Probability Theory I Notes

Instructor: Prof. Konstantinos Spiliopoulos

Student: Wancheng Lin

Contents

1 Foundations of Probability

Sets and Basic Operations

Sets and Elements

Definition 1.1 (Set)**.** *A set is a collection of objects called elements. We write $x \in A$ if x belongs to A, and $x \notin A$ otherwise.*

Example (Basic Examples of Sets)**.**This example illustrates different kinds of sets:

- $A = \{1, 2, 3\}$ (finite set),
- $\mathbb{N} = \{0, 1, 2, \dots\}$ (countable infinite set),
- $\mathbb{R}$ (uncountable set),
- $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$ (the closed unit disk in $\mathbb{R}^2$).

Definition 1.2 (Subset)**.** *For sets A, B, we say A is a subset of B (denoted $A \subseteq B$) if*

$$\forall x \ (x \in A \implies x \in B).$$

Basic Set Operations

For $A, B \subseteq \Omega$, the usual operations are:

$$A \cup B = \{x : x \in A \text{ or } x \in B\}, \quad A \cap B = \{x : x \in A \text{ and } x \in B\},$$

$$A^c = \Omega \setminus A, \quad A \setminus B = \{x \in A : x \notin B\}, \quad A \triangle B = (A \setminus B) \cup (B \setminus A).$$

Theorem 1.1 (De Morgan's Laws)**.** *For $\{A_n\}_{n=1}^{\infty} \subseteq \Omega$,*

$$\Big(\bigcup_{n=1}^{\infty} A_n\Big)^c = \bigcap_{n=1}^{\infty} A_n^c, \qquad \Big(\bigcap_{n=1}^{\infty} A_n\Big)^c = \bigcup_{n=1}^{\infty} A_n^c.$$

Countability

Countable and Uncountable Sets

Definition 1.3 (Countable)**.** *A set A is countable if either it is finite or there exists a bijection $f : A \to \mathbb{N}$.*

Example (Countable vs Uncountable Sets)**.**We contrast common countable and uncountable sets:

- Countable: $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$.
- Uncountable: $(0, 1), \mathbb{R}$, and the set of infinite binary sequences (Cantor's diagonal argument).

Diagonal Argument for Uncountability

Theorem 1.2 (Cantor's Diagonal Argument)**.** *The set of all infinite binary sequences*

$$\{(a_n)_{n=1}^{\infty} : a_n \in \{0, 1\}\}$$

is uncountable.

Proof. Assume, for contradiction, that the set of infinite 0–1 sequences can be listed as

$$s^{(1)}, s^{(2)}, s^{(3)}, \dots$$

where each $s^{(i)} = (a_1^{(i)}, a_2^{(i)}, a_3^{(i)}, \dots)$.

Now construct a new sequence $t = (b_1, b_2, b_3, \dots)$ by choosing

$$b_n = \begin{cases} 0, & a_n^{(n)} = 1, \\ 1, & a_n^{(n)} = 0. \end{cases}$$

That is, b_n is defined to be different from the n-th entry of the n-th sequence. By construction, t differs from $s^{(n)}$ in the n-th coordinate for every n. Thus t is not in the list, contradicting the assumption that all sequences were listed. Therefore, the set of infinite binary sequences is uncountable. □

Remark. The same diagonalization argument shows that the interval $(0, 1)$ is uncountable, since each binary sequence corresponds to the binary expansion of a number in $(0, 1)$.

