MATH 507A, GRADUATE PROBABILITY I, FALL 2020

STEVEN HEILMAN

Abstract. These notes are based upon those of T. Tao, available here, R. Durrett's book, available here, and A. Dembo's notes, available here.

Contents

Date: March 3, 2022

1. Review of Measure Theory

1.1. **Measurable Spaces.** Measure theory is the foundation of probability theory. Probability theory deals with measures such that the measure of the universal set is 1. Measure theory then allows us to assign probabilities to events that could possibly occur, so that the probability of a set is the measure of that set. Unfortunately, it is impossible in general to have a measure on all subsets of a set. For this reason, measure theory is a nontrivial endeavor.

Let us recall some definitions from measure theory. An **algebra of sets** $\mathcal{F}$ is a set of subsets of a set Ω such that $\emptyset, \Omega \in \mathcal{F}$ and such that $\mathcal{F}$ is closed under finite union and complement. That is, (i) if $A, B \in \mathcal{F}$, then $A \cup B \in \mathcal{F}$ and (ii) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$. One can check that $\mathcal{F}$ is then closed with respect to finite intersection (since $A \cap B = (A^c \cup B^c)^c$, difference (since $A \setminus B = A \cap B^c$), and symmetric difference (since$A \Delta B = (A \setminus B) \cup (B \setminus A)$.

Definition 1.1 (**σ-algebra, Measurable space**)**.** A **σ-algebra** $\mathcal{F}$ in Ω (or **σ-field**) is an algebra closed under countable union. So, if $A_1, A_2, \ldots \in \mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$. A **measurable space** is a pair $(\Omega, \mathcal{F})$ where Ω is a set and $\mathcal{F}$ is a σ-algebra in Ω. Elements of $\mathcal{F}$ are called **measurable sets** in the measurable space $(\Omega, \mathcal{F})$.

If $\mathcal{F}, \mathcal{F}'$ are two σ-algebras in $\mathcal{F}$, we say that $\mathcal{F}$ is **coarser** than $\mathcal{F}'$ (or $\mathcal{F}'$ is **finer** than $\mathcal{F}$) if $\mathcal{F} \subseteq \mathcal{F}'$, so that every set in $\mathcal{F}$ is in $\mathcal{F}'$.

Remark 1.2. From De Morgan's Laws, if $\mathcal{F}$ is a σ-algebra, and if $A_1, A_2, \ldots \in \mathcal{F}$, then $\cap_{i=1}^{\infty} A_i \in \mathcal{F}$.

In probability theory, the σ-algebra represents all things that could possibly happen. For this reason, sets in $\mathcal{F}$ are called **events**. And below we will assign probabilities to events.

Example 1.3. If Ω is any set then $\{\emptyset, \Omega\}$ is a σ-algebra. And $\{\emptyset, \Omega\}$ is the coarsest σ-algebra on Ω.

Example 1.4. If Ω is any set then $2^{\Omega} = \{A \colon A \subseteq \Omega\}$ is a σ-algebra. And 2^{Ω} is the finest σ-algebra on Ω. We sometimes call 2^{Ω} the **discrete σ-algebra**.

Example 1.5 (**Generated σ-algebra**)**.** It follows from Definition 1.1 that if $\{\mathcal{F}_i\}_{i \in I}$ is a collection of σ-algebras on Ω, then $\cap_{i \in I} \mathcal{F}_i$ is also a σ-algebra on Ω. If $\mathcal{A}$ is a collection of subsets of Ω, we then define the σ-algebra **generated by** $\mathcal{A}$, denoted $\sigma(\mathcal{A})$, to be the intersection of all σ-algebras containing $\mathcal{A}$. (This intersection is nonempty, since 2^{Ω} contains $\mathcal{A}$.) Equivalently, $\sigma(\mathcal{A})$ is the coarsest σ-algebra such that every set in $\mathcal{A}$ is measurable.

Example 1.6 (**Borel σ-algebra**)**.** Let $n \geq 1$ and let $\Omega := \mathbb{R}^n$ or $\Omega := \mathbb{C}^n$. The **Borel σ-algebra** is the σ-algebra generated by the open sets of Ω. Measurable subsets in the Borel σ-algebra are called **Borel sets**. Unfortunately, some subsets of $\mathbb{R}^n$ are not Borel sets.

More generally, we could define the Borel sets on any any locally compact, Hausdorff, σ-compact topological space Ω. (A σ-compact space Ω can be written as a countable union of compact sets.)

Exercise 1.7. This exercises gives a strategy for proving a property for a generated σ-algebra.

Let $\mathcal{A}$ be a collection of subsets of a set Ω, and let $p(A)$ be a property of subsets A of Ω, so that $p(A)$ is true or false for each $A \in \Omega$. Assume the following:

- $p(\emptyset)$ is true.
- $p(A)$ is true for all $A \in \mathcal{A}$.
- If $A \subseteq \Omega$ is such that $p(A)$ is true, then $p(A^c)$ is also true.
- If $A_1, A_2, \ldots \subseteq \Omega$ are such that $p(A_i)$ is true for all $i \geq 1$, then $p(\cup_{i=1}^{\infty} A_i)$ is true.

Show that $p(A)$ is true for all $A \in \sigma(\mathcal{A})$. (Hint: what can one say about $\{A \subseteq \Omega : p(A) \text{ is true}\}$?

Example 1.8 (**Product σ-algebra**)**.** Let $(\Omega_i, \mathcal{F}_i)_{i \in I}$ be a collection of measurable spaces. We define the **product σ-algeba**, denoted $\prod_{i \in I} \mathcal{F}_i$, on the product space $\prod_{i \in I} \Omega_i$, to be the σ-algebra generated by the basic cylinder sets. If $j \in I$ is fixed, then a basic cylinder set is a set of the form

$$\{(x_i)_{i \in I} \in \prod_{i \in I} \Omega_i \colon x_j \in A_j \text{ for some fixed } A_j \in \mathcal{F}_j\}.$$

So, if $I = \{1, 2\}$, then $\mathcal{F}_1 \times \mathcal{F}_2$ is generated by sets of the form $A_1 \times \Omega_2$ and $\Omega_1 \times A_2$ where $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$.

It turns out that $\mathcal{F}_1 \times \mathcal{F}_2$ is also generated by the sets $A_1 \times A_2$ where $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$, but this is no longer true when I is uncountable.

Exercise 1.9. Let $n \geq 1$. Show that the Borel σ-algebra on $\mathbb{R}^n$ is generated by sets of the form $A_1 \times \cdots \times A_n$ where $A_i \subseteq \mathbb{R}$ is a Borel set for every $1 \leq i \leq n$.

1.2. **Measurable Functions.**

Definition 1.10 (**Random Variable**)**.** A function $X \colon \Omega \to S$ between two measurable spaces $(\Omega, \mathcal{F})$, $(S, \mathcal{B})$ is said to be **measurable** if $X^{-1}(B) \in \mathcal{F}$ for every $B \in \mathcal{B}$. (Recall that $X^{-1}(B) = \{\omega \in \Omega \colon X(\omega) \in B\}$.) A measurable function X is sometimes called a **random variable**.

If $n \geq 1$ and if $(S, \mathcal{B})$ is $\mathbb{R}^n$ with the Borel σ-algebra, then X is sometimes called a **random vector**. Some authors refer to a **random variable** only as a function $X \colon \Omega \to [-\infty, \infty]$, where $[-\infty, \infty]$ has the Borel σ-algebra.

Remark 1.11. The composition of measurable functions is measurable.

Exercise 1.12. Let $n, m \geq 1$. Let $f \colon \mathbb{R}^n \to \mathbb{R}^m$ be a continuous function. Show that f is measurable (if $\mathbb{R}^n, \mathbb{R}^m$ each have the Borel σ-algebra.)

Exercise 1.13. Let $X_1 \colon \Omega \to S_1, \ldots, X_n \colon \Omega \to S_n$ be measurable functions. Show that the joint function $(X_1, \ldots, X_n) \colon \Omega \to S_1 \times \cdots \times S_n$ defined by

$$(X_1, \ldots, X_n)(\omega) := (X_1(\omega), \ldots, X_n(\omega)), \qquad \forall\, \omega \in \Omega$$

is measurable.

Remark 1.14. So, if $F \colon S_1 \times \cdots \times S_n \to S$ is measurable, then $F(X_1, \ldots, X_n)$ is measurable. In particular, if X_1, X_2 are two real-valued random variables, then $X_1 + X_2$ is a random variable, $X_1 \cdot X_2$ is a random variable, etc.

1.3. **Measures.**

Definition 1.15 (**Probability Space**)**.** Let $(\Omega, \mathcal{F})$ be a measurable space. A **measure** μ is a function $\mu \colon \mathcal{F} \to [0, \infty]$ such that

- $\mu(\emptyset) = 0$.
- If $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint, then

$$\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i).$$

A **probability measure** on Ω is a measure μ such that $\mu(\Omega) = 1$. A **probability space** is a triple $(\Omega, \mathcal{F}, \mu)$ where μ is a probability measure on the measurable space $(\Omega, \mathcal{F})$.

Some property $p(\omega)$ that holds for all $\omega \in \Omega$ except for a set of measure zero is said to hold **almost everywhere** (abbreviated a.e.) or **almost surely** (abbreviate a.s.) or **for almost every** ω.

We will typically use the notation $\mathbf{P}$ for a probability measure.

Example 1.16 (**Discrete Probability Space**)**.** Let Ω be a finite or countably infinite set. For any $\omega \in \Omega$ let $a_\omega \geq 0$. Assume that $\sum_{\omega \in \Omega} a_\omega = 1$. For any $A \subseteq \Omega$, define

$$\mathbf{P}(A) := \sum_{\omega \in A} a_\omega.$$

Then $\mathbf{P}$ is a probability measure on $(\Omega, 2^\Omega)$.

Example 1.17 (**Lebesgue Measure**)**.** Let $\Omega = \mathbb{R}$. Then there exists a unique measure m on the Borel σ-algebra of $\mathbb{R}$ such that $m([a, b]) = b - a$ for every $-\infty \leq a \leq b \leq \infty$. This measure m is called **Lebesgue measure** (restricted to the Borel σ-algebra). (Recall that Lebesgue measure is typically defined on a σ-algebra that is larger than the Borel σ-algebra.)

More generally, for any $n \geq 1$, there exists a unique measure m_n on the Borel σ-algebra of $\mathbb{R}^n$ such that

$$m_n([a_1, b_1] \times \cdots \times [a_n, b_n]) = (b_1 - a_1) \cdots (b_n - a_n),$$

for all $-\infty \leq a_1 \leq b_1 \leq \infty, \ldots, -\infty \leq a_n \leq b_n \leq \infty$. Lebesgue measure and other commonly encountered measures are proven to exist by the following Theorem.

Recall that a nonnegative completely additive set function $\nu \colon \mathcal{F} \to [0, \infty]$ on an algebra $\mathcal{F}$ of sets satisfies $\nu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \nu(A_i)$ whenever $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint and $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$. Also we say ν is σ-finite if Ω satisfies $\Omega = \cup_{i=1}^{\infty} A_i$ for some $A_1, A_2, \ldots \in \mathcal{F}$ such that $\nu(A_i) < \infty$ for all $i \geq 1$.

Theorem 1.18 (**Carathéodory Extension Theorem**). *Let $\mathcal{F}$ be an algebra of subsets of a nonempty set Ω, and let ν be a nonnegative completely additive set function on $\mathcal{F}$ that is σ-finite. Then there exists a measure μ on $\sigma(\mathcal{F})$ such that $\mu(A) = \nu(A)$ for all $A \in \mathcal{F}$.*

Proof Sketch. Let $\mathcal{F}$ be an algebra, ν a nonnegative completely additive set function on $\mathcal{F}$ with $\nu(\Omega) < \infty$. Let $\mathcal{U}$ be the class of all countable unions of sets in $\mathcal{F}$, and let $\mathcal{K}$ be the class of countable intersections of sets in $\mathcal{F}$. We want to define an "outer measure" and "inner measure" from ν on subsets of Ω. This can be done unambiguously for any $E \subseteq \Omega$:

$$\mu^*(E) := \inf_{U \supseteq E, U \in \mathcal{U}} \mu^*(U) \qquad \text{and} \qquad \mu_*(E) := \sup_{K \subseteq E, K \in \mathcal{K}} \mu_*(K),$$

where $\mu^*(U) := \lim_{n\to\infty} \nu(A_n)$, where $\cup_{n=1}^{\infty} A_n = U$ and $A_1 \subseteq A_2 \subseteq \cdots$, $A_n \in \mathcal{F}\ \forall\ n \geq 1$, and $\mu_*(K) := \lim_{n\to\infty} \nu(C_n)$ where $\cap_{n=1}^{\infty} C_n = K$ and $C_1 \supseteq C_2 \supseteq \cdots$, $C_n \in \mathcal{F}\ \forall\ n \geq 1$.

We then define $\mathcal{B}$ to be the set of subsets where μ^* and μ_* agree (in general, we only have $\mu_*(E) \leq \mu^*(E)$ for a subset $E \subseteq \Omega$). It turns out that $\mathcal{B}$ is a σ-algebra, containing $\mathcal{U}$ and $\mathcal{K}$ (and $\mathcal{F}$). Moreover, μ^* is a measure on $\mathcal{B}$. Therefore, $\mathcal{B}$ contains the smallest σ-algebra $\sigma(\mathcal{F})$ containing $\mathcal{F}$, and μ^* is a measure on $\sigma(\mathcal{F})$. This proves existence. Uniqueness follows since any other extension μ' has to agree with μ^* on $\mathcal{U}$ and $\mathcal{K}$ (by definition of $\mathcal{U}, \mathcal{K}$). Thus, for any $E \in \mathcal{B}$ and for any $K \subseteq E \subseteq U$ such that $K \in \mathcal{K}$ and $U \in \mathcal{U}$, we have

$$\mu_*(K) = \mu'(K) \leq \mu'(E) \leq \mu'(U) = \mu^*(U).$$

So, taking $\sup_{K \in \mathcal{K}:\ K \subseteq E}$ and $\inf_{U \in \mathcal{U}:\ E \subseteq U}$ (and using the definition of $\mathcal{B} \supseteq \sigma(\mathcal{F})$) gives our desired result (with our special hypothesis $\nu(\Omega) < \infty$).

To handle the σ-finite case, one writes $\Omega = \cup_{i=1}^{\infty} A_i$ with $\nu(A_i) < \infty$ and $A_i \in \mathcal{F}$ for all $i \geq 1$. The above results applies to each of $A_1, A_2, \ldots$, and a "piecing together" argument concludes the proof. □

Exercise 1.19. Let μ be a measure on a measurable space $(\Omega, \mathcal{F})$. Using the axioms for a measure, show:

- (Monotonicity) If $A \subseteq B$ are measurable, then $\mu(A) \leq \mu(B)$.
- (Subadditivity) If $A_1, A_2, \ldots$ are measurable (but not necessarily disjoint), then
$$\mu(\cup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \mu(A_n).$$
- (Continuity from below) If $A_1 \subseteq A_2 \subseteq \cdots$ are measurable, then $\mu(\cup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \mu(A_n)$.
- (Continuity from above) If $A_1 \supseteq A_2 \supseteq \cdots$ are measurable and if $\mu(A_1) < \infty$, then $\mu(\cap_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \mu(A_n)$. Then, find a measurable space $(\Omega, \mathcal{F})$ and measurable subsets $B_1 \supseteq B_2 \supseteq \cdots$ such that $\mu(\cap_{n=1}^{\infty} B_n) \neq \lim_{n\to\infty} \mu(B_n)$.

Exercise 1.20. Let $(\Omega, \mathcal{F})$ be a measurable space. Let $[-\infty, \infty]$ have the Borel σ-algebra.

- Let $X \colon \Omega \to [-\infty, \infty]$. Show that X is measurable if and only if the sets $\{\omega \in \Omega \colon X(\omega) \leq t\}$ are measurable for all $t \in [-\infty, \infty]$.
- Let $X, Y \colon \Omega \to [-\infty, \infty]$. Show that $X = Y$ if and only if $\{\omega \in \Omega \colon X(\omega) \leq t\} = \{\omega \in \Omega \colon Y(\omega) \leq t\}$ for all $t \in [-\infty, \infty]$.
- Let $X_1, X_2, \ldots \colon \Omega \to [-\infty, \infty]$ be measurable. Show that $\sup_{n\geq 1} X_n, \inf_{n\geq 1} X_n$, $\limsup_{n\to\infty} X_n$, and $\liminf_{n\to\infty} X_n$ are all measurable.

Definition 1.21 (**Almost Sure Convergence**). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Let $X_1, X_2, \ldots : \Omega \to [-\infty, \infty]$ be measurable. We say that $X_1, X_2, \ldots$ **converges almost surely** if

$$\liminf_{n\to\infty} X_n = \limsup_{n\to\infty} X_n$$

almost surely. In this case, we define $\lim_{n\to\infty} X_n$ by

$$\lim_{n\to\infty} X_n := \liminf_{n\to\infty} X_n = \limsup_{n\to\infty} X_n.$$

Note that $\lim_{n\to\infty} X_n$ is well defined except on a set of measure 0.

Exercise 1.22. Let μ be a probability measure on $\mathbb{R}$, where $\mathbb{R}$ has the Borel σ-algebra. (Then μ is a Stieltjes measure.) Define the **distribution function** $F\colon \mathbb{R} \to [0,1]$ associated to μ by

$$F(t) := \mu((-\infty, t]) = \mu(\{x \in \mathbb{R}\colon\ -\infty < x \le t\}), \qquad \forall\, t \in \mathbb{R}.$$

Show the following properties of F:

- F is nondecreasing.
- $\lim_{t\to-\infty} F(t) = 0$ and $\lim_{t\to\infty} F(t) = 1$.
- F is right continuous, i.e. $F(t) = \lim_{s\to t^+} F(s)$ for all $t \in \mathbb{R}$.

Remark 1.23. The converse of Exercise 1.22 holds. That is if $F\colon \mathbb{R} \to [0,1]$ satisfies the three properties from Exercise 1.22, then there exists a unique probability measure μ on $\mathbb{R}$ such that F is the distribution function of μ.

If a distribution function is given, a random variable exists with that given distribution function.

Corollary 1.24 (**Construction of a Random Variable**). *Let $F\colon \mathbb{R} \to [0,1]$ satisfy the three properties from Exercise 1.22. Then there exists a random variable X on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ such that*

$$F(t) = \mathbf{P}(X \le t) = \mathbf{P}(\{\omega \in \Omega\colon X(\omega) \le t\}), \qquad \forall\, t \in \mathbb{R}.$$

Proof. Use $\Omega = \mathbb{R}$ with the Borel σ-algebra, and let $\mathbf{P}$ be the probability measure on $\mathbb{R}$ associated to F known to exist by Remark 1.23. Let $X\colon \mathbb{R} \to \mathbb{R}$ so that $X(t) = t$ for all $t \in \mathbb{R}$. Then by the definitions of X and F, $\mathbf{P}(X \le t) = \mathbf{P}((-\infty, t]) = F(t)$, $\forall\, t \in \mathbb{R}$. □

Exercise 1.25. Let $F\colon \mathbb{R} \to [0,1]$ satisfy the three properties from Exercise 1.22. Show that there exists a random variable X on $(0,1)$ with the Borel σ-algebra such that

$$F(t) = \mathbf{P}(X \le t), \qquad \forall\, t \in \mathbb{R}.$$

Here $\mathbf{P}$ is Lebesgue measure on $(0,1)$. (Hint: consider $X(t) := \sup\{y \in \mathbb{R}\colon F(y) < t\}$. Then X is an inverse of F.)

Definition 1.26 (**Cumulative Distribution Function**). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Let $X\colon \Omega \to \mathbb{R}$ be a random variable. The **cumulative distribution function** of X, denoted $F\colon \mathbb{R} \to [0,1]$, is the function

$$F(t) := \mathbf{P}(X \le t), \qquad \forall\, t \in \mathbb{R}.$$

In the more general setting $X\colon \Omega \to S$, the **distribution** of X (or the **law** of X) is the probability measure μ_X defined for any measurable $A \subseteq S$ by

$$\mu_X(A) := \mathbf{P}(X \in A) = \mathbf{P}(\{\omega \in \Omega\colon X(\omega) \in A\}).$$

From Exercise 1.22 applied to μ_X when $X\colon \Omega \to \mathbb{R}$, the cumulative distribution function of X satisfies the three properties of Exercise 1.22.

In many interesting cases, the distribution of X is absolutely continuous with respect to Lebesgue measure. In this case, by Lebesgue's Fundamental Theorem of Calculus, there exists a Lebesgue integrable function $f\colon \mathbb{R} \to [0,\infty)$, called the **density function** of X, such that $F(t) = \int_{-\infty}^{t} f(x)dx$ for all $t \in \mathbb{R}$. Moreover, f is the derivative of F almost everywhere (with respect to Lebesgue measure on $\mathbb{R}$). By Exercise 1.22, $\int_{-\infty}^{\infty} f(x)dx = 1$.

Example 1.27. Let $-\infty < a < b < \infty$. We say that X is **uniformly distributed** in $[a,b]$ if X has density function f such that $f(x) = 1/(b-a)$ for all $x \in [a,b]$ and $f(x) = 0$ otherwise. Then X has the following cumulative distribution function.

$$F(t) = \begin{cases} 0 & \text{, if } t < a \\ (t-a)/(b-a) & \text{, if } a \le t \le b \\ 1 & \text{, if } t > b. \end{cases}$$

Example 1.28. We say that X is a **standard Gaussian** (or standard normal) if X has density $f(x) = e^{-x^2/2}/\sqrt{2\pi}$, $\forall\, x \in \mathbb{R}$.

Definition 1.29. We say two random variables $X\colon \Omega \to \mathbb{R}, Y\colon S \to \mathbb{R}$ are **equal in distribution**, denoted $X \overset{d}{=} Y$, if X, Y have the same cumulative distribution function.

Exercise 1.30. Let X be a random variable with cumulative distribution function $F\colon \mathbb{R} \to [0,1]$. Show:

- $\mathbf{P}(X < t) = \lim_{s \to t^-} F(s)$.
- $\mathbf{P}(X = t) = F(t) - \lim_{s \to t^-} F(s)$.

So, $\mathbf{P}(X = t) = 0$ for all $t \in \mathbb{R}$ if and only if F is continuous.

Example 1.31. Let X be a random variable such that $\mathbf{P}(X = 0) = 1$. Then $F(t) = 0$ when $t < 0$ and $F(t) = 1$ when $t \ge 0$.

Exercise 1.32. Let μ be a probability measure on $\mathbb{R}^n$, where $\mathbb{R}^n$ has the Borel σ-algebra. Define the **distribution function** $F\colon \mathbb{R}^n \to [0,1]$ associated to μ by

$$\begin{aligned} F(t_1,\dots,t_n) &:= \mu((-\infty,t_1] \times \cdots \times (-\infty,t_n]) \\ &= \mu(\{(x_1,\dots,x_n) \in \mathbb{R}^n\colon\ -\infty < x_i \le t_i,\ \forall\, 1 \le i \le n\}), \qquad \forall\, t_1,\dots,t_n \in \mathbb{R}. \end{aligned}$$

Show the following properties of F:

- F is nondecreasing. ($F(t_1,\dots,t_n) \le F(t_1',\dots,t_n')$ whenever $t_i \le t_i'$ $\forall\ 1 \le i \le n$.)
- $\lim_{t_1,\dots,t_n \to -\infty} F(t_1,\dots,t_n) = 0$ and $\lim_{t_1,\dots,t_n \to \infty} F(t_1,\dots,t_n) = 1$.
- F is right continuous, i.e. $F(t_1,\dots,t_n) = \lim_{(s_1,\dots,s_n) \to (t_1,\dots,t_n)^+} F(s_1,\dots,s_n)$ for all $t_1,\dots,t_n \in \mathbb{R}$, where the limit restricts that $s_i \ge t_i$ $\forall\ 1 \le i \le n$.
- If $t_{i,0} \le t_{i,1}$ $\forall\ 1 \le i \le n$, then

$$\sum_{(\omega_1,\dots,\omega_n) \in \{0,1\}^n} (-1)^{\omega_1+\cdots+\omega_n+n} F(t_{1,\omega_1},\dots,t_{n,\omega_n}) \ge 0.$$

Remark 1.33. The converse of Exercise 1.32 holds. That is if $F\colon \mathbb{R}^n \to [0,1]$ satisfies the four properties from Exercise 1.32, then there exists a probability measure μ on $\mathbb{R}^n$ such that F is the distribution function of μ.

Corollary 1.34 (**Construction of Several Random Variables**). *Let* $F\colon \mathbb{R}^n \to [0,1]$ *satisfy the four properties from Exercise 1.32. Then there exist random variables* $X_1, \ldots, X_n$ *on a measurable space* $(\Omega, \mathcal{F})$ *such that*

$$F(t_1, \ldots, t_n) = \mathbf{P}(X_1 \leq t_1, \ldots, X_n \leq t_n), \qquad \forall\, t_1, \ldots, t_n \in \mathbb{R}.$$

1.4. **Expected Value, Integration.** If $A \subseteq \Omega$ is a measurable subset of a measurable space, we define the **indicator function** of A, denoted $1_A\colon \Omega \to \{0,1\}$ by

$$1_A(\omega) := \begin{cases} 1 & \text{, if } \omega \in A \\ 0 & \text{, if } \omega \notin A. \end{cases}$$

Definition 1.35 (**Expected Value**). Let $X\colon \Omega \to [0,\infty]$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. We say X is an unsigned **simple function** if $\exists\ m \geq 1$, $\exists\ a_1, \ldots, a_m \in [0,\infty]$ and $\exists$ disjoint measurable $A_1, \ldots, A_m \subseteq \Omega$ such that $X = \sum_{i=1}^m a_i 1_{A_i}$. We define the **expected value** of X (or the **mean** of X), denoted $\mathbf{E}X$, by

$$\mathbf{E}X := \sum_{i=1}^m a_i \mathbf{P}(A_i).$$

If $X\colon \Omega \to [0,\infty]$ is any random variable, we define the expected value of X to be

$$\mathbf{E}X := \sup_{m \geq 1} \sup_{\substack{0 \leq Y(\omega) \leq X(\omega),\, \forall\, \omega \in \Omega \\ Y = \sum_{i=1}^m a_i 1_{A_i} \text{ unsigned simple}}} \mathbf{E}Y.$$

If $X\colon \Omega \to [-\infty,\infty]$ satisfies $\mathbf{E}\,|X| < \infty$, we define the expected value of X to be

$$\mathbf{E}X := \mathbf{E}\max(X,0) - \mathbf{E}\max(-X,0).$$

If $X\colon \Omega \to \mathbb{C}$ satisfies $\mathbf{E}\,|X| < \infty$, we define the expected value of X to be

$$\mathbf{E}X := \mathbf{E}\mathrm{Re}(X) + \sqrt{-1}\mathbf{E}\mathrm{Im}(X).$$

If $\mathbf{E}\,|X| < \infty$, we say that X is **absolutely integrable**. When $\mathbf{E}X$ exists we sometimes write $\mathbf{E}X = \int_\Omega X d\mathbf{P}$ to match analytic notation and emphasize dependence of $\mathbf{E}X$ on $\mathbf{P}$.

Exercise 1.36. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a finite or countable probability space. If $X\colon \Omega \to [0,\infty]$ is a random variable with $\mathbf{E}\,|X| < \infty$, show that

$$\mathbf{E}X = \sum_{\omega \in \Omega} X(\omega)\mathbf{P}(\omega).$$

Exercise 1.37. Let X, Y be random variables such that $X, Y \geq 0$ or $\mathbf{E}\,|X|, \mathbf{E}\,|Y| < \infty$. For any $a \in [-\infty,\infty]$, define $0 \cdot a := 0$. Show:

- $\mathbf{E}(X+Y) = \mathbf{E}X + \mathbf{E}Y$ and if $c \in \mathbb{R}$, then $\mathbf{E}(cX) = c\mathbf{E}X$.
- If $\mathbf{P}(X = Y) = 1$, then $\mathbf{E}X = \mathbf{E}Y$.
- $\mathbf{E}\,|X| \geq 0$ with equality only when $X = 0$ almost surely.
- If $X \leq Y$ almost surely, then $\mathbf{E}X \leq \mathbf{E}Y$.

Exercise 1.38 (**Inclusion-Exclusion Formula**). Let $A_1, \ldots, A_n \subseteq \Omega$ be events. Then:

$$\mathbf{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathbf{P}(A_i) - \sum_{1 \leq i < j \leq n} \mathbf{P}(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n} \mathbf{P}(A_i \cap A_j \cap A_k)$$
$$\cdots + (-1)^{n+1} \mathbf{P}(A_1 \cap \cdots \cap A_n).$$

To prove this formula, show that $1_{\cup_{i=1}^n A_i} = 1 - \prod_{i=1}^n (1 - 1_{A_i})$ and then take expected values of both sides.

1.5. **Inequalities.**

Exercise 1.39. Let $\phi\colon \mathbb{R} \to \mathbb{R}$. We say that ϕ is **convex** if, for any $x, y \in \mathbb{R}$ and for any $t \in [0,1]$, we have

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y).$$

Let $\phi\colon \mathbb{R} \to \mathbb{R}$. Show that ϕ is convex if and only if: for any $y \in \mathbb{R}$, there exists a constant a and there exists a function $L\colon \mathbb{R} \to \mathbb{R}$ defined by $L(x) = a(x-y) + \phi(y)$, $x \in \mathbb{R}$, such that $L(y) = \phi(y)$ and such that $L(x) \leq \phi(x)$ for all $x \in \mathbb{R}$. (In the case that ϕ is differentiable, the latter condition says that ϕ lies above all of its tangent lines.)

(Hint: Suppose ϕ is convex. If x is fixed and y varies, show that $\frac{\phi(y)-\phi(x)}{y-x}$ increases as y increases. Draw a picture. What slope a should L have at x?)

Exercise 1.40 (**Jensen's Inequality**)**.** Let $X\colon \Omega \to [-\infty, \infty]$ be a random variable. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be convex. Assume that $\mathbf{E}\,|X| < \infty$ and $\mathbf{E}\,|\phi(X)| < \infty$. Then

$$\phi(\mathbf{E}X) \leq \mathbf{E}\phi(X).$$

(Hint: use Exercise 1.39 with $y := \mathbf{E}X$.) Deduce the **triangle inequality**:

$$|\mathbf{E}X| \leq \mathbf{E}\,|X|\,.$$

Exercise 1.41 (**Markov's Inequality**)**.** Let $X\colon \Omega \to [-\infty, \infty]$ be a random variable. Then

$$\mathbf{P}(|X| \geq t) \leq \frac{\mathbf{E}\,|X|}{t}, \qquad \forall\, t > 0.$$

(Hint: multiply both sides by t and use monotonicity of $\mathbf{E}$.)

Corollary 1.42. *If n is a positive integer, then*

$$\mathbf{P}(|X| \geq t) \leq \frac{\mathbf{E}\,|X|^n}{t^n}, \qquad \forall\, t > 0.$$

Proof. From Markov's Inequality, Exercise 1.41,

$$\mathbf{P}(|X| \geq t) = \mathbf{P}(|X|^n \geq t^n) \leq \frac{\mathbf{E}\,|X|^n}{t^n}, \qquad \forall\, t > 0.$$

□

We refer to $\mathbf{E}\,|X|^n$ as the n^{th} **moment** of X.

Definition 1.43 (**Variance**)**.** Let $X\colon \Omega \to [-\infty, \infty]$ be a random variable with $\mathbf{E}\,|X| < \infty$ and $\mathbf{E}X^2 < \infty$. We define the **variance** of X, denoted $\mathrm{var}(X)$, to be

$$\mathrm{var}(X) := \mathbf{E}(X - \mathbf{E}X)^2 = \mathbf{E}X^2 - (\mathbf{E}X)^2.$$

Exercise 1.44. Combining Jensen's Inequality with the Monotone Convergence Theorem below, Theorem 1.54, show that if $\mathbf{E}X^2 < \infty$, then $\mathbf{E}\,|X| < \infty$, so $\mathbf{E}X \in \mathbb{R}$.

Exercise 1.45. Let $a, b \in \mathbb{R}$ and let $X\colon \Omega \to [-\infty, \infty]$ be a random variable with $\mathbf{E}X^2 < \infty$. Show that

$$\mathrm{var}(aX + b) = a^2 \mathrm{var}(X).$$

Then, let X be a standard Gaussian. Show that $\mathbf{E}X = 0$ and $\mathrm{var}(X) = 1$.

Finally, show that the quantity $\mathbf{E}(X - t)^2$ is minimized for $t \in \mathbb{R}$ uniquely when $t = \mathbf{E}X$.

Replacing X by $X - \mathbf{E}X$ and taking $n = 2$ in Corollary 1.42 gives:

Corollary 1.46 (Chebyshev's Inequality). *Let $X : \Omega \to [-\infty, \infty]$ be a random variable with $\mathbf{E}X^2 < \infty$. Then*

$$\mathbf{P}(|X - \mathbf{E}X| \geq t) \leq \frac{\mathrm{var}(X)}{t^2}, \qquad \forall t > 0.$$

(By Exercise 1.44, $\mathbf{E}X \in \mathbb{R}$.)

Corollary 1.42 shows that, if large moments of X are finite, then $\mathbf{P}(X > t)$ decays rapidly. Sometimes, we can even get exponential decay on $\mathbf{P}(X > t)$, if we make the rather strong assumption that $\mathbf{E}e^{rX}$ is finite for some $r > 0$. Note that, by the power series expansion of the exponential, $\mathbf{E}e^{rX} < \infty$ assumes that an infinite sum of the moments of X is finite.

Exercise 1.47 (The Chernoff Bound). Let $X : \Omega \to [-\infty, \infty]$ be a random variable. Show that, for any $r, t > 0$,

$$\mathbf{P}(X > t) \leq e^{-rt}\mathbf{E}e^{rX}.$$

If $1 \leq p < \infty$, and if $X : \Omega \to [-\infty, \infty]$ is a random variable, denote the **L_p-norm** of X as $\|X\|_p := (\mathbf{E}|X|^p)^{1/p}$ and denote the **L_∞-norm** of X as $\|X\|_\infty := \inf\{c > 0 : \mathbf{P}(|X| \leq c) = 1\}$.

Theorem 1.48 (Hölder's Inequality). *Let $X, Y : \Omega \to \mathbb{R}$ be random variables. Let $1 \leq p \leq \infty$, and let q be dual to p (so $1/p + 1/q = 1$). Then*

$$\mathbf{E}|XY| \leq \|X\|_p \|Y\|_q .$$

*In particular, the case $p = q = 2$ recovers the **Cauchy-Schwarz** inequality:*

$$\mathbf{E}|XY| \leq (\mathbf{E}X^2)^{1/2}(\mathbf{E}Y^2)^{1/2}.$$

Proof. By scaling, we may assume $\|X\|_p = \|Y\|_q = 1$ (zeros and infinities being trivial). Also, the case $p = 1, q = \infty$ follows from the triangle inequality, so we assume $1 < p < \infty$. From concavity of the log function, we have the pointwise inequality

$$|X(\omega)Y(\omega)| = (|X(\omega)|^p)^{1/p}(|Y(\omega)|^q)^{1/q} \leq \frac{1}{p}|X(\omega)|^p + \frac{1}{q}|Y(\omega)|^q , \qquad \forall \omega \in \Omega$$

which upon integration gives the result. □

Theorem 1.49 (Triangle Inequality). *Let $X, Y : \Omega \to \mathbb{R}$ be random variables. Let $1 \leq p \leq \infty$. Then*

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p , \quad 1 \leq p \leq \infty$$

Proof. The case $p = \infty$ follows from the scalar triangle inequality, so assume $1 \leq p < \infty$. By scaling, we may assume $\|X\|_p = 1 - t$, $\|Y\|_p = t$, for some $t \in (0, 1)$ (zeros and infinities being trivial). Define $V := X/(1 - t)$, $W := Y/t$. Then by convexity of $x \mapsto |x|^p$ on $\mathbb{R}$,

$$|(1 - t)V(\omega) + tW(\omega)|^p \leq (1 - t)|V(\omega)|^p + t|W(\omega)|^p , \qquad \forall \omega \in \Omega$$

which upon integration completes the proof. □

Exercise 1.50. Let $X, Y : \Omega \to \mathbb{R}$ be random variables. Let $0 < p < 1$ and let $\|X\|_p := (\mathbf{E}|X|^p)^{1/p}$. Show that there exists $c(p) > 0$ such that $\|X + Y\|_p \leq c(p)(\|X\|_p + \|Y\|_p)$. In particular, it suffices to choose $c(p) = 2^{1/p}$. (Hint: a pointwise inequality should imply that $\|X + Y\|_p^p \leq \|X\|_p^p + \|Y\|_p^p$.)

Exercise 1.51. Let $X\colon \Omega \to [-\infty,\infty]$ be a random variable. Show that the function $p \mapsto \|X\|_p$ is nondecreasing on the domain $p \in (0,\infty]$. So, if $\|X\|_p$ is finite for some value of p, then it is finite for all smaller values of p. (Hint: approximate X by bounded random variables, and then by apply the Monotone Convergence Theorem.)

Exercise 1.52 (Paley-Zygmund Inequality). Let X be a nonnegative random variable with $\mathbf{E}X^2 < \infty$. Let $0 \le t \le 1$. Then

$$\mathbf{P}(X > t\,\mathbf{E}X) \ge (1-t)^2 \frac{(\mathbf{E}X)^2}{\mathbf{E}X^2}.$$

(Hint: Apply the Cauchy-Schwarz inequality to $X1_{\{X>t\mathbf{E}X\}}$.)

Exercise 1.53 (Logarithmic Convexity of L_p-Norms). Let X be a real-valued random variable. Let $0 < p_1 < p < p_2 \le \infty$, and define $0 \le t \le 1$ by $\frac{1}{p} = \frac{1-t}{p_1} + \frac{t}{p_2}$. Then

$$\|X\|_p \le \|X\|_{p_1}^{(1-t)} \|X\|_{p_2}^{t}.$$

1.6. **Integral Convergence Theorems.**

Theorem 1.54 (Monotone Convergence Theorem). *Let $0 \le X_1 \le X_2 \le \cdots$ be a monotone increasing sequence of functions on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Then*

$$\mathbf{E} \lim_{n\to\infty} X_n = \lim_{n\to\infty} \mathbf{E}X_n.$$

(By monotonicity, $\lim_{n\to\infty} X_n(\omega) \in [0,\infty]$ exists for every $\omega \in \Omega$.)

Lemma 1.55 (Borel-Cantelli Lemma). *Let $A_1, A_2, \ldots$ be events with $\sum_{n=1}^\infty \mathbf{P}(A_n) < \infty$. Let $B := \{\sum_{n=1}^\infty 1_{A_n} = \infty\}$, so that B is the event that infinitely many of the events occur. Then $\mathbf{P}(B) = 0$.*

Proof. From the Monotone Convergence Theorem, Theorem 1.54,

$$\mathbf{E}\sum_{n=1}^\infty 1_{A_n} = \mathbf{E}\lim_{m\to\infty}\sum_{n=1}^m 1_{A_n} = \lim_{m\to\infty}\sum_{n=1}^m \mathbf{E}1_{A_n} = \sum_{n=1}^\infty \mathbf{P}(A_n) < \infty.$$

So, from Continuity of the Probability Law, Exercise 1.19, and Markov's Inequality,

$$0 \le \mathbf{P}(B) = \mathbf{P}\left(\sum_{n=1}^\infty 1_{A_n} = \infty\right) = \lim_{t\to\infty} \mathbf{P}\left(\sum_{n=1}^\infty 1_{A_n} \ge t\right) \le \lim_{t\to\infty} \frac{\sum_{n=1}^\infty \mathbf{P}(A_n)}{t} = 0.$$

(Note that B is measurable by Exercise 1.20) □

Theorem 1.56 (Fatou's Lemma). *Let $X_1, X_2, \ldots$ be nonnegative random variables on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Then*

$$\mathbf{E}\liminf_{n\to\infty} X_n \le \liminf_{n\to\infty} \mathbf{E}X_n$$

In particular, if $X := \lim_{n\to\infty} X_n$ exists almost surely, then

$$\mathbf{E}X \le \liminf_{n\to\infty} \mathbf{E}X_n.$$

Theorem 1.57 (Dominated Convergence Theorem). *Let $X_1, X_2, \ldots : \Omega \to \mathbb{C}$ be random variables on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ that converge almost surely. Assume that Y is a nonnegative random variable with $\mathbf{E}Y < \infty$ and $|X_n| \le Y$ almost surely, $\forall\, n \ge 1$. Then*

$$\mathbf{E}\lim_{n\to\infty} X_n = \lim_{n\to\infty} \mathbf{E}X_n.$$

Corollary 1.58 (**Bounded Convergence Theorem**). *Let $X_1, X_2, \ldots : \Omega \to \mathbb{C}$ be random variables on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ that converge almost surely. Let $c > 0$. Assume that $|X_n| \leq c$ almost surely, for every $n \geq 1$. Then*

$$\mathbf{E} \lim_{n\to\infty} X_n = \lim_{n\to\infty} \mathbf{E} X_n.$$

Theorem 1.59 (**Convergence Theorem with Bounded Moment**). *Let $X_1, X_2, \ldots : \Omega \to \mathbb{C}$ be random variables on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ that converge almost surely to a random variable X. Assume $\exists\ 0 < \varepsilon, c < \infty$ such that $\mathbf{E}\,|X_n|^{1+\varepsilon} \leq c$, $\forall\ n \geq 1$. Then*

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

Proof. Let $t > 0$. Define $X_n^{(t)} := X_n 1_{|X_n| \leq t}$ and $X^{(t)} := X 1_{|X| \leq t}$. Then $X_1^{(t)}, X_2^{(t)}, \ldots$ converges almost surely to $X^{(t)}$, so the Bounded Convergence Theorem, Corollary 1.58 implies

$$\lim_{n\to\infty} \mathbf{E} X_n^{(t)} = \mathbf{E} X^{(t)}.$$

Also, using the inequality $|x - t| \leq |x/t|^{\varepsilon} |x|$ valid for any $x > t$ or using the inequality $|x - (-t)| \leq |x/t|^{\varepsilon} |x|$ valid for any $x < -t$,

$$\left|X_n - X_n^{(t)}\right| \leq t^{-\varepsilon} |X_n|^{1+\varepsilon}.$$

So, taking expected values and applying the triangle inequality,

$$\left|\mathbf{E} X_n - \mathbf{E} X_n^{(t)}\right| \leq \mathbf{E}\left|X_n - X_n^{(t)}\right| \leq t^{-\varepsilon} \mathbf{E}\,|X_n|^{1+\varepsilon} \leq t^{-\varepsilon} c.$$

Applying similar reasoning to X and using Fatou's Lemma, Theorem 1.56,

$$\left|\mathbf{E} X - \mathbf{E} X^{(t)}\right| \leq \mathbf{E}\left|X - X^{(t)}\right| \leq t^{-\varepsilon} \mathbf{E}\,|X|^{1+\varepsilon} \leq t^{-\varepsilon} c.$$

Combining the above with the scalar triangle inequality,

$$|\mathbf{E} X - \mathbf{E} X_n| \leq \left|\mathbf{E} X - \mathbf{E} X^{(t)}\right| + \left|\mathbf{E} X^{(t)} - \mathbf{E} X_n^{(t)}\right| + \left|\mathbf{E} X_n^{(t)} - \mathbf{E} X_n\right|.$$

$$\limsup_{n\to\infty} |\mathbf{E} X - \mathbf{E} X_n| \leq 2t^{-\varepsilon} c.$$

Letting $t \to \infty$ concludes the result. □

Theorem 1.60 (**Change of Variables**). *Let $X \colon \Omega \to S$ be a random variable. Let $f \colon S \to \mathbb{C}$ be measurable (where $\mathbb{C}$ has the Borel σ-algebra). Assume $f \geq 0$ or $\mathbf{E}\,|f(X)| < \infty$. Then*

$$\mathbf{E} f(X) = \int_S f(x) d\mu_X(x).$$

Remark 1.61. In particular, if $S = \mathbb{R}$, and if $0 < p < \infty$, then

$$\mathbf{E}\,|X| = \int_{\mathbb{R}} |x|\, d\mu_X(x), \qquad \mathbf{E}\,|X|^p = \int_{\mathbb{R}} |x|^p\, d\mu_X(x).$$

And if $X \geq 0$ or $\mathbf{E}\,|X| < \infty$, then

$$\mathbf{E} X = \int_{\mathbb{R}} x d\mu_X(x).$$

So, computing expected values can reduce to computing integrals on the real line. That is, we can change variables to integrate over $\mathbb{R}$, where $d\mu_X$ emulates the Jacobian factor from the usual change of variables formula.

Proof. Suppose there exists $A \subseteq S$ measurable such that $f = 1_A$. Then by Definition 1.26

$$\mathbf{E}f(X) = \mathbf{P}(X \in A) = \mu_X(A) = \int_S 1_A(x) d\mu_X(x).$$

So, the Theorem holds for $f = 1_A$. The Theorem then holds for simple functions by linearity. Then, given any $f\colon S \to [0, \infty)$, and given any $n \geq 1$, let $f_n\colon S \to \mathbb{R}$ so that $f_n(s)$ is $f(s)$ rounded down to the largest multiple of $1/n$ less than $f(s)$ and n. Then $f_1, f_2, \ldots$ increases monotonically to f, so the Theorem then holds for f by the Monotone Convergence Theorem, Theorem 1.54. When $f\colon S \to \mathbb{R}$ satisfies $\mathbf{E}\,|f(X)| < \infty$, we have $\mathbf{E}\max(f(X), 0) < \infty$ and $\mathbf{E}\max(-f(X), 0) < \infty$. And the Theorem holds for $\max(f(X), 0)$ and $\max(-f(X), 0)$. Subtracting these identities,

$$\begin{aligned}\mathbf{E}f(X) &= \mathbf{E}\max(f(X), 0) - \mathbf{E}\max(-f(X), 0) \\ &= \int_S (\max(f(x), 0) - \max(-f(x), 0)) d\mu_X(x) = \int_S f(x) d\mu_X(x).\end{aligned}$$

The case $f\colon S \to \mathbb{C}$ follows from the case $f\colon S \to \mathbb{R}$ by taking real and imaginary parts. □

Remark 1.62. If μ_X is absolutely continuous with respect to Lebesgue measure, then X has density $g\colon \mathbb{R} \to [0, \infty)$ and $d\mu_X(x) = g(x)dx$. If additionally $S = \mathbb{R}$, the Change of Variables formula becomes

$$\mathbf{E}f(X) = \int_{\mathbb{R}} f(x)g(x)dx.$$

Example 1.63. Let X be a uniformly distributed random variable in $[0, 1]$. If $0 < p < \infty$,

$$\mathbf{E}X^p = \int_0^1 x^p dx = \frac{1}{p+1}.$$

Exercise 1.64 (**Stein Identity**)**.** Let X be a standard Gaussian random variable, so that X has density $x \mapsto e^{-x^2/2}/\sqrt{2\pi}$, $\forall\ x \in \mathbb{R}$. Let $g\colon \mathbb{R} \to \mathbb{R}$ be a continuously differentiable function such that g and g' have polynomial volume growth. That is, $\exists\ a, b > 0$ such that $|g(x)|, |g'(x)| \leq a(1 + |x|)^b$, $\forall\ x \in \mathbb{R}$. Prove the **Stein identity**

$$\mathbf{E}Xg(X) = \mathbf{E}g'(X).$$

Using this identity, recursively compute $\mathbf{E}X^k$ for any positive integer k.

Alternatively, for any $t > 0$, show that $\mathbf{E}e^{tX} = e^{t^2/2}$, i.e. compute the **moment generating function** of X. Then, using $\frac{d^k}{dt^k}|_{t=0}\mathbf{E}e^{tX} = \mathbf{E}X^k$ and using the power series expansion of the exponential, compute $\mathbf{E}X^k$ directly from the identity $\mathbf{E}e^{tX} = e^{t^2/2}$.

1.7. **Product Measures, Independence.**

Exercise 1.65 (**Finite Product Measure**)**.** Let $(\Omega_i, \mathcal{F}_i, \mu_i)$ be probability spaces for any $1 \leq i \leq n$. Show that there exists a unique probability measure, denoted $\prod_{i=1}^n \mu_i$ on $(\prod_{i=1}^n \Omega_i, \prod_{i=1}^n \mathcal{F}_i)$ (where the latter measurable space is defined in Example 1.8) such that

$$\Big(\prod_{i=1}^n \mu_i\Big)\Big(\prod_{i=1}^n A_i\Big) = \prod_{i=1}^n \mu_i(A_i), \qquad \forall\, A_1 \in \mathcal{F}_1, \ldots, A_n \in \mathcal{F}_n.$$

Theorem 1.66 (**Fubini's Theorem**). *Let $(\Omega_1, \mathcal{F}_1, \mu_1)$, $(\Omega_2, \mathcal{F}_2, \mu_2)$ be probability spaces. Let $(\Omega_1 \times \Omega_1, \mathcal{F}_1 \times \mathcal{F}_2, \mu_1 \times \mu_2)$ be the product measure space, defined in Example 1.8. Let f be a measurable function on $\Omega_1 \times \Omega_2$ such that either (i) $f\colon \Omega_1 \times \Omega_2 \to [0, \infty]$ or (ii) $f\colon \Omega_1 \times \Omega_2 \to \mathbb{C}$ and $\int_{\Omega_1 \times \Omega_2} |f|\, d(\mu_1 \times \mu_2) < \infty$. Then*

$$\begin{aligned}\int_{\Omega_1 \times \Omega_2} f d(\mu_1 \times \mu_2) &= \int_{\Omega_1} \Big(\int_{\Omega_2} f(\omega_1, \omega_2) d\mu_2(\omega_2) \Big) d\mu_1(\omega_1) \\ &= \int_{\Omega_2} \Big(\int_{\Omega_1} f(\omega_1, \omega_2) d\mu_1(\omega_1) \Big) d\mu_2(\omega_2).\end{aligned}$$

In particular, the function $\omega_1 \mapsto f(\omega_1, \omega_2)$ is measurable for all $\omega_2 \in \Omega_2$, the function $\omega_1 \mapsto \int_{\Omega_2} f(\omega_1, \omega_2) d\mu_2(\omega_2)$ is measurable for all $\omega_2 \in \Omega_2$, the function $\omega_2 \mapsto f(\omega_1, \omega_2)$ is measurable for all $\omega_1 \in \Omega_2$, and the function $\omega_1 \mapsto \int_{\Omega_2} f(\omega_1, \omega_2) d\mu_2(\omega_2)$ is measurable for all $\omega_2 \in \Omega_2$. And if $\int_{\Omega_1 \times \Omega_2} |f|\, d(\mu_1 \times \mu_2) < \infty$, the previous four functions are absolutely integrable almost everywhere.

The Fubini Theorem can be proven via the Monotone Class Lemma.

Definition 1.67 (**Monotone Class**). A collection $\mathcal{F}$ of subsets of Ω is called a **monotone class** when

- If $A_1 \subseteq A_2 \subseteq \cdots$ are sets in $\mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.
- If $A_1 \supseteq A_2 \supseteq \cdots$ are sets in $\mathcal{F}$, then $\cap_{i=1}^{\infty} A_i \in \mathcal{F}$.

Lemma 1.68 (**Monotone Class Lemma**). *Let $\mathcal{F}$ be an algebra of sets on Ω. Then $\sigma(\mathcal{F})$ is the smallest monotone class containing $\mathcal{F}$.*

Definition 1.69 (**Independent Random Variables**). We say a collection $\big(X_i\colon \Omega \to S_i\big)_{i \in I}$ of random variables is **independent** if the distribution of $(X_i)_{i \in I}$ is the product of the distributions of the X_i. That is, for any finite $J \subseteq I$ and for any measurable sets $A_i \subseteq S_i$, $i \in J$, we have

$$\mathbf{P}\big(\bigcap_{i \in J} \{X_i \in A_i\}\big) = \prod_{i \in J} \mathbf{P}(X_i \in A_i).$$

Remark 1.70. In order for the definition of independence to make sense, the random variables must have the same domain.

Remark 1.71. It follows from the above definition that a finite set of random variables $X_1\colon \Omega \to S_1, \ldots, X_m\colon \Omega \to S_m$ is independent if, for any measurable sets $A_i \subseteq S_i$, where $1 \leq i \leq m$, we have

$$\mathbf{P}\big(\bigcap_{i=1}^{m} \{X_i \in A_i\}\big) = \prod_{i=1}^{m} \mathbf{P}(X_i \in A_i).$$

If $(X_i\colon \Omega \to S_i)_{i \in I}$ is a collection of independent random variables, and if for any finite $K \subseteq I$ we denote $X_K := (X_i)_{i \in K}$, then for any finite disjoint subsets $K_1, \ldots, K_n \subseteq I$ and for any measurable sets $B_i \subseteq \prod_{j \in K_i} S_j$ where $1 \leq i \leq n$, we have

$$\mathbf{P}\big(\bigcap_{i=1}^{n} \{X_{K_i} \in B_i\}\big) = \prod_{i=1}^{n} \mathbf{P}(X_{K_i} \in B_i).$$

That is, the random variables $X_{K_1}, \dots, X_{K_n}$ are independent. So, if $F_i \colon \prod_{j \in K_i} S_j \to T_i$ are measurable for all $1 \leq i \leq n$, then the random variables $F_1(X_{K_1}), \dots, F_n(X_{K_n})$ are independent.

For example, if X_1, X_2, X_3 are independent, then $F_1(X_1, X_2)$ and $F_2(X_3)$ are independent.

