PROBABILITY THEORY LECTURE NOTES

JOHN PIKE

These lecture notes were written for MATH 650 at Bridgewater State University in the Fall semester of 2025, and the schedule on the following page reflects that semester. Much of the material and structure (as well as some of the language) comes directly from the course text, *A First Look at Rigorous Probability Theory* by Jeffrey Rosenthal, as well as *Probability: Theory and Examples* by Rick Durrett and *Real Analysis: Modern Techniques and Their Applications* by Gerald Folland. These notes likely contain typos and mistakes. All such errors are mine and corrections are greatly appreciated.

Day 1: To sub-σ-fields

Day 2: Through Theorem 2.1

Day 3: Finished Section 2

Day 4: Through Example 3.5

Day 5: To Proposition 3.10

Day 6: Through Theorem 3.12

Day 7: Through Borels regular

Day 8: To Proposition 4.3

Day 9: Sick Day

Day 10: Through Theorem 4.9

Day 11: Through independent collections

Day 12: To Example 6.1

Day 13: Through Theorem 6.5

Day 14: Finished Section 6

Day 15: Through Corollary 7.6

Day 16: Through Fubini-Tonelli

Day 17: To WLLN

Day 18: Through Example 8.11

Day 19: Through first half of Theorem 9.2

Day 20: To Example 9.8

Day 21: To Lemma 10.1

Day 22: Through Fact 10.4

Day 23: Through definition of conditional expectation

Day 24: Through Example 11.10

Day 25: Finished Section 11

Day 26: Student Presentations

Day 27: Student Presentations

1 Introduction

Probability Spaces

A *probability space* is a measure space $(\Omega, \mathcal{F}, P)$ with $P(\Omega) = 1$.

The *sample space* Ω can be any set and is generally thought of as the collection of all possible outcomes of some experiment or all possible states of some system. Elements of Ω are referred to as *elementary outcomes*. The general idea is that we know all outcomes that could occur in principle, but not which one actually does.

The *σ-field* (or σ-algebra) $\mathcal{F} \subseteq 2^\Omega$ satisfies

(1) $\mathcal{F}$ is nonempty
(2) $E \in \mathcal{F} \Rightarrow E^C \in \mathcal{F}$
(3) For any countable collection $\{E_i\}_{i \in I} \subseteq \mathcal{F}$, $\bigcup_{i \in I} E_i \in \mathcal{F}$.

(Since $\bigcap_{i \in I} E_i = \left(\bigcup_{i \in I} E_i^C\right)^C$, $\mathcal{F}$ is also closed under countable intersections.)

Elements of $\mathcal{F}$ are called *events*, and can be regarded as sets of elementary outcomes about which one can say something meaningful. Before the experiment has been performed, a meaningful statement about $E \in \mathcal{F}$ is $P(E)$. Afterward, a meaningful statement is whether or not E occurred—that is, whether the experiment resulted in an outcome $\omega \in E$.

The *probability measure* $P : \mathcal{F} \to [0, 1]$ satisfies

(1) $P(\Omega) = 1$
(2) For any countable disjoint collection $\{E_i\}_{i \in I}$, $P\left(\bigsqcup_{i \in I} E_i\right) = \sum_{i \in I} P(E_i)$.

The interpretation is that $P(A)$ represents the chance that event A occurs (though there is no general consensus about what that actually means).

Example 1.1. Rolling a fair die: $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = 2^\Omega$, $P(E) = \frac{|E|}{6}$.

Example 1.2. Flipping a (possibly biased) coin: $\Omega = \{H, T\}$, $\mathcal{F} = 2^\Omega = \big\{\emptyset, \{H\}, \{T\}, \{H, T\}\big\}$, P satisfies $P(\{H\}) = p$ and $P(\{T\}) = 1 - p$ for some $p \in [0, 1]$. (If $p \in \{0, 1\}$, then the outcome is guaranteed in advance; deterministic processes also fit within the probability framework.)

Example 1.3. Random point in the unit interval: $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}_{[0,1]} =$ Borel Sets, $P =$ Lebesgue measure. The experiment here is to pick a real number between 0 and 1 uniformly at random. Uniformity corresponds to translation invariance, which is the primary defining property of Lebesgue measure. Indeed, one can fully characterize this uniform probability measure by requiring that for each $0 \leq a \leq b \leq 1$, $P([a, b]) = b - a$. Observe that each outcome $x \in [0, 1]$ has $P(\{x\}) = P([x, x]) = x - x = 0$, so the experiment must result in the realization of an outcome with probability zero.

Example 1.4. Standard normal distribution: $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}$, $P(E) = \frac{1}{\sqrt{2\pi}} \int_E e^{-\frac{x^2}{2}} dx$.

Example 1.5. Poisson distribution with rate $\lambda > 0$: $\Omega = \mathbb{N}_0$, $\mathcal{F} = 2^\Omega$, $P(E) = e^{-\lambda} \sum_{k \in E} \frac{\lambda^k}{k!}$.

Why Measure Theory

Historically, probability was defined in terms of a finite number of equally likely outcomes (Example 1.1) so that $|\Omega| < \infty$, $\mathcal{F} = 2^\Omega$, and $P(E) = \frac{|E|}{|\Omega|}$.

When the sample space is countably infinite (Example 1.5), or finite but the outcomes are not necessarily equally likely (Example 1.2), one can speak of probabilities in terms of weighted outcomes by taking a function $p : \Omega \to [0, 1]$ with $\sum_{\omega\in\Omega} p(\omega) = 1$ and setting $P(E) = \sum_{\omega\in E} p(\omega)$.

For most practical purposes, this can be generalized to the case where $\Omega \subseteq \mathbb{R}$ by taking a weighting function $f : \Omega \to [0, \infty)$ with $\int_\Omega f(x)\,dx = 1$ and setting $P(E) = \int_E f(x)\,dx$ (Examples 1.3 and 1.4), but one must be careful since the integral is not defined for all sets E; see Example 1.7.

Those who have taken undergraduate probability will recognize p and f as pmfs and pdfs, respectively. In measure theoretic terms, $f = \frac{dP}{dm}$ is the *Radon-Nikodym derivative* of P with respect to Lebesgue measure. Similarly, $p = \frac{dP}{dc}$ where c is counting measure on Ω.

Measure theory provides a unifying framework in which these ideas can be made rigorous, and it enables further extensions to more general sample spaces and probability functions.

Also, note that in the formal axiomatic construction of probability, there is absolutely no mention of chance, propensity, credence, etc., so we can use the theory without worrying about any philosophical issues.

Random Variables and Expectation

Given a measurable space $(S, \mathcal{G})$, we define an $(S, \mathcal{G})$-*valued random variable* to be a measurable function $X : \Omega \to S$. This just means that for any $E \in \mathcal{G}$, $X^{-1}(E) \in \mathcal{F}$.

In this class, the unqualified term "random variable" will refer to the case $(S, \mathcal{G}) = (\mathbb{R}, \mathcal{B})$. (The collection of Borel sets, $\mathcal{B}$, is the smallest σ-field containing all of the open subsets of $\mathbb{R}$.)

We typically think of X as an observable, or a measurement to be taken after the experiment has been performed.

An extremely useful example is given by choosing any $A \in \mathcal{F}$ and defining the indicator function,

$$1_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \in A^C \end{cases}.$$

Note that if $(\Omega,\mathcal{F}, P)$ is a probability space and X is an $(S, \mathcal{G})$-valued random variable, then X induces the pushforward probability measure $\mu = P \circ X^{-1}$ on $(S, \mathcal{G})$. Frequently, we will abuse notation and write $P(X \in B) = P\big(X^{-1}(B)\big) = P(\{\omega \in \Omega : X(\omega) \in B\})$ for $\mu(B)$.

X also induces the sub-σ-field $\sigma(X) = \big\{X^{-1}(E) : E \in \mathcal{G}\big\} \subseteq \mathcal{F}$. If we think of Ω as the possible outcomes of an experiment and X as a measurement to be performed, then $\sigma(X)$ represents the insight that measurement will afford us. (If we learn that $X(\omega) \in E$, then we know $\omega \in X^{-1}(E)$.)

In contrast with other areas of measure theory, in probability we are often interested in various *sub-σ-fields* $\mathcal{F}_0 \subseteq \mathcal{F}$, which we think of in terms of information content.

For instance, if the experiment is rolling a six-sided die (Example 1.1), then $\mathcal{F}_0 = \big\{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\big\}$ represents the information concerning the parity of the value rolled. If we only had access to this data, it would not be meaningful to talk about the die landing on a number greater than 3 since $\{4, 5, 6\} \notin \mathcal{F}_0$.

In this case, one might choose instead to model the experiment as $\Omega_0 = \{\text{even}, \text{odd}\}$, $\mathcal{F}_0 = 2^\Omega$, $P_0(\{\text{even}\}) = P_0(\{\text{odd}\}) = \frac{1}{2}$. There are often many valid models for an experiment; probability theory just tells us how to proceed once we've settled on one that is deemed appropriate based on other empirical/practical/theoretical considerations.

Note that the probability space $(\Omega, \mathcal{F}, P)$ from Example 1.1 *extends* $(\Omega_0, \mathcal{F}_0, P_0)$ in the sense that there is a measurable surjection $\pi : (\Omega, \mathcal{F}) \to (\Omega_0, \mathcal{F}_0)$ with the property that $P\big(\pi^{-1}(E)\big) = P_0(E)$ for all $E \in \mathcal{F}_0$—namely $\pi(j) = \begin{cases} \text{even}, & 2 \mid j \\ \text{odd}, & 2 \nmid j \end{cases}$.

We often want to be able to add new details and sources of randomness on the fly, so one takes it as a general rule that probability should only study concepts and perform operations that are preserved by extensions of the underlying space. For instance, probabilities of events or set operations like unions, intersections, and complements are probabilistic concepts, but the equality (as sets) of two events is not, nor is their cardinality.

The *expectation* (or *mean* or *expected value*) of a real-valued random variable X on $(\Omega, \mathcal{F}, P)$ is defined as $E[X] = \int_\Omega X(\omega)\, dP(\omega)$ whenever the integral is well-defined.
Expectation is generally interpreted as a weighted average which gives the 'best guess' for the value of the random quantity X.

We will study random variables and their expectations in greater detail soon. For now, the point is that many familiar objects from undergraduate probability can be rigorously and simply defined using the language of measure theory.

That said, it should be emphasized that probability is not just the study of measure spaces with total mass one. As useful and necessary as the rigorous analytic foundations are, it is equally important to cultivate a probabilistic way of thinking whereby one conceptualizes problems in terms of coin tossing, card shuffling, particle trajectories, and so forth.

Example 1.6. Probabilities aggregate like masses and areas, so it is natural that the first probability measures one sees are discrete (characterized by mass functions) or absolutely continuous (characterized by density functions). There are also 'singular continuous measures,' but the need for a general theory is evident without considering anything so exotic.
For instance, consider the mixture of the Unif$(0,1)$ and Pois(1) measures where one chooses an element of $[0,1] \cup \mathbb{N}$ via the following procedure: Flip a fair coin. If it comes up heads, pick a real number uniformly from $[0,1]$. If it comes up tails, pick a nonnegative integer n with probability $e^{-1}/n!$.
The measure describing this experiment is supported on an uncountable set and assigns positive probability to some individual outcomes, so it is neither discrete nor continuous.

Example 1.7. When discussing the uniform distribution on $[0,1]$, we detailed how to find the probability that a point is chosen in some subinterval of $[0,1]$, but one can imagine asking about events that are not so clearly expressed in those terms, like "What is the probability that a rational number is chosen?" or "What is the probability that a number is chosen with no 1's in its base 3 representation?"

For the first of these, we recall that $P(\{x\}) = 0$ for all $x \in [0,1]$. Since $\mathbb{Q}$ is countable, we conclude that $P(\text{rational}) = \sum_{x \in \mathbb{Q}_{[0,1]}} P(\{x\}) = 0$.

For the second, the event in question is uncountable (since one can bijectively map oneless ternary representations to binary representations by halving the digits), so we don't get an easy answer from countable additivity.

However, we can observe that not having a 1 in the first ternary digit means that a point was not chosen from $(1/3, 2/3)$. Not having a one in the second digit means a point was not chosen from $(1/9, 2/9)$ or $(7/9, 8/9)$ either. In general, no 1 in the n^{th} place precludes the chosen point from lying in the open middle third of the 2^{n-1} intervals of length $1/3^n$ that have not already been ruled out. The set of forbidden points thus has probability $\sum_{n=1}^{\infty} \frac{2^{n-1}}{3^n} = \frac{1}{2} \cdot \frac{2/3}{1-2/3} = 1$, so the probability of choosing a point in the *Cantor set* is 0.

An example of a subset of $[0, 1]$ which has no well-defined probability under this measure is given by the following construction:

Define an equivalence relation on $[0, 1)$ by $x \sim y$ if $x - y \in \mathbb{Q}$.

Using the axiom of choice, let $E \subseteq [0, 1)$ consist of exactly one point from each equivalence class.

For $q \in \mathbb{Q}_{[0,1)}$, define $E_q = E + q \pmod 1$. By construction $E_q \bigcap E_r = \emptyset$ for $r \neq q$ and $\bigcup_{q \in \mathbb{Q}_{[0,1)}} E_q = [0, 1)$.

Thus, by countable additivity, we must have

$$1 = m\left([0, 1)\right) = m\left(\bigsqcup_{q \in \mathbb{Q}_{[0,1)}} E_q \right) = \sum_{q \in \mathbb{Q}_{[0,1)}} m(E_q).$$

However, Lebesgue measure is translation invariant, so $m(E_q) = m(E)$ for all q.

We see that $m(E)$ is not well-defined as $m(E) = 0$ implies $1 = 0$ and $m(E) > 0$ implies $1 = \infty$.

The existence of non-measurable sets can be proved using slightly weaker assumptions than the axiom of choice (such as the Boolean prime ideal theorem), but it has been shown that the existence of non-measurable sets is not provable in Zermelo-Fraenkel alone.

In three or more dimensions, the Banach-Tarski paradox shows that in ZFC, there is no finitely additive measure defined on all subsets of Euclidean space that is invariant under translation and rotation.

(The paradox is that one can cut a unit ball into five pieces and reassemble them using only rigid motions to obtain two disjoint unit balls.)

2 First Properties

We will delve into the technicalities of constructing probability spaces presently, but first let's explore some consequences of the definition to better understand the general framework.

Probability Measures

The following simple facts are extremely useful and will be employed frequently throughout this course.

Theorem 2.1. *Let P be a probability measure on $(\Omega, \mathcal{F})$.*

(i) Complements For any $A \in \mathcal{F}$, $P(A^C) = 1 - P(A)$.

(ii) Monotonicity For any $A, B \in \mathcal{F}$ with $A \subseteq B$, $P(A) \leq P(B)$.

(iii) Subadditivity For any countable collection $\{E_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$, $P\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} P(E_i)$.

(iv) Continuity from below If $A_i \nearrow A$ (i.e. $A_1 \subseteq A_2 \subseteq \ldots$ and $\bigcup_{i=1}^{\infty} A_i = A$), then $\lim_{n\to\infty} P(A_n) = P(A)$.

(v) Continuity from above If $A_i \searrow A = \bigcap_{i=1}^{\infty} A_i$, then $\lim_{n\to\infty} P(A_n) = P(A)$.

Proof.

For (i), $1 = P(\Omega) = P(A \sqcup A^C) = P(A) + P(A^C)$ by countable additivity.

For (ii), $P(B) = P\left(A \sqcup (B \setminus A)\right) = P(A) + P(B \setminus A) \geq P(A)$.

For (iii), we "disjointify" the sets by defining $F_1 = E_1$ and $F_i = E_i \setminus \left(\bigcup_{j=1}^{i-1} E_j\right)$ for $i > 1$, and observe that $\bigcup_{i=1}^{n} F_i = \bigcup_{i=1}^{n} E_i$ for all $n \in \mathbb{N} \cup \{\infty\}$. Since $F_i \subseteq E_i$ for all i, we have

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = P\left(\bigsqcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} P(F_i) \leq \sum_{i=1}^{\infty} P(E_i).$$

For (iv), set $B_1 = A_1$ and $B_i = A_i \setminus A_{i-1}$ for $i > 1$, and note that the B_i's are disjoint with $\bigcup_{i=1}^{n} B_i = A_n$ and $\bigcup_{i=1}^{\infty} B_i = A$. Then

$$\begin{aligned} P(A) = P\left(\bigsqcup_{i=1}^{\infty} B_i\right) &= \sum_{i=1}^{\infty} P(B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i) \\ &= \lim_{n\to\infty} P\left(\bigsqcup_{i=1}^{n} B_i\right) = \lim_{n\to\infty} P(A_n). \end{aligned}$$

For (v), if $A_1 \supseteq A_2 \supseteq \ldots$ and $A = \bigcap_{i=1}^{\infty} A_i$, then $A_1^C \subseteq A_2^C \subseteq \ldots$ and $A^C = \left(\bigcap_{i=1}^{\infty} A_i\right)^C = \bigcup_{i=1}^{\infty} A_i^C$, so it follows from (i) and (iv) that

$$P(A) = 1 - P(A^C) = 1 - \lim_{n\to\infty} P(A_n^C) = \lim_{n\to\infty} \left(1 - P(A_n^C)\right) = \lim_{n\to\infty} P(A_n). \qquad \square$$

We leave it as an easy exercise to show that one also has the union rule $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Sigma Algebras

We now review some some basic facts about σ-fields. Our first observation is immediate from the definition.

Proposition 2.2. *If $\{\mathcal{F}_i\}_{i\in I}$ is a collection of σ-algebras on Ω, then $\bigcap_{i\in I}\mathcal{F}_i$ is also a σ-algebra on Ω.*

It follows from Proposition 2.2 that for any collection of sets $\mathcal{C}\subseteq 2^\Omega$, there is a smallest σ-algebra containing $\mathcal{C}$—namely, the intersection of all σ-algebras containing $\mathcal{C}$. This is called the *σ-algebra generated by $\mathcal{C}$* and is denoted by $\sigma(\mathcal{C})$.

Note that if $\mathcal{F}$ is a σ-algebra and $\mathcal{C}\subseteq\mathcal{F}$, then $\sigma(\mathcal{C})\subseteq\mathcal{F}$.

An important class of examples are the Borel σ-algebras: If $(X,\mathcal{T})$ is a topological space, then $\mathcal{B}_X=\sigma(\mathcal{T})$ is called the *Borel σ-algebra*.

Recall that in $\mathbb{R}^n$ with the standard topology, $U\subseteq\mathbb{R}$ is open if for every $\mathbf{x}\in U$, there is an $\varepsilon=\varepsilon(\mathbf{x})>0$ such that the ball $B_\varepsilon(\mathbf{x})=\{\mathbf{y}\in\mathbb{R}^n:\|\mathbf{x}-\mathbf{y}\|<\varepsilon\}$ is contained in U.

Lemma 2.3. *Every open subset of $\mathbb{R}$ is a countable disjoint union of open intervals.*

Proof. Define an equivalence relation on the open set $U\subseteq\mathbb{R}$ by $x\sim y$ if $\big(\min\{x,y\},\max\{x,y\}\big)\subseteq U$.

The equivalence class containing $x\in U$ is thus the maximal open subinterval of U containing x.

(This relation clearly symmetric in x and y, and we always have $(\min\{x,x\},\max\{x,x\})=(x,x)=\emptyset\subseteq U$, so it's reflexive as well. To see that it's transitive, note that if $x,y,z\in U$ with $x\sim y$ and $y\sim z$, then: $x\le y\le z$ implies $(x,z)=(x,y)\cup\{y\}\cup(y,z)\subseteq U$; $x\le z\le y$ implies $(x,z)\subseteq(x,y)\subseteq U$; and similarly for $y\le x\le z$, $y\le z\le x$, $z\le y\le x$, $z\le x\le y$.)

Let $\mathcal{I}$ denote the set of equivalence classes under $\sim$. Then the elements of $\mathcal{I}$ are disjoint open subintervals of U and every $x\in U$ belongs to some $I\in\mathcal{I}$. Moreover, each $I\in\mathcal{I}$ contains a rational, so $\mathcal{I}$ is countable. It follows that $U=\bigsqcup_{I\in\mathcal{I}}I$ is an example of the asserted decomposition. □

Theorem 2.4. *The Borel σ-algebra for $\mathbb{R}$ with the standard topology is generated by each of the following:*

- **a.** *The finite open intervals $\mathcal{E}_1=\{(a,b):a,b\in\mathbb{R}\text{ with }a<b\}$*
- **b.** *The finite closed intervals $\mathcal{E}_2=\{[a,b]:a,b\in\mathbb{R}\text{ with }a<b\}$*
- **c.** *The finite half-open intervals $\mathcal{E}_3=\{(a,b]:a,b\in\mathbb{R}\text{ with }a<b\}$ or $\mathcal{E}_4=\{[a,b):a,b\in\mathbb{R}\text{ with }a<b\}$*
- **d.** *The open rays $\mathcal{E}_5=\{(a,\infty):a\in\mathbb{R}\}$ or $\mathcal{E}_6=\{(-\infty,b):b\in\mathbb{R}\}$*
- **e.** *The closed rays $\mathcal{E}_7=\{[a,\infty):a\in\mathbb{R}\}$ or $\mathcal{E}_8=\{(-\infty,b]:b\in\mathbb{R}\}$*

Proof. $\mathcal{B}$ is the smallest σ-algebra containing the open subsets of $\mathbb{R}$. Since open intervals are open sets, $\mathcal{B}$ contains the σ-algebra generated by the open intervals. Since every open set is a countable union of open intervals, the σ-algebra generated by the open intervals contains $\mathcal{B}$ as well, hence the two are equal.

As every open interval is finite or a countable union of finite open intervals—$(a,\infty)=\bigcup_{n=0}^\infty(a+n,a+n+2)$, for example—we conclude that $\sigma(\mathcal{E}_1)=\mathcal{B}$.

Likewise, the complement of a finite closed interval is the union of two open intervals, so $\sigma(\mathcal{E}_2)=\mathcal{B}$.

Similarly, $(a,b]=\bigcap_{n=1}^\infty(a,b+\frac{1}{n})$, so $\mathcal{E}_3\subseteq\sigma(\mathcal{E}_1)=\mathcal{B}$ and thus $\sigma(\mathcal{E}_3)\subseteq\mathcal{B}$; and $(a,b)=\bigcup_{n=1}^\infty(a,b-\frac{1}{n}]$, so $\mathcal{E}_1\subseteq\sigma(\mathcal{E}_3)$ and thus $\mathcal{B}=\sigma(\mathcal{E}_1)\subseteq\sigma(\mathcal{E}_3)$.

The other cases are similar. □

Remark 2.5. For any $S \subseteq [8]$, $\mathcal{B}$ contains every set in $\mathcal{E}_S = \bigcup_{i\in S} \mathcal{E}_i$ and for any $i \in S$, $\mathcal{B} = \sigma(\mathcal{E}_i) \subseteq \sigma(\mathcal{E}_S)$, so the Borels are generated by any union of the $\mathcal{E}_i$'s as well. Also, the density of $\mathbb{Q}$ in $\mathbb{R}$ enables us to take any of the above collections restricted to have rational endpoints if we so desire. For example, given $a < b$, there exist rational sequences $a_n \searrow a$ and $b_n \nearrow b$ so that $(a,b) = \bigcup_{n=1}^{\infty}(a_n, b_n)$. It follows that $\sigma\left(\{(a,b) : a, b \in \mathbb{Q}\}\right) = \sigma(\mathcal{E}_1) = \mathcal{B}$. This is nice since it allows one to work with countable generating sets.

Our main technical result about σ-algebras is Dynkin's π-λ Theorem.

Definition. A nonempty collection of sets $\mathcal{P} \subseteq 2^{\Omega}$ is called a *π-system* if $A, B \in \mathcal{P}$ implies $A \cap B \in \mathcal{P}$.

Definition. A collection of sets $\mathcal{L} \subseteq 2^{\Omega}$ is called a *λ-system* if

(1) $\Omega \in \mathcal{L}$
(2) If $A, B \in \mathcal{L}$ and $A \subseteq B$, then $B \setminus A \in \mathcal{L}$
(3) If $A_n \in \mathcal{L}$ with $A_n \nearrow A$, then $A \in \mathcal{L}$

Theorem 2.6. *If $\mathcal{P}$ is a π-system and $\mathcal{L}$ is a λ-system with $\mathcal{P} \subseteq \mathcal{L}$, then $\sigma(\mathcal{P}) \subseteq \mathcal{L}$.*

Proof. We begin by observing that the intersection of any number of λ-systems is a λ-system, so for any collection $\mathcal{C}$, there is a smallest λ-system $\ell(\mathcal{C})$ containing $\mathcal{C}$. Thus it will suffice to show that $\ell(\mathcal{P})$ is a σ-algebra since then $\sigma(\mathcal{P}) \subseteq \ell(\mathcal{P}) \subseteq \mathcal{L}$.

Moreover, λ-systems that are closed under intersections are σ-algebras—$E^C = \Omega \setminus E$, $E \cup F = (E^C \cap F^C)^C$, and $\bigcup_{k=1}^{n} E_k \nearrow \bigcup_{k=1}^{\infty} E_k$—so we need only demonstrate that $\ell(\mathcal{P})$ is a π-system.

To this end, let $A \in \ell(\mathcal{P})$ and define $\mathcal{L}_A = \{E : A \cap E \in \ell(\mathcal{P})\}$. Since $A \cap \Omega = A \in \ell(\mathcal{P})$, we have $\Omega \in \mathcal{L}_A$. Also, if $E, F \in \mathcal{L}_A$ with $E \subseteq F$, then $A \cap E \subseteq A \cap F$ are in $\ell(\mathcal{P})$, so $A \cap (F \setminus E) = (A \cap F) \setminus (A \cap E) \in \ell(\mathcal{P})$, showing that $\mathcal{L}_A$ is closed under subset differences as well. Finally, if $E_1 \subseteq E_2 \subseteq \cdots$ is a sequence of sets in $\mathcal{L}_A$ with $E = \bigcup_{n=1}^{\infty} E_n$, then $A \cap E_1 \subseteq A \cap E_2 \subseteq \cdots$ is a sequence in $\ell(\mathcal{P})$, hence $A \cap E = \bigcup_{n=1}^{\infty}(A \cap E_n) \in \ell(\mathcal{P})$. We have thus shown that $\mathcal{L}_A$ is a λ-system for every $A \in \ell(\mathcal{P})$.

Now $\mathcal{P}$ is a π-system, so if $C \in \mathcal{P}$, then $\mathcal{P} \subseteq \mathcal{L}_C$, hence $\ell(\mathcal{P}) \subseteq \mathcal{L}_C$. It follows that if $B \in \ell(\mathcal{P})$, then $B \in \mathcal{L}_C$, so $B \cap C \in \ell(\mathcal{P})$. As this is true for all $C \in \mathcal{P}$, we see that for every $B \in \ell(\mathcal{P})$, $\mathcal{P} \subseteq \mathcal{L}_B$, hence $\ell(\mathcal{P}) \subseteq \mathcal{L}_B$. This completes the proof since $A, B \in \ell(\mathcal{P})$ implies $A \in \mathcal{L}_B$ and thus $A \cap B \in \ell(\mathcal{P})$. □

It is not especially important to commit the details of the preceding argument to memory, but it worth seeing once and you should definitely know the statement of the theorem. Though it seems a bit obscure upon first encounter, its use in probability is ubiquitous.

In typical applications, we show that a property holds on a π-system that we know generates the σ-algebra of interest. We then show that the collection of all sets for which the property holds is a λ-system in order to conclude that it holds on the entire σ-algebra.

Example 2.7. Suppose P_1 and P_2 are probabilities on $(\mathbb{R}, \mathcal{B})$ with $P_1\big((-\infty, a]\big) = P_2\big((-\infty, a]\big)$ for all $a \in \mathbb{R}$. Since the right-closed rays form form a π-system that generates the Borel sets, we can conclude that P_1 and P_2 agree on $\mathcal{B}$ by showing that $\{A \subseteq \mathcal{B} : P_1(A) = P_2(A)\}$ is a λ-system.

To see that this is so, note that $P_1(\Omega) = 1 = P_2(\Omega)$; if $P_1(A) = P_2(A)$ and $P_1(B) = P_2(B)$ with $A \subseteq B$, then $P_1(B \setminus A) = P_1(B) - P_1(A) = P_2(B) - P_2(A) = P_2(B \setminus A)$; and if $A_n \nearrow A$ with $P_1(A_n) = P_2(A_n)$ for all n, then $P_1(A) = \lim_{n\to\infty} P_1(A_n) = \lim_{n\to\infty} P_2(A_n) = P_2(A)$.

3 Constructing Probability Spaces

Now that we have some familiarity working with probability spaces, we turn our attention to the problem of constructing them.

We begin by dispensing with the easy case: If Ω is any countable set, its outcomes can be enumerated as $\Omega = \{\omega_1, \omega_2, \ldots\}$. Given any sequence of nonnegative numbers $\{p_k\}_{k=1}^{\infty}$ with $\sum_{k=1}^{\infty} p_k = 1$, we can define a probability measure P on $(\Omega, 2^{\Omega})$ by $P(E) = \sum_{k:\omega_k \in E} p_k$.
Conversely, given any probability Q on $(\Omega, 2^{\Omega})$, the sequence $\{q_k\}_{k=1}^{\infty}$ defined by $q_k = Q\{\omega_k\}$ is nonnegative and sums to 1, so this fully characterizes such discrete probability measures.

Note that if Ω is countably infinite, it cannot support a uniform probability since $P\{\omega\} = c$ for all $\omega \in \Omega$ implies $1 = P(\Omega) = \sum_{\omega\in\Omega} P\{\omega\} = \sum_{k=1}^{\infty} c$. If $c = 0$, this gives $1 = 0$, and if $c > 0$, it gives $1 = \infty$.

To treat uncountable sample spaces, we take inspiration from the case of the uniform distribution on $[0, 1]$. There, we said that the probability is characterized by $P\big([a, b]\big) = b - a$ for all $0 \le a \le b \le 1$.
Observe that this implies $P\{a\} = P([a, a]) = 0$, so $(a, b]$, $[a, b)$, and (a, b) also are assigned probability $b - a$. (For instance, $b - a = P\big([a, b]\big) = P\big((a, b] \sqcup \{a\}\big) = P\big((a, b]\big) + P\{a\} = P\big((a, b]\big)$.)

As such, we at least know how to define P on $\mathcal{J} = \{J \subseteq [0, 1] : J \text{ is an interval}\}$. As $\mathcal{J}$ is not closed under unions, it is not a σ-algebra, but it is a start.

Definition. A collection of subsets $\mathcal{S}$ of Ω is called a *semialgebra* if

(1) $\emptyset, \Omega \in \mathcal{S}$
(2) $S_1, S_2 \in \mathcal{S}$ implies $S_1 \cap S_2 \in \mathcal{S}$
(3) If $S \in \mathcal{S}$, there exist disjoint $T_1, \ldots, T_n \in \mathcal{S}$ with $S^C = \bigsqcup_{k=1}^{n} T_k$

Note that parts (1) and (2) imply that semialgebras are π-systems.

Example 3.1. For any nonempty interval $I \subseteq \mathbb{R}$, $\mathcal{J}_I = \{J \subseteq I : J \text{ is an interval}\}$ is a semialgebra. Indeed, $\emptyset = (a, a]$ and $\Omega = I$ belong to $\mathcal{J}_I$, the intersection of two intervals is an interval, and the complement of an interval is an interval or disjoint union of two intervals.

Example 3.2. On $\mathbb{R}$, we can define the collection of *h-intervals* by $\mathcal{H} = \big\{(a, b] : -\infty \le a \le b \le \infty\big\}$ with the understanding $(a, \infty] = (a, \infty)$. Arguing as in the previous example, one easily checks that $\mathcal{H}$ is a semialgebra.

Example 3.3. If $\mathcal{S}_1$ and $\mathcal{S}_2$ are semialgebras over Ω_1 and Ω_2, then $\mathcal{R} = \big\{A_1 \times A_2 : A_k \in \mathcal{S}_k\big\}$ is a semialgebra over $\Omega_1 \times \Omega_2$: $\mathcal{R}$ certainly contains $\emptyset$ and $\Omega_1 \times \Omega_2$; $A_1 \times A_2, B_1 \times B_2 \in \mathcal{R}$ implies $(A_1 \times A_2) \cap (B_1 \times B_2) = (A_1 \cap B_1) \times (A_2 \cap B_2) \in \mathcal{R}$; and if $A \times B \in \mathcal{R}$, there exist $\{S_k\}_{k=1}^{m} \subseteq \mathcal{S}_1$, $\{T_\ell\}_{\ell=1}^{n} \subseteq \mathcal{S}_2$ with $A^C = \bigsqcup_{k=1}^{m} S_k$ and $B^C = \bigsqcup_{\ell=1}^{n} T_\ell$, hence

$$(A\times B)^C = (A^C \times B) \bigsqcup (A \times B^C) \bigsqcup (A^C \times B^C) = \left(\bigsqcup_{k=1}^{m} (S_k \times B)\right) \bigsqcup \left(\bigsqcup_{\ell=1}^{n} (A \times T_\ell)\right) \bigsqcup \left(\bigsqcup_{k=1}^{m} \bigsqcup_{\ell=1}^{n} (S_k \times T_\ell)\right).$$

Iterating this procedure shows that if $\mathcal{S}_k$ is a semialgebra over Ω_k for $k = 1, \ldots, n$, then the collection of *rectangles* $\big\{A_1 \times \cdots \times A_n : A_k \in \mathcal{S}_k\big\}$ is a semialgebra over $\Omega_1 \times \cdots \times \Omega_n$.
In particular, $\big\{J_1 \times \cdots \times J_d : \text{each } J_k \text{ is an interval}\big\}$ (or an h-interval) is a semialgebra over $\mathbb{R}^d$.

Example 3.4. If $\mathcal{S}$ is a semialgebra over Ω and $A \subseteq \Omega$, then $\mathcal{S}_A = \{S \cap A : S \in \mathcal{S}\}$ is a semialgebra over A since $A = \Omega \cap A$ and $\emptyset = \emptyset \cap A$ belong to $\mathcal{S}_A$; $A_1, A_2 \in \mathcal{S}_A$ implies that there exist $S_1, S_2 \in \mathcal{S}$ with $A_i = S_i \cap A$ so that $A_1 \cap A_2 = (S_1 \cap A) \cap (S_2 \cap A) = (S_1 \cap S_2) \cap A \in \mathcal{S}_A$; and if $B = S \cap A \in \mathcal{S}_A$, then $B^C = A \setminus B = A \setminus S = A \cap S^C = A \cap (\bigsqcup_{k=1}^n T_k) = \bigsqcup_{k=1}^n (T_k \cap A) \in \mathcal{S}_A$. (Note that $\mathcal{S}_A \subseteq \mathcal{S}$ iff $A \in \mathcal{S}$.)

Definition. A (probability) *protomeasure* on a semialgebra $\mathcal{S}$ over Ω is a function $P_0 : \mathcal{S} \to [0,1]$ with

(1) $P_0(\emptyset) = 0$ and $P_0(\Omega) = 1$
(2) If $S_1, \ldots, S_n \in \mathcal{S}$ are disjoint with $S = \bigsqcup_{i=1}^n S_i \in \mathcal{S}$, then $P_0(S) = \sum_{i=1}^n P_0(S_i)$
(3) If $S_1, S_2 \ldots \in \mathcal{S}$ with $S = \bigcup_{i=1}^\infty S_i \in \mathcal{S}$, then $P_0(S) \leq \sum_{i=1}^\infty P(S_i)$

(The countable subadditivity condition lets us replace finite additivity with finite superadditivity if desired; nonnegativity and finite additivity make $P_0(\emptyset) = 0$ redundant.)

Example 3.5. Our length function $\lambda_0(J_{a,b}) = b - a$ for any interval $J_{a,b}$ with endpoints $a \leq b$ is a protomeasure on $\mathcal{J}$.

Indeed, for any $0 \leq a \leq b \leq 1$, $\lambda_0(J_{a,b}) = b - a \in [0,1]$, $\lambda_0(\emptyset) = \lambda_0\big((a,a)\big) = 0$, and $\lambda_0([0,1]) = 1$.

Now suppose that $J_{a_1,b_1}, \ldots, J_{a_n,b_n}$ are disjoint intervals with $J = \bigcup_{k=1}^n J_{a_k,b_k}$ also an interval. By reindexing if need be, we can assume that $a_1 \leq \cdots \leq a_n$. Since the subintervals are disjoint, we must have that $a_{k+1} \geq b_k$, and since J is an interval as well, we must have that $a_{k+1} \leq b_k$. It follows that

$$\lambda_0(J) = b_n - a_1 = \sum_{k=1}^{n-1}(a_{k+1} - a_k) + (b_n - a_n) = \sum_{k=1}^{n-1}(b_k - a_j) + (b_n - a_n) = \sum_{k=1}^{n} \lambda_0(J_{a_k,b_k}),$$

hence λ_0 is finitely additive.

Now suppose that $J_{a_1,b_1}, J_{a_2,b_2}, \ldots$ is a sequence of intervals with $\bigcup_{k=1}^\infty J_{a_k,b_k} = J_{a,b}$. Fix $\varepsilon > 0$ and define $J_k = (\alpha_k, \beta_k)$ with $\alpha_k = a_k - \frac{\varepsilon}{2^{k+1}}$ and $\beta_k = b_k + \frac{\varepsilon}{2^{k+1}}$ so that $\lambda_0(J_k) = \lambda_0(J_{a_k,b_k}) + \frac{\varepsilon}{2^k}$ and $\bigcup_{k=1}^\infty J_k \supseteq [a + \frac{\varepsilon}{2}, b - \frac{\varepsilon}{2}]$. Compactness dictates that there exist $\{k(1), \ldots, k(n)\} \subseteq \mathbb{N}$ such that $[a + \frac{\varepsilon}{2}, b - \frac{\varepsilon}{2}] \subseteq \bigcup_{i=1}^n J_{k(i)}$, $\alpha_{k(1)} < \cdots < \alpha_{k(n)}$, and $\alpha_{k(i+1)} < \beta_{k(i)} < \beta_{k(i+1)}$ for $i < n$—otherwise, $J_{k(i)}$ or $J_{k(i+1)}$ could be omitted—so

$$\begin{aligned}
\sum_{k=1}^\infty \lambda_0(J_{a_k,b_k}) &\geq \sum_{i=1}^n \lambda_0(J_{a_{k(i)},b_{k(i)}}) = \sum_{i=1}^n [\lambda_0(J_{k(i)}) - \tfrac{\varepsilon}{2^{k(i)}}] \\
&\geq \sum_{i=1}^n (\beta_{k(i)} - \alpha_{k(i)}) - \sum_{k=1}^\infty \tfrac{\varepsilon}{2^k} \geq \beta_{k(n)} - \alpha_{k(n)} + \sum_{i=1}^{n-1} (\alpha_{k(i+1)} - \alpha_{k(i)}) - \varepsilon \\
&= \beta_{k(n)} - \alpha_{k(1)} - \varepsilon \geq (b - \tfrac{\varepsilon}{2}) - (a + \tfrac{\varepsilon}{2}) - \varepsilon = \lambda_0(J_{a,b}) - 2\varepsilon.
\end{aligned}$$

As ε was arbitrary, we see that λ_0 is countably subadditive as well.

Definition. A function $F : \mathbb{R} \to \mathbb{R}$ which is nondecreasing (so $x < y$ implies $F(x) \leq F(y)$) and right-continuous (so $F(x) = \lim_{y \to x^+} F(y)$) is called a *distribution function*.

Monotonicity ensures that $\alpha = \lim_{x \to -\infty} F(x)$ and $\beta = \lim_{x \to \infty} F(x)$ exist in $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. If $\alpha = 0$ and $\beta = 1$, we say that F is a *probability distribution function*.