Countability of the Rational Numbers

Theorem 1.3. *The set of rational numbers $\mathbb{Q}$ is countable.*

Proof. Every rational number can be expressed as $\frac{p}{q}$ with $p \in \mathbb{Z}$ and $q \in \mathbb{N}$. Consider the infinite array

$$\begin{array}{cccc} \frac{1}{1} & \frac{2}{1} & \frac{3}{1} & \cdots \\ \frac{1}{2} & \frac{2}{2} & \frac{3}{2} & \cdots \\ \frac{1}{3} & \frac{2}{3} & \frac{3}{3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

We can enumerate these fractions by traversing the array diagonally:

$$\frac{1}{1}, \frac{2}{1}, \frac{1}{2}, \frac{1}{3}, \frac{2}{2}, \frac{3}{1}, \frac{4}{1}, \frac{3}{2}, \frac{2}{3}, \frac{1}{4}, \ldots$$

This procedure produces a sequence containing every positive rational number infinitely many times. By discarding duplicates (e.g. $\frac{2}{2} = 1$, $\frac{3}{3} = 1$, etc.), we obtain a bijection between $\mathbb{N}$ and $\mathbb{Q}^+$, the set of positive rationals.

Finally, we can interleave negatives:

$$0, 1, -1, \tfrac{1}{2}, -\tfrac{1}{2}, 2, -2, \tfrac{2}{3}, -\tfrac{2}{3}, \ldots$$

to cover all of $\mathbb{Q}$.

Hence, $\mathbb{Q}$ is countable. □

Remark. This construction illustrates that while $\mathbb{Q}$ is dense in $\mathbb{R}$, it can still be arranged into a list. In contrast, $\mathbb{R}$ is uncountable by Cantor's diagonal argument.

Sequences of Sets and Set Limits

Sequences of Sets

Definition 1.4 (Monotone Sequences of Sets)**.** *Consider that*

- ***Increasing:*** $A_1 \subseteq A_2 \subseteq \ldots$. *Then*
$$\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n.$$
- ***Decreasing:*** $A_1 \supseteq A_2 \supseteq \ldots$. *Then*
$$\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n.$$

Limit Inferior and Limit Superior of Sets

Definition 1.5 (Limit Inferior of Sets)**.** *Given a sequence of sets $(A_n)_{n\geq 1}$, the limit inferior is*

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m = \{\, x : \exists n \text{ s.t. } \forall m \geq n,\ x \in A_m \,\}.$$

This is the set of points that eventually belong to all A_m, i.e. belong to all but finitely many sets.

Definition 1.6 (Limit Superior of Sets)**.** *The limit superior is*

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \{\, x : x \in A_m \text{ for infinitely many } m \,\}.$$

This is the set of points that occur infinitely often in the sequence.

Remark. Keep in mind that

- $\liminf_{n\to\infty} A_n \subseteq \limsup_{n\to\infty} A_n$ always holds.
- $\bigcap_{m=n}^{\infty} A_m$ is increasing in n, while $\bigcup_{m=n}^{\infty} A_m$ is decreasing in n.

Definition 1.7 (Limit of Sets)**.** *If* $\liminf A_n = \limsup A_n$*, we say the sequence* (A_n) *converges, and define*

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

Example. Let

$$A_{2n} = \left[\tfrac{1}{n},\, 7 - \tfrac{1}{n}\right], \qquad A_{2n+1} = \left(-\tfrac{1}{n},\, 7 + \tfrac{1}{n}\right).$$

Then

$$\liminf_{n\to\infty} A_n = (0, 7), \qquad \limsup_{n\to\infty} A_n = [0, 7].$$

Remark. The operations lim inf and lim sup for sets should not be confused with lim inf and lim sup for real sequences. For sets, they capture eventual and infinitely-often membership, not numerical bounds.

Algebras and σ-Algebras

Motivation. Probability theory formalizes randomness through a measure space $(\Omega, \mathcal{F}, P)$. The role of $\mathcal{F}$ is subtle yet foundational: it encodes *what events are observable*. We therefore begin with the structures that specify which subsets of Ω are legitimate "events."

Power Set and Algebras of Sets

Definition 1.8 (Power Set)**.** *For a given sample space* Ω*, the power set*

$$2^{\Omega} := \{A : A \subseteq \Omega\}$$

is the collection of all subsets of Ω*. It represents the largest possible collection of events.*

Example. If $\Omega = \{1, 2, 3\}$, then

$$2^{\Omega} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

When $|\Omega| = n$, the cardinality satisfies $|2^{\Omega}| = 2^n$.