Proposition 1.72. *Let $X, Y \colon \Omega \to \mathbb{C}$ be independent random variables. If either condition $X, Y \geq 0$ or $\mathbf{E}\,|XY| < \infty$ or $\mathbf{E}\,|X|\,, \mathbf{E}\,|Y| < \infty$ holds, then*

$$\mathbf{E}(XY) = \mathbf{E}X\mathbf{E}Y.$$

More generally, if $X, Y \colon \Omega \to S$ are independent random variables, if $F, G \colon S \to \mathbb{C}$ are measurable, and if either $F(X), G(Y) \geq 0$ or $\mathbf{E}\,|F(X)G(Y)| < \infty$ or $\mathbf{E}\,|F(X)|\,, \mathbf{E}\,|G(Y)| < \infty$, then

$$\mathbf{E}(F(X)G(Y)) = \mathbf{E}F(X)\mathbf{E}G(Y).$$

Proof. By Theorem 1.60 for the random variable (X, Y), Definition 1.69, and Theorem 1.66,

$$\mathbf{E}F(X)G(Y) = \int_{S \times S} F(x)G(y) d\mu_{X,Y}(x,y) = \int_{S \times S} F(x)G(y) d\mu_X(x) d\mu_Y(y)$$

$$= \int_S \Big(\int_S F(x)G(y) d\mu_X(x) \Big) d\mu_Y(y) = \int_S F(x) d\mu_X(x) \cdot \int_S G(y) d\mu_Y(y) = \mathbf{E}F(X) \cdot \mathbf{E}G(Y).$$

In the last line, we used Theorem 1.60. Fubini's Theorem was justified when $F(X), G(Y) \geq 0$ or $\mathbf{E}\,|F(X)G(Y)| < \infty$. In the case $\mathbf{E}\,|F(X)|\,, \mathbf{E}\,|G(Y)| < \infty$, the above equality applied to $|F|$ and $|G|$ shows that $\mathbf{E}\,|F(X)G(Y)| = \mathbf{E}\,|F(X)|\,\mathbf{E}\,|G(Y)| < \infty$. So, when $\mathbf{E}\,|F(X)|\,, \mathbf{E}\,|G(Y)| < \infty$ the application of Fubini's Theorem in the above equalities also holds for F and G themselves. □

Exercise 1.73. Let $X \colon \Omega \to \mathbb{R}$ be a random variable (as usual $\mathbb{R}$ has the Borel σ-algebra). Show that X is independent of itself if and only if X is almost surely constant.

Also, show that a constant random variable is independent of any other random variable.

Exercise 1.74. Let $X_1, \dots, X_n$ be discrete random variables (i.e. they take values in finite or countable spaces $S_1, \dots, S_n$ with their discrete σ-algebras). Show that $X_1, \dots, X_n$ are independent if and only if:

$$\mathbf{P}\big(\bigcap_{i=1}^n \{X_i = x_i\}\big) = \prod_{i=1}^n \mathbf{P}(X_i = x_i), \qquad \forall\, x_1 \in S_1, \dots, x_n \in S_n.$$

Exercise 1.75. Show that $X_1, \dots, X_n \colon \Omega \to \mathbb{R}$ are independent if and only if:

$$\mathbf{P}\big(\bigcap_{i=1}^n \{X_i \leq x_i\}\big) = \prod_{i=1}^n \mathbf{P}(X_i \leq x_i), \qquad \forall\, x_1, \dots, x_n \in \mathbb{R}.$$

Exercise 1.76. Let V be a finite-dimensional vector space over a finite field $\mathbb{F}$. Let X be a random variable uniformly distributed in V. Let $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$ be a non-degenerate bilinear form on V (if $v \in V$ satisfies $\langle v, w \rangle = 0$ for all $w \in V$, then $v = 0$.). Let $v_1, \dots, v_n$ be non-zero vectors in V. Show that the random variables $\langle X, v_1 \rangle, \dots, \langle X, v_n \rangle$ are independent if and only if the vectors $v_1, \dots, v_n$ are linearly independent.

Exercise 1.77. Give an example of three random variables $X, Y, Z \colon \Omega \to [-\infty, \infty]$ that are pairwise independent (any two of the random variables X, Y, Z are independent of each other), but such that X, Y, Z are not independent. (Hint: Exercise 1.76 might be helpful.)

Exercise 1.78. Let $X\colon \Omega \to \mathbb{R}^n$ be a random variable with the **standard Gaussian distribution**:

$$\mathbf{P}(X \in A) := \int_A e^{-(x_1^2+\cdots+x_n^2)/2} dx (2\pi)^{-n/2}, \qquad \forall\, A \subseteq \mathbb{R}^n \text{ measurable.}$$

Let $v_1, \ldots, v_m$ be vectors in $\mathbb{R}^n$. Let $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be the standard inner product on $\mathbb{R}^n$, so that $\langle x, y \rangle := \sum_{i=1}^n x_i y_i$ for any $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Show that the random variables $\langle X, v_1 \rangle, \ldots, \langle X, v_m \rangle$ are independent if and only if the vectors $v_1, \ldots, v_m$ are pairwise orthogonal.

We say that a family of events $(A_i)_{i \in I}$ are **independent** if their indicator random variables $(1_{A_i})_{i \in I}$ are independent. One can show this is equivalent to: for any finite subset $J \subseteq I$,

$$\mathbf{P}\Big(\bigcap_{i \in J} A_i\Big) = \prod_{i \in J} \mathbf{P}(A_i).$$

Exercise 1.79.

- Show that two events A, B are independent if and only if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$
- Find events A, B, C such that $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)$, but such that A, B, C are not independent.
- Find events A, B, C that are pairwise independent (so that any two of A, B, C are independent), but such that A, B, C are not independent.

From Corollary 1.24, if $X_1\colon \Omega_1 \to \mathbb{R}, \ldots, X_n\colon \Omega_n \to \mathbb{R}$ are random variables, then there exists a single probability space $\mathbb{R}^n$ with the Borel σ-algebra and with probability measure $\prod_{i=1}^n \mu_{X_i}$ where the random variables $X_1, \ldots, X_n$ can be realized as independent random variables. In particular, for any $1 \le i \le n$, define

$$Y_i(\omega_1, \ldots, \omega_n) := \omega_i, \qquad \forall\, (\omega_1, \ldots, \omega_n) \in \mathbb{R}^n.$$

Note that $\mu_{X_i} = \mu_{Y_i}$, $\mathbf{P}(X_i \le t) = \mathbf{P}(Y_i \le t)$ for all $1 \le i \le n$ and for all $t \in \mathbb{R}$, and $Y_1, \ldots, Y_n$ are independent by Definition 1.69. Unfortunately, the sample space $\mathbb{R}^n$ depends on n, which is undesirable since we will often consider sums of sequences of random variables as $n \to \infty$. That is, in order to construct an infinite sequence of independent random variables, we should have a single sample space such that all of the random variables have that sample space as their domain. Such a sample space, and a measure, are constructed in the following Theorem.

Theorem 1.80 (**Kolmogorov Extension Theorem, Special Case**). *For any $n \ge 1$, suppose we are given μ_n a probability measure on $\mathbb{R}^n$ with the Borel σ-algebra such that*

$$\mu_{n+1}\big((a_1, b_1] \times \cdots \times (a_n, b_n] \times \mathbb{R}\big) = \mu_n((a_1, b_1] \times \cdots \times (a_n, b_n]), \qquad \forall\, a_1 \le b_1, \ldots, a_n \le b_n.$$

Then there exists a unique probability measure μ_∞ on $\mathbb{R}^{\mathbb{N}}$ (with the product σ-algebra defined in Example 1.8) such that

$$\mu_\infty\big((\omega_i)_{i \in \mathbb{N}} \colon \omega_i \in (a_i, b_i],\, \forall\, 1 \le i \le n\big) = \mu_n((a_1, b_1] \times \cdots \times (a_n, b_n]),\, \forall\, a_1 \le b_1, \ldots, a_n \le b_n.$$

Corollary 1.81 (**Existence of Independent Random Variables**). *Let $X_1\colon \Omega_1 \to \mathbb{R}$, $X_2\colon \Omega_2 \to \mathbb{R}, \ldots$ be a sequence of random variables. Then, there exists a probability space $(\Omega, \mathcal{F}, \mu_\infty)$ and there exists a sequence of random variables $Y_1, Y_2, \ldots$ on Ω such that $Y_1, Y_2, \ldots$ are independent, and such that $\mu_{X_i} = \mu_{Y_i}$ for all $i \ge 1$.*

Proof. From Theorem 1.80, we use $\Omega := \mathbb{R}^{\mathbb{N}}$ and $\mu_n := \prod_{i=1}^n \mu_{X_i}$ for any $n \geq 1$. For any $i \geq 1$, define $Y_i \colon \Omega \to \mathbb{R}$ by

$$Y_i(\omega_1, \omega_2, \ldots) := \omega_i, \qquad \forall\, (\omega_1, \omega_2, \ldots) \in \mathbb{R}^{\mathbb{N}}.$$

Then for any $j \geq 1$ and $t \in \mathbb{R}$, the definition of μ_{Y_j} says

$$\begin{aligned}\mu_{Y_j}((-\infty, t]) &= \mu_\infty(Y_j \leq t) = \mu_\infty\big((\omega_i)_{i\in\mathbb{N}} \colon \omega_j \leq t\big) \\ &= \Big(\prod_{i=1}^{j} \mu_{X_i}\Big)(\mathbb{R}^{n-1} \times (-\infty, t]) = \mu_{X_j}((-\infty, t]).\end{aligned}$$

So, μ_{Y_j} and μ_{X_j} agree on half open intervals. It follows that $\mu_{X_j} = \mu_{Y_j}$ by Exercise 1.7. Finally, to see the independence, note that if $n \geq 1$ and if $A_1, \ldots, A_n \subseteq \mathbb{R}$ are measurable, then by Theorem 1.80, the definition of $Y_1, \ldots, Y_n$, and the definition of μ_n,

$$\begin{aligned}\mu_{Y_1,\ldots,Y_n}(\prod_{i=1}^n A_i) &= \mu_\infty(Y_1 \in A_1, \ldots, Y_n \in A_n) = \mu_n(Y_1 \in A_1, \ldots, Y_n \in A_n) \\ &= \mu_n(\omega_1 \in A_1, \ldots, \omega_n \in A_n) = \mu_n(\prod_{i=1}^n A_i) = \prod_{i=1}^n \mu_{X_i}(A_i) = \prod_{i=1}^n \mu_{Y_i}(A_i).\end{aligned}$$

□

In certain situations, e.g. when we want to construct a sequence of non-independent random variables, we may not be able to change the sample space of the given random variables in this way. That is, we may just need to construct a measure on an arbitrary product of measure spaces. Unfortunately, this is not always possible. In order for the measure to exist, we need an additional assumption on each of the measure spaces in the product.

Definition 1.82 (**Standard Borel Space**)**.** A measurable space $(\Omega, \mathcal{F})$ is a **standard Borel space** if it is isomorphic as a measurable space to $[0, 1]$ with the Borel σ-algebra. That is, $\exists$ a bijection $f \colon \Omega \to [0, 1]$ such that f and f^{-1} are measurable.

Theorem 1.83 (**Kolmogorov Extension Theorem**)**.** *Let $(\Omega_i, \mathcal{F}_i, \mu_i)_{i\in I}$ be a collection of probability spaces such that $(\Omega_i, \mathcal{F}_i)_{i\in I}$ are standard Borel spaces. For any J, K such that $K \subseteq J \subseteq I$, let $\pi_{J\to K} \colon \prod_{i\in J} \Omega_i \to \prod_{i\in K} \Omega_i$ be the usual coordinate projection. For any finite $J \subseteq I$, let μ_J be a probability measure on $(\prod_{i\in J} \Omega_i, \prod_{i\in J} \mathcal{F}_i)$ such that the following compatibility condition holds for any $K \subseteq J$*

$$\mu_J(\pi_{J\to K}^{-1}(A_K)) = \mu_K(A_K), \qquad \forall\, A_K \in \prod_{i\in K} \mathcal{F}_i.$$

Then there exists a unique probability measure μ_I on the measurable space $(\prod_{i\in I} \Omega_i, \prod_{i\in I} \mathcal{F}_i)$ such that, for all finite $J \subseteq I$,

$$\mu_I(\pi_{I\to J}^{-1}(A_J)) = \mu_J(A_J), \qquad \forall\, A_J \in \prod_{i\in J} \mathcal{F}_i.$$

Thankfully, a large class of standard Borel spaces exists, so that we can apply Theorem 1.83 without difficulty.

Lemma 1.84. *Let (S, d) be a complete separable metric space with metric $d\colon S \times S \to [0, \infty)$. (So S has a countable dense set.) Let Ω be a Borel subset of S. Then Ω with the Borel σ-algebra is a standard Borel space.*

For a proof and related discussion, see e.g. here.

Exercise 1.85. Let $\varepsilon_1, \varepsilon_2, \cdots \in \{0, 1\}$ be random variables that are independent and identically distributed copies of the Bernoulli random variable with expectation $1/2$, so that $\mathbf{P}(\varepsilon_n = 1) = \mathbf{P}(\varepsilon_n = 0) = 1/2$ for all $n \geq 1$.

- Show that the random variable $\sum_{n=1}^{\infty} 2^{-n}\varepsilon_n$ is uniformly distributed on the unit interval $[0, 1]$.
- Show that the random variable $\sum_{n=1}^{\infty} 2 \cdot 3^{-n}\varepsilon_n$ is uniformly distributed on the standard middle third Cantor set (where the Cantor set's center is $1/2$.)
- Let μ be a probability measure on $\mathbb{R}$. The Fourier Transform of μ at $\xi \in \mathbb{R}$ is defined by $\widehat{\mu}(\xi) := \int_{\mathbb{R}} e^{ix\xi} d\mu(x)$. where $i = \sqrt{-1}$. For example, if μ is uniform on $[-1/2, 1/2]$, then

$$\widehat{\mu}(\xi) = \int_{-1/2}^{1/2} e^{ix\xi} dx = \frac{e^{i\xi/2} - e^{-i\xi/2}}{i\xi} = \frac{2\sin(\xi/2)}{\xi}, \qquad \forall\, \xi \neq 0.$$

  Using the first item, find an expression for $\sin(\xi)/\xi$ in terms of an infinite product of cosines. (Hint: if a random variable X has distribution μ_X, then $\widehat{\mu_X}(\xi) = \mathbf{E}e^{iX\xi}$ for any $\xi \in \mathbb{R}$. So the Fourier transform of the sum of independent random variables is the product of the Fourier transforms.) Similarly, find an expression for the Fourier transform of the uniform measure on the middle third Cantor set (when the Cantor set's center is $0 \in \mathbb{R}$) in terms of an infinite product of cosines.

Theorem 1.60 reduces computing expected values of functions f of a random variable $X\colon \Omega \to \mathbb{R}$ to integrating on the real line with respect to μ_X. If the function f is absolutely continuous, then we can change Theorem 1.60 by "integrating by parts" as follows.

Theorem 1.86 (Integration by parts). *Let $X\colon \Omega \to [0, \infty)$ be a random variable. Let $f\colon \mathbb{R} \to [0, \infty)$ be an absolutely continuous function (with respect to Lebesgue measure on $\mathbb{R}$) with $f(0) = 0$. Then f has an almost everywhere derivative (with respect to Lebesgue measure). Assume $f' \geq 0$ almost everywhere. Then*

$$\mathbf{E}f(X) = \int_0^{\infty} f'(t)\mathbf{P}(X > t)dt.$$

Proof.

$$\begin{aligned}
\mathbf{E}f(X) &= \int_0^\infty f(x)d\mu_X(x), \text{ by Theorem 1.60,} \\
&= \int_0^\infty \int_0^x f'(t)dtd\mu_X(x), \text{ by Lebesgue's Fundamental Theorem of Calculus} \\
&= \int_0^\infty \int_0^\infty 1_{[0,x)}(t)f'(t)dtd\mu_X(x) \\
&= \int_0^\infty \int_0^\infty 1_{(t,\infty)}(x)d\mu_X(x)f'(t)dt, \text{ by Fubini's Theorem, Theorem 1.66} \\
&= \int_0^\infty f'(t)\mathbf{P}(X > t)dt.
\end{aligned}$$

□

Example 1.87. Let $X\colon \Omega \to \mathbb{R}$ be a random variable. Let $0 < p < \infty$. Then

$$\mathbf{E}\,|X|^p = \int_0^\infty pt^{p-1}\mathbf{P}(X > t)dt.$$

Exercise 1.88. Let X be a random variable taking nonnegative integer values. Show that

$$\mathbf{E}X = \sum_{n=1}^\infty \mathbf{P}(X \geq n).$$

Exercise 1.89 (MAX-CUT). The probabilistic method is a very useful way to prove the existence of something satisfying some properties. This method is based upon the following elementary statement: If $\alpha \in \mathbb{R}$ and if a random variable $X\colon \Omega \to \mathbb{R}$ satisfies $\mathbf{E}X \geq \alpha$, then there exists some $\omega \in \Omega$ such that $X(\omega) \geq \alpha$. We will demonstrate this principle in this exercise.

Let $G = (V, E)$ be an undirected graph on the vertices $V = \{1, \ldots, n\}$ so that the edge set E is a subset of unordered pairs $\{i, j\}$ such that $i, j \in V$ and $i \neq j$. Let $S \subseteq V$ and denote $S^c := V \setminus S$. We refer to (S, S^c) as a cut of the graph G. The goal of the MAX-CUT problem is to maximize the number of edges going between S and S^c over all cuts of the graph G.

Prove that there exists a cut (S, S^c) of the graph such that the number of edges going between S and S^c is at least $|E|/2$. (Hint: define a random $S \subseteq V$ such that, for every $i \in V$, $\mathbf{P}(i \in S) = 1/2$, and the events $1 \in S, 2 \in S, \ldots, n \in S$ are all independent. If $\{i, j\} \in E$, show that $\mathbf{P}(i \in S, j \notin S) = 1/4$. So, what is the expected number of edges $\{i, j\} \in E$ such that $i \in S$ and $j \notin S$?)

1.8. **Kolmogorov's Zero-One Law.**

Definition 1.90 (σ-algebra generated by a random variable). Let $X\colon (\Omega, \mathcal{F}) \to (S, \mathcal{B})$ be a random variable. Define the **σ-algebra generated by** X, denoted $\sigma(X)$, to be the σ-algebra generated by

$$\{X \in B\colon B \in \mathcal{B}\} = \{X^{-1}(B)\colon B \in \mathcal{B}\} = \{\{\omega \in \Omega\colon X(\omega) \in B\}\colon B \in \mathcal{B}\}.$$

Equivalently, $\sigma(X)$ is the smallest (coarsest) σ-algebra such that X is measurable. More generally, given a collection of random variables $(X_i)_{i\in I}\colon (\Omega, \mathcal{F}) \to (S, \mathcal{B})$, define $\sigma((X_i)_{i\in I})$

to be the σ-algebra generated by the sets

$$\{X_i^{-1}(B)\colon B \in \mathcal{B}, i \in I\}$$

Equivalently, $\sigma((X_i)_{i\in I})$ is the smallest (coarsest) σ-algebra such that all of the random variables $(X_i)_{i\in I}$ are measurable.

Exercise 1.91. Let $X_1, X_2, \ldots : \Omega \to S$ be random variables. Show that

$$\sigma(X_1, X_2, \ldots) = \sigma(\cup_{i=1}^{\infty}\sigma(X_1, \ldots, X_i)).$$

Definition 1.92. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A collection of σ-algebras $\{\mathcal{F}_i\}_{i\in I} \subseteq \mathcal{F}$ are **independent** if, for any finite $J \subseteq I$, all of the sets $\{A_i\colon i \in J, A_i \in \mathcal{F}_i\}$ are independent. That is, for any finite $J \subseteq I$, and for any sets $\{A_i\colon i \in J, A_i \in \mathcal{F}_i\}$,

$$\mathbf{P}\Big(\bigcap_{i\in J} A_i\Big) = \prod_{i\in J} \mathbf{P}(A_i).$$

Exercise 1.93. Let $(X_i)_{i\in I}$ be a collection of independent random variables. Show that $(X_i)_{i\in I}$ are independent if and only if $(\sigma(X_i))_{i\in I}$ are independent σ-algebras. (Hint: Let $i \in I$ and let $J \subseteq I \setminus \{i\}$ be finite. Are the sets in $\sigma(X_i)$ that are independent of $(\sigma(X_j))_{j\in J}$ a monotone class?)

Exercise 1.94. Let $X_1, X_2, \ldots$ be random variables. Show that $X_1, X_2, \ldots$ are independent if and only if: for every $i \geq 1$, $\sigma(X_{i+1})$ is independent of $\sigma(X_1, \ldots, X_i)$. And the previous cases occur if and only if: for every $i \geq 1$, $\sigma(X_{i+1}, X_{i+2}, \ldots)$ is independent of $\sigma(X_1, \ldots, X_i)$

We define the **tail σ-algebra** of random variables $X_1, X_2, \ldots$ to be

$$\mathcal{T} := \bigcap_{i=1}^{\infty} \sigma(X_i, X_{i+1}, \ldots).$$

Theorem 1.95 (Kolmogorov's Zero-One Law). *Let $X_1, X_2, \ldots$ be independent random variables. Let $A \in \mathcal{T}$. Then $\mathbf{P}(A) \in \{0, 1\}$.*

Proof. For any $i \geq 1$, $\sigma(X_1, \ldots, X_i)$ is independent of $\sigma(X_{i+1}, X_{i+2}, \ldots)$ by Exercise 1.94. So for any $i \geq 1$, by its definition, $\mathcal{T}$ is independent of $\sigma(X_1, \ldots, X_i)$. Fix $C \in \mathcal{T}$. Let $\mathcal{A} := \{A \in \sigma(X_1, X_2, \ldots)\colon A \text{ is independent of } C\}$. As just mentioned, $\mathcal{A} \supseteq \sigma(X_1, \ldots, X_i)$ for every $i \geq 1$, so that $\mathcal{A} \supseteq \cup_{i=1}^{\infty}\sigma(X_1, \ldots, X_i)$. (Note that $\mathcal{F} := \cup_{i=1}^{\infty}\sigma(X_1, \ldots, X_i)$ is an algebra.) We claim that $\mathcal{A}$ is a monotone class. Given this claim, the Monotone Class Lemma, Theorem 1.68 and Exercise 1.91, we conclude that $\mathcal{A} \supseteq \sigma(\mathcal{F}) = \sigma(X_1, X_2, \ldots)$ so that C is independent of $\sigma(X_1, X_2, \ldots)$. By the definition of $C \in \mathcal{T}$, $C \in \sigma(X_1, X_2, \ldots)$, so that C is independent of itself. The only events A independent of themselves have probability 0 or 1, since $\mathbf{P}(A) = \mathbf{P}(A)^2$. So, given the claim, we are done.

We now prove the above claim. Note that $\emptyset, \Omega \in \mathcal{A}$. Let $A, B \in \mathcal{A}$ with $A \subseteq B$. Then $(B \setminus A) \cap C = (B \cap C) \setminus (A \cap C)$, so

$$\begin{aligned}\mathbf{P}((B \setminus A) \cap C) &= \mathbf{P}(B \cap C) - \mathbf{P}(A \cap C) = \mathbf{P}(B)\mathbf{P}(C) - \mathbf{P}(A)\mathbf{P}(C)\\ &= (\mathbf{P}(B) - \mathbf{P}(A))\mathbf{P}(C) = \mathbf{P}(B \setminus A)\mathbf{P}(C).\end{aligned}$$

Therefore, $B \setminus A \in \mathcal{A}$. In particular, $\Omega \setminus A = A^c \in \mathcal{A}$. Since $\mathcal{A}$ is closed under complements, it remains to show that $\mathcal{A}$ is closed under increasing unions. Let $A_1 \subseteq A_2 \subseteq \cdots$ be sets in

$\mathcal{A}$. Using Exercise 1.19 twice,

$$\begin{aligned}\mathbf{P}((\cup_{m=1}^{\infty} A_m) \cap C) &= \mathbf{P}(\cup_{m=1}^{\infty}(A_m \cap C)) = \lim_{m\to\infty} \mathbf{P}(A_m \cap C) \\ &= \lim_{m\to\infty} \mathbf{P}(A_m)\mathbf{P}(C) = \mathbf{P}(\cup_{m=1}^{\infty} A_m)\mathbf{P}(C).\end{aligned}$$

So, $(\cup_{m=1}^{\infty} A_m) \in \mathcal{A}$. That is, $\mathcal{A}$ is a monotone class, as desired. □

Remark 1.96. Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be a sequence of independent random variables. Let $t \in [-\infty, \infty]$. It follows by the definition of $\mathcal{T}$ that the following events are in $\mathcal{T}$

$$\{\lim_{n\to\infty} X_n \text{ exists}\}, \quad \{\limsup_{n\to\infty} X_n > t\}, \quad \{\liminf_{n\to\infty} X_n > t\}.$$

From Kolmogorov's Zero-One Law, Theorem 1.95, all of these events therefore have probability 1 or 0. So, there must exist $a, b \in [-\infty, \infty]$ such that $\liminf_{n\to\infty} X_n = a$ almost surely and $\limsup_{n\to\infty} X_n = b$ almost surely. In the case $a = b$, $X_1, X_2, \ldots$ converges to $a = b$ almost surely, and $\mathbf{P}(\{\lim_{n\to\infty} X_n \text{ exists}\}) = 1$. And in the case $a \neq b$, $X_1, X_2, \ldots$ almost surely does not converge and $\mathbf{P}(\{\lim_{n\to\infty} X_n \text{ exists}\}) = 0$.

Exercise 1.97. Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be a sequence of independent random variables. For any $n \geq 1$, let $S_n := X_1 + \cdots + X_n$. Show the following:

- $\{\lim_{n\to\infty} S_n \text{ exists}\} \in \mathcal{T}$.
- If $t \in [-\infty, \infty]$, then it can occur that $\{\limsup_{n\to\infty} S_n > t\} \notin \mathcal{T}$.
- If $t \in [-\infty, \infty]$ and if $c_1 \leq c_2 \leq \cdots$ is a sequence of real numbers such that $\lim_{n\to\infty} c_n = \infty$, then

$$\{\limsup_{n\to\infty} \frac{S_n}{c_n} > t\} \in \mathcal{T}.$$

It follows from Exercise 1.97 that $\limsup_{n\to\infty} \frac{S_n}{n}$ will be almost surely constant, and $\liminf_{n\to\infty} \frac{S_n}{n}$ will be almost surely constant, if $X_1, X_2, \ldots$ is a sequence of independent random variables.

1.9. **Additional Comments.** The foundations of measure theory were developed in the late 1800s and early 1900s by several mathematicians. In the 1930s, Kolmogorov provided an axiomatic foundation of probability theory via measure theory. Probability theory was often not considered a "serious" subject, perhaps due to its historical affiliation with gambling. Since the 1930s and continuing to the present, more and more subjects embrace probabilistic thinking. Within mathematics itself, analysis, number theory, algebra, combinatorics, etc. all use increasing amounts of probability theory. Randomized algorithms are also used more and more by computer scientists.

2. Laws of Large Numbers

The Laws of Large Numbers and Central Limit Theorem provide limiting statements for sequences of random variables. The exact notions of convergence will depend on the limit theorem. The general goal is to obtain the strongest possible convergence with the weakest possible assumption. Sometimes, the convergence can be upgraded to a stronger notion, but other times this is impossible.

2.1. **Modes of Convergence.** Below are a few of the most commonly encountered notions of convergence of random variables.

Definition 2.1 (**Almost Sure Convergence**)**.** We say random variables $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ converge **almost surely** (or **with probability one**) to a random variable $Y : \Omega \to \mathbb{R}$ if

$$\mathbf{P}(\lim_{n\to\infty} Y_n = Y) = 1.$$

That is, $\mathbf{P}(\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}) = 1$

Definition 2.2 (**Convergence in Probability**)**.** We say that a sequence of random variables $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ **converges in probability** to a random variable $Y : \Omega \to \mathbb{R}$ if: for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}(|Y_n - Y| > \varepsilon) = 0.$$

That is, $\forall\ \varepsilon > 0$, $\lim_{n\to\infty} \mathbf{P}(\omega \in \Omega : |Y_n(\omega) - Y(\omega)| > \varepsilon) = 0$.

Definition 2.3 (**Convergence in Distribution**)**.** We say that real-valued random variables $Y_1, Y_2, \ldots$ **converge in distribution** to a real-valued random variable Y if, for any $t \in \mathbb{R}$ such that $s \mapsto \mathbf{P}(Y \le s)$ is continuous at $s = t$,

$$\lim_{n\to\infty} \mathbf{P}(Y_n \le t) = \mathbf{P}(Y \le t).$$

Note that the random variables are allowed to have different domains.

Definition 2.4 (**Convergence in L_p**)**.** Let $0 < p \le \infty$. We say that random variables $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ **converge in L_p** to $Y : \Omega \to \mathbb{R}$ if $\|Y\|_p < \infty$ and

$$\lim_{n\to\infty} \|Y_n - Y\|_p = 0.$$

(Recall that $\|Y\|_p := (\mathbf{E}\,|Y|^p)^{1/p}$ if $0 < p < \infty$ and $\|X\|_\infty := \inf\{c > 0 : \mathbf{P}(|X| \le c) = 1\}$.)

Exercise 2.5. Let $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ be random variables that converge almost surely to a random variable $Y : \Omega \to \mathbb{R}$. Show that $Y_1, Y_2, \ldots$ converges in probability to Y in the following way.

- For any $\varepsilon > 0$ and for any positive integer n, let

$$A_{n,\varepsilon} := \bigcup_{m=n}^{\infty} \{\omega \in \Omega : |Y_m(\omega) - Y(\omega)| > \varepsilon\}.$$

  Show that $A_{n,\varepsilon} \supseteq A_{n+1,\varepsilon} \supseteq A_{n+2,\varepsilon} \supseteq \cdots$.
- Show that $\mathbf{P}(\cap_{n=1}^{\infty} A_{n,\varepsilon}) = 0$.
- Using Continuity of the Probability Law, deduce that $\lim_{n\to\infty} \mathbf{P}(A_{n,\varepsilon}) = 0$.

Now, show that the converse is false. That is, find random variables $Y_1, Y_2, \ldots$ that converge in probability to Y, but where $Y_1, Y_2, \ldots$ do not converge to Y almost surely.

Exercise 2.6. Let $0 < p \le \infty$. Show that, if $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ converge to $Y : \Omega \to \mathbb{R}$ in L_p, then $Y_1, Y_2, \ldots$ converges to Y in probability.

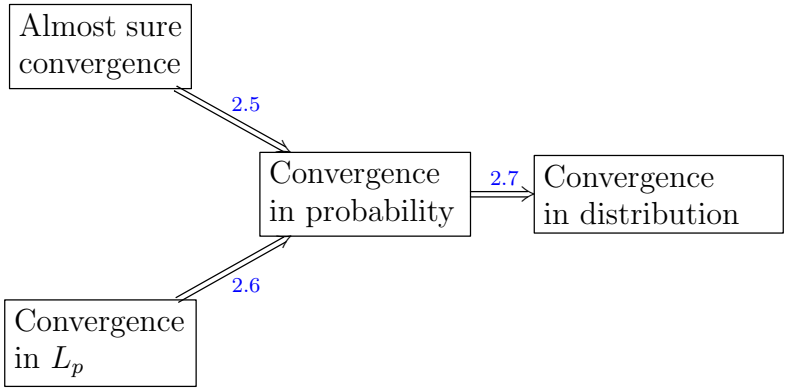
Then, show that the converse is false.

Exercise 2.7. Suppose random variables $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ converge in probability to a random variable $Y : \Omega \to \mathbb{R}$. Prove that $Y_1, Y_2, \ldots$ converge in distribution to Y.

Then, show that the converse is false.

Exercise 2.8. Prove the following statement. Almost sure convergence does not imply convergence in L_2, and convergence in L_2 does not imply almost sure convergence. That is, find random variables that converge in L_2 but not almost surely. Then, find random variables that converge almost surely but not in L_2.

Remark 2.9. The following table summarizes our different notions of convergence of random variables, i.e. the following table summarizes the implications of Exercises 2.6, 2.7 and 2.5.



Exercise 2.10. Let $X, X_1, X_2, \ldots : \Omega \to \mathbb{R}$.

(i) Suppose that $\sum_{i=1}^{\infty} \mathbf{P}(|X_i - X| > \varepsilon) < \infty$ for all $\varepsilon > 0$. Show that $X_1, X_2, \ldots$ converges to X almost surely. Show that the converse does not hold in general.

(ii) Suppose $X_1, X_2, \ldots$ converges to X in probability. Show there is a subsequence $X_{i_1}, X_{i_2}, \ldots$ of $X_1, X_2, \ldots$ such that $X_{i_1}, X_{i_2}, \ldots$ converges to X almost surely. (Here $i_1 < i_2 < \cdots$)

(iii) (Urysohn subsequence principle) Suppose that every subsequence $X_{i_1}, X_{i_2}, \ldots$ of $X_1, X_2, \ldots$ has a further subsequence $X_{i_{j_1}}, X_{i_{j_2}}, \ldots$ that converges to X in probability. Show that $X_1, X_2, \ldots$ also converges to X in probability.

(iv) Suppose $X_1, X_2, \ldots$ converges in probability. Let $F : \mathbb{R} \to \mathbb{R}$ be continuous. Show that $F(X_1), F(X_2), \ldots$ converges in probability to $F(X)$. More generally, suppose $\forall\ 1 \leq j \leq k$, $X_1^{(j)}, X_2^{(j)}, \ldots : \Omega \to \mathbb{R}$ is a sequence of random variables that converge in probability to $X^{(j)}$. Let $F : \mathbb{R}^k \to \mathbb{R}$ be continuous. Show that $F(X_i^{(1)}, \ldots, X_i^{(k)})$ converges in probability to $F(X^{(1)}, \ldots, X^{(k)})$. For example, if $k = 2$, then $X_1^{(1)} + X_1^{(2)}, X_1^{(2)} + X_2^{(2)}, \ldots$ converges in probability to $X^{(1)} + X^{(2)}$, and $X_1^{(1)} \cdot X_1^{(2)}, X_1^{(2)} \cdot X_2^{(2)}, \ldots$ converges in probability to $X^{(1)} \cdot X^{(2)}$.

(v) (Fatou's lemma for convergence in probability) If $X_1, X_2, \ldots : \Omega \to [0, \infty)$ converges in probability to X, show that $\mathbf{E}X \leq \liminf_{n\to\infty} \mathbf{E}X_n$.

(vi) (Dominated convergence in probability) If $X_1, X_2, \ldots$ converge in probability to X, and there exists a random variable $Y : \Omega \to [0, \infty)$ such that, for any $n \geq 1$, $|X_n| \leq Y$ and $\mathbf{E}Y < \infty$, then $\lim_{n\to\infty} \mathbf{E}X_n = \mathbf{E}X$.

2.2. **Limit Theorems with Extra Hypotheses.** In this Section, we prove our first limit theorems under rather strong hypotheses.

We say random variables $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ are **uncorrelated** if $\mathbf{E}X_iX_j = \mathbf{E}X_i\mathbf{E}X_j$ for any $i, j \geq 1$ with $i \neq j$. Recall from Proposition 1.72 that independent random variables are uncorrelated.

Exercise 2.11. Let $X_1, \dots, X_n \colon \Omega \to \mathbb{R}$ be uncorrelated random variables with $\mathbf{E}X_i^2 < \infty$ for any $1 \le i \le n$. Show that

$$\operatorname{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \operatorname{var}(X_i)$$

Exercise 2.12 (L_2 **Weak Law**)**.** Let $\mu, c \in \mathbb{R}$. Let $X_1, X_2, \dots \colon \Omega \to \mathbb{R}$ be uncorrelated random variables with $\mathbf{E}X_i = \mu$ and $\operatorname{var}(X_i) \le c$ for all $i \ge 1$. Then $\frac{X_1+\cdots+X_n}{n}$ converges to μ in L_2 as $n \to \infty$. So, $\frac{X_1+\cdots+X_n}{n}$ converges to μ in probability as $n \to \infty$.

Definition 2.13 (**Identically Distributed**)**.** Let $X, Y \colon \Omega \to \mathbb{R}$ be random variables. We say X, Y are **identically distributed** if X and Y have the same distribution, that is $\mu_X = \mu_Y$. We sometimes refer to independent and identically distributed random variables $X_1, X_2, \dots$ with the abbreviation **i.i.d.**

Proposition 2.14 (**Strong Law of Large Numbers with Finite Fourth Moment**)**.** *Let $X_1, X_2, \dots \colon \Omega \to \mathbb{R}$ be a sequence of independent identically distributed random variables. Let $\mu \in \mathbb{R}$. Assume that $\mu = \mathbf{E}X_1$ and $\mathbf{E}X_1^4 < \infty$. Then*

$$\mathbf{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1.$$

Proof. For any $j \ge 1$, let $Y_j := X_j - \mu$. We are required to show $\mathbf{P}\left(\lim_{n\to\infty} \frac{Y_1+\cdots+Y_n}{n} = 0\right) = 1$. Note that $Y_1, Y_2, \dots$ are independent identically distributed random variables with $\mathbf{E}Y_1 = 0$ and $\mathbf{E}Y_1^4 < \infty$. We compute

$$\mathbf{E}(Y_1 + \cdots + Y_n)^4 = \sum_{1 \le i,j,k,\ell \le n} \mathbf{E}Y_iY_jY_kY_\ell.$$

By independence, terms with $i \ne j = k = \ell$ vanish, since they become $\mathbf{E}Y_iY_jY_kY_\ell = \mathbf{E}Y_i\mathbf{E}Y_j^3 = 0$. Terms with i, j, k, ℓ distinct also vanish, since $\mathbf{E}Y_iY_jY_kY_\ell = \mathbf{E}Y_i\mathbf{E}Y_j\mathbf{E}Y_k\mathbf{E}Y_\ell = 0$. The remaining nonvanishing terms are $i = j = k = \ell$ and the six permutations of $i = j \ne k = \ell$. That is,

$$\mathbf{E}(Y_1 + \cdots + Y_n)^4 = n\mathbf{E}Y_1^4 + 6[n(n-1)/2](\mathbf{E}Y_1^2)^2.$$

By Jensen's Inequality, Exercise 1.40,

$$\mathbf{E}(Y_1 + \cdots + Y_n)^4 \le n\mathbf{E}Y_1^4 + 3n(n-1)\mathbf{E}Y_1^4 \le 4n^2\mathbf{E}Y_1^4. \qquad (*)$$

By Markov's Inequality, Exercise 1.41, for any $t > 0$,

$$\mathbf{P}\left(\left|\frac{Y_1 + \cdots + Y_n}{n}\right| > t\right) \le \frac{\mathbf{E}(Y_1 + \cdots + Y_n)^4}{t^4n^4} \overset{(*)}{\le} \frac{4\mathbf{E}Y_1^4}{t^4n^2}.$$

So $\sum_{n=1}^{\infty} \mathbf{P}(\left|\frac{Y_1+\cdots+Y_n}{n}\right| > t) < \infty$ and by the Borel-Cantelli Lemma 1.55, $\forall\, t > 0$,

$$\mathbf{P}\left(\left|\frac{Y_1 + \cdots + Y_n}{n}\right| > t \text{ for infinitely many } n \ge 1\right) = 0.$$

Since this holds for any $t > 0$, we conclude that $\frac{Y_1+\cdots+Y_n}{n}$ converges almost surely to 0. $\square$

2.3. **Weak Law of Large Numbers.** From the previous section, we see that the weak and strong laws of large numbers follow if we know the random variables have finite second and fourth moments, respectively. In order to weaken these hypotheses, we truncate the random variables. Then, the truncated random variables will automatically have finite moments. And if we do the truncation carefully enough, it will have a negligible effect on the limit theorem in question.

We first state a version of the Weak Law of Large Numbers, where the random variables are allowed to change.

Theorem 2.15 (**Weak Law of Large Numbers for Triangular Arrays**). *For any $n \geq 1$, let $X_{n,1}, \dots, X_{n,n} \colon \Omega \to \mathbb{R}$ be independent random variables. Let $b_1, b_2, \dots$ be a sequence of positive numbers with $\lim_{n\to\infty} b_n = \infty$. For any $1 \leq k \leq n$, define $\overline{X}_{n,k} := X_{n,k} 1_{|X_{n,k}| \leq b_n}$. Assume that*

(i) $\lim_{n\to\infty} \sum_{k=1}^n \mathbf{P}(|X_{n,k}| > b_n) = 0$, *and*
(ii) $\lim_{n\to\infty} b_n^{-2} \sum_{k=1}^n \mathbf{E}\overline{X}_{n,k}^2 = 0$.

Then $b_n^{-1} \sum_{k=1}^n (X_{n,k} - \mathbf{E}\overline{X}_{n,k})$ converges to 0 in probability as $n \to \infty$.

Proof. For any $n \geq 1$, let $S_n := \sum_{k=1}^n X_{n,k}$, $\overline{S}_n := \sum_{k=1}^n \overline{X}_{n,k}$ and let $a_n := \sum_{k=1}^n \mathbf{E}\overline{X}_{n,k}$. Let $\varepsilon > 0$. Using $\mathbf{P}(A) = \mathbf{P}(A \cap B) + \mathbf{P}(A \cap B^c) \leq \mathbf{P}(B) + \mathbf{P}(A \cap B^c)$ for any events A, B,

$$\mathbf{P}(b_n^{-1} |S_n - a_n| > \varepsilon) \leq \mathbf{P}(S_n \neq \overline{S}_n) + \mathbf{P}(b_n^{-1} \left|\overline{S}_n - a_n\right| > \varepsilon)$$

The first term is estimated via the union bound (i.e. subadditivity in Exercise 1.19),

$$\mathbf{P}(S_n \neq \overline{S_n}) \leq \mathbf{P}(\cup_{k=1}^n \{X_{n,k} \neq \overline{X}_{n,k}\}) \leq \sum_{k=1}^n \mathbf{P}(X_{n,k} \neq \overline{X}_{n,k}) = \sum_{k=1}^n \mathbf{P}(|X_{n,k}| > b_n).$$

So, from assumption (i), $\lim_{n\to\infty} \mathbf{P}(S_n \neq \overline{S_n}) = 0$. We bound the second term by Chebyshev's inequality, Corollary 1.46, as in the L_2 Weak Law, Exercise 2.12.

$$\mathbf{P}(b_n^{-1} \left|\overline{S}_n - a_n\right| > \varepsilon) \leq \frac{1}{\varepsilon^2 b_n^2} \operatorname{var}(\overline{S}_n) = \frac{1}{\varepsilon^2 b_n^2} \sum_{k=1}^n \operatorname{var}(\overline{X}_{n,k}) \leq \frac{1}{\varepsilon^2 b_n^2} \sum_{k=1}^n \mathbf{E}\overline{X}_{n,k}^2.$$

We also used Exercise 1.45 twice and Exercise 2.11 (since $X_{n,1}, \dots, X_{n,n}$ are independent, $\overline{X}_{n,1}, \dots, \overline{X}_{n,n}$ are independent by Remark 1.71). Then $\lim_{n\to\infty} \mathbf{P}(b_n^{-1} \left|\overline{S}_n - a_n\right| > \varepsilon) = 0$ by assumption (ii), concluding the proof. □

Theorem 2.16 (**Weak Law of Large Numbers**). *Let $X_1, X_2, \dots \colon \Omega \to \mathbb{R}$ be i.i.d.*

- *Suppose $\lim_{x\to\infty} x\mathbf{P}(|X_1| > x) = 0$. For any $n \geq 1$, let $\mu_n := \mathbf{E}(X_1 1_{|X_1| \leq n})$. Then $\frac{1}{n}(X_1 + \cdots + X_n) - \mu_n$ converges to 0 in probability as $n \to \infty$.*
- *Suppose $\mathbf{E}|X_1| < \infty$. Then $(X_1 + \cdots + X_n)/n$ converges to $\mathbf{E}X_1$ in probability as $n \to \infty$.*

Proof. We apply Theorem 2.15 for $X_{n,k} := X_k$ for any $1 \leq k \leq n$ and $b_n := n$. Assumption (i) holds since

$$\sum_{k=1}^n \mathbf{P}(|X_{n,k}| > b_n) = \sum_{k=1}^n \mathbf{P}(|X_k| > n) = n\mathbf{P}(|X_1| > n).$$

The last quantity converges to 0 as $n \to \infty$ by the present Theorem's assumption. Assumption (ii) holds since Theorem 1.86 gives

$$\mathbf{E}\overline{X}_{n,k}^2 = \int_0^\infty 2t\mathbf{P}(|\overline{X}_{n,k}| > t)dt = \int_0^n 2t\mathbf{P}(|X_k| > t)dt = n^2 \int_0^1 2s\mathbf{P}(|X_1| > sn)ds.$$

$$b_n^{-2} \sum_{k=1}^n \mathbf{E}\overline{X}_{n,k}^2 = \int_0^1 2sn\mathbf{P}(|X_1| > sn)ds.$$

As $n \to \infty$, the last quantity goes to zero, by e.g. the Bounded Convergence Theorem, Theorem 1.58. The first assertion follows. The second assertion now follows from the first. By the Dominated Convergence Theorem, Theorem 1.57,

$$\lim_{x\to\infty} x\mathbf{P}(|X_1| > x) \le \lim_{x\to\infty} \mathbf{E}(|X_1| 1_{|X_1|>x}) = 0.$$

$$\lim_{n\to\infty} \mu_n = \lim_{n\to\infty} \mathbf{E}(X_1 1_{|X_1|\le n}) = \mathbf{E}X_1.$$

So, by the first assertion, $\frac{1}{n}(X_1 + \cdots + X_n) - \mu_n$ converges to 0 in probability. And $\mu_n - \mathbf{E}X_1$ converges to 0 in probability, so $\frac{1}{n}(X_1 + \cdots + X_n) - \mathbf{E}X_1$ converges to 0 in probability by Exercise 2.10(iv). □

Remark 2.17. There exists a random variable $X : \Omega \to [0, \infty)$ such that $\lim_{x\to\infty} x\mathbf{P}(|X| > x) = 0$ but $\mathbf{E}|X| = \infty$. So, the second assertion does not imply the first assertion.

Exercise 2.18. A random variable $X : \Omega \to \mathbb{R}$ is said to be in weak L_1 if

$$\sup_{t>0} t\,\mathbf{P}(|X| > t) < \infty.$$

For example, a Cauchy distributed random variable X has density $f(x) = \frac{1}{\pi(1+x^2)}$ for any $x \in \mathbb{R}$, and X is in weak L_1 while $\mathbf{E}|X| = \infty$.

Show that, if $X_1, X_2, \ldots : \Omega \to (0, \infty)$ are i.i.d. such that X_1 is in weak L_1, then there exist real numbers $a_1, a_2, \ldots$ such that $\lim_{n\to\infty} a_n = \infty$ such that $\frac{1}{a_n}(X_1 + \cdots + X_n)$ converges in probability to 1.

(Hint: If you want to build up your intuition, assume $\mathbf{P}(X_1 > t) = 1/t$ for all $t > 2$, and use $b_n := n \log n$ in the Weak Law for Triangular Arrays.)

(Hint: Let $f(s) := \mathbf{E}X_1 1_{X_1 \le s}$ for any $s > 0$. Note that $f(s)/s = \int_0^s (1/s)\mathbf{P}(X > t)dt = \int_0^1 \mathbf{P}(X > sx)dx \to 0$ as $s \to \infty$ by the Bounded Convergence Theorem. Choose $b_1 \le b_2 \le \cdots$ going to infinity such that $nf(b_n) \le b_n$ for all large $n \ge 1$ as follows. When n is fixed and large, $nf(s)/s$ is larger than 1, and it converges to 0 as $s \to \infty$. Also, $nf(s)/s$ is right continuous in s, so let $b_n := \inf\{s > 0 : nf(s)/s \le 1\}$. Assume $\mathbf{E}X_1 = \infty$. Note that $\lim_{s\to\infty} \frac{f(s)}{s\mathbf{P}(X_1>s)} = \infty$, so $\infty = \lim_{n\to\infty} \frac{f(b_n)}{b_n\mathbf{P}(X_1>b_n)} = \lim_{n\to\infty} \frac{1}{n\mathbf{P}(X_1>b_n)}$, i.e. $\lim_{n\to\infty} n\mathbf{P}(X_1 > b_n) = 0$. Now, use the Weak Law for Triangular arrays. Note that $\lim_{n\to\infty} \frac{b_n}{n} = \lim_{n\to\infty} f(b_n) = \lim_{s\to\infty} f(s) = \infty$, using $\mathbf{E}X_1 = \infty$.)

Exercise 2.19 (Triangular Arrays)**.** For any $n \ge 1$, let $X_{n,1}, \ldots, X_{n,n} : \Omega \to \mathbb{R}$ be a collection of independent random variables, and let $S_n = X_{n,1} + \cdots + X_{n,n}$. Let $\mu \in \mathbb{R}$.

(i) (Weak law) If $\mathbf{E}X_{n,i} = \mu$ for all $1 \le i \le n$ and $\sup_{i,n} \mathbf{E}|X_{n,i}|^2 < \infty$, show that S_n/n converges in probability to μ as $n \to \infty$.

(ii) (Strong law) If $\mathbf{E}X_{n,i} = \mu$ for all $1 \le i \le n$ and $\sup_{i,n} \mathbf{E}|X_{n,i}|^4 < \infty$, show that S_n/n converges almost surely to μ as $n \to \infty$.

Exercise 2.20. For any natural number n and a parameter $0 < p < 1$, define an Erdös-Renyi graph on n vertices with parameter p to be a random graph (V, E) on a (deterministic) vertex set V of n vertices (thus (V, E) is a random variable taking values in the discrete space of all $2^{\binom{n}{2}}$ possible undirected graphs one can place on V) such that the events $\{i, j\} \in E$ for unordered pairs with $i, j \in V$ are independent and each occur with probability p.

For each $n \geq 1$, let (V_n, E_n) be an Erdös-Renyi graph on n vertices with parameter $p = 1/2$ (we do not require the graphs to be independent of each other).

(i) Let $|E_n|$ be the number of edges in (V_n, E_n). Show that $|E_n| / \binom{n}{2}$ converges almost surely to $1/2$ (Hint: use Exercise 2.19.)

(ii) Let $|T_n|$ be the number of triangles in (V_n, E_n) (i.e. the set of unordered triples $\{i, j, k\}$ with $i, j, k \in V_n$ such that $\{i, j\}, \{i, k\}, \{j, k\} \in E_n$), show that $|T_n| / \binom{n}{3}$ converges in probability to $1/8$. (Note: there is not quite enough joint here to directly apply the law of large numbers, so try using the second moment method directly.)

(iii) Show in fact that $|T_n| / \binom{n}{3}$ converges almost surely to $1/8$. (Note: you don't need to compute the fourth moment here.)

Exercise 2.21. For each $n \geq 1$, let $A_n = (a_{ij,n})_{1 \leq i,j \leq n}$ be a random $n \times n$ matrix (i.e. a random variable taking values in the space $\mathbb{R}^{n \times n}$ or $\mathbb{C}^{n \times n}$ of $n \times n$ matrices) such that the entries $a_{ij,n}$ of A_n are independent in i, j and take values in $\{-1, 1\}$ with a probability of $1/2$ each. We do not assume any independence for the sequence $A_1, A_2, \ldots$.

(i) Show that the random variables $\mathrm{Tr} A_n A_n^* / n^2$ are equal to the constant 1, where A_n^* denotes the matrix adjoint (which, in this case, is also the transpose) of A_n and Tr denotes the trace (or sum of the diagonal entries) of a matrix.

(ii) Show that for any natural number $k \geq 1$, the quantities $\mathbf{E}\mathrm{Tr}(A_n A_n^*)^k / n^{k+1}$ are bounded uniformly in $n \geq 1$ (i.e. they are bounded by a quantity C_k that can depend on k but not on n). (It may be helpful to first try $k = 2$ and $k = 3$.)

(iii) Let $\|A_n\|$ denote the operator norm of A_n, and let $\varepsilon > 0$. Show that $\|A_n\| / n^{1/2+\varepsilon}$ converges almost surely to zero, and that $\|A_n\| / n^{1/2-\varepsilon}$ diverges almost surely to infinity. (Hint: use the spectral theorem to relate $\|A_n\|$ with the quantities $\mathrm{Tr}(A_n A_n^*)^k$.

Exercise 2.22. The Cramér random model for the primes is a random subset $\mathcal{P}$ of the natural numbers such that $1 \notin \mathcal{P}$, $2 \in \mathcal{P}$, and the events $n \in \mathcal{P}$ for $n = 3, 4, \ldots$ are independent with $\mathbf{P}(n \in \mathcal{P}) := \frac{1}{\log n}$. Here we used the restriction $n \geq 3$ so that $\frac{1}{\log n} < 1$. This random set of integers $\mathcal{P}$ gives a reasonable way to model the primes $2, 3, 5, 7, \ldots$, since by the Prime Number Theorem, the number of primes less than n is approximately $n / \log n$, so the probability of n being a prime should be about $1 / \log n$. The Cramér random model can provide heuristic confirmations for many conjectures in analytic number theory:

- (Probabilistic prime number theorem) Show that $\frac{1}{x / \log x} |\{n \leq x : n \in \mathcal{P}\}|$ converges almost surely to one as $x \to \infty$.
- (Probabilistic Riemann hypothesis) Let $\varepsilon > 0$. Show that
$$\frac{1}{x^{1/2+\varepsilon}} \Big(|\{n \leq x : n \in \mathcal{P}\}| - \int_2^x \frac{dt}{\log t} \Big)$$
converges almost surely to zero as $x \to \infty$.
- (Probabilistic twin prime conjecture) Show that almost surely, there are an infinite number of elements p of $\mathcal{P}$ such that $p + 2$ also lies in $\mathcal{P}$.