Example 3.6. If F is a probability distribution function, then $\mu_0\big((a,b]\big) = F(b) - F(a)$ is a probability protomeasure on the semialgebra $\mathcal{H}$ of h-intervals. The argument here is pretty much the same as in the previous example—which essentially corresponds to the case $F(x) = x1_{[0,1]}(x) + 1_{(1,\infty)}(x)$—but we need to be a little careful with countable subadditivity.

To this end, suppose that $\bigcup_{k=1}^{\infty}(a_k, b_k] = (a, b]$, where we assume that $a, b \in \mathbb{R}$ to begin with. Given $\varepsilon > 0$, right-continuity ensures there is a $\delta > 0$ with $F(a+\delta) - F(a) < \varepsilon$. Likewise, there exist $\delta_j > 0$ such that $F(b_j + \delta_j) - F(b_j) < \varepsilon/2^j$. The open intervals $(a_j, b_j + \delta_j)$ cover the compact set $[a+\delta, b]$, so there is a finite subcover. By discarding any $(a_j, b_j + \delta_j)$ that is contained in a larger one and reindexing, we may assume $[a+\delta, b] \subseteq \bigcup_{j=1}^{n}(a_j, b_j + \delta_j)$, $a_1 < \cdots < a_n$, and $b_j + \delta_j \in (a_{j+1}, b_{j+1} + \delta_{j+1})$ for $j < n$. It follows that

$$\begin{aligned}
\mu_0\big((a,b]\big) &\leq F(b) - F(a+\delta) + \varepsilon \leq F(b_n + \delta_n) - F(a_1) + \varepsilon \\
&= F(b_n + \delta_n) - F(a_n) + \sum_{j=1}^{n-1}\big[F(a_{j+1}) - F(a_j)\big] + \varepsilon \\
&\leq F(b_n + \delta_n) - F(a_n) + \sum_{j=1}^{n-1}\big[F(b_j + \delta_j) - F(a_j)\big] + \varepsilon \\
&\leq \sum_{j=1}^{n}\big[F(b_j) + \tfrac{\varepsilon}{2^j} - F(a_j)\big] + \varepsilon \leq \sum_{j=1}^{n}\mu_0\big((a_j, b_j]\big) + 2\varepsilon \leq \sum_{j=1}^{\infty}\mu_0\big((a_j, b_j]\big) + 2\varepsilon.
\end{aligned}$$

To treat the case where $(a, b]$ is infinite, observe that the boundary conditions on F ensure there is an $M > 0$ such that $F(-M) < \varepsilon$ and $F(M) > 1 - \varepsilon$. As the preceding argument shows that $I = (\max\{a, -M\}, \min\{b, M\})$ satisfies $\mu_0(I) \leq \sum_{j=1}^{\infty}\mu_0\big((a_j, b_j]\big) + 2\varepsilon$, and we know that $\mu_0\big((a, b]\big) \leq \mu_0(I) + 2\varepsilon$, we conclude that μ_0 is indeed countably subadditive.

Now that we have a nice protomeasure λ_0 on $\mathcal{J}$, a natural first guess would be to try to extend it to $\mathcal{B}_1 := \{\text{countable unions of sets in } \mathcal{J}\}$.

Unfortunately, this is not a σ-algebra. For instance, consider the the Cantor set K from Example 1.7. We know that K^C can be written as a countable union of the open middle-thirds sets $\left(\frac{1}{3}, \frac{2}{3}\right), \left(\frac{1}{9}, \frac{2}{9}\right), \left(\frac{7}{9}, \frac{8}{9}\right), \ldots$ and thus belongs to $\mathcal{B}_1$. However, $(K^C)^C = K \notin \mathcal{B}_1$. Indeed, suppose that $K = \bigcup_{i \in I} J_i$ with I countable and each $J_i \in \mathcal{J}$. If any J_i contained two points $a < b$, then we would have $(a, b) \subseteq J_i$. But this is impossible because choosing n so that $\frac{1}{3^n} < b - a$ shows that there is an element of (a, b) with a 1 in its n^{th} ternary digit. However, the J_i cannot all be singletons either as this would imply that K was countable.

Instead, we remain patient and first consider the set $\mathcal{B}_0 := \big\{\text{finite unions of sets in } \mathcal{J}\big\}$. The following proposition will make it easier to extend our protomeasure to $\mathcal{B}_0$.

Proposition 3.7. *If $\mathcal{S}$ is a semialgebra, $\big\{$finite unions of sets in $\mathcal{S}\big\} = \big\{$finite disjoint unions of sets in $\mathcal{S}\big\}$.*

Proof. If $S_1, \ldots, S_m \in \mathcal{S}$, we can define

$$R_i = S_i \setminus \bigcup_{j=1}^{i-1} S_j = \bigcap_{j=1}^{i-1}(S_i \cap S_j^C) = \bigcap_{j=1}^{i-1}\left(S_i \cap \bigsqcup_{k=1}^{n_j} T_{j,k}\right) = \bigcap_{j=1}^{i-1}\left[\bigsqcup_{k=1}^{n_j}(S_i \cap T_{j,k})\right]$$

with each $S_i \cap T_{j,k} \in \mathcal{S}$. As this in turn can be written as a finite disjoint union of finite intersections of sets in $\mathcal{S}$, we see that $\bigcup_{i=1}^{m} S_i = \bigsqcup_{i=1}^{m} R_i$ can be expressed as a finite disjoint union of sets in $\mathcal{S}$. The other inclusion is immediate. □

Definition. A collection of subsets of Ω is called an *algebra* if it contains $\emptyset$ and is closed under complements and finite unions.

Algebras are also closed under finite intersections since $A_1, \ldots, A_n \in \mathcal{A}$ implies $\bigcap_{i=1}^{n} A_i = \left(\bigcup_{i=1}^{n} A_i^C\right)^C \in \mathcal{A}$.

Proposition 3.8. *If $\mathcal{S}$ is a semialgebra over Ω, then $\overline{\mathcal{S}} = \{$finite disjoint unions of sets in $\mathcal{S}\}$ is an algebra.*

Proof. In view of Proposition 3.7, we can equivalently define $\overline{\mathcal{S}} = \{$finite unions of sets in $\mathcal{S}\}$. With this characterization, we readily check that $\overline{\mathcal{S}}$ contains $\emptyset$, and if $A_i = \bigcup_{j=1}^{n_i} S_{i,j}$ with each $S_{i,j} \in \mathcal{S}$ for $i = 1, \dots, n$, then $\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} \bigcup_{j=1}^{n_i} S_{i,j}$ is a finite union of sets in $\mathcal{S}$ and thus belongs to $\overline{\mathcal{S}}$. Finally, suppose that $A = \bigcup_{k=1}^{m} S_k$ with $\{S_k\}_{k=1}^{m} \subseteq \mathcal{S}$. Then $A^C = \bigcap_{k=1}^{m} S_i^C = \bigcap_{k=1}^{m} \bigsqcup_{\ell=1}^{n_i} T_{k,\ell}$ with each $T_{k,\ell} \in \mathcal{S}$. By the distributivity property of intersections over unions, A^C is a finite union of finite intersections of sets in $\mathcal{S}$ and thus belongs to $\overline{\mathcal{S}}$. □

Example 3.9. Suppose Γ is an infinite set. Then $\mathcal{A} = \{A \subseteq \Gamma : A \text{ or } A^C \text{ is finite}\}$ is an algebra. To check that this is the case, we first observe that $\emptyset$ is finite and so belongs to $\mathcal{A}$. Next, $A \in \mathcal{A}$ implies $A = (A^C)^C$ or A^C is finite; in either case $A^C \in \mathcal{A}$. Finally if $A, B \in \mathcal{A}$, either A and B are both finite, so $A \cup B$ is finite, or at least one of A, B has a finite complement, so $(A \cup B)^C = A^C \cap B^C \subseteq A^C, B^C$ is finite.

$\mathcal{A}$ is not necessarily a σ-algebra, though. For instance, if $\Gamma = \mathbb{N}$ and $A_k = \{2k\} \in \mathcal{A}$, then $\bigcup_{k=1}^{\infty} A_k = 2\mathbb{N} \notin \mathcal{A}$.

Definition. A (probability) *premeasure* on an algebra $\mathcal{A}$ is a function $\widetilde{P} : \mathcal{A} \to [0, 1]$ satisfying $\widetilde{P}(\emptyset) = 0$, $\widetilde{P}(\Omega) = 1$, and for any countable disjoint collection $\{A_i\}_{i \in I} \subseteq \mathcal{A}$ such that $\bigsqcup_{i \in I} A_i \in \mathcal{A}$, $\widetilde{P}\left(\bigsqcup_{i \in I} A_i\right) = \sum_{i \in I} \widetilde{P}(A_i)$.

Note that since $\mathcal{A}$ is only an algebra, the union of an infinite collection of sets in $\mathcal{A}$ need not belong to $\mathcal{A}$; we just want to ensure that our premeasure is countably additive in those cases where it does.

Also, algebras are closed under complements and finite unions/intersections so nonnegativity and the countable additivity condition implies that premeasures are monotone and countably subadditive (when applicable) by the arguments in Theorem 2.1.

Proposition 3.10. *Suppose P_0 is a protomeasure on the semialgebra $\mathcal{S}$ and $\overline{\mathcal{S}}$ is the algebra generated by $\mathcal{S}$. Define the set function $\widetilde{P}$ on $\mathcal{A}$ by $\widetilde{P}\left(\bigsqcup_{i=1}^{n} S_i\right) = \sum_{i=1}^{n} P_0(S_i)$. Then $\widetilde{P}$ is a premeasure on $\overline{\mathcal{S}}$ with $\widetilde{P}|_{\mathcal{S}} = P_0$.*

Proof. To see that $\widetilde{P}$ is well-defined, suppose that $A = \bigsqcup_{i=1}^{m} S_i \in \mathcal{A}$ can also be written as $A = \bigsqcup_{j=1}^{n} R_j$, Then $S_i = S_i \cap A = S_i \cap \left(\bigsqcup_{j=1}^{n} R_j\right) = \bigsqcup_{j=1}^{n} (S_i \cap R_j)$ with $S_i \cap R_j \in \mathcal{S}$. Similarly, $R_j = \bigsqcup_{i=1}^{m} (S_i \cap R_j)$, so finite additivity of P_0 gives

$$\begin{aligned}\sum_{i=1}^{m} P_0(S_i) &= \sum_{i=1}^{m} P_0\left(\bigsqcup_{j=1}^{n} (S_i \cap R_j)\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} P_0(S_i \cap R_j) \\ &= \sum_{j=1}^{n} \sum_{i=1}^{m} P_0(S_i \cap R_j) = \sum_{j=1}^{n} P_0\left(\bigsqcup_{i=1}^{m} (S_i \cap R_j)\right) = \sum_{j=1}^{n} P_0(R_j).\end{aligned}$$

Also, $\widetilde{P}(S) = P_0(S)$ for all $S \in \mathcal{S}$ by definition, and monotonicity then implies $0 = \widetilde{P}(\emptyset) \leq \widetilde{P}(A) \leq \widetilde{P}(\Omega) = 1$ for all $A \in \overline{\mathcal{S}}$.

Next, if $A_1, \dots, A_m \in \overline{\mathcal{S}}$ are disjoint, then finite additivity of P_0 shows that, in obvious notation,

$$\widetilde{P}\left(\bigsqcup_{i=1}^{m} A_i\right) = \widetilde{P}\left(\bigsqcup_{i=1}^{m} \bigsqcup_{j=1}^{n_i} S_{i,j}\right) = \sum_{i,j} P_0(S_{i,j}) = \sum_{i=1}^{m} \widetilde{P}(A_i).$$

Finally, suppose that $\{A_i\}_{i\in\mathbb{N}}$ is a countable collection of disjoint sets in $\overline{\mathcal{S}}$ with $A = \bigsqcup_{i\in\mathbb{N}} A_i \in \overline{\mathcal{S}}$. Since $A_i = \bigsqcup_{j=1}^{n_i} S_{i,j}$ with $S_{i,j} \in \mathcal{S}$, hence $\sum_{i\in\mathbb{N}} \widetilde{P}(A_i) = \sum_{i\in\mathbb{N}} \sum_{j=1}^{n_i} P_0(S_{i,j})$, we may assume without loss of generality that each $A_i \in \mathcal{S}$.

As $A = \bigsqcup_{k=1}^{n} T_k$ with $T_k \in \mathcal{S}$, countable subadditivity of P_0 gives $P_0(T_k) = P_0\big(\bigsqcup_{i\in\mathbb{N}}(A_i \cap T_k)\big) \leq \sum_{i\in\mathbb{N}} P_0(A_i \cap T_k)$ and finite additivity of P_0 gives $P_0(A_i) = P_0\left(\bigsqcup_{k=1}^{n}(A_i \cap T_k)\right) = \sum_{k=1}^{n} P_0(A_i \cap T_k)$, thus

$$\widetilde{P}(A) = \sum_{k=1}^{n} P_0(T_k) \leq \sum_{k=1}^{n}\sum_{i=1}^{\infty} P_0(A_i \cap T_k) = \sum_{i=1}^{\infty}\sum_{k=1}^{n} P_0(A_i \cap T_k) = \sum_{i=1}^{\infty} P_0(A_i) = \sum_{i=1}^{\infty} \widetilde{P}(A_i)$$

where the interchange of summation is justified by the nonnegativity of P_0.

For the reverse inequality, let $B_n = \bigsqcup_{i=1}^{n} A_i$ and $C_n = A \cap B_n^C$. Then $A_1, \ldots, A_n, C_n \in \overline{\mathcal{S}}$ are disjoint, so finite additivity of $\widetilde{P}$ implies $\widetilde{P}(A) = \widetilde{P}(A_1) + \cdots + \widetilde{P}(A_n) + \widetilde{P}(C_n) \geq \widetilde{P}(A_1) + \cdots + \widetilde{P}(A_n)$. Letting $n \to \infty$ yields $\widetilde{P}(A) \geq \sum_{i=1}^{\infty} \widetilde{P}(A_i)$. □

The next step is to extend our premeasure to a set function defined on all of 2^Ω. We will generally lose countable additivity in the process, but we'll cross that bridge when we come to it.

Definition. A (probability) *outer measure* on Ω is a function $P^* : 2^\Omega \to [0,1]$ satisfying $P^*(\emptyset) = 0$ and $P^*(\Omega) = 1$; $P^*(A) \leq P^*(B)$ whenever $A \subseteq B$; and for any countable collection $\{A_i\}_{i\in I}$ of sets in Ω, $P^*\left(\bigcup_{i\in I} A_i\right) \leq \sum_{i\in I} P^*(A_i)$.

Proposition 3.11. *Suppose that $\widetilde{P}$ is a premeasure on an algebra $\mathcal{A}$ and define for each $E \subseteq \Omega$*

$$P^*(E) = \inf\left\{\sum_{i=1}^{\infty} \widetilde{P}(A_i) : A_i \in \mathcal{A} \text{ and } E \subseteq \bigcup_{i=1}^{\infty} A_i\right\}.$$

Then P^* is an outer measure on Ω.

Proof. Note that we can take $A_i = \emptyset$ for $i > n$, so the infimum includes sums over finite covers as well.

Also, $\widetilde{P}$ is nonnegative and every set in 2^Ω is covered by Ω, so P^* is well-defined with $0 \leq P^*(E) \leq \widetilde{P}(\Omega) = 1$ for all $E \subseteq \Omega$.

Now $\emptyset \subseteq \emptyset$, and countable subadditivity of $\widetilde{P}$ ensures that if $\Omega \subseteq \bigcup_{i=1}^{\infty} A_i$, then $\widetilde{P}(\Omega) \leq \sum_{i=1}^{\infty} \widetilde{P}(A_i)$—the union is necessarily Ω in this case and thus belongs to $\mathcal{A}$. It follows that $P^*(\emptyset) \leq \widetilde{P}(\emptyset) = 0$ and $P^*(\Omega) \geq \widetilde{P}(\Omega) = 1$ as well, hence P^* behaves appropriately on $\emptyset, \Omega$.

Next, if $A \subseteq B$, then any cover of B also covers A, so $P^*(A) \leq P^*(B)$.

Finally, let $\varepsilon > 0$ and consider any countable collection $\{A_i\}_{i=1}^{\infty}$. For each $i \in \mathbb{N}$, there's a countable cover $\{B_{i,j}\}_{j=1}^{\infty}$ of A_i with $\sum_{j=1}^{\infty} \widetilde{P}(B_{i,j}) \leq P^*(A_i) + \varepsilon/2^i$. It follows that $\bigcup_{i=1}^{\infty} A_i \subseteq \bigcup_{i=1}^{\infty}\bigcup_{j=1}^{\infty} B_{i,j}$, hence

$$P^*\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \widetilde{P}(B_{i,j}) \leq \sum_{i=1}^{\infty}\left[P^*(A_i) + \varepsilon/2^i\right] = \sum_{i=1}^{\infty} P^*(A_i) + \varepsilon.$$ □

Definition. If P^* is an outer measure on Ω, we say that $A \subseteq \Omega$ is *P^*-measurable* if for each $E \subseteq \Omega$,

$$P^*(E) = P^*(E \cap A) + P^*(E \cap A^C).$$

Note that P^* is countably subadditive, so one only needs to check that $P^*(E) \geq P^*(E \cap A) + P^*(E \cap A^C)$.

While this notion of 'splitting events nicely' is not super obvious or intuitive, the following *Carathéodory extension theorem* shows that it is extremely useful.

Theorem 3.12. *If P^* is an outer measure on Ω, then the collection $\mathcal{M}$ of P^*-measurable sets is a σ-algebra and the restriction of P^* to $\mathcal{M}$ is a measure.*

Proof. $\emptyset \in \mathcal{M}$ since $P^*(E) = 0 + P^*(E \cap \Omega) = P^*(E \cap \emptyset) + P^*(E \cap \emptyset^C)$, and $\mathcal{M}$ is closed under complements since the definition of P^*-measurability is symmetric in A and A^C.

Next, if $A, B \in \mathcal{M}$ and $E \subseteq \Omega$, subadditivity and $A \cup B = (A \cap B) \cup (A \cap B^C) \cup (A^C \cap B)$ gives

$$\begin{aligned} P^*(E) &= P^*(E \cap A) + P^*(E \cap A^C) \\ &= \big[P^*\big((E \cap A) \cap B\big) + P^*\big((E \cap A) \cap B^C\big) + P^*\big((E \cap A^C) \cap B\big)\big] + P^*\big((E \cap A^C) \cap B^C\big) \\ &\geq P^*\big(E \cap (A \cup B)\big) + P^*\big(E \cap (A \cup B)^C\big), \end{aligned}$$

thus $A \cup B \in \mathcal{M}$. This shows that $\mathcal{M}$ is an algebra, so it remains only to establish closure under countable unions, which we may assume to be disjoint.

Given any disjoint sequence $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{M}$, define $B = \bigsqcup_{i=1}^{\infty} A_i$ and $B_n = \bigsqcup_{i=1}^{n} A_i$ for $n \in \mathbb{N}$. Then

$$P^*(E \cap B_n) = P^*(E \cap B_n \cap A_n) + P^*(E \cap B_n \cap A_n^C) = P^*(E \cap A_n) + P^*(E \cap B_{n-1}).$$

As $P^*(E \cap B_1) = P^*(E \cap A_1)$ and $P^*(E \cap B_n) = \sum_{i=1}^{n} P^*(E \cap A_i)$ implies

$$P^*(E \cap B_{n+1}) = P^*(E \cap A_{n+1}) + P^*(E \cap B_n) = \sum_{i=1}^{n+1} P^*(E \cap A_i),$$

it follows from the principle of induction that $P^*(E \cap B_n) = \sum_{i=1}^{n} P^*(E \cap A_i)$ for all n.

Since $B_n \in \mathcal{M}$ and $B_n \subseteq B$, we have

$$P^*(E) = P^*(E \cap B_n) + P^*(E \cap B_n^C) \geq \sum_{i=1}^{n} P^*(E \cap A_i) + P^*(E \cap B^C),$$

so letting $n \to \infty$ yields

$$P^*(E) \geq \sum_{i=1}^{\infty} P^*(E \cap A_i) + P^*(E \cap B^C) \geq P^*\left(\bigcup_{i=1}^{\infty}(E \cap A_i)\right) + P^*(E \cap B^C) = P^*(E \cap B) + P^*(E \cap B^C).$$

This shows that $B \in \mathcal{M}$,and taking $E = B$ in the preceding gives

$$P^*(B) = \sum_{i=1}^{\infty} P^*(B \cap A_i) + P^*(B \cap B^C) = \sum_{i=1}^{\infty} P^*(A_i),$$

so P^* is countably additive on $\mathcal{M}$ and thus defines a measure. □

Proposition 3.13. *If $\widetilde{P}$ is a premeasure on $\mathcal{A}$ and P^* is the induced outer measure from Proposition 3.11, then $P^*(E) = \widetilde{P}(E)$ for all $E \in \mathcal{A}$, and every set in $\mathcal{A}$ is P^*-measurable.*

Proof. If $E \in \mathcal{A}$ and $E \subseteq \bigcup_{i=1}^{\infty} A_i$ with each $A_i \in \mathcal{A}$, we can define $B_n = E \bigcap \left(A_n \setminus \bigcup_{i=1}^{n-1} A_i\right)$. Then the B_n's are disjoint elements of $\mathcal{A}$ whose union is E, so, since $B_n \subseteq A_n$, $\widetilde{P}(E) = \sum_{n=1}^{\infty} \widetilde{P}(B_n) \leq \sum_{n=1}^{\infty} \widetilde{P}(A_n)$. This shows that $\widetilde{P}(E) \leq P^*(E)$, and the reverse inequality is immediate.

To see that $\mathcal{A} \subseteq \mathcal{M}$, let $A \in \mathcal{A}$, $E \subseteq \Omega$, and $\varepsilon > 0$. Then there is a sequence $\{B_n\}_{n=1}^{\infty}$ in $\mathcal{A}$ with $E \subseteq \bigcup_{n=1}^{\infty} B_n$ and $\sum_{n=1}^{\infty} \widetilde{P}(B_n) \leq P^*(E) + \varepsilon$.

Since $\widetilde{P}(B_n) = \widetilde{P}(B_n \cap A) + \widetilde{P}(B_n \cap A^C)$ by finite additivity of $\widetilde{P}$, we have

$$P^*(E) + \varepsilon \geq \sum_{n=1}^{\infty} \widetilde{P}(B_n) = \sum_{n=1}^{\infty} \widetilde{P}(B_n \cap A) + \sum_{n=1}^{\infty} \widetilde{P}(B_n \cap A^C) \geq P^*(E \cap A) + P^*(E \cap A^C).$$

As ε was arbitrary, A is P^*-measurable. □

It took some work, but we have finally arrived at our goal! Though there were a lot of details to attend to, the basic idea was pretty easy:

Start with a reasonable set function that behaves as you intend on a manageable collection of sets (protomeasure on a semialgebra); extend it in the obvious way to a slightly larger collection with an eye to countable additivity (premeasure on the generated algebra); use this to define an 'approximate measure' on all of 2^Ω (the induced outer measure); and then restrict this to get a genuine measure on a σ-algebra containing all the sets you started out with.

Theorem 3.14. *If P_0 is a protomeasure on a semialgebra $\mathcal{S}$, then there is a unique measure P on $\sigma(\mathcal{S})$ such that $P(S) = P_0(S)$ for all $S \in \mathcal{S}$.*

Proof. Extend P_0 to $\overline{\mathcal{S}}$ as in Proposition 3.10 to get a premeasure $\widetilde{P}$ on $\overline{\mathcal{S}}$ that agrees with P_0 on $\mathcal{S}$. Then extend this to the outer measure P^* defined in Proposition 3.11. The collection $\mathcal{M}$ of P^* measurable sets is a σ-algebra containing $\overline{S}$ (and thus $\mathcal{S}$ and thus $\sigma(\mathcal{S})$) with $P^*(S) = \widetilde{P}(S) = P_0(S)$ for all $S \in \mathcal{S}$. As P^* defines a measure on $\mathcal{M}$, $P = P^*|_{\sigma(\mathcal{S})}$ is as asserted.

For uniqueness, suppose that Q is another measure on $\sigma(S)$ with $Q(S) = P_0(S)$ for all $S \in \mathcal{S}$. Since $\mathcal{S}$ is a π-system and, arguing as in Example 2.7, $\mathcal{L} = \{E : P(E) = Q(E)\}$ is a λ-system containing $\mathcal{S}$, Theorem 2.6 ensures that $\sigma(\mathcal{S}) \subseteq \mathcal{L}$—that is, $P(E) = Q(E)$ for all $E \in \sigma(\mathcal{S})$. □

Observe that the Carathéodory construction actually gave us a measure $\overline{P}$ on $\mathcal{M}$ that extends P_0. In general, $\mathcal{M}$ will be larger than $\sigma(\mathcal{S})$.

To see this, we first note that if $A \in \mathcal{M}$ has the property that $P^*(A) = 0$, then for any $E \subseteq \Omega$ we have

$$P^*(E) \leq P^*(E \cap A) + P^*(E \cap A^C) \leq P^*(A) + P^*(E) = P^*(E),$$

so $\mathcal{M}$ contains all sets with outer measure 0.

Since P^* is monotone, this shows that if $B \in \mathcal{M}$ is a *null set* (so $\overline{P}(B) = 0$) and $A \subseteq B$, then $A \in \mathcal{M}$.

Definition. A measure $\overline{P}$ whose domain contains all subsets of null sets is said to be *complete*.

Theorem 3.15. *Suppose $(\Omega, \mathcal{F}, P)$ is a probability space and define $\mathcal{N} = \{N \in \mathcal{F} : P(N) = 0\}$, $\overline{\mathcal{F}} = \{E \cup F : E \in \mathcal{F}$ and $F \subseteq N$ for some $N \in \mathcal{N}\}$. Then $\overline{\mathcal{F}}$ is a σ-algebra and there is a unique extension $\overline{P}$ of P that defines a complete measure on $\overline{\mathcal{F}}$.*

Proof. Since $\mathcal{F}$ and $\mathcal{N}$ are closed under countable unions, so is $\overline{\mathcal{F}}$. Now suppose that $E \cup F \in \overline{\mathcal{F}}$, with $F \subseteq N \in \mathcal{N}$. We can assume that $E \cap N = \emptyset$—otherwise, replace F and N with $F \setminus E$ and $N \setminus E$.

Then $E \cup F = (E \cup N) \cap (F \cup N^C)$ since $x \in E \cup F$ implies $x \in F \subseteq (E \cup N), (F \cup N^C)$ or $x \in E \subseteq (E \cup N), (F \cup N^C)$, and every $x \in (E \cup N) \cap (F \cup N^C)$ is either in N and thus in F or in N^C and thus in E. This means that $(E \cup F)^C = (E \cup N)^C \cup (F \cup N^C)^C = (E \cup N)^C \cup (N \setminus F)$ with $(E \cup N)^C \in \mathcal{F}$ and $N \setminus F \subseteq N \in \mathcal{N}$. Thus $\overline{\mathcal{F}}$ is closed under complements as well.

Now given $E \in \mathcal{F}$, $F \subseteq N \in \mathcal{N}$, define $\overline{P}(E \cup F) = P(E)$. This is well-defined because if $E_1 \cup F_1 = E_2 \cup F_2$ with $E_i \in \mathcal{F}$ and $F_i \subseteq N_i \in \mathcal{N}$, then $E_1 \subseteq E_2 \cup N_2$ and $E_2 \subseteq E_1 \cup N_1$, so $P(E_1) \leq P(E_2) + P(N_2) = P(E_2)$ and $P(E_2) \leq P(E_1) + P(N_1) = P(E_1)$. $\overline{P}$ is complete because every $F \subseteq N \in \mathcal{N}$ can be written as $F = \emptyset \cup F \in \overline{\mathcal{F}}$ (and we have $\overline{P}(F) = P(\emptyset) = 0$).

If Q is any other measure on $\overline{\mathcal{F}}$ that agrees with P on $\mathcal{F}$, then for each $E \cup F \in \overline{\mathcal{F}}$ with $F \subseteq N \in \mathcal{N}$,

$$Q(E \cup F) \leq Q(E) + Q(F) = P(E) + Q(F) \leq P(E) + Q(N) = P(E) = \overline{P}(E \cup F)$$

and

$$\overline{P}(E \cup F) = P(E) = Q(E) \leq Q(E) + Q(F \setminus E) = Q(E \cup F). \qquad \square$$

In light of the preceding, we can upgrade our extension theorem so that it gives a complete measure on $\overline{\sigma(\mathcal{S})}$. This is often convenient, and we will feel free to do so when it is.

For example, the (infinite) measure m on the Borel σ-field generated from the protomeasure $\mu_0\big((a,b]\big) = b - a$ on the semialgebra $\mathcal{H}$ is called the *Lebesgue measure* on $\mathcal{B}$. Its completion gives the Lebesgue σ-algebra $\mathcal{L} := \overline{\mathcal{B}}$.

As the Cantor set K is null with respect to Lebesgue measure, every one of its subsets belongs to $\mathcal{L}$. Since K is uncountable, this produces more Lebesgue sets than the cardinality of the continuum.

However, one can show that $|\mathcal{B}| = |\mathbb{R}|$, so $\mathcal{L}$ is indeed much bigger.

(Basically, one starts with a nice generating set like $B_0 = \big\{(a,b) : a, b \in \mathbb{Q}\big\}$, then takes all complements and countable unions/intersections to get the larger set B_1. Then one takes complements and countable unions/intersections of these sets to get B_2, and so forth. However, one must perform this recursion over all countable ordinals to eventually arrive at $\mathcal{B}$, so the argument involves transfinite induction.)

For the purposes of this class, we can get all the measures on Borel/Lebesgue sets that we need by looking at protomeasures arising from distribution functions on the semialgebra of h-intervals.

One nice thing about these measures is that they are *inner* and *outer regular* in the sense that for all $E \in \mathcal{L}$,

$$\begin{aligned} P(E) &= \sup\{P(K) : K \subseteq E \text{ and } K \text{ is compact}\} \\ &= \inf\{P(U) : E \subseteq U \text{ and } U \text{ is open}\}. \end{aligned}$$

The proof is not too difficult, but in the interest of time, we leave it to independent pursuit.

Example 3.16. If $E \subseteq \mathbb{R}$ is countable and $p : \mathbb{R} \to [0,1]$ satisfies $\sum_{\omega \in E} p(\omega) = 1$ and $p(\omega) = 0$ for $\omega \in E^C$, one can check that $F(x) = \sum_{\omega \leq x} p(\omega)$ is a probability distribution function and thus defines a protomeasure on $\mathcal{H}$ which extends to the measure P on $\mathcal{B}$ defined by $P(A) = \sum_{\omega \in A} p(\omega)$. (These sums are well-defined since p vanishes outside of a countable set.) This takes care of all of the discrete distributions on $\mathbb{R}$.

Example 3.17. If $f : \mathbb{R} \to [0, \infty)$ is integrable with $\int_{-\infty}^{\infty} f(x)\,dx = 1$, then $F(x) = \int_{-\infty}^{x} f(t)\,dt$ is easily seen to be a probability distribution function and thus defines a protomeasure μ_0 on $\mathcal{H}$ which extends to a measure μ on $\mathcal{B} = \sigma(\mathcal{H})$ with the property that $\mu\big((a,b]\big) = \int_a^b f(x)\,dx$ for all $-\infty \leq a \leq b \leq \infty$. These are precisely the absolutely continuous distributions on $\mathbb{R}$ that you studied in undergraduate probability.

Example 3.18. Define $F : \mathbb{R} \to [0, \infty)$ by $F(x) = 0$ for $x \leq 0$, $F(x) = 1$ for $x \geq 1$, $F(x) = \frac{1}{2}$ for $\frac{1}{3} \leq x \leq \frac{2}{3}$, $F(x) = \frac{1}{4}$ for $\frac{1}{9} \leq x \leq \frac{2}{9}$, $F(x) = \frac{3}{4}$ for $\frac{7}{9} \leq x \leq \frac{8}{9}$, and, in general, for x in a closed middle third from the Cantor construction, let $F(x)$ be the average value of F at the endpoints of the interval from which it was removed as defined in the previous step.

(More concretely but less intuitively, suppose $x \in [0, 1]$ has ternary expansion $x = \sum_{n=1}^{\infty} \frac{a_n}{3^n}$. Define $N = \inf\{n : a_n = 1\} \in \mathbb{N} \cup \{\infty\}$ and set $b_n = \frac{a_n}{2}$ for $n < N$, $b_N = 1$ if $N < \infty$. Then $F(x) = \sum_{n=1}^{N} \frac{b_n}{2^n}$.)

This is called the *devil's staircase* and it is not too hard to show that it is a distribution function. However, for every x in the complement of the Cantor set, F is differentiable with $F'(x) = 0$, so it does not arise as the integral of any $f : \mathbb{R} \to [0, \infty)$.

In a sense, these are all of the cases that can arise.

Definition. A measure μ on $(\mathbb{R}, \mathcal{B})$ is said to be *discrete* if there is a countable set $E \subseteq \mathbb{R}$ with $\mu(E^C) = 0$. μ is *continuous* if $\mu\{x\} = 0$ for all $x \in \mathbb{R}$.

Proposition 3.19. *Any finite Borel measure can be uniquely decomposed as $\mu = \mu_d + \mu_c$ where μ_d is discrete and μ_c is continuous.*

Proof. Let $A = \{x \in \mathbb{R} : \mu\{x\} > 0\}$. For any countable $B \subseteq A$, $\sum_{x \in B} \mu\{x\} = \mu(B) < \infty$ by countable additivity and finiteness.

Therefore, $A_k = \{x \in \mathbb{R} : \mu\{x\} > k^{-1}\}$ is finite for all $k \in \mathbb{N}$, hence $A = \bigcup_{k=1}^{\infty} A_k$ is a countable union of finite sets and thus countable.

The result follows by defining $\mu_d(E) = \mu(E \cap A)$, $\mu_c(E) = \mu(E \cap A^C)$. □

Definition. If μ and ν are measures on $(S, \mathcal{G})$, then we say that ν is *absolutely continuous* with respect to μ (and write $\nu \ll \mu$) if $\nu(A) = 0$ for all $A \in \mathcal{G}$ with $\mu(A) = 0$.

We say that μ and ν are *mutually singular* (and write $\mu \perp \nu$) if there exist $E, F \in \mathcal{G}$ such that $E \cap F = \emptyset$, $E \cup F = S$, and $\mu(F) = 0 = \nu(E)$.

The Lebesgue-Radon-Nikodym theorem shows that if μ and ν are σ-finite measures, then $\nu = \rho + \lambda$ with $\rho \ll \mu$, $\lambda \perp \mu$.

(Moreover, there is a function $f : \Omega \to [0, \infty)$ such that $\rho(A) = \int_A f \, d\mu$ for all $A \in \mathcal{F}$. If μ is Lebesgue measure on $\mathbb{R}$, then for all practical purposes, this is just the usual Riemann integral $\int_A f(x) \, dx$.)

Thus we can first decompose any Borel probability measure as $P = \mu_d + \mu_c$ with μ_d discrete and μ_c continuous.

Then we write $\mu_c = \mu_{ac} + \mu_{sc}$ with μ_{ac} absolutely continuous with respect to Lebesgue measure and μ_{sc} singular with respect to Lebesgue measure.

It follows that every Borel probability measure can be written as a convex combination of discrete, absolutely continuous, and singular continuous probability measures.

4 Random Variables

Having carefully constructed a rich set of probability spaces in the previous section, we will typically just take it as given going forward that there is an underlying space $(\Omega, \mathcal{F}, P)$ on which we are working.

Our next task is to enrich this basic framework by introducing a few more fundamental constructs.

Definition. A (real-valued) *random variable* on a probability space $(\Omega, \mathcal{F}, P)$ is a function $X : \Omega \to \mathbb{R}$ such that for all $E \in \mathcal{B}$, $X^{-1}(E) \in \mathcal{F}$.

Example 4.1. For any $c \in \mathbb{R}$, the constant function $X \equiv c$ is a random variable since for each $A \in \mathcal{B}$, $X^{-1}(A) = \Omega$ if $c \in A$ and $X^{-1}(A) = \emptyset$ if $c \notin A$, both of which belong to $\mathcal{F}$.

Likewise, if $E \in \mathcal{F}$, then $X = 1_E$ is a random variable since for every $A \in \mathcal{B}$, $X^{-1}(A) \in \{\emptyset, A, A^C, \Omega\}$ depending on whether A contains 0, 1, neither, or both.

In general, if $(S, \mathcal{G})$ is a *measurable space* (so $\mathcal{G}$ is a σ-field over S), a map $X : \Omega \to S$ is *measurable* if $X^{-1}(G) \in \mathcal{F}$ for all $G \in \mathcal{G}$, and we say that X is an $(S, \mathcal{G})$-valued random variable. If the target σ-field is understood, we often employ the slight abuse of notation $X \in \mathcal{F}$ to indicate that X is $\mathcal{F}$-$\mathcal{G}$ measurable.

Theorem 4.2. *If $\mathcal{A}$ generates $\mathcal{G}$ (in the sense that $\mathcal{G}$ is the smallest σ-algebra containing $\mathcal{A}$) and $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{A}$, then X is an $(S, \mathcal{G})$-valued random variable.*

Proof. Because $X^{-1}\left(\bigcup_i E_i\right) = \bigcup_i X^{-1}(E_i)$ and $X^{-1}\left(E^C\right) = X^{-1}(E)^C$, $\mathcal{E} = \left\{E \subseteq S : X^{-1}(E) \in \mathcal{F}\right\}$ is a σ-algebra. Thus, since $\mathcal{A} \subseteq \mathcal{E}$ and $\mathcal{A}$ generates $\mathcal{G}$, $\mathcal{G} \subseteq \mathcal{E}$, hence X is measurable. □

The fact that inverses commute with set operations also shows that for any map $X : \Omega \to S$, if $\mathcal{G}$ is a σ-algebra on S, then $\sigma(X) = \{X^{-1}(E) : E \in \mathcal{G}\}$ is a σ-algebra on Ω (called the *σ-algebra generated by X*). By construction, it's the smallest σ-algebra on Ω that makes X an $(S, \mathcal{G})$-valued random variable.

We will mostly be concerned with the case $(S, \mathcal{G}) = (\mathbb{R}, \mathcal{B})$, but abstract definitions can be easier to work with since they tend to push extraneous details into the background.

Also, even if the focus is primarily on $\mathbb{R}$-valued random variables, it is often convenient to consider *random vectors* when stating and proving theorems. This corresponds to the case $(S, \mathcal{G}) = (\mathbb{R}^d, \mathcal{B}^d)$, where the Borel σ-field on $\mathbb{R}^d$ is generated, for example, by $\mathcal{R} = \left\{(a_1, b_1] \times \cdots \times (a_d, b_d] : -\infty < a_i < b_i < \infty\right\}$; see the appendix for a careful proof.

Proposition 4.3. *If $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous, then f is $\mathcal{B}^n$-$\mathcal{B}^m$ measurable.*

Proof. Continuity means that $f^{-1}(U)$ is open in $\mathbb{R}^n$ for every open set $U \in \mathbb{R}^m$. Since every open subset of $\mathbb{R}^n$ is contained in $\mathcal{B}^n$ and the open subsets of $\mathbb{R}^m$ generate $\mathcal{B}^m$, the assertion follows from Theorem 4.2.

