_n \sim \mathrm{Ber}_\pm(\frac{1}{2})$ independently. Fix an $\alpha > 0$. Let $X_n = \frac{\xi_n}{n^\alpha}$. Then $\mathbf{E}\, X_n = 0$ and $\operatorname{Var} X_n = \frac{1}{n^{2\alpha}}$. Thus if $\alpha > \frac{1}{2}$, the series converges almost surely.

  Note that the alternating series $\sum_{n=1}^\infty \frac{(-1)^n}{n^\alpha}$ converges for all $\alpha > 0$.

Proof. $\sum X_n$ converges iff $(S_n)_n$ is Cauchy.

$(t_n)_n$ is not Cauchy iff there exists an $\varepsilon > 0$ such that for any $N \in \mathbb{N}$, there exists a $k \in \mathbb{N}$ such that $|t_{N+k} - t_N| \geq \varepsilon$. Then

$$\{(S_n)_n \text{ is not Cauchy}\} = \bigcup_{m=1}^\infty \bigcap_{N=1}^\infty \bigcup_{k=1}^\infty \{|S_{N+k} - S_N| \geq \frac{1}{m}\}$$

To show that this has probability 0 is the same as showing that

$$\bigcap_{N=1}^\infty \bigcup_{k=1}^\infty \{|S_{N+k} - S_N| \geq \frac{1}{m}\} \text{ has probability } 0$$

for each $m \geq 1$. It suffices to show that for each $m \geq 1$,

$$\mathbf{P}\left(\bigcup_{k=1}^\infty \{|S_{N+k} - S_N| \geq \frac{1}{m}\}\right) \to 0 \text{ as } N \to \infty.$$

We fix $\varepsilon = \frac{1}{m}$ and $N \geq 1$ and compute this probability. This is the same as

$$\mathbf{P}\left\{\sup_{k\geq 1} |S_{N+k} - S_N| \geq \varepsilon\right\}.$$

By Kolmogorov's maximal inequality,

$$\mathbf{P}\left\{\sup_{k\geq 1} |S_{N+k} - S_N| \geq \varepsilon\right\} \leq \frac{2}{\varepsilon^2} \sum_{j=N+1}^\infty \operatorname{Var}(X_j) \to 0.$$

This is by reducing to the finite case

$$\mathbf{P}\left\{\sup_{k\geq 1} |S_{N+k} - S_N| \geq \varepsilon\right\} \leq \lim_{M\uparrow\infty} \mathbf{P}\left\{\max_{1\leq k\leq M} |S_{N+k} - S_N| \geq \varepsilon\right\}.$$

This proves the result. ∎

Theorem III.29 (Kolmogorov's maximal inequality)**.** *Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with zero means and finite variances. Let $S_k = Y_1 + \cdots + Y_k$. Then*

$$\mathbf{P}\left\{\max_{1\leq k\leq n} |S_k| \geq t\right\} \leq \frac{\sum_{k=1}^n \operatorname{Var}(Y_k)}{t^2} \text{ for } t > 0.$$

Chebyshev's inequality gives

$$\mathbf{P}\{S_n \geq t\} \leq \frac{\mathbf{E}[S_n^2]}{t^2} = \frac{\sum_{k=1}^n \operatorname{Var} Y_k}{t^2}.$$

The maximum of S_k could be much larger than S_n.

Lecture 17. Tuesday October 8

Definition III.30 (convergence)**.** Let $(X_n)_n$ and X be random variables on $(\Omega, \mathcal{F}, \mathbf{P})$. We say that

(1) (almost sure convergence) $X_n \xrightarrow{\text{a.s.}} X$

(2) (convergence in probability) $X_n \xrightarrow{\mathsf{P}} X$

(3) (convergence in distribution) $X_n \xrightarrow{\text{d}} X$

(4) (L_p convergence) $X_n \xrightarrow{L^p} X$ $(p \geq 0)$

Proposition III.31. *(1) Almost sure convergence implies convergence in probability.*

(2) Convergence in probability implies convergence in distribution.