- (Probabilistic Goldbach conjecture) Show that almost surely, all but finitely many natural numbers n are expressible as the sum of two elements of $\mathcal{P}$.

Exercise 2.23. This exercise proves the Hardy-Ramanujan Theorem. This theorem, with probabilistic proof due to Turán, says that a typical large $n \in \mathbb{N}$ has about $\log\log n$ distinct prime factors. Unlike the previous exercise, the probabilistic proof here proves a rigorous result about primes.

Let $\mathcal{P} \subseteq \mathbb{N}$ denote the set of prime numbers (in this exercise $\mathcal{P}$ is deterministic, not random). When $p \in \mathcal{P}$ and $n \in \mathbb{N}$, we use the notation $p|n$ to denote "p divides n," i.e. n/p is a positive integer. Let $x \geq 100$ with $x \in \mathbb{N}$ (so that $\log\log x \geq 1$), and let N be a natural number that is uniformly distributed in $\{1, 2, \ldots, x\}$. Assume Mertens' theorem

$$\sum_{p \in \mathcal{P}:\ p \leq x} \frac{1}{p} = \log\log x + O(1).$$

- Show that the random variable $\sum_{p \in \mathcal{P}:\ p \leq x^{1/10}} 1_{p|N}$ has mean $\log\log x + O(1)$ and variance $O(\log\log x)$. (Hint: up to reasonable errors, compute the means, variances and covariances of the random variables $1_{p|N}$.)
- For any $n \in \mathbb{N}$, let $f(n)$ denote the number of distinct prime factors of n. Show that $\frac{f(N)}{\log\log N}$ converges to 1 in probability as $x \to \infty$. (Hint: first show that $f(N) = \sum_{p \in \mathcal{P}:\ p \leq x^{1/10}} 1_{p|N} + O(1)$.) More precisely, show that

$$\frac{f(N) - \log\log N}{g(N)\sqrt{\log\log N}}$$

converges in probability to zero as $x \to \infty$, whenever $g\colon \mathbb{N} \to \mathbb{R}$ is any function satisfying $\lim_{n\to\infty} g(n) = \infty$.

2.4. **Strong Law of Large Numbers.** From Chebyshev's Inequality, Corollary 1.46, and Exercise 2.11, if $X_1, \ldots, X_n\colon \Omega \to \mathbb{R}$ are independent random variables with mean zero, then for any $t > 0$,

$$\mathbf{P}(|X_1 + \cdots + X_n| > t) \leq t^{-2}\mathrm{var}(X_1 + \cdots + X_n) = t^{-2}(\mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n))$$

We used this inequality in our proof of the L_2 Weak Law, Exercise 2.12, and Theorem 2.15. To prove the Strong Law of Large Numbers, we use the following stronger version of this inequality, where a maximum appears on the left side.

Theorem 2.24 (**Kolmogorov Maximal Inequality**). *Let $X_1, X_2, \ldots\colon \Omega \to \mathbb{R}$ be independent random variables with $\mathbf{E}X_i = 0$ and $\mathbf{E}X_i^2 < \infty$ for all $i \geq 1$. Then for any $t > 0$, and for any $k > 0$,*

$$\mathbf{P}\left(\max_{1 \leq n \leq k} |X_1 + \cdots + X_n| \geq t\right) \leq \frac{\mathrm{var}(X_1) + \cdots + \mathrm{var}(X_k)}{t^2}.$$

Proof. Let $t > 0$. For any $n \geq 1$, define $S_n := X_1 + \cdots + X_n$. For any $n \geq 1$, let A_n be the event that $|S_n| \geq t$ and $|S_j| < t$ for all $1 \leq j < n$. Then $A_1, \ldots, A_k$ are disjoint, and $\cup_{n=1}^k A_n = \{\max_{1 \leq n \leq k} |S_n| \geq t\}$. So, using $\mathbf{P}(\cup_{n=1}^k A_n) \leq \sum_{n=1}^k \mathbf{P}(A_n)$ and $\sum_{n=1}^k \mathrm{var}(X_n) = \mathbf{E}S_k^2$, it suffices to show that

$$\sum_{n=1}^k \mathbf{P}(A_n) \leq \frac{\mathbf{E}S_k^2}{t^2}. \qquad (*)$$

When A_n occurs, we have $1 \leq \frac{1}{t^2} S_n^2$. Therefore,

$$\mathbf{P}(A_n) = \mathbf{E} 1_{A_n} \leq \mathbf{E} 1_{A_n} \frac{1}{t^2} S_n^2, \qquad \forall\, 1 \leq n \leq k.$$

Below, we will show that

$$\mathbf{E} 1_{A_n} S_n^2 \leq \mathbf{E} 1_{A_n} S_k^2, \qquad \forall\, 1 \leq n \leq k. \qquad (**)$$

Then $(**)$ implies $(*)$, since the disjointness of the sets $A_1, \ldots, A_k$ implies $\sum_{n=1}^k 1_{A_n} \leq 1$, so

$$\sum_{n=1}^k \mathbf{P}(A_n) \leq \sum_{n=1}^k \mathbf{E} 1_{A_n} \frac{1}{t^2} S_n^2 \overset{(**)}{\leq} \frac{1}{t^2} \mathbf{E} \sum_{n=1}^k 1_{A_n} S_k^2 \leq \frac{1}{t^2} \mathbf{E} S_k^2.$$

We now prove $(**)$. Let $1 \leq n \leq k$. Then, squaring both sides of $S_k = S_n + (S_k - S_n)$,

$$\begin{aligned} S_k^2 &= S_n^2 + (X_{n+1} + \cdots + X_k)^2 + 2S_n(X_{n+1} + \cdots + X_k) \\ &\geq S_n^2 + 2S_n(X_{n+1} + \cdots + X_k). \end{aligned}$$

Multiplying by 1_{A_n} and taking expected values,

$$\mathbf{E} S_k^2 1_{A_n} \geq \mathbf{E} S_n^2 1_{A_n} + 2\mathbf{E}[1_{A_n} S_n (X_{n+1} + \cdots + X_k)].$$

So, $(**)$ follows by showing the last term is zero. Note that $X_{n+1}, \ldots, X_k$ are independent of S_n, and $X_{n+1}, \ldots, X_k$ are independent of 1_{A_n}, since 1_{A_n} only depends on $X_1, \ldots, X_n$. Therefore, by Proposition 1.72,

$$\mathbf{E}[1_{A_n} S_n (X_{n+1} + \cdots + X_k)] = \mathbf{E}(1_{A_n} S_n) \cdot \mathbf{E}(X_{n+1} + \cdots + X_k) = 0.$$

The proof of $(**)$ is therefore complete. The Theorem follows. □

Theorem 2.25 (Convergence of Random Series). *Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be independent random variables with $\mathbf{E}X_i = 0$ and $\mathbf{E}X_i^2 < \infty$ for all $i \geq 1$. Assume that*

$$\sum_{i=1}^{\infty} \operatorname{var}(X_i) < \infty.$$

Then $\sum_{i=1}^n X_i$ converges almost surely as $n \to \infty$.

Proof. From the Kolmogorov Maximal Inequality, Theorem 2.24, and continuity of $\mathbf{P}$, Exercise 1.19,

$$\mathbf{P}\left(\sup_{m<n<\infty} |X_{m+1} + \cdots + X_n| \geq t \right) \leq \frac{\sum_{n=m+1}^{\infty} \operatorname{var}(X_n)}{t^2}, \qquad \forall\, t > 0.$$

For any $n \geq 1$, let $S_n := \sum_{i=1}^n X_i$. We have shown

$$\mathbf{P}\left(\sup_{m<n<\infty} |S_n - S_m| \geq t \right) \leq \frac{\sum_{n=m+1}^{\infty} \operatorname{var}(X_n)}{t^2}, \qquad \forall\, t > 0.$$

Let A be the event: $\sup_{n>m} |S_n - S_m| \geq t \;\forall\, m \geq 1$. Then, A can be written as the decreasing intersection $A = \cap_{m=1}^{\infty} \{\sup_{n>m} |S_n - S_m| \geq t\}$. So, by continuity of $\mathbf{P}$, Exercise 1.19,

$$\mathbf{P}(A) = \lim_{m\to\infty} \mathbf{P}\left(\sup_{n>m} |S_n - S_m| \geq t \right) \leq \lim_{m\to\infty} \frac{\sum_{n=m+1}^{\infty} \operatorname{var}(X_n)}{t^2} = 0.$$

Since $\mathbf{P}(A) = 0$, $\mathbf{P}(A^c) = 1$. That is, with probability 1, for any $t > 0$, there exists $m \geq 1$ such that $\sup_{n>m} |S_n - S_m| < t$. That is, with probability 1, the sequence $S_1, S_2, \ldots$ is a Cauchy sequence, so that $\lim_{n\to\infty} S_n$ exists. □

Theorem 2.26 (**Strong Law of Large Numbers**). *Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be i.i.d. random variables with $\mathbf{E}|X_1| < \infty$. Then $\frac{X_1+\cdots+X_n}{n}$ converges almost surely to $\mathbf{E}X_1$ as $n \to \infty$.*

Proof. For any $j \geq 1$, let $Y_j := X_j - \mathbf{E}X_j$. Note that $Y_1, Y_2, \ldots$ are i.i.d. and $\mathbf{E}Y_1 = 0$. We are required to show that $\frac{Y_1+\cdots+Y_n}{n}$ converges to 0 almost surely. Since $Y_1, Y_2, \ldots$ are identically distributed, Theorem 1.86 gives

$$\sum_{n=1}^{\infty} \mathbf{P}(|Y_n| > n) = \sum_{n=1}^{\infty} \mathbf{P}(|Y_1| > n) \leq \int_0^{\infty} \mathbf{P}(|Y_1| > t)dt = \mathbf{E}|Y_1| < \infty.$$

So, the Borel-Cantelli Lemma, Lemma 1.55, says that $|Y_n| > n$ for infinitely many $n \geq 1$ occurs with probability 0. For any $n \geq 1$, let $S_n := \sum_{m=1}^n Y_m$ and $\overline{S}_n := \sum_{m=1}^n Y_m 1_{|Y_m|\leq m}$. Then $S_n/n - \overline{S}_n/n = \frac{1}{n}\sum_{m=1}^n Y_m 1_{|Y_m|>m}$ goes to zero almost surely as $n \to \infty$, since on a set of probability 1, the sum $\sum_{m=1}^n Y_m 1_{|Y_m|>m}$ has only a finite number of nonzero terms (regardless of what n is).

So, it suffices to show that $\overline{S}_n/n$ converges to 0 almost surely. Instead of showing this directly, we first show that a decaying (harmonic) average of the terms $Y_m 1_{|Y_m|\leq m}$ is finite. And in order to apply Theorem 2.25, we need to subtract the mean from these random variables. For any $m \geq 1$, let $Z_m := Y_m 1_{|Y_m|\leq m} - \mathbf{E}Y_m 1_{|Y_m|\leq m}$. Then $Z_1, Z_2, \ldots$ are independent, mean zero random variables (since $Y_1, Y_2, \ldots$ are independent), and using Exercise 1.45,

$$\begin{aligned}\sum_{m=1}^{\infty} \mathrm{var}(Z_m/m) &= \sum_{m=1}^{\infty} m^{-2}\mathrm{var}(Z_m) \leq \sum_{m=1}^{\infty} m^{-2}\mathbf{E}Y_m^2 1_{|Y_m|\leq m} = \mathbf{E}\Big(Y_1^2 \sum_{m=1}^{\infty} m^{-2} 1_{|Y_1|\leq m}\Big) \\ &= \mathbf{E}\Big(Y_1^2 \sum_{m\geq |Y_1|} m^{-2}\Big) \leq \mathbf{E}\Big(Y_1^2 \frac{10}{|Y_1|}\Big) = 10\,\mathbf{E}|Y_1| < \infty.\end{aligned}$$

In the pentultimate inequality we used integral comparison in the form $\sum_{m\geq y}^{\infty} \frac{1}{m^2} \leq \frac{10}{y}$, $\forall\, y > 0$. So, by Theorem 2.25, $\sum_{m=1}^{\infty} Z_m/m$ converges almost surely. By Kronecker's Lemma, Exercise 2.27 below, $\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^n Z_m = 0$ almost surely. That is, $\overline{S}_n/n - \frac{1}{n}\sum_{m=1}^n \mathbf{E}Y_m 1_{|Y_m|\leq m}$ converges to 0 almost surely, as $n \to \infty$. Recalling that $\mathbf{E}Y_m = 0$ for any $m \geq 1$, we have

$$\frac{1}{n}\sum_{m=1}^n \mathbf{E}Y_m 1_{|Y_m|\leq m} = -\frac{1}{n}\sum_{m=1}^n \mathbf{E}Y_m 1_{|Y_m|>m} = -\mathbf{E}\Big(Y_1 \frac{1}{n}\sum_{m=1}^n 1_{|Y_1|>m}\Big).$$

The last quantity goes to 0 as $n \to \infty$ by the Dominated Convergence Theorem, Theorem 1.57. In conclusion, $\overline{S}_n/n$ converges to 0 almost surely. □

Exercise 2.27 (**Kronecker's Lemma**). Let $y_1, y_2, \ldots$ be a sequence of real numbers. Let $0 < b_1 \leq b_2 \leq \cdots$ be a sequence of real numbers that goes to infinity. Assume that $\lim_{n\to\infty} \sum_{m=1}^n y_m$ exists. Then $\lim_{n\to\infty} \frac{1}{b_n}\sum_{m=1}^n b_m y_m = 0$. (Hint: if $s_n := \sum_{m=1}^n y_m$, then the summation by parts formula implies that $\frac{1}{b_n}\sum_{m=1}^n b_m y_m = s_n - \frac{1}{b_n}\sum_{m=1}^{n-1}(b_{m+1} - b_m)s_m$.)

Remark 2.28. The Strong Law of Large Numbers implies the Weak Law of Large Numbers by Exercise 2.5.

Remark 2.29. A Monte Carlo simulation takes n independent samples from some random distribution and then sums the sample results and divides by n. The Strong Law of Large Numbers guarantees that this averaging procedure converges to the average value as n becomes large. Similarly, if we independently sample a population with a poll, this corresponds to randomly sampling some distribution. The Strong Law of Large Numbers guarantees that the average poll results converges to the average value as n becomes large, *regardless of the population size*.

Exercise 2.30 (**Renewal Theory**)**.** Let $t_1, t_2, \ldots$ be positive, independent identically distributed random variables. Let $\mu \in \mathbb{R}$. Assume $\mathbf{E}t_1 = \mu$. For any positive integer j, we interpret t_j as the lifetime of the j^{th} lightbulb (before burning out, at which point it is replaced by the $(j+1)^{st}$ lightbulb). For any $n \geq 1$, let $T_n := t_1 + \cdots + t_n$ be the total lifetime of the first n lightbulbs. For any positive integer t, let $N_t := \min\{n \geq 1 \colon T_n \geq t\}$ be the number of lightbulbs that have been used up until time t. Show that N_t/t converges almost surely to $1/\mu$ as $t \to \infty$. (Hint: if c, t are positive integers, then $\{N_t \leq ct\} = \{T_{ct} \geq t\}$. Apply the Strong Law to T_{ct}.)

Exercise 2.31 (**Playing Monopoly Forever**)**.** Let $t_1, t_2, \ldots$ be independent random variables, all of which are uniform on $\{1, 2, 3, 4, 5, 6\}$. For any positive integer j, we think of t_j as the result of rolling a single fair six-sided die. For any $n \geq 1$, let $T_n = t_1 + \cdots + t_n$ be the total number of spaces that have been moved after the n^{th} roll. (We think of each roll as the amount of moves forward of a game piece on a very large Monopoly game board.) For any positive integer t, let $N_t := \min\{n \geq 1 \colon T_n \geq t\}$ be the number of rolls needed to get t spaces away from the start. Using Exercise 2.30, show that N_t/t converges almost surely to $2/7$ as $t \to \infty$.

Exercise 2.32 (**Random Numbers are Normal**)**.** Let X be a uniformly distributed random variable on $(0, 1)$. Let X_1 be the first digit in the decimal expansion of X. Let X_2 be the second digit in the decimal expansion of X. And so on.

- Show that the random variables $X_1, X_2, \ldots$ are uniform on $\{0, 1, 2, \ldots, 9\}$ and independent.
- Fix $m \in \{0, 1, 2, \ldots, 9\}$. Using the Strong Law of Large Numbers, show that with probability one, the fraction of appearances of the number m in the first n digits of X converges to $1/10$ as $n \to \infty$.

(Optional): Show that for any ordered finite set of digits of length k, the fraction of appearances of this set of digits in the first n digits of X converges to 10^{-k} as $n \to \infty$. (You already proved the case $k = 1$ above.) That is, a randomly chosen number in $(0, 1)$ is normal. On the other hand, if we just pick some number such that $\sqrt{2} - 1$, then it may not be easy to say whether or not that number is normal.

(As an optional exercise, try to explicitly write down a normal number. This may not be so easy to do, even though a random number in $(0, 1)$ satisfies this property!)

Exercise 2.33 (**Cheap Law of the Iterated Logarithm**)**.** Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be independent random variables with mean zero and variance one. The Strong Law of Large Numbers says that $\frac{1}{n}(X_1 + \cdots + X_n)$ converges almost surely to zero (if the random variables are also identically distributed). The Central Limit Theorem says that $\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$ converges in distribution to a standard Gaussian random variable (if the random variables

are also identically distributed). But what happens if we divide by some function of n in between $n^{1/2}$ and n ? This Exercise gives a partial answer to this question.

Let $\varepsilon > 0$. Show that

$$\frac{X_1 + \cdots + X_n}{n^{1/2}(\log n)^{(1/2)+\varepsilon}}$$

converges to zero almost surely as $n \to \infty$. (Hint: Re-do the proof of the Strong Law of Large Numbers, but divide by $n^{1/2}(\log n)^{(1/2)+\varepsilon}$ instead of n. You don't need to do any truncation.)

Exercise 2.34. Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be i.i.d. random variables with $\mathbf{E}\,|X_1| < \infty$. Then $\frac{X_1+\cdots+X_n}{n}$ converges in L_1 to $\mathbf{E}X_1$.

(Hint: From the Strong Law, we already know that $\frac{X_1+\cdots+X_n}{n}$ converges almost surely to $\mathbf{E}X_1$. So, conclude using the Vitali Convergence Theorem, Theorem 6.44.)

2.5. **Concentration for Product Measures.** In certain cases, we can make rather strong conclusions about the distribution of sums of i.i.d. random variables, improving upon the laws of large numbers.

Theorem 2.35 (Hoeffding Inequality/ Large Deviation Estimate). *Let $X_1, X_2, \ldots$ be independent identically distributed random variables with $\mathbf{P}(X_1 = 1) = \mathbf{P}(X_1 = -1) = 1/2$. Let $a_1, a_2, \ldots \in \mathbb{R}$. Then, for any $n \geq 1$,*

$$\mathbf{P}\Big(\sum_{i=1}^n a_i X_i \geq t\Big) \leq e^{-\frac{t^2}{2\sum_{i=1}^n a_i^2}}, \qquad \forall\, t \geq 0.$$

Consequently,

$$\mathbf{P}\Big(\Big|\sum_{i=1}^n a_i X_i\Big| \geq t\Big) \leq 2e^{-\frac{t^2}{2\sum_{i=1}^n a_i^2}}, \qquad \forall\, t \geq 0.$$

Proof. By dividing $a_1, \ldots, a_n$ by a constant, we may assume $\sum_{i=1}^n a_i^2 = 1$. Let $\alpha > 0$. Using the (exponential) moment method as in Markov's inequality, Corollary 1.42, and $\alpha t \geq 0$,

$$\mathbf{P}(\sum_{i=1}^n a_i X_i \geq t) = \mathbf{P}(e^{\alpha \sum_{i=1}^n a_i X_i} \geq e^{\alpha t}) \leq e^{-\alpha t}\mathbf{E}e^{\alpha \sum_{i=1}^n a_i X_i} = e^{-\alpha t}\prod_{i=1}^n \mathbf{E}e^{\alpha a_i X_i}.$$

The last equality used independence of $X_1, X_2, \ldots$ and Proposition 1.72. Using an explicit computation and Exercise 2.36,

$$\mathbf{E}e^{\alpha a_i X_i} = (1/2)(e^{\alpha a_i} + e^{-\alpha a_i}) = \cosh(\alpha a_i) \leq e^{\alpha^2 a_i^2/2}, \qquad \forall\, i \geq 1.$$

In summary, for any $t \geq 0$

$$\mathbf{P}(\sum_{i=1}^n a_i X_i \geq t) \leq e^{-\alpha t}e^{\alpha^2 \sum_{i=1}^n a_i^2/2} = e^{-\alpha t + \alpha^2/2}.$$

Since $\alpha > 0$ is arbitrary, we choose α to minimize the right side. This minimum occurs when $\alpha = t$, so that $-\alpha t + \alpha^2/2 = -t^2/2$, giving the first desired bound. The final bound follows by writing $\mathbf{P}(|\sum_{i=1}^n a_i X_i| \geq t) = \mathbf{P}(\sum_{i=1}^n a_i X_i \geq t) + \mathbf{P}(-\sum_{i=1}^n a_i X_i \geq t)$ and then applying the first inequality twice. □

Exercise 2.36. Show that $\cosh(x) \leq e^{x^2/2}$, $\forall\, x \in \mathbb{R}$.

In particular, Hoeffding's inequality implies that

$$\mathbf{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^n X_i\Big| \geq t\Big) \leq 2e^{-nt^2/2}, \qquad \forall\, t \geq 0.$$

This inequality is much stronger than either Markov's or Cheyshev's inequality, since they only respectively imply that

$$\mathbf{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^n X_i\Big| \geq t\Big) \leq \frac{1}{t}, \quad \mathbf{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^n X_i\Big| \geq t\Big) \leq \frac{1}{nt^2}, \qquad \forall\, t \geq 0.$$

Note also that Hoeffding's inequality gives a quantitative bound for any fixed $n \geq 1$, unlike the (non-quantitative) limit theorems which only hold as $n \to \infty$.

Exercise 2.37 (Chernoff Inequality). Let $0 < p < 1$. Let $X_1, X_2, \ldots$ be independent identically distributed random variables with $\mathbf{P}(X_1 = 1) = p$ and $\mathbf{P}(X_1 = 0) = 1 - p$ for any $i \geq 1$. Then for any $n \geq 1$

$$\mathbf{P}\Big(\frac{1}{n}\sum_{i=1}^n X_i \geq t\Big) \leq e^{-np}\Big(\frac{ep}{t}\Big)^{tn}, \qquad \forall\, t \geq p.$$

Prove the same estimate for $\mathbf{P}(\frac{1}{n}\sum_{i=1}^n X_i \leq t)$ for any $t \leq p$. (Hint: $1 + x \leq e^x$ for any $x \in \mathbb{R}$, so $1 + (e^\alpha - 1)p \leq e^{(e^\alpha - 1)p}$.)

Exercise 2.38. We return to the Erdös-Renyi random graph $G = (V, E)$ on n vertices with parameter $0 < p < 1$ from Exercise 2.20. Define $d := p(n-1)$.

- Show that d is the expected degree of each vertex in G. (The degree of a vertex $v \in V$ is the number of vertices connected to v by an edge in E.)
- Show that there exists a constant $c > 0$ such that the following holds. Assume $p \geq \frac{c \log n}{n}$. Then with probability larger than .9, all vertices of G have degrees in the range $(.9d, 1.1d)$. (Hint: first consider a single vertex, then use the union bound over all vertices.)

Exercise 2.39 (Khintchine Inequality). Let $0 < p < \infty$. Then there exist constants $A_p, B_p \in (0, \infty)$ such that the following holds.

Let $X_1, X_2, \ldots$ be independent identically distributed random variables with $\mathbf{P}(X_1 = 1) = \mathbf{P}(X_1 = -1) = 1/2$. Let $a_1, a_2, \ldots \in \mathbb{R}$. Then

$$A_p \Big\|\sum_{i=1}^n a_i X_i\Big\|_p \leq \Big\|\sum_{i=1}^n a_i X_i\Big\|_2 = (\sum_{i=1}^n a_i^2)^{1/2} \leq B_p \Big\|\sum_{i=1}^n a_i X_i\Big\|_p.$$

So, all L_p (quasi)-norms of $\sum_{i=1}^n a_i X_i$ are comparable.

(In Banach space terminology, there is an isomorphic copy of the Banach space ℓ_2 inside any space $L_p[0,1]$; e.g. we can use $X_i(t) := \operatorname{sign}\sin(2^i \pi t)$ for any $t \in [0,1]$, $i \geq 1$.)

(Hint: For the A_p inequality, use Hoeffding's inequality and "Integration by Parts," Theorem 1.86, obtaining $A_p \leq \sqrt{p} A$ for some fixed $A > 0$. For the B_p inequality with $0 < p < 2$, apply Logarithmic Convexity of L_p norms, Exercise 1.53, in the form $\|X\|_2^2 \leq \|X\|_p^{2(1-\theta)} \|X\|_4^{2\theta}$, then apply the A_4 inequality to get $\|X\|_2^{2(1-\theta)} \leq A_p \|X\|_p^{2(1-\theta)}$.)

2.6. **Additional Comments.** A version of the Law of Large Numbers was stated as early as the 1500s. In the 1700s and 1800s, various laws of large numbers were proved with weaker and weaker hypotheses. For example, the L_2 Weak Law was known to Chebyshev in 1867. The Strong Law of Large Numbers might have first been proven in 1930 by Kolmogorov.

If the random variables have infinite mean, then the Strong Law cannot hold.

Exercise 2.40. Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be i.i.d. with $\mathbf{E}|X_1| = \infty$. Then $\mathbf{P}(|X_n| > n \text{ for infinitely many } n \geq 1) = 1$. And $\mathbf{P}(\lim_{n\to\infty} \frac{X_1+\cdots+X_n}{n} \in (-\infty, \infty)) = 0$. (Hint: show $\sum_{n=1}^{\infty} \mathbf{P}(|X_n| > n) = \infty$, then apply the second Borel-Cantelli Lemma. Write $\frac{S_n}{n} - \frac{S_{n+1}}{n+1} = \frac{S_n}{n(n+1)} - \frac{X_{n+1}}{n+1}$, and consider what happens to both sides on the set where $\lim_{n\to\infty} \frac{S_n}{n} \in \mathbb{R}$.)

Also, unfortunately the strong law cannot hold for triangular arrays.

Exercise 2.41. Let X be a random variable taking values in the natural numbers with $\mathbf{P}(X = n) = \frac{1}{\zeta(3)} \frac{1}{n^3}$, where $\zeta(3) := \sum_{m=1}^{\infty} \frac{1}{m^3}$.

- Show that X is absolutely integrable.
- For any $n \geq 1$, let $X_{n,1}, \ldots, X_{n,n} \colon \Omega \to \mathbb{R}$ be independent copies of X. Show that the random variables $\frac{X_{n,1}+\cdots+X_{n,n}}{n}$ are almost surely unbounded. (Hint: for any constant c, show that $\frac{X_{n,1}+\cdots+X_{n,n}}{n} > c$ occurs with probability at least ε/n for some $\varepsilon > 0$ depending on c. Then use the second Borel-Cantelli lemma.)

Exercise 2.42 (**Second Borel-Cantelli Lemma**). Let $A_1, A_2, \ldots$ be independent events with $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$. Then $\mathbf{P}(A_n \text{ occurs for infinitely many } n \geq 1) = 1$. (Hint: using $1 - x \leq e^{-x}$ for any $x \in \mathbb{R}$, show $\mathbf{P}(\cap_{n=s}^{t} A_n^c) \leq \exp(-\sum_{n=s}^{t} \mathbf{P}(A_n))$, let $t \to \infty$ to conclude $\mathbf{P}(\cup_{n=s}^{\infty} A_n) = 1$ for all $s \geq 1$, then let $s \to \infty$.)

The above proof of the Strong Law of Large Numbers follows a general philosophy in analysis and probability. If one desires an almost sure convergence result, then one needs to prove a maximal inequality. In fact, in certain cases this is an equivalence, formalized as Nikishin's Theorem. For example, Nikishin's Theorem implies the following.

Let $\mathbb{T} := \mathbb{R}/2\pi\mathbb{Z}$. Let $\mathbf{P}$ denote the uniform probability law on $\mathbb{T}$. For any $1 \leq p \leq \infty$, we denote $L_p(\mathbb{T}) := \{f \colon \mathbb{T} \to \mathbb{R} \colon \mathbf{E}|f|^p < \infty\}$. We also recall that any $f \in L_p(\mathbb{T})$ has an associated Fourier series $\sum_{n\in\mathbb{Z}} \widehat{f}(n) e^{-in\theta}$ where $\theta \in \mathbb{T}$, $i = \sqrt{-1}$ and $\widehat{f}(n) := \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{inx} dx$.

Theorem 2.43 (**A Corollary of Nikishin's Theorem**). *Let $1 \leq p \leq 2$. Then the following are equivalent*

(i) *For every $f \in L_p(\mathbb{T})$, $\lim_{m\to\infty} \sum_{n=-m}^{m} \widehat{f}(n) e^{-in\theta} = f$ almost surely.*

(ii) *For every $f \in L_p(\mathbb{T})$, $\sup_{m>0} |\sum_{n=-m}^{m} \widehat{f}(n) e^{-in\theta}| < \infty$ almost surely.*

(iii) *The map $f \mapsto T(f) := \sup_{m>0} |\sum_{n=-m}^{m} \widehat{f}(n) e^{-in\theta}|$ is of weak type (p,p). (There exists a constant $c > 0$ such that $\mathbf{P}(Tf > t) \leq ct^{-p}\mathbf{E}|f|^p$, $\forall\ f \in L_p(\mathbb{T})$ and $\forall\ t > 0$.)*

The famous Carleson-Hunt Theorem showed that (i),(ii),(iii) are true for any $1 < p \leq 2$, and it was known already to Kolmogorov in 1926 that (i),(ii),(iii) are false when $p = 1$. Note that Kolmogorov's Maximal Inequality, Theorem 2.24, is in some sense, of weak type $(2,2)$.

Exercise 2.33 was a weak version of the following.

Theorem 2.44 (Law of the Iterated Logarithm). *Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be i.i.d. random variables with mean zero and variance one. Then, almost surely,*

$$\limsup_{n\to\infty} \frac{X_1 + \cdots + X_n}{\sqrt{2n \log\log n}} = 1.$$

Concentration of measure, Section 2.5, will be revisited in Section 7 for non-product measures.

3. Central Limit Theorems

In Section 2.5 we made some conclusions about the distribution of sums of specific i.i.d. random variables. In this section, we will try to investigate the distribution of general sums of i.i.d. random variables as the number of terms in the sum becomes large. Our effort will culminate in three different proofs of the so-called Central Limit Theorem. This Theorem was apparently called "Central" since it is so fundamental to probability and statistics, and mathematics more generally. We first try to guess how to get a limiting distribution from a sum of i.i.d. random variables.

Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be i.i.d. random variables with mean zero and variance 1. From the Strong Laws of Large Numbers, $\frac{1}{n}(X_1 + \cdots + X_n)$ converges to 0 almost surely (and in probability). Also, as shown in Exercise 2.33, for any $\varepsilon > 0$, $\frac{1}{\sqrt{n}(\log n)^{(1/2)+\varepsilon}}(X_1 + \cdots + X_n)$ converges to 0 almost surely (and in probability). From these results, it is still unclear what value $X_1 + \cdots + X_n$ "typically" takes. For example, if $\mathbf{P}(X_1 = 1) = \mathbf{P}(X_1 = -1) = 1/2$, then $\lim_{n\to\infty} \mathbf{P}(X_1 + \cdots + X_n = 0) = 0$. (What is the exact probability that $\mathbf{P}(X_1 + \cdots + X_n = 0)$?) In order to see what values $X_1 + \cdots + X_n$ "typically" takes, we need to divide by a constant smaller than $\sqrt{n \log n}$

Consider $\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$. Dividing by $\sqrt{n}$ is quite natural since $\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$ has mean zero and variance 1 by Exercise 2.11. So, we expect that the most typical values of $X_1 + \cdots + X_n$ occur in some range $(-a\sqrt{n}, a\sqrt{n})$ for some $a > 0$.

Dividing by anything other than $\sqrt{n}$ will not work correctly. For example, if $g\colon \mathbb{N} \to (0, \infty)$ satisfies $\lim_{n\to\infty} g(n) = \infty$, then it follows from Chebyshev's inequality, Corollary 1.46, that $\frac{1}{g(n)\sqrt{n}}(X_1 + \cdots + X_n)$ converges to 0 in probability. And if $\mathbf{E}\,|X_1|^4 < \infty$, we showed in the proof of Proposition 2.14 that

$$\mathbf{E}\Big(\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)\Big)^4 = 3 + O(\mathbf{E}X_1^4/n).$$

The Paley-Zygmund Inequality, Exercise 1.52 then implies that, for any $0 < \varepsilon < 1$,

$$\mathbf{P}\Big(\Big|\frac{X_1 + \cdots + X_n}{\sqrt{n}}\Big| > \varepsilon\Big) = \mathbf{P}\Big(\Big|\frac{X_1 + \cdots + X_n}{\sqrt{n}}\Big|^2 > \varepsilon^2\Big) \geq (1-\varepsilon^2)^2 \frac{1}{3 + O(\mathbf{E}X_1^4/n)}.$$

In particular, $\frac{g(n)}{\sqrt{n}}(X_1 + \cdots + X_n)$ does not converge in any sensible way as $n \to \infty$. In summary, in order to see what values $X_1 + \cdots + X_n$ typically takes, we must divide by $\sqrt{n}$.

Unfortunately, we cannot hope for $\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$ to converge almost surely or in probability. So, we have to look for a different notion of convergence.

Proposition 3.1. *Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be i.i.d. random variables with mean zero, variance 1 and $\mathbf{E}X_1^4 < \infty$. Then as $n \to \infty$, $\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$ does not converge almost surely*

or in probability to any random variable; moreover no subsequence of $\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$ *converges almost surely or in probability as* $n \to \infty$.

Proof. We argue by contradiction. For any $n \geq 1$, let $S_n := X_1 + \cdots + X_n$. For the sake of contradiction, suppose there exists a subsequence $S_{n_j}/\sqrt{n_j}$ that converges almost surely or in probability to a random variable $Y \colon \Omega \to \mathbb{R}$. By taking a further subsequence, we may assume that $S_{n_{j_k}}/\sqrt{n_{j_k}}$ converges almost surely to Y, by Exercise 2.10(ii). Note that $S_{n_{j_k}}/\sqrt{n_{j_k}}$ has mean zero and variance 1. Also, as shown in Proposition 2.14, the fourth moment of $S_{n_{j_k}}/\sqrt{n_{j_k}}$ is uniformly bounded as $k \to \infty$. So, Theorem 1.59 implies that Y also has mean zero and variance 1. But as shown in Exercise 1.97, Y is measurable in the tail σ-field $\mathcal{T}$, so that Y is almost surely constant. Therefore, Y has variance 0, a contradiction. □

Despite the failure of convergence almost surely or in probability, it turns out that $\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$ does converge in distribution to a Gaussian random variable. Before stating the Central Limit Theorem, we discuss convergence in distribution. First, note that convergence in distribution only involves the cumulative distribution functions of the random variables. So, we can discuss convergence in distribution for random variables on different sample spaces. We begin by presenting convergence in distribution in a general context.

3.1. **Convergence in Distribution.**

Definition 3.2 (**Vague Convergence of Measures**)**.** Let $\mu, \mu_1, \mu_2, \ldots$ be a sequence of finite measures on $\mathbb{R}$ (i.e. $\mu(\mathbb{R}), \mu_n(\mathbb{R}) < \infty$ for all $n \geq 1$). We say that $\mu_1, \mu_2, \ldots$ **converges vaguely** (or **converges weakly**, or **converges in the weak* topology**) to μ if, for any continuous compactly supported function $g \colon \mathbb{R} \to \mathbb{R}$,

$$\lim_{n\to\infty} \int_{\mathbb{R}} g(x) d\mu_n(x) = \int_{\mathbb{R}} g(x) d\mu(x).$$

In functional analysis, there is a subtle but important distinction between weak and weak* convergence, though this difference of terminology seems to be ignored in the probability literature.

As we will show below, convergence in distribution of random variables $X_1, X_2, \ldots$ to a random variable X is equivalent to $\mu_{X_1}, \mu_{X_2}, \ldots$ converging vaguely to μ_X.

Proposition 3.3. *Let* $X, X_1, X_2, \ldots$ *be random variables with values in* $\mathbb{R}$*. Then the following are equivalent*

- $X_1, X_2, \ldots$ *converges in distribution to* X*.*
- $\mu_{X_1}, \mu_{X_2}, \ldots$ *converges vaguely to* μ_X*.*

Proof. Assume that $X_1, X_2, \ldots$ converges in distribution to X. Let $g \colon \mathbb{R} \to \mathbb{R}$ be a continuous compactly supported function. Then g is uniformly continuous. So, if $\varepsilon > 0$, there exist $t_1 < \cdots < t_m$ and $c_1, \ldots, c_m \in \mathbb{R}$ such that $g_\varepsilon(t) := \sum_{i=1}^{m-1} c_i 1_{(t_i, t_{i+1}]}(t)$ satisfies $|g_\varepsilon(t) - g(t)| < \varepsilon$ for all $t \in \mathbb{R}$. Since $F_X \colon \mathbb{R} \to [0,1]$ is monotone increasing and bounded, any point of discontinuity of F_X is a jump discontinuity. So, F_X has at most a countable set of points of discontinuity. Therefore, $t_1 < \cdots < t_m$ can be chosen to all be points of continuity of F_X.

By Theorem 1.86,

$$\left|\mathbf{E}g(X) - \sum_{i=1}^{m-1} c_i\Big(F_X(t_{i+1}) - F_X(t_i)\Big)\right| = |\mathbf{E}g(X) - \mathbf{E}g_\varepsilon(X)| \le \mathbf{E}\,|g(X) - g_\varepsilon(X)| \le \varepsilon.$$

The same holds replacing X with any of $X_1, X_2, \ldots$. So, applying the triangle inequality,

$$\begin{aligned}
&\limsup_{n\to\infty} |\mathbf{E}g(X_n) - \mathbf{E}g(X)| \\
&\quad\le \limsup_{n\to\infty} |\mathbf{E}g(X_n) - \mathbf{E}g_\varepsilon(X_n)| + |\mathbf{E}g_\varepsilon(X_n) - \mathbf{E}g_\varepsilon(X)| + |\mathbf{E}g_\varepsilon(X) - \mathbf{E}g(X)| \\
&\quad\le 2\varepsilon + \limsup_{n\to\infty} \sum_{i=1}^{m-1} |c_i|\,|F_{X_n}(t_{i+1}) - F_X(t_{i+1}) - [F_{X_n}(t_i) - F_X(t_i)]| = 2\varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary $\lim_{n\to\infty} \mathbf{E}g(X_n) = \mathbf{E}g(X)$ as desired.

Now, suppose for any continuous, compactly supported $g\colon \mathbb{R} \to \mathbb{R}$, $\lim_{n\to\infty} \mathbf{E}g(X_n) = \mathbf{E}g(X)$. Let $t \in \mathbb{R}$ be a point of continuity of F_X. Then, for any $\varepsilon > 0$, there exists $\delta > 0$ such that if $|s - t| < 2\delta$, then $|F_X(s) - F_X(t)| < \varepsilon$. By continuity of the probability law, let $m > 0$ such that $\mathbf{P}(|X| > m) < \varepsilon$. By choice of δ, ε we have $\mathbf{P}(|X - t| < \delta) < \varepsilon$. Let $g\colon \mathbb{R} \to [0,1]$ so that $g = 0$ on $(-\infty, -2m]$, $g = 1$ on $(-m, t-\delta]$, $g = 0$ on (t, ∞) and g is linear otherwise. Then

$$\begin{aligned}
\mathbf{E}g(X) &= \mathbf{E}g(X)(1_{-2m<X\le -m} + 1_{-m<X\le t-\delta} + 1_{t-\delta<X\le t}) \\
&= O(\varepsilon) + F_X(t-\delta) + O(\varepsilon) = F_X(t) + O(\varepsilon).
\end{aligned}$$

Since $\lim_{n\to\infty} \mathbf{E}g(X_n) = \mathbf{E}g(X)$, there exists $n_0 = n_0(\varepsilon) > 0$ such that, for all $n > n_0$, $\mathbf{E}g(X_n) = F_X(t) + O(\varepsilon)$. By the definition of g,

$$\mathbf{P}(X_n \le t) \ge \mathbf{E}g(X_n) \ge F_X(t) - O(\varepsilon), \qquad \forall\, n > n_0(\varepsilon).$$

Repeating the above with g where $g = 1$ on $(t+\delta, m]$ and $g = 0$ on $(-\infty, t] \cup [2m, \infty)$ gives

$$\mathbf{P}(X_n > t) \ge 1 - F_X(t) - O(\varepsilon), \qquad \forall\, n > n_0(\varepsilon).$$

Combining these inequalities gives

$$F_{X_n}(t) = F_X(t) + O(\varepsilon), \qquad \forall\, n > n_0(\varepsilon).$$

Letting $\varepsilon \to 0^+$ concludes the proof. □

Proposition 3.3 avoids a subtle technical issue. A sequence of probability measures on $\mathbb{R}$ can converge to a measure μ on $\mathbb{R}$ that is **not** a probability measure. For example, if $\mu_n(A) := 1$ if $n \in A$ and $\mu_n(A) := 0$ if $n \notin A$, then $\mu_1, \mu_2, \ldots$ are probability measures that converge vaguely to the measure μ such that $\mu(\mathbb{R}) = 0$. That is, some mass can be "lost" in the limit as $n \to \infty$. Still, any sequence of probability measures does have a subsequence that converges vaguely to some measure μ with $\mu(\mathbb{R}) \le 1$. This is guaranteed by some general theorems from analysis. Below we denote $C_c(\mathbb{R})$ as the set of continuous compactly supported functions $f\colon \mathbb{R} \to \mathbb{R}$.

Theorem 3.4 (Riesz Representation Theorem). *Let ℓ be a positive linear functional on $C_c(\mathbb{R})$ (so that $\ell(f) \ge 0$ for any $f \in C_c(\mathbb{R})$ with $f \ge 0$, and $\sup_{f\in C_c(\mathbb{R})\colon |f(x)|\le 1, \forall\, x\in\mathbb{R}} |\ell(f)| <$*

∞). *Then there exists a unique regular Borel measure μ on $\mathbb{R}$ such that*

$$\ell(f) = \int_{\mathbb{R}} f d\mu, \qquad \forall\, f \in C_c(\mathbb{R}).$$

That is, the dual $C_c(\mathbb{R})^*$ of $C_c(\mathbb{R})$ is the space of finite signed measures on $\mathbb{R}$.

Theorem 3.5 (**Alaoglu Theorem/ Banach-Alaoglu**). *Let X be a normed linear space. Then the unit ball $B_{X^*} = \{x^* \in X^* \colon \|x^*\| \leq 1\}$ of X^* is weak* compact. (That is, any sequence in B_{X^*} has a subsequence that converges in the weak* topology. Here X^* is the set of linear functions $x^* \colon X \to \mathbb{R}$ with $\|x^*\| < \infty$ and $\|x^*\| := \sup_{x \in X \colon \|x\| \leq 1} |x^*(x)|$)*

The combination of Alaoglu's Theorem (proven in Theorem 8.3) and the Riesz Representation Theorem gives Helly's selection theorem.

Theorem 3.6 (**Helly's Selection Theorem**). *Let $\mu_1, \mu_2, \ldots$ be a sequence of probability measures on $\mathbb{R}$. Then there exists a subsequence $\mu_{n_1}, \mu_{n_2}, \ldots$ that vaguely converges to a measure μ on $\mathbb{R}$ with $\mu(\mathbb{R}) \leq 1$.*

As mentioned above, some mass can "escape" to infinity, in which case $\mu(\mathbb{R})$ can be strictly less than 1. The next Lemma shows that mass "escaping" to infinity is the only obstruction to a sequence of probability measures converging vaguely to another probability measure.

Lemma 3.7. *Let $\mu_1, \mu_2, \ldots$ be a sequence of probability measures on $\mathbb{R}$. Then any subsequential limit of the sequence (with respect to vague convergence) is a probability measure if and only if $\mu_1, \mu_2, \ldots$ is **tight**: $\forall\ \varepsilon > 0$, $\exists\ m = m(\varepsilon) > 0$ such that*

$$\limsup_{n \to \infty} (1 - \mu_n([-m, m])) \leq \varepsilon.$$

Exercise 3.8. Let $X, X_1, X_2, \ldots$ and let $Y, Y_1, Y_2, \ldots$ be random variables with values in $\mathbb{R}$.

(i) Assume that X is constant almost surely. Show that $X_1, X_2, \ldots$ converges to X in distribution if and only if $X_1, X_2, \ldots$ converges to X in probability.

(ii) Prove Lemma 3.7.

(iii) Suppose that $X_1, X_2, \ldots$ converges in distribution to X. Show there exist random variables $Z, Z_1, Z_2, \ldots \colon \Omega \to \mathbb{R}$ such that $\mu_Z = \mu_X$, $\mu_{Z_n} = \mu_{X_n}$ for any $n \geq 1$, and such that $Z_1, Z_2, \ldots$ converges almost surely to Z. (Hint: use the sample space $\Omega = [0, 1]$ and argue as in Exercise 1.25.)

(iv) (Slutsky's Theorem) Suppose $X_1, X_2, \ldots$ converges in distribution to X and $Y_1, Y_2, \ldots$ converges in probability to Y. Assume Y is constant almost surely. Show that $X_1 + Y_1, X_2 + Y_2, \ldots$ converges in distribution to $X + Y$. Show also that $X_1 Y_1, X_2 Y_2, \ldots$ converges in distribution to XY. (Hint: either use (iii) or use (ii) to control error terms.) What happens if Y is not constant almost surely?

(v) (Fatou's lemma) If $g \colon \mathbb{R} \to [0, \infty)$ is continuous, and if $X_1, X_2, \ldots$ converges in distribution to X, show that $\liminf_{n \to \infty} \mathbf{E} g(X_n) \geq \mathbf{E} g(X)$.

(vi) (Bounded convergence) If $g \colon \mathbb{R} \to \mathbb{C}$ is continuous and bounded, and if $X_1, X_2, \ldots$ converges in distribution to X, show that $\lim_{n \to \infty} \mathbf{E} g(X_n) = \mathbf{E} g(X)$.

(vii) (Dominated convergence) If $X_1, X_2, \ldots \colon \Omega \to \mathbb{R}$ converges in distribution to X, and if there exists a random variable $Y \colon \Omega \to [0, \infty)$ with $|X_n| \leq Y$ for all $n \geq 1$ and $\mathbf{E} Y < \infty$, show that $\lim_{n \to \infty} \mathbf{E} X_n = \mathbf{E} X$.

Exercise 3.9 (Portmanteau Theorem). Show that the two properties in Proposition 3.3 are also equivalent to the following three statements:

- For any closed $K \subseteq \mathbb{R}$, $\limsup_{n\to\infty} \mathbf{P}(X_n \in K) \leq \mathbf{P}(X \in K)$.
- For any open $U \subseteq \mathbb{R}$, $\liminf_{n\to\infty} \mathbf{P}(X_n \in U) \leq \mathbf{P}(X \in U)$.
- For any Borel set $E \subseteq \mathbb{R}$ whose topological boundary ∂E satisfies $\mathbf{P}(X \in \partial E) = 0$, $\lim_{n\to\infty} \mathbf{P}(X_n \in E) = \mathbf{P}(X \in E)$.

(Hint: Urysohn's Lemma might be helpful.)

3.2. **Independent Sums and Convolution.** Proposition 3.12 shows that the sum of i.i.d. random variables with density reduces to the repeated convolution of the density function with itself. This interpretation is explored further in Exercise 3.14.

Definition 3.10 (Convolution). Let $g, h\colon \mathbb{R} \to \mathbb{R}$ be measurable functions. The **convolution** of g and h, denoted $g * h$, is the function $g * h\colon \mathbb{R} \to \mathbb{R}$ defined by

$$(g * h)(t) := \int_{-\infty}^{\infty} g(x)h(t-x)dx, \qquad \forall\, t \in \mathbb{R}.$$

In order for this quantity to be well-defined almost surely (with respect to Lebesgue measure on $\mathbb{R}$), we assume that $\int_{\mathbb{R}} |g(x)|\, dx < \infty$ and $\int_{\mathbb{R}} |h(x)|\, dx < \infty$.

Exercise 3.11. Let $f, g, h\colon \mathbb{R} \to \mathbb{R}$ be measurable functions. Assume that $\int_{\mathbb{R}} |f(x)|\, dx$, $\int_{\mathbb{R}} |g(x)|\, dx < \infty$ and $\int_{\mathbb{R}} |h(x)|\, dx < \infty$. Show that $\int_{-\infty}^{\infty} |(g * h)(t)|\, dt < \infty$. Consequently, $(g * h)(t) \in \mathbb{R}$ almost surely for $t \in \mathbb{R}$ (with respect to Lebesgue measure on $\mathbb{R}$).

Then, show that convolution is associative and commutative. That is, $g * h = h * g$ and $f * (g * h) = (f * g) * h$ almost surely.

Proposition 3.12. *Let $X, Y\colon \Omega \to \mathbb{R}$ be two independent random variables with densities $f_X, f_Y\colon \mathbb{R} \to [0, \infty)$, respectively. (Recall that $\int_{\mathbb{R}} f_X(x)dx = 1 < \infty$ and $\int_{\mathbb{R}} f_Y(y)dy = 1 < \infty$). Then $X + Y$ has density $f\colon \mathbb{R} \to [0, \infty)$ given by*

$$f(t) = (f_X * f_Y)(t), \qquad \forall\, t \in \mathbb{R}.$$

Proof. Let $t \in \mathbb{R}$. Using independence and Fubini's Theorem, Theorem 1.66

$$\begin{aligned}\mathbf{P}(X + Y \leq t) &= \int_{\{(x,y)\in\mathbb{R}^2\colon\, x+y\leq t\}} d\mu_{X,Y}(x, y) = \int_{\{(x,y)\in\mathbb{R}^2\colon\, x+y\leq t\}} d\mu_X(x)d\mu_Y(y) \\ &= \int_{\{(x,y)\in\mathbb{R}^2\colon\, x+y\leq t\}} f_X(x)f_Y(y)dxdy = \int_{x=-\infty}^{x=\infty}\int_{y=-\infty}^{y=t-x} f_X(x)f_Y(y)dydx.\end{aligned}$$

Changing variables and using Fubini's Theorem again,

$$\begin{aligned}\mathbf{P}(X + Y \leq t) &= \int_{x=-\infty}^{x=\infty} \Big(\int_{z=-\infty}^{z=t} f_Y(z - x)dz\Big) f_X(x)dx \\ &= \int_{z=-\infty}^{z=t} \Big(\int_{x=-\infty}^{x=\infty} f_Y(z - x)f_X(x)dx\Big)dz = \int_{z=-\infty}^{z=t} (f_X * f_Y)(z)dz.\end{aligned}$$

From Definition 1.26, $X + Y$ has density $f_X * f_Y$, as desired. □

Exercise 3.13. Using convolution, show that if X, Y are standard Gaussian random variables, then $aX + bY$ is a Gaussian random variable with mean 0 and variance $a^2 + b^2$.

Exercise 3.14. Let X, Y, Z be independent and uniformly distributed on $[0, 1]$. Note that f_X is not a continuous function.

Using convolution, compute f_{X+Y}. Draw f_{X+Y}. Note that f_{X+Y} is a continuous function, but it is not differentiable at some points.

Using convolution, compute f_{X+Y+Z}. Draw f_{X+Y+Z}. Note that f_{X+Y+Z} is a differentiable function, but it does not have a second derivative at some points.

Make a conjecture about how many derivatives $f_{X_1+\cdots+X_n}$ has, where $X_1, \ldots, X_n$ are independent and uniformly distributed on $[0, 1]$. You do not have to prove this conjecture. The idea of this exercise is that convolution is a kind of average of functions. And the more averaging you do, the more derivatives $f_{X_1+\cdots+X_n}$ has. Lastly, $f_{X_1+\cdots+X_n}$ should resemble a Gaussian density when n becomes large. So, we should be able to guess at a formulation of the Central Limit Theorem, at least for i.i.d. random variables with density.

Exercise 3.15. Construct two random variables X, Y such that X and Y are each uniformly distributed on $[0, 1]$, and such that $\mathbf{P}(X + Y = 1) = 1$.

Then construct two random variables W, Z such that W and Z are each uniformly distributed on $[0, 1]$, and such that $W + Z$ is uniformly distributed on $[0, 2]$.

(Hint: there is a way to do each of the above problems with about one line of work. That is, there is a way to solve each problem without working very hard.)

3.3. **Fourier Transform/ Characteristic Function.** The quickest way to prove the Central Limit Theorem is to take the Fourier transform of the distribution function of the sum of i.i.d. random variables. We now develop the tools needed for such a proof.