(The same proof works for continuous maps between arbitrary topological spaces equipped with their Borel σ-fields.) □

Proposition 4.4. *If $(S_1, \mathcal{G}_1), (S_2, \mathcal{G}_2), (S_3, \mathcal{G}_3)$ are measurable spaces and $f : S_1 \to S_2$, $g : S_2 \to S_3$ are measurable maps, then $g \circ f : S_1 \to S_3$ is measurable.*

Proof. Given any $G \in \mathcal{G}_3$, measurability of g ensures that $g^{-1}(G) \in \mathcal{G}_2$, so measurability of f implies $(g \circ f)^{-1}(G) = f^{-1}\left(g^{-1}(G)\right) \in \mathcal{G}_1$. □

Corollary 4.5. *If $f : \mathbb{R} \to \mathbb{R}$ is continuous and X is a random variable, then $Y = f(X)$ is a random variable as well.*

Theorem 4.6. *If $X_1, \dots, X_n$ are $\mathbb{R}$-valued random variables, then so are $S_n = \sum_{k=1}^n X_k$ and $V_n = \prod_{k=1}^n X_k$.*

Proof. We first observe that the map $\mathbb{X} : \Omega \to \mathbb{R}^d$ given by $\mathbb{X}(\omega) = \big(X_1(\omega), \dots, X_n(\omega)\big)$ is measurable.
Indeed, $\mathcal{B}^d$ is generated by rectangles of the form $(a_1, b_1] \times \cdots \times (a_d, b_d]$ and

$$\mathbb{X}^{-1}\big((a_1, b_1] \times \cdots \times (a_d, b_d]\big) = \big\{\omega \in \Omega : \big(X_1(\omega), \dots, X_n(\omega)\big) \in (a_1, b_1] \times \cdots \times (a_d, b_d]\big\} = \bigcap_{k=1}^n X_k^{-1}\big((a_k, b_k]\big),$$

which is a finite intersection of sets in $\mathcal{F}$ (since each X_k is measurable) and thus belongs to $\mathcal{F}$.
Next, we note that the maps $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ defined by $f\big((x_1, \dots, x_n)\big) = x_1 + \cdots + x_n$ and $g\big((x_1, \dots, x_n)\big) = x_1 \cdots x_n$ are continuous and thus measurable by Proposition 4.3.
(The projection functions $\pi_k(\mathbf{x}) = x_k$ are continuous for each k, so f can be expressed as a sum of continuous functions and g as a product.)
Therefore, $S_n = f(\mathbb{X})$ and $V_n = g(\mathbb{X})$ are compositions of measurable functions, and the claim follows from Proposition 4.4. □

Remark 4.7. Note that if X is measurable with respect to $\mathcal{F}$ and $\widetilde{\mathcal{F}}$ is any σ-field containing $\mathcal{F}$, then X is automatically measurable with respect to $\widetilde{\mathcal{F}}$. In particular, we can always consider the completion of the source σ-field.
However, enlarging the target σ-field provides more opportunities for maps to fail to be measurable. For example, one can construct continuous functions from $\mathbb{R}$ to $\mathbb{R}$ that are not $\mathcal{L}$-$\mathcal{L}$ measurable. Similarly, if $f : (R, \mathcal{E}) \to (S, \mathcal{F}_1)$ and $g : (S, \mathcal{F}_2) \to (T, \mathcal{G})$ are measurable, $g \circ f : (R, \mathcal{E}) \to (T, \mathcal{G})$ need not be unless $\mathcal{F}_2 \subseteq \mathcal{F}_1$. This is why we use the Borel rather than Lebesgue σ-algebra in our definition of $\mathbb{R}$-valued random variables.

It is sometimes convenient to allow random variables to assume the values $\pm\infty$, and we observe that almost all of our results generalize easily to $(\mathbb{R}^*, \mathcal{B}^*)$ where $\mathbb{R}^* = \mathbb{R} \cup \{\pm\infty\}$ and $\mathcal{B}^* = \big\{E \subseteq \overline{\mathbb{R}} : E \cap \mathbb{R} \in \mathcal{B}\big\}$, which is generated, for example, by rays of the form $[-\infty, a)$ with $a \in \mathbb{R}$.

Theorem 4.8. *If $X_1, X_2, \dots$ are random variables, then so are*

$$\inf_{n\in\mathbb{N}} X_n, \quad \sup_{n\in\mathbb{N}} X_n, \quad \liminf_{n\to\infty} X_n, \quad \limsup_{n\to\infty} X_n.$$

Proof. For any $a \in \mathbb{R}$, the infimum of a sequence is strictly less than a if and only if some term is strictly less than a, hence

$$\big\{\inf_{n\in\mathbb{N}} X_n < a\big\} = \bigcup_{n\in\mathbb{N}} \big\{X_n < a\big\} \in \mathcal{F}.$$

Since $\big\{[-\infty, a) : a \in \mathbb{R}\big\}$ generates $\mathcal{B}^*$, we conclude that $\inf_{n\in\mathbb{N}} X_n$ is measurable.
To see that $\sup_{n\in\mathbb{N}} X_n$ is a random variable, note that $\sup_{n\in\mathbb{N}} X_n = -\inf_{n\in\mathbb{N}} -X_n$ and $x \mapsto -x$ is measurable.
Arguing as in the first case, $\inf_{m\geq n} X_m$ is measurable for all $m \in \mathbb{N}$, so it follows from the second case that

$$\liminf_{n\to\infty} X_n = \sup_{n\in\mathbb{N}} \left(\inf_{m\geq n} X_m\right)$$

is a random variable. The lim sup case is similar. □

It follows from Theorem 4.8 that

$$\left\{\lim_{n\to\infty} X_n \text{ exists}\right\} = \left\{\liminf_{n\to\infty} X_n = \limsup_{n\to\infty} X_n\right\} = \left\{\liminf_{n\to\infty} X_n - \limsup_{n\to\infty} X_n = 0\right\}$$

is measurable since it is the preimage of $\{0\} \in \mathcal{B}$ under the map $(\liminf_{n\to\infty} X_n) - (\limsup_{n\to\infty} X_n)$, which is the difference of measurable functions and thus measurable.

When $P\{\lim_{n\to\infty} X_n \text{ exists}\} = 1$, we say that the sequence $\{X_n\}$ converges *almost surely* to $X := \limsup_{n\to\infty} X_n$, and write $X_n \to X$ a.s.

Distributions

Every random variable induces a probability measure μ on $\mathbb{R}$ (called its *distribution*) by

$$\mu(A) = P\left(X^{-1}(A)\right)$$

for all $A \in \mathcal{B}$.

To check that μ is a probability measure, note that since X is a function, if $A_1, A_2, \ldots \in \mathcal{B}$ are disjoint, then so are $\{X \in A_1\}, \{X \in A_2\}, \ldots \in \mathcal{F}$, hence

$$\mu\left(\bigcup_i A_i\right) = P\left(\{X \in \bigcup_i A_i\}\right) = P\left(\bigcup_i \{X \in A_i\}\right) = \sum_i P\{X \in A_i\} = \sum_i \mu(A_i).$$

The distribution of a random variable X is usually described in terms of its *distribution function*

$$F(x) = P(X \le x) = \mu\big((-\infty, x]\big).$$

In cases where confusion may arise, we will emphasize dependence on the random variable using subscripts—i.e. μ_X, F_X.

Theorem 4.9. *If F is the distribution function of a random variable X, then*

(i) F is nondecreasing

(ii) F is right-continuous (so $\lim_{x\to a^+} F(x) = F(a)$ for all $a \in \mathbb{R}$)

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

(iv) If $F(x^-) = \lim_{y\to x^-} F(y)$, then $F(x^-) = P(X < x)$

(v) $P(X = x) = F(x) - F(x^-)$

Proof.

For (i), if $x \le y$, then $\{X \le x\} \subseteq \{X \le y\}$, so $F(x) = P(X \le x) \le P(X \le y) = F(y)$ by monotonicity.

For (ii), observe that if $x \searrow a$, then $\{X \le x\} \searrow \{X \le a\}$, and apply continuity from above.

For (iii), we have $\{X \le x\} \searrow \emptyset$ as $x \searrow -\infty$ and $\{X \le x\} \nearrow \mathbb{R}$ as $x \nearrow \infty$.

For (iv), $\{X \le y\} \nearrow \{X < x\}$ as $y \nearrow x$. (Note that the limit exists since F is monotone.)

For (v), $\{X = x\} = \{X \le x\} \setminus \{X < x\}$. □

It is perhaps worth observing that if D is the set of discontinuity points of a distribution function F, then monotonicity and right-continuity ensure that $\{(F(d^-), F(d)) : d \in D\}$ is a collection of disjoint open intervals. As each must contain a rational number, D is necessarily countable.

Theorem 4.10. *If $F : \mathbb{R} \to \mathbb{R}$ satisfies properties (i), (ii), and (iii) from Theorem 4.9, then it is the distribution function of some random variable.*

Proof. Let $\Omega = (0,1)$, $\mathcal{F} = \mathcal{B}_{(0,1)}$, $P =$ Lebesgue measure, and define $X : (0,1) \to \mathbb{R}$ by

$$X(\omega) = F^{-1}(\omega) := \inf\{y \in \mathbb{R} : F(y) \geq \omega\}.$$

Note that properties (i) and (iii) ensure that X is well-defined.

To see that F is indeed the distribution function of X, it suffices to show that

$$\{\omega : X(\omega) \leq x\} = \{\omega : \omega \leq F(x)\}$$

for all $x \in \mathbb{R}$, as this implies

$$P(X \leq x) = P\{\omega : X(\omega) \leq x\} = P\{\omega : \omega \leq F(x)\} = F(x)$$

where the final equality uses the definition of Lebesgue measure and the fact that $F(x) \in [0,1]$.

Now if $\omega \leq F(x)$, then $x \in \{y \in \mathbb{R} : F(y) \geq \omega\}$, so $X(\omega) = \inf\{y \in \mathbb{R} : F(y) \geq \omega\} \leq x$.

This shows that $\{\omega : \omega \leq F(x)\} \subseteq \{\omega : X(\omega) \leq x\}$.

To establish the reverse inclusion, observe that if $\omega > F(x)$, then properties (i) and (ii) imply that there is an $\varepsilon > 0$ such that $F(x) \leq F(x+\varepsilon) < \omega$.

Since F is nondecreasing, $x + \varepsilon$ is a lower bound for $\{y \in \mathbb{R} : F(y) \geq \omega\}$, hence $X(\omega) \geq x + \varepsilon > x$.

Therefore, $\{\omega : \omega \leq F(x)\}^C \subseteq \{\omega : X(\omega) \leq x\}^C$ and thus $\{\omega : X(\omega) \leq x\} \subseteq \{\omega : \omega \leq F(x)\}$. □

Theorem 4.10 shows that any function satisfying properties (i) - (iii) gives rise to a random variable X, and thus to a probability measure μ, the distribution of X.

Example 2.7 shows that this measure is uniquely determined—that is, if two random variables have the same distribution function, then they have the same distribution.

To summarize, every random variable induces a probability measure on $(\mathbb{R}, \mathcal{B})$, every probability measure defines a function satisfying properties (i)-(iii) in Theorem 4.9, and every such function uniquely determines a probability measure.

Consequently, it is equivalent to give the distribution or the distribution function of a random variable.

However, one should be aware that distributions/distribution functions do not determine random variables, even neglecting differences on null sets.

Definition. If X and Y are defined on a common probability space and $P(X = Y) = 1$, we say that $X = Y$ *almost surely* (or a.s. for short).

Definition. When two random variables X and Y have the same distribution function, we say that they are *equal in distribution* and write $X =_d Y$.

If $X = Y$ a.s., then $X =_d Y$ since for any $E \in \mathcal{B}$,

$$\begin{aligned}\mu_X(E) &= P(X \in E, X = Y) + P(X \in E, X \neq Y) = P(X \in E, X = Y) \\ &= P(Y \in E, X = Y) = P(Y \in E, X = Y) + P(Y \in E, X \neq Y) = \mu_Y(E)\end{aligned}$$

where the commas denote intersections of events.

The converse is not true though. For example, if X is uniform on $[-1, 1]$—so $\mu_X = \frac{1}{2} m|_{[-1,1]}$—then $-X$ also has distribution μ_X, but $-X \neq X$ a.s.

Moreover, random variables can be equal in distribution even if they are defined on different probability spaces.

5 Independence

Another fundamental concept in probability theory is independence. Heuristically, two objects are independent if information concerning one of them does not contribute to one's knowledge about the other.

In order to motivate a general and rigorous formulation of this notion, it's helpful to first recall the classical notion of conditional probability.
(A more sophisticated version involves conditioning on sub-σ-fields and replacing probabilities with random variables, but we'll have to wait a while to be able to develop that approach.)

Suppose that we have a reasonable model of some experiment, encoded by the probability space $(\Omega, \mathcal{F}, P)$, and then somehow became convinced that the experiment will result in an outcome belonging to $F \in \mathcal{F}$. For reasons that will soon become clear, let us suppose moreover that $P(F) > 0$. How should we go about updating our model to account for this additional information?

One way to proceed would be to collapse our sample space to F, restrict our σ-field to $\{E \cap F : E \in \mathcal{F}\}$, and normalize our probability measure so that F is assigned probability 1. Effectively, this will result in the same definition we are about to derive, but in fact, it is sufficient (and desirable from the perspective of extending probability spaces, etc.) to retain $(\Omega, \mathcal{F})$ and only modify P.

Let's denote the updated probability measure by P_F. Other than satisfying the definition of a probability on $(\Omega, \mathcal{F})$, it is natural to require that

(1) $P_F(F) = 1$ (since we know that F occurs)
(2) For any events $A, B \in \mathcal{F}$ with $A, B \subseteq F$ and $P(B) > 0$, $P_F(A)/P_F(B) = P(A)/P(B)$ (since we have learned nothing about the relative probabilities of A and B)

Since $1 = P_F(\Omega) = P_F(F) + P_F(F^C) = 1 + P_F(F^C)$, we must have that $P_F(F^C) = 0$ and thus $P_F(G) = 0$ for any $G \subseteq F^C$ by monotonicity.
It follows that for any $E \in \mathcal{F}$,

$$P_F(E) = P_F(E \cap F) + P_F(E \cap F^C) = P_F(E \cap F) = \frac{P_F(E \cap F)}{P_F(F)} = \frac{P(E \cap F)}{P(F)}.$$

It is a simple exercise to check that this definition of P_F does indeed give a valid probability on $(\Omega, \mathcal{F})$.

Now if knowledge of the occurrence of F was completely uninformative about the probability we should assign to E, then we would have $P(E) = P_F(E) = \frac{P(E \cap F)}{P(F)}$, hence $P(E \cap F) = P(E)P(F)$.

It turns out that this product formulation of independence is the right choice since it is simple to work with, generalizes nicely, and sidesteps the problem of potential division by 0.

Formally, let us say that

- Two events A and B are independent if $P(A \cap B) = P(A)P(B)$.
- Two random variables X and Y are independent if $P(X \in E, Y \in F) = P(X \in E)P(Y \in F)$ for all $E, F \in \mathcal{B}$. (That is, if the events $\{X \in E\}$ and $\{Y \in F\}$ are independent.)
- Two sub-σ-fields $\mathcal{F}_1$ and $\mathcal{F}_2$ are independent if for all $A \in \mathcal{F}_1$, $B \in \mathcal{F}_2$, the events A and B are independent.

Observe that if $A \in \mathcal{F}$ has $P(A) = 0$ or $P(A) = 1$, then A is independent of every $B \in \mathcal{F}$.
This also implies that if X is a.s. constant, then X is independent of every $Y \in \mathcal{F}$ and that $\{\emptyset, \Omega\}$ is independent of every sub-σ-field.

This is all well and good so far, but we immediately run into problems if we try to naively extend these multiplication rules to more than two events/random variables/sub-σ-fields.

For instance, if A, B, C are events with $P(A \cap B) \neq P(A)P(B)$ and $P(C) = 0$, then we have $P(A \cap B \cap C) = 0 = P(A)P(B)P(C)$, but it wouldn't make sense to say that $\{A, B, C\}$ is an independent collection of events since A and B are dependent.

On the other hand, just because any pair of events in a collection is independent, it does not follow that the entire collection should be regarded as such. As an example, consider the experiment where two fair coins are flipped and set $A = \{\text{1st coin heads}\}$, $B = \{\text{2nd coin heads}\}$, $C = \{\text{both coins same}\}$. It's straightforward to check that $P(A \cap B) = P(A \cap C) = P(B \cap C) = \frac{1}{4} = P(A)P(B) = P(A)P(C) = P(B)P(C)$, but $P(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8} = P(A)P(B)P(C)$.

As such, we say that an infinite collection of objects is independent if every finite subcollection is, where

- Events $A_1, ..., A_n \in \mathcal{F}$ are independent if for any $I \subseteq [n]$, we have
$$P\Big(\bigcap_{i \in I} A_i\Big) = \prod_{i \in I} P(A_i).$$
- Random variables $X_1, ..., X_n \in \mathcal{F}$ are independent if for any choice of $E_i \in \mathcal{B}_i$, $i = 1, ..., n$, we have
$$P(X_1 \in E_1, ..., X_n \in E_n) = \prod_{i=1}^{n} P(X_i \in E_i).$$
- Sub-σ-fields $\mathcal{F}_1, ..., \mathcal{F}_n$ are independent if for any choice of $A_i \in \mathcal{F}_i$, $i = 1, ..., n$, we have
$$P\Big(\bigcap_{i=1}^{n} A_i\Big) = \prod_{i=1}^{n} P(A_i).$$

Note that σ-algebras and random variables are implicitly subject to the same subcollection constraint as events since special cases of the definition include taking $A_i = \Omega$, $E_i = \mathbb{R}$ for $i \in I^C$.

One can show that independence of events is a special case of independence of random variables (via indicators), which in turn is a special case of independence of sub-σ-fields (via the generated σ-fields).

We will take as our running definition of independence, the further generalization:

Definition. Given a probability space $(\Omega, \mathcal{F}, P)$, collections of events $\mathcal{C}_1, ..., \mathcal{C}_n \subseteq \mathcal{F}$ are *independent* if for all $I \subseteq [n]$,
$$P\Big(\bigcap_{i \in I} A_i\Big) = \prod_{i \in I} P(A_i)$$
whenever $A_i \in \mathcal{C}_i$ for each $i \in I$.
An infinite collection of subsets of $\mathcal{F}$ is independent if every finite subcollection is.

Note that if $\mathcal{C}_1, ..., \mathcal{C}_n$ are independent and we set $\overline{\mathcal{C}_i} = \mathcal{C}_i \cup \{\Omega\}$, then $\overline{\mathcal{C}_1}, ..., \overline{\mathcal{C}_n}$ are independent as well. In this case, the independence criterion reduces to $P\Big(\bigcap_{i=1}^{n} A_i\Big) = \prod_{i=1}^{n} P(A_i)$ for any choice of $A_i \in \overline{\mathcal{C}_i}$.

These definitions seem to require us to check an impractical number of cases to determine whether a given collection of objects is independent. The following results are useful for simplifying this task.

Theorem 5.1. *Suppose that $\mathcal{C}_1, ..., \mathcal{C}_n$ are independent collections of events. If each $\mathcal{C}_i$ is a π-system, then the sub-σ-algebras $\sigma(\mathcal{C}_1), ..., \sigma(\mathcal{C}_n)$ are independent.*

Proof. Because Ω is independent of every event and $\sigma(\mathcal{C}_i) = \sigma(\overline{\mathcal{C}_i})$, we can assume without loss of generality that $\Omega \in \mathcal{C}_i$ for all i so that we need only consider intersections/products over $[n]$.

Let $A_2, ..., A_n$ be events with $A_i \in \mathcal{C}_i$, set $F = \bigcap_{i=2}^n A_i$, and set $\mathcal{L} = \{A \in \mathcal{F} : P(A \cap F) = P(A)P(F)\}$.

Since $P(\Omega \cap F) = P(F) = P(\Omega)P(F)$, we have that $\Omega \in \mathcal{L}$.

Now suppose that $A, B \in \mathcal{L}$ with $A \subseteq B$. Then

$$\begin{aligned} P\left((B \setminus A) \cap F\right) &= P\left((B \cap F) \setminus (A \cap F)\right) = P(B \cap F) - P(A \cap F) \\ &= P(B)P(F) - P(A)P(F) = \left(P(B) - P(A)\right) P(F) = P(B \setminus A)P(F), \end{aligned}$$

hence $(B \setminus A) \in \mathcal{L}$.

Finally, let $B_1, B_2, ... \in \mathcal{L}$ with $B_n \nearrow B$. Then $(B_n \cap F) \nearrow (B \cap F)$, so

$$P(B \cap F) = \lim_{n\to\infty} P(B_n \cap F) = \lim_{n\to\infty} P(B_n)P(F) = P(B)P(F),$$

so $B \in \mathcal{L}$ as well.

Therefore, $\mathcal{L}$ is a λ-system, so, since $\mathcal{C}_1$ is a π-system contained in $\mathcal{L}$ by assumption, the π-λ Theorem shows that $\sigma(\mathcal{C}_1) \subseteq \mathcal{L}$.

Because $A_2, ..., A_n$ were arbitrary members of $\mathcal{C}_2, ..., \mathcal{C}_n$, we conclude that $\sigma(\mathcal{C}_1), \mathcal{C}_2, ..., \mathcal{C}_n$ are independent.

Repeating this argument for $\mathcal{C}_2, \mathcal{C}_3, ..., \mathcal{C}_n, \sigma(\mathcal{C}_1)$ shows that $\sigma(\mathcal{C}_2), \mathcal{C}_3, ..., \mathcal{C}_n, \sigma(\mathcal{C}_1)$ are independent, and $n-2$ more iterations completes the proof. □

Corollary 5.2. *Random variables $X_1, ..., X_n$ are independent if*

$$P(X_1 \leq x_1, ..., X_n \leq x_n) = \prod_{i=1}^n P(X_i \leq x_i) \text{ for all } x_1, ..., x_n \in \mathbb{R}.$$

Proof. Let $\mathcal{C}_i = \{\{X_i \leq x\} : x \in \mathbb{R}\}$ for $i = 1, ..., n$.

Since $\{X_i \leq x\} \cap \{X_i \leq y\} = \{X_i \leq x \wedge y\}$, the $\mathcal{C}_i$'s are π-systems, so $\sigma(\mathcal{C}_1), ..., \sigma(\mathcal{C}_n)$ are independent by Theorem 5.1.

Because $\{(-\infty, x] : x \in \mathbb{R}\}$ generates $\mathcal{B}$, $\sigma(\mathcal{C}_i) = \sigma(X_i)$, and the result follows. □

Since the converse of Corollary 5.2 is true by definition, independence of random variables $X_1, ..., X_n$ is equivalent to the condition that their joint cdf factors as a product of the marginals cdfs.

One can prove analogous results for density and mass functions using the same basic ideas.

If $X_1, ..., X_n$ are independent random variables and $f_1, ..., f_n : \mathbb{R} \to \mathbb{R}$ are measurable, then $f(X_1), ..., f(X_n)$ are independent random variables since for any choice of $B_i \in \mathcal{B}_i$,

$$\begin{aligned} P\left(f_1(X_i) \in B_1, ..., f_n(X_n) \in B_n\right) &= P\left(X_1 \in f_1^{-1}(B_1), ..., X_n \in f_n^{-1}(B_n)\right) \\ &= \prod_{i=1}^n P\left(X_i \in f_i^{-1}(B_i)\right) = \prod_{i=1}^n P\left(f_i(X_i) \in B_i\right). \end{aligned}$$

With the help of Theorem 5.1, we can prove the stronger result that functions of disjoint sets of independent random variables are independent.

Lemma 5.3. *Suppose $\mathcal{F}_{i,j}$, $1 \leq i \leq n$, $1 \leq j \leq m(i)$, are independent sub-σ-algebras and let $\mathcal{G}_i = \sigma\left(\bigcup_j \mathcal{F}_{i,j}\right)$. Then $\mathcal{G}_1, ..., \mathcal{G}_n$ are independent.*

Proof. Let $\mathcal{C}_i = \left\{\bigcap_j A_{i,j} : A_{i,j} \in \mathcal{F}_{i,j}\right\}$.

If $\bigcap_j A_{i,j}, \bigcap_j B_{i,j} \in \mathcal{C}_i$, then $\left(\bigcap_j A_{i,j}\right) \cap \left(\bigcap_j B_{i,j}\right) = \bigcap_j (A_{i,j} \cap B_{i,j}) \in \mathcal{C}_i$, so $\mathcal{C}_i$ is a π-system and Theorem 5.1 guarantees that $\sigma(\mathcal{C}_1), ..., \sigma(\mathcal{C}_n)$ are independent.

Because $F \in \bigcup_j \mathcal{F}_{i,j}$ implies $F \in \mathcal{F}_{i,k}$ for some k and thus $F = \Omega \cap \cdots \cap \Omega \cap F \cap \Omega \cap \cdots \cap \Omega \in \mathcal{C}_i$, we have that $\bigcup_j \mathcal{F}_{i,j} \subseteq \mathcal{C}_i$, so $\mathcal{G}_i = \sigma\left(\bigcup_j \mathcal{F}_{i,j}\right) \subseteq \sigma(\mathcal{C}_i)$. Consequently, $\mathcal{G}_1, ..., \mathcal{G}_n$ are independent. □

Corollary 5.4. *If $X_{i,j}$, $1 \leq i \leq n$, $1 \leq j \leq m(i)$, are independent random variables and the functions $f_i : \mathbb{R}^{m(i)} \to \mathbb{R}$ are measurable, then $f_1(X_{1,1}, ..., X_{1,m(1)}), ..., f_n(X_{n,1}, ..., X_{n,m(n)})$ are independent.*

Proof. Let $\mathcal{F}_{i,j} = \sigma(X_{i,j})$. Since $f_i(X_{i,1}, ..., X_{i,m(i)})$ is measurable with respect to $\mathcal{G}_i = \sigma\left(\bigcup_j \mathcal{F}_{i,j}\right)$, the result follows from Lemma 5.3. □

6 Expectation

Random variables often enable us to summarize or simplify probability experiments by attaching a single real number to each outcome.

In many cases we can obtain a further useful reduction by distilling the random variable down to a fixed number that represents our 'best guess' for the value it takes. This notion of typicality or centrality is formally encoded in terms of the random variable's *expected value* (or *expectation*).

If X is a *discrete* random variable (so its range is countable), we can encode its distribution via the *probability mass function* $p_X(x) = P(X = x)$. In this case, the expectation is just the probability-weighted average of the values it takes, $E[X] = \sum_{x\in\mathbb{R}} x\, p_X(x)$.

Since all summands are nonnegative and only countably many are nonzero, this sum is well-defined. If we further assume that Ω itself is countable, this is equivalent to $E[X] = \sum_{\omega\in\Omega} X(\omega)P\{\omega\}$.

In undergraduate probability, this is generalized to the case where X is *absolutely continuous* (so there is a function $f_X : \mathbb{R} \to [0,\infty)$, called the *probability density function* of X, such that $F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)\,dt$) by declaring that $E[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx$.

The intuition is that integrals are the continuous version of sums. To flesh out this line of reasoning, observe that if f_X is continuous, then $P(x - \frac{\varepsilon}{2} \le X \le x + \frac{\varepsilon}{2}) = \int_{x-\varepsilon/2}^{x+\varepsilon/2} f_X(t)\,dt \approx \varepsilon f_X(x)$ for $\varepsilon > 0$ sufficiently small, so the density function is playing a similar role to the mass function.

If we also have that $X \in [a,b]$ a.s., then we can imitate the definition of expectation for discrete random variables by taking a tagged partition $a = x_0 \le t_1 \le x_1 \le \cdots \le x_{n-1} \le t_n \le x_n = b$ and forming the sum $\sum_{k=1}^{n} t_k P(x_{k-1} \le X \le x_k) \approx \sum_{k=1}^{n} t_k f_X(t_k)(x_k - x_{k-1})$, which will converge to $\int_a^b x f_X(x)\,dx$ as the mesh of the partition tends to 0. If X is unbounded, taking $a \to -\infty$, $b \to \infty$ gives $E[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx$ whenever the improper integral converges.

For general X, we could play the same game by defining $E[X]$ as the limit as $\max_{1\le k\le n}(x_k - x_{k-1}) \to 0$ of

$$\sum_{k=1}^{n} t_k P(x_{k-1} \le X \le x_k) = \sum_{k=1}^{n} t_k [F_X(x_k) - F_X(x_{k-1})].$$

This gives the expectation of X as the Riemann-Stieltjes integral of x with respect to F_X. Of course, there are a lot of details to attend to if we want to make it rigorous.

Rather than pursue this approach, we will consider a theory of expectation/integration that is better suited to proving general theorems, taking limits, and accommodating a larger class of integrands. The key idea is to partition the codomain rather than the domain when approximating the integral by a sum.

Definition. If X is a real-valued random variable defined on a probability space $(\Omega, \mathcal{F}, P)$, we define $E[X] = \int_\Omega X(\omega)\,dP(\omega)$ (or just $\int X\,dP$ for short), provided this integral exists in $\mathbb{R}^*$.

To construct the above integral, we first consider the case where X is the indicator of some $A \in \mathcal{F}$. Since $X(\omega) = 1_A(\omega)$ equals 1 on A and 0 on A^C, it's natural to define $\int X\,dP = P(A)$.

Next, we extend this to linear combinations of indicators in the obvious way. That is, if $X = \sum_{k=1}^{n} x_k 1_{A_k}$, we would like to say that $\int X\,dP = \sum_{k=1}^{n} x_k P(A_k)$.

However, such representations are not necessarily unique (e.g. $21_A = \frac{3}{4}1_A + \frac{5}{4}1_A$ and $1_{\{a\}} + 21_{\{b\}} = 1_{\{a,b\}} + 1_{\{b\}}$), so we need to be careful.

To this end, we say that X is *simple* if there is a finite partition of the sample space $\Omega = \bigsqcup_{k=1}^n A_k$ and real numbers $x_1, \dots, x_n$ such that $X = \sum_{k=1}^n x_k 1_{A_k}$.

Equivalently, X is simple if $\text{Range}(X) = \{X(\omega) : \omega \in \Omega\}$ is finite because if $\text{Range}(X) = \{x_1, \dots, x_n\}$, we have the canonical representation $X = \sum_{k=1}^n x_k 1_{A_k}$ where $A_k = \{\omega \in \Omega : X(\omega) = x_k\}$.

Example 6.1. It's easy to check that for any disjoint events $A, B \in \mathcal{F}$, $1_{A \sqcup B} = 1_A + 1_B$, and for any events $A, B \in \mathcal{F}$, $1_{A \cap B} = 1_A 1_B$.

It follows that if $X = \sum_{j=1}^m x_j 1_{A_j}$ and $Y = \sum_{k=1}^n y_k 1_{B_k}$ are simple, then so are $X + Y$ and XY.

Indeed, we can write $A_j = \bigsqcup_{k=1}^n (A_j \cap B_k)$ and $B_k = \bigsqcup_{j=1}^m (A_j \cap B_j)$, so that $\Omega = \bigsqcup_{(j,k) \in [m] \times [n]} (A_j \cap B_k)$,

$$X + Y = \sum_{j=1}^m x_j \sum_{k=1}^n 1_{A_j \cap B_k} + \sum_{k=1}^n y_k \sum_{j=1}^m 1_{A_j \cap B_k} = \sum_{j=1}^m \sum_{k=1}^n (x_j + y_k) 1_{A_j \cap B_k},$$

and

$$XY = \left(\sum_{j=1}^m x_j 1_{A_j} \right) \left(\sum_{k=1}^n y_k 1_{B_k} \right) = \sum_{j=1}^m \sum_{k=1}^n x_j y_k 1_{A_j} 1_{B_j} = \sum_{j=1}^m \sum_{k=1}^n x_j y_k 1_{A_j \cap B_k}.$$

(Of course this also follows from the finite range characterization of simple functions.)

Now for a simple random variable $X = \sum_{k=1}^n x_k 1_{A_k}$, we define $\int X \, dP = \sum_{k=1}^n x_k P(A_k)$.

Even with the requirement that the constituent events partition the sample space, there may still be many ways to express a simple function. For instance, $1_\Omega = 1_A + 1_{A^c}$ for any $A \in \mathcal{F}$.

However, by appealing to the aforementioned canonical representation, we see that $\sum_{j=1}^m a_j 1_{A_j} = \sum_{k=1}^n b_k 1_{B_k}$ implies $\sum_{j=1}^m a_j P(A_j) = \sum_{k=1}^n b_k P(B_k)$. (Just union together any events having the same weights.)

Example 6.2. Let $(\Omega, \mathcal{F}, P)$ be Lebesgue measure on $[0, 1]$ and consider the simple random variables

$$X(\omega) = \begin{cases} 2, & \omega \le 1/3 \\ 5, & \omega > 1/3 \end{cases}, \qquad Y(\omega) = \begin{cases} 2, & \omega \in \mathbb{Q} \\ 4, & \omega = 1/\sqrt{2} \\ 6, & \omega \notin \mathbb{Q}, \omega < \frac{1}{4} \\ 8, & \text{else} \end{cases}, \qquad Z \equiv c.$$

One readily checks that $E[X] = 2 \cdot \frac{1}{3} + 5 \cdot \frac{2}{3} = 4$, $E[Y] = 2 \cdot 0 + 4 \cdot 0 + 6 \cdot \frac{1}{4} + 8 \cdot \frac{3}{4} = \frac{15}{2}$, and $E[Z] = E[c 1_{[0,1]}] = c$.

Proposition 6.3. *If $X = \sum_{j=1}^m x_j 1_{A_j}$ and $Y = \sum_{k=1}^n y_k 1_{B_k}$ are simple, then $E[aX + bY] = aE[X] + bE[Y]$ for any $a, b \in \mathbb{R}$.*

Proof. Arguing as in Example 6.1 shows that

$$\begin{aligned} E[aX + bY] &= \sum_{j=1}^m \sum_{k=1}^n (a x_k + b y_k) P(A_j \cap B_k) \\ &= a \sum_{j=1}^m x_k \sum_{k=1}^n P(A_j \cap B_k) + b \sum_{k=1}^n y_k \sum_{j=1}^m P(A_j \cap B_k) \\ &= a \sum_{j=1}^m x_k P(A_j) + b \sum_{k=1}^n y_k P(B_k) = aE[X] + bE[Y]. \end{aligned}$$

□

An immediate corollary is that $E[\sum_{k=1}^n x_k 1_{A_k}] = \sum_{k=1}^n x_k P(A_k)$ for any collection of sets $\{A_k\}_{k=1}^n$, regardless of whether they partition the sample space, just as we had hoped for originally.

Linearity also shows that expectation/integration is monotone for simple random variables: If $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$, then $Y - X$ is a nonnegative simple random variable, so $E[Y] - E[X] = E[Y - X] \geq 0$.
It follows that if X is a simple random variable, then $E[X] = \max\big\{E[Y] : Y \text{ is simple with } Y \leq X\big\}$.

This suggests a means of extending our definition of expectation to nonnegative random variables. Namely, if $X \geq 0$, we declare that

$$E[X] = \sup\big\{E[Y] : Y \text{ is simple with } Y \leq X\big\}.$$

This is well-defined, though possibly infinite, and it does agree with our previous definitions in the event that X is simple. Moreover, monotonicity is baked right into the definition.

The following result shows that we can approximate nonnegative random variable from below by simple random variables, providing a more concrete way of understanding the supremum.

Theorem 6.4. *If X is a nonnegative random variable, then there is a sequence $\{X_n\}_{n=1}^\infty$ of simple functions with $0 \leq X_1 \leq X_2 \leq ... \leq X$ such that $X_n \to X$ pointwise, and the convergence is uniform on any set on which X is bounded.*

Proof. For $n = 1, 2, ...$ and $k = 0, 1, ..., 4^n - 1$, define

$$A_{n,k} = X^{-1}\left(\left(\frac{k}{2^n}, \frac{k+1}{2^n}\right]\right) \text{ and } A_n = X^{-1}\left((2^n, \infty]\right),$$

and set

$$X_n = \sum_{k=0}^{4^n-1} \frac{k}{2^n} 1_{A_{n,k}} + 2^n 1_{A_n}.$$

By construction, $0 \leq X - X_n \leq 2^{-n}$ on $A_n^C = \big\{\omega : X(\omega) \leq 2^n\big\} \searrow \emptyset$, and the result follows. □

In order to leverage the preceding, we need to be able to pass limits through integrals. For this, we appeal to the following *monotone convergence theorem*.

Theorem 6.5. *If $X_1 \leq X_2 \leq \cdots$ are nonnegative random variables converging pointwise to X, then* $\lim_{n\to\infty} E[X_n] = E[X]$.

Proof. X is a random variable by Theorem 4.8, and monotonicity ensures $E[X_1] \leq E[X_2] \leq \cdots \leq E[X]$, hence $\lim_{n\to\infty} E[X_n]$ exists in $\mathbb{R}^*$ and is bounded above by $E[X]$.
As such, we need only show that $\lim_{n\to\infty} E[X_n] \geq E[X]$. According to our definition of expectation of nonnegative random variables, this will follow upon demonstrating that $\lim_{n\to\infty} E[X_n] \geq E[Y]$ for any simple $Y \leq X$.
Let $Y = \sum_{k=1}^m y_k 1_{A_k}$ be bounded above by X, choose $\varepsilon > 0$, and set $A_{k,n} = \big\{\omega \in A_k : X_n(\omega) \geq y_k - \varepsilon\big\}$.
Then $E[X_n] \geq \sum_{k=1}^m (y_k - \varepsilon) P(A_{k,n})$, and since $A_{k,n} \nearrow A_k$ as $n \to \infty$, continuity from below gives

$$\lim_{n\to\infty} E[X_n] \geq \lim_{n\to\infty} \sum_{k=1}^m (y_k - \varepsilon) P(A_{k,n}) = \sum_{k=1}^m (y_k - \varepsilon) P(A_k) = E[Y] - \varepsilon$$

completing the proof. □

By construction, expectation is unchanged if we modify a random variable on an event having probability 0, so one can weaken the assumption to $X_n \nearrow X$ a.s. in the monotone convergence theorem.

Also, once we extend our definition to general random variables, we can weaken the nonnegativity condition to $E[X_1] > -\infty$ by applying Theorem 6.5 $0 \le X_n - X_1 \nearrow X - X_1$. (If $E[X_1] = \infty$, then $E[X] = \infty$.)

Similarly, replacing X_n, X with $-X_n, -X$, we see that if $E[X_1] < \infty$ and $X_n \searrow X$ a.s., then $E[X_n] \to E[X]$.

However, monotonicity of some sort is important. For instance, if $(\Omega, \mathcal{F}, P)$ is Lebesgue measure on $[0,1]$, then $X_n(\omega) = n1_{(0,\frac{1}{n})}(\omega) \to 0$ for all ω, but $E[X_n] = 1$ for all n.

As an example of the MCT in action, we derive the following 'layer cake representation' for the expectation of $\mathbb{N}_0$-valued random variables.

Proposition 6.6. *If* $\text{Range}(X) \subseteq \mathbb{N}_0$, *then* $E[X] = \sum_{k=1}^{\infty} P(X \ge k)$.

Proof. The simple random variables $X_n = \sum_{k=0}^{n} k1_{\{X=k\}}$ increase to X, so

$$E[X] = \lim_{n\to\infty} E[X_n] = \lim_{n\to\infty} \sum_{k=0}^{n} kP(X=k) = \sum_{k=1}^{\infty} kP(X=k),$$

hence

$$\sum_{k=1}^{\infty} P(X \ge k) = \sum_{k=1}^{\infty}\sum_{j=0}^{\infty} P(X = k+j) = \sum_{\ell=1}^{\infty} \ell P(X=\ell) = E[X]. \qquad \square$$

To complete our construction of the integral/expectation, suppose that X is any random variable and define $X^+ = \max\{X, 0\}$, $X^- = \max\{-X, 0\}$. (Equivalently, if $E = X^{-1}\big([0,\infty)\big)$, $X^+ = X1_E$ and $X^- = -X1_{E^C}$.) Then X^+ and X^- are nonnegative random variables and thus have well-defined expectations. If at least one of them is finite, then we define $E[X] = E[X^+] - E[X^-]$.