Definition 1.9 (Algebra of Sets)**.** *A collection* $\mathcal{A} \subseteq 2^{\Omega}$ *is called an algebra on* Ω *if:*

1. $\Omega \in \mathcal{A}$,
2. $A \in \mathcal{A} \implies A^c \in \mathcal{A}$,
3. $A, B \in \mathcal{A} \implies A \cup B \in \mathcal{A}$.

Remark. An algebra is closed under complements and finite unions, hence also under finite intersections. It can be viewed as a "mini-universe" of sets stable under finite Boolean operations. However, it is too small to handle countable operations — a crucial feature for probability limits and random sequences.

Definition 1.10 (σ-Algebra)**.** *An algebra* $\mathcal{A}$ *is called a* σ*-algebra if it is closed under countable unions:*

$$A_1, A_2, \cdots \in \mathcal{A} \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}.$$

Remark. Closure under countable unions (and complements) implies closure under countable intersections via De Morgan's laws. A σ-algebra is thus a mathematically consistent framework for countable limiting operations — the heart of measure theory.

Example (Borel σ-algebra)**.**On $\mathbb{R}$, define

$$\mathcal{B}(\mathbb{R}) := \sigma\big(\{(a,b) : a < b\}\big),$$

the smallest σ-algebra containing all open intervals. It is the canonical domain for defining Lebesgue and probability measures on $\mathbb{R}$.

Constructing σ-Algebras

Goal. Given a primitive collection of sets (e.g. intervals, rectangles, or events), how do we enlarge it minimally to a full σ-algebra suitable for measure definition?

We next introduce three key structures—π-systems, monotone classes, and Dynkin systems—whose interplay underlies Carathéodory's extension theorem and the uniqueness theorems of measure theory.

π-systems, Monotone Classes, and Dynkin Systems

Definition 1.11 (π-system)**.** *A collection $\mathcal{A}$ of subsets of Ω is called a π-system if it is closed under finite intersections:*

$$A, B \in \mathcal{A} \implies A \cap B \in \mathcal{A}.$$

Definition 1.12 (Monotone Class)**.** *A collection $\mathcal{M}$ of subsets of Ω is a monotone class if:*

- *$\forall$ increasing sequences (A_n) in $\mathcal{M}$, $\bigcup_{n=1}^{\infty} A_n \in \mathcal{M}$;*
- *$\forall$ decreasing sequences (A_n) in $\mathcal{M}$, $\bigcap_{n=1}^{\infty} A_n \in \mathcal{M}$.*

Definition 1.13 (Dynkin System (or λ-system))**.** *A collection $\mathcal{D}$ of subsets of Ω is a Dynkin system if:*

1. *$\Omega \in \mathcal{D}$,*
2. *$A, B \in \mathcal{D}$ with $A \subseteq B$ implies $B \setminus A \in \mathcal{D}$,*
3. *If (A_n) are disjoint and each $A_n \in \mathcal{D}$, then $\bigcup_n A_n \in \mathcal{D}$.*

Remark. Dynkin systems generalize σ-algebras by relaxing closure under arbitrary intersections. A λ-system that is also a π-system is automatically a σ-algebra, since closure under finite intersections plus disjoint countable unions yields full countable union closure.

Generated σ-Algebra and Minimal Closure Principles

Definition 1.14 (Generated σ-Algebra)**.** *For any collection $\mathcal{A} \subseteq 2^{\Omega}$, the σ-algebra generated by $\mathcal{A}$ is*

$$\sigma(\mathcal{A}) := \bigcap \{\mathcal{B} : \mathcal{A} \subseteq \mathcal{B},\ \mathcal{B} \text{ is a } \sigma\text{-algebra}\}.$$

It is the smallest σ-algebra containing $\mathcal{A}$.