Definition 3.16 (**Fourier Transform/ Characteristic Function**)**.** Let μ be a probability measure on $\mathbb{R}$. Let $i := \sqrt{-1}$. The **Fourier Transform** of μ at $t \in \mathbb{R}$ is defined by

$$\widehat{\mu}(t) := \int_{\mathbb{R}} e^{ixt} d\mu(x).$$

If $X \colon \Omega \to \mathbb{R}$ is a random variable, we define the **characteristic function** of X, or the **Fourier transform** of μ_X, denoted $\phi_X \colon \mathbb{R} \to \mathbb{R}$, by

$$\phi_X(t) := \mathbf{E}e^{itX} = \int_{\mathbb{R}} e^{itx} d\mu_X(x) = \widehat{\mu_X}(t), \qquad \forall\, t \in \mathbb{R}.$$

In particular, if X has density $f \colon \mathbb{R} \to [0, \infty)$, then $\phi_X(t) = \int_{\mathbb{R}} e^{itx} f(x) dx$ is the Fourier transform of the function f.

Note that $\phi_X(t)$ exists for every $t \in \mathbb{R}$ since $|e^{itx}| \leq 1$ for all $x, t \in \mathbb{R}$.

From the power series expansion of the exponential, we have the following formal power series expansion for ϕ_X:

$$\phi_X(t) \sim \sum_{n=0}^{\infty} \frac{(it)^n}{n!} \mathbf{E}X^n.$$

ϕ_X is actually equal to this power series expansion in the following settings.

Exercise 3.17. Let $k \geq 1$ be an integer. Let $X \colon \Omega \to \mathbb{R}$ be a random variable with finite k^{th} moment: $\mathbf{E}\,|X|^k < \infty$. Show that $\phi_X(t)$ is k-times continuously differentiable in t, and

$$\frac{d^k}{dt^k}|_{t=0} \phi_X(t) = i^k \mathbf{E}X^k.$$

In particular, we get the Taylor expansion

$$\phi_X(t) = \sum_{n=0}^{k} \frac{(it)^n}{n!} \mathbf{E}X^n + o(|t|^k), \qquad \forall\, t \in \mathbb{R}.$$

Let $f, g\colon \mathbb{R} \to \mathbb{R}$. We use the notation $f(t) = o(g(t))$, $\forall\ t \in \mathbb{R}$ to denote $\lim_{t\to 0} \left|\frac{f(t)}{g(t)}\right| = 0$.

Exercise 3.18. Assume that $X\colon \Omega \to \mathbb{R}$ is a **subgaussian** random variable, i.e. $\exists\ a, b > 0$ such that

$$\mathbf{P}(|X| > t) \leq ae^{-bt^2}, \qquad \forall\, t \in \mathbb{R}.$$

Show that ϕ_X is equal to its Taylor series:

$$\phi_X(t) = \sum_{n=0}^{\infty} \frac{(it)^n}{n!} \mathbf{E}X^n, \qquad \forall\, t \in \mathbb{R}.$$

Also, show that the Taylor series converges uniformly on any closed interval.

The Fourier transform converts convolution to multiplication (see Proposition 8.6(d)). Likewise, the characteristic function transforms sums of independent random variables into products of characteristic functions.

Exercise 3.19. Let $X, Y\colon \Omega \to \mathbb{R}$ be independent random variables. Show that

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t), \qquad \forall\, t \in \mathbb{R}.$$

The following Theorem allows us to restate the Central Limit Theorem in terms of convergence of characteristic functions.

Theorem 3.20 (Lévy Continuity Theorem, Special Case). *Let $X, X_1, X_2, \ldots$ be real-valued random variables (possibly on different sample spaces). The following are equivalent.*

- *For every $t \in \mathbb{R}$, $\lim_{n\to\infty} \phi_{X_n}(t) = \phi_X(t)$.*
- *$X_1, X_2, \ldots$ converges in distribution to X.*

Proof. The second condition implies the first by Exercise 3.8(vi).

Now, assume the first condition holds. Let $g\colon \mathbb{R} \to \mathbb{R}$ be a Schwartz function (for any integers $j, k \geq 1$, g is k times continuously differentiable and there exists $c_{j,k} \in \mathbb{R}$ such that $|g^{(k)}(x)| \leq \frac{c_{jk}}{1+|x|^j}$, $\forall\, x \in \mathbb{R}$.) The Fourier Inversion Formula, Theorem 8.9, implies that

$$g(X_n) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-iX_n y} \widehat{g}(y) dy.$$

where $\widehat{g}(y) = \int_{\mathbb{R}} e^{ixy} g(x) dx$ for all $y \in \mathbb{R}$. From the Fubini Theorem 1.66,

$$\mathbf{E}g(X_n) = \frac{1}{2\pi} \int_{\mathbb{R}} \mathbf{E}e^{-iX_n y} \widehat{g}(y) dy = \frac{1}{2\pi} \int_{\mathbb{R}} \phi_{X_n}(-y) \widehat{g}(y) dy.$$

Similarly, $\mathbf{E}g(X) = \frac{1}{2\pi} \int_{\mathbb{R}} \phi_X(-y) \widehat{g}(y) dy$. So, $\lim_{n\to\infty} \mathbf{E}g(X_n) = \mathbf{E}g(X)$ by the Dominated Convergence Theorem, Theorem 1.57 (and Proposition 8.7(c)). Since any continuous, compactly supported function g can be uniformly approximated by Schwartz functions in the L_∞ norm (by e.g. replacing g with $g * \phi_\varepsilon$, where $\phi_\varepsilon(x) = \varepsilon^{-1} e^{-x^2/(2\varepsilon^2)}/\sqrt{2\pi}$, letting $\varepsilon \to 0^+$ and applying Proposition 8.5(d)), the identity $\lim_{n\to\infty} \mathbf{E}g(X_n) = \mathbf{E}g(X)$ holds for any continuous, compactly supported $g\colon \mathbb{R} \to \mathbb{R}$. We then conclude by Proposition 3.3. □

Remark 3.21. In particular, if $Y = X_1 = X_2 = \cdots$, the above Theorem implies that if $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}$, then $\mu_X = \mu_Y$.

Exercise 3.22 (Lévy Continuity Theorem). Let $X, X_1, X_2, \ldots$ be real-valued random variables (possibly on different sample spaces). Assume that, $\forall\, t \in \mathbb{R}$, $\phi(t) := \lim_{n\to\infty} \phi_{X_n}(t)$ exists. Then the following are equivalent.

(i) ϕ is continuous at 0.
(ii) $\mu_{X_1}, \mu_{X_2}, \ldots$ is tight. ($\forall\ \varepsilon > 0$, $\exists\ m = m(\varepsilon) > 0$ such that $\limsup_{n\to\infty}(1 - \mu_{X_n}([-m, m])) \le \varepsilon$.)
(iii) There exists a random variable X such that $\phi_X = \phi$.
(iv) $X_1, X_2, \ldots$ converges in distribution to X.

(Hint: Use Lemma 3.7 to get from (ii) to other conditions.)

3.4. **Three Proofs of the Central Limit Theorem.** From Exercise 3.14, we could guess that the Central Limit Theorem could be true, since the repeated convolution of the same function looks more and more like a Gaussian density. Also, as we have seen from Characteristic Functions in Exercise 3.19, they transform convolution into multiplication (see also Proposition 8.6(d)). Since multiplication is easier to understand than convolution itself, it is then natural to examine the characteristic functions of sums of independent random variables.

Theorem 3.23 (Central Limit Theorem). *Let $X_1, X_2, \ldots$ be real-valued, independent, identically distributed random variables. Let Z be a standard Gaussian random variable. Let $\mu, \sigma \in \mathbb{R}$ with $\sigma > 0$. Assume that $\mathbf{E}X_1 = \mu$ and $\mathrm{var}(X_1) = \sigma^2$. Then, as $n \to \infty$, $\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ converges in distribution to Z. That is, for any $t \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le t\right) = \int_{-\infty}^{t} e^{-x^2/2}\frac{dx}{\sqrt{2\pi}}.$$

Remark 3.24. The random variable $\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ has mean zero and variance 1, just like the standard Gaussian Z. So, the normalizations of $X_1 + \cdots + X_n$ we have chosen are sensible.

Exercise 3.25. Let $f, g, h\colon \mathbb{R} \to \mathbb{C}$. We use the notation $f(s) = o(g(s))\ \forall\ s \in \mathbb{R}$ to denote $\lim_{s\to 0}\left|\frac{f(s)}{g(s)}\right| = 0$. For example, if $f(s) = s^3\ \forall\ s \in \mathbb{R}$, then $f(s) = o(s^2)$, since $\lim_{s\to 0}\left|\frac{f(s)}{s^2}\right| = \lim_{s\to 0}|s| = 0$. Show: (i) if $f(s) = o(g(s))$ and if $h(s) = o(g(s))$, then $(f + h)(s) = o(g(s))$. (ii) If c is any nonzero constant, then $o(cg(s)) = o(g(s))$. (iii) $\lim_{s\to 0} g(s)o(1/g(s)) = 0$. (iv) $\lim_{s\to 0} o(g(s))/g(s) = 0$. (v) $o(g(s) + o(g(s))) = o(g(s))$.

Proof using Fourier Transform. For any $j \ge 1$, let $Y_j := (X_j - \mu)/\sigma$. Then $Y_1, Y_2, \ldots$ are independent and identically distributed by Remark 1.71, $\mathbf{E}Y_j = 0$ and $\mathbf{E}Y_j^2 = 1$, $\forall\ j \ge 1$. We will show that $\lim_{n\to\infty} \mathbf{P}(\frac{Y_1+\cdots+Y_n}{\sqrt{n}} \le t) = \mathbf{P}(Z \le t)$, $\forall\ t \in \mathbb{R}$. From Theorem 3.20 and Proposition 8.7, it suffices to show:

$$\lim_{n\to\infty} \mathbf{E}e^{it\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = \mathbf{E}e^{itZ} = e^{-t^2/2}, \qquad \forall\, t \in \mathbb{R}.$$

From Exercise 3.19,

$$\mathbf{E}e^{it\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = \prod_{j=1}^{n} \mathbf{E}e^{itY_j/\sqrt{n}} = (\mathbf{E}e^{itY_1/\sqrt{n}})^n.$$

By Exercise 3.17 with $k = 2$, and using $\mathbf{E}Y_1 = 0$ and $\mathbf{E}Y_1^2 = 1$,

$$\mathbf{E}e^{itY_1/\sqrt{n}} = 1 + \frac{it}{\sqrt{n}}\mathbf{E}Y_1 - \frac{t^2}{2n}\mathbf{E}Y_1^2 + o(t^2/n) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right).$$

Therefore,

$$\mathbf{E}e^{it\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n.$$

Taking logarithms, using $\log(1+x) = x + o(x)$ for $-1 < x < 1$, and using Exercise 3.25,

$$\log \mathbf{E}e^{it\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = n \log\left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right) = -\frac{t^2}{2} + n \cdot o\left(\frac{t^2}{n}\right).$$

Letting $n \to \infty$ and using Exercise 3.25(iii) completes the proof. □

Proof using Lindeberg replacement, assuming finite third moment. As in the proof above, we may assume $\mathbf{E}X_1 = 0$ and $\mathbf{E}X_1^2 = 1$. Let $g\colon \mathbb{R} \to \mathbb{R}$ be a Schwartz function. Fix $n \geq 1$. Consider $\mathbf{E}g((X_1+\cdots+X_n)/\sqrt{n})$. We will show this quantity is close to $\mathbf{E}g(Z)$ by replacing one term of $X_1, \ldots, X_n$ at a time with an independent standard Gaussian. Let $Z_1, Z_2, \ldots$ be independent standard Gaussian random variables, independent of $X_1, X_2, \ldots$. Note that $(Z_1 + \cdots + Z_n)/\sqrt{n}$ is a standard Gaussian by Exercise 3.13. We write a telescoping sum:

$$\begin{aligned}
&\left|\mathbf{E}g\Big(\frac{X_1+\cdots+X_n}{\sqrt{n}}\Big) - \mathbf{E}g(Z)\right| = \left|\mathbf{E}g\Big(\frac{X_1+\cdots+X_n}{\sqrt{n}}\Big) - \mathbf{E}g\Big(\frac{Z_1+\cdots+Z_n}{\sqrt{n}}\Big)\right| \\
&= \left|\sum_{j=1}^n \Big(\mathbf{E}g\Big(\frac{X_1+\cdots+X_j+Z_{j+1}+\cdots+Z_n}{\sqrt{n}}\Big) - \mathbf{E}g\Big(\frac{X_1+\cdots+X_{j-1}+Z_j+\cdots+Z_n}{\sqrt{n}}\Big)\Big)\right| \\
&\leq \sum_{j=1}^n \left|\mathbf{E}g\Big(\frac{X_1+\cdots+X_j+Z_{j+1}+\cdots+Z_n}{\sqrt{n}}\Big) - \mathbf{E}g\Big(\frac{X_1+\cdots+X_{j-1}+Z_j+\cdots+Z_n}{\sqrt{n}}\Big)\right|. \quad (*)
\end{aligned}$$

We control each term in the sum separately. Write g in its third order Taylor expansion as $g(y+t) = g(y) + tg'(y) + t^2g''(y)/2 + O(|t|^3)$. Fix $1 \leq j \leq n$. Using $Y := (X_1+\cdots+X_{j-1}+Z_{j+1}+\cdots+Z_n)/\sqrt{n}$, and $T_1 = X_j/\sqrt{n}$ and also $T_2 = Z_j/\sqrt{n}$, then taking expected values,

$$\begin{aligned}
&\left|\mathbf{E}g\Big(\frac{X_1+\cdots+X_j+Z_{j+1}+\cdots+Z_n}{\sqrt{n}}\Big) - \mathbf{E}g\Big(\frac{X_1+\cdots+X_{j-1}+Z_j+\cdots+Z_n}{\sqrt{n}}\Big)\right| \\
&= |\mathbf{E}g(Y+T_1) - \mathbf{E}g(Y+T_2)| \\
&= \left|\frac{\mathbf{E}X_jg'(Y)}{\sqrt{n}} - \frac{\mathbf{E}Z_jg'(Y)}{\sqrt{n}} + \frac{1}{2n}(\mathbf{E}X_j^2g''(Y) - \mathbf{E}Z_j^2g''(Y)) + O\Big(\mathbf{E}\frac{|X_j|^3}{n^{3/2}}\Big) + O\Big(\mathbf{E}\frac{|Z_j|^3}{n^{3/2}}\Big)\right|.
\end{aligned}$$

Since X_j is independent of Y, Proposition 1.72 implies that $\mathbf{E}X_jg'(Y) = \mathbf{E}X_j\mathbf{E}g'(Y) = 0$. Similarly, $\mathbf{E}Z_jg'(Y) = 0$. The next terms also cancel since $\mathbf{E}X_j^2 = \mathbf{E}Z_j^2 = 1$. In summary,

$$\left|\mathbf{E}g\Big(\frac{X_1+\cdots+X_j+Z_{j+1}+\cdots+Z_n}{\sqrt{n}}\Big) - \mathbf{E}g\Big(\frac{X_1+\cdots+X_{j-1}+Z_j+\cdots+Z_n}{\sqrt{n}}\Big)\right| \leq O\Big(\frac{1+\mathbf{E}|X_1|^3}{n^{3/2}}\Big).$$

Substituting this back into $(*)$, we get

$$\left|\mathbf{E}g\Big(\frac{X_1+\cdots+X_n}{\sqrt{n}}\Big)-\mathbf{E}g(Z)\right| \le O\Big(\frac{1+\mathbf{E}\,|X_1|^3}{\sqrt{n}}\Big), \qquad \forall\, n\ge 1.$$

Letting $n\to\infty$ shows that $\lim_{n\to\infty}\mathbf{E}g(\frac{X_1+\cdots+X_n}{\sqrt{n}})=\mathbf{E}g(Z)$. Since this holds for any Schwartz function $g\colon\mathbb{R}\to\mathbb{R}$, it then holds for any continuous compactly supported function g. We then conclude by Proposition 3.3. □

Remark 3.26. The Lindeberg argument has the advantage of providing a quantitative bound for the convergence that occurs in the Central Limit Theorem. In particular, we showed that, there exists an absolute constant $c>0$ such that, if $X_1, X_2, \dots$ are i.i.d. with mean zero, variance 1 and $\mathbf{E}\,|X_1|^3<\infty$, then for any three times continuously differentiable compactly supported function $g\colon\mathbb{R}\to\mathbb{R}$,

$$\left|\mathbf{E}g\Big(\frac{X_1+\cdots+X_n}{\sqrt{n}}\Big)-\mathbf{E}g(Z)\right| \le c\cdot\sup_{x\in\mathbb{R}}|g'''(x)|\,\Big(\frac{1+\mathbf{E}\,|X_1|^3}{\sqrt{n}}\Big)$$

Exercise 3.27 (Weak Berry-Esséen theorem)**.** Let $X, X_1, X_2, \dots$ be i.i.d. real-valued random variables with mean zero, variance 1 and with $\mathbf{E}\,|X|^3<\infty$. Let Z be a standard Gaussian random variable.

(i) Show that for any compactly supported $g\colon\mathbb{R}\to\mathbb{R}$ with three continuous derivatives, and for any $n\ge 1$,

$$\mathbf{E}g\Big(\frac{X_1+\cdots+X_n}{\sqrt{n}}\Big)=\mathbf{E}g(Z)+O(n^{-1/2}\sup_{x\in\mathbb{R}}|g'''(x)|\,\mathbf{E}|X|^3),$$

where the implied constant does not depend on g, n or on any of the random variables $X, X_1, X_2, \dots$.

(ii) Show that, for any $n\ge 1$, and for any $t\in\mathbb{R}$

$$\mathbf{P}\Big(\frac{X_1+\cdots+X_n}{\sqrt{n}}\le t\Big)=\mathbf{P}(Z\le t)+O(n^{-1/2}\mathbf{E}|X|^3)^{1/4},$$

where the implied constant does not depend on n, t or on any of the random variables $X, X_1, X_2, \dots$.

(Hint: for the second item, consider $g=1_{[-\infty,t]}*\phi_\varepsilon$, where $\phi(x)=e^{-x^2/2}/\sqrt{2\pi}$ and $\phi_\varepsilon(x)=\varepsilon^{-1}\phi(x/\varepsilon)$ for any $x\in\mathbb{R}$ and for appropriately chosen $\varepsilon>0$.)

Recall from Exercise 1.64 that a Schwartz function $g\colon\mathbb{R}\to\mathbb{R}$ and a standard Gaussian Z satisfy $\mathbf{E}Zg(Z)=\mathbf{E}g'(Z)$. The next theorem shows that this equality can characterize how far a random variable is from being a standard Gaussian.

Theorem 3.28 (**Stein Continuity Theorem**)**.** *Let $X_1, X_2, \dots$ be a sequence of real-valued random variables with $\sup_{n\ge1}\mathbf{E}\,|X_n|^2<\infty$. Let Z be a standard Gaussian random variable. Then the following are equivalent.*

(i) *For any differentiable function $g\colon\mathbb{R}\to\mathbb{R}$ such that g and g' are bounded and continuous,*

$$\lim_{n\to\infty}\mathbf{E}(X_ng(X_n)-g'(X_n))=0.$$

(ii) *$X_1, X_2, \dots$ converges in distribution to Z.*

Proof. First, assume (ii) occurs. Since g' is bounded, $\lim_{n\to\infty} \mathbf{E}g'(X_n) = \mathbf{E}g'(Z)$ by Exercise 3.8(vi). To prove (i), it then suffices by Exercise 1.64 to show that $\lim_{n\to\infty} \mathbf{E}X_n g(X_n) = \mathbf{E}Zg(Z)$. For the purpose of proving this equality, we may assume by Theorem 1.60 and Exercise 3.8(iii) that $X_1, X_2, \ldots$ converges almost surely to Z. By assumption,

$$\sup_{n\geq 1} \mathbf{E}\left|X_n g(X_n)\right|^2 \leq \sup_{n\geq 1} \mathbf{E}X_n^2 \cdot \sup_{x\in\mathbb{R}} \left|g(x)\right|^2 < \infty.$$

Theorem 1.59 then implies that $\lim_{n\to\infty} \mathbf{E}X_n g(X_n) = \mathbf{E}Zg(Z)$, proving (i).

Now, assume (i). By Proposition 3.3, it suffices to show that $\lim_{n\to\infty} \mathbf{E}\phi(X_n) - \mathbf{E}\phi(Z) = 0$ for any continuous compactly supported $\phi\colon \mathbb{R}\to\mathbb{R}$. It further suffices to assume that ϕ is a Schwartz function bounded by 1 in absolute value. We now claim it suffices to find a function $g\colon \mathbb{R}\to\mathbb{R}$ with both g, g' bounded such that g satisfies the ODE

$$\phi(x) - \mathbf{E}\phi(Z) = g'(x) - xg(x), \qquad \forall\, x\in\mathbb{R}. \qquad (*)$$

To see that this implies (ii), let $x = X_n$, take expected values, then let $n\to\infty$ to get

$$\lim_{n\to\infty} \mathbf{E}\phi(X_n) - \mathbf{E}\phi(Z) = \lim_{n\to\infty} \mathbf{E}(g'(X_n) - X_n g(X_n)) \overset{(i)}{=} 0.$$

In order to solve the ODE $(*)$, we use the method of integrating factors to get

$$g(x) = e^{x^2/2}\int_{-\infty}^{x} e^{-y^2/2}(\phi(y) - \mathbf{E}\phi(Z))dy = -e^{x^2/2}\int_{x}^{\infty} e^{-y^2/2}(\phi(y) - \mathbf{E}\phi(Z))dy,\ \forall\, x\in\mathbb{R}. \ (**)$$

The last equality follows since $\int_{-\infty}^{\infty} e^{-y^2/2}(\phi(y) - \mathbf{E}\phi(Z))dy/\sqrt{2\pi} = 0$. It remains to show that g, g' are bounded. If $x, y\in\mathbb{R}$ we write

$$e^{-y^2/2} = e^{-(y-x+x)^2/2} = e^{-x^2/2}e^{-(y-x)^2/2}e^{-x(y-x)} \leq e^{-x^2/2}e^{-x(y-x)}. \qquad (\ddagger)$$

Using $(\ddagger)$ when $y > x$ and $x > 1$, the second equality of $(**)$ implies that $|g(x)| \leq 2/|x|$ since $\int_x^\infty e^{-x(y-x)}dy = \int_0^\infty e^{-xy}dy = 1/x$. Using $(\ddagger)$ with $y < x$ and $x < -1$, the first equality of $(**)$ implies that $|g(x)| \leq 2/|x|$. Either equality of $(*)$ implies $|g(x)| \leq 5$ when $-1 < x < 1$. In summary,

$$|g(x)| \leq \frac{10}{1+|x|}, \qquad \forall\, x\in\mathbb{R}.$$

So, by $(*)$, g' and g are bounded by 20 in absolute value, as desired. □

Remark 3.29. Changing variables in $(**)$ shows that, for any $x\in\mathbb{R}$,

$$g(x) = -e^{x^2/2}\int_0^\infty e^{-(y+x)^2/2}(\phi(y+x) - \mathbf{E}\phi(Z))dy = -\int_0^\infty e^{-y^2/2}e^{-yx}(\phi(y+x) - \mathbf{E}\phi(Z))dy.$$

Differentiating in x and repeating the above argument (using $\sup_{y\in\mathbb{R}} |ye^{-y^2/2}| \leq 1$) gives

$$|g'(x)| \leq 10\cdot\frac{1+\sup_{y\in\mathbb{R}}|\phi'(y)|}{1+|x|}, \qquad \forall\, x\in\mathbb{R}.$$

Proof of Central Limit Theorem using Stein's Method. As in the proofs above, we may assume $\mathbf{E}X_1 = 0$ and $\mathbf{E}X_1^2 = 1$. Let $\phi\colon \mathbb{R}\to[-1,1]$ be a Schwartz function. Let Z be a standard Gaussian. We additionally assume that $\mathbf{E}\left|X_1\right|^3 < \infty$. For any $n\geq 1$, let

$Y_n := (X_1 + \cdots + X_n)/\sqrt{n}$. We will show the following Berry-Esseén type quantitative estimate for any $n \geq 1$:

$$\mathbf{E}\phi(Y_n) = \mathbf{E}\phi(Z) + O(\mathbf{E}\,|X_1|^3/\sqrt{n}) \cdot (1 + \sup_{y\in\mathbb{R}} |\phi'(y)|).$$

Beginning with $(*)$ from the proof of Theorem 3.28, we get

$$\mathbf{E}\phi(Y_n) - \mathbf{E}\phi(Z) = \mathbf{E}\Big(g'(Y_n) - Y_n g(Y_n)\Big).$$

We write $Y_n g(Y_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i g(Y_n)$, and we consider each individual term in the sum. For any $n \geq 1$, let $Y_{n,i} := Y_n - X_i/\sqrt{n}$. By the fundamental theorem of calculus applied to g,

$$\mathbf{E}X_i g(Y_n) = \mathbf{E}\Big(X_i g(Y_{n,i}) + \frac{1}{\sqrt{n}} X_i^2 g'(Y_{n,i} + \frac{T}{\sqrt{n}} X_i)\Big), \qquad (**)$$

where T is uniformly distributed in $[0,1]$ and independent of $X_1, \ldots, X_n$. By independence $\mathbf{E}[X_i g(Y_{n,i})] = \mathbf{E}X_i \mathbf{E}g(Y_{n,i}) = 0$. Combining the above,

$$\begin{aligned}
\mathbf{E}\phi(Y_n) - \mathbf{E}\phi(Z) = \mathbf{E}\Big(g'(Y_n) - Y_n g(Y_n)\Big) &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\Big(g'(Y_n) - \sqrt{n} X_i g(Y_n)\Big) \\
&\overset{(**)}{=} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\Big(g'(Y_n) - X_i^2 g'(Y_{n,i} + \frac{T}{\sqrt{n}} X_i)\Big) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\Big(g'(Y_n) - g'(Y_{n,i}) - X_i^2\Big[g'(Y_{n,i} + \frac{T}{\sqrt{n}} X_i) - g'(Y_{n,i})\Big]\Big). \qquad (***)
\end{aligned}$$

In the last line, we used independence to get $\mathbf{E}g'(Y_{n,i}) = \mathbf{E}X_i^2 \mathbf{E}g'(Y_{n,i}) = \mathbf{E}X_i^2 g'(Y_{n,i})$. Using now Remark 3.29, and defining $c := 10(1 + \sup_{y\in\mathbb{R}} |\phi'(y)|)$, we have $|(xg(x))'| \leq 2c$ for all $x \in \mathbb{R}$. Differentiating $(*)$, we get $|g''(x)| \leq 3c$ for all $x \in \mathbb{R}$. So, $(***)$ can be rewritten as

$$\mathbf{E}\phi(Y_n) - \mathbf{E}\phi(Z) = \frac{1}{n}\sum_{i=1}^n \frac{1}{\sqrt{n}} \mathbf{E}O(c\,|X_i| + c\,|X_i|^3) = \frac{c}{\sqrt{n}} O(\mathbf{E}\,|X_1|^3),$$

using the Fundamental Theorem of Calculus. □

Exercise 3.30. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a Schwartz function. Let Z be a standard Gaussian random variable. In applications of Stein's method, it is sometimes more convenient to take another derivative of Stein's identity, resulting in the following Ornstein-Uhlenbeck identities.

- $\mathbf{E}[\phi''(Z) - Z\phi'(Z)] = 0$.
- If $h\colon \mathbb{R} \to \mathbb{R}$ is a Schwartz function, then the function

$$g(x) := \int_0^1 \frac{1}{2t}[\mathbf{E}h(x\sqrt{t} + Z\sqrt{1-t}) - \mathbf{E}h(Z)]dt, \qquad \forall\, x \in \mathbb{R},$$

  is a solution of the differential equation

$$h(x) - \mathbf{E}h(Z) = g''(x) - xg'(x), \qquad \forall\, x \in \mathbb{R}.$$

Exercise 3.31. Using the Central Limit Theorem, prove the Weak Law of Large Numbers (assume the random variables have mean zero and variance one).

The Lindeberg replacement argument and Stein's Method make quantitative estimates for the distributions of *nonlinear* functions of independent random variables. For this reason, these methods have gained renewed interest in the last two decades. For example, if we consider a matrix of i.i.d. random variables, then one is naturally led to consider nonlinear functions of the matrix entries when considering e.g. the eigenvalues (or singular values) of the matrix.

The Fourier analytic proof of the Central limit Theorem, though rather elegant, seems only able to analyze the distribution of linear functions of independent random variables.

3.5. **Additional Comments.** The Central Limit Theorem was described by de Moivre in 1733 and again by Laplace in 1785 and 1812, where the Fourier Transform was used. In 1901, Lyapunov proved the Central Limit Theorem under an assumption similar to $\mathbf{E}\,|X_1|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$. The Central Limit Theorem under the assumption of a finite (truncated) second moment was proven by Lindeberg in 1920. This result was extended by Feller in 1935, also with contributions by Lévy in the same year.

Theorem 3.32 (**Lindeberg Central Limit Theorem for Triangular Arrays**). *Let $j_1, j_2, \dots$ be a sequence of natural numbers with $\lim_{n\to\infty} j_n = \infty$. For any $n \geq 1$, let $X_{n,1}, \dots, X_{n,n} \colon \Omega_n \to \mathbb{R}$ be independent with mean zero and finite variance. (Note e.g. that $X_{3,1}$ and $X_{2,2}$ might not be independent, and the sample space is allowed to change as n changes.) Define*

$$\sigma_n^2 := \sum_{k=1}^{j_n} \mathrm{Var}(X_{n,k}), \qquad \forall\, n \geq 1.$$

Assume that $\sigma_n > 0$ for all $n \geq 1$. If, for any $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \frac{1}{\sigma_n^2} \sum_{k=1}^{j_n} \mathbf{E}(|X_{n,k}|^2 1_{|X_{n,k}|>\varepsilon\sigma_n}) = 0, \qquad (*)$$

then the random variables $\frac{X_{n,1}+\cdots+X_{n,n}}{\sigma_n}$ converge in distribution to a standard Gaussian random variable.

The Lindeberg condition $(*)$ implies the Feller condition

$$\lim_{n\to\infty} \frac{1}{\sigma_n^2} \max_{1\leq k\leq n} \mathbf{E}|X_{n,k}|^2 = 0.$$

It was shown by Feller that if the above assumptions hold (without $(*)$) and if the Feller condition holds, then the Lindeberg condition $(*)$ is necessary and sufficient for $\frac{X_{n,1}+\cdots+X_{n,n}}{\sigma_n}$ to converge in distribution to a standard Gaussian random variable. The combined result is sometimes known as the Lindeberg-Feller theorem.

Berry and Esseén separately gave an error bound for the Central Limit Theorem in the early 1940s. Above, we only proved weaker versions of this theorem, though the methods discussed above can be used to prove the stronger statement below.

Theorem 3.33 (**Berry-Esseén**). *There exists $c > 0$ such that the following holds. Let $X_1, X_2, \dots$ be i.i.d. real-valued random variables with mean zero, variance 1 and $\mathbf{E}\,|X_1|^3 <$*

∞. *Let Z be a standard Gaussian random variable. Then for any $n \geq 1$,*

$$\sup_{t\in\mathbb{R}} \left|\mathbf{P}(X_1 + \cdots + X_n/\sqrt{n} < t) - \mathbf{P}(Z < t)\right| \leq c \cdot \frac{\mathbf{E}\,|X_1|^3}{\sqrt{n}}.$$

With the assumption of more bounded moments, an asymptotic expansion can be written, with explicit dependence on t, for the difference $|\mathbf{P}(X_1 + \cdots + X_n/\sqrt{n} < t) - \mathbf{P}(Z < t)|$. This expansion is called the Edgeworth Expansion; see Feller, Vol. 2, XVI.4.(4.1).

One may ask for general conditions under which the average of any i.i.d. random variables have a limiting distribution, with moment assumptions different than the Central Limit Theorem. Necessary and sufficient conditions are described in the following Theorem.

Theorem 3.34. *Let $X_1, X_2, \ldots$ be i.i.d. real-valued random variables. Assume there exists a function $h\colon [0,\infty) \to (0,\infty)$ such that, for any $x > 0$, $\lim_{x\to\infty} L(tx)/L(x) = 1$. Assume also there exists $\theta \in [0,1]$ and $\alpha \in (0,2)$such that*

- $\lim_{x\to\infty} \mathbf{P}(X_1 > x)/\mathbf{P}(|X_1| > x) = \theta$,
- $\mathbf{P}(|X_1| > x) = x^{-\alpha}L(x)$, $\forall\ x > 0$.

For any $n \geq 1$, define

$$a_n := \inf\{x > 0\colon P(|X_1| > x) \leq 1/n\}, \qquad b_n := \mathbf{E}(X_1 1_{|X_1|\leq a_n}).$$

Then $\frac{X_1+\cdots+X_n-a_n}{b_n}$ converges in distribution to a random variable Y as $n \to \infty$

Exercise 3.35. Show that there exists a nonzero random variable X such that, if $X_1, X_2, \ldots$ are i.i.d. copies of X, then $\frac{X_1+\cdots+X_n}{n}$ is equal in distribution to X, for any $n \geq 1$. (Optional: can you write out an explicit formula for the density of X?) (Hint: take the Fourier transform.)

Show that there exists a nonzero random variable X such that, if $X_1, X_2, \ldots$ are i.i.d. copies of X, then $\frac{X_1+\cdots+X_n}{n^2}$ is equal in distribution to X, for any $n \geq 1$.

By projection the random variables onto one-dimensional lines, the following Central Limit Theorem in $\mathbb{R}^d$ can be proven from the corresponding result in $\mathbb{R}$.

Theorem 3.36 (Central Limit Theorem in $\mathbb{R}^d$). *Let $X^{(1)}, X^{(2)}, \ldots$ be i.i.d. $\mathbb{R}^d$-valued random variables. Let $\mu \in \mathbb{R}^d$. (We write a random variable in its components as $X^{(n)} = (X_1^{(n)}, \ldots, X_d^{(n)}) \in \mathbb{R}^d$.) Assume $\mathbf{E}X^{(n)} = \mu$ for all $n \geq 1$, and for any $1 \leq i, j \leq d$, all of the covariances*

$$a_{ij} := \mathbf{E}((X_i^{(1)} - \mathbf{E}X_i^{(1)})(X_j^{(1)} - \mathbf{E}X_j^{(1)})).$$

are finite. Then as $n \to \infty$, $\frac{X^{(1)}+\cdots+X^{(n)}-n\mu}{\sqrt{n}}$ converges weakly to a Gaussian random vector $Z = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$ with covariance matrix $(a_{ij})_{1\leq i,j\leq d}$.

Remark 3.37. By definition, a random vector $Z = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$ is **Gaussian** if, for any $v_1, \ldots, v_d \in \mathbb{R}$, the random variable $\sum_{i=1}^d v_i Z_i$ is a Gaussian random variable. Equivalently, for any $v \in \mathbb{R}^d$, the random variable $\langle v, Z\rangle$ is a Gaussian random variable. The covariance matrix $(a_{ij})_{1\leq i,j\leq d}$ of Z is defined by

$$a_{ij} := \mathbf{E}((Z_i - \mathbf{E}Z_i)(Z_j - \mathbf{E}Z_j)).$$

Exercise 3.38. Let $Z = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$ be a Gaussian random vector.

- Show that the covariance matrix $(a_{ij})_{1\le i,j\le d}$ of Z is symmetric, positive semidefinite. That is, for any $v \in \mathbb{R}^d$, we have
$$v^T a v = \sum_{i,j=1}^{d} v_i v_j a_{ij} \ge 0.$$
- Given any symmetric positive semidefinite matrix $(b_{ij})_{1\le i,j\le d}$, show that there exists a Gaussian random vector Z such that the covariance matrix of Z is $(b_{ij})_{1\le i,j\le d}$. (Hint: write the matrix b in its Cholesky decomposition $b = rr^*$, where r is a $d \times d$ real matrix. Let $e^{(1)}, \dots, e^{(d)}$ be the rows of r. Let $X_1, \dots, X_d$ be independent standard Gaussian random variables. Let $X := (X_1, \dots, X_d)$. Define $Z_i := \langle X, e^{(i)} \rangle$ for any $1 \le i \le d$.)

Both Stein's Method and the Lindeberg Replacement argument have gained renewed interest in the last two decades. For a survey on Stein's Method, see Chatterjee06 or Chatterjee14. For some general results using the Lindeberg replacement method, see MOO05, Chatterjee07 or TaoVu12.

4. Random Walks

In the Strong Law of Large Numbers and Central Limit Theorem, we investigated the distribution of a sum of i.i.d. random variables $X_1, X_2, \dots$. In this section, we change our perspective and investigate what values the sum takes. For example, a basic question is, " Is $X_1 + \cdots + X_n = 0$ for finitely many $n \ge 1$? "

Our first result in this direction is a generalization of Kolmogorov's Zero-One Law, Theorem 1.95. For technical reasons, we use an explicit construction of the sample space Ω in this section.

Definition 4.1 (**Random Walk**). Let $\mathbb{N} := \{1, 2, 3, \dots\}$ and let $d \ge 1$. Let $\Omega := (\mathbb{R}^d)^{\mathbb{N}}$. Let $X \colon \mathbb{R}^d \to \mathbb{R}^d$ be a random variable. We construct i.i.d. copies of X using Theorem 1.80 and Corollary 1.81. The probability measure $\prod_{j=1}^{\infty} \mu_X$ exists on Ω. So, for any $(\omega_1, \omega_2, \dots) \in \Omega$, let $X_j(\omega_1, \omega_2, \dots) := \omega_j$. Then $X_1, X_2, \dots$ are i.i.d. copies of X.

Let $x \in \mathbb{R}^d$. Let $X_0 := x$. For any $n \ge 0$, let $S_n := X_0 + \cdots + X_n$. We call the sequence of random variables $S_0, S_1, \dots$ a **random walk** on $\mathbb{R}^d$ started at x.

4.1. Limiting Behavior.

Definition 4.2 (**Exchangeable σ-algebra**). A finite permutation of $\mathbb{N}$ is a bijective map $\pi \colon \mathbb{N} \to \mathbb{N}$ such that $\pi(j) \ne j$ for only finitely many $j \in \mathbb{N}$. For any $\omega = (\omega_1, \omega_2, \dots) \in \Omega$, we define $\pi\omega \in \Omega$ so that $(\pi\omega)_j := \omega_{\pi(j)}$ for any $j \in \mathbb{N}$. For any $A \subseteq \Omega$, we define $\pi^{-1}A := \{\omega \in \Omega \colon \pi\omega \in A\}$. An event $A \subseteq \Omega$ is **permutable** if, for any finite permutation π, we have $\pi^{-1}A = A$. The collection of all permutable events, denoted $\mathcal{E}$, is a σ-algebra referred to as the **exchangeable** σ-algebra of $X_1, X_2, \dots$.

As usual, we equip Ω with the product σ-algebra defined in Example 1.8. So, an event $A \subseteq \Omega$ is defined to be an element of this σ-algebra. Note that the product σ-algebra is equal to $\sigma(X_1, X_2, \dots)$, by the definition of $X_1, X_2, \dots$.

Since $X_j(\omega) = \omega_j$ for any $j \in \mathbb{N}$, and for any $\omega \in \Omega$, note that
$$X_j(\pi\omega) = \omega_{\pi(j)} = X_{\pi(j)}(\omega).$$

In this way, applying π to Ω permutes the random variables $X_1, X_2, \ldots$. So, an event $A \subseteq \Omega$ is permutable if it is not affected by a finite permutation of the random variables $X_1, X_2, \ldots$.

Example 4.3. Let B be a measurable subset of $\mathbb{R}^d$. Let $c_1 < c_2 < \cdots$. Consider the events

$$\{\omega \in \Omega \colon S_n(\omega) \in B \text{ for infinitely many } n \geq 1\}.$$

$$\{\omega \in \Omega \colon \limsup_{n\to\infty} S_n(\omega)/c_n \geq 1\}.$$

If π is a finite permutation, then there exists $n \in \mathbb{N}$ such that $\pi(j) = j$ for all $j \geq n$, so that $S_j(\pi\omega) = S_j(\omega)$ for all $j \geq n$, and for all $\omega \in \Omega$. So, both of the above events are in $\mathcal{E}$.

Proposition 4.4. *The exchangeable σ-algebra strictly contains the tail σ-algebra: $\mathcal{E} \supsetneq \mathcal{T}$.*

Proof. Let $A \in \mathcal{T}$ and let π be a finite permutation. Let $n \geq 1$ so that $\pi(j) = j$ for all $j \geq n$. By definition of $\mathcal{T}$, $A \in \sigma(X_n, X_{n+1}, \ldots)$. By the definition of n, $\pi^{-1}B = B$ for any $B \in \sigma(X_n, X_{n+1}, \ldots)$. Therefore, $\pi^{-1}A = A$, so that $A \in \mathcal{E}$. Strict containment follows since the second event of Example 4.3 is not in $\mathcal{T}$ by Exercise 1.97 using $c_n = 1$ for all $n \geq 1$. (The first event of Example 4.3 is also not in $\mathcal{E}$.) □

Since $\mathcal{E} \supseteq \mathcal{T}$, Theorem 4.5 generalizes Kolmogorov's Zero-One Law, Theorem 1.95.

Theorem 4.5 (Hewitt-Savage Zero-One Law). *Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}^d$ be i.i.d. Let $A \in \mathcal{E}$. Then $\mathbf{P}(A) \in \{0, 1\}$.*

Proof. Let $A \in \mathcal{E}$. As in Kolmogorov's Zero-One Law, Theorem 1.95, we will show A is independent of itself.

Note that $A \in \sigma(X_1, X_2, \ldots)$. From Exercise 1.91 and recalling the sketched proof of the Extension Theorem, Theorem 1.18, for any $n \geq 1$ there exists $A_n \in \sigma(X_1, \ldots, X_n)$ such that

$$\lim_{n\to\infty} \mathbf{P}(A_n \Delta A) = 0. \qquad (*)$$

By the definition of Ω, $\forall\ n \geq 1$, $\exists$ a measurable set $B_n \subseteq (\mathbb{R}^d)^n$ such that $A_n = \{\omega \in \Omega \colon (\omega_1, \ldots \omega_n) \in B_n\}$. Fix $n \geq 1$ and let $\pi = \pi_n \colon \mathbb{N} \to \mathbb{N}$ be the finite permutation

$$\pi(j) := \begin{cases} j + n & \text{, if } 1 \leq j \leq n \\ j - n & \text{, if } n + 1 \leq j \leq 2n \\ j & \text{, if } j \geq 2n + 1. \end{cases}$$

Since $X_1, X_2, \ldots$ are i.i.d., $\mathbf{P}$ is permutation invariant, so that

$$\mathbf{P}(\omega \in \Omega \colon \omega \in A_n \Delta A) = \mathbf{P}(\omega \in \Omega \colon \pi\omega \in A_n \Delta A). \qquad (**)$$

We rewrite the last event. Since A is permutable, $\{\omega \in \Omega \colon \pi\omega \in A\} = \{\omega \in \Omega \colon \omega \in A\}$. Also, by the definition of B_n and π, $\{\omega \in \Omega \colon \pi\omega \in A_n\} = \{\omega \in \Omega \colon (\omega_{n+1}, \ldots, \omega_{2n}) \in B_n\} =: A'_n$. Combining these observations, $(**)$ becomes

$$\mathbf{P}(A_n \Delta A) = \mathbf{P}(A'_n \Delta A). \qquad (***)$$

For any event C, we have $|\mathbf{P}(C) - \mathbf{P}(A)| \leq \mathbf{P}(C \Delta A)$, so $(*)$ and $(***)$ imply that $\lim_{n\to\infty} \mathbf{P}(A_n) = \lim_{n\to\infty} \mathbf{P}(A'_n) = \mathbf{P}(A)$. So, since A_n, A'_n are independent,

$$\lim_{n\to\infty} \mathbf{P}(A_n \cap A'_n) = \lim_{n\to\infty} \mathbf{P}(A_n)\mathbf{P}(A'_n) = \mathbf{P}(A)^2. \qquad (\ddagger)$$

We now investigate the left side. For any events B, C, we have $B \setminus C \subseteq (B \setminus A) \cup (A \setminus C)$, with a similar containment for $C \setminus B$, together implying that $B \Delta C \subseteq (B \Delta A) \cup (A \Delta C)$. So, by subadditivity of $\mathbf{P}$,

$$\mathbf{P}(A_n \Delta A) + \mathbf{P}(A \Delta A'_n) \geq \mathbf{P}(A_n \Delta A'_n) = \mathbf{P}(A_n \cup A'_n) - \mathbf{P}(A_n \cap A'_n) \geq \mathbf{P}(A_n) - \mathbf{P}(A_n \cap A'_n) \geq 0.$$

Letting $n \to \infty$, $(*)$ and $(***)$ imply that the left side tends to zero. So,

$$\lim_{n \to \infty} \mathbf{P}(A_n \cap A'_n) = \lim_{n \to \infty} \mathbf{P}(A_n) \overset{(*)}{=} \mathbf{P}(A).$$

So, $(\ddagger)$ says that $\mathbf{P}(A) = \mathbf{P}(A)^2$, as desired. □

Theorem 4.6. *Let $S_0, S_1, \ldots$ be a random walk on $\mathbb{R}$ with $S_0 = 0$. Exactly one of the following four conditions holds with probability one.*

(i) $S_n = 0$ *for all* $n \geq 1$.
(ii) $\lim_{n \to \infty} S_n = \infty$.
(iii) $\lim_{n \to \infty} S_n = -\infty$.
(iv) $-\infty = \liminf_{n \to \infty} S_n$ *and* $\limsup_{n \to \infty} S_n = \infty$.

Proof. From Example 4.3 and Theorem 4.5, $\exists\, c \in [-\infty, \infty]$ such that, with probability one, $\limsup_{n \to \infty} S_n = c$. For any $n \geq 1$, define $S'_n := S_n - X_1$. Then $S'_2, S'_3, \ldots$ each have the same distribution as $S_1, S_2, \ldots$ so with probability one, $\limsup_{n \to \infty} S'_n = c$. That is, $c = c - X_1$ with probability 1. If $\mathbf{P}(X_1 = 0) = 1$, then (i) occurs. From now on, we assume $\mathbf{P}(X_1 = 0) < 1$, so that $c = \infty$ or $c = -\infty$. Arguing similarly for $c' := \liminf_{n \to \infty} S_n \in \{-\infty, \infty\}$, there are four possible values of the ordered pair (c, c'), though $c' \leq c$ eliminates the case $c' = \infty$ and $c = -\infty$. The remaining three cases are (ii),(iii) and (iv). □

Exercise 4.7. Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be i.i.d. In each of the cases below, show that with probability one, $-\infty = \liminf_{n \to \infty} S_n$ and $\limsup_{n \to \infty} S_n = \infty$.

- The distribution μ_{X_1} is symmetric about 0 (i.e. $\mu_{-X_1} = \mu_{X_1}$) and $\mathbf{P}(X_1 = 0) < 1$.
- $\mathbf{E}X_1 = 0$ and $\mathbf{E}X_1^2 \in (0, \infty)$. (Hint: use the Central Limit Theorem.)

For example, when $\mathbf{P}(X_1 = 1) = \mathbf{P}(X_1 = -1) = 1/2$ and $S_0 = 0$, show that with probability one, $S_0, S_1, \ldots$ takes every integer value infinitely many times.

4.2. **Stopping Times.**

Definition 4.8 (**Stopping Time**). Let $S_0, S_1, \ldots$ be a random walk on $\mathbb{R}^d$. Let $\mathcal{F}_0 := \{\emptyset, \Omega\}$. For any $n \geq 1$, denote $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$. We say that $N \colon \Omega \to \{0, 1, \ldots\} \cup \{\infty\}$ is a **stopping time** if, for any $n \in \{0, 1, \ldots\}$, $\{N = n\} \in \mathcal{F}_n$.

For example, if the stopping time takes the value 3, then the event $\{N = 3\}$ only uses "information" about X_1, X_2, X_3. From our definition, $\{N = 0\} \in \{\emptyset, \Omega\}$.

The stopping time N depends on the random walk, but this dependence should be clear in a given example, so we de-emphasize this dependence in the notation for N.

For a real-world example of a stopping time, suppose $S_0, S_1, \ldots$ is a random walk that describes the price of a stock. Suppose the starting price S_0 of the stock is \$100 and you instruct your stock broker to sell the stock when its price reaches either \$110 or \$90. That is, define the stopping time $N = \min\{n \geq 1 \colon S_n \geq 110 \text{ or } S_n \leq 90\}$. Then N is a stopping time. (We define $\min(\emptyset) := \infty$.) So, we can intuitively think of a stopping time as a stock-trading strategy, since selling the stock at time n can only use information up to time n.

More generally, if A is a measurable subset of $\mathbb{R}^d$, the **hitting time** of A is defined as

$$N := \min\{n \geq 1 \colon S_n \in A\}.$$

And N is a stopping time since, for any $n \geq 1$,

$$\{N = n\} = \{S_1 \in A^c, \ldots, S_{n-1} \in A^c, S_n \in A\} \in \mathcal{F}_n.$$

Exercise 4.9. Let M, N be stopping times for a random walk $S_0, S_1, \ldots$. Show that $\max(M, N)$ and $\min(M, N)$ are stopping times. In particular, if $n \geq 0$ is fixed, then $\max(M, n)$ and $\min(M, n)$ are stopping times

For any $k \geq 1$, define the shift map $\theta^k \colon (\mathbb{R}^d)^{\mathbb{N}} \to (\mathbb{R}^d)^{\mathbb{N}}$ by

$$(\theta^k \omega)_n := \omega_{n+k}, \qquad \forall\, \omega = (\omega_1, \omega_2, \ldots) \in (\mathbb{R}^d)^{\mathbb{N}}.$$

Suppose T is a stopping time. Define $T_1 := T$ and for any $n \geq 2$, inductively define $T_n \colon \Omega \to \mathbb{N} \cup \{\infty\}$ so that, $\forall\, \omega \in \Omega = (\mathbb{R}^d)^{\mathbb{N}}$,

$$T_n(\omega) := \begin{cases} T_{n-1}(\omega) + T(\theta^{T_{n-1}}\omega) & , \text{if } T_{n-1}(\omega) < \infty \\ \infty & , \text{if } T_{n-1}(\omega) = \infty. \end{cases}$$

Example 4.10. Let $T = T_1 := \min\{n \geq 1 \colon S_n = 0\} = \min\{n \geq 1 \colon \omega_1 + \cdots + \omega_n = 0\}$ be the time of the first visit of the random walk to 0. Then

$$T(\theta^{T_1}\omega) = \min\{n \geq 1 \colon (\theta^{T_1}\omega)_1 + \cdots + (\theta^{T_1}\omega)_n = 0\} = \min\{n \geq 1 \colon \omega_{T_1+1} + \cdots + \omega_{T_1+n} = 0\}.$$

$$T_2(\omega) = T_1(\omega) + T(\theta^{T_1}\omega) = \min\{n > T_1 \colon \omega_1 + \cdots + \omega_n = 0\}, \qquad \forall\, \omega \in \Omega = (\mathbb{R}^d)^{\mathbb{N}}.$$

That is, T_2 is the second time that the random walk returns to 0. More generally, for any $k \geq 1$, T_k is the k^{th} time that the random walk returns to 0.

Lemma 4.11. *Let T be a stopping time. For any $n \geq 1$, $\mathbf{P}(T_n < \infty) = [\mathbf{P}(T < \infty)]^n$.*

Proof. The case $\{T = 0\} = \Omega$ is easy, so assume $\{T = 0\} = \emptyset$. We induct on n. The case $n = 1$ follows since $T = T_1$. Suppose the case $n - 1$ holds. By the definition of T_n,

$$\mathbf{P}(T_n < \infty) = \mathbf{P}(T_{n-1} < \infty, T(\theta^{T_{n-1}}) < \infty).$$

For any $m \geq 1$, we have

$$\mathbf{P}(T_{n-1} = m, T(\theta^{T_{n-1}}) < \infty) = \mathbf{P}(T_{n-1} = m, T(\theta^m) < \infty).$$

Since T_{n-1} is a stopping time, $\{T_{n-1} = m\} \in \mathcal{F}_m = \sigma(X_1, \ldots, X_m)$. Note that $T(\theta^m)$ is measurable in $\sigma(X_{m+1}, X_{m+2}, \ldots)$ since $T \geq 1$. So, these two events are independent, i.e.

$$\mathbf{P}(T_{n-1} = m, T(\theta^{T_{n-1}}) < \infty) = \mathbf{P}(T_{n-1} = m)\mathbf{P}(T(\theta^m) < \infty) = \mathbf{P}(T_{n-1} = m)\mathbf{P}(T < \infty).$$

The last equality used that $\mathbf{P}$ is invariant under the shift θ^m for any $m \geq 1$, since $X_1, X_2, \ldots$ are i.i.d. Summing over all $m \geq 1$ gives

$$\mathbf{P}(T_n < \infty) = \mathbf{P}(T_{n-1} < \infty, T(\theta^{T_{n-1}}) < \infty) = \mathbf{P}(T_{n-1} < \infty)\mathbf{P}(T < \infty).$$

Iterating this equality now concludes the proof. □

The reasoning of Lemma 4.11 implies the following.

Exercise 4.12. Let $S_0, S_1, \ldots$ be a random walk with $S_0 = 0$. Let X be the number of times the random walk takes the value 0. Let $T := \min\{n \geq 1 \colon S_n = 0\}$.

- X is a geometric random variable with success probability $\mathbf{P}(T = \infty)$.

- $\mathbf{E}X = \frac{1}{\mathbf{P}(T=\infty)}$. (Here we interpret $1/0$ as ∞.)

(Hint: $\{X = k\} = \{T_{k-1} < \infty, T_k = \infty\} = \{T_{k-1} < \infty, T_k - T_{k-1} = \infty\}$.)