If $E[X^+]$ and $E[X^-]$ are finite, so are $E[X]$ and $E\,|X| = E[X^+] + E[X^-]$, and we say that X is *integrable*. If $E[X^+] = \infty$ and $E[X^-] < \infty$, then $E[X] = \infty$; and if $E[X^+] < \infty$ and $E[X^-] = \infty$, then $E[X] = -\infty$. This is all well and good in terms of defining the expected value, but we will see that integrability is often a desirable feature.

And that's how you build the integral! Start with indicators, extend linearly to simple functions, then to nonnegative functions by monotone convergence, and finally to arbitrary functions by considering positive and negative parts.

Many useful properties of expectation can be established by following this four-step procedure (or fewer depending on where you start).

Proposition 6.7. *If* $X \le Y$ *and* $E[X], E[Y]$ *exist in* $\mathbb{R}^*$, *then* $E[X] \le E[Y]$.

Proof. If $0 \le X \le Y$, then

$$E[X] = \sup\big\{E[Z] : Z \le X \text{ is simple}\big\} \le \sup\big\{E[Z] : Z \le Y \text{ is simple}\big\} = E[Y].$$

For general $X \le Y$, we must have $X^+ \le Y^+$ and $X^- \ge Y^-$, so

$$E[X] = E[X^+] - E[X^-] \le E[Y^+] - E[Y^-] = E[Y]$$

by our result for nonnegative random variables. $\square$

Similarly, we know that expectation is linear for simple random variables.

If X is nonnegative and $c \geq 0$, then for any sequence of simple $X_n \nearrow X$, we have that $cX_n \nearrow cX$, so $E[cX] = \lim_{n\to\infty} E[cX_n] = \lim_{n\to\infty} cE[X_n] = cE[X]$.

Thus if $E[X]$ exists and $c \geq 0$, then

$$E[cX] = E[(cX)^+] - E[(cX)^- = cE[X^+] - cE[X^-] = cE[X].$$

For $c < 0$, we have that $(cX)^\pm = |c|\, X^\mp$, so

$$E[cX] = E\left[|c|\, X^-\right] - E\left[|c|\, X^+\right] = |c|\left(E\left[X^-\right] - E\left[X^+\right]\right) = c\left(E\left[X^+\right] - E\left[X^-\right]\right) = cE[X]$$

in this case as well.

If X and Y are nonnegative random variables, Theorem 6.4 ensures the existence of sequences of simple $X_n \nearrow X$ and $Y_n \nearrow Y$, so Theorems 6.5 and 6.3 tell us that

$$E[X+Y] = \lim_{n\to\infty} E[X_n + Y_n] = \lim_{n\to\infty} E[X_n] + \lim_{n\to\infty} E[Y_n] = E[X] + E[Y].$$

Finally, if $Z = X + Y$ for integrable X, Y, then $Z^+ - Z^- = Z = (X^+ - X^-) + (Y^+ - Y^-)$, hence $Z^+ + X^- + Y^- = Z^- + X^+ + Y^+$. Applying our result for nonnegative random variables shows that $E[Z^+] + E[X^-] + E[Y^-] = E[Z^-] + E[X^+] + E[Y^+]$, hence

$$E[Z] = E[Z^+] - E[Z^-] = E[X^+] - E[X^-] + E[Y^+] - E[Y^-] = E[X] + E[Y].$$

(Integrability of X and Y guarantees that $X^\pm$, $Y^\pm$, and $Z^\pm$ all have finite expectations.)

There are additional cases where certain combinations of $E[X^+], E[X^-], E[Y^+], E[Y^-]$ are infinite and analogous results hold, but we'll content ourselves with the following consequence of the above computations.

Proposition 6.8. *For any integrable X, Y and any $a, b \in \mathbb{R}$, $E[aX + bY] = aE[X] + bE[Y]$.*

At this point, we recall that random variables X, Y are independent if $\{X \in A\}$ and $\{Y \in B\}$ are independent events for all $A, B \in \mathcal{B}$.

Proposition 6.9. *If X and Y are independent and nonnegative/integrable, then $E[XY] = E[X]E[Y]$.*

Proof. If $X, Y \geq 0$, let $X_n = \sum_{j=1}^{4^n} x_{n,j} 1_{A_{n,j}}$ and $Y_n = \sum_{k=1}^{4^n} y_{n,k} 1_{B_{n,k}}$ be as in Theorem 6.4—so $x_{n,j} = \frac{j}{2^n}$, $A_{n,j} = X^{-1}\left(\left(\frac{j}{2^n}, \frac{j+1}{2^n}\right]\right)$ for $j < 4^n$, $x_{n,4^n} = 2^n$, and $A_{n,4^n} = X^{-1}\left((2^n, \infty]\right)$; similarly for Y_n.

Then $\{A_{n,j}\}$ and $\{B_{n,k}\}$ are independent and $X_n Y_n = \sum_{j=1}^{4^n} \sum_{k=1}^{4^n} x_{n,j} y_{n,k} 1_{A_{n,j} \cap B_{n,k}}$ increases to XY, so

$$\begin{aligned} E[XY] &= \lim_{n\to\infty} E[X_n Y_n] = \lim_{n\to\infty} \sum_{j=1}^{4^n} \sum_{k=1}^{4^n} x_{n,j} y_{n,k} P(A_{n,j} \cap B_{n,k}) \\ &= \lim_{n\to\infty} \sum_{j=1}^{4^n} \sum_{k=1}^{4^n} x_{n,j} y_{n,k} P(A_{n,j}) P(B_{n,k}) = \lim_{n\to\infty} \sum_{j=1}^{4^n} x_{n,j} P(A_{n,j}) \cdot \sum_{k=1}^{4^n} y_{n,k} P(B_{n,k}) \\ &= \lim_{n\to\infty} E[X_n] \lim_{n\to\infty} E[Y_n] = E[XY]. \end{aligned}$$

If X and Y are independent and integrable, then $\{X^-, X^+\}$ is independent of $\{Y^-, Y^+\}$, so

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

□

This factorization formula immediately extends to any finite number of independent random variables because Corollary 5.4 ensures that if $X_1, \ldots, X_n$ are independent, then X_1 and $X_2 \cdots X_n$ are as well, hence

$$E[X_1 \cdots X_n] = E[X_1]E[X_2 \cdots X_n] = E[X_1] \cdots E[X_n]$$

by the implicit induction hypothesis.

Likewise, if $X_1, \ldots, X_n$ are independent and $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$ are measurable, then $f_1(X_1), \ldots, f_n(X_n)$ are independent, hence

$$E\Big[\prod_{k=1}^{n} f_k(X_k)\Big] = \prod_{k=1}^{n} E[f_k(X_k)].$$

Before closing out this first section on expectation, we remark that while the monotone convergence theorem is a great tool for constructing the integral, there are several other results for interchanging limits and integration that will come in handy down the line.

We begin with *Fatou's lemma*, which is interesting in its own right and will enable any easy derivation of what is perhaps the most useful of these limit theorems.

Lemma 6.10. *If $X_n \geq C$ for all $n \in \mathbb{N}$ and some $C \in \mathbb{R}$, then*

$$E[\liminf_{n\to\infty} X_n] \leq \liminf_{n\to\infty} E[X_n].$$

Proof. Let $Y_n = \inf_{m \geq n} X_m$ so that $Y = \lim_{n\to\infty} Y_n = \liminf_{n\to\infty} X_n$. Then $X_n \geq Y_n$ and $Y_n \nearrow Y \geq C$, so monotonicity and monotone convergence yield

$$\liminf_{n\to\infty} E[X_n] \geq \liminf_{n\to\infty} E[Y_n] = E[Y] = E[\liminf_{n\to\infty} X_n]. \qquad \square$$

By replacing X_n with $-X_n$ and using the fact that $\liminf_{n\to\infty}(-x_n) = -\limsup_{n\to\infty} x_n$, we see that if the X_n are uniformly bounded above, then $E[\limsup_{n\to\infty} X_n] \geq \limsup_{n\to\infty} E[X_n]$.

We are now in a position to establish the celebrated *dominated convergence theorem.*

Theorem 6.11. *If $X, X_1, X_2, \ldots$ are random variables with $X_n \to X$ a.s. and Y is an integrable random variable with $|X_n| \leq Y$ for all n, then*

$$\lim_{n\to\infty} E[X_n] = E[X].$$

Proof. By assumption, $Y + X_n$ is a nonnegative random variable, so Fatou tells us that

$$E[Y] + E[X] = E[Y + X] \leq \liminf_{n\to\infty} E[Y + X_n] = E[Y] + \liminf_{n\to\infty} E[X_n],$$

hence $E[X] \leq \liminf_{n\to\infty} E[X_n]$.

Similarly, $Y - X_n$ is a nonnegative random variable, so

$$E[Y] - E[X] = E[Y - X] \leq \liminf_{n\to\infty} E[Y - X_n] = E[Y] - \limsup_{n\to\infty} E[X_n].$$

It follows that $E[X] \geq \limsup_{n\to\infty} E[X_n] \geq \liminf_{n\to\infty} E[X_n] \geq E[X]$, so the inequalities are all equalities and the theorem has been proved. $\square$

When $Y \equiv C$ for some constant $C > 0$, the above is sometimes termed the *bounded convergence theorem.*

As an initial illustration of the utility of this result, we provide a partial converse to our factorization rule for independent random variables.

Proposition 6.12. *X and Y are independent if $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ for all bounded continuous functions f and g.*

Proof. Given any $x, y \in \mathbb{R}$, define

$$f_n(t) = \begin{cases} 1, & t \le x \\ 1 - n(t-x), & x < t \le x + \frac{1}{n}, \\ 0, & t > x + \frac{1}{n} \end{cases} \quad g_n(t) = \begin{cases} 1, & t \le y \\ 1 - n(t-y), & y < t \le y + \frac{1}{n}. \\ 0, & t > y + \frac{1}{n} \end{cases}$$

Then bounded convergence and the assumptions give

$$\begin{aligned} P(X \le x, Y \le y) &= E\left[\lim_{n\to\infty} f_n(X) g_n(Y)\right] = \lim_{n\to\infty} E\left[f_n(X)\right] E\left[g_n(Y)\right] \\ &= E\left[\lim_{n\to\infty} f_n(X)\right] \left[\lim_{n\to\infty} g_n(Y)\right] = P(X \le x) P(Y \le y). \end{aligned}$$

□

7 Further Properties

Higher Moments

A nice thing about expectation is that it describes some feature of a random variable's distribution in terms of a single real number, the *mean* $\mu = E[X]$, which can be interpreted as the distribution's 'center of mass' or the 'average value' taken by the random variable.

Of course, there are many other quantities that can capture different features of the distribution.

Definition. For $k \in \mathbb{N}$, the k^{th} *moment* of a random variable X is $m_k = E[X^k]$, provided this expectation exists. If X has finite mean $\mu = m_1$, its k^{th} *central moment* is $c_k = E[(X-\mu)^k]$.

After the mean, the most important of these is the second central moment, or *variance*, $\mathrm{Var}(X) = E[(X-\mu)^2]$. The variance measures the dispersion or spread of a distribution: If $\mathrm{Var}(X)$ is small, most of the mass of the distribution is concentrated around $E[X]$, whereas a large variance implies that there is a reasonable likelihood that the random variable takes values further from its mean.

Note that by linearity of expectation, we have the 'shortcut formula'

$$\mathrm{Var}(X) = E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - E[X]^2.$$

A similar calculation shows that if $E[X^2] < \infty$, then for any $a \in \mathbb{R}$,

$$\begin{aligned} E[(X-a)^2] &= E[\big((X-\mu) + (\mu - a)\big)^2] \\ &= E[(X-\mu)^2) + 2(\mu - a)E[X-\mu] + (\mu-a)^2 = \mathrm{Var}(X) + (\mu-a)^2, \end{aligned}$$

which is clearly minimized at $a = \mu$. This is the sense in which $\mu = E[X]$ is our best guess for X: it's the projection of X onto the space of constant functions from Ω to $\mathbb{R}$.

Often, one describes the variance of a random variable X in terms of its *standard deviation* $\sigma_X = \sqrt{\mathrm{Var}(X)}$. The two encode the same information since the variance is nonnegative, but the standard deviation is nice because it's expressed in the same 'units' as X. For instance, one easily checks that $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$ for all $a, b \in \mathbb{R}$, so $\sigma_{aX+b} = |a|\,\sigma_X$.

A related quantity captures the (linear) association between two random variables X_1 and X_2 having finite means μ_1 and μ_2, the *covariance* $\mathrm{Cov}(X_1, X_2) = E[(X_1-\mu_1)(X_2-\mu_2)]$, and we have the resemblant formula

$$\mathrm{Cov}(X_1, X_2) = E[X_1X_2 - \mu_2X_1 - \mu_1X_2 + \mu_1\mu_2] = E[X_1X_2] - E[X_1]E[X_2].$$

Whereas the variance is necessarily nonnegative, the covariance is not so constrained. It's positive if, on average, X_1 and X_2 are simultaneously larger than their means or simultaneously smaller, and it's negative if X_1 tends to exceed its mean when X_2 subceeds its mean and vice versa.

The covariance is small when the association between the two is small—knowing that $X_1 > \mu_1$ does not provide much information about whether $X_2 > \mu_2$.

Since the size of the covariance may be largely due to the individual dispersion of the two variates, one often normalizes it to obtain the *correlation*

$$\rho(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}}.$$

The Cauchy-Schwarz inequality ensures that $\rho(X_1, X_2)$ always has absolute value at most 1, and one can verify that $|\rho(X_1, X_2)| = 1$ precisely when $X_2 = aX_1 + b$, with the sign determined by that of a.

In the case where X_1 and X_2 are independent, we have $E[X_1X_2] = E[X_1]E[X_2]$, so $\mathrm{Cov}(X_1, X_2) = 0$. However, two random variables may have covariance 0 (in which case we say they are *uncorrelated*), but still be dependent.

Example 7.1. Suppose that X has first and third moment equal to 0. In this case, the third moment is the third central moment, or *skewness* of the distribution, which measures the asymmetry of μ_X. (For symmetric distributions, $W =_d -W$, we have $E[W^{2k+1}] = E[(-W)^{2k+1}] = -E[W^{2k+1}]$, so the odd order moments vanish when they exist.) In this case X and $Y = X^2$ are clearly dependent, but $\mathrm{Cov}(X, Y) = E[X^3] - E[X]E[X^2] = 0$. This is because the covariance/correlation is really measuring the linear relationship between the variates.

Given random variables $X_1, \ldots, X_n$ defined on a common probability space, we have by linearity of the expected value that $E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n]$.
If $E[X_k] = \mu_k$ is finite for each k, then

$$\begin{aligned}
\mathrm{Var}(X_1 + \cdots + X_n) &= E\left[\left(\sum_{k=1}^n X_k - \sum_{k=1}^n \mu_k\right)^2\right] = E\left[\left(\sum_{k=1}^n (X_k - \mu_k)\right)^2\right] \\
&= E\left[\sum_{k=1}^n (X_k - \mu_k)^2 + \sum_{i \neq j} (X_i - \mu_i)(X_j - \mu_j)\right] \\
&= \sum_{k=1}^n E[(X_k - \mu_k)^2] + 2 \sum_{1 \leq i < j \leq n} E[(X_i - \mu_i)(X_j - \mu_j)] \\
&= \sum_{k=1}^n \mathrm{Var}(X_k) + 2 \sum_{i<j} \mathrm{Cov}(X_i, X_j).
\end{aligned}$$

Thus if $X_1, \ldots, X_n$ are pairwise independent/uncorrelated, then the variance of their sum is the sum of their variances.

Inequalities

In many applications, it is sufficient to obtain general bounds on the moments of a random variable, and there are a number of nice results for doing so. To state our first, we need some facts about convex functions.

Definition. A function $\varphi : \mathbb{R} \to \mathbb{R}$ is convex if for all $x_1, x_2 \in \mathbb{R}$, $\lambda \in [0, 1]$, we have

$$\varphi\left(\lambda x_1 + (1 - \lambda)x_2\right) \leq \lambda\varphi(x_1) + (1 - \lambda)\varphi(x_2).$$

Thus φ is convex if the secant line connecting $\big(x_1, \varphi(x_1)\big)$ to $\big(x_2, \varphi(x_2)\big)$ always lies above the graph of φ.

Example 7.2. If φ is twice-differentiable, convexity is equivalent to $\varphi''(x) > 0$ for all x. To see this, note that a second order Taylor expansion around any $x_0 \in \mathbb{R}$ gives $\varphi(x) = \varphi(x_0) + \varphi'(x_0)(x - x_0) + \frac{1}{2}\varphi''(x^*)(x - x_0)^2$ for some x^* between x and x_0; in particular, $\varphi(x) \geq \varphi(x_0) + \varphi'(x_0)(x - x_0)$.
For any $x_1, x_2 \in \mathbb{R}$, taking $x_0 = \lambda x_1 + (1 - \lambda)x_2$ yields

$$\begin{aligned}
\varphi(x_1) &\geq \varphi(x_0) + \varphi'(x_0)(x_1 - x_0), \\
\varphi(x_2) &\geq \varphi(x_0) + \varphi'(x_0)(x_2 - x_0),
\end{aligned}$$

hence

$$\begin{aligned}\lambda\varphi(x_1)+(1-\lambda)\varphi(x_2) &\geq \lambda\varphi(x_0)+\lambda\varphi'(x_0)(x_1-x_0)+(1-\lambda)\varphi(x_0)+(1-\lambda)\varphi'(x_0)(x_2-x_0)\\ &= \varphi(x_0)+\varphi'(x_0)[\lambda x_1+(1-\lambda)x_2-x_0]=\varphi(\lambda x_1+(1-\lambda)x_2).\end{aligned}$$

Thus, for example, $f(x)=x^2$ and $g(x)=e^x$ are both convex functions.

In the case of a twice-differentiable convex function φ, the inequality $\varphi(x)\geq\varphi(x_0)+\varphi'(x_0)(x-x_0)$ says that the graph of φ always lies strictly above the tangent line to that graph at any point.
It turns out (and is not especially hard to prove), that an analogous statement is true for any convex function.

Fact 7.3. *If φ is convex, then for any $c\in\mathbb{R}$, there is a linear function $\ell(x)$ which satisfies $\ell(c)=\varphi(c)$ and $\ell(x)\leq\varphi(x)$ for all $x\in\mathbb{R}$.*

It is now a short step to derive the remarkably utile *Jensen's inequality*.

Theorem 7.4. *If φ is a convex function and X is a random variable, then*

$$\varphi\left(E[X]\right)\leq E\left[\varphi(X)\right]$$

whenever the expectations exist.

Proof. Fact 7.3 gives the existence of a function $\ell(x)=ax+b$ which satisfies $\ell\left(E[X]\right)=\varphi\left(E[X]\right)$ and $\ell(x)\leq\varphi(x)$ for all $x\in\mathbb{R}$.
By monotonicity and linearity, we have

$$E[\varphi(X)]\geq E[\ell(X)]=E[aX+b]=aE[X]+b=\ell\left(E[X]\right)=\varphi\left(E[X]\right). \qquad \square$$

The *triangle inequality*, $E\left|X\right|\geq\left|E[X]\right|$, is an important special case of Theorem 7.4.

To state our next result, *Hölder's inequality*, we define the *L^p-norm* of a random variable by $\left\|X\right\|_p=E\left[\left|X\right|^p\right]^{\frac{1}{p}}$ for $p\in[1,\infty)$ and $\left\|X\right\|_\infty=\inf\{M:P\left(\left|X\right|>M\right)=0\}$.

Theorem 7.5. *If $p,q\in[1,\infty]$ with $\frac{1}{p}+\frac{1}{q}=1$ (where $\frac{1}{\infty}:=0$), then*

$$\left\|XY\right\|_1\leq\left\|X\right\|_p\left\|Y\right\|_q.$$

Proof. We first note that the result holds trivially if the right-hand side is infinity, and if $\left\|X\right\|_p=0$ or $\left\|Y\right\|_q=0$, then $\left|XY\right|=0$ a.s.
Accordingly, we may assume that $0<\left\|X\right\|_p,\left\|Y\right\|_q<\infty$. In fact, since constants factor out of L^p-norms, it suffices to establish the result when $\left\|X\right\|_p=\left\|Y\right\|_q=1$.
Also, the case $p=\infty$, $q=1$ (and symmetrically) is immediate since $\left|X\right|\leq\left\|X\right\|_\infty$ a.s., thus

$$E\left|XY\right|\leq E\left[\left\|X\right\|_\infty\left|Y\right|\right]=\left\|X\right\|_\infty E\left|Y\right|=\left\|X\right\|_\infty\left\|Y\right\|_1.$$

Accordingly, we will assume henceforth that $p,q\in(1,\infty)$.
Now fix $y\geq0$, and define the function $\varphi:[0,\infty)\to\mathbb{R}$ by $\varphi(x)=\frac{x^p}{p}+\frac{y^q}{q}-xy$.
Since $\varphi'(x)=x^{p-1}-y$ and $\varphi''(x)=(p-1)x^{p-2}>0$ for $x>0$, φ attains its minimum at $x_0=y^{\frac{1}{p-1}}$.

Thus, as the conjugacy of p and q implies that $q = \left(1 - \frac{1}{p}\right)^{-1} = \frac{p}{p-1} = \frac{1}{p-1} + 1$, we have that

$$\varphi(x) \geq \varphi(x_0) = \frac{x_0^p}{p} + \frac{y^q}{q} - x_0 y = \frac{y^{\frac{p}{p-1}}}{p} + \frac{y^q}{q} - y^{\frac{1}{p-1}+1} = y^q \left(\frac{1}{p} + \frac{1}{q}\right) - y^q = 0$$

for all $x > 0$. It follows that $\frac{x^p}{p} + \frac{y^q}{q} \geq xy$ for every $x, y \geq 0$.
In particular, taking $x = |X|$, $y = |Y|$, and integrating, we have

$$\begin{aligned} E|XY| &= \int |X|\,|Y|\,dP \leq \frac{1}{p}\int |X|^p\,dP + \frac{1}{q}\int |Y|^q\,dP \\ &= \frac{\|X\|_p^p}{p} + \frac{\|Y\|_q^q}{q} = \frac{1}{p} + \frac{1}{q} = 1 = \|X\|_p \|Y\|_q . \end{aligned}$$

□

Some useful corollaries of Hölder's inequality are:

Corollary 7.6 (Cauchy-Schwarz)**.** $E|XY| \leq \sqrt{E[X^2]\,E[Y^2]}$.

Alternate Proof. For all $t \in \mathbb{R}$,

$$0 \leq E\left[(|X| + t|Y|)^2\right] = E\left[X^2\right] + 2tE|XY| + t^2 E\left[Y^2\right] = q(t),$$

thus the quadratic $q(t)$ has at most one real root, so its discriminant satisfies

$$(2E|XY|)^2 - 4E\left[X^2\right]E\left[Y^2\right] \leq 0.$$

□

Corollary 7.7. *For any random variable X and any $1 \leq r < s \leq \infty$, $\|X\|_r \leq \|X\|_s$.*

Proof. For $s = \infty$, we have $|X|^r \leq \|X\|_\infty^r$ a.s., hence

$$\|X\|_r^r = \int |X|^r\,dP \leq \int \|X\|_\infty^r\,dP = \|X\|_\infty^r .$$

For $s < \infty$, apply Holder's inequality to X^r and 1 with $p = \frac{s}{r}$, $q = \frac{s}{s-r}$ to get

$$\|X\|_r^r = E\left[|X|^r\right] \leq \|X^r\|_p \|1\|_q = \left(\int |X^r|^{\frac{s}{r}}\,dP\right)^{\frac{r}{s}} = \|X\|_s^r .$$

□

Our last big result is *Chebychev's inequality*, which is both simple and surprisingly practical.

Theorem 7.8. *For any nonnegative random variable X and any $a > 0$,*

$$P(X \geq a) \leq \frac{E[X]}{a}.$$

Proof. Let $A = \{\omega : X(\omega) \geq a\}$. Then

$$aP(X \geq a) = a\int 1_A dP \leq \int X 1_A dP \leq \int X dP = E[X].$$

□

Corollary 7.9. *For any $(S, \mathcal{G})$-valued random variable X and any measurable function $\varphi : S \to [0, \infty)$,*

$$P(\varphi(X) \geq a) \leq \frac{E[\varphi(X)]}{a}.$$

Some important cases of Corollary 7.9 for real-valued X are

- $\varphi(x) = |x|$: to control the probability that an integrable random variable is large.
- $\varphi(x) = (x - E[X])^2$: to control the probability that a random variable with finite variance is far from its mean.
- $\varphi(x) = e^{tx}$: to establish exponential decay for random variables with moment generating functions (*concentration inequalities*).

Change of Variables

The main practical use of the distribution of a random variable is that it enables us to transfer questions about X from the abstract space $(\Omega, \mathcal{F})$ to the more familiar $(\mathbb{R}, \mathcal{B})$. This is especially helpful for computing expectations, connecting our general theory back to the setting of undergraduate probability.

For discrete random variables, we easily recover the definition of expectation as a sum against the pmf:
If $\text{Range}(X) = \{x_1, \ldots, x_n\}$, then $X = \sum_{k=1}^n x_k 1_{A_k}$ with $A_k = X^{-1}(x_k)$, so our definition for simple functions gives $E[X] = \sum_{k=1}^n x_k P(A_k) = \sum_{k=1}^n x_k P(X = x_k)$;
If X is nonnegative with range $\{x_1, x_2, \ldots\}$, then $X_n = \sum_{k=1}^n x_k 1_{A_k} \nearrow X$, so monotone convergence gives $E[X] = \lim_{n\to\infty} E[X_n] = \lim_{n\to\infty} \sum_{k=1}^n x_k P(X = x_k) = \sum_{k=1}^\infty x_k P(X = x_k)$;
If X has countable range and is integrable, then X^+ and X^- are nonnegative with countable ranges and finite expectation, so

$$E[X] = E[X^+] - E[X^-] = \sum_{j:x_j \geq 0} x_j P(X = x_j) - \sum_{k:x_k < 0} |x_k| \, P(X = x_k) = \sum_{k=1}^\infty x_k P(X = x_k)$$

where absolute convergence justifies rearranging the summands.

In order to deal with more complicated cases, we record the following change of variables theorem, which allows us to compute expectations by integrating functions of a random variable against its distribution.

Theorem 7.10. *Let X be a random variable taking values in the measurable space $(S, \mathcal{G})$, and let $\mu = P \circ X^{-1}$ be the pushforward measure on $(S, \mathcal{G})$.*
If f is a measurable function from $(S, \mathcal{G})$ to $(\mathbb{R}, \mathcal{B})$ such that $f \geq 0$ or $E\,|f(X)| < \infty$, then

$$E[f(X)] = \int_S f(s)\, d\mu(s).$$

Proof. We will proceed by verifying the result in increasingly general cases paralleling the construction of the integral.
To begin with, let $B \in \mathcal{G}$ and $f = 1_B$. Then

$$E[f(X)] = E[1_B(X)] = P(X \in B) = \mu(B) = \int_S 1_B(s)\, d\mu(s) = \int_S f(s)\, d\mu(s).$$

Now suppose that $f = \sum_{i=1}^n a_i 1_{B_i}$ is a simple function. Then by linearity and the previous case,

$$E[f(X)] = \sum_{i=1}^n a_i E[1_{B_i}(X)] = \sum_{i=1}^n a_i \int_S 1_{B_i}(s)\, d\mu(s) = \int_S f(s)\, d\mu(s).$$

If $f \geq 0$, then Theorem 6.4 gives a sequence of simple functions $\phi_n \nearrow f$, so the previous case and two applications of the monotone convergence give

$$E[f(X)] = \lim_{n\to\infty} E[\phi_n(X)] = \lim_{n\to\infty} \int_S \phi_n(s)\, d\mu(s) = \int_S f(s)\, d\mu(s).$$

Finally, suppose that $E\,|f(X)| < \infty$, and set $f^+(x) = \max\{f(x), 0\}$, $f^-(x) = \max\{-f(x), 0\}$. Then $f^+, f^- \geq 0$, $f = f^+ - f^-$, and $E[f(X)^+], E[f(X)^-] \leq E\,|f(X)| < \infty$, so it follows from the previous result and linearity that

$$E[f(X)] = E[f^+(X)] - E[f^-(X)] = \int_S f^+(s)\, d\mu(s) - \int_S f^-(s)\, d\mu(s) = \int_S f(s)\, d\mu(s). \qquad \square$$

If X has density f, then $\mu(A) = \int_A d\mu(x) = \int_A f(x)\, dx$ (where we use the notation $\int_A d\nu = \int 1_A d\nu$), so Theorem 7.10 shows that for any measurable $g : \mathbb{R} \to \mathbb{R}$ with $g \geq 0$ a.s. or $\int |g|\, d\mu < \infty$,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx.$$

An immediate corollary of Theorem 7.10 is that the expectation of a random variable is determined by its distribution.

Corollary 7.11. *Let X and Y be random variables (possibly defined on different probability spaces). Then $X =_d Y$ if and only if $E[f(X)] = E[f(Y)]$ for all measurable f for which the expectations exist in $\mathbb{R}$.*

Proof. Suppose $X =_d Y$ and let μ denote their common distribution. Then for all such f, we have $E[f(X)] = \int_{-\infty}^{\infty} f(x)\, d\mu(x) = E[f(Y)]$.
Conversely, if $E[f(X)] = E[f(Y)]$ for all f, taking $f = 1_A$ for any $A \in \mathcal{B}$ gives $P(X \in A) = E[1_A(X)] = E[1_A(Y)] = P(Y \in A)$, hence $X =_d Y$. $\square$

Product Measure

If $X_1, \ldots, X_n$ are random variables defined on a common probability space $(\Omega, \mathcal{F}, P)$, we define their *joint distribution* as the probability $\mu_{X_1,\ldots,X_n}(A) := P\big((X_1, \ldots, X_n) \in A\big)$ on $(\mathbb{R}^n, \mathcal{B}^n)$.
Since $\mathcal{B}^n$ is generated by $\big\{(-\infty, x_1] \times \cdots \times (-\infty, x_n] : x_1, \ldots, x_n \in \mathbb{R}\big\}$ and P is countably additive, Theorem 3.14 ensures that there is a unique measure μ on $\mathcal{B}^n$ (or its completion) satisfying

$$\mu\big((-\infty, x_1] \times \cdots \times (-\infty, x_n]\big) = P(X_1 \leq x_1, \ldots, X_n \leq x_n) =: F_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$$

for all $x_1, \ldots, x_n \in \mathbb{R}$.
That is, the joint distribution $\mu_{X_1,\ldots,X_n}$ is uniquely determined by the joint cdf $F_{X_1,\ldots,X_n}$.

Note that the *marginal distribution* of X_i can be recovered as

$$\mu_{X_i}(A) = P(X_1 \in \mathbb{R}, \ldots, X_i \in A, \ldots, X_n \in \mathbb{R}) = \mu_{X_1,\ldots,X_n}(\mathbb{R} \times \cdots \times A \times \cdots \times \mathbb{R}),$$

and independence of $X_1, \ldots, X_n$ is equivalent to the statement that the joint distribution factors as

$$\mu_{X_1,\ldots,X_n}(A_1 \times \cdots \times A_n) = \mu_{X_1}(A_1) \cdots \mu_{X_n}(A_n)$$

for all $A_1, \ldots, A_n \in \mathcal{B}$.

Our next result will allow us to construct a probability space supporting finitely many independent random variables with specified marginals.

Proposition 7.12. *Given probability spaces $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$, there exists a unique measure $\mu_1 \times \mu_2$ on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ that satisfies $(\mu_1 \times \mu_2)(A \times E) = \mu_1(A)\mu_2(E)$ for all $A \in \mathcal{F}_1$, $E \in \mathcal{F}_2$.*

Proof. Since σ-algebras are semialgebras, Example 3.3 shows that $\mathcal{R} = \{A \times E : A \in \mathcal{F}_1, E \in \mathcal{F}_2\}$ is a semialgebra, and Proposition 11.27 shows it generates $\mathcal{F}_1 \otimes \mathcal{F}_2$.

Define $\nu : \mathcal{R} \to [0, \infty)$ by $\nu(A \times E) = \mu_1(A)\mu_2(E)$.

In light of Theorem 3.14, the result will follow if we can prove that for any countable disjoint union of sets $\{A_i \times E_i\}_{i \in I}$ in $\mathcal{R}$ such that $A \times E = \bigcup_{i \in I}(A_i \times E_i) \in \mathcal{R}$, we have $\nu(A \times E) = \sum_{i \in I} \nu(A_i \times E_i)$.

(Clearly $\nu(\emptyset) = 0$ and $\nu(\Omega_1 \times \Omega_2) = 1$, and countable additivity implies both finite additivity and countable subadditivity, so ν will be a protomeasure if this condition obtains.)

To see that this is so, observe that for all $(x, y) \in \Omega_1 \times \Omega_2$,

$$1_A(x)1_E(y) = 1_{A \times E}\big((x, y)\big) = \sum_{i \in I} 1_{A_i \times E_i}\big((x, y)\big) = \sum_{i \in I} 1_{A_i}(x)1_{E_i}(y).$$

Consequently,

$$\begin{aligned}
\mu_1(A)1_E(y) &= \int_{\Omega_1} 1_A(x)1_E(y)\, d\mu_1(x) = \int_{\Omega_1} \sum_{i \in I} 1_{A_i}(x)1_{E_i}(y)\, d\mu_1(x) \\
&= \sum_{i \in I} \int_{\Omega_1} 1_{A_i}(x)1_{E_i}(y)\, d\mu_1(x) = \sum_{i \in I} \left(\int_{\Omega_1} 1_{A_i}(x)\, d\mu_1(x) \right) 1_{E_i}(y) \\
&= \sum_{i \in I} \mu_1(A_i)1_{E_i}(y).
\end{aligned}$$

The interchange of summation and integration in the second line is justified by the monotone convergence theorem.

Integrating against μ_2 then gives

$$\begin{aligned}
\nu(A \times E) = \mu_1(A)\mu_2(E) &= \int_{\Omega_2} \mu_1(A)1_E(y)\, d\mu_2(y) = \int_{\Omega_2} \sum_{i \in I} \mu_1(A_i)1_{E_i}(y)\, d\mu_2(y) \\
&= \sum_{i \in I} \mu_1(A_i) \int_{\Omega_2} 1_{E_i}(y)\, d\mu_2(y) = \sum_{i \in I} \mu_1(A_i)\mu_2(E_i) = \sum_{i \in I} \nu(A_i \times E_i).
\end{aligned}$$

□

By induction, given $(\Omega_1, \mathcal{F}_1, P_1), \ldots, (\Omega_n, \mathcal{F}_n, P_n)$, we see that there is a unique probability $P_1 \times \cdots \times P_n$ on $\left(\prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{F}_i \right)$ satisfying $P(A_1 \times \cdots \times A_n) = \prod_{i=1}^n P(A_i)$. The projections $X_i\big((\omega_1, \ldots, \omega_n)\big) = \omega_i$ are clearly independent with $\mu_{X_i} = \mu_i$.

The main tool for integrating against product measure is the *Fubini-Tonelli theorem*, the proof of which is relegated to the appendix. (Tonelli's theorem treats nonnegative functions and Fubini extends to those which are integrable. In practice, one applies Tonelli to $|f|$ in order to then deduce Fubini.)

Fact 7.13. *Suppose that $(R, \mathcal{F}, \mu)$ and $(S, \mathcal{G}, \nu)$ are probability spaces. If a measurable function $f : R \times S \to \mathbb{R}$ is nonnegative or integrable, then*

$$\int_{R \times S} f\, d(\mu \times \nu) = \int_S \left(\int_R f(x, y)\, d\mu(x) \right) d\nu(y) = \int_R \left(\int_S f(x, y)\, d\nu(y) \right) d\mu(x).$$

In order to build an infinite sequence of independent random variables with given distribution functions, we need to perform the above construction on the infinite product space

$$\mathbb{R}^{\mathbb{N}} = \big\{(\omega_1, \omega_2, ...) : \omega_i \in \mathbb{R}\big\} = \big\{\text{functions } \omega : \mathbb{N} \to \mathbb{R}\big\}.$$

The product σ-algebra $\mathcal{B}^{\mathbb{N}}$ is generated by *cylinder sets* of the form

$$\big\{\omega \in \mathbb{R}^{\mathbb{N}} : \omega_i \in (a_i, b_i] \text{ for } i = 1, ..., n\big\},$$

and the random variables are the projections $X_i(\omega) = \omega_i$.

(In the definition of cylinders, we take $-\infty \leq a_i \leq b_i \leq \infty$ with the interpretation that $(a_i, \infty] = (a_i, \infty)$. $a_j = b_j$ for any j gives the empty set.)

Clearly, the desired measure should satisfy

$$P\left(\{\omega \in \mathbb{R}^{\mathbb{N}} : \omega_i \in (a_i, b_i] \text{ for } i = 1, ..., n\}\right) = \prod_{i=1}^{n} \left(F_i(b_i) - F_i(a_i)\right)$$

on the cylinders.

To see that we can uniquely extend this to all of $\mathcal{B}^{\mathbb{N}}$, we appeal to the following *Kolmogorov extension theorem*, a proof of which can be found in the appendix.

Fact 7.14. *Suppose that we are given a sequence of probability measures μ_n on $(\mathbb{R}^n, \mathcal{B}^n)$ which are* consistent *in the sense that*

$$\mu_{n+1}\left((a_1, b_1] \times \cdots \times (a_n, b_n] \times \mathbb{R}\right) = \mu_n\left((a_1, b_1] \times \cdots \times (a_n, b_n]\right).$$

Then there is a unique probability measure P on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}})$ with

$$P\left(\{\omega \in \mathbb{R}^{\mathbb{N}} : \omega_i \in (a_i, b_i], i = 1, ..., n\}\right) = \mu_n\left((a_1, b_1] \times \cdots \times (a_n, b_n]\right).$$

In particular, given distribution functions $F_1, F_2, ...$, if we define the μ_n's by the condition

$$\mu_n\left((a_1, b_1] \times \cdots \times (a_n, b_n]\right) = \prod_{i=1}^{n} \left(F_i(b_i) - F_i(a_i)\right),$$

then the projections $X_n(\omega) = \omega_n$ are independent with $P(X_n \leq x) = F_n(x)$.

8 Convergence in Probability and the WLLN

Large number laws give conditions for the arithmetic average of repeated observations to converge in certain senses. Among other things, they justify and formalize our intuitive notions of probability as representing some kind of measure of long-term relative frequency.

Convergence in Probability and L^p

Definition. A sequence of random variables $X_1, X_2, ...$ is said to *converge to X in probability* if for every $\varepsilon > 0$, $\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$. In this case, we write $X_n \to_p X$.

Note that if $X_n \to_p X$, then $\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1$ for all $\varepsilon > 0$, while $X_n \to X$ a.s. implies that $P(\lim_{n\to\infty} |X_n - X| < \varepsilon) = 1$ for all $\varepsilon > 0$. The following proposition and example demonstrate the importance of the placement of the limit in the two definitions.

Proposition 8.1. *If $X_n \to X$ a.s., then $X_n \to_p X$.*

Proof. Let $\varepsilon > 0$ be given and define

$$A_n = \bigcup_{m\geq n} \{\, |X_m - X| > \varepsilon\}, \qquad A = \bigcap_{n=1}^{\infty} A_n,$$

$$E = \{\omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\}.$$

Since $A_1 \supseteq A_2 \supseteq ...$, continuity from above implies that $P(A) = \lim_{n\to\infty} P(A_n)$.
Now if $\omega \in A$, then for every $n \in \mathbb{N}$, there is an $m \geq n$ with $|X_m(\omega) - X(\omega)| > \varepsilon$, so $\lim_{n\to\infty} X_n(\omega) \neq X(\omega)$, and thus $A \subseteq E$.
Because we also have the inclusion $\{|X_n - X| > \varepsilon\} \subseteq A_n$, monotonicity gives

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) \leq \lim_{n\to\infty} P(A_n) = P(A) \leq P(E) = 0$$

where the final equality is the definition of almost sure convergence. □

Example 8.2 (Scanning Interval)**.** On the interval $[0, 1)$ with Lebesgue measure, define

$$X_1 = 1_{[0,1)}, X_2 = 1_{[0,\frac{1}{2})}, X_3 = 1_{[\frac{1}{2},1)}, ..., X_{2^n+k} = 1_{[\frac{k}{2^n},\frac{k+1}{2^n})}, ...$$

It is straightforward that $X_n \to_p 0$—for any $\varepsilon > 0$, $m \geq 2^n$ implies $P(|X_m - 0| > \varepsilon) \leq \frac{1}{2^n}$—but $\lim_{n\to\infty} X_n(\omega)$ does not exist for any ω since there are infinitely many values of n with $X_n(\omega) = 1$ and infinitely many values with $X_n(\omega) = 0$.