Example. Let $\Omega = \{1,2,3,4\}$ and $\mathcal{A} = \{\{1\}\}$. Then $\sigma(\mathcal{A}) = \{\emptyset, \{1\}, \{2,3,4\}, \Omega\}$.

Example (Borel σ-algebra)**.**Let $\mathcal{A} = \{(a,b) : a < b\}$ on $\mathbb{R}$. Then $\sigma(\mathcal{A}) = \mathcal{B}(\mathbb{R})$, the domain of the Lebesgue measure.

Filtrations: Growing Information Over Time

Example (Filtration Generated by Coin Tosses)**.**Let $\Omega = \{X = (x_1, x_2, \dots) : x_i \in \{0,1\}\}$. Define $\mathcal{F}_n =$ the information after n coin flips:

$$\mathcal{F}_1 = \sigma(x_1), \quad \mathcal{F}_2 = \sigma(x_1, x_2), \quad \dots$$

Concretely,

$$\mathcal{F}_1 = \{\{X : x_1 = 0\}, \{X : x_1 = 1\}, \emptyset, \Omega\}.$$

Each $\mathcal{F}_n$ refines the previous one, and $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$.

Remark. This increasing family $(\mathcal{F}_n)$ is called a *filtration*. It models how information is progressively revealed — the foundation of martingale theory and stochastic processes.

Remark (Some Insight)**.**Generated σ-algebras are ubiquitous in probability:

- $\sigma(X)$ denotes the information revealed by a random variable X.
- $\sigma(X_1, \ldots, X_n)$ captures the joint information of a vector.
- Filtrations $(\mathcal{F}_t)_{t \geq 0}$ describe information flow in time.

Thus, the notion of "generated σ-algebra" underlies both static and dynamic probability models.

Probability Space and Basic Properties

Definition 1.15 (Probability Space)**.** *A probability space is a triple $(\Omega, \mathcal{F}, P)$ where:*

- Ω — *the sample space;*
- $\mathcal{F}$ — *a σ-algebra of events;*
- $P : \mathcal{F} \to [0, 1]$ — *a measure satisfying:*
  1. $P(\Omega) = 1$*;*
  2. $P(A) \geq 0$ *for all* $A \in \mathcal{F}$*;*
  3. *(Countable additivity) for disjoint* A_i*,*

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

Remark (Basic Properties)**.** 1. $P(\emptyset) = 0$, $P(A^c) = 1 - P(A)$.

2. (Inclusion–Exclusion) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
3. (Subadditivity) $P(\bigcup_i A_i) \leq \sum_i P(A_i)$.
4. (Continuity from below/above)

$$A_n \uparrow A \implies P(A_n) \uparrow P(A), \quad A_n \downarrow A \implies P(A_n) \downarrow P(A).$$

Historical Note and Perspective

Remark (Historical Remark)**.**The concept of σ-algebra arose from Lebesgue's 1901 work on integration, and was later formalized by Kolmogorov (1933) as the basis for probability axioms. The π–λ theorem and monotone class theorem, introduced by Dynkin and Doob, form the bridge between algebraic structure and measure extension — a central pillar of modern stochastic analysis.

Probability Measure and Probability Space

Probability Measures

Definition 1.16 (Probability Measure)**.** *A function $P : \mathcal{A} \to [0, 1]$ on a σ-algebra $\mathcal{A}$ is a probability measure if*

1. $P(\Omega) = 1$*,*
2. *(Countable additivity) For disjoint* $A_i \in \mathcal{A}$*,*

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

Definition 1.17 (Probability Space)**.** *A triple $(\Omega, \mathcal{A}, P)$ is a probability space, where*

- Ω *is the sample space,*
- $\mathcal{A}$ *is a σ-algebra of events,*
- P *is a probability measure.*

Properties of Probability

Proposition 1.1 (Basic Properties)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Then for all $A, B, A_i \in \mathcal{F}$:*

(i) $P(\emptyset) = 0, \quad P(A^c) = 1 - P(A)$.