Recall that a geometric random variable X with success probability $0 < p < 1$ satisfies $\mathbf{P}(X = k) = p(1-p)^{k-1}$ for any integer $k \geq 1$, so by e.g. Theorem 1.60,

$$\begin{aligned}\mathbf{E}X &= \sum_{k=1}^{\infty} kp(1-p)^{k-1} = -p\frac{d}{dt}\Big|_{t=p} \sum_{k=1}^{\infty}(1-t)^k = -p\frac{d}{dt}\Big|_{t=p} \frac{1-t}{t} \\ &= -p\frac{-t-(1-t)}{t^2}\Big|_{t=p} = \frac{p}{p^2} = \frac{1}{p}.\end{aligned}$$

4.3. **Recurrence and Transcience.**

Definition 4.13. Let $S_0, S_1, \ldots$ be a random walk on $\mathbb{R}^d$ started at $x \in \mathbb{R}^d$. Suppose X_1 is a discrete random variable. Let $T^{(x)} := \min\{n \geq 1 \colon S_n = x\}$. We say that x is **recurrent** if $\mathbf{P}(T^{(x)} < \infty) = 1$ and **transient** if $\mathbf{P}(T^{(x)} < \infty) < 1$.

Theorem 4.14. *Let $S_0, S_1, \ldots$ be a random walk in $\mathbb{R}^d$ started at $x = 0$. Let $T := \min\{n \geq 1 \colon S_n = 0\}$. Then the following are equivalent*

(i) $\mathbf{P}(T < \infty) = 1$.
(ii) $\mathbf{P}(S_n = 0 \text{ for infinitely many } n \geq 1) = 1$.
(iii) $\sum_{n=0}^{\infty} \mathbf{P}(S_n = 0) = \infty$.

If additionally $S_1, S_2, \ldots$ only takes values in $\mathbb{Z}^d$, then (i),(ii),(iii) are equivalent to:

(iv)
$$\infty = \lim_{s \to 1^-} \int_{[-\pi,\pi]^d} \frac{1}{1 - s\phi(y)} dy.$$

Here $i = \sqrt{-1}$, $\phi(y) := \mathbf{E}e^{i\langle y, X_1\rangle}$, $\forall\ y \in \mathbb{R}^d$, and for any $x = (x_1, \ldots, x_d), y = (y_1, \ldots, y_d) \in \mathbb{R}^d$, we define $\langle x, y\rangle := \sum_{j=1}^d x_j y_j$.

Proof. Let X be the number of times the random walk takes the value 0. Then

$$X = \sum_{n=0}^{\infty} 1_{S_n=0} = \sum_{n=0}^{\infty} 1_{T_n<\infty}, \qquad T_0 := 0,$$

$$\mathbf{E}X = \sum_{n=0}^{\infty} \mathbf{P}(S_n = 0) = \sum_{n=0}^{\infty} \mathbf{P}(T_n < \infty). \qquad (*)$$

If (i) occurs, then $\mathbf{P}(T_n < \infty) = 1$ for every $n \geq 1$ by Lemma 4.11, so that (ii) occurs. If (ii) occurs, then $\mathbf{P}(T_n < \infty) = 1$ for every $n \geq 1$, so the second equality of $(*)$ shows that (iii) occurs. If (iii) occurs, then (i) occurs by Exercise 4.12. So, (i),(ii),(iii) are equivalent.

By $(*)$, it remains to show that the right side of (iv) is equal to $(2\pi)^d \mathbf{E}X$. Recall that $\int_{-\pi}^{\pi} e^{im\theta} d\theta = 0$ for any nonzero $m \in \mathbb{Z}$, while $\int_{-\pi}^{\pi} e^{i0\theta} d\theta = 2\pi$. Therefore, for any $n \geq 0$,

$$1_{S_n=0} = \int_{[-\pi,\pi]^d} e^{i\langle y, S_n\rangle} \frac{dy}{(2\pi)^d}.$$

Taking expected values of both sides,

$$\mathbf{P}(S_n = 0) = \int_{[-\pi,\pi]^d} \mathbf{E}e^{i\langle y, S_n\rangle} \frac{dy}{(2\pi)^d}. \qquad (**)$$

Recalling $S_n = X_1 + \cdots + X_n$ and using that $X_1, \ldots, X_n$ are i.i.d., we have $\mathbf{E}e^{i\langle y, S_n\rangle} = \prod_{j=1}^n \mathbf{E}e^{i\langle y, X_j\rangle} = (\phi(y))^n$. So, multiplying both sides of $(**)$ by s^n and summing over $n \geq 0$,

$$\sum_{n=0}^{\infty} s^n \mathbf{P}(S_n = 0) = \int_{[-\pi,\pi]^d} \sum_{n=0}^{\infty} (s\phi(y))^n \frac{dy}{(2\pi)^d} = \int_{[-\pi,\pi]^d} \frac{1}{1 - s\phi(y)} \frac{dy}{(2\pi)^d}.$$

(Since $|\phi(y)| \leq 1 \ \forall \ y \in \mathbb{R}^d$, if $|s| < 1$, then $|s\phi(y)| < 1 \ \forall \ y \in \mathbb{R}^d$.) Letting $s \to 1^-$, the left side increases monotonically to $\mathbf{E}X$ by $(*)$, so the limit of the right side exists as well. $\square$

Definition 4.15 (Simple Random Walk). For any $1 \leq j \leq d$, let $e_j \in \mathbb{R}^d$ be the vector with a 1 in the j^{th} entry and zeros in all other entries, so that $e_1, \ldots, e_d$ is the standard basis of $\mathbb{R}^d$. Let X be a random variable so that $\mathbf{P}(X = e_j) = \mathbf{P}(X = -e_j) = 1/(2d)$ for all $1 \leq j \leq d$. Let $X_1, X_2, \ldots$ be i.i.d. copies of X. The random walk $S_n := X_1 + \cdots + X_n$, $\forall$ $n \geq 1$ with $S_0 := 0$ is called the **simple random walk** on $\mathbb{Z}^d$.

The Simple Random Walk is the most basic random walk. It may be surprising that the transcience/recurrence of this random walk depends on d. Note that each point in the integer grid $\mathbb{Z}^d$ has $2d$ locations to move to at each step of the walk. And when d is large, there are more ways for the random walk to wander away from the origin.

Theorem 4.16. *Simple Random Walk is recurrent when $d \leq 2$ and transient when $d \geq 3$.*

Proof. It suffices to check whether or not condition (iv) holds of Theorem 4.14. For any $y \in \mathbb{R}^d$, we have

$$\phi(y) = \mathbf{E}e^{i\langle y, X_1\rangle} = \frac{1}{2d} \sum_{j=1}^{d} [e^{iy_j} + e^{-iy_j}] = \frac{1}{d} \sum_{j=1}^{d} \cos(y_j) = 1 + \frac{1}{d} \sum_{j=1}^{d} [-1 + \cos(y_j)].$$

For any $z \in [-\pi, \pi]$, we have $z^2/4 \leq 1 - \cos(z) \leq z^2$ by e.g. taking derivatives and using the Fundamental Theorem of Calculus. Therefore, for any $y \in \mathbb{R}^d$,

$$-\frac{1}{d} \sum_{j=1}^{d} y_j^2 \leq \frac{1}{d} \sum_{j=1}^{d} [-1 + \cos(y_j)] \leq -\frac{1}{4d} \sum_{j=1}^{d} y_j^2.$$

So, for any $y \in \mathbb{R}^d$, and for any $0 < s < 1$,

$$1 - s + s\frac{1}{4d} \sum_{j=1}^{d} y_j^2 \leq 1 - s\phi(y) \leq 1 - s + s\frac{1}{d} \sum_{j=1}^{d} y_j^2.$$

Letting $s \to 1^-$, and noting that the integrand increases monotonically in a neighborhood of $y = 0$ while remaining bounded outside this neighborhood,

$$(d/4) \int_{[-\pi,\pi]^d} \frac{1}{\sum_{j=1}^d y_j^2} dy \leq \lim_{s \to 1^-} \int_{[-\pi,\pi]^d} \frac{1}{1 - s\phi(y)} dy \leq d \int_{[-\pi,\pi]^d} \frac{1}{\sum_{j=1}^d y_j^2} dy.$$

And $\int_{[-\pi,\pi]^d} \frac{1}{\sum_{j=1}^d y_j^2} dy = \infty$ if and only if $d \leq 2$, by e.g. changing to polar coordinates. $\square$

Exercise 4.17. Give a combinatorial proof that the simple random walk $S_0, S_1, \ldots$ on $\mathbb{Z}^d$ is recurrent for $d \leq 2$. That is, estimate $\mathbf{P}(S_n = 0) \approx n^{-d/2}$ when n is large and $d \leq 2$, and conclude $\sum_{n=0}^{\infty} \mathbf{P}(S_n = 0) = \infty$ for $d \leq 2$. (Hint: use Stirling's Formula, Proposition 8.10)

Exercise 4.18. Show that if the Simple Random Walk on $\mathbb{Z}^d$ is recurrent, then this random walk takes every value in $\mathbb{Z}^d$ infinitely many times. And if the Simple Random Walk on $\mathbb{Z}^d$ is transient, then this random walk takes any fixed value in $\mathbb{Z}^d$ only finitely many times.

Exercise 4.19. Let $0 < p < 1$. Consider the random walk on $\mathbb{Z}$ such that $\mathbf{P}(X_1 = 1) = p$ and $\mathbf{P}(X_1 = -1) = 1 - p$. Show that the corresponding random walk $S_0, S_1, \ldots$ is transient when $p \neq 1/2$.

Exercise 4.20. Let $S_0, S_1, \ldots$ and $S'_0, S'_1, \ldots$ be independent simple random walks on $\mathbb{Z}^d$. Let $N := \sum_{n,m \geq 0} 1_{S_n = S'_m}$ be the number of pairs of intersections of these two random walks. For any $y \in \mathbb{R}^d$, let $\phi(y) := \mathbf{E}e^{i\langle y, X_1 \rangle}$.

- Show $\mathbf{E}N = \lim_{s \to 1^-} \int_{[-\pi,\pi]^d} \frac{1}{|1 - s\phi(y)|^2} \frac{dy}{(2\pi)^d}$. (Hint: consider $\mathbf{E}e^{i\langle y, (S_n - S'_m) \rangle}$.)
- For what $d \geq 1$ is $\mathbf{E}N < \infty$?
- Let $C := \{S_n \colon n \geq 0\} \cap \{S'_n \colon n \geq 0\}$ be the intersection set of the two independent random walks. Let $|C|$ denote the cardinality of C. Show that if the simple random walk on $\mathbb{Z}^d$ is transient, then $\mathbf{P}(N = \infty) = 1$ if and only if $\mathbf{P}(|C| = \infty) = 1$. (Hint: $N = \sum_{x \in C} N_x N'_x$ where $N_x := \sum_{n \geq 0} 1_{S_n = x}$ is the number of visits of the first random walk to x.) In the recurrent case $d = 1, 2$, Exercise 4.18 implies that $\mathbf{P}(|C| = \infty) = 1$. For any $d \geq 1$, note that $N < \infty$ implies $|C| < \infty$. It can also be shown that $\mathbf{P}(N < \infty) \in \{0, 1\}$, $\mathbf{P}(|C| = \infty) \in \{0, 1\}$, and that $\mathbf{P}(N < \infty) = 1$ if and only if $\mathbf{E}N < \infty$ (you don't have to show these things). In summary, $\mathbf{P}(|C| = \infty) = 1$ if and only if $\mathbf{E}N = \infty$.
- Hypothesize what happens to $\mathbf{E}N$ when we instead consider the tuples of intersections of $k > 2$ independent simple random walks in $\mathbb{R}^d$. (You don't have to prove your hypothesis.)

The following proposition will be derived from a more general result, Theorem 6.51 below.

Proposition 4.21 (Wald's Equations). *Let $X_1, X_2, \ldots \colon \Omega \to \mathbb{R}$ be i.i.d. Let N be a stopping time. Let $S_0, S_1, \ldots$ be the corresponding random walk with $S_0 := 0$.*

- *If $\mathbf{E}N < \infty$, and $\mathbf{E}|X_1| < \infty$, then $\mathbf{E}S_N = \mathbf{E}X_1 \mathbf{E}N$.*
- *If $\mathbf{E}X_1 = 0, \mathbf{E}X_1^2 < \infty$ and $\mathbf{E}N < \infty$, then $\mathbf{E}S_N^2 = \mathbf{E}X_1^2 \mathbf{E}N$.*

Example 4.22. Suppose $\mathbf{P}(X_1 = 1) = \mathbf{P}(X_1 = -1) = 1/2$. Let $a, b \in \mathbb{Z}$ with $a < 0 < b$. Let $N := \min\{n \geq 1 \colon S_n \notin (a, b)\}$. We first check that $\mathbf{E}N < \infty$. If $x \in \mathbb{Z} \cap (a, b)$ and if $S_n = x$ for some $n \geq 1$, then with probability at least $2^{-(b-a)}$, the random walk exits the interval (a, b) in time $b - a$. That is, $\mathbf{P}(N > (b - a)) \leq (1 - 2^{-(b-a)})$. We claim that for any $n \geq 1$,

$$\mathbf{P}(N > n(b - a)) \leq (1 - 2^{-(b-a)})^n. \qquad (*)$$

If $\{X_1 = x_1, \ldots, X_{(n-1)(b-a)} = x_{(n-1)(b-a)}\} \subseteq \{N > (n-1)(b-a)\}$ for some $x_1, \ldots, x_{(n-1)(b-a)} \in \mathbb{Z}$, then by the above reasoning

$$\begin{aligned} &\mathbf{P}(N > n(b - a), X_1 = x_1, \ldots, X_{(n-1)(b-a)} = x_{(n-1)(b-a)}) \\ &\quad \leq (1 - 2^{-(b-a)}) \mathbf{P}(X_1 = x_1, \ldots, X_{(n-1)(b-a)} = x_{(n-1)(b-a)}). \end{aligned}$$

Summing over all $x_1, \ldots, x_{(n-1)(b-a)}$ such that $\{X_1 = x_1, \ldots, X_{(n-1)(b-a)} = x_{(n-1)(b-a)}\} \subseteq \{N > (n-1)(b-a)\}$, we get

$$\mathbf{P}(N > n(b-a), N > (n-1)(b-a)) = \mathbf{P}(N > n(b-a)) \leq (1-2^{-(b-a)}) \mathbf{P}(N > (n-1)(b-a)).$$

Iterating this inequality proves $(*)$. Then $(*)$ implies $\mathbf{E}N < \infty$ by Theorem 1.86.

The first part of Proposition 4.21 says $\mathbf{E}S_N = 0$. Note that S_N only takes two values, a and b, so $\mathbf{E}S_N$ is straightforward to compute directly. Let $c := \mathbf{P}(S_N = a)$. Then

$$0 = \mathbf{E}S_N = ca + (1-c)b.$$

Solving for c we get

$$c = \mathbf{P}(S_N = a) = \frac{b}{b-a}, \qquad \mathbf{P}(S_N = b) = \frac{-a}{b-a}.$$

The second part of Proposition 4.21 says $\mathbf{E}S_N^2 = \mathbf{E}N$. Once again, S_N^2 only takes two values, so

$$\mathbf{E}N = \mathbf{E}S_N^2 = ca^2 + (1-c)b^2 = \frac{a^2b - ab^2}{b-a} = ab\frac{a-b}{b-a} = -ab.$$

Exercise 4.23. Let $1/2 < p < 1$. Consider the random walk on $\mathbb{Z}$ such that $\mathbf{P}(X_1 = 1) = p$ and $\mathbf{P}(X_1 = -1) = 1-p$. Let $S_0, S_1, \ldots$ be the corresponding random walk with $S_0 := 0$. Let $N := \min\{n \geq 1 \colon S_n > 0\}$. Using Wald's equation for $\min(N, n)$ and then letting $n \to \infty$, show that $\mathbf{E}N = 1/\mathbf{E}X_1 = 1/(2p-1)$.

4.4. **Additional Comments.** The term "random walk" was first proposed by Karl Pearson in 1905 in a letter to *Nature*. In this letter, Pearson proposed model of mosquito infestation of a forest. At each time step, a single mosquito moves a fixed length at a randomly chosen angle. Pearson asked for the distribution of the mosquitoes in the forest after a long time has passed. Rayleigh answered the letter, since he had solved a similar problem in 1880 for the modeling of sound waves in a heterogeneous material. A sound wave traveling through a material can be modeled as summing a sequence of vectors of constant amplitude but random phase, i.e. a sum of the form $\sum_{j=1}^n e^{iY_j}$, where $Y_1, Y_2, \ldots$ are real-valued and independent.

In 1900, Bachelier proposed random walks as a model for stock prices, and he also related random walks to the continuous diffusion of heat. Apparently unaware of other related works, around 1905 Einstein published his work on Brownian motion, i.e. the path of a dust particle in the air pushed in random directions by collision with gas molecules. Einstein modeled this behavior with a random walk. Smoluchowski published results similar to Einstein in 1906.

Random Walks are some of the most basic stochastic processes. They are used to model random phenomena in many scientific fields. The Simple Random Walk is essentially a discrete version of Brownian Motion.

Our presentation above focused on random walks where X_1 is discrete. In the case that X_1 is not discrete, if $S_0, S_1, \ldots$ is a random walk with $S_0 := 0$, then $x \in \mathbb{R}^d$ is called a **recurrent value** for the random walk if, for any $\varepsilon > 0$, $\mathbf{P}(\|S_n - x\| < \varepsilon \text{ for infinitely many } n \geq 1) = 1$. Here $\|(x_1, \ldots, x_d)\| := (\sum_{j=1}^d x_j^2)^{1/2}$. And $x \in \mathbb{R}^d$ is called a **possible value** for the random walk if, for any $\varepsilon > 0$, $\exists\, n \geq 0$ such that $\mathbf{P}(\|S_n - x\| < \varepsilon) > 0$. The random walk is said to be **transient** if it has no recurrent values. Otherwise, the random walk is called **recurrent**. If the random walk is recurrent, it can be shown that the set of recurrent values is equal to the set of possible values, as in Exercise 4.18.

Theorem 4.14 can then be generalized as follows.

Theorem 4.24. *Let $S_0, S_1, \ldots$ be a random walk on $\mathbb{Z}^d$ with $S_0 := 0$. For any $y \in \mathbb{R}^d$, let $\phi(y) := \mathbf{E}e^{i\langle y, X_1\rangle}$, where $i = \sqrt{-1}$.*

(a) *The convergence (or divergence) of* $\sum_{n\geq 0} P(\|S_n\| < \varepsilon)$ *for a single* $\varepsilon > 0$ *is sufficient to prove transience (or recurrence) of the random walk.*
(b) *Let* $\delta > 0$*. Then the random walk is recurrent if and only if*

$$\sup_{0<s<1} \int_{(-\delta,\delta)^d} \mathrm{Re} \frac{1}{1 - s\phi(y)} dy = \infty.$$

5. Conditional Probability and Conditional Expectation

In elementary probability theory, conditional probability and conditional expectation allow a rigorous notion for incorporating previously unknown information into a probability law. If A, B are events and if $\mathbf{P}(B) > 0$, we define the **conditional probability of** A **given** B, denoted $\mathbf{P}(A|B)$, to be

$$\mathbf{P}(A|B) := \mathbf{P}(A \cap B)/\mathbf{P}(B).$$

For example, if $\mathbf{P}$ is uniform on the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$, and if $B = \{2, 4, 6\}$, then $\mathbf{P}(\{1\}|B) = 0$ and $\mathbf{P}(\{2\}|B) = 1/3$.

Let $X \colon \Omega \to [-\infty, \infty]$ be a random variable with $\mathbf{E}\,|X| < \infty$. Note that, if B is fixed, then the function $A \mapsto \mathbf{P}(A|B)$ is itself a probability law on Ω, so we can e.g. define the **conditional expectation** of a random variable X given B, denoted $\mathbf{E}(X|B)$, to be the usual expectation of X with respect to the probability law $\mathbf{P}(\cdot|B)$.

$$\mathbf{E}(X|B) := \mathbf{E}(X1_B)/\mathbf{P}(B).$$

In case $X \geq 0$, we have the equivalent definition $\mathbf{E}(X|B) = \int_0^\infty \mathbf{P}(X > t|B)dt$.

If Z is a discrete random variable, i.e. if Z takes at most countably many values, and if $\mathbf{P}(Z = z) > 0$ for some $z \in \mathbb{R}$, we let $B := \{Z = z\}$ in the above definition to define $\mathbf{E}(X|Z = z)$. By splitting the sample space Ω into countably many disjoint sets $B_1, B_2, \ldots$ such that $\cup_{n=1}^\infty B_n = \Omega$ and $\mathbf{P}(B_n) > 0$ for all $n \geq 1$, we can write

$$\mathbf{P}(A) = \sum_{n=1}^\infty \mathbf{P}(A \cap B_n) = \sum_{n=1}^\infty \mathbf{P}(A|B_n)\mathbf{P}(B_n).$$

$$\mathbf{E}X = \sum_{n=1}^\infty \mathbf{E}(X1_{B_n}) = \sum_{n=1}^\infty \mathbf{E}(X|B_n)\mathbf{P}(B_n). \qquad (*)$$

By breaking up expected values or probabilities into pieces in this way, sometimes the quantities on the right side are easier to compute, allowing computation of the left side.

Exercise 5.1. Prove Wald's first equation. Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be i.i.d. Let N be a stopping time with $\mathbf{E}N < \infty$. Let $S_0 := 0$ and for any $n \geq 1$, let $S_n := X_1 + \cdots + X_n$. Then $\mathbf{E}S_N = \mathbf{E}X_1\mathbf{E}N$. (Hint: condition on N taking fixed values.)

We now restate the identity on the right of $(*)$.

Definition 5.2. Let $X \colon \Omega \to \mathbb{R}$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Let $\mathcal{G} \subseteq \mathcal{F}$ where $\mathcal{G}$ is the σ-algebra of Ω generated by the disjoint sets $B_1, B_2, \ldots$ such that $\cup_{n=1}^\infty B_n = \Omega$ and $\mathbf{P}(B_n) > 0$ for all $n \geq 1$. Define the **conditional expectation of** X **given** $\mathcal{G}$, denoted $\mathbf{E}(X|\mathcal{G})$, to be the random variable on Ω that takes the value $\mathbf{E}(X|B_n)$ on the set B_n for all $n \geq 1$.

Then $\mathbf{E}(X|\mathcal{G})$ takes the value $\mathbf{E}(X|B_n)$ with probability $\mathbf{P}(B_n)$ for all $n \geq 1$, so we can rewrite $(*)$ as

$$\mathbf{E}X = \mathbf{E}(\mathbf{E}(X|\mathcal{G})).$$

Note that $\mathcal{G}$ consists of all disjoint unions of sets $B_1, B_2, \ldots$, and $\mathbf{E}(X|\mathcal{G})$ is constant on each of these sets. Moreover, for any $n \geq 1$,

$$\mathbf{E}(\mathbf{E}(X|\mathcal{G})1_{B_n}) = \mathbf{E}(X|B_n)\mathbf{P}(B_n) = \mathbf{E}(X1_{B_n}).$$

By linearity of $\mathbf{E}$, we conclude that, for any $B \in \mathcal{G}$, we have

$$\mathbf{E}(\mathbf{E}(X|\mathcal{G})1_B) = \mathbf{E}(X1_B).$$

We can in fact turn this identity into a *definition* of $\mathbf{E}(X|\mathcal{G})$, for any σ-algebra $\mathcal{G}$. Compared to the definition above, the definition below does not require the nonempty sets in $\mathcal{G}$ to have positive measure.

Definition 5.3 (**Conditional Expectation**)**.** Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, let $X\colon \Omega \to \mathbb{R}$ be an $\mathcal{F}$-measurable random variable with $\mathbf{E}\,|X| < \infty$. Let $\mathcal{G}$ be a σ-algebra with $\mathcal{G} \subseteq \mathcal{F}$ (so that $\mathcal{G}$ is coarser than $\mathcal{F}$). We define a **conditional expectation of** X **with respect to** $\mathcal{G}$ to be any random variable $Y\colon \Omega \to \mathbb{R}$ such that

- Y is measurable with respect to $\mathcal{G}$. (For any measurable $A \subseteq \mathbb{R}$, $Y^{-1}(A) \in \mathcal{G}$.)
- For any $B \in \mathcal{G}$, $\mathbf{E}(Y1_B) = \mathbf{E}(X1_B)$.

In Proposition 5.5 below, we will show that a Y satisfying the above properties exists and is unique up to measure zero changes to Y, and $\mathbf{E}\,|Y| < \infty$. We therefore denote the conditional expectation of X with respect to $\mathcal{G}$ as $\mathbf{E}(X|\mathcal{G})$.

Definition 5.4. Let μ, ν be measures on a measurable space $(\Omega, \mathcal{F})$. We say that ν is **absolutely continuous** with respect to μ, denoted $\nu \ll \mu$, if whenever $\mu(A) = 0$ for some $A \in \mathcal{F}$, we have $\nu(A) = 0$.

Proposition 5.5. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, let $X\colon \Omega \to \mathbb{R}$ be a random variable with $\mathbf{E}\,|X| < \infty$. Let $\mathcal{G}$ be a σ-algebra with $\mathcal{G} \subseteq \mathcal{F}$. There exists a random variable Y that is the conditional expectation of X given $\mathcal{G}$. Moreover, if Y' is another conditional expectation of X given $\mathcal{G}$, then $\mathbf{P}(Y = Y') = 1$.*

Proof. We first show uniqueness. Let $\varepsilon > 0$. Let $B_\varepsilon := \{Y - Y' > \varepsilon\} \in \mathcal{G}$. Then

$$0 = \mathbf{E}(Y1_{B_\varepsilon}) - \mathbf{E}(X1_{B_\varepsilon}) = \mathbf{E}((Y - Y')1_{B_\varepsilon}) \geq \varepsilon\mathbf{P}(B_\varepsilon).$$

So, $\mathbf{P}(B_\varepsilon) = 0$. Letting $\varepsilon \to 0^+$, we get $\mathbf{P}(Y - Y' > 0) = 0$ by continuity of the probability law. Interchanging the roles of Y, Y' show that $\mathbf{P}(Y' - Y > 0) = 0$ as well.

We now show existence. Assume for now that $X \geq 0$. Let μ denote the restriction of $\mathbf{P}$ to the measurable space $(\Omega, \mathcal{G})$ and also define a measure ν on this space by $\nu(A) := \int_A X d\mathbf{P}$ for any $A \in \mathcal{G}$. Since $\int_\Omega |X|\, d\mathbf{P} < \infty$, $\nu(\Omega) < \infty$. Also, if $\mu(A) = 0$ then $\mathbf{P}(A) = 0$ so $\nu(A) = 0$, so ν is absolutely continuous with respect to μ. By the Radon-Nikodym Theorem, Theorem 8.2, $\exists$ a nonnegative $\mathcal{G}$-measurable random variable Y such that $\nu = Y\mu$. So, for any $A \in \mathcal{G}$,

$$\mathbf{E}X1_A = \int_A X d\mathbf{P} = \nu(A) = \int_A d\nu = \int_A Y d\mu = \int_A Y d\mathbf{P} = \mathbf{E}Y1_A.$$

So, Y is a conditional expectation of X with respect to $\mathcal{G}$ (and $\mathbf{E}Y = \mathbf{E}\,|Y| = \nu(\Omega) < \infty$.)

We now consider the case of general $X\colon \Omega \to \mathbb{R}$. Write $X = \max(X,0) - \max(-X,0) =: X_+ - X_-$. The preceding argument gives Y_+, Y_- that are conditional expectations of X_+, X_- respectively, so $Y := Y_+ - Y_-$ satisfies $\mathbf{E}\,|Y| < \infty$, and for any $A \in \mathcal{G}$,

$$\mathbf{E}X1_A = \mathbf{E}X_+1_A - \mathbf{E}X_-1_A = \mathbf{E}Y_+1_A - \mathbf{E}Y_-1_A = \mathbf{E}Y1_A.$$

□

Exercise 5.6. Let $\Omega = [0,1]$. Let $\mathbf{P}$ be the uniform probability law on Ω. Let $X\colon [0,1] \to \mathbb{R}$ be a random variable such that $X(t) = t^2$ for all $t \in [0,1]$. Let

$$\mathcal{G} = \sigma\{[0,1/4), [1/4,1/2), [1/2,3/4), [3/4,1]\}.$$

Compute explicitly the function $\mathbf{E}(X|\mathcal{G})$. (It should be constant on each of the partition elements.) Draw the function $\mathbf{E}(X|\mathcal{G})$ and compare it to a drawing of X itself.

Now, for every integer $k > 1$, let $s = 2^{-k}$, and let $\mathcal{G}_k := \{[0,s), [s,2s), [2s,3s), \ldots, [1-2s, 1-s), [1-s,1)\}$. Try to draw $\mathbf{E}(X|\mathcal{G}_k)$. Prove that, for every $t \in [0,1]$,

$$\lim_{k\to\infty} \mathbf{E}(X|\mathcal{G}_k)(t) = X(t).$$

Exercise 5.7. Let $X\colon \Omega \to \mathbb{R}$ be a random variable with finite variance, and let $t \in \mathbb{R}$. Consider the function $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(t) = \mathbf{E}(X-t)^2$. Show that the function f is uniquely minimized when $t = \mathbf{E}X$. That is, $f(\mathbf{E}X) < f(t)$ for all $t \in \mathbb{R}$ such that $t \neq \mathbf{E}X$. Put another way, setting t to be the mean of X minimizes the quantity $\mathbf{E}(X-t)^2$ uniquely.

The conditional expectation, being a piecewise version of taking an average, has a similar property. Let $B_1, \ldots, B_k \subseteq \Omega$ such that $B_i \cap B_j = \emptyset$ for all $i, j \in \{1, \ldots, k\}$ with $i \neq j$, and $\cup_{i=1}^k B_i = \Omega$. Write $\mathcal{G} = \sigma\{B_1, \ldots, B_k\}$. Let Y be any other random variable such that, for each $1 \leq i \leq k$, Y is constant on B_i. Show that the quantity $\mathbf{E}(X-Y)^2$ is uniquely minimized by such a Y only when $Y = \mathbf{E}(X|\mathcal{G})$.

Exercise 5.8. Let $\Omega = [0,1]$. Let $\mathbf{P}$ be the uniform probability law on Ω. Let $X\colon [0,1] \to \mathbb{R}$ be a random variable such that $X(t) = t^2$ for all $t \in [0,1]$. For every integer $k > 1$, let $s = 2^{-k}$, let $\mathcal{G}_k := \sigma\{[0,s), [s,2s), [2s,3s), \ldots, [1-2s, 1-s), [1-s,1)\}$, and let $M_k := \mathbf{E}(X|\mathcal{G}_k)$. Show that the increments $M_2 - M_1, M_3 - M_2, \ldots$ are orthogonal in the following sense. For any $i, j \geq 1$ with $i \neq j$,

$$\mathbf{E}(M_{i+1} - M_i)(M_{j+1} - M_j) = 0.$$

This property is sometimes called **orthogonality of martingale increments**.

Remark 5.9. In Definition 5.2, if $Z\colon \Omega \to \mathbb{R}$ is a discrete random variable, i.e. if Z only takes countably many values $z_1, z_2, \ldots \in \mathbb{R}$, and if $\mathcal{G} = \sigma(Z)$ is the σ-algebra generated by $\{\omega \in \Omega\colon Z(\omega) = z_1\}, \{\omega \in \Omega\colon Z(\omega) = z_2\}, \ldots$, then the random variable $\mathbf{E}(X|\mathcal{G})$ is denoted as $\mathbf{E}(X|Z)$. If we use $B_i := \{Z = z_i\}$ for all $i \geq 1$ in equation $(*)$, we get

$$\mathbf{E}X = \sum_{i=1}^{\infty} \mathbf{E}(X|Z = z_i)\mathbf{P}(Z = z_i).$$

We can intuitively think of $\mathcal{G}$ as some amount of information that can change our knowledge of a random variable X. As in the example below, if $\mathcal{G}$ is the coarsest possible σ-algebra, then we know essentially nothing about X, and $\mathbf{E}(X|\mathcal{G})$ is constant almost surely. At the other extreme, if X is $\mathcal{G}$-measurable, then $\mathbf{E}(X|\mathcal{G}) = X$, i.e. we know everything about X.

Example 5.10. In Definition 5.3, suppose X is measurable with respect to $\mathcal{G}$. We can then use $Y := X$ in Definition 5.3. By the uniqueness part of Proposition 5.5, we conclude that $\mathbf{E}(X|\mathcal{G}) = X$.

In Definition 5.3, suppose $\mathcal{G} = \{\emptyset, \Omega\}$ is the coarsest possible σ-algebra on Ω. Then $\mathbf{E}(X|\mathcal{G})$ must be constant almost everywhere, since constant functions are the only $\mathcal{G}$-measurable functions. Choosing $B = \Omega$ in Definition 5.3, we conclude that $\mathbf{E}(X|\mathcal{G}) = \mathbf{E}X$.

Exercise 5.11. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X\colon \Omega \to \mathbb{R}$ be a random variable with $\mathbf{E}\,|X| < \infty$. Let $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ be σ-algebras. Let $\mathcal{H}$ be a σ-algebra that is independent of $\sigma(\sigma(X), \mathcal{G})$. Show that

$$\mathbf{E}(X|\sigma(\mathcal{G}, \mathcal{H})) = \mathbf{E}(X|\mathcal{G}).$$

In particular, if we choose $\mathcal{G} = \{\emptyset, \Omega\}$, we get: if $\mathcal{H}$ is independent of $\sigma(X)$, then $\mathbf{E}(X|\mathcal{H}) = \mathbf{E}X$.

(Hint: Let $G \in \mathcal{G}, H \in \mathcal{H}$, let $Y := \mathbf{E}(X|\mathcal{G})$. Compare $\mathbf{E}(X1_{G\cap H})$ and $\mathbf{E}(Y1_{G\cap H})$. Is the set of $A \in \sigma(\mathcal{G}, \mathcal{H})$ such that $\mathbf{E}(X1_A) = \mathbf{E}(Y1_A)$ a monotone class?)

Proposition 5.12. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X\colon \Omega \to \mathbb{R}$ be a random variable with $\mathbf{E}\,|X| < \infty$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. Then*

- $\mathbf{E}X = \mathbf{E}(\mathbf{E}(X|\mathcal{G}))$
- *If $X \geq 0$ then $\mathbf{E}(X|\mathcal{G}) \geq 0$ almost surely. And if $X > 0$ then $\mathbf{E}(X|\mathcal{G}) > 0$ almost surely. And if*

Proof. The first item follows by choosing $B := \Omega$ in Definition 5.3. Let $Y := \mathbf{E}(X|\mathcal{G})$. For the second item, choose $B := \{\omega \in \Omega\colon Y \leq 0\} \in \mathcal{G}$ in Definition 5.3 to get $0 \leq \mathbf{E}(X1_B) = \mathbf{E}(Y1_B) \leq 0$. Therefore $\mathbf{E}(Y1_{Y\leq 0}) = 0$ so that $Y \geq 0$ almost surely. Also, for any $\varepsilon > 0$, $\varepsilon\mathbf{P}(X > \varepsilon, Y \leq 0) \leq \mathbf{E}(X1_{X>\varepsilon}1_{Y\leq 0}) \leq \mathbf{E}(X1_{Y\leq 0}) = 0$, so $\mathbf{P}(X > 0, Y = 0) = 0$, by continuity of the probability law (as $\varepsilon \to 0^+$). □

Proposition 5.13 (Linearity of Conditional Expectation). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X, Y\colon \Omega \to \mathbb{R}$ be random variables with $\mathbf{E}\,|X|, \mathbf{E}\,|Y| < \infty$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. Then for any $\alpha \in \mathbb{R}$,*

$$\mathbf{E}(\alpha X + Y|\mathcal{G}) = \alpha\mathbf{E}(X|\mathcal{G}) + \mathbf{E}(Y|\mathcal{G}).$$

Proof. Let $V := \mathbf{E}(X|\mathcal{G})$, $W := \mathbf{E}(Y|\mathcal{G})$. Note that $\mathbf{E}\,|\alpha V + W| < \infty$, and for any $B \in \mathcal{G}$,

$$\mathbf{E}(\alpha X + Y)1_B = \alpha\mathbf{E}(X1_B) + \mathbf{E}Y1_B = \alpha\mathbf{E}V1_B + \mathbf{E}W1_B = \mathbf{E}(\alpha V + W)1_B.$$

□

Proposition 5.13 and the second part of Proposition 5.12 imply the following.

Corollary 5.14 (Monotonicity of Conditional Expectation). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X, Y\colon \Omega \to \mathbb{R}$ be random variables with $\mathbf{E}\,|X|, \mathbf{E}\,|Y| < \infty$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. If $X \leq Y$, then*

$$\mathbf{E}(X|\mathcal{G}) \leq \mathbf{E}(Y|\mathcal{G}).$$

Exercise 5.15. Prove Jensen's inequality for the conditional expectation. Let $X\colon \Omega \to \mathbb{R}$ be a random variable and let $\phi\colon \mathbb{R} \to \mathbb{R}$ be convex. Assume $\mathbf{E}\,|X|, \mathbf{E}\,|\phi(X)| < \infty$. Then

$$\phi(\mathbf{E}(X|\mathcal{G})) \leq \mathbf{E}(\phi(X)|\mathcal{G})$$

Conclude that for any $1 \leq p \leq \infty$ we have the following contractive inequality for conditional expectation

$$\|\mathbf{E}(X|\mathcal{G})\|_p \leq \|X\|_p .$$

Proposition 5.16. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X, Y \colon \Omega \to \mathbb{R}$ be random variables with $\mathbf{E}|X|, \mathbf{E}|XY| < \infty$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. Suppose Y is $\mathcal{G}$-measurable. Then*

$$\mathbf{E}(XY|\mathcal{G}) = Y\mathbf{E}(X|\mathcal{G}).$$

Proof. Let $Z := \mathbf{E}(X|\mathcal{G})$. Note that YZ is $\mathcal{G}$-measurable, so we must check

$$\mathbf{E}(XY1_B) = \mathbf{E}(ZY1_B), \qquad (*)$$

for any $B \in \mathcal{G}$. If $Y = 1_A$ for some $A \in \mathcal{G}$, then $A \cap B \in \mathcal{G}$, so by definition of Z,

$$\mathbf{E}(XY1_B) = \mathbf{E}(X1_{A\cap B}) = \mathbf{E}(Z1_{A\cap B}) = \mathbf{E}(ZY1_B).$$

By linearity, $(*)$ holds when Y is a simple function. If $X \geq 0$ then $Z \geq 0$ by Proposition 5.12, so $(*)$ holds for any nonnegative Y by the Monotone Convergence Theorem 1.54. More generally, write $X = \max(X, 0) - \max(-X, 0))$, $Y = \max(Y, 0) - \max(-Y, 0))$, write $|XY| = (\max(X, 0) + \max(-X, 0))(\max(Y, 0) + \max(-Y, 0))$, and note that all four products have finite expected value since $\mathbf{E}|XY| < \infty$. Also by Corollary 5.14, $Z = \mathbf{E}(\max(X, 0)|\mathcal{G}) - \mathbf{E}(\max(-X, 0)|\mathcal{G})$. So, the previous result applied to each of the four products $\max(\pm X, 0) \max(\pm Y, 0)$ concludes the proof. □

Exercise 5.17 (Tower Property)**.** Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X \colon \Omega \to \mathbb{R}$ be a random variable with $\mathbf{E}|X| < \infty$. Let $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$ be σ-algebras. Then $\mathbf{E}(X|\mathcal{H}) = \mathbf{E}(\mathbf{E}(X|\mathcal{G})|\mathcal{H})$.

Exercise 5.18 (Conditional Markov Inequality)**.** Let $p > 0$. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X \colon \Omega \to \mathbb{R}$ be a random variable with $\mathbf{E}|X|^p < \infty$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. For any $A \in \mathcal{F}$, we denote $\mathbf{P}(A|\mathcal{G}) := \mathbf{E}(1_A|\mathcal{G})$.

- Show that, almost surely,

$$\mathbf{E}(|X|^p|\mathcal{G}) = \int_0^\infty pt^{p-1}\mathbf{P}(|X| > t|\mathcal{G})dt.$$

- Deduce a conditional version of Markov's inequality: for any $t > 0$, almost surely,

$$\mathbf{P}(|X| > t|\mathcal{G}) \leq \frac{\mathbf{E}(|X|^p|\mathcal{G})}{t^p}.$$

Exercise 5.19 (Conditional Hölder Inequality)**.** Let $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X, Y \colon \Omega \to \mathbb{R}$ be random variables with $\mathbf{E}|X|^p, \mathbf{E}|Y|^q < \infty$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. Show that, almost surely,

$$\mathbf{E}(|XY||\mathcal{G}) \leq [\mathbf{E}(|X|^p|\mathcal{G})]^{1/p}[\mathbf{E}(|Y|^q|\mathcal{G})]^{1/q}.$$

5.1. **Conditional Expectation as Hilbert Space Projection.** Unlike other sections, in this section capital letters will often not denote random variables.

Definition 5.20. A **real Hilbert space** H is a vector space over $\mathbb{R}$ equipped with a bilinear, symmetric function $\langle \cdot, \cdot \rangle \colon H \times H \to \mathbb{R}$ such that $\langle h, h \rangle \geq 0$ for all $h \in H$ with equality only when $h = 0$, and such that H is complete with respect to the metric $d \colon H \times H \to [0, \infty)$ defined by $d(g, h) := \langle g - h, g - h \rangle^{1/2} =: \|g - h\|$, $\forall\ g, h \in H$. By complete, we mean: for any sequence $h_1, h_2, \ldots \in H$ that is Cauchy ($\forall\ \varepsilon > 0$, $\exists\ n > 0$ such that $\forall\ m \geq n$, $\|h_m - h_n\| < \varepsilon$), there exists $h \in H$ such that $\lim_{n\to\infty} \|h_n - h\| = 0$. The function $\langle \cdot, \cdot \rangle$ is called an **inner product** on H, and the function $\|\cdot\|$ is called the **norm** on H associated to the inner product $\langle \cdot, \cdot \rangle$.

Exercise 5.21. Let H be a Hilbert space. Let $g, h \in H$. Prove the Cauchy-Schwarz inequality

$$|\langle g, h \rangle| \leq \|g\| \, \|h\| .$$

Show also the triangle inequality $\|g + h\| \leq \|g\| + \|h\|$, and the parallelogram law $\|g + h\|^2 + \|g - h\|^2 = 2\|g\|^2 + 2\|h\|^2$.

If $X \colon \Omega \to \mathbb{R}$ is a random variable on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and if $\mathcal{G} \subseteq \mathcal{F}$ is a σ-algebra, then we can interpret $\mathbf{E}(X|\mathcal{G})$ as a Hilbert space projection. In the special case that $\mathcal{G}$ is a σ-algebra generated by a countable set of disjoint sets, it follows immediately from Definition 5.2 that the map $X \mapsto \mathbf{E}(X|\mathcal{G})$ is a projection, i.e. $\mathbf{E}(\mathbf{E}(X|\mathcal{G})|\mathcal{G}) = \mathbf{E}(X|\mathcal{G})$. In the case of more general $\mathcal{G}$, we will make a similar statement below.

Theorem 5.22 (**Hilbert space projections**). *Let H be a Hilbert space, $W \subseteq H$ a closed convex set, $M \subseteq H$ a closed subspace. Define $M^\perp := \{h \in H \colon \langle h, m \rangle = 0, \forall\, m \in M\}$.*

(a) *$\exists\ v \in W$ such that $\|h - v\| = \inf_{w \in W} \|h - w\|$.*
(b) *Every $h \in H$ can be uniquely written as $h = v + p$, for some $v \in M$, $p \in M^\perp$. (We therefore write $H = M \oplus M^\perp$.)*
(c) *$(M^\perp)^\perp = M$.*

The map $h \mapsto v$ is called the linear projection of H onto M (choosing $W := M$ above.)

Proof of (a). Let $a := \inf_{w \in W} \|h - w\|$. Let $w_1, w_2, \ldots \in W$ such that $\lim_{n\to\infty} \|h - w_n\| = a$. The parallelogram law says

$$\|2h - (w_n + w_m)\|^2 + \|w_n - w_m\|^2 = 2(\|h - w_m\|^2 + \|h - w_n\|^2) \to 4a^2 \quad (*)$$

as $m, n \to \infty$. But $\frac{1}{2}(w_n + w_m) \in W$, so $4\left\|h - \frac{1}{2}(w_n + w_m)\right\|^2 \geq 4a^2$, by definition of a. Then from the left side of $(*)$, $\|w_n - w_m\|^2 \to 0$ as $m, n \to \infty$, so $w_1, w_2, \ldots$ is a Cauchy sequence, i.e. $v := \lim_{n\to\infty} w_n$ exists in H. Note that $\|h - v\| \geq a$ by definition of a. Also, by the triangle inequality, $\|h - v\| \leq \|h - w_n\| + \|w_n - v\|$ for all $n \geq 1$. Letting $n \to \infty$ shows $\|h - v\| \leq a$. Therefore, $\|h - v\| = a$.

Proof of (b). First observe that $M^\perp$ is closed and $M \cap M^\perp = \{0\}$ by definition of $M^\perp$. Uniqueness follows since $M \cap M^\perp = \{0\}$, so if $h = v + p = v' + p'$ with $v, v' \in M$ and $p, p' \in M^\perp$, then $v - v' = p' - p \in M \cap M^\perp = \{0\}$. To get existence, use part (a) to find $v \in W := M$ with $\|h - v\| = \inf_{m \in M} \|h - m\|$. Let $m \in M$ with $\|m\| = 1$. Then $v + \langle h - v, m \rangle m \in M$, and by definition of v,

$$\|h - v\|^2 \leq \|h - v - \langle h - v, m \rangle m\|^2 = \|h - v\|^2 - |\langle h - v, m \rangle|^2$$

Thus $\langle h-v, m\rangle = 0$ for all $m \in M$, i.e. $p := h - v \in M^\perp$, so $h = v + (h-v) = v + p$.

Proof of (c). By definition, $M \subseteq M^{\perp\perp}$. For any $h \in M^{\perp\perp}$, apply part (b) to get $h = m + m^\perp$, so $0 = m^\perp + (m-h)$, $m^\perp \in M^\perp, m - h \in M^{\perp\perp}$. Apply part (b) again to $M^\perp$, so $H = M^\perp \oplus M^{\perp\perp}$. By uniqueness of this decomposition for $0 \in H$, we conclude $m^\perp = 0$, $m - h = 0$, so $h = m \in M$, i.e. $M^{\perp\perp} = M$. □

Exercise 5.23. Let H be a Hilbert space, let $M \subseteq H$ a closed subspace, and for any $h \in H$, denote $f(h)$ as the linear projection of H onto M. Show that $h \mapsto f(h)$ is actually a linear projection. That is, verify that $f(\alpha g + h) = \alpha f(g) + f(h)$ and $f(f(h)) = f(h)$ for any $\alpha \in \mathbb{R}$, $g, h \in H$.

Definition 5.24. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. We denote

$$L_2(\Omega, \mathcal{F}, \mathbf{P}) := \{X : \Omega \to \mathbb{R} \quad : X \text{ is } \mathcal{F}-\text{measurable and } \mathbf{E}X^2 < \infty\}.$$

It is well known that $L_2(\Omega, \mathcal{F}, \mathbf{P})$ equipped with the inner product $\langle X, Y\rangle := \mathbf{E}XY$, $\forall$, $X, Y \in L_2(\Omega, \mathcal{F}, \mathbf{P})$, is a Hilbert space. Note that $\|X\| = \langle X, X\rangle^{1/2} = (\mathbf{E}X^2)^{1/2}$. (Strictly speaking, any two random variables $X, Y : \Omega \to \mathbb{R}$ such that $\mathbf{P}(X = Y) = 1$ are identified as the same element of $L_2(\Omega, \mathcal{F}, \mathbf{P})$ if $\mathbf{E}X^2 < \infty$. That is, $L_2(\Omega, \mathcal{F}, \mathbf{P})$ consists of equivalence classes of random variables that are equal almost surely.)

Theorem 5.25 (Completeness of L_2). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Any Cauchy sequence $X_1, X_2, \ldots$ in $L_2(\Omega, \mathcal{F}, \mathbf{P})$ has a subsequence such that $\|X_{k_n} - X_{k_m}\| \le c_{\min\{m,n\}}$ with $\sum_{n=1}^\infty c_n < \infty$. A subsequence with this property is then Cauchy pointwise almost surely. If X denotes the a.e. limit of the subsequence, then the original sequence converges to X in $L_2(\Omega, \mathcal{F}, \mathbf{P})$.*

Proof. Given any Cauchy sequence, we may take a rapidly convergence subsequence (so that, e.g. $\|X_{k_i} - X_{k_j}\| \le 2^{-\max(i,j)}$). For any $n \ge 1$, let $Y_n := |X_1| + \sum_{k=2}^n |X_k - X_{k-1}|$, and define $Y := \lim_{n\to\infty} Y_n$. From our rapid convergence, we have $\|Y_n\| \le c < \infty$ for all $n \ge 1$, so the Monotone Convergence Theorem, Theorem 1.54) gives $\|Y\| \le c$. Thus, Y is finite almost surely, and $\sum_{k=2}^\infty |X_k(\omega) - X_{k-1}(\omega)|$ converges in $\mathbb{R}$ for all $\omega \in \Omega$ (after redefining X_k's on a $\mathbf{P}$ measure zero set). So the telescoping sum $\sum_{k=2}^\infty (X_k(\omega) - X_{k-1}(\omega))$ is absolutely convergent, therefore convergent, therefore $X := \lim_{n\to\infty} X_n$ exists for every $\omega \in \Omega$, and the Dominated Convergence Theorem, Theorem 1.57 (using $|X - X_n| \le Y$ for all $n \ge 1$) shows $\lim_{n\to\infty} \|X - X_n\| = 0$. □

Proposition 5.26 (Conditional Expectation as Projection). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, let $X : \Omega \to \mathbb{R}$ be an $\mathcal{F}$-measurable random variable with $\mathbf{E}X^2 < \infty$. Let $\mathcal{G}$ be a σ-algebra with $\mathcal{G} \subseteq \mathcal{F}$. Define $Z \in L_2(\Omega, \mathcal{G}, \mathbf{P})$ by*

$$\mathbf{E}\,|X - Z|^2 := \inf_{Y \in L_2(\Omega, \mathcal{G}, \mathbf{P})} \mathbf{E}(X - Y)^2. \qquad (*)$$

Then Z satisfies $(*)$ if and only if $\mathbf{E}((X - Z)W) = 0$ for all $W \in L_2(\Omega, \mathcal{G}, \mathbf{P})$.

So, the map $X \mapsto \mathbf{E}(X|\mathcal{G})$ is a linear projection from $L_2(\Omega, \mathcal{F}, \mathbf{P})$ to $L_2(\Omega, \mathcal{G}, \mathbf{P})$. In particular, $\mathbf{E}(X|\mathcal{G})$ exists and is unique.

Proof. $L_2(\Omega, \mathcal{G}, \mathbf{P})$ is itself a Hilbert space by Theorem 5.25. Since $\mathcal{G} \subseteq \mathcal{F}$, $X \in L_2(\Omega, \mathcal{G}, \mathbf{P})$ implies $X \in L_2(\Omega, \mathcal{F}, \mathbf{P})$. So, $L_2(\Omega, \mathcal{G}, \mathbf{P})$ is a closed vector subspace of $L_2(\Omega, \mathcal{F}, \mathbf{P})$. Existence and uniqueness of Z follows from Theorem 5.22(a),(b). Let $W \in L_2(\Omega, \mathcal{G}, \mathbf{P})$ and let

$t \in \mathbb{R}$. By definition of Z,

$$0 \leq \|X - (Z + tW)\|^2 - \|X - Z\|^2 = t^2 \mathbf{E}W^2 - 2t\mathbf{E}((X - Z)W).$$

Since this holds $\forall\, t \in \mathbb{R}$, we must have $\mathbf{E}((X - Z)W) = 0$. Conversely, if $\mathbf{E}((X - Z)W) = 0$ for all $W \in L_2(\Omega, \mathcal{G}, \mathbf{P})$, then letting $Y \in L_2(\Omega, \mathcal{G}, \mathbf{P})$ and using $W := Z - Y$,

$$\begin{aligned}\|X - Y\|^2 &= \|(X - Z) + (Z - Y)\|^2 \\ &= \|X - Z\|^2 + 2\mathbf{E}((X - Z)(Z - Y)) + \|Z - Y\|^2 \geq \|X - Z\|^2.\end{aligned}$$

So, Z must satisfy $(*)$. Finally, if we choose $Y := 1_A$ with $A \in \mathcal{G}$, we have $\mathbf{E}((X - Z)1_A) = 0$, so that $Z = \mathbf{E}(X|\mathcal{G})$. □

We observed the minimization property of conditional expectation discussed above in Exercise 5.7. Similarly, we can generalize the orthogonality property from Exercise 5.8.

Exercise 5.27. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, let $X \colon \Omega \to \mathbb{R}$ be an $\mathcal{F}$-measurable random variable with $\mathbf{E}X^2 < \infty$. Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}$. For any $n \geq 1$, define $M_n := \mathbf{E}(X|\mathcal{F}_n)$. Show that, for any $i, j \geq 1$ with $i \neq j$,

$$\mathbf{E}(M_{i+1} - M_i)(M_{j+1} - M_j) = 0.$$

This property is sometimes called **orthogonality of martingale increments**. (Hint: what do Hilbert space projections say about the random variables $M_1, M_2, \ldots$ and about the differences $M_{n+1} - M_n$?)

5.2. **Conditional Expectation as Regular Conditional Distribution.** There is yet another way to interpret conditional expectation. We first demonstrate a special case of this construction, and we then generalize the construction.

