MA 361: Probability Theory

Naman Mishra

August 2024

Contents

The course

Grading

Lecture 1.
Thursday
August 1

- 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 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 sketch. 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 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.

	$\varnothing, \Omega$	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

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, $\mathcal{G}$ is a σ-algebra and thus $\mathcal{G} \supseteq \mathcal{F}$. ■

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

This is due to Sierpiński and Dynkin.

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 is closed under complements and finite unions.

Theorem II.15 (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.16. *The above extension is unique.*

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

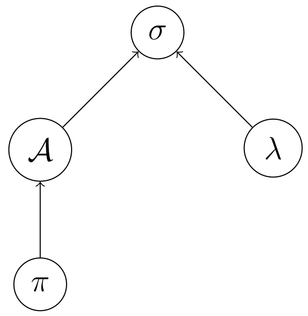


Figure II.1: Heirarchy of systems of sets under inclusion

II.3 Existence of Lebesgue measure

Theorem II.17. *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 $A_n \uparrow A$.

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.18. 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.19. *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.20 (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.21. *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

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

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

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

  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

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

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)

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

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

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

Thus

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

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

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

Lecture 5. Tuesday August 20

Definition II.22. 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.23. *Let $\mu_n, \mu \in P(\mathbb{R})$. Then*

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

Examples.

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

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

  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

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

for all large n. Thus we have

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

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

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

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.24. *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.25. 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 $\{N(\mu, 1)\}_{-16\leq\mu\leq32768}$ is tight.
- $\{N(\mu, \sigma^2)\}_{-10\leq\mu\leq10}$ is not tight, but $\{N(\mu, \sigma^2)\}_{\substack{-10\leq\mu\leq10 \\ 0<\sigma<10}}$ is.
- $\{\delta_{\frac{1}{n}}\}_{n\in\mathbb{N}}$ is tight.

Definition II.26. A set $E \subseteq (X, d)$ is *pre-compact* if its closure $\overline{E}$ is compact.

Theorem II.27. *Any $\mathcal{A} \subseteq P(\mathbb{R})$ is pre-compact iff it is tight.*

We will cover two prerequisites before we prove this theorem.

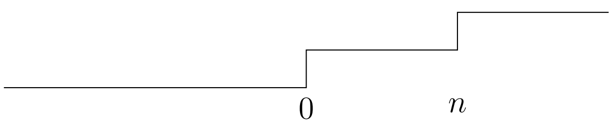
Theorem II.28 (Helly's selection principle)**.** *Let $\mu_n \in P(\mathbb{R})$. Then there is a subsequence $n_1 < n_2 < \ldots$ and an increasing, right continuous function $F \colon \mathbb{R} \to [0, 1]$ such that*

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

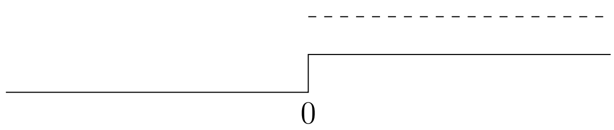
F may be a "defective CDF". It need not go to 0 to the left, nor 1 to the right.

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 $(n_{1,k})_{k\in\mathbb{N}}$, $(n_{2,k})_{k\in\mathbb{N}}$, $(n_{3,k})_{k\in\mathbb{N}}$, $\ldots$.

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

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 $F(x) \leq F(y) \leq F(x) + \varepsilon$. ■

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.27).

Lecture 6. Thursday August 22

Lemma II.29. *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. ($\implies$) 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.

($\impliedby$) 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.27.

Proof of theorem II.27. ($\implies$) 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.

($\impliedby$) 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 $X: \Omega \to \overline{\mathbb{R}}$ such that for each $B \in \mathcal{B}(\bar{R})$, $X^{-1}(B) \in \mathcal{F}$.

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 $X: (\Omega, \mathcal{F}) \to \left(\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}})\right)$ is measurable iff*

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

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) $\{\omega \mid \lim X_n \text{ exists}\}$

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*

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

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

(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*

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

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

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

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

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

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

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

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

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

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 $\mathrm{Var}(X) \geq 0$.

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 $q = \frac{p}{p-1}$.

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

■

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.

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

■

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 $x_1, \ldots, x_n$ on sets $A_1, \ldots, A_n \in \mathcal{F}$, 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

Tutorial 2: hmm

- $\mathbf{E}[X]$ 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

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