(3) L^p convergence implies convergence in probability.

Proposition III.32. *(1) If X is an almost sure constant, then $X_n \xrightarrow{d} X$ implies $X_n \xrightarrow{\mathsf{P}} X$.*

(2) If $(X_n)_n$ are uniformly integrable, then $X_n \xrightarrow{\mathsf{P}} X$ implies $X_n \xrightarrow{L^1} X$.

Proposition III.33. *If convergence in probability is* fast*, i.e.,*

$$\sum_n \mathbf{P}(|X_n - X| > \varepsilon) < \infty \text{ for all } \varepsilon > 0,$$

then $X_n \xrightarrow{\mathsf{P}} X$ implies $X_n \xrightarrow{a.s.} X$.

Definition III.34 (uniform integrability)**.** A family $\{X_\alpha \mid \alpha \in I\}$ of integrable random variables is *uniformly integrable* if for any $\varepsilon > 0$ there exists an $M < \infty$ such that

$$\mathbf{E}[|X_\alpha| \mathbf{1}_{|X_\alpha| > M}] < \varepsilon \text{ for all } \alpha \in I.$$

Remarks.

- If $\mathrm{d}Q_\alpha = |X_\alpha| \,\mathrm{d}\,\mathbf{P}$ (that is, $Q_\alpha(A) = \int_A |X_\alpha| \,\mathrm{d}\,\mathbf{P}$), then uniform integrability is equivalent to the condition that ???
- A single integrable random variable forms a uniformly integrable family.
- Uniform integrability implies the following: given $\varepsilon > 0$, there exists a $\delta > 0$ such that for any $A \in \mathcal{F}$ with $\mathbf{P}(A) < \delta$, we have $\mathbf{E}[|X_\alpha| \mathbf{1}_A] < \varepsilon$ for all $\alpha \in I$. Pf

- If I is finite, uniform integrability holds. (If $\mathbf{E}|X| < \infty$, then $\mathbf{E}[|X|\mathbf{1}_{|X|>M}] \to 0$ as $M \to \infty$.) More generally, if $\{X_\alpha\}_{\alpha\in I}$ is uniformly integrable and $\{Y_\beta\}_{\beta\in J}$ is a finite integrable family, then the union is uniformly integrable.

- The family of pairwise sums of two uniformly integrable families is uniformly integrable.

  Proof. Let $\{X_\alpha\}_\alpha$ and $\{Y_\beta\}_\beta$ be uniformly integrable. Then for any α, β,

  $$|X_\alpha + Y_\beta|\mathbf{1}_{|X_\alpha+Y_\beta|>M} \le 2|X_\alpha|\mathbf{1}_{|X_\alpha|>M} + 2|Y_\beta|\mathbf{1}_{|Y_\beta|>M}. \qquad \blacksquare$$

- If $\{X_\alpha\}_{\alpha\in I}$ are dominated by an integrable random variable Y, that is, $|X_\alpha| \le Y$ almost surely for all α, then $\{X_\alpha\}_\alpha$ is uniformly integrable.

  The converse is not true. For example, consider $X_n = n\mathbf{1}_{[\frac{1}{n},\frac{1}{n-1}]}$ on $([0,1],\mathcal{B},\lambda)$. Then $\{X_n\}_n$ is uniformly integrable, but not dominated by any integrable random variable.

- If $\{X_\alpha\}_\alpha$ is bounded in L^p for some $p > 1$, then it is uniformly integrable.

  $$\sup_\alpha \mathbf{E}[|X_\alpha|^p] < \infty \implies \{X_\alpha\}_\alpha \text{ is uniformly integrable.}$$

  Proof. By Markov's inequality,

  $$\mathbf{E}[|X_\alpha|\mathbf{1}_{|X_\alpha|>M}] \le \frac{\mathbf{E}[|X_\alpha|^p\mathbf{1}_{|X_\alpha|>M}}{M^{p-1}} \le \frac{\mathbf{E}[|X_\alpha|^p]}{M^{p-1}} \to 0$$

  as $M \to \infty$, since the numerator is uniformly bounded. $\blacksquare$

  This is false for $p = 1$. Consider $X_n = n\mathbf{1}_{[0,\frac{1}{n}]}$ on $([0,1],\mathcal{B},\lambda)$. $\{X_n\}_n$ is bounded in L^1 since each expectation is 1. But for any M, $\mathbf{E}[X_n\mathbf{1}_{|X_n|>M}] = 1$ for all $n > M$.