Suppose the random vector $(X, Y) \in \mathbb{R}^2$ has density function $f_{X,Y} \colon \mathbb{R}^2 \to [0, \infty)$, so that $\int_{\mathbb{R}^2} f_{X,Y}(x, y)dxdy = 1$ and $\mathbf{P}((X, Y) \in A) = \int_A f_{X,Y}(x, y)dxdy$ for any measurable $A \subseteq \mathbb{R}^2$. We demonstrate a way to construct the conditional expectation $\mathbf{E}(X|Y) := \mathbf{E}(X|\sigma(Y))$ directly from $f_{X,Y}$.

For any $x \in \mathbb{R}$, let $f_X(x) := \int_{\mathbb{R}} f_{X,Y}(x, y)dy$ be the marginal distribution of X, and for any $y \in \mathbb{R}$, let $f_Y(y) := \int_{\mathbb{R}} f_{X,Y}(x, y)dy$ be the marginal distribution of Y. From Fubini's Theorem, Theorem 1.66, these functions are measurable. For any $x, y \in \mathbb{R}$, define

$$f_{X|Y}(x|y) := \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} & \text{, if } f_Y(y) > 0 \\ f_X(x) & \text{, otherwise.} \end{cases}$$

Then $f_{X|Y}(x|y)$ is a measurable function of x, $\forall\, y \in \mathbb{R}$, and $\int_{\mathbb{R}} f_{X|Y}(x|y)dx = 1$, $\forall\, y \in \mathbb{R}$.

Proposition 5.28. *Suppose the random vector $(X, Y) \in \mathbb{R}^2$ has density function $f_{X,Y} \colon \mathbb{R}^2 \to [0, \infty)$. Let $g \colon \mathbb{R} \to \mathbb{R}$ be measurable with $\mathbf{E}\,|g(X)| < \infty$, and $\forall\, y \in \mathbb{R}$, define $h \colon \mathbb{R} \to \mathbb{R}$ by*

$$h(y) := \begin{cases} \int_{\mathbb{R}} g(x) f_{X|Y}(x|y)dx & \text{, if } \int_{\mathbb{R}} |g(x)|\, f_{X|Y}(x|y)dx < \infty \\ 0 & \text{, otherwise.} \end{cases}$$

Then $h(Y) = \mathbf{E}(g(X)|Y)$.

Proof. As in Fubini's Theorem 1.66, h is measurable. Also, by Fubini's Theorem 1.66, since $\mathbf{E}\,|g(X)| < \infty$, the set $A = \{y \in \mathbb{R}\colon \int_{\mathbb{R}} |g(x)|\, f_{X,Y}(x,y)dx < \infty\}$ satisfies $m(A^c) = 0$, where m denotes Lebesgue measure on $\mathbb{R}$. (Recall that $\mathbf{E}g(X) = \int_{\mathbb{R}^2} g(x) f_{X,Y}(x,y)dxdy$ by Theorem 1.60.) Since $\mathbf{P}(Y \in A) = \int_A f_Y(y)dy$ and $m(A^c) = 0$, we conclude that $\mathbf{P}(Y \in A) = 1$. By the definition of $f_{X|Y}$, Fubini's Theorem and Jensen's inequality,

$$\begin{aligned}\infty > \mathbf{E}\,|g(X)| &= \int_{\mathbb{R}} |g(x)|\, f_X(x)dx \geq \int_R |g(x)| \Big(\int_A f_{X|Y}(x|y) f_Y(y)dy \Big) dx \\ &= \int_A \Big(\int_{\mathbb{R}} |g(x)|\, f_{X|Y}(x|y)dx \Big) f_Y(y)dy \geq \int_A |h(y)|\, f_Y(y)dy = \mathbf{E}\,|h(Y)|\,.\end{aligned}$$

For any Borel measurable $B \subseteq \mathbb{R}$, if we use the definition of h, Fubini's Theorem, and the definition of $f_{X|Y}$,

$$\begin{aligned}\mathbf{E}(h(Y)1_B(Y)) &= \int_{B\cap A} h(y) f_Y(y)dy = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} g(x) f_{X|Y}(x|y)dx \Big) 1_{B\cap A}(y) f_Y(y)dy \\ &= \int_{\mathbb{R}^2} g(x) 1_{B\cap A}(y) f_{X,Y}(x,y)dxdy = \mathbf{E}(g(X)1_B(Y)).\end{aligned}$$

□

From the conditional density $f_{X|Y}$, we have for any $\omega \in \Omega$ a conditional probability measure $\mu_{X|Y}(\cdot,\omega)$ defined for any Borel measurable $B \subseteq \mathbb{R}$ by $\mu_{X|Y}(B,\omega) := \int_B f_{X|Y}(x|Y(\omega))dx$. We observed above that $\mathbf{E}(g(X)|Y) = h(Y) = \int_{\mathbb{R}} g(x) f_{X|Y}(x|Y)dx$. The measure $\mu_{X|Y}(\cdot,\omega)$ represents the distribution of X, if $Y(\omega)$ is fixed. So, using our intuition from Theorem 1.60, we anticipate that $\mathbf{E}(g(X)|Y) = \int_{\mathbb{R}} \mu_{X|Y}(x,Y) g(x)dx$. That is, conditional expectation can be constructed by averaging the family of conditional probability measures $\mu_{X|Y}(\cdot,\omega)$. We generalize this construction, replacing $\sigma(Y)$ by a more general σ-algebra $\mathcal{G}$.

Definition 5.29 (Regular Conditional Distribution). Let $X\colon \Omega \to S$ be a measurable function from $(\Omega,\mathcal{F})$ to $(S,\mathcal{B})$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. A function $\mu_{X|\mathcal{G}}(\cdot,\cdot)\colon \mathcal{B}\times\Omega \to [0,1]$ is called a **regular conditional probability distribution** of X given $\mathcal{G}$ if:

- $\mu_{X|\mathcal{G}}(A,\cdot) = \mathbf{E}(1_{X\in A}|\mathcal{G})$ for every $A \in \mathcal{B}$.
- For any $\omega \in \Omega$, the set function $\mu_{X|\mathcal{G}}(\cdot,\omega)$ is a probability measure.

In the case $S = \Omega$, $\mathcal{B} = \mathcal{F}$ and $X(\omega) = \omega$ for all $\omega \in \Omega$, we call $\mu_{X|\mathcal{G}}$ a **regular conditional probability** on $\mathcal{F}$ given $\mathcal{G}$.

If a regular conditional probability exists, we can construct conditional expectation from it, using a variant of the Change of Variables formula, Theorem 1.60.

Exercise 5.30. Let X be $\mathcal{F}$-measurable and let Y be $\mathcal{G}$-measurable, real-valued random variables, where $\mathcal{G} \subseteq \mathcal{F}$. Let $\mu_{X|\mathcal{G}}$ be a regular conditional probability of X given $\mathcal{G}$. Let $h\colon \mathbb{R}^2 \to \mathbb{R}$ be a Borel measurable function with $\mathbf{E}\,|h(X,Y)| < \infty$. Then, almost surely with respect to $\omega \in \Omega$,

$$\mathbf{E}(h(X,Y)|\mathcal{G})(\omega) = \int_{\mathbb{R}} h(x,Y(\omega))\mu_{X|\mathcal{G}}(x,\omega)dx.$$

In particular, if Y is constant and if $\mathbf{E}\,|X| < \infty$,

$$\mathbf{E}(X|\mathcal{G})(\omega) = \int_{\mathbb{R}} x\mu_{X|\mathcal{G}}(x,\omega)dx.$$

6. Martingales

The simple random walk, and certain functions of other random walks, can be generalized to a class of stochastic processes called martingales. As in the case of random walks, we define the "information known up to time n" using a sequence of σ-algebras known as a filtration. For simplicity, we specialize to real-valued random variables below, though a theory of vector-valued martingales could be given.

Definition 6.1 (**Filtration**)**.** Let $(\Omega, \mathcal{F})$ be a measurable space. A **filtration** is a non-decreasing sequence of σ-algebras $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \mathcal{F}$.

Definition 6.2 (**Adapted Random Variables**)**.** Let $X_0, X_1, \ldots$ be real-valued random variables on a measurable space $(\Omega, \mathcal{F})$. Let $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$ be a filtration. We say that $X_0, X_1, \ldots$ is **adapted** to this filtration if, for every $n \geq 0$, X_n is $\mathcal{F}_n$-measurable (equivalently, $\sigma(X_n) \subseteq \mathcal{F}_n$).

Note that $X_0, X_1, \ldots$ is adapted to $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$ if and only if $\sigma(X_0, X_1, \ldots, X_n) \subseteq \mathcal{F}_n$ for every $n \geq 1$. Therefore,

Definition 6.3 (**Canonical Filtration**)**.** Let $X_0, X_1, \ldots$ be real-valued random variables on a measurable space $(\Omega, \mathcal{F})$. Let $\mathcal{F}_n := \sigma(X_0, X_1, \ldots, X_n)$. Then $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$ is the smallest filtration such that $X_0, X_1, \ldots$ is adapted to this filtration. We therefore refer to this filtration as the **canonical filtration**.

Definition 6.4 (**Martingale**)**.** Let $X_0, X_1, \ldots$ be real-valued random variables on a measurable space $(\Omega, \mathcal{F})$. Let $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$ be a filtration. A **martingale** is a pair $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ such that $X_0, X_1, \ldots$ is adapted to $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$, such that $\mathbf{E}\,|X_n| < \infty$ for all $n \geq 0$, and almost surely,

$$\mathbf{E}(X_{n+1}|\mathcal{F}_n) = X_n, \qquad \forall\, n \geq 0.$$

Remark 6.5. Suppose $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ is a martingale. Let $\mathcal{G}_n \subseteq \mathcal{F}_n$ for all $n \geq 0$ be σ-algebras. Assume that $X_0, X_1, \ldots$ is adapted to $\mathcal{G}_0 \subseteq \mathcal{G}_1 \subseteq \cdots$. From the Tower Property, Exercise 5.17, $\mathbf{E}(X_{n+1}|\mathcal{G}_n) = \mathbf{E}(\mathbf{E}(X_{n+1}|\mathcal{F}_n)|\mathcal{G}_n) = \mathbf{E}(X_n|\mathcal{G}_n) = X_n$, by the containment $\sigma(X_0, X_1, \ldots, X_n) \subseteq \mathcal{G}_n$ and by Example 5.10. So, $((X_n)_{n\geq 0}, (\mathcal{G}_n)_{n\geq 0})$ is a martingale. In particular, if $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ is a martingale, then $X_0, X_1, \ldots$ is also a martingale with respect to the canonical filtration. So, from now on, if we say that $X_0, X_1, \ldots$ is a martingale, without explicitly mentioning a filtration, we mean that $X_0, X_1, \ldots$ is a martingale with respect to the canonical filtration.

Remark 6.6. From Proposition 5.12, if $X_0, X_1, \ldots$ is a martingale, then $\mathbf{E}X_n = \mathbf{E}X_0$ for all $n \geq 0$. That is, the expected value of a martingale is constant in $n \geq 0$.

Example 6.7 (**Simple Random Walk**)**.** The Simple Random Walk on $\mathbb{Z}$ from Definition 4.15 is a martingale. Recall that $X_0 := 0$, $Y_1, Y_2, \ldots$ are i.i.d. with such that $\mathbf{P}(Y_1 = 1) = \mathbf{P}(Y_1 = -1) = 1/2$, and $X_n := Y_1 + \cdots + Y_n$ for any $n \geq 1$. Then $\mathbf{E}\,|X_n| \leq n < \infty$ for all $n \geq 0$. Also, Y_{n+1} is independent of $X_0, \ldots, X_n$ for any $n \geq 0$, so Y_{n+1} is independent of $\mathcal{F}_n := \sigma(X_0, \ldots, X_n)$ by Exercise 1.93. So, by the definition of X_{n+1}, Proposition 5.13, Example 5.10 and Exercise 5.11,

$$\mathbf{E}(X_{n+1}|\mathcal{F}_n) = \mathbf{E}(X_n + Y_{n+1}|\mathcal{F}_n) = \mathbf{E}(X_n|\mathcal{F}_n) + \mathbf{E}(Y_{n+1}|\mathcal{F}_n) = X_n + \mathbf{E}(Y_{n+1}) = X_n.$$

Example 6.8 (**Gambler's Ruin**). Let $0 < p < 1$. Suppose you are playing a game of chance. For each round of the game, with probability p you win \$1 and with probability $1-p$ you lose \$1. Suppose you start with \$50 and you decide to quit playing when you reach either \$0 or \$100. With what probability will you end up with \$100?

Later on, we will answer this question using Martingales and Stopping Times.

Let $Y_1, Y_2, \ldots$ be independent random variables such that $\mathbf{P}(Y_n = 1) =: p$ and $\mathbf{P}(Y_n = -1) = 1 - p =: q \ \forall\ n \geq 1$. Let $Y_0 := 50$. Let $Z_n = Y_0 + \cdots + Y_n$, and let $X_n := (q/p)^{Z_n} \ \forall\ n \geq 0$. Then Z_n denotes the amount of money you have at time $n \leq 50$. For any $n \geq 1$, let $\mathcal{F}_n := \sigma(Y_0, \ldots, Y_n)$. We claim that $X_0, X_1, \ldots$ is a martingale with respect to $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$. Indeed, X_n is $\mathcal{F}_n$-measurable for any $n \geq 0$, $\mathbf{E}\,|X_n| \leq \max((q/p)^{50+n}, (q/p)^{50-n}) < \infty$ for all $n \geq 0$. So, by Proposition 5.16 and Exercise 5.11,

$$\begin{aligned}\mathbf{E}(X_{n+1}|\mathcal{F}_n) &= \mathbf{E}((q/p)^{Z_{n+1}}|\mathcal{F}_n) = (q/p)^{Z_n}\mathbf{E}((q/p)^{Y_{n+1}}|\mathcal{F}_n) = (q/p)^{Z_n}\mathbf{E}((q/p)^{Y_{n+1}}) \\ &= (q/p)^{Z_n}(p(q/p) + q(q/p)^{-1}) = (q/p)^{Z_n}(q+p) = (q/p)^{Z_n} = X_n.\end{aligned}$$

Exercise 6.9 (**Binomial Option Pricing Model**). Let $u, d > 0$. Let $0 < p < 1$. Let $Y_1, Y_2, \ldots$ be independent random variables such that $\mathbf{P}(Y_n = \log u) =: p$ and $\mathbf{P}(Y_n = \log d) = 1 - p \ \forall\ n \geq 1$. Let Z_0 be a fixed constant. Let $Z_n := Y_0 + \cdots + Y_n$, and let $V_n := e^{Z_n} \ \forall\ n \geq 1$. In general, $V_0, V_1, \ldots$ will not be a martingale, but we can e.g. compute $\mathbf{E}V_n$, by modifying $V_0, V_1, \ldots$ to be a martingale.

First, note that if $n \geq 1$, then Z_n has a binomial distribution, in the sense that

$$\mathbf{P}(Z_n = X_0 + i\log u + (n-i)\log d) = \binom{n}{i} p^i (1-p)^{n-i}, \qquad \forall\, 0 \leq i \leq n.$$

For any $n \geq 1$, let $\mathcal{F}_n := \sigma(Y_0, \ldots, Y_n)$. Define

$$r := p(u-d) - 1 + d.$$

Here we chose r so that $p = \frac{1+r-d}{u-d}$. For any $n \geq 0$, define

$$X_n := (1+r)^{-n} V_n.$$

Show that $X_0, X_1, \ldots$ is a martingale with respect to $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$. Consequently,

$$(1+r)^{-n}\mathbf{E}V_n = \mathbf{E}V_0, \qquad \forall\, n \geq 0.$$

Exercise 6.10. Let $M_0, M_1, \ldots$ be a martingale with $\mathbf{E}M_n^2 < \infty$ for all $n \geq 0$. Show that the increments $M_2 - M_1, M_3 - M_2, \ldots$ are orthogonal in the following sense. For any $i, j \geq 1$ with $i \neq j$,

$$\mathbf{E}(M_{i+1} - M_i)(M_{j+1} - M_j) = 0.$$

This property is sometimes called **orthogonality of martingale increments**.

Exercise 6.11. Let X be a real-valued random variable on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Assume $\mathbf{E}\,|X| < \infty$. Let $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \mathcal{F}$ be σ-algebras. For any $n \geq 0$, define $X_n := \mathbf{E}(X|\mathcal{F}_n)$. Show that $X_0, X_1, \ldots$ is a martingale. (Optional challenge question: For any martingale $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$, is there a random variable X with $\mathbf{E}\,|X| < \infty$ such that $X_n = \mathbf{E}(X|\mathcal{F}_n)$ for all $n \geq 0$?)

The definition of stopping time for a random walk generalizes to martingales, with essentially no change.

Definition 6.12 (**Stopping Time**)**.** Let $X_0, X_1, \ldots$ be a martingale with respect to a filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$. We say that $N \colon \Omega \to \{0, 1, 2, \ldots\} \cup \{\infty\}$ is a **stopping time** if, for any $n \in \{0, 1, 2, \ldots\}$, $\{N = n\} \in \mathcal{F}_n$.

For random walks, we always assumed that X_0 is constant almost surely, i.e. $\mathcal{F}_0 = \{\emptyset, \Omega\}$.

Exercise 6.13. Let M, N be stopping times for a martingale $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$. Show that $\max(M, N)$ and $\min(M, N)$ are stopping times. In particular, if $n \geq 0$ is fixed, then $\max(M, n)$ and $\min(M, n)$ are stopping times

Definition 6.14 (**Submartingale, Supermartingale**)**.** Let $X_0, X_1, \ldots$ be real-valued random variables on a measurable space $(\Omega, \mathcal{F})$. Let $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$ be a filtration. A **submartingale** is a pair $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ such that $X_0, X_1, \ldots$ is adapted to $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$, such that $\mathbf{E}\,|X_n| < \infty$ for all $n \geq 0$, and almost surely,

$$\mathbf{E}(X_{n+1}|\mathcal{F}_n) \geq X_n, \qquad \forall\, n \geq 0.$$

A **supermartingale** is a pair $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ such that $X_0, X_1, \ldots$ is adapted to $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$, such that $\mathbf{E}\,|X_n| < \infty$ for all $n \geq 0$, and almost surely,

$$\mathbf{E}(X_{n+1}|\mathcal{F}_n) \leq X_n, \qquad \forall\, n \geq 0.$$

Remark 6.15. If $X_0, X_1, \ldots$ is a submartingale, then $-X_0, -X_1, \ldots$ is a supermartingale. So, any result about submartingales has a corresponding statement for supermartingales. For this reason, we will specialize some statements below to one of these two cases. Moreover, note that $X_0, X_1, \ldots$ is a martingale if and only if it is a supermartingale and a submartingale. In particular, if some statement holds for submartingales, then it also holds for martingales.

Exercise 6.16. Let $X_0, X_1, \ldots$ and let $Y_0, Y_1, \ldots$ be submartingales adapted to the same filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$. Show that $X_0 + Y_0, X_1 + Y_1, \ldots$ is a submartingale adapted to the filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$. Consequently, a sum of supermartingales is a supermartingale, and a sum of martingales is a martingale (when they are adapted to the same filtration).

Exercise 6.17.

(i) Let $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ be a submartingale. Show that, almost surely, $\mathbf{E}(X_n|\mathcal{F}_m) \geq X_m$ for any $n > m$. Consequently, $n \mapsto \mathbf{E}X_n$ is nondecreasing.

(ii) Let $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ be a supermartingale. Show that, almost surely, $\mathbf{E}(X_n|\mathcal{F}_m) \leq X_m$ for any $n > m$. Consequently, $n \mapsto \mathbf{E}X_n$ is nonincreasing.

(iii) Let $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ be a martingale. Let $\phi \colon \mathbb{R} \to \mathbb{R}$ be convex. Assume $\mathbf{E}\,|\phi(X_n)| < \infty$ for all $n \geq 1$. Show that $((\phi(X_n))_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ is a submartingale.

(iv) Let $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ be a submartingale. Let $\phi \colon \mathbb{R} \to \mathbb{R}$ be convex and nondecreasing. Assume $\mathbf{E}\,|\phi(X_n)| < \infty$ for all $n \geq 1$. Show that $((\phi(X_n))_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ is a submartingale.

Example 6.18. Let $1 \leq p < \infty$ and $c \in \mathbb{R}$. Some convex functions $\phi \colon \mathbb{R} \to \mathbb{R}$ applied to the previous exercise include $\phi(x) := |x|^p$, $\phi(x) := \max(x - c, 0)$, $\phi(x) := \max(x, c)$, and $\phi(x) := e^x$, $\forall\, x \in \mathbb{R}$. For example, if $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ is a submartingale, then $((\max(X_n - c, 0))_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ is a submartingale. For another example, if $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ is a supermartingale, then $((\min(X_n, c))_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ is a supermartingale, since $((-X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ is a submartingale and $-\min(-x, c) = \max(x, -c)$ is convex and nonincreasing.

Exercise 6.19 (**Azuma's Inequality**)**.** In this exercise, we prove a generalization of the Hoeffding inequality to martingales. Let $c_1, c_2, \ldots > 0$. Let $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ be a martingale. Assume that $|X_n - X_{n-1}| \leq c_n$ for all $n \geq 1$. Then for any $t > 0$,

$$\mathbf{P}(|X_n - X_0| > t) \leq 2e^{-\frac{t^2}{2\sum_{i=1}^n c_i^2}}.$$

Prove this inequality using the following steps.

- Let $\alpha > 0$. Show that $\mathbf{E}e^{\alpha(X_n - X_0)} = \mathbf{E}[e^{\alpha(X_{n-1}-X_0)}\mathbf{E}(e^{\alpha(X_n - X_{n-1})}|\mathcal{F}_{n-1})]$.
- For any $y \in [-1, 1]$, show that $e^{\alpha c_n y} \leq \frac{1+y}{2}e^{\alpha c_n} + \frac{1-y}{2}e^{-\alpha c_n}$.
- Take the conditional expectation of this inequality when $y = (X_n - X_{n-1})/c_n$.
- Now argue as in Hoeffding's inequality.

Using Azuma's inequality, deduce **McDiarmid's Inequality**. Let $X_1, \ldots, X_n$ be independent real-valued random variables. Let $c_1, c_2, \ldots > 0$. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a measurable function such that, for any $1 \leq m \leq n$,

$$\sup_{x_1,\ldots,x_{m-1},x_m,x_m',x_{m+1},\ldots,x_n\in\mathbb{R}} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{m-1}, x_m', x_{m+1}, \ldots, x_n)| \leq c_m.$$

Then, for any $t > 0$,

$$\mathbf{P}(|f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n)| > t) \leq 2e^{-\frac{t^2}{2\sum_{i=1}^n c_i^2}}.$$

(Note that a linear function f recovers Hoeffding's inequality, Theorem 2.35.)

6.1. **Gambling Strategies.** Suppose you can bet any amount of money you want on a fair coin flip. And the coin can be flipped any number of times, i.e. you can play this game any number of times. If you bet \$$c$ with $c > 0$ and the coin lands heads, then you win \$$c$, but if the coin lands tails, then you lose \$$c$. A naïve strategy to make money off of this game is the following. Just keep doubling your bet until you win. For example, start by betting \$1. If you lose, bet \$2. If you lose that, bet \$4. Then let's say you finally won, then in total you won \$4 and you lost \$3, so you gained \$1 in total. We know that the probability of losing $k > 0$ rounds of this game in a row is 2^{-k}, so it seems like this strategy must win money. However, there are some caveats to this analysis.

First, if your starting bet is \$1, and if you lose twenty rounds of the game in a row, you will be betting over one million dollars. More generally, if you lose k times in a row, you will have to bet \$$2^k$. So, when $k \geq 20$, most people would not be able to continue playing the game, i.e. they would lose all of their money.

Second, *your expected gain from every round of the game is zero.* At each round of the game, no matter what your bet is, your expected earnings are zero. So, it is impossible to win money in this game, in expectation. And indeed, the Law of Large Numbers assures us that when the game is repeated many times, we will earn zero dollars on average, with probability 1.

It turns out that, no matter what betting strategy is chosen in this game, there is still no way to make any money. We will prove this using martingale methods. In fact, these gambling strategies initiated the study of martingales in France in the 1700s.

Let $Y_1, Y_2, \ldots$ each be independent random variables such that $\mathbf{P}(Y_n = 1) = \mathbf{P}(Y_n = -1) = 1/2$ for every $n \geq 0$. For any $n \geq 1$, let $X_n := Y_1 + \cdots + Y_n$. Let $X_0 = 0$. If someone bets one dollar at every round of the game, then their profit is X_n after the n^{th} round of the game. Since $\mathbf{E}Y_1 = 0$, Example 6.7 implies that $X_0, X_1, \ldots$ is a martingale. A gambling

strategy for the n^{th} round of the game can use any information from the previous rounds of the game. Let H_n be the amount of money we bet in the n^{th} round of the game. We formalize our assumption about $H_1, H_2, \ldots$ using the filtration.

Definition 6.20 (Predictable). Let $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$ be a filtration on a measurable space $(\Omega, \mathcal{F})$. We say that a sequence of real-valued random variables $H_1, H_2, \ldots$ is **predictable** (or previsible) if, for any $n \geq 1$, H_n is $\mathcal{F}_{n-1}$-measurable.

When the m^{th} round of the game occurs, we earn $H_m(X_m - X_{m-1})$ dollars. In summary, our wealth W_n at time $n \geq 1$ is then

$$W_n := \sum_{m=1}^{n} H_m(X_m - X_{m-1}), \qquad W_0 := 0.$$

We will now prove that we cannot make money from this game.

Theorem 6.21. *Let $H_1, H_2, \ldots$ be predictable for a filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$.*

- *Let $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ be a martingale. If $\mathbf{E}\,|W_n| < \infty$ for all $n \geq 1$, then $((W_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ is a martingale.*
- *Let $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ be a submartingale (or supermartingale). If $\mathbf{E}\,|W_n| < \infty$ and $H_n \geq 0$ for all $n \geq 1$, then $((W_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ is a submartingale (or supermartingale).*

In order to show that $\mathbf{E}\,|W_n| < \infty$ for all $n \geq 1$, it suffices in either case to have real constants $c_1, c_2, \ldots$ such that $|H_n| \leq c_n$ for all $n \geq 1$, or $\exists\ p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$ such that $\|H_n\|_q < \infty$ for all $n \geq 1$ and $\|X_n\|_p < \infty$ for all $n \geq 0$.

Remark 6.22. The quantity $\sum_{m=1}^{n} H_m(X_m - X_{m-1})$ is a finite version of a stochastic integral. And in fact, there is a corresponding statement to be made about stochastic integrals, namely that you cannot make money off of (continuous-time) supermartingales.

Remark 6.23. Allowing $H_n < 0$ would correspond to betting negative amounts on a supermartingale, so that the gambler could assume the position of the "house." So, $H_n \geq 0$ is a sensible constraint in the second item of Theorem 6.21.

Proof of Theorem 6.21. We first prove the last claim. From the triangle inequality and Hölder's inequality, Theorem 1.48,

$$\mathbf{E}\,|W_n| \leq \sum_{m=1}^{n} \|H_n\|_q \left(\|X_m\|_p + \|X_{m-1}\|_p\right) < \infty.$$

In the case $q = \infty, p = 1$, note that $X_0, X_1, \ldots$ is a (super/sub)martingale, so that $\mathbf{E}\,|X_m| < \infty$ for all $m \geq 0$. Also, since $X_0, X_1, \ldots$ is adapted and $H_1, H_2, \ldots$ is predictable, for any $1 \leq m \leq k \leq n$, $H_m X_k$ is $\mathcal{F}_k$-measurable, hence $\mathcal{F}_n$-measurable since $k \leq n$, so that W_n is $\mathcal{F}_n$-measurable for any $n \geq 0$.

Assume that $X_0, X_1, \ldots$ is a (sub)martingale. Observe that, for any $n \geq 0$

$$W_{n+1} - W_n = H_{n+1}(X_{n+1} - X_n)$$

Since $H_1, H_2, \ldots$ is predictable, we have by Proposition 5.16, for any $n \geq 0$,

$$\mathbf{E}(W_{n+1} - W_n | \mathcal{F}_n) = \mathbf{E}(H_{n+1}(X_{n+1} - X_n)|\mathcal{F}_n) = H_{n+1}\mathbf{E}(X_{n+1} - X_n|\mathcal{F}_n) \geq 0.$$

The last inequality follows since either $X_0, X_1, \ldots$ is a martingale, or it is a submartingale and $H_{n+1} \geq 0$. Note also that $\mathbf{E}\,|H_{n+1}(X_{n+1} - X_n)| = \mathbf{E}\,|W_{n+1} - W_n| < \infty$ by assumption, so the assumption of Proposition 5.16 is justified. Finally, if $X_0, X_1, \ldots$ is a supermartingale, we apply the above argument to the submartingale $-X_0, -X_1, \ldots$. □

From Remark 6.6, a martingale satisfies $\mathbf{E}X_n = \mathbf{E}X_0$ for all $n \geq 0$. In some cases, we can replace n with a stopping time N in this equality. However, this cannot always hold.

Example 6.24. Let $X_0, X_1, \ldots$ be the simple random walk on $\mathbb{Z}$. Note that $\mathbf{E}X_0 = 0$. As shown in Example 6.7, $X_0, X_1, \ldots$ is a martingale. Let $N := \min\{n \geq 1 \colon X_n = 1\}$ be the return time to 1. Then N is a stopping time and $X_N = 1$, so $\mathbf{E}X_N = 1 \neq 0 = \mathbf{E}X_0$.

Remark 6.25. Let $a, b \in \mathbb{R}$. We use the notation $a \wedge b := \min(a, b)$.

Theorem 6.26 (**Optional Stopping Theorem, Version 1**). *Let $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ be a submartingale (or supermartingale, or martingale) and let $T \leq N$ be a stopping times for $(\mathcal{F}_n)_{n\geq 0}$. Then $X_{0\wedge N} - X_{0\wedge T}, X_{1\wedge N} - X_{1\wedge T}, \ldots$ is a submartingale (or supermartingale, or martingale) adapted to $(\mathcal{F}_n)_{n\geq 0}$.*

Consequently, $X_{0\wedge N}, X_{1\wedge N}, \ldots$ is a submartingale (or supermartingale, or martingale) adapted to $(\mathcal{F}_n)_{n\geq 0}$.

So, if $X_0, X_1, \ldots$ is a martingale, then $\mathbf{E}X_{n\wedge N} = \mathbf{E}X_0$ for all $n \geq 0$.

Proof. We may assume that $X_0, X_1, \ldots$ is a submartingale, since if $X_0, X_1, \ldots$ is a supermartingale we apply the submartingale case to $-X_0, -X_1, \ldots$, and the martingale case follows by combining the submartingale and supermartingale cases. We first consider the case $T := 0$. For any $n \geq 1$, define $H_n := 1_{N\geq n} = 1 - 1_{N\leq n-1} = 1 - \sum_{m=0}^{n-1} 1_{N=m}$. Since N is a stopping time, if $m \leq n-1$, $\{N = m\} \in \mathcal{F}_m \subseteq \mathcal{F}_{n-1}$, so that H_n is $\mathcal{F}_{n-1}$-measurable. That is, $H_1, H_2, \ldots$ is predictable. Let $W_0 := 0$ and for any $n \geq 1$, define $W_n := \sum_{m=1}^{n} H_m(X_m - X_{m-1})$. By Theorem 6.21, $W_0, W_1, \ldots$ is a submartingale. By the definition of H_m,

$$W_n = \sum_{m=1}^{n} (1_{\{N\geq m\}})(X_m - X_{m-1}) = \sum_{m=1}^{n} (X_{m\wedge N} - X_{(m-1)\wedge N}) = X_{n\wedge N} - X_0. \qquad (*)$$

In the case of general T, we let $H_n := 1_{N\geq n>T} = 1_{N\geq n} - 1_{T\geq n}$ for any $n \geq 1$. Then $H_1, H_2, \ldots$ is predictable, and for any $n \geq 1$, $W_n := \sum_{m=1}^{n} H_m(X_m - X_{m-1}) = X_{n\wedge N} - X_{n\wedge T}$ (using $(*)$) is a submartingale by Theorem 6.21.

Finally, in the case $T = 0$, since the constant random variable $X_0, X_0, \ldots$ is a submartingale adapted to $(\mathcal{F}_n)_{n\geq 0}$, we add it to the submartingale from $(*)$ to conclude that $X_{0\wedge N}, X_{1\wedge N}, \ldots$ is a submartingale by Exercise 6.16. □

6.2. **Maximal Inequalities and Up-crossing.** As discussed in Section 2.6 and Theorem 2.43, in order to prove pointwise convergence, one often needs a weak-type maximal inequality. Recall e.g. that we used Kolmogorov's Maximal Inequality, Theorem 2.24, in the proof of the Strong Law of Large Numbers. In the setting of martingales, we require a weak type $(1, 1)$ maximal inequality, known as Doob's inequality.

Theorem 6.27 (**Doob's Maximal Inequality**). *Let $X_0, X_1, \ldots$ be a submartingale. Let $t > 0$. Then for any integer $n \geq 0$,*

$$t\mathbf{P}(\max_{0\leq m\leq n} X_m > t) \leq \mathbf{E}X_n 1_{\{\max_{0\leq m\leq n} X_m > t\}} \leq \mathbf{E}\max(0, X_n).$$

Proof. Let $A := \{\max_{0\le m\le n} X_m > t\}$ and let $N := \min\{m \ge 0\colon X_m > t\}$. Then N is a stopping time since $\{N = s\} = \{X_0 \le t, \dots, X_{s-1} \le t, X_s > t\}$ for any $s \ge 0$, so $N \wedge n$ is a stopping time by Exercise 6.13. Since $X_{N\wedge n}1_A = X_N 1_A \ge t1_A$ and $X_{N\wedge n}1_{A^c} = X_n 1_{A^c}$,

$$\mathbf{E}X_{N\wedge n} = \mathbf{E}X_{N\wedge n}(1_A + 1_{A^c}) = \mathbf{E}X_N 1_A + \mathbf{E}X_n 1_{A^c} \ge t\mathbf{P}(A) + \mathbf{E}X_n 1_{A^c}.$$

By Theorem 6.26 for the stopping time $N \wedge n$ and the constant stopping time n, $X_{0\wedge n} - X_{0\wedge N}, X_{1\wedge n} - X_{1\wedge N}, \dots$ is a submartingale, so that $\mathbf{E}X_n - \mathbf{E}X_{n\wedge N} \ge \mathbf{E}(X_{0\wedge n} - X_{0\wedge N}) = 0$. In summary, $\mathbf{E}X_n \ge t\mathbf{P}(A) + \mathbf{E}X_n 1_{A^c}$, so that $\mathbf{E}X_n 1_A \ge t\mathbf{P}(A)$, as desired. The second inequality follows since $X1_B \le \max(X, 0)$ for any real random variable X and for any measurable set B. □

Remark 6.28. Doob's maximal inequality implies Kolmogorov's Maximal Inequality, Theorem 2.24. Let $Y_1, Y_2, \dots$ be independent mean zero real random variables with $\mathbf{E}Y_n^2 < \infty$ for all $n \ge 1$. Let $X_0 := 0$ and let $X_n := (Y_1 + \cdots + Y_n)^2$ for any $n \ge 1$. Then $X_0, X_1, \dots$ is a submartingale by Exercise 6.17(iii). So, Doob's inequality says, for any $t > 0$,

$$\begin{aligned}\mathbf{P}(\max_{1\le m\le n} |Y_1 + \cdots + Y_m| > t) = \mathbf{P}(\max_{1\le m\le n} X_m > t^2) &\le \frac{\mathbf{E}\max(X_n, 0)}{t^2} \\ &= \frac{\mathbf{E}X_n}{t^2} = \frac{\mathrm{Var}(Y_1) + \cdots + \mathrm{Var}(Y_n)}{t^2}.\end{aligned}$$

The weak-type $(1,1)$ inequality of Theorem 6.27 can be interpolated in a standard argument to strong L_p bounds when $p > 1$.

Corollary 6.29 (L_p **Maximal Inequality**). *Let $X_0, X_1, \dots$ be a submartingale. For any $x \in \mathbb{R}$, denote $x_+ := \max(x, 0)$. Let $p > 1$. Then for any integer $n \ge 0$,*

$$\|(\max_{0\le m\le n} X_m)_+\|_p \le \frac{p}{p-1}\|(X_n)_+\|_p.$$

Consequently, if $X_0, X_1, \dots$ is a martingale, then for any $n \ge 0$, $p > 1$,

$$\|\max_{0\le m\le n} |X_m|\,\|_p \le \frac{p}{p-1}\|X_n\|_p.$$

Proof. Denote $X_n^* := \max_{0\le m\le n} X_m$. Using Theorem 1.86, Theorem 6.27, Fubini's Theorem, Theorem 1.66, and Hölder's inequality, Theorem 1.48 with $1/p + 1/q = 1$, so that $q = p/(p-1)$,

$$\begin{aligned}\mathbf{E}\,|(X_n^*)_+|^p &= \int_0^\infty pt^{p-1}\mathbf{P}(X_n^* > t)dt \le \int_0^\infty pt^{p-2}\mathbf{E}X_n 1_{X_n^*>t}dt = \mathbf{E}\Big(X_n\int_0^\infty pt^{p-2}1_{X_n^*>t}dt\Big) \\ &= \mathbf{E}\Big(X_n\int_0^{(X_n^*)_+} pt^{p-2}dt\Big) = \frac{p}{p-1}\mathbf{E}(X_n(X_n^*)_+^{p-1}) \le \frac{p}{p-1}\mathbf{E}(|X_n|\,(X_n^*)_+^{p-1}) \\ &\le \frac{p}{p-1}(\mathbf{E}\,|X_n|^p)^{1/p}(\mathbf{E}(X_n^*)_+^p)^{(p-1)/p}\end{aligned}$$

In the case that X_n^* is bounded, we divide both sides by the right-most term to conclude. The general case then follows by applying the bounded case to $X_n^* \wedge s$, letting $s \to \infty$, and using Monotone Convergence, Theorem 1.54. The final statement follows from the first since $|X_0|, |X_1|, \dots$ is a submartingale from Exercise 6.17(iii). □

Exercise 6.30. In the L_p maximal inequality, the constant $\frac{p}{p-1}$ goes to infinity as $p \to 1^+$. So, one might guess that the L_p maximal inequality does not hold when $p = 1$. (If so, this justifies the need to prove a weaker statement when $p = 1$, i.e. Doob's inequality.) Using the simple random walk on $\mathbb{Z}$, show that the L_p maximal inequality does not hold when $p = 1$. (Hint: use the probabilities from Example 4.22.)

Exercise 6.31. Show that the second part of the L_p maximal inequality cannot hold when $X_0, X_1, \ldots$ is a submartingale. That is, for any $n \geq 1$, find a submartingale $X_0, X_1, \ldots$ such that, for any $p > 1$, $\|\max_{0 \leq m \leq n} |X_m|\|_p > 0$ but such that $\|X_n\|_p = 0$. (Hint: just consider a non-random sequence of numbers.)

In order to prove the almost sure convergence of martingales, we will bound the number of up-crossings of the martingale.

Definition 6.32 (**Up-crossing**)**.** Let $X_1, \ldots, X_n \colon \Omega \to \mathbb{R}$ be a sequence of random variables. Let $a, b \in \mathbb{R}$ with $a < b$. Define the number of up-crossings of the sequence $X_1, \ldots, X_n$ across the interval $[a, b]$ to be $U_n[a, b] \colon \Omega \to \mathbb{Z}$ so that, for any $\omega \in \Omega$, $U_n[a, b](\omega)$ is the largest integer $m > 0$ such that there exist integers $0 \leq s_1 < t_1 < \cdots < s_m < t_m \leq n$ such that $X_{s_i}(\omega) \leq a$ and $X_{t_i}(\omega) \geq b$ for all $1 \leq i \leq m$.

If $X_1, X_2, \ldots$ is an infinite sequence of real-valued random variables, the sequence $U_1[a, b]$, $U_2[a, b], \ldots$ monotonically increases to a random variable denoted $U[a, b]$. We say $U[a, b]$ is the total number of up-crossings of the sequence $X_1, X_2, \ldots$ across $[a, b]$.

Exercise 6.33. Let $x_1, x_2, \ldots$ be a sequence of real numbers. Show that total number of up-crossings of the sequence across the interval $[a, b]$ is finite for any $a, b \in \mathbb{R}$ with $a < b$ if and only if the sequence $x_1, x_2, \ldots$ converges to some $x \in [-\infty, \infty]$. (Here the random variables in the definition of up-crossing are chosen to be constant $X_m := x_m$ for all $m \geq 0$.)

The observation of Doob is that bounding up-crossings in expectation can also prove almost sure convergence of certain (super)martingales.

Lemma 6.34 (**Doob's Up-crossing Inequality**)**.** *Let $X_0, X_1, \ldots$ be a supermartingale. Then for any $a, b \in \mathbb{R}$ with $a < b$,*

$$(b - a)\mathbf{E}U_n[a, b] \leq \mathbf{E}\max(a - X_n, 0) - \mathbf{E}\max(a - X_0, 0).$$

Proof. Let $N_0 := -1$ and for any integer $k \geq 1$, define

$$N_{2k-1} := \min\{m > N_{2k-2} \colon X_m \leq a\}, \qquad N_{2k} := \min\{m > N_{2k-1} \colon X_m \geq b\}.$$

Then $N_0, N_1, \ldots$ are stopping times and for any $k \geq 0, m \geq 1$, $\{N_{2k-1} < m \leq N_{2k}\} = \{N_{2k-1} \leq m - 1\} \cap \{N_{2k} \leq m - 1\}^c \in \mathcal{F}_{m-1}$, so if we define

$$H_m := 1_{N_{2k-1} < m \leq N_{2k} \text{ for some } k \geq 0},$$

then $H_1, H_2, \ldots$ is predictable. Note that if $n \geq 1$,

$$W_n := \sum_{m=1}^{n} H_m(X_m - X_{m-1}) = \sum_{\substack{m \colon N_1 < m \leq N_2 \\ N_3 < m \leq N_4, \cdots}} (X_m - X_{m-1}) = X_{N_2} - X_{N_1} + X_{N_4} - X_{N_3} + \cdots.$$

If $k := U_n[a, b]$, the last term in the sum is either $X_{N_{2k}} - X_{N_{2k-1}}$ or $X_n - X_{N_{2k+1}}$, the latter case corresponding to $N_{2k+1} < n < N_{2k+2}$, in which case $X_n - X_{N_{2k+1}} \geq X_n - a \geq -\max(a - X_n, 0)$

since $X_{N_{2k+1}} \leq a$. Meanwhile, if $X_0 \leq a$, then $N_1 = 0$, so $X_{N_2} - X_{N_1} \geq (b - X_0) = (b-a) + (a - X_0)$. In any case, the first term is at least $(b-a) + \max(a - X_0, 0)$. So,

$$W_n \geq (b-a)U_n[a,b] + \max(a - X_0, 0) - \max(a - X_n, 0).$$

From Theorem 6.21, if $W_0 := 0$, then $W_0, W_1, \ldots$ is a supermartingale, so that $\mathbf{E}W_n \leq 0$. $\square$

In the above proof, if we think of $H_1, H_2, \ldots$ as a gambling strategy, it corresponds to buying low (when the price is below a) and selling high (when the price is above b).

Exercise 6.35. Below, we will use Lemma 6.34 to show that the $U[a,b]$ is finite almost surely for a nonnegative supermartingale. In this exercise, we derive Dubins' up-crossing inequality, an improvement to Doob's result that gives exponential decay of $\mathbf{P}(U[a,b] > t)$.

Let $((X_n^1)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ and $((X_n^2)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ be supermartingales and let N be a stopping time such that $X_N^1 \geq X_N^2$.

(i) (**Switching Principle**) For any $n \geq 0$, define $Y_n := X_n^1 1_{N>n} + X_n^2 1_{N\leq n}$. Show that $((Y_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ is a supermartingale. Show the same conclusion for $Z_n := X_n^1 1_{N\geq n} + X_n^2 1_{N<n}$.

That is, if we use N to "switch" from one supermartingale to another, the random variables do not increase at the time of "switching", then we still have a supermartingale.

(ii) Let $X_0, X_1, \ldots$ be a supermartingale with $X_n \geq 0$ for all $n \geq 0$. Let $a, b \in \mathbb{R}$ with $b > a > 0$. Let $N_0 := -1$ and for any integer $k \geq 1$, define

$$N_{2k-1} := \min\{m > N_{2k-2} \colon X_m \leq a\}, \qquad N_{2k} := \min\{m > N_{2k-1} \colon X_m \geq b\}.$$

Define $V_0, V_1, \ldots$ such that $V_n := 1$ for all $0 \leq n < N_1$, and for any $k \geq 1$,

$$V_n := \begin{cases} (b/a)^{k-1}(X_n/a) & \text{, if } N_{2k-1} \leq n < N_{2k} \\ (b/a)^k & \text{, if } N_{2k} \leq n < N_{2k+1}. \end{cases}$$

Using the switching principle, show by induction on k that for any integer $k \geq 1$, $V_{0\wedge N_k}, V_{1\wedge N_k}, \ldots$ is a supermartingale.

(iii) (Dubins' Inequality) Show that for any $b > a > 0$ and for any integer $t \geq 1$,

$$\mathbf{P}(U[a,b] \geq t) \leq (a/b)^t \mathbf{E} \min(X_0/a, 1).$$

6.3. **Martingale Convergence.** Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Let $1 \leq p < \infty$. Recall that L_p is the set of (almost sure equivalence classes of) real-valued random variables X with finite L_p norm: $\|X\|_p := (\mathbf{E}|X|^p)^{1/p} < \infty$. Recall that a sequence of real-valued random variables $X_1, X_2, \ldots$ converges in L_p to a real-valued random variable X if $\|X\|_p < \infty$ and $\lim_{n\to\infty} \|X_n - X\|_p = 0$. Recall also that in Exercise 2.34 we showed that the Strong Law of Large Numbers holds for convergence in L_1. Below, we will investigate the L_1 convergence of martingales $X_0, X_1, \ldots$ as $n \to \infty$. Since some random walks are martingales, and some martingales are not random walks, the convergence of martingales is related to but distinct from the Laws of Large Numbers.

Since martingales are defined to have finite L_1 norm, it is natural to focus on the L_1 convergence of martingales. That is, we will look for conditions on a martingale $X_0, X_1, \ldots$ such that it converges in L_1 to some real-valued random variable X with $\|X\|_1 < \infty$. By Jensen's inequality, if $\exists\ 1 < p < \infty$ such that $X_0, X_1, \ldots$ converges in L_p to some real-valued

random variable X with $\|X\|_p < \infty$, then $X_0, X_1, \ldots$ converges in L_q to X for all $1 \leq q < p$. So, if we want to show convergence in L_1, it suffices to show convergence in L_p for any $p > 1$.

Some martingales cannot converge in L_1, or in any L_p space. Consider for example the simple random walk $S_0, S_1, \ldots$ on $\mathbb{Z}$. By Khintchine's inequality, Exercise 2.39, for any $1 \leq p < \infty$ there exists $c(p) > 0$ such that $\|S_n\|_p \geq c(p) \cdot \sqrt{n}$ for all $n \geq 1$. Therefore, $S_0, S_1, \ldots$ cannot converge in L_p for any $1 \leq p < \infty$.

On the other hand, some martingales will certainly converge in every L_p space where $1 \leq p < \infty$.

Exercise 6.36. Let $c_1, c_2, \ldots$ be positive constants such that $\sum_{n=1}^{\infty} c_n < \infty$. Let $Y_1, Y_2, \ldots$ be independent random variables such that $\|Y_n\|_\infty \leq c_n$ and $\mathbf{E}Y_n = 0$ for all $n \geq 1$, and let $X_n := Y_1 + \cdots + Y_n$ for all $n \geq 1$ with $X_0 := 0$. Show that $X_0, X_1, \ldots$ is a martingale that converges in every L_p space with $1 \leq p < \infty$. That is, show there exists a real-valued random variable X such that $\lim_{n\to\infty} \|X_n - X\|_p = 0$.

We begin with an almost sure convergence result that makes an additional moment assumption on the supermartingale (or submartingale).

Theorem 6.37 (Doob's Convergence Theorem). *Let $X_0, X_1, \ldots$ be a supermartingale such that $\sup_{n\geq 0} \mathbf{E}\max(-X_n, 0) < \infty$. Then there exists a random variable X such that $X_0, X_1, \ldots$ converges almost surely to X and $\mathbf{E}|X| \leq \liminf_{n\to\infty} \mathbf{E}|X_n| < \infty$.*

Proof. Let $a, b \in \mathbb{R}$ with $a < b$. Then $U_1[a,b], U_2[a,b], \ldots$ monotonically increases to a random variable denoted $U[a,b]$. By Monotone Convergence, Theorem 1.54, $\mathbf{E}U[a,b] = \sup_{n\geq 0} \mathbf{E}U_n[a,b]$. Using Lemma 6.34 and the inequality $\max(a - x, 0) \leq |a| + \max(-x, 0)$ valid for any $x \in \mathbb{R}$,

$$\mathbf{E}U_n[a,b] \leq \frac{1}{b-a}\mathbf{E}\max(a - X_n, 0) \leq \frac{1}{b-a}(|a| + \sup_{m\geq 0} \mathbf{E}\max(-X_m, 0)).$$

So, by assumption, $\sup_{n\geq 0} \mathbf{E}U_n[a,b] < \infty$ so that $\mathbf{E}U[a,b] < \infty$. In particular, $U[a,b]$ is finite almost surely. The event that $X_0, X_1, \ldots$ converges almost surely in $[-\infty, \infty]$ is the complement of

$$\bigcup_{a,b\in\mathbb{Q}:\, a<b} \{\liminf_{n\to\infty} X_n < a < b < \limsup_{n\to\infty} X_n\}.$$

We therefore show this event has probability zero. Since the union is countable, it suffices to show that each such event has probability zero. Let $a, b \in \mathbb{Q}$ with $a < b$. Then $U[a,b]$ is finite almost surely, so indeed $\mathbf{P}(\liminf_{n\to\infty} X_n < a < b < \limsup_{n\to\infty} X_n) = 0$ by Exercise 6.33. So, $X_0, X_1, \ldots$ converges almost surely to some X.

Using the equality $|x| = x + 2\max(-x, 0)$ valid for all $x \in \mathbb{R}$ and the supermartingale property, $\mathbf{E}|X_n| = \mathbf{E}X_n + 2\mathbf{E}\max(-X_n, 0) \leq \mathbf{E}X_0 + 2\mathbf{E}\max(-X_n, 0)$. Since $|X_0|, |X_1|, \ldots$ converges almost surely to $|X|$, we have by Fatou's Lemma, Theorem 1.56,

$$\mathbf{E}|X| \leq \liminf_{n\to\infty} \mathbf{E}|X_n| \leq \mathbf{E}|X_0| + 2\sup_{m\geq 0} \mathbf{E}\max(-X_m, 0) < \infty.$$

□

If we make an additional bounded moment assumption on the martingale, then almost sure and L_1 convergence of the martingale follows from Corollary 6.29 and Theorem 6.37.

Proposition 6.38 (L_p **Convergence**). *Let $X_0, X_1, \ldots$ be a martingale. Let $p > 1$. Assume $\sup_{n\geq 0} \mathbf{E}\,|X_n|^p < \infty$. Then $X_0, X_1, \ldots$ converges almost surely and in L_p. In particular, $X_0, X_1, \ldots$ converges in L_1.*

Proof. Let $n \geq 0$. Then $(\mathbf{E}\max(X_n, 0))^p \leq (\mathbf{E}\,|X_n|)^p \leq \mathbf{E}\,|X_n|^p$ by Jensen's Inequality, so $\sup_{n\geq 0} \mathbf{E}\max(X_n, 0) < \infty$ by our assumption. Then Doob's Convergence Theorem, Theorem 6.37 implies that $X_0, X_1, \ldots$ converges almost surely to a random variable X with $\mathbf{E}\,|X| < \infty$.

From the L_p Maximal Inequality, Corollary 6.29, and Monotone Convergence, Theorem 1.54, $\|\sup_{n\geq 0} |X_n|\,\|_p < \infty$. Since $|X| \leq \sup_{n\geq 0} |X_n|$, we have $\|X\|_p < \infty$. From the triangle inequality, for any $n \geq 0$, $|X_n - X|^p \leq (2\sup_{m\geq 0} X_m)^p$. So, Dominated Convergence, Theorem 1.57, implies that $\lim_{n\to\infty} \|X_n - X\|_p = 0$. □

As we will show further below, the following condition is necessary and sufficient for the convergence of a martingale in L_1.

Definition 6.39 (**Uniform Integrability**). Let H be a collection of random variables on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. We say that H is **uniformly integrable** if

$$\lim_{m\to\infty} \sup_{X\in H} \mathbf{E}\,|X|\,1_{|X|>m} = 0.$$

That is, $\forall\ \varepsilon > 0$, $\exists\ n = n(\varepsilon) > 0$ such that, $\forall\ m \geq n$, $\sup_{X\in H} \mathbf{E}\,|X|\,1_{|X|>m} < \varepsilon$.

In Exercise 6.43 below, it is shown that the assumption of Proposition 6.38 implies that the martingale $\{X_0, X_1, \ldots\}$ is uniformly integrable. So, the conclusion of L_1 convergence in Proposition 6.38 will be subsumed by the main theorem of the section, Theorem 6.47 below.

The following exercise shows that uniform integrability is a relaxed assumption for the Dominated Convergence Theorem.