The preceding shows that convergence in probability is weaker than almost sure convergence. Another mode of convergence that shows up often in probability is with respect to the L^p norm.

Definition. For $p \in (0, \infty]$, a sequence of random variables $X_1, X_2, ...$ is said to *converge to X in L^p* if $\lim_{n\to\infty} \|X_n - X\|_p = 0$. (For $p \in (0, \infty)$, this is equivalent to $E[|X_n - X|^p] \to 0$.)

Proposition 8.3. *For any $1 \leq r < s \leq \infty$, if $X_n \to X$ in L^s, then $X_n \to X$ in L^r.*

Proof. If $X_n \to X$ in L^s, then Corollary 7.7 implies $\|X_n - X\|_r \leq \|X_n - X\|_s \to 0$. □

Proposition 8.4. *If $X_n \to X$ in L^p for $p > 0$, then $X_n \to_p X$.*

Proof. By the previous proposition, we can assume $p < \infty$. For any $\varepsilon > 0$, Chebychev's inequality gives

$$P(|X_n - X| > \varepsilon) = P(|X_n - X|^p > \varepsilon^p) \leq \varepsilon^{-p} E[|X_n - X|^p] \to 0.$$ □

Example 8.5. On $[0,1]$ with Lebesgue measure, define a sequence of random variables by $X_n = n^{\frac{1}{p}} 1_{(0,n^{-1}]}$, $p \in (0,\infty)$. Then $X_n \to 0$ a.s. (and thus in probability) since for all $\omega \in (0,1]$, $X_n(\omega) = 0$ whenever $n > \omega^{-1}$. However, $E[|X_n - 0|^p] = 1$ for all n, so $X_n \not\to 0$ in L^p. Additionally, $\|X_n - 0\|_\infty = n^{\frac{1}{p}}$ diverges.

Proposition 8.4 and Example 8.5 show that L^p convergence is stronger than convergence in probability.

Example 8.5 also shows that almost sure convergence need not imply convergence in L^p (unless one makes additional assumptions such as boundedness or uniform integrability).

Conversely, Example 8.2 shows that L^p convergence does not imply almost sure convergence for $p < \infty$; since $|X| \leq \|X\|_\infty$ a.s., L^∞ convergence implies a.s. (uniform) convergence.

As one can pass limits through continuous functions, it's immediate that $X_n \to X$ a.s. implies $f(X_n) \to f(X)$ a.s. for any continuous $f : \mathbb{R} \to \mathbb{R}$. We will see later that convergence in probability is also preserved by continuous functions. However, L^p convergence need not be. For example, on $[0,1]$ with Lebesgue measure, $X_n = n^{\frac{1}{2}} 1_{(0,n^{-p})}$ converges to 0 in L^p for $p < \infty$, but if $f(x) = x^2$, $\|f(X_n) - f(0)\|_p = 1$ for all n.

Weak Laws of Large Numbers

Theorem 8.6. *Let $X_1, X_2, \ldots$ be uncorrelated random variables with common mean $E[X_i] = \mu$ and uniformly bounded variance $\mathrm{Var}(X_i) \leq C < \infty$, and set $S_n = X_1 + \ldots + X_n$. Then $\frac{1}{n} S_n \to \mu$ in L^2 and in probability.*

Proof. Since $E\left[\frac{1}{n} S_n\right] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$, we see that

$$E\left[\left(\frac{1}{n} S_n - \mu\right)^2\right] = \mathrm{Var}\left(\frac{1}{n} S_n\right) = \frac{1}{n^2} \sum_{i=1}^n \mathrm{Var}(X_i) \leq \frac{nC}{n^2} \to 0$$

as $n \to \infty$, hence $\frac{1}{n} S_n \to \mu$ in L^2. By Proposition 8.4, $\frac{1}{n} S_n \to_p \mu$ as well. □

Specializing to the case where the X_i's are *independent and identically distributed* (or *i.i.d.*), we have the oft-quoted weak law

Corollary 8.7. *If $X_1, X_2, \ldots$ are i.i.d. with mean μ and variance $\sigma^2 < \infty$, then $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ converges in probability to μ.*

The statistical interpretation of Corollary 8.7 is that under mild conditions, if the sample size is sufficiently large, then the sample mean will be close to the population mean with high probability.

The following amusing result can be interpreted as saying that a high-dimensional cube is almost a sphere.

Example 8.8. Let $X_1, X_2, ...$ be independent and uniformly distributed on $[-1,1]$. Then $X_1^2, X_2^2, ...$ are i.i.d. with $E[X_i^2] = \int_{-1}^{1} \frac{x^2}{2} dx = \frac{1}{3}$ and $\text{Var}(X_i^2) \leq E[X_i^4] \leq 1$, so Corollary 8.7 shows that $\frac{1}{n}\sum_{i=1}^n X_i^2$ converges to $\frac{1}{3}$ in probability.

Now given $\varepsilon \in (0,1)$, write $A_{n,\varepsilon} = \left\{x \in \mathbb{R}^n : (1-\varepsilon)\sqrt{\frac{n}{3}} \leq \|x\| \leq (1+\varepsilon)\sqrt{\frac{n}{3}}\right\}$ where $\|x\| = (x_1^2 + ... + x_n^2)^{\frac{1}{2}}$ is the usual Euclidean distance, and let m denote Lebesgue measure. We have

$$\begin{aligned}
\frac{m\left(A_{n,\varepsilon} \cap [-1,1]^n\right)}{2^n} &= P\left((X_1, ..., X_n) \in A_{n,\varepsilon}\right) = P\left((1-\varepsilon)\sqrt{\frac{n}{3}} \leq \sqrt{\sum_{i=1}^n X_i^2} \leq (1+\varepsilon)\sqrt{\frac{n}{3}}\right) \\
&= P\left(\frac{1}{3}(1 - 2\varepsilon + \varepsilon^2) \leq \frac{1}{n}\sum_{i=1}^n X_i^2 \leq \frac{1}{3}(1 + 2\varepsilon + \varepsilon^2)\right) \\
&\geq P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{1}{3}\right| \leq \frac{2\varepsilon - \varepsilon^2}{3}\right),
\end{aligned}$$

so that $\frac{m(A_{n,\varepsilon} \cap [-1,1]^n)}{2^n} \to 1$ as $n \to \infty$. In words, most of the volume of the cube $[-1,1]^n$ comes from $A_{n,\varepsilon}$, which is almost the boundary of the ball centered at the origin with radius $\sqrt{\frac{n}{3}}$.

Our next set of examples concern the limiting behavior of row sums of *triangular arrays*, for which we appeal to the following easy generalization of Theorem 8.6.

Theorem 8.9. *Given a triangular array of integrable random variables, $\{X_{n,k}\}_{n\in\mathbb{N}, k\in[n]}$, let $S_n = \sum_{k=1}^n X_{n,k}$ denote the n^{th} row sum, and write $\mu_n = E[S_n]$, $\sigma_n^2 = \text{Var}(S_n)$. If $\{b_n\}_{n=1}^\infty$ satisfies $\lim_{n\to\infty} \frac{\sigma_n^2}{b_n^2} = 0$, then*

$$\frac{S_n - \mu_n}{b_n} \to_p 0.$$

Proof. By assumption, $E\left[\left(\frac{S_n - \mu_n}{b_n}\right)^2\right] = \frac{\text{Var}(S_n)}{b_n^2} \to 0$ as $n \to \infty$, so the result follows since L^2 convergence implies convergence in probability. □

Example 8.10 (Coupon Collector's Problem)**.** Suppose that there are n distinct types of coupons and each time one obtains a coupon it is, independent of prior selections, equally likely to be any one of the types. We are interested in the number of draws needed to obtain a complete set. To this end, let $T_{n,k}$ denote the number of draws needed to collect k distinct types for $k = 1, ..., n$ and note that $T_{n,1} = 1$. Set $X_{n,1} = 1$ and $X_{n,k} = T_{n,k} - T_{n,k-1}$ for $k = 2, ..., n$ so that $X_{n,k}$ is the number of trials needed to obtain a type different from the first $k-1$. The number of draws needed to obtain a complete set is given by

$$T_n := T_{n,n} = 1 + \sum_{k=2}^n (T_{n,k} - T_{n,k-1}) = 1 + \sum_{k=2}^n X_{n,k}.$$

By construction, $X_{n,2}, ..., X_{n,n}$ are independent with $P(X_{n,k} = m) = \left(\frac{n-k+1}{n}\right)\left(\frac{k-1}{n}\right)^{m-1}$ for $m \in \mathbb{N}$.

Now a random variable X with $P(X = m) = p(1-p)^{m-1}$ is said to be *geometric with success probability p*. A little calculus gives

$$E[X] = \sum_{m=1}^\infty mp(1-p)^{m-1} = p\sum_{m=1}^\infty -\frac{d}{dp}(1-p)^m = -p\frac{d}{dp}\frac{1-p}{p} = \frac{1}{p}$$

and

$$E[X^2] = \sum_{m=1}^{\infty} m^2 p(1-p)^{m-1} = \sum_{m=1}^{\infty} [m(m-1)+m]p(1-p)^{m-1}$$
$$= p(1-p)\sum_{m=1}^{\infty} m(m-1)(1-p)^{m-2} + \sum_{m=1}^{\infty} mp(1-p)^{m-1}$$
$$= p(1-p)\sum_{m=2}^{\infty} \frac{d^2}{dp^2}(1-p)^m + E[X] = p(1-p)\frac{d^2}{dp^2}\frac{(1-p)^2}{p} + \frac{1}{p}$$
$$= \frac{2(1-p)}{p^2} + \frac{1}{p} = \frac{2-p}{p^2},$$

hence

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{1-p}{p^2} \le \frac{1}{p^2}.$$

It follows that

$$E[T_n] = 1 + \sum_{k=2}^{n} E[X_{n,k}] = 1 + \sum_{k=2}^{n} \frac{n}{n-k+1} = 1 + n\sum_{j=1}^{n-1} \frac{1}{j} = n\sum_{j=1}^{n} \frac{1}{j}$$

and

$$\mathrm{Var}(T_n) = \sum_{k=2}^{n} \mathrm{Var}(X_{n,k}) \le \sum_{k=2}^{n} \left(\frac{n}{n-k+1}\right)^2 = n^2 \sum_{j=1}^{n-1} \frac{1}{j^2} \le n^2 \sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{\pi^2 n^2}{6}.$$

Taking $b_n = n\log(n)$ we have $\frac{\mathrm{Var}(T_n)}{b_n^2} \le \frac{\pi^2}{6\log(n)^2} \to 0$, so Theorem 8.9 implies $\frac{T_n - n\sum_{k=1}^{n} k^{-1}}{n\log(n)} \to_p 0$.
Using the inequality

$$\log(n) \le \sum_{k=1}^{n} \frac{1}{k} \le \log(n) + 1$$

(which can be seen by bounding $\log(n) = \int_1^n \frac{dx}{x}$ with the upper Riemann sum $\sum_{k=1}^{n-1} \frac{1}{k} \le \sum_{k=1}^{n} \frac{1}{k}$ and the lower Riemann sum $\sum_{k=2}^{n} \frac{1}{k} = \sum_{k=1}^{n} \frac{1}{k} - 1$), we conclude that $\dfrac{T_n}{n\log(n)} \to_p 1$.

Example 8.11 (Occupancy Problem)**.** Suppose that we drop r_n balls at random into n bins where $\frac{r_n}{n} \to c$. Letting $X_{n,k} = 1\{\text{bin } k \text{ is empty}\}$, the number of empty bins is $X_n = \sum_{k=1}^{n} X_{n,k}$.
It is clear that

$$E[X_n] = \sum_{k=1}^{n} E[X_{n,k}] = \sum_{k=1}^{n} P(\text{bin } k \text{ is empty}) = n\left(\frac{n-1}{n}\right)^{r_n}$$

and

$$E[X_n^2] = E\left[\sum_{k=1}^{n} X_{n,k}^2 + 2\sum_{i<j} X_{n,i}X_{n,j}\right] = \sum_{k=1}^{n} E[X_{n,k}] + 2\sum_{i<j} E[X_{n,i}X_{n,j}]$$
$$= \sum_{k=1}^{n} P(\text{bin } k \text{ is empty}) + 2\sum_{i<j} P(\text{bins } i \text{ and } j \text{ are empty})$$
$$= n\left(\frac{n-1}{n}\right)^{r_n} + 2\binom{n}{2}\left(\frac{n-2}{n}\right)^{r_n} = n\left(1-\frac{1}{n}\right)^{r_n} + n(n-1)\left(1-\frac{2}{n}\right)^{r_n},$$

so

$$\mathrm{Var}(X_n) = E[X_n^2] - E[X_n]^2 = n\left(1-\frac{1}{n}\right)^{r_n} + n(n-1)\left(1-\frac{2}{n}\right)^{r_n} - n^2\left(1-\frac{1}{n}\right)^{2r_n}.$$

Now L'Hospital's rule gives $\lim_{n\to\infty} \frac{\log\left(\frac{n-1}{n}\right)}{n^{-1}} = \lim_{n\to\infty} \frac{n^{-2}}{-n^{-2}} \cdot \frac{n}{n-1} = -1$, so, since $\frac{r_n}{n} \to c$, we have that $\log\left[\left(\frac{n-1}{n}\right)^{r_n}\right] = \frac{r_n}{n} \cdot \frac{\log\left(\frac{n-1}{n}\right)}{n^{-1}} \to -c$ and thus $\left(\frac{n-1}{n}\right)^{r_n} \to e^{-c}$ as $n \to \infty$.

Similarly, $\left(1-\frac{2}{n}\right)^{r_n}$, $\left(1-\frac{1}{n}\right)^{2r_n} \to e^{-2c}$.

Consequently, $\frac{E[X_n]}{n} = \left(\frac{n-1}{n}\right)^{r_n} \to e^{-c}$ and

$$\frac{\mathrm{Var}(X_n)}{n^2} = \frac{\left(1-\frac{1}{n}\right)^{r_n}}{n} + \frac{n(n-1)}{n^2}\left(1-\frac{2}{n}\right)^{r_n} - \left(1-\frac{1}{n}\right)^{2r_n} \to 0 + 1 \cdot e^{-2c} - e^{-2c} = 0$$

as $n \to \infty$, so taking $b_n = n$ in Theorem 8.9 shows that the proportion of empty bins, $\frac{X_n}{n}$, converges to e^{-c} in probability.

We conclude this section with a weakening of the moment assumptions in the classical WLLN. The trick is to use truncation in order to consider cases where we have control over the size and the probability, respectively.

Theorem 8.12. *Suppose that $X_1, X_2, \ldots$ are i.i.d. with $E|X_1| < \infty$. Let $S_n = \sum_{i=1}^n X_i$ and $\mu = E[X_1]$. Then $\frac{1}{n}S_n \to \mu$ in probability.*

Proof. In what follows, the arithmetic average of the first n terms of a sequence of random variables $Y_1, Y_2, \ldots$ will be denoted by $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$.

We first note that, by replacing X_i with $X_i' = X_i - \mu$ if necessary, we may suppose without loss of generality that $E[X_i] = 0$.

Thus we need to show that for given $\varepsilon, \delta > 0$, there is an $N \in \mathbb{N}$ such that $P\left(\left|\overline{X}_n\right| > \varepsilon\right) < \delta$ whenever $n \geq N$.

To this end, we pick $C < \infty$ large enough that $E\left[|X_1| 1\left\{|X_1| > C\right\}\right] < \eta$ for some η to be determined.

(This is possible since $|X_1| 1\left\{|X_1| \leq n\right\} \leq |X_1|$ and $E|X_1| < \infty$, so $\lim_{n\to\infty} E\left[|X_1| 1\left\{|X_1| \leq n\right\}\right] = E|X_1|$ by the dominated convergence theorem, hence $E\left[|X_1| 1\left\{|X_1| > n\right\}\right] = E|X_1| - E\left[|X_1| 1\left\{|X_1| \leq n\right\}\right] \to 0$.)

Now define

$$W_i = X_i 1\left\{|X_i| \leq C\right\} - E\left[X_i 1\left\{|X_i| \leq C\right\}\right]$$
$$Z_i = X_i 1\left\{|X_i| > C\right\} - E\left[X_i 1\left\{|X_i| > C\right\}\right].$$

By assumption, we have that

$$E|Z_i| \leq 2E\left[|X_1| 1\left\{|X_1| > C\right\}\right] < 2\eta,$$

and thus, for every $n \in \mathbb{N}$,

$$E\left|\overline{Z}_n\right| = E\left|\frac{1}{n}\sum_{i=1}^n Z_i\right| \leq \frac{1}{n}\sum_{i=1}^n E|Z_i| \leq 2\eta.$$

Also, the W_i's are i.i.d. with mean zero and satisfy $|W_i| \leq 2C$ by construction, so

$$E\left[\overline{W}_n^2\right] = \frac{1}{n^2}\left(\sum_{i=1}^n E[W_i^2] + \sum_{i\neq j} E[W_i W_j]\right) = \frac{E[W_1^2]}{n} \leq \frac{4C^2}{n},$$

and thus, by Jensen's inequality,

$$E\left[\left|\overline{W}_n\right|\right]^2 \leq E\left[\overline{W}_n^2\right] \leq \frac{4C^2}{n}.$$

Consequently, if $n \geq N := \left\lceil \frac{4C^2}{\eta^2} \right\rceil$, then $E\left|\overline{W}_n\right| \leq \eta$.

Finally, Chebychev's inequality and the fact that

$$\left|\overline{X}_n\right| = \left|\overline{W}_n + \overline{Z}_n\right| \leq \left|\overline{W}_n\right| + \left|\overline{Z}_n\right|$$

imply that for $n \geq N$,

$$P\left(\left|\overline{X}_n\right| > \varepsilon\right) \leq P\left(\left|\overline{W}_n\right| + \left|\overline{Z}_n\right| > \varepsilon\right) \leq \frac{E\left|\overline{W}_n\right| + E\left|\overline{Z}_n\right|}{\varepsilon} < \frac{3\eta}{\varepsilon}.$$

Taking $\eta = \frac{\varepsilon\delta}{3}$ completes the proof. □

Finally, we mention that the weak law can be slightly upgraded to accommodate certain situations involving infinite means. A proof is given in the appendix, along with a fun example.

Fact 8.13. *Let $X_1, X_2, ...$ be i.i.d. with*

$$xP(|X_1| > x) \to 0 \text{ as } x \to \infty.$$

Set $S_n = X_1 + ... + X_n$ and $\mu_n = E\left[X_1 1\left\{|X_1| \leq n\right\}\right]$. Then $\frac{1}{n}S_n - \mu_n \to 0$ in probability.

Given a sequence of events $A_1, A_2, ... \in \mathcal{F}$, we define

$$\limsup_n A_n := \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \{\omega : \omega \text{ is in infinitely many } A_n\},$$

which is often abbreviated as $\{A_n \text{ i.o.}\}$ where "i.o." stands for "infinitely often."

The nomenclature derives from the straightforward identity $\limsup_{n\to\infty} 1_{A_n} = 1_{\limsup_n A_n}$.

One can likewise define $\liminf_n A_n := \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m$, the set of outcomes belonging to all but finitely many of the A_n, but little is gained by doing so since $\liminf_n A_n = \left(\limsup_n A_n^C\right)^C$.

To illustrate the utility of this notion, observe that $X_n \to X$ a.s. if and only if $P(|X_n - X| > \varepsilon \text{ i.o.}) = 0$ for every $\varepsilon > 0$.

Lemma 9.1 (Borel-Cantelli I)**.** *If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n \text{ i.o.}) = 0$.*

Proof. Let $N = \sum_{n=1}^{\infty} 1_{A_n}$ denote the number of events that occur. Tonelli's theorem (or MCT) gives

$$E[N] = \sum_{n=1}^{\infty} E[1_{A_n}] = \sum_{n=1}^{\infty} P(A_n) < \infty,$$

so it must be the case that $N < \infty$ a.s.

(Alternatively, writing $B_n = \bigcup_{m=n}^{\infty} A_m$, we see that $B_n \searrow \bigcap_{n=1}^{\infty} B_n = \limsup_n A_n$, so continuity from above implies $P(A_n \text{ i.o.}) = \lim_{n\to\infty} P(B_n) \leq \lim_{n\to\infty} \sum_{m=n}^{\infty} P(A_n) = 0$.) □

A nice application of the first Borel-Cantelli lemma is

Theorem 9.2. *$X_n \to_p X$ if and only if every subsequence $\{X_{n_m}\}_{m=1}^{\infty}$ has a further subsequence $\{X_{n_{m(k)}}\}_{k=1}^{\infty}$ such that $X_{n_{m(k)}} \to X$ a.s. as $k \to \infty$.*

Proof. Suppose that $X_n \to_p X$ and let $\{X_{n_m}\}_{m=1}^{\infty}$ be any subsequence. Then $X_{n_m} \to_p X$, so for every $k \in \mathbb{N}$, $P\left(|X_{n_m} - X| > \frac{1}{k}\right) \to 0$ as $m \to \infty$. It follows that we can choose a further subsequence $\{X_{n_{m(k)}}\}_{k=1}^{\infty}$ such that $P\left(\left|X_{n_{m(k)}} - X\right| > \frac{1}{k}\right) \leq 2^{-k}$ for all $k \in \mathbb{N}$. Since

$$\sum_{k=1}^{\infty} P\left(\left|X_{n_{m(k)}} - X\right| > \frac{1}{k}\right) \leq 1 < \infty,$$

the first Borel-Cantelli lemma shows that $P\left(\left|X_{n_{m(k)}} - X\right| > \frac{1}{k} \text{ i.o.}\right) = 0$.
Because $\left\{\left|X_{n_{m(k)}} - X\right| > \varepsilon \text{ i.o.}\right\} \subseteq \left\{\left|X_{n_{m(k)}} - X\right| > \frac{1}{k} \text{ i.o.}\right\}$ for every $\varepsilon > 0$, we see that $X_{n_{m(k)}} \to X$ a.s.

To prove the converse, we appeal to the following lemma.

Lemma 9.3. *Let $\{y_n\}_{n=1}^{\infty}$ be a sequence of elements in a topological space. If every subsequence $\{y_{n_m}\}_{m=1}^{\infty}$ has a further subsequence $\{y_{n_{m(k)}}\}_{k=1}^{\infty}$ that converges to y, then $y_n \to y$.*

Proof. If $y_n \nrightarrow y$, then there is an open set $U \ni y$ such that for every $N \in \mathbb{N}$, there is an $n \geq N$ with $y_n \notin U$, hence there is a subsequence $\{y_{n_m}\}_{m=1}^{\infty}$ with $y_{n_m} \notin U$ for all m. By construction, no subsequence of $\{y_{n_m}\}_{m=1}^{\infty}$ can converge to y, and the result follows by contraposition. □

Now if every subsequence of $\{X_n\}_{n=1}^{\infty}$ has a further subsequence that converges to X almost surely, then applying Lemma 9.3 to the sequence $y_n = P(|X_n - X| > \varepsilon)$ for an arbitrary $\varepsilon > 0$ shows that $X_n \to_p X$. $\square$

Since there are sequences that converge in probability but not almost surely (e.g. Example 8.2), it follows from Theorem 9.2 and Lemma 9.3 that a.s. convergence does not come from a topology.

Theorem 9.2 can sometimes be used to upgrade results depending on almost sure convergence.

For example, one can show that the assumptions in Fatou's lemma and the dominated convergence theorem can be weakened to require only convergence in probability.

To get a feel for how this works, we prove

Theorem 9.4. *If f is continuous and $X_n \to_p X$, then $f(X_n) \to_p f(X)$. If f is also bounded, then $E[f(X_n)] \to E[f(X)]$.*

Proof. If $\{X_{n_m}\}$ is a subsequence, then Theorem 9.2 guarantees the existence of a further subsequence $\left\{X_{n_{m(k)}}\right\}$ that converges to X a.s. Since limits commute with continuous functions, this means that $f\left(X_{n_{m(k)}}\right) \to f(X)$ a.s. The other direction of Theorem 9.2 now implies that $f(X_n) \to_p f(X)$.

If f is bounded as well, then the dominated convergence theorem yields $E\left[f\left(X_{n_{m(k)}}\right)\right] \to E[f(X)]$.

Applying Lemma 9.3 to the sequence $y_n = E[f(X_n)]$ establishes convergence in expectation. $\square$

We will now use the first Borel-Cantelli lemma to prove a weak form of the *Strong Law of Large Numbers*.

Theorem 9.5. *Let $X_1, X_2, \ldots$ be i.i.d. with $E[X_1] = \mu$ and $E\left[X_1^4\right] < \infty$. If $S_n = X_1 + \ldots + X_n$, then $\frac{1}{n}S_n \to \mu$ almost surely.*

Proof. By taking $X_i' = X_i - \mu$, we can suppose without loss of generality that $\mu = 0$. Now

$$E\left[S_n^4\right] = E\left[\left(\sum_{i=1}^n X_i\right)\left(\sum_{j=1}^n X_j\right)\left(\sum_{k=1}^n X_k\right)\left(\sum_{l=1}^n X_l\right)\right] = E\left[\sum_{1\le i,j,k,l\le n} X_i X_j X_k X_l\right].$$

By independence, terms of the form $E\left[X_i^3 X_j\right]$, $E\left[X_i^2 X_j X_k\right]$ and $E\left[X_i X_j X_k X_l\right]$ are all zero (since the expectation of the product is the product of the expectations).

The only non-vanishing terms are thus of the form $E\left[X_i^4\right]$ and $E\left[X_i^2 X_j^2\right]$, of which there are n of the former and $3n(n-1)$ of the latter (determined by the $\binom{n}{2}$ ways of picking the indices and the $\binom{4}{2}$ ways of picking which two of the four sums gave rise to the smaller index).

Because $E\left[X_i^2 X_j^2\right] = E\left[X_i^2\right]^2 \le E\left[X_i^4\right]$, we have

$$E\left[S_n^4\right] \le nE\left[X_1^4\right] + 3n(n-1)E\left[X_1^2\right]^2 \le Cn^2$$

where $C = 3E\left[X_1^4\right] < \infty$ by assumption.

It follows from Chebychev's inequality that

$$P\left(\frac{1}{n}|S_n| > \varepsilon\right) = P\left(|S_n|^4 > (n\varepsilon)^4\right) \le \frac{C}{n^2\varepsilon^4},$$

hence

$$\sum_{n=1}^{\infty} P\left(\frac{1}{n}|S_n| > \varepsilon\right) \le C\varepsilon^{-4}\sum_{n=1}^{\infty}\frac{1}{n^2} < \infty.$$

Therefore, $P\left(\frac{1}{n}|S_n| > \varepsilon \text{ i.o.}\right) = 0$ by Borel-Cantelli, so, since $\varepsilon > 0$ was arbitrary, $\frac{1}{n}S_n \to 0$ a.s. $\square$

A much more involved argument based on truncation and passing to a suitable subsequence shows that the assumptions in the strong law can be weakened to $E|X_1| < \infty$.

The converse of the Borel-Cantelli lemma is false without additional assumptions.

Example 9.6. Let $\Omega = [0,1]$, $\mathcal{F} =$ Borel sets, $P =$ Lebesgue measure, and define $A_n = (0, \frac{1}{n})$.
Then $\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$ and $\limsup_{n\to\infty} A_n = \emptyset$.

Lemma 9.7 (Borel-Cantelli II)**.** *If the events $A_1, A_2, \ldots$ are independent, then $\sum_{n=1}^{\infty} P(A_n) = \infty$ implies $P(A_n \text{ i.o.}) = 1$.*

Proof. For each $n \in \mathbb{N}$, the sequence $B_{n,1}, B_{n,2}, \ldots$ defined by $B_{n,k} = \bigcap_{m=n}^{n+k} A_m^C$ decreases to $B_n := \bigcap_{m=n}^{\infty} A_m^C$.
Also, since the A_m's (and thus their complements) are independent, we have

$$\begin{aligned} P(B_{n,k}) &= P\left(\bigcap_{m=n}^{n+k} A_m^C\right) = \prod_{m=n}^{n+k} P\left(A_m^C\right) \\ &= \prod_{m=n}^{n+k} (1 - P(A_m)) \leq \prod_{m=n}^{n+k} e^{-P(A_m)} = e^{-\sum_{m=n}^{n+k} P(A_m)} \end{aligned}$$

where the inequality is due to the Taylor series bound $e^{-x} \geq 1 - x$ for $x \in [0,1]$.
Because $\sum_{m=n}^{\infty} P(A_m) = \infty$ by assumption, it follows from continuity from above that

$$P(B_n) = \lim_{k\to\infty} P(B_{n,k}) \leq \lim_{k\to\infty} e^{-\sum_{m=n}^{n+k} P(A_m)} = 0,$$

hence $P\left(\bigcup_{m=n}^{\infty} A_m\right) = P\left(B_n^C\right) = 1$ for all $n \in \mathbb{N}$.
Since $\bigcup_{m=n}^{\infty} A_m \searrow \limsup_{n\to\infty} A_n = \{A_n \text{ i.o.}\}$, another application of continuity from above gives

$$P\left(A_n \text{ i.o.}\right) = \lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} A_m\right) = 1. \qquad \square$$

Taken together, the Borel-Cantelli lemmas show that if $A_1, A_2, \ldots$ is a sequence of independent events, then the event $\{A_n \text{ i.o.}\}$ occurs either with probability 0 or probability 1.
Thus if $A_1, A_2, \ldots$ are independent, then $P(A_n \text{ i.o.}) > 0$ implies $P(A_n \text{ i.o.}) = 1$.
This is an example of a 0-1 law; we'll see another presently.

It follows from the second Borel-Cantelli lemma that infinitely many independent trials of a random experiment will almost surely result in infinitely many realizations of any event having positive probability.
For example, an infinite string with characters chosen independently and uniformly from a finite alphabet (produced by the proverbial monkey at a typewriter, say) will almost surely contain infinitely many instances of any finite string (like the complete works of Shakespeare in chronological order).

Our next example is a typical application where the two Borel-Cantelli lemmas are used together to obtain results on limits of (suitably scaled) sequences of i.i.d. random variables.

Example 9.8. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. exponential random variables with rate 1 (so that $X_i \geq 0$ with $P(X_i \leq x) = 1 - e^{-x}$). We will show that $\limsup_{n\to\infty} \frac{X_n}{\log(n)} = 1$ a.s.

First observe that

$$P\left(\frac{X_n}{\log(n)} \geq 1\right) = P\left(X_n \geq \log(n)\right) = P\left(X_n > \log(n)\right) = e^{-\log(n)} = \frac{1}{n},$$

so

$$\sum_{n=1}^{\infty} P\left(\frac{X_n}{\log(n)} \geq 1\right) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

Since the X_n's are independent, the second Borel-Cantelli lemma implies that $P\left(\frac{X_n}{\log(n)} \geq 1 \text{ i.o.}\right) = 1$, and we conclude that $\limsup_{n\to\infty} \frac{X_n}{\log(n)} \geq 1$ almost surely.

On the other hand, for any $\varepsilon > 0$,

$$P\left(\frac{X_n}{\log(n)} \geq 1+\varepsilon\right) = P\left(X_n > (1+\varepsilon)\log(n)\right) = \frac{1}{n^{1+\varepsilon}},$$

which is summable, so it follows from the first Borel-Cantelli lemma that $P\left(\frac{X_n}{\log(n)} \geq 1+\varepsilon \text{ i.o.}\right) = 0$.

Since $\varepsilon > 0$ was arbitrary, this means that $\limsup_{n\to\infty} \frac{X_n}{\log(n)} \leq 1$ almost surely, and the claim is proved.

We conclude this section with another famous 0-1 law.

Definition. Given a sequence of random variables $X_1, X_2, ...$, the *tail σ-field* is $\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, ...)$.

Theorem 9.9 (Kolmogorov)**.** *If $X_1, X_2, ...$ are independent and $A \in \mathcal{T}$, then $P(A) \in \{0, 1\}$.*

Proof. We will show that A is independent of itself so that $P(A)^2 = P(A)P(A) = P(A \cap A) = P(A)$.

To do so, we first note that $B \in \sigma(X_1, ..., X_k)$ and $C \in \sigma(X_{k+1}, X_{k+2}, ...)$ are independent.

This follows from Lemma 5.3 if $C \in \sigma(X_{k+1}, ..., X_{k+j})$. Since $\sigma(X_1, ..., X_k)$ and $\bigcup_{j=1}^{\infty} \sigma(X_{k+1}, ..., X_{k+j})$ are π-systems, Theorem 5.1 shows this is true in general.

Next, we observe that $E \in \sigma(X_1, X_2, ...)$ and $F \in \mathcal{T}$ are independent.

If $E \in \sigma(X_1, ..., X_k)$, then this follows from the previous observation since $F \in \mathcal{T} \subseteq \sigma(X_{k+1}, X_{k+2}, ...)$.

Since $\bigcup_{k=1}^{\infty} \sigma(X_1, ..., X_k)$ and $\mathcal{T}$ are π-systems, Theorem 5.1 shows it is true in general.

Because $\mathcal{T} \subseteq \sigma(X_1, X_2, ...)$, the last observation shows that $A \in \mathcal{T}$ is independent of itself. □

Example 9.10. If $B_1, B_2, ... \in \mathcal{B}$, then $\{X_n \in B_n \text{ i.o.}\} \in \mathcal{T}$. Taking $X_n = 1_{A_n}$, $B_n = \{1\}$, we see that $\{X_n \in B_n \text{ i.o.}\} = \{A_n \text{ i.o.}\}$, so if $A_1, A_2, ...$ are independent, then $P\left(A_n \text{ i.o.}\right) \in \{0, 1\}$. Of course, this also follows from the Borel-Cantelli lemmas.

Example 9.11. Let $S_n = X_1 + ... + X_n$. Then

- $\{\lim_{n\to\infty} S_n \text{ exists}\} \in \mathcal{T}$ since convergence of series only depends on their tails.
- $A = \{\limsup_{n\to\infty} S_n > 0\} \notin \mathcal{T}$ in general since the initial terms can effect the sign of the sum.
- If $c_n \to \infty$, then $\left\{\limsup_{n\to\infty} \frac{1}{c_n} S_n > x\right\} \in \mathcal{T}$ for all $x \in \mathbb{R}$ since the contribution from any finite number of terms of S_n will be killed by c_n.

Whereas large number laws treat the first order asymptotics of sums of random variables, central limit theorems describe their fluctuations about these limits. There are a number of CLTs applying to different objects and operating under various assumptions, but the prototypical version says that if $X_1, X_2, \ldots$ are i.i.d. with mean μ and variance $\sigma^2 \in (0, \infty)$, then $S_n = \sum_{i=1}^n X_i$ satisfies

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \Rightarrow Z \sim \mathcal{N}(0,1).$$

The double arrow here denotes *weak convergence* (or convergence in distribution/law), and we say that a sequence $\{X_n\}$ with distribution functions $\{F_n\}$ converges weakly to a random variable X with distribution function F if $\lim_{n\to\infty} F_n(x) = F(x)$ for all x at which F is continuous. [There are a number of other equivalent characterizations, and one can show that convergence in probability (and thus a.s. or in L^p) implies weak convergence, but not conversely in general.]

The classical CLT is typically established by considering *characteristic functions*, defined by $\varphi_X(t) = E[e^{itX}]$. These necessarily exist for all $t \in \mathbb{R}$ and satisfy $\varphi_{a_1X_1+\ldots+a_nX_n}(t) = \prod_{i=1}^n \varphi_{X_i}(a_i t)$ if $X_1, \ldots, X_n$ are independent. Moreover, one can show that if φ_X is continuous at 0, then $X_n \Rightarrow X$ iff $\varphi_{X_n}(t) \to \varphi_X(t)$ pointwise. If X has mean 0 and variance $\sigma^2 < \infty$, a second order Maclaurin expansion yields $\varphi_X(t) = 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2)$, so the characteristic function of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ is given by $\varphi_n(t) = \left(1 - \frac{t^2}{2} + o(t^2/n)\right)^n \to e^{-\frac{t^2}{2}}$, the ch.f. of a standard normal.

It takes a fair amount of work to make all of this rigorous, and though one sees some nice stuff along the way, we will opt here to pursue an alternative approach developed by Charles Stein in the 1960s and 1970s. In addition to requiring less technical machinery, this method is highly adaptable to other situations involving distributional convergence and approximation. Moreover, it does not require as rigid independence assumptions as the characteristic function route and also yields convergence rates; while it's all well and good to know that sample means are normal in the limit, in practice one would like to know how good the normal approximation is for finite n.

Broadly, *Stein's method* refers to a framework based on solutions of certain differential or difference equations for bounding the distance between the distribution of a random variable X and that of a random variable W having some specified target distribution.

The metrics for which this approach is applicable are of the form

$$d_{\mathcal{H}}(\mathscr{L}(X), \mathscr{L}(W)) = \sup_{h \in \mathcal{H}} |E[h(X)] - E[h(W)]|$$

for some suitable class of functions $\mathcal{H}$, and include the Kolmogorov, Wasserstein, and total variation distances as special cases. These arise by taking $\mathcal{H}$ to be the set of indicators of right-closed rays, 1-Lipschitz functions, and indicators of Borel sets, respectively. Convergence in each of these three metrics is strictly stronger than weak convergence (which can be metrized by taking $\mathcal{H}$ to be the set of 1-Lipschitz functions with sup norm at most 1).

The basic idea is to find an operator $\mathcal{A}$ such that $E[(\mathcal{A}f)(X)] = 0$ for all f belonging to some sufficiently large class of functions $\mathcal{F}$ if and only if $\mathscr{L}(X) = \mathscr{L}(W)$.
For example, we will see that $W \sim \mathcal{N}(0,1)$ if and only if $E\left[f'(W) - Wf(W)\right] = 0$ for all Lipschitz f.

If one can then show that for any $h \in \mathcal{H}$, the equation

$$(\mathcal{A}f)(x) = h(x) - E[h(W)]$$

has solution $f_h \in \mathcal{F}$, then upon taking expectations, absolute values, and suprema, they find that

$$d_{\mathcal{H}}(\mathscr{L}(X), \mathscr{L}(W)) = \sup_{h \in \mathcal{H}} |E[h(X)] - E[h(W)]| = \sup_{h \in \mathcal{H}} |E[(\mathcal{A}f_h)(X)]| .$$

Remarkably, it is often easier to work with the right-hand side of this equation and the techniques for analyzing distances between probability distributions in this manner are collectively known as Stein's method.

We begin by establishing a characterizing operator for the standard normal.

Lemma 10.1. *Define the operator $\mathcal{A}$ by*

$$(\mathcal{A}f)(x) = f'(x) - xf(x).$$

If $Z \sim \mathcal{N}(0,1)$, then $E[(\mathcal{A}f)(Z)] = 0$ for all absolutely continuous f with $E|f'(Z)| < \infty$.