- If $X_n \xrightarrow{L^1} X$, then $\{X_n\}_n$ is uniformly integrable.

  Proof. Suppose $X = 0$. Then $\mathbf{E}[|X_n|] \to 0$. Given $\varepsilon > 0$, find an N such that $\mathbf{E}[|X_n|] < \varepsilon$ for all $n > N$. For the first N terms, simply take the maximum M_n such that $\mathbf{E}[|X_n|\mathbf{1}_{|X_n|>M_n}] < \varepsilon$. Thus $\{X_n\}_n$ is uniformly integrable.

  If X is non-zero, then $\{X_n - X\}_n \xrightarrow{L^1} 0$. Thus $\{X_n - X\}_n$ is uniformly integrable. But the singleton $\{X\}$ is uniformly integrable. Thus the pairwise sum $\{X_n\}_n$ is uniformly integrable. $\blacksquare$

Theorem III.35. *Let X_n, X be integrable random variables on $(\Omega, \mathcal{F}, \mathbf{P})$. Then*

$$X_n \xrightarrow{L^1} X \iff \begin{cases} X_n \xrightarrow{\mathsf{P}} X, \text{ and} \\ \{X_n\}_n \text{ is uniformly integrable.} \end{cases}$$

Proof. The forward implication has already been proved.

Let $X_n \xrightarrow{\mathsf{P}} X$ and $\{X_n\}_n$ be uniformly integrable. Define $Y_n := X_n - X$. Then $Y_n \xrightarrow{\mathsf{P}} 0$ and $\{Y_n\}_n$ is uniformly integrable. We need to show that $\mathbf{E}|Y_n| \to 0$.

$$\mathbf{E}|Y_n| = \mathbf{E}[|Y_n|\mathbf{1}_{|Y_n|>M}] + \mathbf{E}[|Y_n|\mathbf{1}_{|Y_n|\le M}].$$

Prove via almost sure subsequences and DCT

The first term can be made arbitrarily small by uniform integrability. For the second term,

$$\begin{aligned} \mathbf{E}[|Y_n|\mathbf{1}_{|Y_n|\le M}] &= \int_0^\infty \mathbf{P}\{|Y_n|\mathbf{1}_{|Y_n|\le M} > t\}\,\mathrm{d}t \\ &= \int_0^M \mathbf{P}\{|Y_n| > t\}\,\mathrm{d}t. \end{aligned}$$

We know that for any $t > 0$, $\mathbf{P}\{|Y_n| > t\} \to 0$ Since probabilities are bounded by 1, we can apply the dominated convergence theorem (on Lebesgue integrals) to get

$$\mathbf{E}[|Y_n|\mathbf{1}_{|Y_n|\le M}] \to 0. \qquad \blacksquare$$

As a corollary of everything,

Corollary III.36. *Suppose $X_n \xrightarrow{a.s.} X$. Then $X_n \xrightarrow{L^1} X$ iff $\{X_n\}_n$ is uniformly integrable.*

III.5 Laws of large numbers

Fix a space $(\Omega, \mathcal{F}, \mathbf{P})$ and a sequence of i.i.d. random variables $X_1, X_2, \ldots$ on this space. Let $S_n = \sum_{i=1}^n X_i$. The weak law of large numbers states that

$$\frac{S_n}{n} \xrightarrow{\mathsf{P}} \mathbf{E}[X_1].$$

The strong law of large numbers states that

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mathbf{E}[X_1].$$

Obviously, the strong law implies the weak law.