Exercise 6.40. Let H be a collection of random variables on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Let $Y\colon \Omega \to [0, \infty)$ with $\mathbf{E}Y < \infty$. Assume that, for all $X \in H$, $|X| \leq Y$. Show that H is uniformly integrable. In particular, if H is any finite set of random variables in L_1, then H is uniformly integrable.

Exercise 6.41. Let H be a collection of random variables on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Show that H is uniformly integrable if and only if the following two conditions hold.

(a) $\sup_{X\in H} \mathbf{E}\,|X| < \infty$.
(b) For any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\sup\{\mathbf{E}\,|X|\,1_A\colon A \in \mathcal{F},\ \mathbf{P}(A) < \delta,\ X \in H\} < \varepsilon.$$

(Hint: when $X \in H$ is fixed, which A with $\mathbf{P}(A) < \delta$ maximizes $\mathbf{E}\,|X|\,1_A$? Also, to show the first item, let $\varepsilon = 1/2$ in the definition of uniform integrability.)

Exercise 6.42. Let H be a collection of random variables on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. The analytic definition of uniform integrability is just the second item of the above exercise. That is, H is uniformly integrable in the analytic sense if and only if condition (b) holds in Exercise 6.41. In the case that $\mathbf{P}$ is non-atomic, show that if condition (b) holds in Exercise 6.41, then condition (a) holds. In summary, if $\mathbf{P}$ is non-atomic, then the probabilistic and analytic definitions of uniform integrability coincide. (We say that $\mathbf{P}$ is non-atomic if for any $A \in \mathcal{F}$ with $\mathbf{P}(A) > 0$ there exists $B \in \mathcal{F}$ with $B \subseteq A$ such that $\mathbf{P}(A) > \mathbf{P}(B) > 0$. An atom for $\mathbf{P}$ is a set $A \in \mathcal{F}$ such that, for any $B \in \mathcal{F}$ with $B \subseteq A$, if $\mathbf{P}(B) < \mathbf{P}(A)$, then

$\mathbf{P}(B) = 0$. In a non-atomic probability space, the following holds and you do not have to prove it: for any $A \in \mathcal{F}$ with $\mathbf{P}(A) > 0$, and for any $t \in \mathbb{R}$ such that $0 < t < \mathbf{P}(A)$, there exists $B \in \mathcal{F}$ with $B \subseteq A$ and $\mathbf{P}(B) = t$.)

Exercise 6.43. Show that condition (a) of Exercise 6.41 is not sufficient to prove uniform integrability. In fact, the unit ball $\{X \in L_1 \colon \|X\|_1 \leq 1\}$ of L_1 is not uniformly integrable, in general. More specifically, find a set of random variables $X_1, X_2, \ldots$ on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ with $\|X_n\|_1 \leq 1$ for all $n \geq 1$, but such that the collection $H := \{X_1, X_2, \ldots\}$ is not uniformly integrable. (Hint: let $\Omega := \{1, 2, 3, \ldots\}$ with the probability measure $\mathbf{P}$ defined by $\mathbf{P}(\{n\}) := 2^{-n}$ for all $n \geq 1$, and choose the X_n to have disjoint supports.)

Now, let $p > 1$. Show that the unit ball $\{X \in L_p \colon \|X\|_p \leq 1\}$ of L_p is uniformly integrable.

Theorem 6.44 (Vitali Convergence Theorem). *Let $X_0, X_1, \ldots$ be real-valued random variables that converge in probability to a random variable X. Then the following are equivalent.*

(i) *The collection of random variables $\{X_0, X_1, \ldots\}$ is uniformly integrable.*
(ii) *$X_0, X_1, \ldots \in L_1$ converges in L_1 to $X \in L_1$.*
(iii) *$X_0, X_1, \ldots \in L_1$ and $\lim_{n\to\infty} \|X_n\|_1 = \|X\|_1 < \infty$.*

Proof. We first show that (i) implies (ii). Fix $m > 0$ and define $\phi_m \colon \mathbb{R} \to \mathbb{R}$ by

$$\phi_m(x) := \begin{cases} -m & \text{, if } x < -m \\ x & \text{, if } -m \leq x \leq m \\ m & \text{, if } x > m. \end{cases}$$

Applying the triangle inequality and $|\phi_m(Y) - Y| = \max(|Y| - m, 0) \leq |Y| \, 1_{|Y|>m}$ for the random variables $Y := X_n$ and $Y := X$,

$$\begin{aligned} \mathbf{E}\,|X_n - X| &\leq \mathbf{E}\,|X_n - \phi_m(X_n)| + \mathbf{E}\,|\phi_m(X_n) - \phi_m(X)| + \mathbf{E}\,|\phi_m(X) - X| \\ &\leq \mathbf{E}\,|X_n| \, 1_{|X_n|>m} + \mathbf{E}\,|\phi_m(X_n) - \phi_m(X)| + \mathbf{E}\,|X| \, 1_{|X|>m}. \end{aligned}$$

Let $\varepsilon > 0$. We can choose $m > 0$ such that $\sup_{n\geq 0} \mathbf{E}\,|X_n| \, 1_{|X_n|>m} < \varepsilon/2$ by assumption (i). By Exercise 6.41(a), $\sup_{n\geq 0} \mathbf{E}\,|X_n| < \infty$, so Fatou's Lemma (Exercise 2.10(v)) implies $\mathbf{E}\,|X| < \infty$, so the third term can be made less than $\varepsilon/2$ by choosing m larger if necessary, by Dominated Convergence, Theorem 1.57. Since ϕ_m is continuous, $\phi_m(X_0), \phi_m(X_1), \ldots$ converges in probability to $\phi_m(X)$ by Exercise 2.10(iv). So, $\forall\ m \geq 1$, $\lim_{n\to\infty} \mathbf{E}\,|\phi_m(X_n) - \phi_m(X)| = 0$ by Exercise 2.10(vi). In summary, $\forall\ \varepsilon > 0$, $\exists\ m > 0$ such that $\limsup_{n\to\infty} \mathbf{E}\,|X_n - X| < \varepsilon$, implying that $\lim_{n\to\infty} \mathbf{E}\,|X_n - X| = 0$.

We now show (ii) implies (iii). This follows by the L_1 triangle inequality (or Jensen's inequality) and the reverse triangle inequality for scalars ($||x| - |y|| \leq |x - y|\ \forall\ x, y \in \mathbb{R}$):

$$|\mathbf{E}(|X_n| - |X|)| \leq \mathbf{E}\,||X_n| - |X|| \leq \mathbf{E}\,|X_n - X|\,.$$

Letting $n \to \infty$ concludes the implication.

We now show (iii) implies (i). Fix $m > 0$ and define $\psi_m \colon [0, \infty) \to \mathbb{R}$ so that $\psi_m(x) := x$ for any $x \in [0, m-1]$, $\psi_m(x) := 0$ for any $x \geq m$, and $\psi_m(x) = (m-1)(m-x)$ when $x \in [m-1, m]$. Let $\varepsilon > 0$. Since $\mathbf{E}\,|X| < \infty$, observe that $\lim_{m\to\infty}(\mathbf{E}\,|X| - \mathbf{E}\psi_m(|X|)) = 0$ by the Dominated Convergence Theorem, Exercise 2.10(vi). So, $\exists\ m > 0$ such that $\mathbf{E}\,|X| - \mathbf{E}\psi_m(|X|) < \varepsilon/3$. Since ψ_m is continuous, $\psi_m(X_0), \psi_m(X_1), \ldots$ converges in probability to $\psi_m(X)$ by Exercise 2.10(iv). And since ψ_m is a bounded function,

$\lim_{n\to\infty} \mathbf{E}\psi_m(|X_n|) = \mathbf{E}\psi_m(|X|)$ by Exercise 2.10(vi). So, $\exists\ r > 0$ such that for all $n > r$, $|\mathbf{E}\psi_m(|X_n|) - \mathbf{E}\psi_m(|X|)| < \varepsilon/3$ and $|\mathbf{E}\,|X_n| - \mathbf{E}\,|X|| < \varepsilon/3$ by (iii). So, by definition of ψ_m,

$$\mathbf{E}\,|X_n|\,1_{|X_n|>m} \le \mathbf{E}\,|X_n| - \mathbf{E}\psi_m(|X_n|) \le \mathbf{E}\,|X| - \mathbf{E}\psi_m(|X|) + 2\varepsilon/3 \le \varepsilon.$$

This holds for fixed $m > 0$ and for all $n > r$, where r can depend on m. Since $\mathbf{E}\,|X_n| < \infty$ for all $0 \le n \le r$, there exists $m' > 0$ such that $\mathbf{E}\,|X_n|\,1_{|X_n|>m'} < \varepsilon$ for all $0 \le n \le r$. So $\mathbf{E}\,|X_n|\,1_{|X_n|>\max(m,m')} < \varepsilon$ for all $n \ge 0$, proving (i). □

Theorem 6.45 (**Submartingale Convergence**). *Let $X_0, X_1, \ldots$ be a submartingale. Then the following are equivalent.*

(i) *The collection of random variables $\{X_0, X_1, \ldots\}$ is uniformly integrable.*
(ii) *$X_0, X_1, \ldots$ converges almost surely and in L_1.*
(iii) *$X_0, X_1, \ldots$ converges in L_1.*

Proof. We first show (i) implies (ii). If $n \ge 0$, $\mathbf{E}\max(X_n, 0) \le \mathbf{E}\,|X_n| \le \sup_{m\ge 0} \mathbf{E}\,|X_m| < \infty$ by Exercise 6.41, so $X_0, X_1, \ldots$ converges almost surely to some X with $\mathbf{E}\,|X| < \infty$ by Doob's Convergence Theorem, Theorem 6.37. Exercise 2.5 and Vitali's Convergence Theorem, Theorem 6.44 then imply $X_0, X_1, \ldots$ converges in L_1 to X. The implication (ii) implies (iii) is clear. We now show (iii) implies (i). From Exercise 2.6, $X_0, X_1, \ldots$ converges in probability to X, so Vitali's Convergence Theorem, Theorem 6.44, implies (i). □

Lemma 6.46. *Let $((X_n)_{n\ge 0}, (\mathcal{F}_n)_{n\ge 0})$ be a martingale. Let X be a real-valued random variable with $\mathbf{E}\,|X| < \infty$. Assume that $X_0, X_1, \ldots$ converges to X in L_1. Then $X_n = \mathbf{E}(X|\mathcal{F}_n)$ for all $n \ge 0$.*

Proof. Let $m > n$. Since $((X_n)_{n\ge 0}, (\mathcal{F}_n)_{n\ge 0})$ is a martingale, $\mathbf{E}(X_m|\mathcal{F}_n) = X_n$, so if $A \in \mathcal{F}_n$, $\mathbf{E}X_m 1_A = \mathbf{E}X_n 1_A$. From the triangle inequality, $|\mathbf{E}X_m 1_A - \mathbf{E}X 1_A| \le \mathbf{E}\,|X_m - X|$, so $\lim_{m\to\infty} \mathbf{E}X_m 1_A = \mathbf{E}X 1_A$. That is, $\mathbf{E}X 1_A = \mathbf{E}X_n 1_A$ for any $A \in \mathcal{F}_n$. That is, $X_n = \mathbf{E}(X|\mathcal{F}_n)$ for all $n \ge 0$, by the definition of conditional expectation, Definition 5.3. □

Below we can finally answer the question posed in Exercise 6.11.

Theorem 6.47 (**Martingale Convergence**). *Let $X_0, X_1, \ldots$ be a martingale. Then the following are equivalent.*

(i) *The collection of random variables $\{X_0, X_1, \ldots\}$ is uniformly integrable.*
(ii) *$X_0, X_1, \ldots$ converges almost surely and in L_1.*
(iii) *$X_0, X_1, \ldots$ converges in L_1.*
(iv) *$\exists$ a real-valued random variable X with $X_n = \mathbf{E}(X|\mathcal{F}_n)$ for any $n \ge 0$ and $\mathbf{E}\,|X| < \infty$.*

Proof. Since martingales are submartingales, Theorem 6.45 shows that (i),(ii) and (iii) are equivalent. By Lemma 6.46, (iii) implies (iv). It remains to show that (iv) implies (i). Let $n \ge 0$. From Jensen's inequality, Exercise 5.15 applied twice, and the definition of conditional expectation (using $\{\mathbf{E}(|X|\,|\mathcal{F}_n) > m\} \in \mathcal{F}_n$), if $m > 0$,

$$\begin{aligned}\mathbf{E}\,|\mathbf{E}(X|\mathcal{F}_n)|\,1_{|\mathbf{E}(X|\mathcal{F}_n)|>m} &\le \mathbf{E}\mathbf{E}(|X|\,|\mathcal{F}_n)1_{|\mathbf{E}(X|\mathcal{F}_n)|>m} \\ &\le \mathbf{E}\mathbf{E}(|X|\,|\mathcal{F}_n)1_{\mathbf{E}(|X||\mathcal{F}_n)>m} = \mathbf{E}\,|X|\,1_{\mathbf{E}(|X||\mathcal{F}_n)>m}. \qquad (*)\end{aligned}$$

From Markov's inequality and Proposition 5.12,

$$\mathbf{P}(\mathbf{E}(|X|\,|\mathcal{F}_n) > m) \le \frac{1}{m}\mathbf{E}\mathbf{E}(|X|\,|\mathcal{F}_n) = \frac{1}{m}\mathbf{E}\,|X|\,.$$

Therefore, $\lim_{m\to\infty}\sup_{n\geq 0}\mathbf{E}\,|X|\,1_{\mathbf{E}(|X||\mathcal{F}_n)>m}=0$ by the Dominated Convergence Theorem, Theorem 1.57. So, uniformly integrability of $\{X_0,X_1,\ldots\}$ follows from $(*)$. □

Exercise 6.48 (Galton-Watson Process). Let $(\xi_{i,n})_{i,n\geq 1}$ be i.i.d. nonnegative integer-valued random variables. Let $Z_0:=1$ and for any $n\geq 0$ define

$$Z_{n+1}:=\begin{cases}\xi_{1,n+1}+\cdots+\xi_{Z_n,n+1} & \text{, if } Z_n>0\\ 0 & \text{, if } Z_n=0.\end{cases}$$

Then $Z_0,Z_1,\ldots$ is an example of a branching process, known as the Galton-Watson process. The intuition is that Z_n is the number of individuals in the n^{th} generation of a family tree, and at each time step, each person has a certain number of offspring. Galton and Watson originally used this process to model the occurrence of last names in human family trees, to see why some names become common while others become extinct.

$\forall\, n\geq 0$, let $\mathcal{F}_n:=\sigma(\xi_{i,m}\colon i\geq 1, 1\leq m\leq n)$, and let $\mu:=\mathbf{E}\xi_{1,1}$. Assume $\mu\in(0,\infty)$.

- Show that $Z_0, Z_1/\mu, Z_2/\mu^2,\ldots$ is a martingale with respect to $\mathcal{F}_0,\mathcal{F}_1,\ldots$. (Hint: write $\mathbf{E}(Z_{n+1}|\mathcal{F}_n)=\sum_{k=1}^{\infty}\mathbf{E}(Z_{n+1}1_{Z_n=k}|\mathcal{F}_n)$.)
- If $\mu<1$, show that $\mathbf{P}(Z_n>0 \text{ for infinitely many } n\geq 0)=0$. Therefore, Z_n/μ^n converges to 0 almost surely as $n\to\infty$. Also, show that the expected total population $\mathbf{E}\sum_{n=0}^{\infty}Z_n$ is finite. That is, extinction occurs as $n\to\infty$ if the average number of offspring is less than 1 for each individual.
- If $\mu=1$, and $\mathbf{P}(\xi_{1,1}=1)<1$, show that $Z_0,Z_1,\ldots$ converges almost surely to 0.

Lemma 6.49. *Let $((X_n)_{n\geq 0},(\mathcal{F}_n)_{n\geq 0})$ be a submartingale and let T be a stopping time. Assume $\sup_{n\geq 0}\mathbf{E}\max(X_n,0)<\infty$. Then $\mathbf{E}\,|X_T|<\infty$. (It is shown in the proof that $X_\infty:=\lim_{n\to\infty}X_n$ exists almost surely.)*

Proof. By Exercise 6.17(iv), $\max(X_0,0),\max(X_1,0),\ldots$ is a submartingale. So, by Theorem 6.26 with stopping times $T\leq N:=\infty$, $\mathbf{E}\max(X_{n\wedge T},0)\leq\mathbf{E}\max(X_n,0)$ for all $n\geq 0$. So, by our assumption, $\sup_{n\geq 0}\mathbf{E}\max(X_{n\wedge T},0)<\infty$. Applying Doob's Convergence Theorem, Theorem 6.37, to the submartingale $X_{0\wedge T},X_{1\wedge T},\ldots$ (by Theorem 6.26), we have almost surely $\lim_{n\to\infty}X_{n\wedge T}=X_T$ and $\mathbf{E}\,|X_T|<\infty$ □

Theorem 6.50 (Optional Stopping Theorem, Version 2). *Let $((X_n)_{n\geq 0},(\mathcal{F}_n)_{n\geq 0})$ be a supermartingale with $X_n\geq 0$ for all $n\geq 0$. Let $T\leq N$ be stopping times. Then*

$$\infty>\mathbf{E}X_T\geq\mathbf{E}X_N.$$

(It is shown in the proof that $X_\infty:=\lim_{n\to\infty}X_n$ exists almost surely.)

Proof. From Theorem 6.26, $0=X_{0\wedge N}-X_{0\wedge T},X_{1\wedge N}-X_{1\wedge T},\ldots$ is a supermartingale, so $\infty>\mathbf{E}X_{n\wedge T}\geq\mathbf{E}X_{n\wedge N}$ for any $n\geq 0$. Since $T\leq N$, $1_{N\geq n}-1_{T\geq n}=1_{N\geq n>T}$ and subtracting $\mathbf{E}X_n1_{T\geq n}$ from both sides,

$$\mathbf{E}X_T1_{T<n}\geq\mathbf{E}X_{n\wedge N}(1_{N<n}+1_{N\geq n})-\mathbf{E}X_n1_{T\geq n}=\mathbf{E}X_N1_{N<n}+\mathbf{E}X_n1_{N\geq n>T}.\qquad(*)$$

Since $X_0,X_1,\ldots$ is a nonnegative supermartingale, Doob's Convergence Theorem, Theorem 6.37 says there exists a random variable X with $\mathbf{E}\,|X|<\infty$ such that $X_0,X_1,\ldots$ converges almost surely to X. By Fatou's Lemma, Theorem 1.56,

$$\liminf_{n\to\infty}\mathbf{E}X_n1_{N\geq n>T}=\liminf_{n\to\infty}\mathbf{E}X_n1_{N\geq n}1_{n>T}\geq\mathbf{E}X1_{N=\infty>T}$$

By Monotone Convergence, Theorem 1.54, $\lim_{n\to\infty} \mathbf{E}X_N 1_{N<n} = \mathbf{E}X_N 1_{N<\infty}$ and similarly $\lim_{n\to\infty} \mathbf{E}X_T 1_{T<n} = \mathbf{E}X_T 1_{T<\infty}$, so $(*)$ implies

$$\mathbf{E}X_T 1_{T<\infty} \geq \mathbf{E}X_N 1_{N<\infty} + \mathbf{E}X 1_{N=\infty>T}.$$

Adding the equality $\mathbf{E}X_T 1_{T=\infty} = \mathbf{E}X_N 1_{N=T=\infty}$, which holds since $T \leq N$, we get $\mathbf{E}X_T \geq \mathbf{E}X_N$. Choosing $T' := 0 \leq T$ in this inequality shows $\infty > \mathbf{E}X_0 \geq \mathbf{E}X_T \geq \mathbf{E}X_N$. □

6.4. **Optional Stopping Theorems.**

Theorem 6.51 (Doob's Optional Stopping Theorem). *Let $((V_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ and $((Y_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ be submartingales with $V_n \leq 0$ for all $n \geq 0$. Let $T \leq N$ be stopping times. Assume that $\{Y_{0\wedge N}, Y_{1\wedge N}, \ldots\}$ is uniformly integrable. Define*

$$X_n := Y_n + V_n, \qquad \forall\, n \geq 0.$$

Also define $X_N 1_{N=\infty} := 1_{N=\infty} \limsup_{n\to\infty} X_n$. Then $\mathbf{E}\,|X_N|, \mathbf{E}\,|X_T| < \infty$, and

$$\mathbf{E}X_N \geq \mathbf{E}X_T \geq \mathbf{E}X_0.$$

Remark 6.52. Setting $V_n := 0$ for all $n \geq 0$, we see that the conclusion of Theorem 6.51 holds for any submartingale $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ such that $\{X_{0\wedge N}, X_{1\wedge N}, \ldots\}$ is uniformly integrable.

More specifically, Theorem 6.51 is commonly applied when $((X_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ is a martingale and $\{X_{0\wedge N}, X_{1\wedge N}, \ldots\}$ is uniformly integrable, giving the conclusion

$$\mathbf{E}X_N = \mathbf{E}X_0.$$

Conditions for $\{X_{0\wedge N}, X_{1\wedge N}, \ldots\}$ being uniformly integrable are given in Proposition 6.53.

Proof. By linearity of expected value, we deal with the Y_n and V_n terms separately. Since $((-V_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ is a nonnegative supermartingale, Theorem 6.50 implies that $\mathbf{E}V_N \geq \mathbf{E}V_T \geq \mathbf{E}V_0 > -\infty$.

It remains to consider the Y_n terms. For any $n \geq 0$, define $U_n := Y_{n\wedge N}$, $Z_n := Y_{n\wedge T}$. From Theorem 6.26 and $T \leq N$, the following three sequences are submartingales with respect to $(\mathcal{F}_n)_{n\geq 0}$: $U_0, U_1, \ldots$, $Z_0, Z_1, \ldots$ and $U_0 - Z_0, U_1 - Z_1, \ldots$. So,

$$\mathbf{E}U_n \geq \mathbf{E}Z_n \geq \mathbf{E}Z_0, \qquad \forall\, n \geq 0 \qquad (*)$$

By assumption, $\{U_0, U_1, \ldots\}$ is uniformly integrable, so there exists a random variable U with $\mathbf{E}\,|U| < \infty$ such that $U_0, U_1, \ldots$ converges almost surely and in L_1 to U, by Theorem 6.45. Since $T \leq N$, $Z_n = U_{n\wedge T}$ for all $n \geq 0$. So, since $\{U_0, U_1, \ldots\}$ is uniformly integrable, $\{Z_0, Z_1, \ldots\}$ is uniformly integrable by Proposition 6.53(iii). By Theorem 6.45 again, there exists a random variable Z with $\mathbf{E}\,|Z| < \infty$ such that $Z_0, Z_1, \ldots$ converges almost surely and in L_1 to Z. Then, almost surely, $Z = \lim_{n\to\infty} Y_{n\wedge T} = Y_T$ and $U = \lim_{n\to\infty} Y_{n\wedge N} = Y_N$, so by $(*)$, and their L_1 convergence, $\mathbf{E}Y_N \geq \mathbf{E}Y_T \geq \mathbf{E}Z_0 = \mathbf{E}Y_0$. □

Proposition 6.53. *Let $Y_0, Y_1, \ldots$ be a sequence of real-valued random variables with $\mathbf{E}\,|Y_n| < \infty$ for all $n \geq 0$. Let T be a stopping time adapted to a filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$. Then $\{Y_{0\wedge T}, Y_{1\wedge T}, \ldots\}$ is uniformly integrable if any one of the following conditions holds.*

(i) $\mathbf{E}T < \infty$, *and* $\exists\ c > 0$ *such that, almost surely,* $\mathbf{E}(|Y_n - Y_{n-1}|\ |\mathcal{F}_{n-1}) \leq c$, $\forall\ n \geq 1$.

(ii) $\{Y_0 1_{T>0}, Y_1 1_{T>1}, \ldots\}$ *is uniformly integrable and* $\mathbf{E}\,|Y_T 1_{T<\infty}| < \infty$.

(iii) $((Y_n)_{n\geq 0}, (\mathcal{F}_n)_{n\geq 0})$ *is a uniformly integrable submartingale (or supermartingale).*

Proof. We first prove (i). From the triangle inequality, for any $n \geq 0$,

$$|Y_{n\wedge T}| \leq Z_n := |Y_0| + \sum_{m=1}^{n\wedge T} |Y_m - Y_{m-1}| = |Y_0| + \sum_{m=1}^{n} |Y_m - Y_{m-1}|\, 1_{T\geq m}.$$

Since $Z_0 \leq Z_1 \leq \cdots$, we have $\sup_{n\geq 0} |Y_{n\wedge T}| \leq \lim_{n\to\infty} Z_n =: Z$. So, if $\mathbf{E}Z < \infty$, we are done by Exercise 6.40. We therefore show $\mathbf{E}Z < \infty$. Since T is a stopping time, $1_{T\geq m} = 1 - 1_{T<m}$ is $\mathcal{F}_{m-1}$-measurable. So, from Propositions 5.12 and 5.16, for any $m \geq 1$,

$$\mathbf{E}(|Y_m - Y_{m-1}| \cdot 1_{T\geq m}) = \mathbf{E}\mathbf{E}(|Y_m - Y_{m-1}|\ |\mathcal{F}_{m-1}) 1_{T\geq m} \leq c\mathbf{P}(T \geq m).$$

Using this inequality, Monotone Convergence, Theorem 1.54, and Exercise 1.88,

$$\mathbf{E}Z \leq \mathbf{E}\,|Y_0| + c\sum_{m=1}^{\infty} \mathbf{P}(T \geq m) = \mathbf{E}\,|Y_0| + c\mathbf{E}T < \infty.$$

We now prove (ii). For any sequence of random variables $X_0, X_1, \ldots$, and for any $n \geq 0$, $|X_{n\wedge T}| \leq |X_T|\, 1_{T<\infty} + |X_n|\, 1_{T>n}$. (As above, define $X_T 1_{T=\infty} := 1_{T=\infty} \limsup_{n\to\infty} X_n$.) By (ii), $\{|Y_0|\, 1_{T>0}, |Y_1|\, 1_{T>1}, \ldots\}$ is uniformly integrable. Let $m > 0$. Using $X_n := Y_n 1_{|Y_n|>m}$ for all $n \geq 0$,

$$\sup_{n\geq 0} \mathbf{E}\,|Y_{n\wedge T}|\, 1_{|Y_{n\wedge T}|>m} \leq \mathbf{E}\,|Y_T|\, 1_{|Y_T|>m} 1_{T<\infty} + \sup_{n\geq 0} \mathbf{E}\,|Y_n|\, 1_{|Y_n|>m} 1_{T>n}.$$

Since $\mathbf{E}\,|Y_n|\, 1_{|Y_n|>m} 1_{T>n} = \mathbf{E}\,|Y_n|\, 1_{T>n} 1_{\{|Y_n| 1_{T>n} > m\}}$, the uniform integrability assumption (ii) and $\mathbf{E}\,|Y_T 1_{T<\infty}| < \infty$ imply that the right side converges to 0 as $m \to \infty$, as desired.

We now prove (iii). By assumption, $\{Y_0 1_{T>0}, Y_1 1_{T>1}, \ldots\}$ is uniformly integrable and $\sup_{n\geq 0} \mathbf{E} \max(Y_n, 0) < \infty$ by Exercise 6.41. Lemma 6.49 then implies that $\mathbf{E}\,|Y_T|\, 1_{T<\infty} < \infty$, so that (iii) reduces to (ii). □

Exercise 6.54. Explain why Example 6.24 does not contradict Doob's Optional Stopping Theorem.

Exercise 6.55 (Gambler's Ruin). We can now finally answer the question posed in Example 6.8. Let $0 < p < 1$. Let $0 \leq a < y_0 < b$ with $a, y_0, b \in \mathbb{Z}$. Let $Y_1, Y_2, \ldots$ be independent random variables such that $\mathbf{P}(Y_n = 1) =: p$ and $\mathbf{P}(Y_n = -1) = 1 - p =: q\ \forall\ n \geq 1$. Let $Y_0 := y_0$. Let $Z_n = Y_0 + \cdots + Y_n$, and let $X_n := (q/p)^{Z_n}\ \forall\ n \geq 0$. For any $n \geq 0$, let $\mathcal{F}_n := \sigma(Y_0, \ldots, Y_n)$. We showed in Example 6.8 that $X_0, X_1, \ldots$ is a martingale with respect to $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$. Let $T = \min\{n \geq 1 \colon Z_n \in \{a, b\}\}$. That is, T is the first time the random walk $Z_0, Z_1, \ldots$ hits either a or b.

- Compute $c := \mathbf{P}(Y_T = a)$, using Doob's Optional Stopping Theorem, when $p \neq 1/2$.
- Compute $\mathbf{E}T$ using Doob's Optional Stopping Theorem, when $p \neq 1/2$. (Hint: $Z_0 - 0(2p-1), Z_1 - (2p-1), \ldots$ is a martingale.)
- Compute $\mathbf{E}\min\{n \geq 1 \colon Z_n = a\}$ when $p < 1/2$. (Hint: let $b \to \infty$.)
- Compute c when $p = 1/2$ using the martingale $Z_0, Z_1, \ldots$.
- Compute $\mathbf{E}T$ when $p = 1/2$. (Hint: if $y_0 = 0$, then $Z_0^2 - 0, Z_1^2 - 1, \ldots$ is a martingale.)

Exercise 6.56. Let $Y_1, Y_2, \ldots$ be independent random variables such that $\mathbf{P}(Y_n = 1) = \mathbf{P}(Y_n = -1) = 1/2\ \forall\ n \geq 1$. Let $Y_0 := 0$. Let $Z_n := Y_0 + \cdots + Y_n$ for any $n \geq 0$. From the previous exercise, one might wonder where the martingale $Z_0^2 - 0, Z_1^2 - 1, \ldots$ came from, and if more like it exist. In this exercise, we compute an infinite family of such martingales.

For any $\alpha \in \mathbb{R}$ and $n \geq 0$, let $X_n := e^{\alpha Z_n - n \log\cosh(\alpha)}$. Show that $X_0, X_1, \ldots$ is a martingale.

Then, using the power series expansion of the exponential function, we have $X_n = \sum_{m=0}^{\infty} \frac{\alpha^m}{m!} M_{m,n}$ for some random variables $M_{1,1}, \ldots$, for any $\alpha \in \mathbb{R}$ and for any $n \geq 0$. It follows that, for any $m \geq 0$, $M_{m,0}, M_{m,1}, \ldots$ is a martingale. For example, using $m = 2$ we get $M_{2,n} = Z_n^2 - n$ for all $n \geq 0$. And using $m = 4$, $M_{4,n} = Z_n^4 - 6nZ_n^2 + 2n + 3n^2$ for all $n \geq 0$. Using this martingale, compute $\mathbf{E}T^2$ when $T := \min\{n \geq 1 \colon Z_n \in \{-b, b\}\}$ and $b > 0$, $b \in \mathbb{Z}$.

Exercise 6.57. Let $0 < p < 1$. Let b be a positive integer. Let $Y_1, Y_2, \ldots$ be independent random variables such that $\mathbf{P}(Y_n = 1) =: p$ and $\mathbf{P}(Y_n = -1) = 1 - p =: q$ $\forall\ n \geq 1$. Let $Y_0 := 0$. Let $Z_n = Y_0 + \cdots + Y_n$, $\forall\ n \geq 0$. Let $T_b := \min\{n \geq 1 \colon Z_n = b\}$. For any $\alpha \in \mathbb{R}$ let $M(\alpha) := \mathbf{E}e^{\alpha Y_1}$. For any $n \geq 0$, let $X_n := e^{\alpha Z_n}(M(\alpha))^{-n}$.

- If $1/2 \leq p < 1$, show that $e^{\alpha b}\mathbf{E}M(\alpha)^{-T_b} = 1$ for all $\alpha > 0$.
- If $1/2 \leq p < 1$ and $0 < s < 1$, show that
$$\mathbf{E}s^{T_1} = \frac{1}{2qs}(1 - \sqrt{1 - 4pqs^2}), \qquad \mathbf{E}s^{T_b} = (\mathbf{E}s^{T_1})^b.$$
- If $0 < p < 1/2$, show that $\mathbf{P}(T_b < \infty) = e^{-\lambda b}$ where $\lambda := \log((1-p)/p) > 0$.
- If $0 < p < 1/2$, show that $Z := 1 + \max_{n \geq 0} Z_n$ is a geometric random variable with success probability $1 - e^{-\lambda}$.

Exercise 6.58 (**Ballot Theorem**)**.** Let a, b be positive integers. Suppose there are c votes cast by c people in an election. Candidate 1 gets a votes and candidate 2 gets b votes. (So $c = a + b$.) Assume $a > b$. The votes are counted one by one. The votes are counted in a uniformly random ordering, and we would like to keep a running tally of who is currently winning. (News agencies seem to enjoy reporting about this number.) Suppose the first candidate eventually wins the election. We ask: with what probability will candidate 1 always be ahead in the running tally of who is currently winning the election? As we will see, the answer is $\frac{a-b}{a+b}$.

To prove this, for any positive integer k, let S_k be the number of votes for candidate 1, minus the number of votes for candidate 2, after k votes have been counted. Then, define $X_k := S_{c-k}/(c-k)$. Show that $X_0, X_1, \ldots$ is a martingale. Then, let T such that $T = \min\{0 \leq k \leq c \colon X_k = 0\}$, or $T = c - 1$ if no such k exists. Apply the Optional Stopping theorem to X_T to deduce the result.

Exercise 6.59. Prove Wald's Equation, Proposition 4.21, using Doob's Optional Stopping Theorem.

6.5. **Additional Comments.** A "martingale" originally referred to the double-your-bet strategy for betting on fair coin flips, discussed in France in the late 1700s. The term "martingale" was introduced to probability theory by Ville in 1939, though the concept was introduced by Lévy in 1934. This term was then used by Doob in 1951. The term "Optional Stopping Theorem" seems to have originated in Doob's 1953 book.

Consider the set $\Omega := \{1, 2, 3, \ldots\}$ with the probability measure $\mathbf{P}$ defined by $\mathbf{P}(\{n\}) := 2^{-n}$ for all $n \geq 1$. For any $n \geq 1$, define $X_n \colon \Omega \to \mathbb{R}$ so that $X_n(n) := 2^n$ and $X_n(m) := 0$ for any $m \in \Omega$ with $m \neq n$. Let H be the collection of random variables $H := \{X_1, X_2, \ldots\}$. Then H is bounded in L_1 since $\mathbf{E}\,|X_n| = 1$ for all $n \geq 1$, but H is not uniformly integrable. For any $m \geq 1$, $\sup_{X \in H} \mathbf{E}\,|X|\,1_{|X|>m} = 1$, so $\lim_{m\to\infty} \sup_{X \in H} \mathbf{E}\,|X|\,1_{|X|>m} = 1 \neq 0$. The

following theorem says that, in some sense, this is the only example of a bounded set $H \subseteq L_1$ that is not uniformly integrable. (From Exercise 6.41, if H is not bounded in L_1, then H is not uniformly integrable.) The Theorem below also restates uniform integrability in terms of compactness.

Theorem 6.60. *Let $H \subseteq L_1$ satisfy $\sup_{X \in H} \mathbf{E}\,|X| < \infty$. Then the following are equivalent.*

- *H is not uniformly integrable.*
- *H is not relatively weakly compact. (There exists a sequence $X_1, X_2, \ldots$ in H such that any subsequence of this sequence does not converge in the weak topology.) (We say $X_1, X_2, \ldots \in L_1$ converges to $X \in L_1$ in the weak topology if, for any $Y \in L_\infty$, $\lim_{n\to\infty} \mathbf{E}X_n Y = \mathbf{E}XY$.)*
- *$\exists\ \varepsilon > 0$ such that, for any integer $n \geq 1$, there exist n disjoint sets $A_1, \ldots, A_n \in \mathcal{F}$ such that, for all $1 \leq m \leq n$,*

$$\sup_{X \in H} \mathbf{E}\,|X|\,1_{A_m} \geq \varepsilon.$$

- *$\exists\ a, b > 0$ such that, for any integer $n \geq 1$, there exist $X_1, \ldots, X_n \in H$ such that, for any $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$,*

$$a \sum_{i=1}^n |\alpha_i| \leq \left\| \sum_{i=1}^n \alpha_i X_i \right\|_1 \leq b \sum_{i=1}^n |\alpha_i|\,.$$

(Said another way, $X_1, \ldots, X_n$ is equivalent up to constant factors to the unit vector basis of $\mathbb{R}^n$ equipped with the 1-norm, $\|(\alpha_1, \ldots, \alpha_n)\|_1 := \sum_{i=1}^n |\alpha_i|$.)

This obstruction to weak compactness does not occur in L_p when $1 < p < \infty$. If $1 < p < \infty$, it is well-known that $L_p^* = L_q$ where $1/p + 1/q = 1$, so that $L_p^{**} = L_p$, so the weak topology and weak* topology coincide on L_p. Therefore, the unit ball $\{X \in L_p\colon \|X\|_p \leq 1\}$ of L_p is weakly compact by Alaoglu's Theorem, Theorem 8.3. So, if $1 < p < \infty$, any bounded set in L_p is relatively weakly compact. Also, as we verified in Exercise 6.43, the unit ball $\{X \in L_p\colon \|X\|_p \leq 1\}$ of L_p is uniformly integrable if $1 < p < \infty$.

7. Some Concentration of Measure

Theorem 7.1 (Hall Marriage Theorem). *Let $G = (V, E)$ be a bipartite graph with vertices $C \cup D = V$. Suppose: for all $A \subseteq C$, $|\{d \in D\colon \{c, d\} \in E, c \in A\}| \geq |C|$. Then there exists a bijection $\phi\colon C \to D$ where $\phi(c) = d$ implies $(c, d) \in E$.*

Proof. We create an inductive procedure to improve any given matching. Begin with any injective, partially defined ϕ as above. Assume $\exists\ c \in C$ unmatched. By assumption, $\exists\ d_1 \in D$ such that $c \sim d_1$. If d_1 is unmatched, the inductive step is done. If not, let $c_1 := \phi^{-1}(d_1)$. By assumption, $\exists\ d_2 \in D$, $d_2 \neq d_1$ with d_2 joined to at least one of c, c_1. If d_2 is unmatched and $c \sim d_2$, the inductive step is done. If d_2 is unmatched and $c_1 \sim d_2$, we "flip" the assignments of ϕ, and we are done. (Let $\phi(c) := d_1, \phi(c_1) := d_2$). If d_2 is matched, let $c_2 := \phi^{-1}(d_2)$, etc. Eventually this process will terminate. □

Theorem 7.2 (Existence and Uniqueness of Haar Measure). *Let (Ω, d) be a compact metric space, $d\colon \Omega \times \Omega \to [0, \infty)$. Let G be a group whose members act as isometries on Ω. Then there exists a regular measure μ on the Borel sets of Ω that is invariant under the action of G. Equivalently, $\int_\Omega f(t)d\mu(t) = \int_\Omega f(gt)d\mu(t)$ for all $g \in G$, and for all continuous*

functions $f\colon \Omega \to \mathbb{R}$. Additionally, if G acts transitively on Ω, then μ is unique, up to multiplication by a constant.

Proof. Let $\varepsilon > 0$. Let N_ε be an ε-net of minimal cardinality (for every $\omega \in \Omega$, $\exists\ x \in N_\varepsilon$ such that $d(x,\omega) < \varepsilon$, and N_ε is as small as possible). For any $f\colon \Omega \to \mathbb{R}$ continuous, define $\mu_\varepsilon = \frac{1}{|N_\varepsilon|}\sum_{t\in N_\varepsilon} f(t)$. Since $\|\mu_\varepsilon\| := \sup_{f\colon \Omega\to\mathbb{R}, \text{continuous}, \|f\|_\infty \le 1} \int_\Omega f(x)\mu_\varepsilon dx \le 1$, and μ is positive, Alaoglu's Theorem 8.3) and Riesz's Theorem 3.4) show that $\exists\ \mu$ a subsequential weak* limit as $\varepsilon \to 0^+$ of μ_ε, with μ a regular Borel measure. (In fact μ is a probability measure since $\mu(1) = 1$). We claim that the choice of N_ε (among other minimal cardinality ε nets) does not change μ. Let $A := \{x_1, \dots, x_n\} \in N_\varepsilon$, $n \le |N_\varepsilon|$. If N'_ε denotes another such minimal net, then $A' := \{x' \in N'_\varepsilon\colon B(x',\varepsilon) \cap B(x,\varepsilon) \neq \emptyset, x \in A\}$ satisfies $|A'| \ge |A|$, where $B(x,\varepsilon) := \{\omega \in \Omega\colon d(\omega,x) < \varepsilon\}$. For if not, $A' \cup (N_\varepsilon \setminus A)$ is an ε net with smaller cardinality than N_ε. (If y is not ε close to $N_\varepsilon \setminus A$, then y is ε close to A, so $\exists\ a \in A, b \in N'_\varepsilon$ with $y \in B(a,\varepsilon) \cap B(b,\varepsilon)$, so $y \in A'$).

For any $x \in N_\varepsilon$, $y \in N'_\varepsilon$ write $x \sim y$ if their ε-neighborhoods intersect. By the Hall marriage theorem, Theorem 7.1), $\exists$ a bijection from $\phi\colon N_\varepsilon \to N'_\varepsilon$ where $\phi(x) = y$ implies $x \sim y$. That is, $d(\phi(x), x) \le 2\varepsilon$. Thus, we can immediately see that $|\mu_\varepsilon(f) - \mu'_\varepsilon(f)| \le \sup_{a,b\in\Omega\colon d(a,b)<2\varepsilon} |f(a) - f(b)|$. This proves G-invariance of μ, since gN_ε is another ε-net.

To show uniqueness, let $G_0 := \{g \in G\colon g\omega = \omega, \forall \omega \in \Omega\}$ and define a metric on G/G_0 by $\rho(g,h) := \sup_{\omega\in\Omega} d(g\omega, h\omega)$. Using the compactness of Ω, a subsequential argument with ε-nets of X "increasing density" shows that a sequence $g_1, g_2, \dots \in G/G_0$ has a subsequence that converges to some isometry g of Ω with respect to ρ. Since G/G_0 is a topological group, we use: if $V \subseteq G/G_0$ is a neighborhood of the identity, then $V \subseteq \overline{V} \subseteq V^2$. This result implies that $g \in G/G_0$, so that $(G/G_0, \rho)$ is a compact metric space.

Since G/G_0 acts on itself, let ν be a Haar measure for G/G_0. For any $f\colon \Omega \to \mathbb{R}$ continuous,

$$\nu(G)\mu(f) = \int_G 1 \int_\Omega f(gt)d\mu(t)d\nu(g) = \int_\Omega \int_G f(gt)d\nu(g)d\mu(t) =: \overline{\nu}(f)\mu(\Omega)$$

For the final equality, fix $t_0 \in \Omega$. Given $t \in \Omega$, invariance and transitivity show that there exists $g' \in G$ such that $g'(t) = t_0$, so $\int_G f(gt)d\nu(g) = \int_G f(gg't)d\nu(g) = \int_G f(gt_0)d\nu(g)$. That is, the inner integral in the last equality does not depend on t. □

Below, for any $n \ge 0$, we denote the n-dimensional sphere in $\mathbb{R}^{n+1}$ centered at the origin with radius 1 as

$$S^n := \{(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1}\colon x_1^2 + \cdots + x_{n+1}^2 = 1\}.$$

Theorem 7.3 (**Spherical Isoperimetric Inequality**). *Among all domains of fixed volume on the sphere, one with minimal boundary volume is the geodesic ball.*

Proof due to Figiel-Lindenstrauss-Milman-1977. Idea: if we start with an optimal set that is not a geodesic ball, we can apply a finite number of symmetrizations to it so that its interior is squished into a smaller region. This gives a contradiction, so we had a ball at the beginning. The technical device of outer radius allows the argument to proceed rigorously.

Let $S^{n-1} \subseteq \mathbb{R}^n$ be the unit sphere centered at the origin, and let $A \subseteq S^{n-1}$ be closed. Given two antipodal points $a, b \in S^{n-1}$, let $\gamma \subseteq S^{n-1}$ be a geodesic joining a and b. We define the symmetrization $\sigma_\gamma(A)$ as follows. For each $y \in \gamma$, let Π_y be the plane containing y, such that Π_y is perpendicular to the line in $\mathbb{R}^n$ connecting a and b. Note that $\Pi_y \cap S^{n-1}$ is a

dilation and translation of S^{n-2}, so let $\mu_{n-2,y}$ be the normalized Haar measure on $\Pi_y \cap S^{n-1}$. We let $\sigma_\gamma(A) \cap \Pi_y$ be a geodesic ball in S^{n-2} with center y, such that $\mu_{n-2,y}(\sigma_\gamma(A) \cap \Pi_y) = \mu_{n-2,y}(A \cap \Pi_y)$. From Fubini's Theorem, Theorem 1.66), we have $\mu_{n-1}(B) = \mu_{n-1}(A)$, where μ_{n-1} is normalized Haar measure on S^{n-1}.

We say $\sigma_\gamma(A)$ is the **symmetrization** of A with respect to γ. Let $r(A) := \min\{r > 0 \colon \exists\, x \in S^{n-1}, A \subseteq \overline{B(x,r)}\}$ be the (outer) **radius** of A. Here $B(x,r) := \{y \in S^{n-1} \colon d(x,y) < r\}$ is the open ball of radius r centered at y on S^{n-1}, and d is the usual metric on S^{n-1}, so that $d(x,y) = \cos^{-1}(\langle x, y\rangle)$ for all $x, y \in S^{n-1}$. The minimum in the definition of $r(A)$ exists by closedness of A and $\overline{B(x,r)}$. We claim that $\sigma_\gamma(A)$ is closed.

To show this, we use the Hausdorff distance on closed sets in S^{n-1}, $\delta(A,B) := \min\{r > 0 \colon A_r \supseteq B, B_r \supseteq A\}$. (Here $A_r := \{x \in S^{n-1} \colon d(x,A) < r\}$.) Let $B := \sigma_\gamma(A)$. Recall that the set of closed sets is a complete metric space with respect to the metric δ. Note that the function $y \mapsto \mu_{n-2,y}(A) = \mu_{n-2,y}(B)$ is upper semicontinuous in $y \in \gamma$, i.e. $\mu_{n-2,y_0}(A) \geq \limsup_{y \to y_0} \mu_{n-2,y}(A)$ when $y, y_0 \in \gamma$. This follows by the definition of the product topology and by the closedness of A. Now, writing S^{n-1} as $S^{n-2} \times [-1,1]/\sim$ where $(x,1) \sim (x',1)$ and $(x,-1) \sim (x',-1)$ $\forall\, x, x' \in S^{n-2}$, we can treat $A \subseteq S^{n-1}$ as a closed set in the product topology of $S^{n-2} \times [-1,1]$. Given $(x,y) \in B^c \subseteq S^{n-1} \times [-1,1]$, we wish to find a box $F \times G \subseteq S^{n-2} \times [-1,1]$ with F, G open, so that $(x,y) \in F \times G$ and $F \times G$ is disjoint from B. Since $B \cap \Pi_y$ is a geodesic ball (which is not all of S^{n-2}), we can find $F \times G$ as required, by the upper semicontinuity of $y \mapsto \mu_{n-2,y}(B)$. (Specifically, our inability to find such a box $F \times G$ would violate this upper semicontinuity.)

Below we also use that $\mu_{n-1}(\cdot)$ is upper semi-continuous with respect to δ, that is if $A^{(1)}, A^{(2)}, \ldots \subseteq S^{n-1}$ satisfy $\lim_{k\to\infty} \delta(A^{(k)}, A) = 0$, then $\mu_{n-1}(A) \geq \limsup_{k\to\infty} \mu_{n-1}(A^{(k)})$. To see this, let $x_k \in A^{(k)}$ for any $k \geq 1$. Since $d(x_k, A) \leq \delta(A^{(k)}, A) \to 0$ as $k \to \infty$, any limit point of the set $\{x_k\}_{k=1}^\infty$ must be contained in A. Therefore, for any fixed $\varepsilon > 0$, there exists $K > 0$ such that $k \geq K$ implies $A^{(k)} \subseteq A_\varepsilon$. Let $\lambda > \mu_{n-1}(A)$. Since μ_{n-1} is a Borel measure, there exists an open set U such that $A \subseteq U$ and $\mu_{n-1}(U) < \lambda$. Since A is compact, $d(A, U^c) > 0$, and there exists $\varepsilon > 0$ such that $A_\varepsilon \subseteq U$. Combining these observations, $\limsup_{k\to\infty} \mu_{n-1}(A^{(k)}) \leq \mu_{n-1}(A_\varepsilon) \leq \mu_{n-1}(U) < \lambda$. Therefore, $\limsup_{k\to\infty} \mu_{n-1}(A^{(k)}) \leq \mu_{n-1}(A)$, as desired.

We are now ready to proceed by inducting on n. For the case $n = 1$, the theorem is clear. We require the following claims, which are proven by induction.

Claim 1: Let $A \subseteq S^{n-1}$ be closed, and define

$$M(A) := \{C \subseteq S^{n-1} \colon C \text{ is closed}, \\ \mu_{n-1}(C) = \mu_{n-1}(A),\ \mu_{n-1}(C_\varepsilon) \leq \mu_{n-1}(A_\varepsilon)\ \forall\, \varepsilon > 0\}$$

Then there is a $B \in M(A)$ with minimal radius, i.e. $\min\{r(C) \colon C \subseteq M(A)\}$ exists.

Claim 2: Let $A \subseteq S^{n-1}$ be closed. Then for every half circle γ, $\sigma_\gamma(A) \in M(A)$.

Claim 3: Let $B \subseteq S^{n-1}$ be a closed set that is not a geodesic ball. There exists a finite family of half circles $\{\gamma_i\}_{i=1}^n \subseteq S^{n-1}$ so that $r(\sigma_{\gamma_n}(\sigma_{\gamma_{n-1}}(\cdots \sigma_{\gamma_1}(B) \cdots))) < r(B)$.

We prove the theorem assuming these claims. By definition of $M(A)$, $B \in M(A)$ and $C \in M(B)$ implies $C \in M(A)$. So, using Claim 2, $B \in M(A)$ and $\sigma_{\gamma_1}(B) \in M(B)$ implies $\sigma_{\gamma_1}(B) \in M(A)$, $\sigma_{\gamma_2}(\sigma_{\gamma_1}(B)) \in M(A)$, etc. Using Claim 3, we therefore see that an element of minimal (outer) radius in $M(A)$ must be a geodesic ball. Claim 1 says that this minimal

element must exist, so $M(A)$ must contain a geodesic ball. The theorem is therefore proven. We now prove the claims.

Proof of Claim 1: $B \mapsto r(B)$ is continuous (with respect to the Hausdorff metric for $B \subseteq S^{n-1}$), so it suffices to show that $M(A)$ is a closed subset in the space of closed subsets of S^{n-1} (since the latter space is compact with respect to δ). Let $B^{(1)}, B^{(2)}, \ldots \in M(A)$ with $\lim_{k\to\infty} \delta(B^{(k)}, B) = 0$ for some $B \subseteq S^{n-1}$, and let $\varepsilon \geq 0$. We will show $B \in M(A)$. For any fixed $\eta > 0$, there exists $K > 0$ such that, if $k \geq K$, then $B \subseteq B^{(k)}_\eta$, so $B_\varepsilon \subseteq B^{(k)}_{\varepsilon+\eta}$. So, for all $k \geq K$, $\mu_{n-1}(B_\varepsilon) \leq \mu_{n-1}(B^{(k)}_{\varepsilon+\eta}) \leq \mu_{n-1}(A_{\varepsilon+\eta})$, since $B^{(1)}, B^{(2)}, \ldots \in M(A)$. Therefore,

$$\mu_{n-1}(B_\varepsilon) \leq \inf_{\eta>0} \mu_{n-1}(A_{\varepsilon+\eta}) = \mu_{n-1}(\cap_{\eta>0} A_{\varepsilon+\eta}) = \mu_{n-1}(A_\varepsilon)$$

So, letting $\varepsilon = 0$, we get $\mu_{n-1}(B) \leq \mu_{n-1}(A)$. Moreover, $\mu_{n-1}(B) \geq \limsup_{k\to\infty} \mu_{n-1}(B^{(k)}) = \mu_{n-1}(A)$, using the upper semicontinuity of $\mu_{n-1}(\cdot)$ mentioned above, and the definition of $B^{(1)}, B^{(2)}, \ldots \in M(A)$. So $B \in M(A)$, as desired.

Proof of Claim 2: Let $A \subseteq S^{n-1}$ be closed and let γ be a half circle on S^{n-1} joining $z \in S^{n-1}$ with $-z$. Let u be the midpoint of γ. As usual, identify $S^{n-2,u} := S^{n-1} \cap \Pi_u$ with S^{n-2}. For any $y \in \gamma, y \neq \pm x$, define a map $\tau_y \colon S^{n-2,y} \to S^{n-2,u}$ by letting $\tau_y(x) := \gamma \cap S^{n-2,u}$ for any $x \in S^{n-2,y}$. (Note that this intersection is a single point). By applying polar coordinates, we see that there exists a function f such that, if $y_1, y_2 \in \gamma$ and if $x_1 \in S^{n-2,y_1}, x_2 \in S^{n-2,y_2}$, we have

$$d(x_1, x_2) = f(y_1, y_2, d(\tau_{y_1}(x_1), \tau_{y_2}(x_2))).$$

Moreover, for y_1, y_2 fixed, f is monotonically increasing with respect to its third argument, $d(\tau_{y_1}(x_1), \tau_{y_2}(x_2)) \leq \pi$.