Proof. Let f be as in the statement of the lemma. Then Fubini's theorem gives

$$\begin{aligned}
E[f'(Z)] &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f'(x) e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} f'(x) e^{-\frac{x^2}{2}} dx + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} f'(x) e^{-\frac{x^2}{2}} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} f'(x) \left(- \int_{-\infty}^{x} y e^{-\frac{y^2}{2}} dy \right) dx + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} f'(x) \left(\int_{x}^{\infty} y e^{-\frac{y^2}{2}} dy \right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} y e^{-\frac{y^2}{2}} \left(- \int_{y}^{0} f'(x) dx \right) dy + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} y e^{-\frac{y^2}{2}} \left(\int_{0}^{y} f'(x) dx \right) dy \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} y e^{-\frac{y^2}{2}} \left(f(y) - f(0) \right) dy + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} y e^{-\frac{y^2}{2}} \left(f(y) - f(0) \right) dy \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y f(y) e^{-\frac{y^2}{2}} dy - f(0) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y e^{-\frac{y^2}{2}} dy \\
&= E[Zf(Z)] - f(0)E[Z] = E[Zf(Z)].
\end{aligned}$$

□

If $\|f'\|_\infty < \infty$, then $E|f'(Z)| < \infty$, and the condition $E[(\mathcal{A}f)(W)] = 0$ whenever $\|f'\|_\infty < \infty$ turns out to be sufficient for $W \sim \mathcal{N}(0,1)$. For this class of functions, Lemma 10.1 is just integration by parts.

Lemma 10.2. *If Φ is the distribution function for the standard normal, then the unique bounded solution to the differential equation*

$$f'(w) - wf(w) = 1_{(-\infty, x]}(w) - \Phi(x)$$

is given by

$$f_x(w) = \begin{cases} \sqrt{2\pi} e^{\frac{w^2}{2}} \left(1 - \Phi(x)\right) \Phi(w), & w \le x \\ \sqrt{2\pi} e^{\frac{w^2}{2}} \Phi(x)(1 - \Phi(w)), & w > x \end{cases} .$$

Moreover, f_x is absolutely continuous with $\|f_x\|_\infty \le \sqrt{\frac{\pi}{2}}$ and $\|f_x'\|_\infty \le 2$.

Proof. Multiplying both sides of the equation $f'(t) - tf(t) = 1_{(-\infty,x]}(t) - \Phi(x)$ by the integrating factor $e^{-\frac{t^2}{2}}$ shows that a bounded solution f_x must satisfy

$$\frac{d}{dt}\left(e^{-\frac{t^2}{2}} f_x(t) \right) = e^{-\frac{t^2}{2}} \left[f_x'(t) - t f_x(t) \right] = e^{-\frac{t^2}{2}} \left[1_{(-\infty,x]}(t) - \Phi(x) \right] ,$$

and integration gives

$$\begin{aligned} f_x(w) &= e^{\frac{w^2}{2}} \int_{-\infty}^{w} e^{-\frac{t^2}{2}} \left(1_{(-\infty,x]}(t) - \Phi(x)\right) dt \\ &= -e^{\frac{w^2}{2}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} \left(1_{(-\infty,x]}(t) - \Phi(x)\right) dt. \end{aligned}$$

When $w \leq x$, we have

$$\begin{aligned} f_x(w) &= e^{\frac{w^2}{2}} \int_{-\infty}^{w} e^{-\frac{t^2}{2}} \left(1_{(-\infty,x]}(t) - \Phi(x)\right) dt = e^{\frac{w^2}{2}} \int_{-\infty}^{w} e^{-\frac{t^2}{2}} \left(1 - \Phi(x)\right) dt \\ &= \sqrt{2\pi} e^{\frac{w^2}{2}} \left(1 - \Phi(x)\right) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{w} e^{-\frac{t^2}{2}} dt = \sqrt{2\pi} e^{\frac{w^2}{2}} \left(1 - \Phi(x)\right) \Phi(w), \end{aligned}$$

and when $w > x$, we have

$$\begin{aligned} f_x(w) &= -e^{\frac{w^2}{2}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} \left(1_{(-\infty,x]}(t) - \Phi(x)\right) dt = -e^{\frac{w^2}{2}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} \left(0 - \Phi(x)\right) dt \\ &= \sqrt{2\pi} e^{\frac{w^2}{2}} \Phi(x) \frac{1}{\sqrt{2\pi}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} dt = \sqrt{2\pi} e^{\frac{w^2}{2}} \Phi(x)(1 - \Phi(w)). \end{aligned}$$

To check boundedness, we first observe that for any $z \geq 0$,

$$\begin{aligned} 1 - \Phi(z) &= \frac{1}{\sqrt{2\pi}} \int_{z}^{\infty} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{(s+z)^2}{2}} ds \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \int_{0}^{\infty} e^{-\frac{s^2}{2}} e^{-sz} ds \leq e^{-\frac{z^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{s^2}{2}} ds = \frac{1}{2} e^{-\frac{z^2}{2}}, \end{aligned}$$

and, by symmetry, for any $z \leq 0$,

$$\Phi(z) = 1 - \Phi\left(|z|\right) \leq \frac{1}{2} e^{-\frac{z^2}{2}}.$$

Since f_x is nonnegative and $f_x(w) = f_{-x}(-w)$, it suffices to show that f_x is bounded above for $x \geq 0$.

If $w > x \geq 0$, then

$$f_x(w) = \sqrt{2\pi} e^{\frac{w^2}{2}} \Phi(x)(1 - \Phi(w)) \leq \sqrt{2\pi} e^{\frac{w^2}{2}} \cdot 1 \cdot \frac{1}{2} e^{-\frac{w^2}{2}} = \sqrt{\frac{\pi}{2}};$$

If $0 < w \leq x$, then

$$\begin{aligned} f_x(w) &= \sqrt{2\pi} e^{\frac{w^2}{2}} \left(1 - \Phi(x)\right) \Phi(w) \\ &\leq \sqrt{2\pi} e^{\frac{w^2}{2}} \cdot \frac{1}{2} e^{-\frac{x^2}{2}} \cdot 1 \leq \sqrt{2\pi} e^{\frac{w^2}{2}} \cdot \frac{1}{2} e^{-\frac{w^2}{2}} = \sqrt{\frac{\pi}{2}}; \end{aligned}$$

and if $w \leq 0 \leq x$, then

$$f_x(w) = \sqrt{2\pi} e^{\frac{w^2}{2}} \left(1 - \Phi(x)\right) \Phi(w) \leq \sqrt{2\pi} e^{\frac{w^2}{2}} \cdot 1 \cdot \frac{1}{2} e^{-\frac{w^2}{2}} = \sqrt{\frac{\pi}{2}}.$$

That f_x is the only bounded solution follows by observing that the homogeneous equation $f'(w) - wf(w) = 0$ has solution $f_h(w) = Ce^{\frac{w^2}{2}}$ for $C \in \mathbb{R}$, so the general solution of our nonhomogeneous linear equation is given by $f_x(w) + Cf_h(w)$, which is bounded if and only if $C = 0$.

Finally, by construction, f_x is differentiable at all points $w \neq x$ with $f_x'(w) = wf_x(w) + 1_{(-\infty,x]}(w) - \Phi(x)$, so that

$$|f_x'(w)| \leq |wf_x(w)| + \left|1_{(-\infty,x]}(w) - \Phi(x)\right| \leq |wf_x(w)| + 1.$$

For $w > 0$,

$$|wf_x(w)| = \left| -we^{\frac{w^2}{2}} \int_w^\infty e^{-\frac{t^2}{2}} \left(1_{(-\infty,x]}(t) - \Phi(x)\right) dt \right| \le we^{\frac{w^2}{2}} \int_w^\infty e^{-\frac{t^2}{2}} \left|1_{(-\infty,x]}(t) - \Phi(x)\right| dt$$
$$\le we^{\frac{w^2}{2}} \int_w^\infty e^{-\frac{t^2}{2}} dt \le we^{\frac{w^2}{2}} \int_w^\infty \frac{t}{w} e^{-\frac{t^2}{2}} dt = e^{\frac{w^2}{2}} \int_w^\infty te^{-\frac{t^2}{2}} dt = e^{\frac{w^2}{2}} e^{-\frac{w^2}{2}} = 1,$$

and for $w < 0$,

$$|wf_x(w)| = |-wf_{-x}(-w)| \le 1,$$

hence $|f_x'(w)| \le |wf_x(w)| + 1 \le 2$.

Since f_x is continuous and differentiable at all points $w \ne x$ with uniformly bounded derivative, it is Lipschitz and thus absolutely continuous. □

An immediate consequence of the preceding lemmas is

Theorem 10.3. *A random variable W has the standard normal distribution if and only if*

$$E[f'(W) - Wf(W)] = 0$$

for all Lipschitz f.

Proof. Lemma 10.1 establishes necessity.

For sufficiency, observe that for any $x \in \mathbb{R}$, taking f_x as in Lemma 10.2 implies

$$|P(W \le x) - \Phi(x)| = \left|E\left[1_{(-\infty,x]}(W) - \Phi(x)\right]\right| = \left|E\left[f_x'(W) - Wf_x(W)\right]\right| = 0.$$ □

The methodology of Lemma 10.2 can be extended to cover test functions other than indicators of half-lines. Indeed, the argument given there shows that for any function $h : \mathbb{R} \to \mathbb{R}$ such that

$$Nh := E[h(Z)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty h(z) e^{-\frac{z^2}{2}} dz$$

exists in $\mathbb{R}$, the differential equation

$$f'(w) - wf(w) = h(w) - Nh$$

has solution

$$(*) \qquad f_h(w) = e^{\frac{w^2}{2}} \int_{-\infty}^w \left(h(t) - Nh\right) e^{-\frac{t^2}{2}} dt.$$

Some fairly tedious computations show that

Fact 10.4. *For any $h : \mathbb{R} \to \mathbb{R}$ such that Nh exists, let f_h be given by $(*)$.*

If h is bounded, then

$$\|f_h\|_\infty \le \sqrt{\frac{\pi}{2}} \|h - Nh\|_\infty, \quad \|f_h'\|_\infty \le 2\|h - Nh\|_\infty.$$

If h is absolutely continuous, then

$$\|f_h\|_\infty \le 2\|h'\|_\infty, \quad \|f_h'\|_\infty \le \sqrt{\frac{2}{\pi}} \|h'\|_\infty, \quad \|f_h''\|_\infty \le 2\|h'\|_\infty.$$

(That the relevant derivatives are defined almost everywhere is part of the statement of Lemma 10.4.)

We can now give bounds on the error in normal approximation for sums of i.i.d. random variables, which will imply the central limit theorem.

We will work in the Wasserstein metric

$$d_W(\mathscr{L}(W), \mathscr{L}(Z)) = \sup_{h \in \mathcal{H}_W} |E[h(W)] - E[h(Z)]|$$

where

$$\mathcal{H}_W = \{h : \mathbb{R} \to \mathbb{R} \text{ such that } |f(x) - f(y)| \leq |x - y| \text{ for all } x, y \in \mathbb{R}\}.$$

If $Z \sim \mathcal{N}(0,1)$, then the preceding analysis shows that

$$d_W(\mathscr{L}(W), \mathscr{L}(Z)) = \sup_{h \in \mathcal{H}_W} |E[f_h'(W) - W f_h(W)]|$$

where f_h is given by $(*)$.

Since Lipschitz functions are absolutely continuous, the second part of Lemma 10.4 applies with $\|h'\|_\infty = 1$.

From these observations and some elementary manipulations we have

Theorem 10.5. *Suppose $X_1, X_2, ..., X_n$ are independent random variables with $E[X_i] = 0$ and $E[X_i^2] = 1$ for all $i = 1, ..., n$. If $W = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ and $Z \sim \mathcal{N}(0,1)$, then*

$$d_W(\mathscr{L}(W), \mathscr{L}(Z)) \leq \frac{3}{n^{\frac{3}{2}}} \sum_{i=1}^n E\left[|X_i|^3\right].$$

Proof. Let f be any differentiable function with f' absolutely continuous, $\|f\|_\infty, \|f'\|_\infty, \|f''\|_\infty < \infty$.

For each $i = 1, ..., n$, set

$$W_i = \frac{1}{\sqrt{n}} \sum_{j \neq i} X_j = W - \frac{1}{\sqrt{n}} X_i.$$

Then X_i and W_i are independent, so $E[X_i f(W_i)] = E[X_i]E[f(W_i)] = 0$.

It follows that

$$E[W f(W)] = E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i f(W)\right] = E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \left(f(W) - f(W_i)\right)\right].$$

Adding and subtracting $E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i (W - W_i) f'(W_i)\right]$ yields

$$\begin{aligned} E[W f(W)] = E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \left(f(W) - f(W_i) - (W - W_i) f'(W_i)\right)\right] \\ + E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i (W - W_i) f'(W_i)\right]. \end{aligned}$$

The independence and unit variance assumptions show that

$$E[X_i(W - W_i)f'(W_i)] = E\left[\frac{1}{\sqrt{n}} X_i^2 f'(W_i)\right] = \frac{1}{\sqrt{n}} E[X_i^2] E[f'(W_i)] = \frac{1}{\sqrt{n}} E[f'(W_i)],$$

so

$$E[W f(W)] = E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \left(f(W) - f(W_i) - (W - W_i) f'(W_i)\right)\right] + E\left[\frac{1}{n} \sum_{i=1}^n f'(W_i)\right],$$

and thus

$$\begin{aligned}
&|E[f'(W) - Wf(W)]| \\
&\quad = \left| E\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right] + E\left[\frac{1}{n}\sum_{i=1}^{n} f'(W_i)\right] - E[f'(W)]\right| \\
&\quad = \left| E\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right] + E\left[\frac{1}{n}\sum_{i=1}^{n} (f'(W_i) - f'(W))\right]\right| \\
&\quad \leq \frac{1}{\sqrt{n}} E\left[\sum_{i=1}^{n} \left|X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right|\right] + \frac{1}{n} E\left[\sum_{i=1}^{n} |f'(W_i) - f'(W)|\right]
\end{aligned}$$

The Taylor expansion (with Lagrange remainder)

$$f(w) = f(z) + f'(z)(w - z) + \frac{f''(\zeta)}{2}(w - z)^2$$

for some ζ between w and z gives the bound

$$|f(w) - f(z) - (w - z)f'(z)| \leq \frac{\|f''\|_\infty}{2}(w - z)^2,$$

so

$$\begin{aligned}
\frac{1}{\sqrt{n}} E\left[\sum_{i=1}^{n} \left|X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right|\right] &\leq \frac{1}{\sqrt{n}} E\left[\sum_{i=1}^{n} \left|X_i \frac{\|f''\|_\infty}{2}(W - W_i)^2\right|\right] \\
&= \frac{\|f''\|_\infty}{2\sqrt{n}} \sum_{i=1}^{n} E\left|X_i \left(\frac{X_i}{\sqrt{n}}\right)^2\right| = \frac{\|f''\|_\infty}{2n^{\frac{3}{2}}} \sum_{i=1}^{n} E\left[|X_i|^3\right].
\end{aligned}$$

Also, the mean value theorem shows that

$$\frac{1}{n} E\left[\sum_{i=1}^{n} |f'(W_i) - f'(W)|\right] \leq \frac{1}{n} E\left[\sum_{i=1}^{n} \left(\|f''\|_\infty |W_i - W|\right)\right] = \frac{\|f''\|_\infty}{n^{\frac{3}{2}}} \sum_{i=1}^{n} E|X_i|.$$

As $1 = E[X_i^2] = E\left[\left(|X_i|^3\right)^{\frac{2}{3}}\right] \leq E\left[|X_i|^3\right]^{\frac{2}{3}}$, we have $E\left[|X_i|^3\right] \geq 1$, so $E|X_i| \leq E\left[|X_i|^3\right]^{\frac{1}{3}} \leq E\left[|X_i|^3\right]$. (The conclusion is trivial if $E\left[|X_i|^3\right] = \infty$.)

Putting all of this together gives

$$\begin{aligned}
|E[f'(W) - Wf(W)]| &\leq \frac{1}{\sqrt{n}} E\left[\sum_{i=1}^{n} \left|X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right|\right] + \frac{1}{n} E\left[\sum_{i=1}^{n} |f'(W_i) - f'(W)|\right] \\
&\leq \frac{\|f''\|_\infty}{2n^{\frac{3}{2}}} \sum_{i=1}^{n} E\left[|X_i|^3\right] + \frac{\|f''\|_\infty}{n^{\frac{3}{2}}} \sum_{i=1}^{n} E|X_i| \leq \frac{3\|f''\|_\infty}{2n^{\frac{3}{2}}} \sum_{i=1}^{n} E\left[|X_i|^3\right],
\end{aligned}$$

and the result follows since

$$d_W(\mathscr{L}(W), \mathscr{L}(Z)) = \sup_{h \in \mathcal{H}_W} |E[f_h'(W) - Wf_h(W)]|$$

and $\|f_h''\|_\infty \leq 2\|h'\|_\infty = 2$ for all $h \in \mathcal{H}_W$. □

Of course the mean zero, variance one condition is just the usual normalization in the CLT and so imposes no real loss of generality. If the random variables have uniformly bounded third moments, then Theorem 10.5 gives a rate of order $n^{-\frac{1}{2}}$ which is the best possible.

11 Conditional Expectation

Let $(\Omega, \mathcal{F}, P)$ be a probability space and consider any $A, B \in \mathcal{F}$ with $P(B) > 0$. In undergraduate probability, we define the probability of A conditional on B as $P(A\,|B) = P(A \cap B)/P(B)$.

The idea is that if we learn B has occurred, then we must update our probability measure to account for this information. Our new measure, P_B, should satisfy $P_B(B) = 1$ (since we know B has occurred) and, for any $E, F \in \mathcal{F}$ with $E, F \subseteq B$, $P_B(E)P(F) = P(E)P_B(F)$ (since we have learned nothing about the relative likelihoods of events contained in B). It follows that for any $A \in \mathcal{F}$,

$$P_B(A) = P_B(A \cap B) + P_B(A \cap B^C) = P_B(A \cap B) = \frac{P_B(A \cap B)}{P_B(B)} = \frac{P(A \cap B)}{P(B)}.$$

(The second equality is because we must have $1 = P_B(\Omega) = P_B(B) + P_B(B^C) = 1 + P_B(B^C)$, so monotonicity dictates that events contained in B^C have probability 0 under P_B.)

When thinking about conditional probability, it can be instructive to take a step back and think of a second observer with access to partial information. Here we interpret $(\Omega, \mathcal{F}, P)$ as describing a random system whose chance of being in state $\omega \in \Omega$ is governed by P. $\mathcal{F}$ represents the possible conclusions that can be drawn about the state of the system: All that can be said is whether it lies in A for each $A \in \mathcal{F}$.
Now suppose that the observer has performed a measurement that tells her if B holds for some $B \in \mathcal{F}$ with $P(B) \in (0, 1)$. If she found out that B is true, her assessment of the probability of $A \in \mathcal{F}$ would be $P(A\,|B)$. If she found that B is false, she would evaluate the probability of A as $P\big(A\,|B^C\big)$. Thus, from our point of view, her description of the probability of A is given by the random variable

$$X_A(\omega) = \begin{cases} P\,(A\,|B)\,, & \omega \in B \\ P\big(A\,|B^C\big), & \omega \notin B \end{cases}.$$

This is ultimately the kind of idea we are trying to capture with conditional expectation.

The typical development in elementary treatments of probability is to apply the definition of $P(A\,|B)$ to the events $\{X = x\}$ and $\{Y = y\}$ for discrete random variables X, Y in order to define the conditional mass function of X given that $Y = y$ as $p_X(x\,|Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$. One then extrapolates to absolutely continuous X and Y by replacing mass functions with densities (which is problematic in that it treats pdfs as probabilities and raises issues concerning conditioning on null events). Finally, conditional expectation is defined in terms of integrating against the conditional pmfs/pdfs.

In what follows, we will need a more sophisticated theory of conditioning that avoids some of the pitfalls, paradoxes, and limitations of the framework sketched out above. Rather than try to arrive at the proper definition by way of more familiar concepts, we will begin with a formal definition and then work through a variety of examples and related results in order to provide motivation, build intuition, and make connections with ideas from elementary probability.

Definition. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ a random variable with $E\,|X| < \infty$, and $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra. We define $E[X\,|\mathcal{G}]$, the *conditional expectation of X given $\mathcal{G}$*, to be any random variable Y satisfying

(i) $Y \in \mathcal{G}$ (i.e. Y is measurable with respect to $\mathcal{G}$)

(ii) $\int_A Y\,dP = \int_A X\,dP$ for all $A \in \mathcal{G}$

If Y satisfies (i) and (ii), we say that Y is a *version* of $E[X\,|\mathcal{G}]$.

Our most immediate order of business is to show that this definition makes good mathematical sense by proving existence and uniqueness theorems.

To streamline this task, we first take a moment to establish integrability for random variables that fit the definition so we may manipulate various quantities of interest with impunity.

Lemma 11.1. *If Y satisfies conditions (i) and (ii) in the definition of $E[X\,|\mathcal{G}]$, then it is integrable.*

Proof. Letting $A = \{Y \geq 0\} \in \mathcal{G}$, condition (ii) implies

$$\int_A Y\,dP = \int_A X\,dP \leq \int_A |X|\,dP,$$
$$\int_{A^C} (-Y)\,dP = -\int_{A^C} Y\,dP = -\int_{A^C} X\,dP = \int_{A^C} (-X)\,dP \leq \int_{A^C} |X|\,dP.$$

It follows that

$$E\,|Y| = \int_A Y\,dP + \int_{A^C} (-Y)\,dP \leq \int_A |X|\,dP + \int_{A^C} |X|\,dP = E\,|X| < \infty. \qquad \square$$

The following existence proof gives an interpretation of conditional expectation in terms of Radon-Nikodym derivatives. (Recall from Theorem 11.25 that if μ and ν are σ-finite measures on $(S, \mathcal{S})$ with $\nu \ll \mu$, then there is a measurable function $f : S \to \mathbb{R}$ such that $\nu(A) = \int_A f\,d\mu$ for all $A \in \mathcal{S}$. $f = \frac{d\nu}{d\mu}$ is called the Radon-Nikodym derivative of ν with respect to μ.)

Theorem 11.2. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ a random variable with $E\,|X| < \infty$, and $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra. There exists a random variable Y satisfying*

(i) $Y \in \mathcal{G}$

(ii) $\int_A Y\,dP = \int_A X\,dP$ *for all* $A \in \mathcal{G}$

Proof. First suppose that $X \geq 0$. Define $\nu(A) = \int_A X dP$ for $A \in \mathcal{G}$. Then $P|_{\mathcal{G}}$ and ν are finite measures on $(\Omega, \mathcal{G})$. (That ν is countably additive is an easy application of the DCT.) Moreover, ν is clearly absolutely continuous with respect to P. The Radon-Nikodym theorem therefore implies that there is a function $\frac{d\nu}{dP} \in \mathcal{G}$ such that

$$\int_A X\,dP = \nu(A) = \int_A \frac{d\nu}{dP}\,dP.$$

It follows that $Y = \frac{d\nu}{dP}$ is a version of $E[X\,|\mathcal{G}]$.

For general X, write $X = X^+ - X^-$ and let $Y_1 = E\,[X^+\,|\mathcal{G}]$, $Y_2 = E\,[X^-\,|\mathcal{G}]$. Then $Y = Y_1 - Y_2$ is integrable and $\mathcal{G}$-measurable, so for all $A \in \mathcal{G}$,

$$\int_A Y\,dP = \int_A Y_1\,dP - \int_A Y_2\,dP = \int_A X^+\,dP - \int_A X^-\,dP = \int_A X\,dP. \qquad \square$$

Theorem 11.3. *Y is unique up to null sets.*

Proof. Suppose that Y' is also a version of $E[X\,|\mathcal{G}]$.

Condition (ii) implies that

$$\int_A Y'\,dP = \int_A X\,dP = \int_A Y\,dP$$

for all $A \in \mathcal{G}$.

By condition (i), the event $A_\varepsilon = \{Y - Y' \geq \varepsilon\}$ is in $\mathcal{G}$ for all $\varepsilon > 0$, hence

$$0 = \int_{A_\varepsilon} Y \, dP - \int_{A_\varepsilon} Y' \, dP = \int_{A_\varepsilon} (Y - Y') \, dP \geq \varepsilon P\left(Y - Y' \geq \varepsilon\right).$$

It follows that $Y \leq Y'$ a.s. Interchanging the roles of Y and Y' in the preceding argument shows that $Y' \leq Y$ a.s. as well, and the proof is complete. □

Proposition 11.4. *If Y is a version of $E[X\,|\mathcal{G}]$ and $Y' \in \mathcal{G}$ with $Y = Y'$ a.s., then Y' is also a version of $E[X|\mathcal{G}]$.*

Proof. Since Y and Y' are $\mathcal{G}$-measurable, $E = \{\omega : Y(\omega) \neq Y'(\omega)\} \in \mathcal{G}$. Since $P(E) = 0$, we see that for any $B \in \mathcal{G}$,

$$\begin{aligned}\int_B X \, dP = \int_B Y \, dP &= \int_{B\cap E} Y \, dP + \int_{B\setminus E} Y \, dP = \int_{B\setminus E} Y \, dP \\ &= \int_{B\setminus E} Y' \, dP = \int_{B\setminus E} Y' \, dP + \int_{B\cap E} Y' \, dP = \int_B Y' \, dP.\end{aligned}$$

□

Lemma 11.1, Theorem 11.3, and Proposition 11.4 combine to tell us that conditional expectation is unique as an element of $L^1(\Omega, \mathcal{G}, P)$. Just as elements of L^p spaces are really equivalence classes of functions (rather than specific functions) in classical analysis, conditional expectations are equivalence classes of random variables. Here versions play the role of specific functions.

Often we will omit the "almost sure" qualification when speaking of relations between conditional expectations, but it is important to keep this issue in mind.

In light of Proposition 11.4, we can often work with convenient versions of $E[X\,|\mathcal{G}]$ when we need to make use of pointwise results.

Examples

Intuitively, sub-σ-algebras represent (potentially available) information—for each $A \in \mathcal{G}$ we can ask whether or not A has occurred. From this perspective, we can think of $E[X\,|\mathcal{G}]$ as giving the 'best guess' for the value of X given the information in $\mathcal{G}$. The following examples are intended to clarify this view.

Example 11.5. If $X \in \mathcal{G}$, then our heuristic suggests that $E[X\,|\mathcal{G}] = X$ since if we know X, then our best guess is X itself. This clearly fulfills the definition as X always satisfies condition (ii) and condition (i) is met by assumption.

Since constants are measurable with respect to any σ-algebra, taking $X = c$ shows that $E[c\,|\mathcal{G}] = c$.

Example 11.6. At the other extreme, suppose that X is independent of $\mathcal{G}$—that is, for all $A \in \mathcal{G}$, $B \in \mathcal{B}$, $\{X \in B\}$ and A are independent events. In this case, $\mathcal{G}$ tells us nothing about X, so our best guess is $E[X]$. As a constant, $E[X]$ automatically satisfies condition (i).

To see that (ii) holds as well, note that for any $A \in \mathcal{G}$,

$$\int_A E[X] \, dP = E[X]P(A) = E[X]E[1_A] = E[X1_A] = \int_A X \, dP$$

by independence.

In particular, ordinary expectation corresponds to conditional expectation w.r.t. $\mathcal{G} = \{\Omega, \emptyset\}$.

Example 11.7. We now expand upon our introductory example: Suppose that $\Omega_1, \Omega_2, ...$ is a countable partition of Ω into disjoint measurable sets, each having positive probability (e.g. B and B^C). Let $\mathcal{G} = \sigma(\Omega_1, \Omega_2, ...)$. We claim that $E[X\,|\mathcal{G}] = P(\Omega_i)^{-1}E[X;\Omega_i]$ on Ω_i. The interpretation is that $\mathcal{G}$ tells us which Ω_i contains the outcome, and given that information, our best guess for X is its average over Ω_i.

To verify our claim, note that

$$E[X\,|\mathcal{G}](\omega) = \sum_i \frac{E[X;\Omega_i]}{P(\Omega_i)} 1_{\Omega_i}(\omega)$$

is $\mathcal{G}$-measurable since each Ω_i belongs to $\mathcal{G}$. Also, since each $A \in \mathcal{G}$ is a countable disjoint union of the Ω_i's, it suffices to check condition (ii) on the elements of the partition. But this is trivial as

$$\int_{\Omega_i} P(\Omega_i)^{-1}E[X;\Omega_i]\,dP = E[X;\Omega_i] = \int_{\Omega_i} X\,dP.$$

If we make the obvious definition $P(A\,|\mathcal{H}) = E[1_A\,|\mathcal{H}]$, then the above says that

$$P(A\,|\mathcal{G}) = P(\Omega_i)^{-1}\int_{\Omega_i} 1_A\,dP = \frac{P(A\cap\Omega_i)}{P(\Omega_i)} \text{ on } \Omega_i.$$

Example 11.8. Conditioning on a random variable can be seen as a special case of our definition by taking $E[X\,|Y\,] = E[X\,|\sigma(Y)]$. To see how this compares with the definition given in undergraduate probability, suppose that X and Y are discrete with joint pmf $p_{X,Y}$ and marginals p_X, p_Y. Then $\sigma(Y)$ is generated by the countable partition $\{Y = y\}_{y\in\text{Range}(Y)}$, so the previous example shows that if $E\,|X| < \infty$, then

$$E[X|Y] = P(Y=y)^{-1}E[X;\{Y=y\}] = \frac{1}{P(Y=y)}\sum_x xP(X=x,Y=y) = \sum_x x\frac{p_{X,Y}(x,y)}{p_Y(y)}$$

on $\{Y = y\}$.

Example 11.9. Similarly, suppose that X and Y are jointly absolutely continuous with joint density $f_{X,Y}$ and marginals f_X, f_Y. Suppose for simplicity that $f_Y(y) > 0$ for all $y \in \mathbb{R}$. In this case, if $E\,|g(X)| < \infty$, then $E[g(X)\,|Y\,] = h(Y)$ where

$$h(y) = \int g(x)\frac{f_{X,Y}(x,y)}{f_Y(y)}dx.$$

The Doob-Dynkin lemma shows that $E[g(X)\,|Y\,] \in \sigma(Y)$. To see that the second criterion is satisfied, recall that every $A \in \sigma(Y)$ is of the form $A = \{Y \in B\}$ for some $B \in \mathcal{B}$, and a change of variables gives

$$\begin{aligned}\int_{\{Y\in B\}} h(Y)\,dP &= \int_B h(y)f_Y(y)\,dy = \int 1_B(y)\left(\int g(x)\frac{f_{X,Y}(x,y)}{f_Y(y)}dx\right)f_Y(y)\,dy \\ &= \int\int g(x)1_B(y)f_{X,Y}(x,y)\,dxdy = E[g(X)1_B(Y)] = \int_{\{Y\in B\}} g(X)\,dP.\end{aligned}$$

Note that the condition $f_Y > 0$ is actually unnecessary since the above proof only needs h to satisfy

$$h(y)f_Y(y) = \int g(x)f_{X,Y}(x,y)\,dx,$$

so h can take on any value at those y with $f_Y(y) = 0$. (Since $f_Y(y) = \int f_{X,Y}(x,y)\,dx$ and $f_{X,Y} \geq 0$, the right-hand side of the above equation will also be 0 at such y.)

Example 11.10. Suppose that X and Y are independent and φ satisfies $E\,|\varphi(X,Y)| < \infty$. Then

$$E[\varphi(X,Y)\,|X\,] = g(X)$$

where $g(x) = E[\varphi(x,Y)]$.

As in the previous example, condition (i) is satisfied by Doob-Dynkin, and condition (ii) can be verified by letting μ and ν denote the distributions of X and Y, respectively, and computing

$$\begin{aligned}\int_{\{X\in B\}} g(X)\,dP &= \int_B g(x)\,d\mu(x) = \int 1_B(x)\left(\int \varphi(x,y)\,d\nu(y)\right)d\mu(x)\\ &= \int\int 1_B(x)\varphi(x,y)\,d(\mu\times\nu)(x,y) = \int 1_B(X)\varphi(X,Y)\,dP = \int_{\{X\in B\}} \varphi(X,Y)\,dP.\end{aligned}$$

Properties

Many of the properties of ordinary expectation carry over to conditional expectation as they are ultimately facts about integrals.

Proposition 11.11 (Linearity)**.** $E\,[aX+Y\,|\mathcal{G}\,] = aE\,[X\,|\mathcal{G}\,] + E\,[Y\,|\mathcal{G}\,]$

Proof. Sums and constant multiples of $\mathcal{G}$-measurable functions are $\mathcal{G}$-measurable, and for any $A\in\mathcal{G}$

$$\begin{aligned}\int_A \left(aE\,[X\,|\mathcal{G}\,] + E\,[Y\,|\mathcal{G}\,]\right)dP &= a\int_A E\,[X\,|\mathcal{G}\,]\,dP + \int_A E\,[Y\,|\mathcal{G}\,]\,dP\\ &= a\int_A X\,dP + \int_A Y\,dP = \int_A (aX+Y)\,dP.\end{aligned}$$

□

Proposition 11.12 (Monotonicity)**.** *If $X\le Y$, then $E[X\,|\mathcal{G}\,] \le E[Y\,|\mathcal{G}\,]$.*

Proof. By assumption, we have

$$\int_A E[X\,|\mathcal{G}\,]\,dP = \int_A X\,dP \le \int_A Y\,dP = \int_A E[Y\,|\mathcal{G}\,]\,dP$$

for all $A\in\mathcal{G}$. For any $\varepsilon>0$, $A_\varepsilon = \{\omega : E[X\,|\mathcal{G}\,] - E[Y\,|\mathcal{G}\,] \ge \varepsilon\} \in \mathcal{G}$, so

$$\varepsilon P(A_\varepsilon) \le \int_{A_\varepsilon} \left(E[X\,|\mathcal{G}\,] - E[Y\,|\mathcal{G}\,]\right)dP = \int_{A_\varepsilon} E[X\,|\mathcal{G}\,]\,dP - \int_{A_\varepsilon} E[Y\,|\mathcal{G}\,]\,dP \le 0.$$

It follows that $E[X\,|\mathcal{G}\,] \le E[Y\,|\mathcal{G}\,]$ a.s. □

Proposition 11.13 (Monotone Convergence)**.** *If $X_n \ge 0$ and $X_n \nearrow X$, then $E[X_n\,|\mathcal{G}\,] \nearrow E[X\,|\mathcal{G}\,]$.*

Proof. By monotonicity, $0 \le E[X_n\,|\mathcal{G}\,] \le E[X_{n+1}\,|\mathcal{G}\,] \le E[X\,|\mathcal{G}\,]$ for all n. (The inequalities are almost sure, but we can work with versions of the conditional expectations where they hold pointwise.) Since bounded nondecreasing sequences of reals converge to their limit superior, there is a random variable Y with $E[X_n\,|\mathcal{G}\,] \nearrow Y$.

Moreover, $Y\in\mathcal{G}$ as it is the limit of $\mathcal{G}$-measurable functions.

Finally, applying the ordinary MCT to $E[X_n\,|\mathcal{G}]1_B \nearrow Y1_B$, invoking the definition of conditional expectation, and then applying the MCT to $X_n 1_B \nearrow X1_B$ shows that

$$\int_B Y\,dP = \lim_{n\to\infty}\int_B E[X_n\,|\mathcal{G}]\,dP = \lim_{n\to\infty}\int_B X_n\,dP = \int_B X\,dP$$

for all $B \in \mathcal{G}$, hence Y is a version of $E[X\,|\mathcal{G}]$. □

Note that since we have established a conditional MCT, conditional versions of Fatou and dominated convergence follow from the usual arguments.

The final analogue we will consider is a conditional form of Jensen's inequality. It is fairly straightforward to derive conditional variants of other familiar theorems using these examples as templates.

Proposition 11.14 (Jensen)**.** *If φ is convex and $E\,|X|\,, E\,|\varphi(X)| < \infty$, then*

$$\varphi\left(E[X\,|\mathcal{G}]\right) \leq E[\varphi(X)\,|\mathcal{G}].$$

Proof. When we proved the original Jensen inequality, we established that if φ is convex, then for every $c \in \mathbb{R}$, there is a linear function $l_c(x) = a_c x + b_c$ such that $l_c(c) = \varphi(c)$ and $l_c(x) \leq \varphi(x)$ for all $x \in \mathbb{R}$.
Let $S = \{(a_r, b_r)\}_{r\in\mathbb{Q}}$. Then S is countable with $ax + b \leq \varphi(x)$ for all $x \in \mathbb{R}$, $(a, b) \in S$. Moreover, since $\mathbb{Q}$ is dense in $\mathbb{R}$ and convex functions are continuous, we have $\varphi(x) = \sup_{(a,b)\in S} ax + b$ for all $x \in \mathbb{R}$.
Monotonicity and linearity imply that

$$E[\varphi(X)\,|\mathcal{G}] \geq E[aX + b\,|\mathcal{G}] = aE[X\,|\mathcal{G}] + b \text{ a.s.}$$

whenever $(a, b) \in S$.
As S is countable, the event $A = \{E[\varphi(X)\,|\mathcal{G}] \geq aE[X\,|\mathcal{G}] + b \text{ for all } (a, b) \in S\}$ has full probability.
Thus with probability one, we have

$$E[\varphi(X)\,|\mathcal{G}] \geq \sup_{(a,b)\in S} aE[X\,|\mathcal{G}] + b = \varphi\left(E[X\,|\mathcal{G}]\right).$$ □

One use for conditional expectation is as an intermediary for computing ordinary expectations. This is justified by the *law of total expectation*:

Proposition 11.15. $E\left[E[X\,|\mathcal{G}]\right] = E[X]$.

Proof. Taking $A = \Omega$ in the definition of $E[X\,|\mathcal{G}]$ yields

$$E[X] = \int_\Omega X\,dP = \int_\Omega E[X\,|\mathcal{G}]\,dP = E\left[E[X\,|\mathcal{G}]\right].$$ □

As an example of the utility of the preceding observation, we prove

Proposition 11.16. *Conditional expectation is a contraction in L^p, $p \geq 1$.*

Proof. Since $\varphi(x) = |x|^p$ is convex, Proposition 11.14 implies that $|E[X\,|\mathcal{G}]|^p \leq E\left[|X|^p\,|\mathcal{G}\right]$.
Taking expectations and appealing to Proposition 11.15 gives

$$E\left[|E[X\,|\mathcal{G}]|^p\right] \leq E\left[E\left[|X|^p\,|\mathcal{G}\right]\right] = E\left[|X|^p\right].$$ □

Proposition 11.15 is actually a special case of the 'tower property' of conditional expectation.

This result is one of the more useful theorems about conditional expectation and is often summarized as "The smaller σ-algebra always wins."

Theorem 11.17. *If $\mathcal{G}_1 \subseteq \mathcal{G}_2$, then*

$$E\left[E[X\,|\mathcal{G}_1]\,|\mathcal{G}_2\right] = E\left[E[X\,|\mathcal{G}_2]\,|\mathcal{G}_1\right] = E[X\,|\mathcal{G}_1].$$

Proof. Since $E[X\,|\mathcal{G}_1] \in \mathcal{G}_1 \subseteq \mathcal{G}_2$, Example 11.5 shows that $E\left[E[X\,|\mathcal{G}_1]\,|\mathcal{G}_2\right] = E[X\,|\mathcal{G}_1]$.
To see that $E\left[E[X\,|\mathcal{G}_2]\,|\mathcal{G}_1\right] = E[X\,|\mathcal{G}_1]$, we observe that $E[X\,|\mathcal{G}_1] \in \mathcal{G}_1$ and for any $A \in \mathcal{G}_1 \subseteq \mathcal{G}_2$,

$$\int_A E[X\,|\mathcal{G}_1]\,dP = \int_A X\,dP = \int_A E[X\,|\mathcal{G}_2]\,dP. \qquad \square$$

Proposition 11.15 is the case $\mathcal{G}_1 = \{\Omega, \emptyset\}$, $\mathcal{G}_2 = \mathcal{G}$.