> If these were not true, we would have to rework probability.
> — Prof. Manjunath Krishnapur

Proof attempt. Assume X_k's have finite variance σ^2. WLOG take $\mu = 0$ (else replace X_k with $X_k - \mu$). Then $\mathbf{E}[X_k] = 0$ and $\mathrm{Var}(X_k) = \sigma^2$. Then S_n has mean 0 and variance $n\sigma^2$.

$$\mathbf{E}[S_n^2] = \sum_i \mathbf{E}[X_i^2] + \sum_{i \neq j} \mathbf{E}[X_i X_j] = n\sigma^2.$$

This is the "square root law". Thus

$$\mathbf{E}[(S_n/n)^2] = \frac{\sigma^2}{n} \to 0.$$

Since L^2 convergence implies convergence in probability, we have

$$\frac{S_n}{n} \xrightarrow{\mathsf{P}} 0.$$

If $n_1 < n_2 < \dots$ is a subsequence such that $\sum_k \frac{1}{n_k} < \infty$, then

$$\mathbf{P}\{|S_{n_k}/n_k| > \delta\} \leq \frac{\sigma^2}{\delta^2 n_k}$$

is summable, so $\frac{S_{n_k}}{n_k} \xrightarrow{\text{a.s.}} 0$.

Alternatively, if we assume a higher moment condition, say fourth moment, then

$$\begin{aligned}
\mathbf{E}[S_n^4] &= \sum_{i,j,k,l} \mathbf{E}[X_i X_j X_k X_l] \\
&= \sum_i \mathbf{E}[X_i^4] + 3 \sum_{i \neq j} \mathbf{E}[X_i^2]\, \mathbf{E}[X_j^2] \\
&= nB + 3n(n-1)\sigma^4
\end{aligned}$$

Thus

$$\begin{aligned}
\mathbf{P}\{|S_n/n| \geq \delta\} &\leq \frac{\mathbf{E}[S_n^4]}{(n\delta)^4} \\
&\leq
\end{aligned}$$

■

Lecture 20. Thursday October 17

Theorem III.37 (strong law of large numbers)**.** *Let $X_1, X_2, \dots$ be i.i.d with $\mathbf{E}[X_i] = 0$. Then $\frac{S_n}{n} \xrightarrow{a.s.} 0$*

Proof. First reduce this to the case when all $X_i \geq 0$.

If $\lambda > 1$ and $n_k = \lfloor \lambda^k \rfloor$, then

$$\frac{S_{n_k}}{n_k} \xrightarrow{\text{a.s.}} \mathbf{E}[X_1].$$

If $X_i \geq 0$ and $n_k \leq n < n_{k+1}$, then

$$S_{n_k} \leq S_n \leq S_{n_{k+1}}, \text{ so } \frac{S_{n_k}}{n_{k+1}} \leq \frac{S_n}{n} \leq \frac{S_{n_{k+1}}}{n_k}.$$

Thus for all $\lambda > 1$,

$$\frac{\mu}{\lambda} \le \liminf \frac{S_n}{n} \le \limsup \frac{S_n}{n} \le \lambda\mu \text{ a.s.}$$

Take intersection over $\lambda = 1 + \frac{1}{j}$ to get

$$\lim \frac{S_n}{n} = \mu \text{ a.s.} \qquad \blacksquare$$

III.5.1 Two extensions in different directions

(1) Let $X_1, X_2, \dots$ be i.i.d. with $\mathbf{E}[X_i] = 0$. For which α does $\frac{S_n}{n^\alpha} \xrightarrow{\text{a.s.}} 0$? For which f does $\frac{S_n}{f(n)} \xrightarrow{\text{a.s.}} 0$? What more assumptions are needed?

(2) In the same setting with any more assumptions necessary, what is the rate of convergence of $\frac{S_n}{n}$? How does $\mathbf{P}\left\{|\frac{S_n}{n}| > \delta\right\}$ go to 0? Chebyshev gives a $\frac{1}{n}$ convergence assuming second moment. Can we do better?