For every $y_1, y_2 \in \gamma, \varepsilon > 0$ (with $d(y_1, y_2) < \varepsilon$) there is an $\eta(y_1, y_2, \varepsilon)$ so that, for every $C \subseteq S^{n-2,y_1}$, we have

$$C_\varepsilon \cap S^{n-2,y_2} = \tau_{y_2}^{-1}((\tau_{y_1} C)_{\eta(y_1,y_2,\varepsilon)}) \qquad (*)$$

To see this, it suffices to consider the case that $C = \{x_1\}$. Then

$$\begin{aligned} C_\varepsilon \cap S^{n-2,y_2} &= \{x_2 \in S^{n-2,y_2} \colon d(x_1, x_2) < \varepsilon\} \\ &= \{x_2 \in S^{n-2,y_2} \colon f(y_1, y_2, d(\tau_{y_1}(x_1), \tau_{y_2}(x_2))) < \varepsilon\} \\ &= \{x_2 \in S^{n-2,y_2} \colon d(\tau_{y_1}(x_1), \tau_{y_2}(x_2)) < \eta\} \end{aligned}$$

Here η is determined by the existence and monotonicity of f. (If $d(y_1, y_2) \geq \varepsilon$, then $C_\varepsilon \cap S^{n-2,y_2} = \emptyset$.) Note that the subscript ε on the left of $(*)$ denotes an ε neighborhood in S^{n-1}, whereas the subscript η on the right of $(*)$ denotes an η neighborhood in S^{n-2}. Let $A^y := A \cap S^{n-2,y}$. By fixing $y_2 = y$ and varying $y_1 = z$ in $(*)$, we have

$$\tau_y((A_\varepsilon)^y) = \cup_{\{z\in\gamma \colon d(z,y)<\varepsilon\}} (\tau_z(A^z))_{\eta(z,y,\varepsilon)} \qquad (**)$$

Substituting $B := \sigma_\gamma(A)$ gives

$$\tau_y((B_\varepsilon)^y) = \cup_{\{z\in\gamma \colon d(z,y)<\varepsilon\}} (\tau_z(B^z))_{\eta(z,y,\varepsilon)} \qquad (\dagger)$$

By definition of B, $\tau_z(B^z)$ is a geodesic ball in $S^{n-2,u}$ $\forall$ $z \in \gamma$, and $\mu_{n-2,u}(\tau_z(B^z)) = \mu_{n-2,u}(\tau_z(A^z))$. So, the induction hypothesis (i.e. the full theorem) says

$$\mu_{n-2,u}((\tau_z(B^z))_{\eta(z,y,\varepsilon)}) \leq \mu_{n-2,u}((\tau_z(A^z))_{\eta(z,y,\varepsilon)}) \qquad (\ddagger)$$

for admissible y, z, ε. Since the sets on the right side of (†) are all $(n-2)$-dimensional geodesic balls with the same center, we have

$$\begin{aligned}\mu_{n-2,u}(\tau_y(B_\varepsilon)^y) &= \sup_{z\in\gamma:\ d(z,y)\le\varepsilon} \mu_{n-2,u}((\tau_z(B^z))_{\eta(z,y,\varepsilon)}) \\ &\le \sup_{z\in\gamma:\ d(z,y)\le\varepsilon} \mu_{n-2,u}((\tau_z(A^z))_{\eta(z,y,\varepsilon)}) \quad \text{, from } (\ddagger) \\ &\le \mu_{n-2,u}(\tau_y(A_\varepsilon)^y) \quad \text{, from } (**)\end{aligned}$$

Re-writing this inequality, we see that for every $y \in \gamma, y \neq \pm x$ we have

$$\mu_{n-2,y}((B_\varepsilon)^y) \le \mu_{n-2,y}((A_\varepsilon)^y)$$

So by Fubini's Theorem, Theorem 1.66), we can integrate this inequality to get $\mu_{n-1}(B_\varepsilon) \le \mu_{n-1}(A_\varepsilon)$, so that $B \in M(A)$ as desired.

Proof of Claim 3: Let $B \subseteq S^{n-1}$ be closed, and suppose B is not a geodesic ball. Let $r = r(B)$ as above, and let $u \in S^{n-1}$ be such that $B \subseteq \overline{B(u,r)}$. Let γ be a half circle with midpoint u, so that we will symmetrize with respect to γ, leaving $\overline{B(u,r)}$ fixed. Since B is not a geodesic ball, $E := B^c \cap \partial B(u,r) \neq \emptyset$.

We need two observations. First, any symmetrization σ_γ does not decrease the set E. That is, $E \subseteq (\sigma_\gamma(B))^c \cap \partial B(u,r)$. Second, we can find symmetrizations that increase E. To see the second claim, let $G \subseteq \partial B(u,r)$ be a relatively open set. Given any $x \in \partial B(u,r) \setminus G$, there exists a relatively open set $G_x \subseteq \partial B(u,r)$ and γ_x such that $x \in G_x$, and $G_x \cap \sigma_{\gamma_x}(B) = \emptyset$. To construct γ_x, consider the straight line ℓ (in $\mathbb{R}^n$) between x and some point $y \in B^c \cap \partial B(u,r)$ (which exists since B is not a ball). Let P reflect $\partial B(u,r)$ across a hyperplane perpendicular to ℓ and intersecting ℓ at its midpoint. Then, let G_x be a small ball (in $\partial B(u,r)$) around x disjoint from G, such that $PG_x \subseteq B^c \cap \partial B(u,r)$ (which is possible since B is closed). Observe that G_x does what we claimed above. Also note that G_x, γ_x depend on x and G, but not on B.

Now, apply the above observations to B and $G := B^c \cap \partial B(u,r)$ to produce γ_1, G_{x_1}. Then, apply these same observations to $\sigma_{\gamma_1}(B)$ and $G := \sigma_{\gamma_1}(B)^c \cap \partial B(u,r)$ to produce γ_2 and G_{x_2}, and so on. By compactness of S^{n-1} (using a cover by $\{G_{x_i}\}_{i\ge1}$), after a finite number of symmetrizations we have $\sigma_{\gamma_n}(\cdots\sigma_{\gamma_1}(B)\cdots)$ disjoint from $\partial B(u,r)$. Therefore, $r(\sigma_{\gamma_n}(\cdots\sigma_{\gamma_1}(B)\cdots)) < r(B)$. □

As an application, we prove the following concentration of measure result. Note that the exponential dependence on n implies that almost all of a high dimensional sphere is close to any given set of Haar measure 1/2. Put another way, a high dimensional sphere has a "large waist."

Theorem 7.4 (**Concentration of measure on the sphere**). *Let μ be the normalized Haar measure on S^{n+1} (using Theorem 7.2). Let $A \subseteq S^{n+1}$, let $\varepsilon > 0$, and define $A_\varepsilon := \{x \in S^{n+1} : \exists y \in S^{n+1}$ with $d_{S^{n+1}}(x,y) \le \varepsilon\}$. If $\mu(A) \ge 1/2$ then $\mu(A_\varepsilon) \ge 1 - \sqrt{\frac{\pi}{8}} e^{-\varepsilon^2 n/2}$.*

Proof. By Theorem. 7.3, it suffices to prove this claim for geodesic balls, i.e. it suffices to analyze the quantity

$$\mu(B(\pi/2+\varepsilon)) = \frac{\int_{-\pi/2}^{\varepsilon} \cos^n(t)dt}{\int_{-\pi/2}^{\pi/2} \cos^n(t)dt}.$$

For any $n \geq 1$, let $I_n := \int_0^{\pi/2} \cos^n(t)dt$. Changing variables and using $\cos(t) \leq e^{-t^2/2}$, valid for any $0 \leq t \leq \pi/2$ (which follows since $f(t) := \log\cos t$ satisfies $f''(t) = -1/\cos^2(t) \leq -1$ for all $0 \leq t \leq \pi/2$),

$$\begin{aligned}
1 - \mu(B(\pi/2+\varepsilon)) &= \int_\varepsilon^{\pi/2} \cos^n(t)\frac{dt}{2I_n} = \frac{1}{\sqrt{n}}\int_{\varepsilon\sqrt{n}}^{(\pi/2)\sqrt{n}} \cos^n(t/\sqrt{n})\frac{dt}{2I_n} \\
&\leq \frac{1}{\sqrt{n}}\int_{\varepsilon\sqrt{n}}^{(\pi/2)\sqrt{n}} e^{-t^2/2}\frac{dt}{2I_n} \leq \frac{1}{\sqrt{n}}e^{-\varepsilon^2 n/2}\int_0^{(\pi/2-\varepsilon)\sqrt{n}} e^{-t^2/2}\frac{dt}{2I_n} \\
&\leq \frac{1}{\sqrt{n}}e^{-\varepsilon^2 n/2}\int_0^{\infty} e^{-t^2/2}\frac{dt}{2I_n} = \frac{1}{2\sqrt{n}I_n}e^{-\varepsilon^2 n/2}\sqrt{\pi/2}.
\end{aligned}$$

Integration by parts shows that $I_n = \frac{n-1}{n}I_{n-2}$. Since $(n-1)/\sqrt{n(n-2)} \geq 1$ for any $n \geq 3$, we get $\sqrt{n}I_n \geq \sqrt{n-2}I_{n-2}$ for any $n \geq 3$, so that

$$\sqrt{n}I_n \geq \min(I_1, \sqrt{2}I_2) = \min(1, \sqrt{2}\pi/4) = 1, \qquad \forall\, n \geq 1$$

In summary, $1 - \mu(B(\pi/2+\varepsilon)) \leq e^{-\varepsilon^2 n/2}\sqrt{\pi/8}$. □

Theorem 7.4 implies a corresponding statement for Lipschitz functions. That is, Lipschitz functions on high-dimensional spheres are typically close to their average value.

For any $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we denote $\|x\| := (x_1^2 + \cdots + x_n^2)^{1/2}$.

Theorem 7.5 (**Concentration of measure, Lipschitz function form**)). *Let $f\colon S^{n+1} \to \mathbb{R}$. Suppose that for all $x, y \in S^{n+1}$, $|f(x) - f(y)| \leq \|x - y\|$, so that f is 1-Lipschitz. Let μ denote normalized Haar measure on S^{n+1}, by Theorem 7.2. Then for all $\varepsilon > 0$,*

$$\mu\left(x \in S^{n+1}\colon \left|f(x) - \int_{S^{n+1}} f(y)d\mu(y)\right| \geq \varepsilon\right) \leq \sqrt{\frac{\pi}{2}}e^{-n\varepsilon^2/4}.$$

Proof. Let $m \in \mathbb{R}$ such that $\mu(x \in S^{n+1}\colon f(x) \leq m) \geq 1/2$ and $\mu(x \in S^{n+1}\colon f(x) \geq m) \geq 1/2$. Let $C := \{x \in S^{n+1}\colon f(x) \leq m\}$. Then $x \in C_\varepsilon$ if and only if $\exists\, y \in C$ with $\|x - y\|_2 \leq \varepsilon$. Since f is 1-Lipschitz, $|f(x) - f(y)| \leq \varepsilon$, so that $f(x) \leq m + \varepsilon$. Taking the contrapositive,

$$\{x \in S^{n+1}\colon f(x) > m + \varepsilon\} \subseteq S^{n+1} \setminus A_\varepsilon$$

So, from Thm. 7.4, since $\mu(C) \geq 1/2$, we have

$$\mu\left(x \in S^{n+1}\colon f(x) > m + \varepsilon\right) \leq \sqrt{\pi/8}e^{-n\varepsilon^2/2}.$$

Similarly, $\mu\left(x \in S^{n+1}\colon f(x) < m - \varepsilon\right) \leq \sqrt{\pi/8}e^{-n\varepsilon^2/2}$. In conclusion,

$$\mu\left(x \in S^{n+1}\colon |f(x) - m| > \varepsilon\right) \leq 2\sqrt{\pi/8}e^{-n\varepsilon^2/2}. \qquad (*)$$

It remains to replace m with $\int_{S^{n+1}} f(y)d\mu(y)$. Consider $\mu \times \mu$ on $S^{n+1} \times S^{n+1}$. Observe

$$\begin{aligned}
&(\mu\times\mu)\left((x,y) \in S^{n+1}\times S^{n+1}\colon |f(x) - f(y)| \geq \varepsilon\right) \\
&\quad\leq (\mu\times\mu)\left(\{|f(x) - m| \geq \varepsilon/2\} \cup \{|f(y) - m| \geq \varepsilon/2\}\right) \\
&\quad\leq 2\mu(|f(x) - m| \geq \varepsilon/2) \leq 4\sqrt{\pi/8}e^{-n\varepsilon^2/2} \quad \text{, from } (*)
\end{aligned}$$

Let $\lambda > 0$. Then from Theorem 1.60, if $\lambda^2 := n/4$,

$$\begin{aligned}
&\int_{S^{n+1}\times S^{n+1}} e^{\lambda^2(f(x)-f(y))^2} d\mu(x)d\mu(y) \\
&\quad = \int_0^\infty 2\lambda^2 t e^{\lambda^2 t^2} (\mu\times\mu)\left((x,y)\in S^{n+1}\times S^{n+1}\colon |f(x)-f(y)|\geq t\right) dt \\
&\quad \leq 4\sqrt{\pi/8}\int_0^\infty \lambda^2 t e^{\lambda^2 t^2} e^{-nt^2/2} dt = \sqrt{\pi/8}\int_0^\infty tne^{-nt^2/4} dt = 2\sqrt{\pi/8} = \sqrt{\pi/2}.
\end{aligned}$$

So, for this λ, Jensen's inequality in y, Theorem 1.40, implies that

$$\sqrt{\pi/2} \geq \int_{S^{n+1}\times S^{n+1}} e^{\lambda^2(f(x)-f(y))^2} d\mu(x)d\mu(y) \geq \int_{S^{n+1}} e^{\lambda^2(f(x)-\int_{S^{n+1}} f(y)d\mu(y))^2} d\mu(x).$$

Finally, by Chebyshev's inequality,

$$\begin{aligned}
\mu(x\in S^{n+1}\colon |f(x)-\int_{S^{n+1}} f d\mu| \geq \varepsilon) &= \mu(x\in S^{n+1}\colon e^{\lambda^2|f(x)-\int_{S^{n+1}} f(y)d\mu(y)|^2} \geq e^{\lambda^2\varepsilon^2}) \\
&\leq e^{-\lambda^2\varepsilon^2}\int_{S^{n+1}} e^{\lambda^2\left|f(x)-\int_{S^{n+1}} f(y)d\mu(y)\right|^2} d\mu(x) \leq \sqrt{\pi/2}e^{-\lambda^2\varepsilon^2}.
\end{aligned}$$

□

The Johnson-Lindenstrauss lemma says that the pairwise distances between n vectors in Euclidean space can be almost preserved by almost all linear projections into $O(\log n)$ dimensional Euclidean space.

Theorem 7.6 (**Johnson-Lindenstrauss**). *Let $x^{(1)},\ldots,x^{(n)}\in\mathbb{R}^m$. Let $\varepsilon > 0$. Then there exists a linear function $T\colon \mathbb{R}^m \to \mathbb{R}^{O(\varepsilon^{-2}\log n)}$ such that*

$$\left\|x^{(i)}-x^{(j)}\right\| \leq \left\|T(x^{(i)})-T(x^{(j)})\right\| \leq (1+\varepsilon)\left\|x^{(i)}-x^{(j)}\right\|, \qquad \forall\, 1\leq i,j\leq n.$$

One proves this via the probabilistic method. By concentration of measure, a random projection does what we require.

Proof. Let μ denote normalized Haar measure on S^{n-1} and let ν be normalized Haar measure on $O(n)$, the group of orthogonal $n\times n$ real matrices, by Theorem 7.2. Let $P\colon \mathbb{R}^n\to\mathbb{R}^n$ be the orthogonal projection such that $P(z_1,\ldots,z_n) := (z_1,\ldots,z_k,0,\ldots,0)$ for all $(z_1,\ldots,z_n)\in\mathbb{R}^n$. Fix $x_0\in S^{n-1}$. Suppose U is uniformly distributed in $O(n)$ and X is uniformly distributed in S^{n-1}. Observe that Ux_0 and X have the same distribution. To see this, let $A\subseteq S^{n-1}$ and define $\widetilde{\mu}(A) := \nu(U\in O(n)\colon Ux_0\in A)$. Note that $\widetilde{\mu}$ is $O(n)$ invariant, so apply Theorem 7.2. Now, define

$$E := \int_{S^{n-1}} \|Px\|\, d\mu(x) = \int_{O(n)} \|PUx_0\|\, d\nu(U).$$

We will eventually show that $E\geq 10^{-2}\sqrt{k/n}$. Observe

$$\begin{aligned}
\int_{S^{n-1}} \|Px\|^2\, d\mu(x) &= \int_{S^{n-1}} \Big(\sum_{i=1}^k x_i^2\Big) d\mu(x) \\
&= k\int_{S^{n-1}} x_1^2 d\mu(x) = \frac{k}{n}\int_{S^{n-1}} \Big(\sum_{i=1}^n x_i^2\Big) d\mu(x) = k/n. \qquad (*)
\end{aligned}$$

Now, we use Theorems 1.86 and 7.5 for the 1-Lipschitz function $x \mapsto \|Px\|$,

$$\begin{aligned}
&\int_{S^{n-1}} \|Px\|^4 \, d\mu(x) = \int_0^\infty 4u^3 \mu(\|Px\| \geq u) du \\
&\leq \int_0^{2E} 4u^3 du + \int_{2E}^\infty 4u^3 \mu(|\,\|Px\| - E| > u/2) du \\
&\leq 16E^4 + \sqrt{\frac{\pi}{2}} \int_{2E}^\infty u^3 e^{-nu^2/16} du \leq 16E^4 + \frac{8}{n^2} \int_{2E\sqrt{n}}^\infty v^3 e^{-v^2/16} dv \quad \text{, setting } v = u\sqrt{n} \\
&\leq 16E^4 + 8n^{-2} \int_0^\infty v^3 e^{-v^2/16} dv \leq 16E^4 + 10^3 n^{-2} \leq 16E^4 + 10^3 k^2 n^{-2} \\
&\leq 10^4 \left(\int_{S^{n-1}} \|Px\|_2^2 \, d\mu(x) \right)^2 \quad \text{, using Jensen's inequality and } (*).
\end{aligned}$$

So, if $Z := \|Px\|_2$ is a random variable, we have shown that $\mathbb{E}Z^4 < c(\mathbb{E}Z^2)^2$ where $c := 10^4$. So, using Hölder's Inequality, Theorem 1.48, for $p = 3/2$, $q = 3$,

$$\mathbb{E}Z^2 = \mathbb{E}(Z^{2/3} Z^{4/3}) \leq (\mathbb{E}Z)^{2/3} (\mathbb{E}Z^4)^{1/3} \leq (\mathbb{E}Z)^{2/3} c^{1/3} (\mathbb{E}Z^2)^{2/3}.$$

Using this inequality and $(*)$,

$$\mathbb{E}Z \geq c^{-1/2} \sqrt{\mathbb{E}Z^2} \geq 10^{-2} \sqrt{k/n}. \qquad (**)$$

In summary, $E \geq 10^{-2}\sqrt{k/n}$ for E defined above. Now, by uniqueness of Haar measure, Theorem 7.2, Theorem 7.5, and using $E \geq 10^{-2}\sqrt{k/n}$, for any $\varepsilon > 0$, and for any $x_0 \in S^{n-1}$,

$$\begin{aligned}
&\nu\left(U \in O(n) \colon \left|\, \left\|U^{-1}PUx_0\right\|_2 - E\right| \geq \varepsilon E\right) \\
&\quad = \mu\left(x \in S^{n-1} \,|\, \|Px\|_2 - E| \geq \varepsilon E\right) \leq \sqrt{\frac{\pi}{2}} e^{-n\varepsilon^2 E^2/4} \leq 2e^{-10^{-5} k\varepsilon^2}.
\end{aligned}$$

Let $x^{(1)}, \ldots, x^{(n)}$ be n points in $\mathbb{R}^n$. If $k \geq 10^6 \varepsilon^{-2} \log n$, the union bound shows that

$$\nu\left(U \in O(n) \colon \exists\, i \neq j \colon \left| \left\| U^{-1}PU \left(\frac{x^{(i)} - x^{(j)}}{\|x^{(i)} - x^{(j)}\|} \right) \right\|_2 - E \right| \geq \varepsilon E \right) \leq \binom{n}{2} 2e^{-10^{-5} k\varepsilon^2} < 1.$$

For any $1 \leq i \leq n$, define $y_i := U^{-1}PUx^{(i)}/(E(1-\varepsilon))$. Then $\exists\, U \in O(n)$ such that

$$1 \leq \left\| \frac{y^{(i)} - y^{(j)}}{\|x^{(i)} - x^{(j)}\|} \right\| \leq \frac{1+\varepsilon}{1-\varepsilon} \leq 1 + 3\varepsilon, \qquad \forall\, 1 \leq i, j \leq n.$$

So, our required embedding is $T(x^{(i)}) := y^{(i)}$ for all $1 \leq i \leq n$. Note that T is linear. (In fact, if we choose k to be slightly larger, then the probability becomes exponentially small, so essentially all U satisfies our desired property, hence essentially all linear projections $T \colon \mathbb{R}^n \to \mathbb{R}^{O(\varepsilon^{-2} \log n)}$ satisfy our desired property.) $\square$

8. Appendix: Results from Analysis

Theorem 8.1 (**Riesz Representation Theorem, Hilbert space version**). *Let $\ell\colon H \to \mathbb{R}$ be a continuous linear functional on a Hilbert space H. Then $\exists$ unique $v \in H$ such that $\ell(u) = \langle u, v\rangle$ for all $u \in H$.*

Proof. Uniqueness is clear. For existence, if $\ell = 0$ take $v = 0$. Otherwise let $M = \{u \in H\colon \ell(u) = 0\}$. Observe that M is a closed subspace and $M \neq H$. So we can let $w \neq 0$, $w \in M^\perp$, via Theorem 5.22(b). Then $\ell(w) \neq 0$. Let $v = (\overline{\ell(w)}/\|w\|^2)w$. Then $\ell(u - (\ell(u)/\ell(w))w) = 0$, so $u - (\ell(u)/\ell(w))w \in M$, and $v \in M^\perp$ so

$$\langle u, v\rangle = \left\langle u - \left(u - \frac{\ell(u)}{\ell(w)}w\right), v\right\rangle = \left\langle \frac{\ell(u)}{\ell(w)}w, \frac{\overline{\ell(w)}}{\|w\|^2}w\right\rangle = \ell(u)$$

□

Theorem 8.2. ***(Radon-Nikodym Theorem)*** *Let $(\Omega, \mathcal{F}, \mu)$ be a σ-finite measure space, and let ν be a σ-finite measure on $\mathcal{A}$ with $\nu \ll \mu$. Then $\exists$ $f \geq 0$ measurable with $\nu(E) = \int_E f d\mu$ for all $E \in \mathcal{A}$, and f is unique up to a set of μ measure zero.*

Proof. (von Neumann, finite ν, μ) One can check that $\ell(g) := \int_\Omega g d\nu$ is well-defined on $L_2(\Omega, \mathcal{F}, \mu + \nu)$. By duality (i.e. Theorem 8.1), $\exists$ $\phi \in L_2(\Omega, \mathcal{F}, \mu + \nu)$ such that $\ell(g) = \int_\Omega g\phi d(\mu + \nu)$. Let $E := \{\phi \geq 1\}$. Examining $\nu(E) = \ell(1_E)$, shows $\mu(E) = 0$, so $\nu(E) = 0$, so $\phi < 1$, ν-almost everywhere, hence μ-almost everywhere. So, since $d\nu = \phi d\mu + \phi d\nu$, we have $d\nu = \frac{\phi}{1-\phi} d\mu$. □

A **normed linear space** H is a vector space (over $\mathbb{R}$ or $\mathbb{C}$) with a **norm** $\|\cdot\|$. A norm is a function $\|\cdot\| : H \to [0, \infty)$ such that $\|h\| \geq 0$ for all $h \in H$, with equality if and only if $h = 0$, $\|\alpha h\| = |\alpha| \, \|j\|$ for all $h \in H$ and for all scalars α, and $\|h + g\| \leq \|h\| + \|g\|$ for all $h, g \in H$. Using the norm, we see that $d(h, g) := \|h - g\|$ is a metric $d\colon H \times H \to [0, \infty)$, whose open balls define the metric topology on H. We refer to this topology as the norm topology, or strong topology. Using the triangle inequality, one can show that a normed linear space is also a topological vector space. A **Banach space** is a normed linear space that is complete with respect to the norm topology.

A **linear functional** h^* on H is a linear map from H to scalars ($\mathbb{C}$ or $\mathbb{R}$) with $\|h^*\| < \infty$. The space of linear functionals is called the **dual space** of H, and is denoted by H^*. The norm of $h^* \in H^*$ is given by $\|h^*\| := \sup_{h \in H\colon \|h\| \leq 1} |h^*(h)|$.

We define the **weak* topology** on H^* via its basis of neighborhoods of $h_0^* \in H^*$ given by

$$U(h_0^*, \varepsilon, h_1, \ldots, h_n) := \{h^* \in H^*\colon \, |h^*(h_j) - h_0^*(h_j)| < \varepsilon \; \forall \, 1 \leq j \leq n\}.$$

Here $\varepsilon > 0$, $n \geq 1$, $h_1, \ldots, h_n \in H$.

Theorem 8.3. ***(Alaoglu Theorem/ Banach-Alaoglu)*** *Let H be a normed linear space. Then the unit ball $B_{H^*} = \{h^* \in H^*\colon \, \|h^*\| \leq 1\}$ of H^* is compact in the weak* topology.*

Proof. Let A be the set of scalar valued functions ξ on H with $\|\xi(h)\| \leq \|h\|$ for all $h \in H$. Equivalently, $A = \prod_{h \in H} B_h$ where $B_h := \{\lambda \in \{\text{scalars}\}\colon \, |\lambda| \leq \|h\|\}$. Then A with the product topology is compact by Tychonoff's Theorem. By the definition of the product topology, a basic open neighborhood of some $\xi_0 \in A$ is $\{\xi \in A\colon \, |\xi(h_j) - \xi_0(h_j)| < \varepsilon, \forall \, 1 \leq j \leq n\}$ for some $h_1, \ldots, h_n \in H$. Now for fixed $h \in H$, the projection map $\xi \mapsto \xi(h)$ is

continuous, from A (with the product topology) to scalars. (Given ξ in the inverse image of a small open interval, ξ is contained in an open set in A). Consider the natural embedding $B_{X^*} \subseteq A$. By the definition of the weak* topology, the topology induced by $B_{X^*} \subseteq A$ is exactly the weak* topology.

Putting everything together, let $g, h \in H$, α, β scalars, and observe: $\xi(\alpha g + \beta h) - \alpha\xi(g) - \beta\xi(h)$ is a continuous function of ξ, from A to scalars (since it is a composition of continuous functions). Therefore, $\{\xi \in A\colon \xi(\alpha x + \beta y) - \alpha\xi(x) - \beta\xi(y) = 0\}$ is closed in A (being an inverse image of zero). Therefore,

$$B_{X^*} = \bigcap_{\substack{g,h\in H,\\ \alpha,\beta\in\{\text{scalars}\}}} \{\xi \in A\colon \xi(\alpha x + \beta y) - \alpha\xi(x) - \beta\xi(y) = 0\}$$

is closed in the compact set A. □

Let $f, g\colon \mathbb{R} \to \mathbb{C}$ be measurable. For any $1 \leq p < \infty$, in this section we denote $\|f\|_p := (\int_{\mathbb{R}} |f(x)|^p\, dx)^{1/p}$ and $\|f\|_\infty := \inf\{c > 0\colon |f(x)| \leq c \text{ almost everywhere}\}$.

Theorem 8.4. ***(Minkowski's Inequality)*** *Let $1 \leq p \leq \infty$, and let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be measurable. Then*

$$\left\| \int_{\mathbb{R}} f(x,y)dx \right\|_{p,dy} \leq \int_{\mathbb{R}} \|f(x,y)\|_{p,dy}\, dx.$$

In particular, the integrand on the right is measurable, so if the right side is finite, then $\int_{\mathbb{R}} f(x,y)dx$ is defined for almost every $y \in \mathbb{R}$.

Proof. The right side is unchanged by replacing f with $|f|$, so without loss of generality we assume $f\colon \mathbb{R}^2 \to [0,\infty)$. The case $p = 1$ follows from Fubini's Theorem, Theorem 1.66. If $1 < p < \infty$, measurability follows from Fubini's Theorem, and the inequality follows from Fubini's Theorem and the Hölder inequality for y, Theorem 1.48 (for Lebesgue measure), with exponents p, p' (using $(p-1)p' = p$).

$$\begin{aligned}
\int_{\mathbb{R}} \left| \int_{\mathbb{R}} f(x,y)dx \right|^p dy &= \int_{\mathbb{R}} \left| \int_{\mathbb{R}} f(x,y)dx \right|^{p-1} \left| \int_{\mathbb{R}} f(x',y)dx' \right| dy \\
&= \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} f(x',y) \left| \int_{\mathbb{R}} f(x,y)dx \right|^{p-1} dy \Big) dx' \\
&\leq \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} |f(x',y)|^p\, dy \Big)^{1/p} \Big(\int_{\mathbb{R}} | \int_{\mathbb{R}} f(x,y)dx|^{p'(p-1)} dy \Big)^{1/p'} dx' \\
&= \int_{\mathbb{R}} \|f(x',y)\|_{p,dy}\, dx' \cdot \Big(\int_{\mathbb{R}} | \int_{\mathbb{R}} f(x,y)dx|^p dy \Big)^{1/p'}.
\end{aligned}$$

If the right-most term is nonnegative and finite, we divide both sides by it to conclude, using $1 - 1/p' = 1/p$. If the right-most term is zero, there is nothing to prove. In the case that f is the indicator function of a rectangle, the right-most term is finite, so the Theorem holds in this case. The Monotone Convergence Theorem, Theorem 1.54, then implies that the Theorem holds for more general functions f.

The case $p = \infty$ takes more work. Measurability follows by approximating f by simple functions, and using that the limit of measurable functions is measurable. We then use

duality. Let $g\colon \mathbb{R} \to [0,\infty)$ be measurable with $\int_{\mathbb{R}} g(y)dy \le 1$. Then by Fubini's Theorem and Hölder's inequality for y, Theorem 1.48 (for Lebesgue measure)

$$\int_{\mathbb{R}} g(y)\Big(\int_{\mathbb{R}} f(x,y)dx\Big)dy = \int_{\mathbb{R}}\Big(\int_{\mathbb{R}} f(x,y)g(y)dy\Big)dx \le \int_{\mathbb{R}} \|f(x,y)\|_{\infty,dy}\,dx. \qquad (*)$$

From the Reverse Hölder inequality, if $h\colon \mathbb{R} \to \mathbb{R}$ is measurable, then

$$\|h\|_\infty = \sup_{\substack{g\colon \mathbb{R}\to[0,\infty)\\ \int_{\mathbb{R}} g(y)dy \le 1}} \int_{\mathbb{R}} g(x)h(x)dx.$$

So, taking the supremum over such g in $(*)$, $\big\|\int_{\mathbb{R}} f(x,y)dx\big\|_{\infty,dy} \le \int_{\mathbb{R}} \|f(x,y)\|_{\infty,dy}\,dx$. □

We say $f\colon \mathbb{R} \to \mathbb{R}$ is a **Schwartz function** if, for any integers $j,k \ge 1$, f is k times continuously differentiable and there exists $c_{j,k} \in \mathbb{R}$ such that

$$\big|f^{(k)}(x)\big| \le \frac{c_{jk}}{1+|x|^j}, \qquad \forall\, x \in \mathbb{R}.$$

Proposition 8.5 (Properties of Convolution on $\mathbb{R}$). *Let $1 \le p \le \infty$, let p' with $1/p + 1/p' = 1$. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ with $\int_{\mathbb{R}} |\phi(x)|\,dx < \infty$, let $\varepsilon > 0$ and define $\phi_\varepsilon(x) := \frac{1}{\varepsilon}\phi(x/\varepsilon)$ for any $x \in \mathbb{R}$ and $c := \int_{\mathbb{R}} \phi(x)dx$. Let $f,g\colon \mathbb{R} \to \mathbb{R}$ be Schwartz functions.*

(a) *For any $1 \le p < \infty$, $\lim_{\varepsilon\downarrow 0} \|\phi_\varepsilon * f - cf\|_p = 0$.*
(b) $\lim_{\varepsilon\to 0^+} \|\phi_\varepsilon * f - cf\|_\infty = 0$.
(c) *For any $x \in \mathbb{R}$, $\lim_{\varepsilon\to 0^+}(\phi_\varepsilon * f)(x) = cf(x)$ (using only that f is bounded, continuous).*
(d) *The convergence in (c) is uniform on $\mathbb{R}$ (using only that f is uniformly continuous).*
(e) *$\forall\, m \ge 1$, $f * g$ is m times continuously differentiable, and $(f*g)^{(m)} = f^{(m)} * g$.*

Proof of (a),(b):

$$\begin{aligned}
\|\phi_\varepsilon * f - cf\|_p &= \left\| \int_{\mathbb{R}} \phi_\varepsilon(y)(f(x-y) - f(x))dy \right\|_{p,dx} \\
&\le \int_{\mathbb{R}} |\phi_\varepsilon(y)|\,\|f(x-y) - f(x)\|_{p,dx}\,dy \quad \text{, by Theorem. 8.4} \\
&= \int_{\mathbb{R}} |\phi(y)|\,\|f(x-\varepsilon y) - f(x)\|_{p,dx}\,dy, \text{ changing variables.}
\end{aligned}$$

The y-integrand is bounded by $2\|f\|_p \int_{\mathbb{R}} |\phi(y)|\,dy < \infty$ and by $|\phi(y)|\,|\varepsilon y|\,\|f'\|_\infty$ by the Fundamental Theorem of Calculus. Since f is Schwartz, the latter quantity is bounded, so it goes to zero pointwise as $\varepsilon \to 0$. So, the Dominated Convergence Theorem, Theorem 1.57, implies (a) and (b).

Proof of (c): Arguing as in (a) (taking absolute values, changing variables, and applying Dominated Convergence),

$$|(\phi_\varepsilon * f)(x) - cf(x)| \le \int_{\mathbb{R}} |\phi(y)|\,|f(x-\varepsilon y) - f(x)|\,dy \to 0.$$

Proof of (d): Let $\eta > 0$. Choose $m > 0$ so that $2\|f\|_\infty \int_{|y|>m} |\phi(y)| \le \eta$. Choose $\delta > 0$ by uniform continuity of f so that for any $x \in \mathbb{R}$, if $|u| \le \delta$ then $|f(x+u) - f(x)| \le \eta/\|\phi\|_1$.

Then for any $0 < \varepsilon \le \delta/m$ and for any $x \in \mathbb{R}$, if $|y| \le m$, then $|f(x-\varepsilon y) - f(x)| \le \eta / \|\phi\|_1$. So, continuing the calculation of (c), and applying the definition of m,

$$\int_{\mathbb{R}} |\phi(y)|\,|f(x-\varepsilon y)-f(x)|\,dy = \int_{\{y\in\mathbb{R}:\,|y|>m\}} (\cdots) + \int_{\{y\in\mathbb{R}:\,|y|\le m\}} (\cdots)$$
$$\le 2\,\|f\|_\infty \int_{\{y\in\mathbb{R}:\,|y|>m\}} |\phi(y)|\,dy + \int_{\{y\in\mathbb{R}:\,|y|\le m\}} |\phi(y)|\,\frac{\eta}{\|\phi\|_1} \le \eta+\eta = 2\eta.$$

Proof of (e): Let $h > 0$ and $x \in \mathbb{R}$. Then

$$\left|\frac{(f*g)(x+h)-(f*g)(x)}{h} - (f'*g)(x)\right| \le \left\|\frac{f(x+h)-f(x)}{h} - f'(x)\right\|_{\infty,dx} \|g\|_1$$
$$\le \left\|\frac{1}{h}\int_x^{x+h} (x+h-t)f''(t)dt\right\|_{\infty,dx} \|g\|_1 \le |h|\,\|f''\|_\infty\,\|g\|_1\,.$$

Since f is a Schwartz function, $\|f''\|_\infty < \infty$, so the case $m = 1$ follows by letting $h \to 0^+$. The case of larger m follows by iteration. □

Let $f\colon \mathbb{R} \to \mathbb{R}$ with $\int_{\mathbb{R}} |f(x)|\,dx < \infty$. For any $\xi \in \mathbb{R}$, we define

$$\widehat{f}(\xi) = \mathcal{F}(f)(\xi) := \int_{\mathbb{R}} e^{ix\xi} f(x)dx.$$

Then $\widehat{f}\colon \mathbb{R} \to \mathbb{R}$ is called the **Fourier Transform** of f.

Proposition 8.6 (**Properties of Fourier Transform**). *Let f, g be Schwartz functions. Let $\xi \in \mathbb{R}$ and let $\lambda > 0$.*

(a) $|\widehat{f}(\xi)| \le \int_{\mathbb{R}} |f(x)|\,dx$, $\forall\ \xi \in \mathbb{R}$.
(b) $\mathcal{F}[f(x-h)](\xi) = e^{i\xi h}\widehat{f}(\xi)$, $\mathcal{F}[e^{ixh}f(x)](\xi) = \widehat{f}(\xi+h)$, $\forall\ h \in \mathbb{R}$.
(c) $\mathcal{F}[f(x/\lambda)](\xi) = \lambda\widehat{f}(\lambda\xi)$.
(d) $\widehat{(f*g)} = \widehat{f}\,\widehat{g}$
(e) $\partial\widehat{f}/\partial\xi = \mathcal{F}(ixf(x))$
(f) $\mathcal{F}[f'](\xi) = -i\xi\widehat{f}(\xi)$.
(g) $\int_{\mathbb{R}} f(x)\widehat{g}(x)dx = \int_{\mathbb{R}} \widehat{f}(x)g(x)dx$.

Proof of (a): $|\widehat{f}(\xi)| = \left|\int_{\mathbb{R}} e^{ix\xi}f(x)dx\right| \le \int_{\mathbb{R}} |f(x)|\,dx$.
Proof of (b): By the change of variables formula, if $\xi \in \mathbb{R}$,

$$\mathcal{F}[f(x-h)](\xi) = \int_{\mathbb{R}} e^{ix\xi}f(x-h)dx = e^{ixh}\int_{\mathbb{R}} e^{ix\xi}f(x)dx = e^{ixh}\widehat{f}(\xi).$$

$$\mathcal{F}[e^{ixh}f(x)](\xi) = \int_{\mathbb{R}} e^{ix(\xi+h)}f(x)dx = \widehat{f}(\xi+h).$$

Proof of (c): By the change of variables formula,

$$\mathcal{F}[f(x/\lambda)](\xi) = \int_{\mathbb{R}} e^{ix\xi}f(x/\lambda)dx = \lambda\int_{\mathbb{R}} e^{ix\xi\lambda}f(x)dx = \lambda\widehat{f}(\xi\lambda).$$

Proof of (d): Applying Fubini's Theorem, Theorem 1.66, and part (b) give

$$\int_{\mathbb{R}} e^{ix\xi}\left(\int_{\mathbb{R}} f(x-y)g(y)dy\right)dx = \int_{\mathbb{R}}\int_{\mathbb{R}} e^{ix\xi}f(x-y)dxg(y)dy$$
$$\stackrel{(b)}{=} \int_{\mathbb{R}} e^{i\xi y}\widehat{f}(\xi)g(y)dy = \widehat{f}(\xi)\int_{\mathbb{R}} e^{i\xi y}g(y)dy = \widehat{f}(\xi)\widehat{g}(\xi).$$

Proof of (e): Let $h > 0$. Using part (b) and the Dominated Convergence Theorem 1.57,

$$\frac{\widehat{f}(\xi+h)-\widehat{f}(\xi)}{h} \stackrel{(b)}{=} \mathcal{F}\left[\left(\frac{e^{ixh}-1}{h}\right)f(x)\right](\xi) \to \mathcal{F}[ixf(x)](\xi)\,, \text{as } h\to 0.$$

We now justify the use of the Dominated Convergence Theorem. By the Mean Value Theorem, $\left|\mathrm{Re}(e^{ixh}-1)/h\right| = |(\cos(xh)-1)/h| \le |x|$ and $\left|\mathrm{Im}(e^{ixh}-1)/h\right| = |(\sin(xh)-1)/h| \le |x|$, so $\left|(e^{ixh}-1)/h\right| \le 2\,|x|$ and $\left|f(x)(e^{ixh}-1)/h\right| \le 2\,|x|\,|f(x)|$.
Proof of (f): Integrating by parts and then using that f is a Schwartz function

$$\mathcal{F}[f'(x)](\xi) = \lim_{N\to\infty}\int_{-N}^{N} f'(x)e^{ix\xi}dx = \lim_{N\to\infty} -\int_{-N}^{N} f(x)(i\xi)e^{ix\xi}dx = -i\xi\widehat{f}(\xi).$$

Proof of (g): Apply Fubini's Theorem 1.66. □

Proposition 8.7. *Let f, g be Schwartz functions. Let $\xi \in \mathbb{R}$.*

(a) $\mathcal{F}[e^{-x^2/2}](\xi) = \sqrt{2\pi}e^{-\xi^2/2}$.
(b) $\lim_{\xi\to\infty}\widehat{f}(\xi) = 0$.
(c) *$\widehat{f}$ is a Schwarz function.*

Proof. Let $\xi \in \mathbb{R}$. Completing the square, and then shifting the contour in the complex plane,

$$\int_{\mathbb{R}} e^{-x^2/2+ix\xi}dx = e^{-\xi^2/2}\int_{\mathbb{R}} e^{-(x-i\xi)^2/2}dx = e^{-\xi^2/2}\int_{\mathbb{R}} e^{-x^2/2}dx = \sqrt{2\pi}e^{-\xi^2/2}.$$

Now, let $\phi(x) := e^{-x^2/2}/\sqrt{2\pi}$ for any $x \in \mathbb{R}$ and denote $\phi_\varepsilon(x) := \varepsilon^{-1}\phi(x/\varepsilon)$ for any $x \in \mathbb{R}$. Note that $\int_{\mathbb{R}}\phi_\varepsilon(x)dx = 1$. From Proposition 8.6(a),(d) and Proposition 8.5(a),

$$\left|\widehat{\phi_\varepsilon}(\xi)\widehat{f}(\xi) - \widehat{f}(\xi)\right| = \left|\widehat{\phi_\varepsilon * f}(\xi) - \widehat{f}(\xi)\right| \le \int_{\mathbb{R}} |\phi_\varepsilon * f(x) - f(x)|\,dx \to 0,$$

as $\varepsilon \to 0$. Combining this statement with Proposition 8.6(c) and part (a) of the current Proposition, $e^{-\varepsilon^2\xi^2/2}\widehat{f}(\xi)$ converges to $\widehat{f}(\xi)$ uniformly over all $\xi \in \mathbb{R}$, as $\varepsilon \to 0$. Since $\widehat{f}$ itself is bounded by Proposition 8.6(a), $e^{-\varepsilon^2\xi^2/2}\widehat{f}(\xi)$ vanishes at $\xi = \infty$, for every $\varepsilon > 0$. So, the uniform convergence implies that $\widehat{f}(\xi)$ also vanishes as $\xi \to \infty$, proving (b).

To prove (c), note that repeated application of Proposition 8.6 shows that $\widehat{f}$ is k times differentiable for any $k \ge 1$, since f is a Schwartz function. And part (b) of the current Proposition says that $f^{(k)}$ vanishes at infinity for any $k \ge 1$, so repeated application of Proposition 8.6(f) shows that $\widehat{f}$ is a Schwartz function. □

Exercise 8.8. Give an alternate proof of the fact $\mathcal{F}[e^{-x^2/2}](\xi) = \sqrt{2\pi}e^{-\xi^2/2}$ using the following strategy:

- Let $g(\xi) := (2\pi)^{-1/2}\mathcal{F}[e^{-x^2/2}](\xi)$. Show that $g'(\xi) = -\xi g(\xi)$ for all $\xi \in \mathbb{R}$.

- Deduce that $(d/d\xi)(g(\xi)e^{\xi^2/2}) = 0$.
- Finally, conclude that $g(\xi) = e^{-\xi^2/2}$.

Theorem 8.9 (**Fourier Inversion**). *Let $f\colon \mathbb{R} \to \mathbb{R}$ be a Schwartz function. Then*

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-ix\xi}\widehat{f}(\xi)d\xi, \qquad \forall\, x \in \mathbb{R}.$$

Proof. let $\phi(x) := e^{-x^2/2}/\sqrt{2\pi}$ for any $x \in \mathbb{R}$ and denote $\phi_\varepsilon(x) := \varepsilon^{-1}\phi(x/\varepsilon)$ for any $x \in \mathbb{R}$. Note that $\int_{\mathbb{R}} \phi_\varepsilon(x)dx = 1$. By Proposition 8.6(c) and Proposition 8.7(a), $\mathcal{F}[\phi](\xi) = e^{-\xi^2/2}$, $\mathcal{F}[\phi_\varepsilon](\xi) = e^{-\varepsilon^2\xi^2/2}$, and $\mathcal{F}(\mathcal{F}(\phi_\varepsilon)) = 2\pi\phi_\varepsilon$. So, using Theorem 8.6(g), we get

$$2\pi\int_{\mathbb{R}} f(x)\phi_\varepsilon(x)dx = \int_{\mathbb{R}} \widehat{f}(\xi)e^{-\varepsilon^2\xi^2/2}d\xi. \qquad (*)$$

Using this equality for $f(x+y)$, applying Theorem 8.6(b), and using $\phi_\varepsilon(-y) = \phi_\varepsilon(y)\ \forall\, y \in \mathbb{R}$,

$$\frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(\xi)e^{-ix\xi}e^{-\varepsilon^2\xi^2/2}d\xi \overset{(*)}{=} \int_{\mathbb{R}} f(x+y)\phi_\varepsilon(y)dy = \int_{\mathbb{R}} f(x-y)\phi_\varepsilon(y)dy = (\phi_\varepsilon * f)(x).$$

As $\varepsilon \to 0$, the left side converges to $\frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(\xi)e^{ix\xi}d\xi$ by the Dominated Convergence Theorem 1.57. And the right side tends to f uniformly in x by Proposition 8.5(d). So $f(x) = \frac{1}{2\pi}\int \widehat{f}(\xi)e^{-ix\xi}d\xi$ almost everywhere in $x \in \mathbb{R}$, hence everywhere since f is Schwartz. □

Lemma 8.10 (**Stirling's Formula**). *Let $n \in \mathbb{N}$. Then $n! \sim \sqrt{2\pi n}n^n e^{-n}$. That is,*

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}n^n e^{-n}} = 1.$$

Proof. We prove the weaker estimate that $\exists\ c \in \mathbb{R}$ such that

$$n! = (1 + O(1/n))e^{1-c}\sqrt{n}n^n e^{-n}. \qquad (*)$$

Note that $\log(n!) = \sum_{m=1}^n \log m$. We use integral comparison for this sum. On the interval $[m, m+1]$ the function $x \mapsto \log x$ has second derivative $O(1/m^2)$. So, Taylor expansion (i.e. the trapezoid rule) gives

$$\int_m^{m+1} \log x dx = \frac{1}{2}\log(m+1) + \frac{1}{2}\log m + O(1/m^2).$$

$$\int_1^n \log x dx = \sum_{m=1}^{n-1}\int_m^{m+1} \log x dx = \sum_{m=1}^{n-1} \log m + \frac{1}{2}\log n + c + O(1/n).$$

Since $\int_1^n \log x dx = n(\log(n) - 1) + 1$, $\log(n!) = \sum_{m=1}^n \log m$, exponentiating proves $(*)$. □

9. Appendix: Notation

Let n, m be a positive integers. Let A, B be sets contained in a universal set Ω.

$\mathbb{N} = \{1, 2, \ldots\}$ denotes the set of natural numbers

$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ denotes the set of integers

$\mathbb{Q} = \{a/b \colon a, b, \in \mathbb{Z}, b \neq 0\}$ denotes the set of rational numbers

$\mathbb{R}$ denotes the set of real numbers

$\mathbb{C} = \{a + b\sqrt{-1} \colon a, b \in \mathbb{R}\}$ denotes the set of complex numbers

$\in$ means "is an element of." For example, $2 \in \mathbb{R}$ is read as "2 is an element of $\mathbb{R}$."

$\forall$ means "for all"

$\exists$ means "there exists"

$\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n) \colon x_i \in \mathbb{R} \, \forall \, 1 \leq i \leq n\}$

$f \colon A \to B$ means f is a function with domain A and range B. For example,

$f \colon \mathbb{R}^2 \to \mathbb{R}$ means that f is a function with domain $\mathbb{R}^2$ and range $\mathbb{R}$

$\emptyset$ denotes the empty set

$A \subseteq B$ means $\forall \, a \in A$, we have $a \in B$, so A is contained in B

$A \setminus B := \{a \in A \colon a \notin B\}$

$A^c := \Omega \setminus A$, the complement of A in Ω

$A \cap B$ denotes the intersection of A and B

$A \cup B$ denotes the union of A and B

$A \Delta B := (A \setminus B) \cup (B \setminus A)$

$\mathbf{P}$ denotes a probability law on Ω

Let $a_1, \ldots, a_n$ be real numbers. Let n be a positive integer.

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_{n-1} + a_n.$$

$$\prod_{i=1}^{n} a_i = a_1 \cdot a_2 \cdots a_{n-1} \cdot a_n.$$

$\min(a_1, a_2)$ denotes the minimum of a_1 and a_2.

$\max(a_1, a_2)$ denotes the maximum of a_1 and a_2.

The min of a set of nonnegative real numbers is the smallest element of that set. We also define $\min(\emptyset) := \infty$.

Let $z \in \mathbb{C}$, so that $z = a + b\sqrt{-1}$ for some $a, b \in \mathbb{R}$.

$\mathrm{Re}(z) := a$ denotes the real part of z.

$\mathrm{Im}(z) := b$ denotes the imaginary part of z.

Let $X\colon \Omega \to \mathbb{R}$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mu)$.

$$\begin{aligned}
\mathbf{E}(X) &\text{ denotes the expected value of } X \\
\|X\|_p &:= (\mathbf{E}\,|X|^p)^{1/p}, \text{ denotes the } L_p\text{-norm of } X \text{ when } 1 \le p < \infty \\
\|X\|_\infty &:= \inf\{c > 0\colon \mathbf{P}(|X| \le c) = 1\}, \text{ denotes the } L_\infty\text{-norm of } X \\
\operatorname{var}(X) &= \mathbf{E}(X - \mathbf{E}(X))^2, \text{ the variance of } X \\
\sigma_X &= \sqrt{\operatorname{var}(X)}, \text{ the standard deviation of } X
\end{aligned}$$

Let $A \subseteq \Omega$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. Let $Y\colon \Omega \to \mathbb{R}$. Assume $\mathbf{E}\,|X| < \infty$. Let $\sigma(Y)$ denote the σ-algebra generated by Y.

$\mathbf{E}(X|A) := \mathbf{E}(X1_A)/\mathbf{P}(A)$ denotes the expected value of X conditioned on the event A.

$\mathbf{E}(X|\mathcal{G})$ denotes the conditional expectation of X given $\mathcal{G}$.

$\mathbf{E}(X|Y) := \mathbf{E}(X|\sigma(Y))$ denotes the conditional expectation of X given Y.

$1_A\colon \Omega \to \{0, 1\}$, denotes the indicator function of A, so that

$$1_A(\omega) = \begin{cases} 1 & \text{, if } \omega \in A \\ 0 & \text{, otherwise.} \end{cases}$$

Let H be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$. Let $h \in H$.

$$\|h\| := \langle h, h \rangle^{1/2}, \text{ denotes the norm of } h$$

Let X be a random variable on a sample space Ω, so that $X\colon \Omega \to \mathbb{R}$. Let $\mathbf{P}$ be a probability law on Ω. Let $x, t \in \mathbb{R}$. Let $i := \sqrt{-1}$.

$$\begin{aligned}
F_X(x) &= \mathbf{P}(X \le x) = \mathbf{P}(\{\omega \in \Omega\colon X(\omega) \le x\}) \\
&\quad \text{the Cumulative Distibution Function of } X. \\
M_X(t) &= \mathbf{E}e^{tX} \text{ denotes the Moment Generating Function of } X \text{ at } t \in \mathbb{R} \\
\phi_X(t) &= \mathbf{E}e^{itX} \text{ denotes the Characteristic Function (or Fourier Transform) of } X \text{ at } t \in \mathbb{R}
\end{aligned}$$

We define the **tail σ-algebra** of random variables $X_1, X_2, \ldots$ to be

$$\mathcal{T} := \bigcap_{i=1}^{\infty} \sigma(X_i, X_{i+1}, \ldots).$$

We let $\mathcal{E}$ denote the **exchangeable σ-algebra**.

Let $g, h\colon \mathbb{R} \to \mathbb{R}$. Let $t \in \mathbb{R}$.

$$(g * h)(t) = \int_{-\infty}^{\infty} g(x)h(t - x)dx \text{ denotes the convolution of } g \text{ and } h \text{ at } t \in \mathbb{R}$$

Let $f, g\colon \mathbb{R} \to \mathbb{C}$. We use the notation $f(t) = o(g(t))$, $\forall\ t \in \mathbb{R}$ to denote $\lim_{t\to 0} \left|\frac{f(t)}{g(t)}\right| = 0$. Let $A \subseteq \mathbb{R}$ and let $f, g\colon A \to \mathbb{C}$. We use the notation $f(t) = O(g(t))$ to denote that $\exists\ c > 0$ such that $|f(t)| \leq c\,|g(t)|$ for all $t \in A$.

USC Mathematics, Los Angeles, CA
E-mail address: stevenmheilman@gmail.com