(ii) If $A \subseteq B$, then $P(A) \leq P(B)$.

(iii) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

(iv) (Boole's inequality)

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) \leq \sum_{i=1}^{\infty} P(A_i).$$

Proof. (i) Since $\Omega = A \cup A^c$ (disjoint), we have

$$P(\Omega) = P(A) + P(A^c) = 1,$$

hence $P(A^c) = 1 - P(A)$. Taking $A = \Omega$ gives $P(\emptyset) = 0$.

(ii) If $A \subseteq B$, then $B = A \cup (B \setminus A)$ disjoint. Thus

$$P(B) = P(A) + P(B \setminus A) \geq P(A).$$

(iii) Trick: decompose $A \cup B$ into disjoint parts. Write

$$A \cup B = (A \setminus B) \dot{\cup} (B \setminus A) \dot{\cup} (A \cap B).$$

Therefore

$$P(A \cup B) = P(A \setminus B) + P(B \setminus A) + P(A \cap B).$$

But $P(A) = P(A \setminus B) + P(A \cap B)$ and $P(B) = P(B \setminus A) + P(A \cap B)$. Adding yields the inclusion–exclusion identity:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

(iv) (Boole's inequality). The spirit of this proof is a *disjointification trick*. Define

$$B_1 = A_1, \qquad B_2 = A_2 \setminus A_1, \qquad B_3 = A_3 \setminus (A_1 \cup A_2), \quad \ldots$$

Then the B_i are disjoint, and

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i.$$

Hence by countable additivity,

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(B_i).$$

Since $B_i \subseteq A_i$, we have $P(B_i) \leq P(A_i)$. Thus

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) \leq \sum_{i=1}^{\infty} P(A_i).$$

□

Remark (Spirit of the Trick)**.**is

- The inclusion–exclusion formula relies on *splitting overlapping sets into disjoint pieces*. This avoids over-counting.
- Boole's inequality relies on constructing a *disjoint cover* B_i of $\cup_i A_i$ by "peeling off" previously counted parts. This is a standard probability trick: make things disjoint to apply additivity.

Theorem 1.4 (Continuity from Below and Above)**.** *Let $\{A_n\}$ be a monotone sequence of sets.*

1. *If* $A_1 \subseteq A_2 \subseteq \dots$, *then*

$$\lim_{n\to\infty} P(A_n) = P\Big(\bigcup_{n=1}^{\infty} A_n\Big).$$

2. *If* $A_1 \supseteq A_2 \supseteq \dots$, *then*

$$\lim_{n\to\infty} P(A_n) = P\Big(\bigcap_{n=1}^{\infty} A_n\Big).$$

Proof Idea. For (i), define disjoint increments $B_1 = A_1$, $B_{n+1} = A_{n+1} \setminus A_n$. Then $A = \cup_n A_n = \cup_n B_n$ disjoint, and

$$P(A_n) = \sum_{k=1}^{n} P(B_k), \quad P(A) = \sum_{k=1}^{\infty} P(B_k).$$

Thus $\lim_n P(A_n) = P(A)$. For (ii), apply the result to complements: A_n^c is increasing, and use $P(A) = 1 - P(A^c)$. □

Proposition 1.2 (Liminf and Limsup Inequalities)**.** *For any sequence* $\{A_n\} \subseteq \mathcal{F}$,

$$P(\liminf A_n) \le \liminf_{n\to\infty} P(A_n) \le \limsup_{n\to\infty} P(A_n) \le P(\limsup A_n).$$