The second criterion in our definition of conditional expectation can be expressed in more probabilistic language as $E[Y1_A] = E[X1_A]$ for all $A \in \mathcal{G}$. One sometimes sees the alternative criterion $E[YZ] = E[XZ]$ for all bounded $Z \in \mathcal{G}$. The equivalence of the two conditions follows from the usual four-step procedure for building general integrals from integrals of indicators. We will stick with our original definition as it is easier to check.

The following theorem generalizes this alternative definition.

Theorem 11.18. *If $W \in \mathcal{G}$ and $E\,|X|\,, E\,|WX| < \infty$, then $E[WX\,|\mathcal{G}] = WE[X\,|\mathcal{G}]$.*

Proof. $WE[X\,|\mathcal{G}] \in \mathcal{G}$ by assumption, so we need only check the second criterion.
We first suppose that $W = 1_B$ for some $B \in \mathcal{G}$. Then for all $A \in \mathcal{G}$,

$$\begin{aligned}\int_A WE[X\,|\mathcal{G}]\,dP &= \int_A 1_B E[X\,|\mathcal{G}]\,dP = \int_{A\cap B} E[X\,|\mathcal{G}]\,dP \\ &= \int_{A\cap B} X\,dP = \int_A 1_B X\,dP = \int_A WX\,dP.\end{aligned}$$

By linearity, we see that the condition $\int_A WE[X\,|\mathcal{G}]\,dP = \int_A WX\,dP$ also holds when W is a simple function.
Now if $W, X \geq 0$, we can take a sequence of simple functions $W_n \nearrow W$ and use the MCT to conclude that

$$\begin{aligned}\int_A WE[X\,|\mathcal{G}]\,dP &= \lim_{n\to\infty} \int_A W_n E[X\,|\mathcal{G}]\,dP \\ &= \lim_{n\to\infty} \int_A W_n X\,dP = \int_A WX\,dP.\end{aligned}$$

The general result follows by splitting W and X into positive and negative parts. $\square$

Our last theorem about conditional expectation gives a geometric interpretation for square integrable X. Namely, noting that $L^2(\mathcal{F}) = \{Y \in \mathcal{F} : E[Y^2] < \infty\}$ is a Hilbert space and $L^2(\mathcal{G})$ is a closed subspace of $L^2(\mathcal{F})$, we will show that if $X \in L^2(\mathcal{F})$, then $E[X\,|\mathcal{G}]$ is the orthogonal projection of X onto $L^2(\mathcal{G})$.

Theorem 11.19. *If $E[X^2] < \infty$, then $E[X\,|\mathcal{G}]$ minimizes the mean square error $E[(X-Y)^2]$ amongst all $Y \in \mathcal{G}$.*

Proof. To begin, we note that if $Z \in L^2(\mathcal{G})$, then $E\,|ZX| < \infty$ by the Cauchy-Schwarz inequality, so Theorem 11.18 implies $ZE[X\,|\mathcal{G}] = E[ZX\,|\mathcal{G}]$.

Taking expected values gives

$$E\,[ZE[X\,|\mathcal{G}]] = E\,[E[ZX\,|\mathcal{G}]] = E[ZX],$$

showing that

$$E\,[Z\,(X - E[X\,|\mathcal{G}])] = E[ZX] - E\,[ZE[X\,|\mathcal{G}]] = 0$$

for $Z \in L^2(\mathcal{G})$.

Thus for any $Y \in L^2(\mathcal{G})$, if we set $Z = E[X\,|\mathcal{G}] - Y$, then we have

$$\begin{aligned} E\left[(X-Y)^2\right] &= E\left[((X - E[X\,|\mathcal{G}]) + Z)^2\right] \\ &= E\left[(X - E[X\,|\mathcal{G}])^2\right] + 2E\,[Z\,(X - E[X\,|\mathcal{G}])] + E[Z^2] \\ &= E\left[(X - E[X\,|\mathcal{G}])^2\right] + E[Z^2]. \end{aligned}$$

(Proposition 11.16 ensures $E[X\,|\mathcal{G}] \in L^2(\mathcal{G})$, so $Z = E[X\,|\mathcal{G}] - Y \in L^2(\mathcal{G})$ as well.)

It follows that $E\left[(X-Y)^2\right]$ is minimized over $L^2(\mathcal{G})$ when $E[X\,|\mathcal{G}] - Y = Z = 0$.

To see that $E[X\,|\mathcal{G}]$ minimizes the MSE over $L^0(\mathcal{G})$, we make use of the inequality

$$(a+b)^2 \leq (a+b)^2 + (a-b)^2 = 2a^2 + 2b^2.$$

If $Y \in \mathcal{G}$ is such that $E[(X-Y)^2] = \infty$, then it certainly doesn't minimize the MSE since $E[X\,|\mathcal{G}] \in L^2(\mathcal{G})$ with

$$E\left[(X - E[X\,|\mathcal{G}])^2\right] \leq 2E[X^2] + 2E\left[E[X\,|\mathcal{G}]^2\right] < \infty,$$

and if $E[(X-Y)^2] < \infty$, then

$$E[Y^2] = E[((Y-X)+X)^2] \leq 2E[(X-Y)^2] + 2E[X^2] < \infty. \qquad \square$$

Remark. In some treatments of conditional expectation, the Radon-Nikodym approach is bypassed entirely by first defining $E[X\,|\mathcal{G}]$ for $X \in L^2(\mathcal{F})$ in terms of projection onto $L^2(\mathcal{G})$, and then extending the definition to $X \in L^1(\mathcal{G})$ using approximating sequences of square integrable random variables. An upshot of this strategy is that one can then prove the Radon-Nikodym theorem using martingales!

Example 11.20. Define the *conditional variance* of X given $\mathcal{G}$ as

$$\mathrm{Var}(X\,|\mathcal{G}) = E\left[(X - E[X\,|\mathcal{G}])^2\,|\mathcal{G}\right] = E[X^2\,|\mathcal{G}] - E[X\,|\mathcal{G}]^2.$$

Adding

$$\begin{aligned} E\,[\mathrm{Var}(X\,|\mathcal{G})] &= E\left[E[X^2\,|\mathcal{G}]\right] - E\left[E[X\,|\mathcal{G}]^2\right] \\ &= E[X^2] - E\left[E[X\,|\mathcal{G}]^2\right] \end{aligned}$$

to

$$\begin{aligned} \mathrm{Var}\,(E[X\,|\mathcal{G}]) &= E\left[(E[X\,|\mathcal{G}] - E[X])^2\right] \\ &= E\left[E[X\,|\mathcal{G}]^2\right] - 2E[X]E\,[E[X\,|\mathcal{G}]] + E[X]^2 \\ &= E\left[E[X\,|\mathcal{G}]^2\right] - E[X]^2 \end{aligned}$$

yields the *law of total variance*

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = E\left[\mathrm{Var}(X\,|\mathcal{G})\right] + \mathrm{Var}\left(E[X\,|\mathcal{G}]\right).$$

The idea is that $X \in L^2(\mathcal{F})$ can be decomposed into its projection onto $L^2(\mathcal{G})$ and the mean-zero error term $W = X - E[X\,|\mathcal{G}]$. The variability in X that is not explained by $\mathcal{G}$ comes from W, which has variance

$$\mathrm{Var}(W) = E\left[\left(X - E[X\,|\mathcal{G}]\right)^2\right] = E\left[E\left[\left(X - E[X\,|\mathcal{G}]\right)^2|\mathcal{G}\right]\right] = E\left[\mathrm{Var}(X\,|\mathcal{G})\right].$$

Appendix

For the sake of completeness, we collect here some proofs of results that were left as facts to be accepted in the main body of these notes. They appear in the order they were mentioned, and so are sometimes anachronistic in terms of the concepts upon which they depend.

Regularity

Throughout this subsection, suppose that μ is a complete measure on $(\mathbb{R}, \mathcal{M})$ arising from a distribution function F. That is, for all $E \in \mathcal{M}$,

$$\begin{aligned}\mu(E) &= \inf\Big\{\sum_{j=1}^{\infty}[F(b_j) - F(a_j)] : E \subseteq \bigcup_{j=1}^{\infty}(a_j, b_j]\Big\} \\ &= \inf\Big\{\sum_{j=1}^{\infty}\mu\big((a_j, b_j]\big) : E \subseteq \bigcup_{i=1}^{\infty}(a_j, b_j]\Big\}.\end{aligned}$$

Lemma 11.21. *For any $E \in \mathcal{M}$, $\mu(E) = \inf\Big\{\sum_{j=1}^{\infty}\mu\big((a_j, b_j)\big) : E \subseteq \bigcup_{i=1}^{\infty}(a_j, b_j)\Big\}$.*

Proof. Let $\nu(E)$ denote the infimum appearing in the statement of the lemma. If $E \subseteq \bigcup_{j=1}^{\infty}(a_j, b_j)$, set $\ell_j = b_j - a_j$ and let $I_{j,k} = \big(b_j - \frac{\ell_j}{2^{k-1}}, b_j - \frac{\ell_j}{2^k}\big)$. Then $(a_j, b_j) = \bigsqcup_{k=1}^{\infty} I_{j,k}$, so $E \subseteq \bigcup_{j,k\in\mathbb{N}} I_{j,k}$ and $\sum_{j=1}^{\infty}\mu\big((a_j, b_j)\big) = \sum_{j,k\in\mathbb{N}}\mu(I_{j,k}) \geq \mu(E)$, hence $\nu(E) \geq \mu(E)$.

Conversely, given $\varepsilon > 0$, there exist $\{(a_j, b_j]\}_{j=1}^{\infty}$ with $E \subseteq \bigcup_{j=1}^{\infty}(a_j, b_j]$ and $\mu(E) \geq \sum_{j=1}^{\infty}\mu\big((a_j, b_j]\big) - \varepsilon$. Also, for each j, there exists $\delta_j > 0$ such that $F(b_j + \delta_j) - F(b_j) < \frac{\varepsilon}{2^j}$. Thus $E \subseteq \bigcup_{j=1}^{\infty}(a_j, b_j + \delta_j)$ and

$$\sum_{j=1}^{\infty}\mu\big((a_j, b_j + \delta_j)\big) \leq \sum_{j=1}^{\infty}\mu\big((a_j, b_j]\big) + \varepsilon \leq \mu(E) + 2\varepsilon$$

so that $\nu(E) \leq \mu(E)$ as well. □

Theorem 11.22. *For all $E \in \mathcal{M}$,*

$$\begin{aligned}\mu(E) &= \inf\big\{\mu(U) : U \text{ is open and } E \subseteq U\big\} \\ &= \sup\big\{\mu(K) : K \text{ is compact and } K \subseteq E\big\}.\end{aligned}$$

Proof. If U is open and $E \subseteq U$, then $\mu(U) \geq \mu(E)$. On the other hand, Lemma 2.3 ensures that we can write $U = \bigcup_{j=1}^{\infty}(a_j, b_j)$ so that $\mu(U) \leq \sum_{j=1}^{\infty}\mu\big((a_j, b_j)\big)$. Invoking Lemma 11.21 establishes the first claim.

For the second, suppose first that E is bounded. If E is also closed, then its compact so the second claim is immediate. Otherwise, given $\varepsilon > 0$, we can choose an open set $U \supseteq \overline{E} \setminus E$ such that $\mu(U) \leq \mu(\overline{E} \setminus E) + \varepsilon$. Then $K = \overline{E} \setminus U$ is a compact subset of E with $E \setminus K = E \cap (\overline{E} \cap U^C)^C = E \cap (\overline{E}^C \cup U) = E \cap U$, hence

$$\begin{aligned}\mu(K) &= \mu(E) - \mu(E \cap U) \\ &= \mu(E) - [\mu(U) - \mu(U \setminus E)] \\ &\geq \mu(E) - \mu(U) + \mu(\overline{E} \setminus E) \geq \mu(E) - \varepsilon.\end{aligned}$$

Finally, if E is unbounded, write $E_j = E \cap (j, j+1]$ for $j \in \mathbb{Z}$. The preceding ensures that there is a compact $K_j \subseteq E_j$ with $\mu(K_j) \geq \mu(E_j) - \frac{\varepsilon}{2^{|j|}}$. Then $H_n = \bigcup_{j=-n}^{n} K_j$ is a compact subset of E with $\mu(H_n) \geq \sum_{j=-n}^{n}\mu(E_j) - 3\varepsilon = \mu\left(\bigcup_{j=-n}^{n} E_j\right) - 3\varepsilon$. Since $\mu(E) = \lim_{n\to\infty}\mu\left(\bigcup_{j=-n}^{n} E_j\right)$, there is an $N \in \mathbb{N}$ with $\mu\left(\bigcup_{j=-N}^{N} E_j\right) \geq \mu(E) - \varepsilon$, hence $\mu(H_N) \geq \mu(E) - 4\varepsilon$. □

In light of the preceding, we see that if μ is a probability distribution on $\mathbb{R}$, then for any $E \in \mathcal{B}$, $\varepsilon > 0$, there is an open set $U \supseteq E$ with $0 \leq \mu(U) - \mu(B) < \varepsilon/2$. Since U is a countable union of disjoint open intervals, $\lim_{n\to\infty} \sum_{k=1}^{n} \mu\big((a_k, b_k)\big) = \mu(U) \leq 1$, so there is an $N \in \mathbb{N}$ with $\mu(U) \leq \sum_{k=1}^{N} \mu\big((a_k, b_k)\big) + \varepsilon/2$. Consequently, $\left|\mu(B) - \sum_{k=1}^{N} \mu\big((a_k, b_k)\big)\right| < \varepsilon$. Thus every Borel set is almost a finite union of open intervals!

Lebesgue-Radon-Nikodym

Our first step in establishing the Lebesgue-Radon-Nikodym theorem amounts to a derivation of the *Hahn decomposition* of a finite signed measure. [One can build up a theory of $\mathbb{C}$-valued measures by considering real and imaginary parts and then positive and negative parts. As we have no need for such generality in this course, we'll just state our results in terms of ordinary (positive) measures.]

Proposition 11.23. *Suppose μ and ν are finite measures on $(S, \mathcal{G})$. Then there is a set $P \in \mathcal{G}$ such that $\mu(A) \geq \nu(A)$ if $A \subseteq P$ and $\nu(B) \geq \mu(B)$ if $B \subseteq P^C$.*

Proof. Write $\eta = \mu - \nu$ and say that $F \in \mathcal{G}$ is positive for η if $\eta(B) \geq 0$ for all $B \subseteq F$. Our aim is thus to find a set $P \in \mathcal{G}$ such that P is positive for η and P^C is positive for $-\eta$. (In this case, we say that (P, P^C) is a *Hahn decomposition* for η.)

We first claim that for any $A \in \mathcal{G}$, $\varepsilon > 0$, there is a $B \subseteq A$ such that $\eta(B) \geq \eta(A)$ and $\eta(E) > -\varepsilon$ for all $E \subseteq B$. If not, there is an $E_1 \subseteq A$ such that $\eta(E_1) \leq -\varepsilon$. Since $\eta(A \setminus E_1) = \eta(A) - \eta(E_1) \geq \eta(A)$, there is an $E_2 \in A \setminus E_1$ with $\eta(E_2) < -\varepsilon$. Continuing thusly gives a sequence $\{E_n\}$ of disjoint sets with $\mu(E_n) - \nu(E_n) = \eta(E_n) < -\varepsilon$ for all n. But then for every $N \in \mathbb{N}$, $E_N = \bigsqcup_{n=1}^{N} E_n$ has the property that $\mu(E_N) - \nu(E_N) = \sum_{n=1}^{N} [\mu(E_n) - \nu(E_n)] < -N\varepsilon$, hence $\nu(\Omega) \geq \nu(E_N) > \mu(E_N) + N\varepsilon > N\varepsilon$, contradicting the assumption that ν is finite.

Next, given $A \in \mathcal{G}$, let $A_1 = A$ and inductively pick $A_n \subseteq A_{n-1}$ such that $\eta(A_n) \geq \eta(A_{n-1})$ and $\nu(B) > -\frac{1}{n}$ for all $B \subseteq A_n$. Let $F = \bigcap_{n=1}^{\infty} A_n$. Then F is positive for η and continuity from above implies $\eta(F) = \lim_{n\to\infty} \eta(A_n) \geq \eta(A)$.

Now let $\alpha = \sup\{\eta(A) : A \in \mathcal{G}\}$. Then the preceding ensures we can choose a sequence $\{P_n\}$ in $\mathcal{G}$ such that each P_n is positive for η and $\lim_{n\to\infty} \eta(P_n) = \alpha$. Continuity from below shows that $P = \bigcup_{n=1}^{\infty} P_n$ is positive with $\eta(P) = \alpha$. Moreover, $\eta(B) \leq 0$ for all $B \subseteq P^C$ since otherwise we would have $\eta(P \cup B) = \eta(P) + \eta(B) > \alpha$.

Finally, we observe that if P' is any other set having this property, then $P' \setminus P \subseteq P'$ and $P' \setminus P \subseteq P^C$, so $0 \leq \eta(P' \setminus P) \leq 0$, and likewise for $P \setminus P'$. It follows that $\mu(P \Delta P') - \nu(P \Delta P') = \eta(P \Delta P') = 0$, so this decomposition is unique up to sets on which μ and ν agree. □

Proposition 11.23 gives us the following useful characterization of mutual singularity.

Lemma 11.24. *Suppose μ and ν are finite measures on $(S, \mathcal{G})$. Either $\mu \perp \nu$ or there exist $\varepsilon > 0$, $E \in \mathcal{G}$ with $\mu(E) > 0$ and $\nu(B) \geq \varepsilon\mu(B)$ for all $B \subseteq E$.*

Proof. For each $n \in \mathbb{N}$ let (P_n, P_n^C) be a Hahn decomposition for $\eta_n = \nu - n^{-1}\mu$. Set $P = \bigcup_{n=1}^{\infty} P_n$. Then $P^C = \bigcap_{n=1}^{\infty} P_n^C$ satisfies $\nu(B) - n^{-1}\mu(B) \leq 0$ and thus $0 \leq \nu(B) \leq n^{-1}\mu(B)$ for all $B \subseteq P^C$, $n \in \mathbb{N}$. In particular, $\nu(P^C) = 0$. If $\mu(P) = 0$, then $\mu \perp \nu$. Otherwise, $\mu(P_n) > 0$ for some $n \in \mathbb{N}$ and $\nu(B) - n^{-1}\mu(B) \geq 0$, hence $\nu(B) \geq n^{-1}\mu(B)$ for all $B \subseteq P_n$. □

At this point, we observe that if μ is a finite measure on $(S,\mathcal{G})$ and $f : S \to [0,\infty)$ is integrable, then $\nu(E) = \int_E f\,d\mu := \int f1_E\,d\mu$ is also a finite measure. Indeed, if $E = \bigsqcup_{k=1}^{\infty} E_k$, writing $F_n = \bigsqcup_{k=1}^{n} E_k$, we have that $F_n \nearrow E$, so $f1_{F_n} \nearrow f1_E$ pointwise. Monotone convergence and linearity then give

$$\begin{aligned}\nu(E) &= \int f1_E\,d\mu = \int \lim_{n\to\infty} f1_{F_n}\,d\mu = \lim_{n\to\infty}\int f1_{F_n}\,d\mu \\ &= \lim_{n\to\infty}\int f\sum_{k=1}^{n} 1_{E_k}\,d\mu = \lim_{n\to\infty}\sum_{k=1}^{n}\int f1_{E_k}\,d\mu = \sum_{k=1}^{\infty}\nu(E_k).\end{aligned}$$

Moreover, if $\mu(E) = 0$, then $f1_E = 0$ μ-a.s., hence $\nu(E) = \int f1_E\,d\mu = 0$—that is, $\nu \ll \mu$.

The *Radon-Nikodym theorem* asserts that the converse is true—if $\nu \ll \mu$, then $d\nu = f\,d\mu$ in the sense that $\nu(E) = \int_E f\,d\mu$ for all $E \in \mathcal{F}$. We call $f = \frac{d\nu}{d\mu}$ the *Radon-Nikodym derivative* of ν with respect to μ. Additionally, the *Lebesgue decomposition theorem* says that we can uniquely express ν as a sum of measures that are absolutely continuous/singular, with respect to μ. The following theorem combines these statements.

Theorem 11.25. *If μ and ν are finite measures on $(S,\mathcal{G})$, then there exist unique measures ρ,λ with $\rho \ll \mu$, $\lambda \perp \mu$, and $\nu = \rho + \lambda$. Moreover, there is an integrable $\varphi : S \to [0,\infty)$ with $\rho(E) = \int_E \varphi\,d\mu$ for all $E \in \mathcal{G}$.*

Proof. Define $\mathscr{F} = \left\{f : S \to [0,\infty) : \int_E f\,du \le \nu(E) \text{ for all } E \in \mathcal{G}\right\}$. $\mathscr{F}$ is nonempty since it contains $f \equiv 0$. Also, if $f,g \in \mathscr{F}$, then so is $h = \max\{f,g\}$: If $A = \{x \in S : f(x) > g(x)\}$, then $h = f1_A + g1_{A^C}$ satisfies $\int_E h\,d\mu = \int_{E\cap A} f\,d\mu + \int_{E\cap A^C} g\,d\mu \le \nu(E\cap A) + \nu(E\cap A^C) = \nu(E)$.

Let $s = \sup\left\{\int f\,d\mu : f \in \mathscr{F}\right\}$, so that $s \le \nu(S)$, and choose a sequence $\{f_n\}$ in $\mathscr{F}$ with $\int f_n\,d\mu \to s$. Define $g_n = \max\{f_1,\dots,f_n\}$ and $\varphi = \sup_n f_n$. Then $g_n \in \mathscr{F}$, g_n increases pointwise to φ, and $\int g_n\,d\mu \ge \int f_n\,d\mu$ for all n. It follows that $\varphi \in \mathscr{F}$ and, by monotone convergence, $\int \varphi\,d\mu = \lim_{n\to\infty}\int g_n\,d\mu = s$.

Setting $\rho(E) = \int_E \varphi\,d\mu$, we have that $\rho \ll \mu$ and $\rho(E) \le \nu(E)$ for all $E \in \mathcal{G}$. This shows that $\lambda := \nu - \rho$ is a (positive) measure on $(S,\mathcal{G})$. Moreover, it must be the case that $\lambda \perp \mu$ because otherwise there would be some $F \in \mathcal{G}$, $\varepsilon > 0$ with $\mu(F) > 0$ and $\nu(B) - \rho(B) = \lambda(B) \ge \varepsilon\mu(B)$ for all $B \subseteq F$. But then $f = \varphi + \varepsilon 1_F$ satisfies $\int_E f\,d\mu = \rho(E\cap F^C) + [\rho(E\cap F) + \varepsilon\mu(F\cap E)] \le \nu(E\cap F^C) + \nu(E\cap F) = \nu(E)$, hence $f \in \mathscr{F}$, and $\int f\,d\mu = s + \varepsilon\mu(F) > s$, a contradiction.

For uniqueness, suppose $\nu = \rho' + \lambda'$ is another such decomposition. Then there are sets $E_1, E_2 \in \mathcal{G}$ with $\mu(E_1) = \mu(E_2) = 0$ and $\lambda(E_1^C) = \lambda'(E_2^C) = 0$. Writing $E = E_1 \cup E_2$, so that $E^C = E_1^C \cap E_2^C$, we see that $\mu(E) = 0$ (hence $\rho(E) = \rho'(E) = 0$) and $\lambda(E^C) = \lambda'(E^C) = 0$. It follows that for any $A \in \mathcal{G}$, $\lambda(A) = \lambda(A\cap E) = \nu(A\cap E) = \lambda'(A\cap E) = \lambda'(A)$ and $\rho(A) = \rho(A\cap E^C) = \nu(A\cap E^C) = \rho'(A\cap E^C) = \rho'(A)$. □

$(S,\mathcal{G},\mu)$ is *σ-finite* if S is a countable union of sets having finite μ-measure; Lebesgue measure on $\mathbb{R}$ is such an example. If μ and ν are σ-finite measures on $(S,\mathcal{G})$, we can write $S = \bigsqcup_{j=1}^{\infty} S_j$ with $\mu(S_j),\nu(S_j) < \infty$ for all j. Applying the preceding to the restrictions of μ and ν to each S_j and then recombining the pieces shows that the finiteness assumption can be weakened to σ-finiteness.

Riemann-Lebesgue

Suppose f is an $\mathbb{R}$-valued function defined on a bounded interval $[a,b]$. Let $\mathcal{P} = \{x_0,x_1,\dots,x_n\}$ be a *partition* of $[a,b]$ (so $a = x_0 < x_1 < \cdots < x_n = b$), denote its *mesh* by $|\mathcal{P}| = \max_{1\le k\le n}(x_k - x_{k-1})$, and let $\mathcal{T} = \{t_1,\dots,t_n\}$ be a sequence of *tags* for $\mathcal{P}$ (so $t_k \in [x_{k-1},x_k]$). The associated *Riemann sum* is given by

$$\mathcal{R}(f,\mathcal{P},\mathcal{T}) = \sum_{k=1}^{n} f(t_k)(x_k - x_{k-1}).$$

We say that f is *Riemann integrable* over $[a,b]$ if there exists $I \in \mathbb{R}$ such that for every $\varepsilon > 0$, there is a $\delta > 0$ such that any tagged partition $(\mathcal{P}, \mathcal{T})$ with $|\mathcal{P}| < \delta$ satisfies $|\mathcal{R}(f, \mathcal{P}, \mathcal{T}) - I| < \varepsilon$. In this case, we write $I = \int_a^b f(t)\,dt$, the *Riemann integral* of f over $[a,b]$.

Alternatively, given a partition $\mathcal{P}$ of $[a,b]$, define $M_j = \sup_{t \in [x_{j-1}, x_j]} f(t)$ and $m_j = \inf_{t \in [x_{j-1}, x_j]} f(t)$. Define the upper and lower *Darboux sums* by

$$\mathcal{U}(f, \mathcal{P}) = \sum_{j=1}^{n} M_j (x_j - x_{j-1}), \qquad \mathcal{V}(f, \mathcal{P}) = \sum_{j=1}^{n} m_j (x_j - x_{j-1}).$$

Necessarily $\mathcal{V}(f, \mathcal{P}) \leq \mathcal{R}(f, \mathcal{P}, \mathcal{T}) \leq \mathcal{U}(f, \mathcal{P})$ for all $(\mathcal{P}, \mathcal{T})$. Moreover, if $\mathcal{P}'$ refines $\mathcal{P}$ in the sense that $\mathcal{P} \subseteq \mathcal{P}'$, then one has $\mathcal{V}(f, \mathcal{P}) \leq \mathcal{V}(f, \mathcal{P}') \leq \mathcal{U}(f, \mathcal{P}') \leq \mathcal{U}(f, \mathcal{P})$. It can be shown that Riemann integrability is equivalent to $\sup_{\mathcal{P}} \mathcal{V}(f, \mathcal{P}) = \inf_{\mathcal{P}} \mathcal{U}(f, \mathcal{P})$, in which case the common value is $\int_a^b f(t)\,dt$.

Theorem 11.26. *Let f be a bounded function on the finite interval $[a,b]$ and let m denote Lebesgue measure.*

(1) If f is Riemann integrable, then it's Lebesgue measurable (and thus integrable since it's bounded), and $\int_a^b f(x)\,dx = \int_{[a,b]} f\,dm$.

(2) f is Riemann integrable iff $m(\{x \in [a,b] : f \text{ is discontinuous at } x\}) = 0$.

Proof. Suppose f is Riemann integrable. For each partition $\mathcal{P} = \{x_0, \ldots, x_n\}$ of $[a,b]$, define $g_{\mathcal{P}} = \sum_{j=1}^{n} m_j 1_{[x_{j-1}, x_j]}$ and $G_{\mathcal{P}} = \sum_{k=1}^{n} M_j 1_{[x_{j-1}, x_j]}$ where $m_j = \inf_{t \in [x_{j-1}, x_j]} f(t)$ and $M_j = \sup_{t \in [x_{j-1}, x_j]} f(t)$. By assumption, there is a nested sequence of partitions $\mathcal{P}_1 \subseteq \mathcal{P}_2 \subseteq \cdots$ such that $\lim_{k\to\infty} \mathcal{V}(f, \mathcal{P}_k) = \lim_{k\to\infty} \mathcal{U}(f, \mathcal{P}_k) = \int_a^b f(t)\,dt$. Let $G = \lim_{k\to\infty} G_{\mathcal{P}_k}$ and $g = \lim_{k\to\infty} g_{\mathcal{P}_k}$. Then $g \leq f \leq G$, and by dominated convergence, $\int G\,dm = \int_a^b f(t)\,dt = \int g\,dm$. It follows that $\int (G - g)\,dm = 0$, so $G = g$ almost everywhere, hence $f = G$ a.e. Since G is a limit of simple functions, f is measurable with $\int f\,dm = \int G\,dm = \int_a^b f(x)\,dx$.

To prove the second statement, define $H(x) = \limsup_{y \to x} f(y)$ and $h(x) = \liminf_{y \to x} f(y)$. Note that f is continuous at x iff $H(x) = h(x)$. Also, using the notation from the first part, $H = G$ and $h = g$ on a set of full measure, hence both are measurable/integrable with $\int_{[a,b]} H\,dm = \int G\,dm$ and $\int_{[a,b]} h\,dm = \int g\,dm$. The result follows since f is integrable iff $\int G\,dm = \int g\,dm$ iff $H - h \geq 0$ vanishes outside a null set. $\square$

Product σ-fields

Definition. Given an indexed collection of measurable spaces $\{(S_\alpha, \mathcal{G}_\alpha)\}_{\alpha \in A}$, the *product* σ-field, $\bigotimes_{\alpha \in A} \mathcal{G}_\alpha$, on $S = \prod_{\alpha \in A} S_\alpha$ is generated by $\left\{\pi_\alpha^{-1}(G_\alpha) : G_\alpha \in \mathcal{G}_\alpha, \alpha \in A\right\}$ where $\pi_\alpha : S \to S_\alpha$ is projection onto the α coordinate.

The product σ-algebra is thus the smallest σ-algebra for which the projections are measurable. This is because we want a function taking values in the product space to be measurable precisely when its components are.

Proposition 11.27. *If A is countable, then $\bigotimes_{\alpha \in A} \mathcal{G}_\alpha$ is generated by the rectangles $\left\{\prod_{\alpha \in A} G_\alpha : G_\alpha \in \mathcal{G}_\alpha\right\}$. If, in addition, $\mathcal{G}_\alpha$ is generated by $\mathcal{E}_\alpha$ for every $\alpha \in A$, then $\bigotimes_{\alpha \in A} \mathcal{G}_\alpha$ is generated by $\left\{\prod_{\alpha \in A} E_\alpha : E_\alpha \in \mathcal{E}_\alpha\right\}$.*

Proof. If $G_\alpha \in \mathcal{G}_\alpha$, then $\pi_\alpha^{-1}(G_\alpha) = \prod_{\beta \in A} G_\beta$ where $G_\beta = S_\beta$ for all $\beta \neq \alpha$, hence

$$\sigma\left(\left\{\pi_\alpha^{-1}(G_\alpha) : G_\alpha \in \mathcal{G}_\alpha, \alpha \in A\right\}\right) \subseteq \sigma\Big(\Big\{\prod_{\alpha \in A} G_\alpha : G_\alpha \in \mathcal{G}_\alpha\Big\}\Big).$$

On the other hand, $\prod_{\alpha\in A} G_\alpha = \bigcap_{\alpha\in A} \pi_\alpha^{-1}(G_\alpha)$, so

$$\sigma\Big(\Big\{\prod_{\alpha\in A} G_\alpha : G_\alpha \in \mathcal{G}_\alpha\Big\}\Big) \subseteq \sigma\left(\left\{\pi_\alpha^{-1}(G_\alpha) : G_\alpha \in \mathcal{G}_\alpha, \alpha \in A\right\}\right).$$

The second statement will follow from the above argument once we show that $\bigotimes_{\alpha\in A} \mathcal{G}_\alpha$ is generated by $\mathcal{F}_1 = \left\{\pi_\alpha^{-1}(E_\alpha) : E_\alpha \in \mathcal{E}_\alpha, \alpha \in A\right\}$. To this end, observe that $\mathcal{F}_1 \subseteq \left\{\pi_\alpha^{-1}(G_\alpha) : G_\alpha \in \mathcal{G}_\alpha, \alpha \in A\right\}$ by construction, so $\sigma(\mathcal{F}_1) \subseteq \bigotimes_{\alpha\in A} \mathcal{G}_\alpha$.

Conversely, arguing as in the proof of Theorem 4.2 shows that for each $\alpha \in A$, $\left\{E \subseteq S_\alpha : \pi_\alpha^{-1}(E) \in \sigma(\mathcal{F}_1)\right\}$ is a σ-algebra containing $\mathcal{E}_\alpha$ (and thus $\mathcal{G}_\alpha$), so $\pi_\alpha^{-1}(E) \in \sigma(\mathcal{F}_1)$ for all $E \in \mathcal{G}_\alpha$, and we conclude that $\sigma\left(\left\{\pi_\alpha^{-1}(G_\alpha) : G_\alpha \in \mathcal{G}_\alpha, \alpha \in A\right\}\right) \subseteq \sigma(\mathcal{F}_1)$ as well. □

Proposition 11.28. *If $S_1, \ldots, S_n$ are separable metric spaces and $S = \prod_{i=1}^n S_i$ is equipped with the product metric, then $\mathcal{B}_X = \bigotimes_{i=1}^n \mathcal{B}_{S_i}$.*

Proof. By Proposition 11.27, $\bigotimes_{i=1}^n \mathcal{B}_{S_i}$ is generated by $\left\{\pi_i^{-1}(U_i) : U_i \text{ is open in } S_i, i \in [n]\right\}$. Since each $\pi_i^{-1}(U_i)$ is open in S, $\bigotimes_{i=1}^n \mathcal{B}_{S_i} \subseteq \mathcal{B}_S$.

For the reverse inclusion, suppose $D_i = \{x_{i,k}\}_{k=1}^\infty$ is a countable dense set in S_i for each $i \in [n]$ and let $\mathcal{E}_i = \left\{B_r(x_{i,k}) : r \in \mathbb{Q}^+, k \in \mathbb{N}\right\}$ be the collection of balls of rational radii centered at the $x_{i,k}$. For each open $U \subseteq S$, $x \in U$, there exist $x_{i,k_i} \in D_i$, $i \in [n]$, and $r \in \mathbb{Q}^+$ such that $x \in \prod_{i=1}^n B_r(x_{i,k_i}) \subseteq U$. Accordingly, U can be written as a union of such products. As each $\mathcal{E}_i$ is countable, this union must be as well, and we conclude that $\mathcal{B}_S \subseteq \bigotimes_{i=1}^n \mathcal{B}_{S_i}$. □

Corollary 11.29. *If $\mathcal{E}$ is any of the collections from Theorem 2.4, then $\mathcal{B}^d$ is generated, by the semialgebra of rectangles $\mathcal{R} = \left\{J_1 \times \cdots \times J_d : J_k \in \mathcal{E}\right\}$.*

Convex Functions

Lemma 11.30. *If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex, then for all $x < y < z$,*

$$\frac{\varphi(y) - \varphi(x)}{y - x} \leq \frac{\varphi(z) - \varphi(x)}{z - x} \leq \frac{\varphi(z) - \varphi(y)}{z - y}.$$

Proof. Writing $\lambda = \frac{y-x}{z-x} \in (0,1)$, we have $y = \lambda z + (1-\lambda)x$, so it follows from convexity that $\varphi(y) \leq \lambda\varphi(z) + (1-\lambda)\varphi(x)$, and thus

$$\varphi(y) - \varphi(x) \leq \lambda\left(\varphi(z) - \varphi(x)\right) = \frac{y - x}{z - x}\left(\varphi(z) - \varphi(x)\right).$$

Dividing by $y - x > 0$ gives the first inequality.

Similarly, setting $\mu = \frac{z-y}{z-x} = 1 - \lambda \in (0,1)$, we have $y = \mu x + (1-\mu)z$, so $\varphi(y) \leq \mu\varphi(x) + (1-\mu)\varphi(z)$, and thus

$$\varphi(y) - \varphi(z) \leq \mu\left(\varphi(x) - \varphi(z)\right) = \frac{z - y}{z - x}\left(\varphi(x) - \varphi(z)\right),$$

hence

$$\frac{\varphi(z) - \varphi(y)}{z - y} \geq \frac{\varphi(z) - \varphi(x)}{z - x}.$$

□

Proof of Fact 7.3. For any $h > 0$, taking $x = c - h$, $y = c$, $z = c + h$ in Lemma 11.30, it follows from the outer inequality that that

$$\frac{\varphi(c) - \varphi(c-h)}{h} \leq \frac{\varphi(c+h) - \varphi(c)}{h}.$$

Also, for any $0 < h_1 < h_2$, we have $c - h_2 < c - h_1 < c$, so the second inequality in Lemma 11.30 shows that $\frac{\varphi(c)-\varphi(c-h_2)}{h_2} \leq \frac{\varphi(c)-\varphi(c-h_1)}{h_1}$.

Similarly, since $c < c + h_1 < c + h_2$, the first inequality in Lemma 11.30 shows that $\frac{\varphi(c+h_2)-\varphi(c)}{h_2} \geq \frac{\varphi(c+h_1)-\varphi(c)}{h_1}$.

Consequently, the one-sided derivatives exist and satisfy

$$\varphi_l'(c) := \lim_{h\to 0^+} \frac{\varphi(c) - \varphi(c-h)}{h} \leq \lim_{h\to 0^+} \frac{\varphi(c+h) - \varphi(c)}{h} := \varphi_r'(c).$$

Now let $a \in [\varphi_l'(c), \varphi_r'(c)]$ and define the linear function $\ell(x) = a(x - c) + \varphi(c)$. Clearly, $\ell(c) = \varphi(c)$.

To see that $\ell(x) \leq \varphi(x)$ for all $x \in \mathbb{R}$, note that if $x < c$, then $x = c - k$ for some $k > 0$, so

$$\ell(x) - \varphi(x) = a(x - c) + \varphi(c) - \varphi(c - k) = -k\left(a - \frac{\varphi(c) - \varphi(c-k)}{k}\right) \leq 0$$

since $\frac{\varphi(c)-\varphi(c-k)}{k} \leq \varphi_l'(c) \leq a$ by monotonicity. The $x > c$ case is similar. □

Fubini-Tonelli

Lemma 11.31. *Suppose that $(R, \mathcal{F}, \mu)$ and $(S, \mathcal{G}, \nu)$ are probability spaces. For $E \in \mathcal{F} \otimes \mathcal{G}$, $x \in R$, $y \in S$, define $E_x = \{y \in S : (x, y) \in E\}$, $E_y = \{x \in R : (x, y) \in E\}$. Then*

(1) $E_x \in \mathcal{G}$, $E_y \in \mathcal{F}$ for all $x \in R$, $y \in S$
(2) The maps $x \mapsto \nu(E_x)$, $y \mapsto \mu(E_y)$ are measurable functions on R and S, respectively,
(3) $(\mu \times \nu)(E) = \int_R \nu(E_x)\, d\mu(x) = \int_S \mu(E_y)\, d\nu(y)$ for all $E \in \mathcal{F} \otimes \mathcal{G}$.

Proof. Let $\mathcal{L}$ be the set of all $E \in R \times S$ for which the conclusions of the lemma obtain.

If $A \in \mathcal{F}$,$B \in \mathcal{G}$, then $(A \times B)_x = B$ for $x \in A$ and $(A \times B)_x = \emptyset$ for $x \in A^C$, so $(A \times B)_x \in \mathcal{G}$ for all $x \in R$ and $\nu\big((A \times B)_x\big) = \nu(B)1_A(x)$ is a measurable function of x with

$$(\mu \times \nu)(A \times B) = \mu(A)\nu(B) = \nu(B) \int_R 1_A(x)\, d\mu(x) = \int_R \nu\big((A \times B)_x\big)\, d\mu(x).$$

The analogous claims hold for the y-slices, so the π-system $\mathcal{R} = \{A \times B : A \in \mathcal{F}, B \in \mathcal{G}\}$, which contains $R \times S$ and generates $\mathcal{F} \otimes \mathcal{G}$, is contained in $\mathcal{L}$.