Recall the Möbius function

$$\mu(n) = \begin{cases} 1 & n = p_1 p_2 \dots p_{2k} \text{ distinct primes,} \\ -1 & n = p_1 p_2 \dots p_{2k+1} \text{ distinct primes,} \\ 0 & n \text{ is not square-free.} \end{cases}$$

Claim (Riemann hypothesis)**.** $\sum_{i=1}^n \mu(i) = O(n^{1/2+\varepsilon})$ *for all* $\varepsilon > 0$.

We will prove this in the next lecture. Let $X_1, X_2, \dots$ be i.i.d. $\text{Ber}_\pm(\frac{1}{2})$. Since X_i is bounded, all moments exist.

Let $S_n = X_1 + \cdots + X_n$. We have seen

$$\begin{aligned} &\mathbf{E}[S_n^2] = n \\ \implies &\mathbf{P}\left\{\left|\frac{S_n}{n}\right| \ge \delta\right\} \le \frac{1}{n\delta^2}. \\ &\mathbf{E}[S_n^4] = 3n(n-1) \\ \implies &\mathbf{P}\left\{\left|\frac{S_n}{n}\right| \ge \delta\right\} \le \frac{3n(n-1)+n}{n^4\delta^4}. \end{aligned}$$

More generally,

$$\mathbf{P}\left\{\left|\frac{S_n}{n^\alpha}\right| \ge \delta\right\} \le \frac{3n(n-1)+n}{n^{4\alpha}\delta^4} \le \frac{C}{\delta^4 n^{4\alpha-2}}.$$

This is summable for $\alpha > \frac{3}{4}$. What if we use higher moments?

$$\begin{aligned} \mathbf{E}[S_n^6] &= 5!!n(n-1)(n-2) + O(n^2) \\ \implies \mathbf{P}\left\{\left|\frac{S_n}{n}\right| \geq \delta\right\} &\leq \frac{5!!n(n-1)(n-2) + O(n^2)}{n^6\delta^6} \\ &\leq \frac{C}{\delta^6 n^{6\alpha-3}}, \end{aligned}$$

which is summable for $\alpha > \frac{2}{3}$.

$$\begin{aligned} \mathbf{E}[S_n^8] &= 7!!n(n-1)(n-2)(n-3) + O(n^3) \\ \implies \mathbf{P}\left\{\left|\frac{S_n}{n}\right| \geq \delta\right\} &\leq \frac{7!!n(n-1)(n-2)(n-3) + O(n^3)}{n^8\delta^8} \\ &\leq \frac{C}{\delta^8 n^{8\alpha-4}}, \end{aligned}$$

which is summable for $\alpha > \frac{5}{8}$. More generally,

$$\begin{aligned} \mathbf{E}[S_n^{2p}] &= (2p-1)!!n^p + O(n^{p-1}) \\ \implies \mathbf{P}\left\{\left|\frac{S_n}{n}\right| \geq \delta\right\} &\leq \frac{C_p}{\delta^{2p} n^{2p\alpha-p}}, \end{aligned}$$

Which is summable for $\alpha > \frac{1+p}{2p}$.

Thus for any $\alpha > \frac{1}{2}$, $\frac{S_n}{n^\alpha} \xrightarrow{\text{a.s.}} 0$.

Exercise III.38. *$\alpha = \frac{1}{2}$ does not work.*

We can do better using Hoeffding's inequality. Recall that if $|X_i| \leq d_i$ are independent with zero mean, then

$$\mathbf{P}\{S_n \geq t\} \leq \exp\left(-\frac{t^2}{2\sum d_i^2}\right).$$

Thus

$$\mathbf{P}\left\{\left|\frac{S_n}{n^\alpha}\right| \geq \delta\right\} \leq 2\exp\left(-\frac{n^{2\alpha}\delta^2}{2n}\right) = 2\exp\left(-\frac{\delta^2}{2}n^{2\alpha-1}\right).$$

This is summable for $\alpha > \frac{1}{2}$. All the moment crunching in one shot. Do better!

$$\mathbf{P}\left\{\left|\frac{S_n}{\sqrt{nh(n)}}\right| \geq \delta\right\} \leq 2\exp\left(-\frac{\delta^2}{2}h(n)\right)$$

This is summable for $h(n) \geq (\log n)^{1+\varepsilon}$. Thus

$$\frac{S_n}{\sqrt{n(\log n)^{1+\varepsilon}}} \xrightarrow{\text{a.s.}} 0.$$

In fact, this proof works for all bounded X_i.