If $\lim A_n$ *exists (i.e.* $\liminf A_n = \limsup A_n$*), then*

$$\lim_{n\to\infty} P(A_n) = P(\lim A_n).$$

Spirit. This result captures the *compatibility of limits and probability*. - The trick: rewrite $\liminf A_n$ and $\limsup A_n$ using $\cap\cup$ and $\cup\cap$, then apply continuity from below/above. - The inequality arises because $\liminf A_n \subseteq$ "eventually in A_n" $\subseteq \limsup A_n$. □

Countable vs. Uncountable Sets in Probability Spaces

Example (Countable Probability Space)**.**Let $\Omega = \mathbb{N}$, $\mathcal{F} = 2^{\mathbb{N}}$, and define

$$P(\{k\}) = 2^{-k}, \qquad k \in \mathbb{N}.$$

Since $\sum_{k=1}^{\infty} 2^{-k} = 1$, this defines a probability measure. For instance, the probability of choosing an even number is

$$P(\{\text{even}\}) = \sum_{n=1}^{\infty} P(\{2n\}) = \sum_{n=1}^{\infty} 2^{-2n} = \frac{1}{4} \cdot \frac{1}{1-\frac{1}{4}} = \tfrac{1}{3}.$$

Example.(Lebesgue Measure on $[0,1]$) Let $\Omega = [0,1]$, $\mathcal{F}$ the Borel σ-algebra, and $P((a,b)) = b-a$. Then for any point $x \in [0,1]$,

$$P(\{x\}) = \lim_{n\to\infty} P\big((x - \tfrac{1}{n}, x + \tfrac{1}{n}) \cap [0,1]\big) = 0.$$

Thus every singleton has measure zero.

More generally, if $A = \{a_n\}_{n=1}^{\infty} \subseteq [0,1]$ is countable, then

$$P(A) = \sum_{n=1}^{\infty} P(\{a_n\}) = 0.$$

In particular,

$$P(\mathbb{Q} \cap [0,1]) = 0.$$

Remark (Countable vs. Uncountable Additivity)**.**While countable sets in $[0,1]$ always have probability zero, the uncountable union

$$[0,1] = \bigcup_{x\in[0,1]} \{x\}$$

satisfies

$$P([0,1]) = 1 \quad \neq \quad \sum_{x\in[0,1]} P(\{x\}) = 0.$$

This illustrates why probability measures are *countably additive*, but not uncountably additive.

Conditioning and Independence

Conditional Probability

Definition 1.18 (Conditional Probability)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. For $A, B \in \mathcal{F}$ with $P(B) > 0$, the conditional probability of A given B is*

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Independence of Events

Definition 1.19 (Independence)**.** *Two events $A, B \in \mathcal{F}$ are independent if*

$$P(A \cap B) = P(A)P(B),$$

equivalently, $P(A \mid B) = P(A)$ (when $P(B) > 0$) and $P(B \mid A) = P(B)$ (when $P(A) > 0$).

Remark. If $P(A) = 0$ or $P(A) = 1$, then A is independent of every $B \in \mathcal{F}$.

Proposition 1.3 (Closure Properties)**.** *If A and B are independent, then so are:*

$$A \text{ and } B^c, \qquad A^c \text{ and } B, \qquad A^c \text{ and } B^c.$$

Proof. Suppose $P(A \cap B) = P(A)P(B)$. Then

$$P(A \cap B^c) = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B^c).$$

Similarly for the other cases. □

Independence of Families of Events

Definition 1.20 (Mutual Independence)**.** *A collection $\{A_k : 1 \le k \le n\}$ is independent if*

$$P\left(\bigcap_{k=1}^{m} A_{i_k}\right) = \prod_{k=1}^{m} P(A_{i_k}), \quad \forall\, 1 \le i_1 < i_2 < \cdots < i_m \le n.$$

Definition 1.21 (Pairwise Independence)**.** *A collection $\{A_k : 1 \le k \le n\}$ is pairwise independent if*

$$P(A_i \cap A_j) = P(A_i)P(A_j), \quad \forall i \ne j.$$

Remark. Mutual independence $\implies$ pairwise independence, but the converse is not true. Pairwise independence is strictly weaker.