If $E, F \in \mathcal{L}$ with $E \subseteq F$, then $E_x \subseteq F_x$ and $(F \setminus E)_x = F_x \setminus E_x$, so $\nu\big((F \setminus E)_x\big) = \nu(F_x) - \nu(E_x)$ is a difference of measurable functions with

$$\begin{aligned}(\mu \times \nu)(F \setminus E) &= (\mu \times \nu)(F) - (\mu \times \nu)(E) \\ &= \int_R \nu(F_x)\, d\mu(x) - \int_R \nu(E_x)\, d\mu(x) \\ &= \int_R [\nu(F_x) - \nu(E_x)]\, d\mu(x) = \int_R \nu\big((F \setminus E)_x\big)\, d\mu(x).\end{aligned}$$

Repeating the above with y shows that $\mathcal{L}$ is closed under subset differences.

Finally, if $\{E_k\} \subseteq \mathcal{L}$ with $E_k \nearrow E$, then $(E_k)_x \nearrow E_x$, so $\nu\big(E_x\big) = \lim_{n\to\infty} \nu\big((E_k)_x\big)$ is a limit of measurable functions with

$$(\mu \times \nu)(E) = \lim_{n\to\infty} (\mu \times \nu)(E_k) = \lim_{n\to\infty} \int_R \nu\big((E_k)_x\big)\, d\mu(x) = \int_R \nu\big(E_x\big)\, d\mu(x)$$

by monotone convergence. As the same holds for y, we see that $\mathcal{L}$ is a λ-system and the proof is complete. □

Proof of Fact 7.13. Given probability spaces $(R, \mathcal{F}, \mu)$ and $(S, \mathcal{G}, \nu)$, suppose first that $f(x,y) = 1_E\big((x,y)\big)$ for some $E \in \mathcal{F} \otimes \mathcal{G}$. Then Lemma 11.31 implies

$$\begin{aligned}\int_{R\times S} f\, d(\mu \times \nu) = (\mu \times \nu)(E) &= \int_S \mu(E_y)\, d\nu(y) = \int_S \left(\int_R 1_B(x,y)\, d\mu(x) \right) d\nu(y) \\ &= \int_R \nu(E_x)\, d\mu(x) = \int_R \left(\int_S f(x,y)\, d\nu(y) \right) d\mu(x).\end{aligned}$$

This extends to simple functions by linearity, then nonnegative functions by monotone convergence. That proves Tonelli's theorem and Fubini follows by considering positive and negative parts. □

Kolmogorov Extension

Proof of Fact 7.14. Let $\{\mu_n\}_{n=1}^\infty$ be a consistent sequence of probability measures, let $\mathcal{S}$ be the semialgebra of cylinder sets, and define $P_0 : \mathcal{S} \to [0,1]$ by

$$P_0\left(\{\omega \in \mathbb{R}^{\mathbb{N}} : \omega_i \in (a_i, b_i], 1 \le i \le n\}\right) = \mu_n\left((a_1, b_1] \times \cdots \times (a_n, b_n]\right).$$

Next, let $\overline{\mathcal{S}}$ be the algebra generated by $\mathcal{S}$ and define $\widetilde{P} : \overline{\mathcal{S}} \to [0,1]$ by $\widetilde{P}\left(\bigsqcup_{k=1}^n S_k\right) = \sum_{k=1}^n P_0(S_k)$ for $S_1, ..., S_n$ disjoint sets in $\mathcal{S}$. It's easy to check that, P_0 and thus $\widetilde{P}$ is finitely additive.

As $\overline{\mathcal{S}}$ generates $\mathcal{B}^{\mathbb{N}}$, it suffices to show that $\widetilde{P}$ is countably additive and thus a premeasure on $\overline{\mathcal{S}}$. We do so by proving that if $\{B_n\}_{n=1}^\infty$ is a sequence of sets in $\overline{\mathcal{S}}$ with $B_n \searrow \emptyset$, then $\widetilde{P}(B_n) \searrow 0$.

Indeed if $\{A_i\}_{i=1}^\infty$ is a countable collection of disjoint sets in $\overline{\mathcal{S}}$ such that $A = \bigcup_{i=1}^\infty A_i \in \overline{\mathcal{S}}$, then for any $n \in \mathbb{N}$, $B_n = \bigcup_{i=n}^\infty A_i = A \backslash \bigcup_{i=1}^{n-1} A_i$ belongs to the algebra $\overline{\mathcal{S}}$, so finite additivity gives $\widetilde{P}(A) = \sum_{i=1}^{n-1} \widetilde{P}(A_i) + \widetilde{P}(B_n)$.

To further simplify our task, let $\mathcal{F}_n$ be the sub-σ-algebra of $\mathcal{B}^{\mathbb{N}}$ consisting of all sets of the form $E = E^* \times \mathbb{R} \times \mathbb{R} \times \cdots$ with $E^* \in \mathcal{B}^n$. We'll use this asterisk notation throughout to denote the '$\mathcal{B}^n$ component' of sets in $\mathcal{F}^n$.

We begin by showing that we may assume without loss of generality that $B_n \in \mathcal{F}_n$ for all n.

To see this, note that $B_n \in \overline{\mathcal{S}}$ implies that there is a $j(n) \in \mathbb{N}$ such that $B_n \in \mathcal{F}_k$ for all $k \ge j(n)$. Let $k(1) = j(1)$ and $k(n) = k(n-1) + j(n)$ for $n \ge 2$. Then $k(1) < k(2) < \cdots$ and $B_n \in \mathcal{F}_{k(n)}$ for all n. Define $\widetilde{B}_i = \mathbb{R}^{\mathbb{N}}$ for $i < k(1)$ and $\widetilde{B}_i = B_n$ for $k(n) \le i < k(n+1)$. Then $\widetilde{B}_n \in \mathcal{F}_n$ for all n and the collections $\{B_n\}$ and $\left\{\widetilde{B}_n\right\}$ differ only in that the latter possibly includes $\mathbb{R}^{\mathbb{N}}$ and repeats sets. The assertion follows since $\widetilde{B}_n \searrow \emptyset$ if and only if $B_n \searrow \emptyset$ and $\widetilde{P}\left(\widetilde{B}_n\right) \searrow 0$ if and only if $\widetilde{P}\left(B_n\right) \searrow 0$.

Now suppose that $\widetilde{P}(B_n) \ge \delta > 0$ for all n. (Since $\widetilde{P}$ is monotone and $B_n \searrow \emptyset$, this is equivalent to $\widetilde{P}(B_n) \not\to 0$.) We will derive a contradiction by approximating the B_n^* from within by compact sets and then using a diagonal argument to obtain $\bigcap_n B_n \neq \emptyset$.

Since B_n is nonempty and belongs to $\overline{\mathcal{S}} \cap \mathcal{F}_n$, we can write

$$B_n = \bigcup_{k=1}^{K(n)} \{\omega : \omega_i \in (a_{i,k}, b_{i,k}], i = 1, ..., n\} \text{ where } -\infty \leq a_{i,k} < b_{i,k} \leq \infty.$$

By a continuity from below argument, we can find a set $E_n \subseteq B_n$ of the form

$$E_n = \bigcup_{k=1}^{K(n)} \left\{\omega : \omega_i \in [\widetilde{a}_{i,k}, \widetilde{b}_{i,k}], i = 1, ..., n\right\}, \quad -\infty < \widetilde{a}_{i,k} < \widetilde{b}_{i,k} < \infty,$$

with $\mu_n\left(B_n^* \setminus E_n^*\right) \leq \frac{\delta}{2^{n+1}}$.

Let $F_n = \bigcap_{m=1}^n E_m$. Since $B_n \subseteq B_m$ for any $m \leq n$, we have

$$B_n \setminus F_n = B_n \cap \left(\bigcup_{m=1}^n E_m^C\right) = \bigcup_{m=1}^n \left(B_n \cap E_m^C\right) \subseteq \bigcup_{m=1}^n \left(B_m \cap E_m^C\right),$$

hence

$$\mu_n(B_n^* \setminus F_n^*) \leq \sum_{m=1}^n \mu_m(B_m^* \setminus E_m^*) \leq \frac{\delta}{2}.$$

As $\mu_n(B_n^*) = \widetilde{P}(B_n) \geq \delta$, this means that $\mu_n(F_n^*) \geq \frac{\delta}{2}$, hence F_n^* is nonempty.

Moreover, E_n^* is a finite union of closed and bounded rectangles, so

$$F_n^* = E_n^* \cap (E_{n-1}^* \times \mathbb{R}) \cap \cdots \cap (E_1 \times \mathbb{R}^{n-1})$$

is compact.

For each $m \in \mathbb{N}$, choose some $\omega^m \in F_m$. As $F_m \subseteq F_1$, ω_1^m (the first coordinate of ω^m) is in F_1^*.

By compactness, we can find a subsequence $m(1, j) \geq j$ such that $\omega_1^{m(1,j)}$ converges to a limit $\theta_1 \in F_1^*$.

For $m \geq 2$, $F_m \subseteq F_2$, so $(\omega_1^m, \omega_2^m) \in F_2^*$. Because F_2^* is compact, we can find a subsequence of $\{m(1, j)\}$, which we denote by $m(2, j)$, such that $\omega_2^{m(2,j)}$ converges to a limit θ_2 with $(\theta_1, \theta_2) \in F_2^*$.

In general, we can find a subsequence $m(n, j)$ of $m(n-1, j)$ such that $\omega_n^{m(n,j)}$ converges to θ_n with $(\theta_1, ..., \theta_n) \in F_n^*$.

Finally, define the sequence $\omega(i) = \omega^{m(i,i)}$. Then $\omega(i)$ is a subsequence of each $\omega^{m(i,j)}$, so $\lim_{i\to\infty} \omega(i)_k = \theta_k$ for all k. Since $(\theta_1, ..., \theta_n) \in F_n^*$ for all n, $\theta = (\theta_1, \theta_2, ...) \in F_n$ for all n, hence

$$\theta \in \bigcap_{n=1}^\infty F_n \subseteq \bigcap_{n=1}^\infty B_n,$$

a contradiction! □

Generalized WLLN

Theorem 11.32. *For each $n \in \mathbb{N}$, let $X_{n,1}, ..., X_{n,n}$ be independent. Let $\{b_n\}_{n=1}^\infty$ be a sequence of positive numbers with $\lim_{n\to\infty} b_n = \infty$ and let $\widetilde{X}_{n,k} = X_{n,k} 1\{|X_{n,k}| \leq b_n\}$. Suppose that as $n \to \infty$*

(1) $\sum_{k=1}^n P(|X_{n,k}| > b_n) \to 0$

(2) $b_n^{-2} \sum_{k=1}^n E\left[\widetilde{X}_{n,k}^2\right] \to 0.$

If we let $S_n = \sum_{k=1}^n X_{n,k}$ and $a_n = \sum_{k=1}^n E\left[\widetilde{X}_{n,k}\right]$, then $\frac{S_n - a_n}{b_n} \to_p 0$.

Proof. Let $\widetilde{S}_n = \sum_{k=1}^n \widetilde{X}_{n,k}$. By partitioning the event $\left\{\left|\frac{S_n - a_n}{b_n}\right| > \varepsilon\right\}$ according to whether or not $S_n = \widetilde{S}_n$, we see that

$$P\left(\left|\frac{S_n - a_n}{b_n}\right| > \varepsilon\right) \leq P(S_n \neq \widetilde{S}_n) + P\left(\left|\frac{\widetilde{S}_n - a_n}{b_n}\right| > \varepsilon\right).$$

To estimate the first term, we observe that

$$P(S_n \neq \widetilde{S}_n) \leq P\left(\bigcup_{k=1}^n \{X_{n,k} \neq \widetilde{X}_{n,k}\}\right) \leq \sum_{k=1}^n P\left(X_{n,k} \neq \widetilde{X}_{n,k}\right) = \sum_{k=1}^n P\left(|X_{n,k}| > b_n\right) \to 0.$$

For the second, we use Chebychev's inequality, $E\left[\widetilde{S}_n\right] = a_n$, the independence of the $\widetilde{X}_{n,k}$'s, and our second assumption to obtain

$$\begin{aligned} P\left(\left|\frac{\widetilde{S}_n - a_n}{b_n}\right| > \varepsilon\right) &\leq \varepsilon^{-2} E\left[\left(\frac{\widetilde{S}_n - a_n}{b_n}\right)^2\right] = \varepsilon^{-2} b_n^{-2} \mathrm{Var}(\widetilde{S}_n) \\ &= \varepsilon^{-2} b_n^{-2} \sum_{k=1}^n \mathrm{Var}\left[\widetilde{X}_{n,k}^2\right] \leq \varepsilon^{-2}\left(b_n^{-2} \sum_{k=1}^n E\left[\widetilde{X}_{n,k}^2\right]\right) \to 0. \end{aligned}$$

□

Theorem 11.32 was so easy to prove because we assumed exactly what we needed. Essentially, these are the correct hypotheses for the weak law, but they are a little clunky so we usually talk about special cases that take a nicer form.

To prove our weak law for i.i.d. sequences, we need the following simple generalization of Proposition 6.6.

Lemma 11.33 (Layer cake representation)**.** *If $Y \geq 0$ and $p > 0$, then*

$$E\left[Y^p\right] = \int_0^\infty p y^{p-1} P(Y > y)\, dy.$$

Proof. Tonelli's theorem gives

$$\begin{aligned} \int_0^\infty p y^{p-1} P(Y > y)\, dy &= \int_0^\infty p y^{p-1} \left(\int_\Omega 1\{Y > y\}\, dP\right) dy \\ &= \int_\Omega \left(\int_0^\infty p y^{p-1} 1\{y < Y\}\, dy\right) dP \\ &= \int_\Omega \left(\int_0^Y p y^{p-1}\, dy\right) dP = \int_\Omega Y^p dP = E\left[Y^p\right]. \end{aligned}$$

□

We now have all the necessary ingredients for a

Proof of Fact 8.13. We will apply Theorem 11.32 with $X_{n,k} = X_k$ and $b_n = n$ (hence $a_n = n\mu_n$).

The first assumption is satisfied since

$$\sum_{k=1}^n P\left(|X_{n,k}| > n\right) = nP(|X_1| > n) \to 0.$$

For the second assumption, we have $\widetilde{X}_{n,k} = X_k 1\{|X_k| \leq n\}$, so we must show that

$$\frac{1}{n} E\left[\widetilde{X}_{n,1}^2\right] = \frac{1}{n^2} \sum_{k=1}^n E\left[\widetilde{X}_{n,k}^2\right] \to 0.$$

Lemma 11.33 tells us that

$$E\left[\widetilde{X}_{n,1}^2\right] = \int_0^\infty 2yP\left(\left|\widetilde{X}_{n,1}\right| > y\right) dy \leq \int_0^n 2yP\left(|X_1| > y\right) dy$$

since $P\left(\left|\widetilde{X}_{n,1}\right| > y\right) = 0$ for $y > n$ and $P\left(\left|\widetilde{X}_{n,1}\right| > y\right) = P\left(|X_1| > y\right) - P\left(|X_1| > n\right)$ for $y \leq n$, so we will be done once we prove

$$\frac{1}{n}\int_0^n 2yP\left(|X_1| > y\right) dy \to 0.$$

To see that this is the case, note that since $2yP\left(|X_1| > y\right) \to 0$ as $y \to \infty$, for any $\varepsilon > 0$, there is an $N \in \mathbb{N}$ such that $2yP\left(|X_1| > y\right) < \varepsilon$ whenever $y \geq N$. Because $2yP\left(|X_1| > y\right) < 2N$ for $y < N$, we see that for all $n > N$,

$$\begin{aligned}\frac{1}{n}\int_0^n 2yP\left(|X_1| > y\right) dy &= \frac{1}{n}\int_0^N 2yP\left(|X_1| > y\right) dy + \frac{1}{n}\int_N^n 2yP\left(|X_1| > y\right) dy \\ &\leq \frac{1}{n}\int_0^N 2N dy + \frac{1}{n}\int_N^n \varepsilon dy = \frac{2N^2}{n} + \frac{n-N}{n}\varepsilon,\end{aligned}$$

hence

$$\limsup_{n\to\infty} \frac{1}{n}\int_0^n 2yP\left(|X_1| > y\right) dy \leq \limsup_{n\to\infty} \frac{2N^2}{n} + \frac{n-N}{n}\varepsilon = \varepsilon,$$

and the result follows since ε was arbitrary. □

Example 11.34 (The St. Petersburg Paradox)**.** Suppose that I offered to pay you 2^j dollars if it takes j flips of a fair coin for the first head to appear. That is, your winnings are given by the random variable X with $P(X = 2^j) = 2^{-j}$ for $j \in \mathbb{N}$. How much would you pay to play the game n times? The paradox is that $E[X] = \sum_{j=1}^\infty 2^j \cdot 2^{-j} = \infty$, but most sensible people would not pay anywhere near \$40 a game.
Using Theorem 11.32, we will show that a fair price for playing n times is $\$\log_2(n)$ per play, so that one would need to play about a trillion rounds to reasonably expect to break even at \$40 a play.

Proof. To cast this problem in terms of Theorem 11.32, we will take $X_1, X_2, ...$ to be independent random variables which are equal in distribution to X and set $X_{n,k} = X_k$. Then $S_n = \sum_{k=1}^n X_k$ denotes your total winnings after n games. We need to choose b_n so that

$$nP(X > b_n) = \sum_{k=1}^n P(X_{n,k} > b_n) \to 0,$$

$$\frac{n}{b_n^2} E\left[X^2 1\left\{X \leq b_n\right\}\right] = b_n^{-2} \sum_{k=1}^n E\left[\left(X_{n,k} 1\left\{|X_{n,k}| \leq b_n\right\}\right)^2\right] \to 0.$$

To this end, let $m(n) = \log_2(n) + K(n)$ where $K(n)$ is such that $m(n) \in \mathbb{N}$ and $K(n) \to \infty$ as $n \to \infty$.
If we set $b_n = 2^{m(n)} = n2^{K(n)}$, we have

$$nP(X > b_n) = n \sum_{i=m(n)+1}^\infty 2^{-i} = n2^{-m(n)} = 2^{-K(n)} \to 0$$

and

$$E\left[X^2 1\left\{X \leq b_n\right\}\right] = \sum_{i=1}^{m(n)} 2^{2i} \cdot 2^{-i} = 2^{m(n)+1} - 2 \leq 2b_n,$$

so that

$$\frac{n}{b_n^2} E\left[X^2 1\left\{|X| \leq b_n\right\}\right] \leq \frac{2n}{b_n} = 2^{-K(n)+1} \to 0.$$

Since

$$a_n = \sum_{k=1}^{n} E\left[X_{n,k} 1\left\{|X_{n,k}| \le b_n\right\}\right] = nE\left[X1\left\{X \le b_n\right\}\right] = n\sum_{i=1}^{m(n)} 2^i \cdot 2^{-i} = nm(n),$$

Theorem 11.32 gives

$$\frac{S_n - n\log_2(n) - nK(n)}{n2^{K(n)}} \to_p 0.$$

If we take $K(n) \le \log_2\left(\log_2(n)\right)$, then the conclusion holds with $n\log_2(n)$ in the denominator, so we get

$$\frac{S_n}{n\log_2(n)} \to_p 1. \qquad \square$$

SLLN

Theorem 11.35 (Strong Law of Large Numbers)**.** *Suppose that $X_1, X_2, \ldots$ are pairwise independent and identically distributed with $E|X_1| < \infty$. Let $S_n = \sum_{k=1}^{n} X_k$ and $\mu = E[X_1]$. Then $\frac{1}{n}S_n \to \mu$ almost surely as $n \to \infty$.*

Proof. We begin by noting that $X_k^+ = \max\{X_k, 0\}$ and $X_k^- = \max\{-X_k, 0\}$ satisfy the theorem's assumptions, so, since $X_k = X_k^+ - X_k^-$, we may suppose without loss of generality that the X_k's are nonnegative.

Next, we observe that it suffices to consider truncated versions of the X_k's:

Claim 11.36. If $Y_k = X_k 1\left\{X_k \le k\right\}$ and $T_n = \sum_{k=1}^{n} Y_k$, then $\frac{1}{n}T_n \to \mu$ a.s. implies $\frac{1}{n}S_n \to \mu$ a.s.

Proof. Lemma 11.33 and the fact that $G(t) = P\left(X_1 > t\right)$ is nonincreasing imply

$$\sum_{k=1}^{\infty} P(X_k \ne Y_k) = \sum_{k=1}^{\infty} P\left(X_k > k\right) = \sum_{k=1}^{\infty} P\left(X_1 > k\right) \le \int_0^{\infty} P\left(X_1 > t\right) dt = E\left|X_1\right| < \infty,$$

so the first Borel-Cantelli lemma gives $P(X_k \ne Y_k \text{ i.o.}) = 0$. Thus for all ω in a set of probability one, $\sup_n |S_n(\omega) - T_n(\omega)| < \infty$, hence $\dfrac{S_n}{n} - \dfrac{T_n}{n} \to 0$ a.s. and the claim follows. $\square$

The truncation step should not be too surprising as it is generally easier to work with bounded random variables. The reason that we reduced the problem to the $X_k \ge 0$ case is that this assures that the sequence $T_1, T_2, \ldots$ is nondecreasing.

Our strategy going forward will be to prove convergence along a cleverly chosen subsequence and then exploit monotonicity to handle intermediate values.

Specifically, for $\alpha > 1$, let $k(n) = \lfloor \alpha^n \rfloor$, the greatest integer less than or equal to α^n.

Chebychev's inequality and Tonelli's theorem give

$$\begin{aligned}\sum_{n=1}^{\infty} P\left(\left|T_{k(n)} - E\left[T_{k(n)}\right]\right| > \varepsilon k(n)\right) &\le \sum_{n=1}^{\infty} \frac{\operatorname{Var}\left(T_{k(n)}\right)}{\varepsilon^2 k(n)^2} = \varepsilon^{-2}\sum_{n=1}^{\infty} k(n)^{-2}\sum_{m=1}^{k(n)} \operatorname{Var}\left(Y_m\right) \\ &= \varepsilon^{-2}\sum_{m=1}^{\infty} \operatorname{Var}\left(Y_m\right) \sum_{n:k(n)\ge m} k(n)^{-2} \le \varepsilon^{-2}\sum_{m=1}^{\infty} E\left[Y_m^2\right] \sum_{n:\alpha^n \ge m} \lfloor \alpha^n \rfloor^{-2}.\end{aligned}$$

Since $\lfloor \alpha^n \rfloor \geq \frac{1}{2}\alpha^n$ for $n \geq 1$ (by casing out according to α^n smaller or bigger than 2),

$$\sum_{n:\alpha^n \geq m} \lfloor \alpha^n \rfloor^{-2} \leq 4 \sum_{n \geq \log_\alpha m} \alpha^{-2n} \leq 4\alpha^{-2\log_\alpha m} \sum_{n=0}^{\infty} \alpha^{-2n} = 4(1-\alpha^{-2})^{-1} m^{-2},$$

hence

$$\begin{aligned}\sum_{n=1}^{\infty} P\left(\left|T_{k(n)} - E\left[T_{k(n)}\right]\right| > \varepsilon k(n)\right) &\leq \varepsilon^{-2} \sum_{m=1}^{\infty} E\left[Y_m^2\right] \sum_{n:\alpha^n \geq m} [\alpha^n]^{-2} \\ &\leq 4(1-\alpha^{-2})^{-1}\varepsilon^{-2} \sum_{m=1}^{\infty} \frac{E\left[Y_m^2\right]}{m^2}.\end{aligned}$$

Claim 11.37. $\displaystyle\sum_{m=1}^{\infty} \frac{E[Y_m^2]}{m^2} < \infty.$

Proof. By Lemma 11.33,

$$E\left[Y_m^2\right] = \int_0^\infty 2yP(Y_m > y)dy = \int_0^m 2yP(Y_m > y)dy \leq \int_0^m 2yP(X_1 > y)dy,$$

so Tonelli's theorem gives

$$\sum_{m=1}^{\infty} \frac{E[Y_m^2]}{m^2} \leq \sum_{m=1}^{\infty} m^{-2} \int_0^m 2yP(X_1 > y)dy = 2\int_0^\infty \left(y \sum_{m>y} m^{-2}\right) P(X_1 > y)dy.$$

Since $\displaystyle\int_0^\infty P(X_1 > y)dy = E[X_1] < \infty$, we will be done if we can show that $\displaystyle y\sum_{m>y} m^{-2}$ is uniformly bounded. To see that this is the case, observe that

$$y\sum_{m>y} m^{-2} \leq \sum_{m=1}^{\infty} m^{-2} = \frac{\pi^2}{6} < 2$$

for $y \in [0,1]$, and for $j \geq 2$,

$$\sum_{m=j}^{\infty} m^{-2} \leq \int_{j-1}^{\infty} x^{-2}dx = (j-1)^{-1},$$

so

$$y\sum_{m>y} m^{-2} = y\sum_{m=\lfloor y \rfloor + 1}^{\infty} m^{-2} \leq \frac{y}{\lfloor y \rfloor} \leq 2$$

for $y > 1$. □

It follows that $\displaystyle\sum_{n=1}^{\infty} P\left(\left|T_{k(n)} - E\left[T_{k(n)}\right]\right| > \varepsilon k(n)\right) < \infty$, so, since $\varepsilon > 0$ is arbitrary, the first Borel-Cantelli lemma implies that $\dfrac{T_{k(n)} - E\left[T_{k(n)}\right]}{k(n)} \to 0$ a.s.

Now $\lim\limits_{k\to\infty} E[Y_k] = E[X_1]$ by the dominated convergence theorem, so $\lim_{n\to\infty} \dfrac{E\left[T_{k(n)}\right]}{k(n)} = E[X_1]$.

Thus we have shown that $\dfrac{T_{k(n)}}{k(n)} \to \mu$ almost surely.

Finally, if $k(n) \leq m < k(n+1)$, then

$$\frac{k(n)}{k(n+1)} \cdot \frac{T_{k(n)}}{k(n)} = \frac{T_{k(n)}}{k(n+1)} \leq \frac{T_m}{m} \leq \frac{T_{k(n+1)}}{k(n)} = \frac{T_{k(n+1)}}{k(n+1)} \cdot \frac{k(n+1)}{k(n)}$$

since T_n is nondecreasing.

Because $\dfrac{k(n+1)}{k(n)} = \dfrac{\lfloor \alpha^{n+1} \rfloor}{\lfloor \alpha^n \rfloor} \to \alpha$ as $n \to \infty$, we see that

$$\frac{\mu}{\alpha} \le \liminf_{n\to\infty} \frac{T_m}{m} \le \limsup_{n\to\infty} \frac{T_m}{m} \le \alpha\mu,$$

and we're done since $\alpha > 1$ is arbitrary. □

Stein Bounds

We begin with a bound on the *complementary error function*, $1 - \Phi(w) = \frac{1}{\sqrt{2\pi}} \int_w^\infty e^{-\frac{t^2}{2}}\, dt$.

Lemma 11.38. *For all $w > 0$,*

$$\frac{w}{w^2+1} e^{-\frac{w^2}{2}} \le \int_w^\infty e^{-\frac{t^2}{2}}\, dt \le \frac{1}{w} e^{-\frac{w^2}{2}}.$$

Proof. The upper bound follows by observing that

$$\int_w^\infty e^{-\frac{t^2}{2}} dt \le \int_w^\infty \frac{t}{w} e^{-\frac{t^2}{2}} dt = w^{-1} \int_{w^2/2}^\infty e^{-u}\, du = w^{-1} e^{-w^2/2}.$$

For the lower bound, define $g(w) = \int_w^\infty e^{-\frac{t^2}{2}}\, dt - \frac{w}{1+w^2} e^{-\frac{w^2}{2}}$. One easily checks that $g(0) = \sqrt{\pi/2}$, $g'(w) = -e^{-\frac{w^2}{2}} - \frac{1-2w^2-w^4}{1+2w^2+w^4} e^{-\frac{w^2}{2}} < 0$, and $\lim_{w\to\infty} g(w) = 0$, so it must be the case that $g(w) \ge 0$ for all $w \ge 0$. (Note that the lower bound holds trivially for $w \le 0$.) □

Proof of Fact 10.4. Define $\widetilde{h}(w) = h(w) - Nh$ and $c_0 = \left\|\widetilde{h}\right\|_\infty$, and let $c_1 = \|h'\|_\infty$ if h is absolutely continuous and $c_1 = \infty$ otherwise. Since $\widetilde{h}$ and f_h are unchanged if h is replaced by $h - h(0)$, we may assume that $h(0) = 0$. Thus $|h(t)| \le c_1 |t|$ and $|Nh| \le c_1 E|Z| \le c_1\sqrt{2/\pi}$.

We begin by bounding the sup norm of

$$\begin{aligned} f_h(w) &= e^{\frac{w^2}{2}} \int_{-\infty}^w \widetilde{h}(t) e^{-\frac{t^2}{2}}\, dt \\ &= -e^{\frac{w^2}{2}} \int_w^\infty \widetilde{h}(t) e^{-\frac{t^2}{2}}\, dt. \end{aligned}$$

Applying the upper bound from Lemma 11.38 shows that for all $w > 0$

$$\frac{d}{dw} e^{\frac{w^2}{2}} \int_w^\infty e^{-\frac{t^2}{2}}\, dt = w e^{\frac{w^2}{2}} \int_w^\infty e^{-\frac{t^2}{2}}\, dt - e^{\frac{w^2}{2}} e^{-\frac{w^2}{2}} \le 0.$$

It follows that $e^{\frac{w^2}{2}} \int_{|w|}^\infty e^{-\frac{t^2}{2}}\, dt$ is minimized at $w = 0$, in which case its value is $\int_0^\infty e^{-\frac{t^2}{2}}\, dt = \sqrt{\pi/2}$. Consequently,

$$\begin{aligned} |f_h(w)| &\le \begin{cases} e^{\frac{w^2}{2}} \int_{-\infty}^w \left|\widetilde{h}(t)\right| e^{-\frac{t^2}{2}}\, dt, & w \le 0 \\ e^{\frac{w^2}{2}} \int_w^\infty \left|\widetilde{h}(t)\right| e^{-\frac{t^2}{2}}\, dt, & w \ge 0 \end{cases} \\ &\le e^{\frac{w^2}{2}} \min\left\{ c_0 \int_{|w|}^\infty e^{-\frac{t^2}{2}}\, dt, c_1 \int_{|w|}^\infty \left(t + \sqrt{2/\pi}\right) e^{-\frac{t^2}{2}}\, dt \right\} \\ &\le \min\left\{ \sqrt{\pi/2} c_0, 2c_1 \right\}. \end{aligned}$$

Next we consider $f_h'(w) = wf_h(w) + \widetilde{h}(w)$. By our previous analysis, when $w \geq 0$, we have

$$\begin{aligned}|f_h'(w)| &\leq \left|\widetilde{h}(w)\right| + \left|we^{\frac{w^2}{2}} \int_w^\infty \widetilde{h}(t)e^{-\frac{t^2}{2}}\,dt\right| \\ &\leq c_0 + c_0 we^{\frac{w^2}{2}} \int_w^\infty e^{-\frac{t^2}{2}}\,dt \leq 2c_0.\end{aligned}$$

A nearly identical argument gives $|f_h'(w)| \leq 2c_0$ for $w < 0$, completing the proof of the claims for h bounded.

If h is absolutely continuous, then

$$\begin{aligned}h(t) - Nh &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty [h(t) - h(x)]e^{-\frac{x^2}{2}}\,dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t \int_x^t h'(u)e^{-\frac{x^2}{2}}\,dudx - \frac{1}{\sqrt{2\pi}} \int_t^\infty \int_t^x h'(u)e^{-\frac{x^2}{2}}\,dudx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t \int_{-\infty}^u h'(u)e^{-\frac{x^2}{2}}\,dxdu - \frac{1}{\sqrt{2\pi}} \int_t^\infty \int_u^\infty h'(u)e^{-\frac{x^2}{2}}\,dxdu \\ &= \int_{-\infty}^t h'(u)\Phi(u)\,du - \int_t^\infty h'(u)\big(1 - \Phi(u)\big)\,du,\end{aligned}$$

so

$$\begin{aligned}\frac{1}{\sqrt{2\pi}} \int_{-\infty}^w [h(t) - Nh]e^{-\frac{t^2}{2}}\,dt &= \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^w \int_{-\infty}^t h'(u)\Phi(u)e^{-\frac{t^2}{2}}\,dudt - \int_{-\infty}^w \int_t^\infty h'(u)\big(1 - \Phi(u)\big)e^{-\frac{t^2}{2}}\,dudt\right) \\ &= \int_{-\infty}^w h'(u)\Phi(u) \left(\frac{1}{\sqrt{2\pi}} \int_u^w e^{-\frac{t^2}{2}}\,dt\right) du - \int_{-\infty}^w h'(u)\big(1 - \Phi(u)\big) \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{t^2}{2}}\,dt\right) du \\ &\qquad - \int_w^\infty h'(u)\big(1 - \Phi(u)\big) \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^w e^{-\frac{t^2}{2}}\,dt\right) du \\ &= \int_{-\infty}^w h'(u)\Phi(u)[\Phi(w) - \Phi(u)]\,du - \int_{-\infty}^w h'(u)\big(1 - \Phi(u)\big)\Phi(u)\,du \\ &\qquad - \int_w^\infty h'(u)\big(1 - \Phi(u)\big)\Phi(w)\,du \\ &= -\int_{-\infty}^w h'(u)\Phi(u)\big(1 - \Phi(w)\big)\,du - \int_w^\infty h'(u)\big(1 - \Phi(u)\big)\Phi(w)\,du,\end{aligned}$$

hence

$$f_h(w) = -\sqrt{2\pi}e^{\frac{w^2}{2}} \left(\big(1 - \Phi(w)\big) \int_{-\infty}^w h'(u)\Phi(u)\,du + \Phi(w) \int_w^\infty h'(u)\big(1 - \Phi(u)\big)\,du\right).$$

Combining these representations yields

$$\begin{aligned}f_h'(w) &= wf_h(w) + h(w) - Nh \\ &= -\sqrt{2\pi}we^{\frac{w^2}{2}} \left(\big(1 - \Phi(w)\big) \int_{-\infty}^w h'(u)\Phi(u)\,du + \Phi(w) \int_w^\infty h'(u)\big(1 - \Phi(u)\big)\,du\right) \\ &\qquad + \int_{-\infty}^w h'(u)\Phi(u)\,du - \int_w^\infty h'(u)\big(1 - \Phi(u)\big)\,du \\ &= \left[1 - \sqrt{2\pi}we^{\frac{w^2}{2}}\big(1 - \Phi(w)\big)\right] \int_{-\infty}^w h'(u)\Phi(u)\,du \\ &\qquad - \left[1 + \sqrt{2\pi}we^{\frac{w^2}{2}}\Phi(w)\right] \int_w^\infty h'(u)\big(1 - \Phi(u)\big)\,du.\end{aligned}$$

Since integration by parts shows that

$$\int_{-\infty}^{w} \Phi(u)\,du = w\Phi(w) - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{w} ue^{-\frac{u^2}{2}}\,du = w\Phi(w) + \frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}},$$
$$\int_{w}^{\infty} [1-\Phi(u)]\,du = -w\big(1-\Phi(w)\big) + \int_{w}^{\infty} ue^{-\frac{u^2}{2}}\,du = -w\big(1-\Phi(w)\big) + \frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}},$$

we conclude that

$$|f_h'(w)| \le \|h'\|_\infty \sup_{w\in\mathbb{R}} \left[\left|1-\sqrt{2\pi}we^{\frac{w^2}{2}}\big(1-\Phi(w)\big)\right| \left(w\Phi(w) + \frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}}\right) \right.$$
$$\left. + \left|1+\sqrt{2\pi}we^{\frac{w^2}{2}}\Phi(w)\right| \left(\frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}} - w\big(1-\Phi(w)\big)+\right)\right].$$

By casing out according to the sign of w and differentiating, one can show that the term in the brackets attains its maximum value of $\sqrt{2/\pi}$ at $w=0$.

It remains only to derive the second derivative bound in the absolutely continuous case. For this, note that

$$\begin{aligned} f_h''(w) &= \frac{d}{dw}\Big[wf_h(w) + \widetilde{h}(w)\Big] \\ &= f_h(w) + wf_h'(w) + h'(w) \\ &= (1+w^2)f_h(w) + w\widetilde{h}(w) + h'(w). \end{aligned}$$

Using our previous estimates,

$$\begin{aligned} &(1+w^2)f_h(w) + w\widetilde{h}(w) \\ &\quad= -\sqrt{2\pi}e^{\frac{w^2}{2}}(1+w^2)\left(\big(1-\Phi(w)\big)\int_{-\infty}^{w} h'(u)\Phi(u)\,du + \Phi(w)\int_{w}^{\infty} h'(u)\big(1-\Phi(u)\big)\,du\right) \\ &\qquad + w\int_{-\infty}^{w} h'(u)\Phi(u)\,du - w\int_{w}^{\infty} h'(u)\big(1-\Phi(u)\big)\,du \\ &\quad= \Big[w - \sqrt{2\pi}e^{\frac{w^2}{2}}(1+w^2)\big(1-\Phi(w)\big)\Big]\int_{-\infty}^{w} h'(u)\Phi(u)\,du \\ &\qquad - \Big[w + \sqrt{2\pi}e^{\frac{w^2}{2}}(1+w^2)\Phi(w)\Big]\int_{w}^{\infty} h'(u)\big(1-\Phi(u)\big)\,du. \end{aligned}$$

Now the lower bound in Lemma 11.38 ensures that $w - \sqrt{2\pi}e^{\frac{w^2}{2}}(1+w^2)\big(1-\Phi(w)\big) \le 0$ for all w, and this in turn implies $w + \sqrt{2\pi}e^{\frac{w^2}{2}}(1+w^2)\Phi(w) \ge 0$ for all w.

The triangle inequality thus gives

$$\begin{aligned} \Big|(1+w^2)f_h(w) + w\widetilde{h}(w)\Big| &\le c_1\left|w - \sqrt{2\pi}e^{\frac{w^2}{2}}(1+w^2)\big(1-\Phi(w)\big)\right|\int_{-\infty}^{w}\Phi(u)\,du \\ &\qquad + c_1\left|w + \sqrt{2\pi}e^{\frac{w^2}{2}}(1+w^2)\Phi(w)\right|\int_{w}^{\infty}\big(1-\Phi(u)\big)\,du \\ &= c_1\left(-w + \sqrt{2\pi}e^{\frac{w^2}{2}}(1+w^2)\big(1-\Phi(w)\big)\right)\left(w\Phi(w) + \frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}}\right) \\ &\qquad + c_1\left(w + \sqrt{2\pi}e^{\frac{w^2}{2}}(1+w^2)\Phi(w)\right)\left(-w\big(1-\Phi(w)\big) + \frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}}\right) \\ &= c_1\left(-w + \sqrt{2\pi}e^{\frac{w^2}{2}}(1+w^2)\big(1-\Phi(w)\big)\right)\left(w\Phi(w) + \frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}}\right) \\ &\qquad + c_1\left(w + \sqrt{2\pi}e^{\frac{w^2}{2}}(1+w^2)\Phi(w)\right)\left(-w\big(1-\Phi(w)\big) + \frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}}\right). \end{aligned}$$

A little arithmetic shows that this final expression is equal to c_1, so

$$|f_h''(w)| \leq \left|(1+w^2)f_h(w) + w\widetilde{h}(w)\right| + |h(w)| \leq 2c_0$$

for all w as desired. $\square$