Fact III.39 (Khinchin's law of the iterated logarithm)**.** *Let $X_1, X_2, \dots$ be i.i.d. with mean 0 and variance 1. Then*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n \log\log n}} = \sqrt{2} \text{ a.s.}$$

Remark. Khinchin only proved this for Bernoullis. The general case is due to Hartman and Wintner.

Suppose $A_1, A_2, \dots$ are independent events. Then

$$\mathbf{P}\{A_n \text{ i.o}\} = \begin{cases} 0 & \text{if } \sum \mathbf{P}(A_n) < \infty, \\ 1 & \text{if } \sum \mathbf{P}(A_n) = \infty. \end{cases}$$

Let $B_1, B_2, \dots$ be such that

$$\begin{aligned} B_1 &= A_1, \\ B_2 = B_3 &= A_2, \\ B_4 = B_5 = B_6 &= A_3, \\ &\vdots \end{aligned}$$

Then $\{B_n \text{ i.o.}\} = \{A_n \text{ i.o.}\}$. Borel-Cantelli gives that if $\sum n\,\mathbf{P}(A_n) < \infty$, then B_n occur infinitely often with probability 0. This is a weaker conclusion that Borel-Cantelli on the A_n's.

Khinchin proved his thoerem by reverse engineering this. S_n's barely change with neighbouring n's. Khinchin managed to create blocks of n's where S_n's are almost constant and independent of each other.

We want bounds for $\mathbf{P}\Big\{\Big|\frac{S_n}{n} - \mu\Big| \geq \delta\Big\}$.

Example. Let X_i be iid $\text{Ber}(\frac{1}{2})$. Fix a $k \in n+1$. Then $\mathbf{P}\{S_n = k\} = \binom{n}{k}\frac{1}{2^n}$ By the Strirling approximation,

$$m! \sim \sqrt{2\pi m}\Big(\frac{m}{e}\Big)^m.$$

That is, $\frac{m!}{\sqrt{2\pi} m^{m+\frac{1}{2}} e^{-m}} \to 1$. Assume $1 \ll k \ll n$ ($n \to \infty$ and $n - k \to \infty$).

Then

$$\begin{aligned}\mathbf{P}\{S_n = k\} &\sim \frac{1}{\sqrt{2\pi}2^n} \frac{n^{n+\frac{1}{2}}e^{-n}}{k^{k+\frac{1}{2}}e^{-k}(n-k)^{n-k+\frac{1}{2}}e^{-(n-k)}} \\ &\sim \frac{\sqrt{n}}{\sqrt{2\pi}\sqrt{k}\sqrt{n-k}} \frac{n^n}{k^k(n-k)^{n-k}2^n} \\ &= \frac{\sqrt{n}}{\sqrt{2\pi k(n-k)}} e^{-n\left[\log 2 + \frac{n}{k}\log\frac{n}{k} + \left(1-\frac{k}{n}\right)\log\left(1-\frac{k}{n}\right)\right]}\end{aligned}$$

Let $H(p) = -p\log p - (1-p)\log(1-p)$. Then $\mathbf{P}\{S_n = k\} = C_{n,k}e^{-nH(\frac{k}{n})}$, where $C_{n,k}$ is polynomially bounded in n and k.

If $\frac{1}{2} < p < 1$, then

$$\mathbf{P}\{\frac{S_n}{n} \geq p\} = \sum_{k=np} \mathbf{P}\{S_n = k\}.$$

This is lower-bounded by the first term, and upper-bounded by n times the first term. Taking logarithms,