Example (Pairwise Independent but not Mutually Independent)**.** Consider two fair coin flips with sample space $\Omega = \{00, 01, 10, 11\}$ and uniform probability $P(x) = 1/4$. Define events:

$A = \{01, 11\}$ ("2nd coin is head"), $B = \{10, 11\}$ ("1st coin is head"), $C = \{01, 10\}$ ("exactly one head").

We check:

$$P(A \cap B) = P(\{11\}) = \tfrac{1}{4} = P(A)P(B), \quad P(A \cap C) = \tfrac{1}{4} = P(A)P(C), \quad P(B \cap C) = \tfrac{1}{4} = P(B)P(C).$$

So (A, B, C) are pairwise independent. However,

$$P(A \cap B \cap C) = 0 \ne \tfrac{1}{4} = P(A)P(B)P(C),$$

so they are not mutually independent.

Definition 1.22 (Countable Independence)**.** *A countable family $\{A_k : k \ge 1\}$ is independent if every finite subfamily is independent.*

Law of Total Probability

Definition 1.23 (Partition)**.** *A collection* $\{H_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ *is a partition of* Ω *if*

$$H_i \cap H_j = \emptyset \quad (i \neq j), \qquad \bigcup_{i=1}^{\infty} H_i = \Omega.$$

Theorem 1.5 (Law of Total Probability)**.** *If* $\{H_i\}_{i=1}^{\infty}$ *is a partition with* $P(H_i) > 0$*, then for any* $A \in \mathcal{F}$*,*

$$P(A) = \sum_{i=1}^{\infty} P(A \mid H_i) P(H_i).$$

Proof. We can write $A = \bigcup_{i=1}^{\infty} (A \cap H_i)$, a disjoint union. Then

$$P(A) = \sum_{i=1}^{\infty} P(A \cap H_i) = \sum_{i=1}^{\infty} P(A \mid H_i) P(H_i).$$

□

Example. Roll a fair die, and then flip a coin as many times as the die shows. Let $A =$ "total number of heads is 3." Partition by die outcome: $H_i = \{\text{die shows } i\}$.

$$P(A \mid H_i) = \begin{cases} 0, & i < 3, \\ \binom{i}{3} \left(\frac{1}{2}\right)^i, & i \geq 3. \end{cases}$$

Since $P(H_i) = 1/6$,

$$P(A) = \sum_{i=3}^{6} \binom{i}{3} \left(\tfrac{1}{2}\right)^i \cdot \tfrac{1}{6}.$$

Remark (Spirit of the Law)**.**The law of total probability decomposes a complicated event A into simpler conditional pieces along a partition $\{H_i\}$. It is the foundation for *Bayes' Theorem* and for reasoning under uncertainty: probabilities are consistent across different levels of information.

2 Random Variables

Measurable Functions and Random Elements

Definition 2.1 (Random Element)**.** *Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ be measurable spaces. A mapping $X : \Omega \to S$ is called a random element. If $S = \mathbb{R}$ with $\mathcal{S} = \mathcal{B}(\mathbb{R})$, then X is a random variable.*

Definition 2.2 (Measurable Function)**.** *$X : \Omega \to \mathbb{R}$ is measurable if*

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F} \quad \forall A \in \mathcal{B}(\mathbb{R}).$$

Example (Indicator Function)**.**If $A \in \mathcal{F}$, then

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

is a random variable.

Proposition 2.1 (Closure)**.** *If $X_1, \ldots, X_n$ are random variables and $g : \mathbb{R}^n \to \mathbb{R}$ is Borel measurable, then $g(X_1, \ldots, X_n)$ is a random variable.*

Proposition 2.2 (Limits)**.** *If $X_n \to X$ almost surely and each X_n is a random variable, then X is a random variable.*

Equivalent Random Variables and Measurability

Definition 2.3 (Equivalent Random Variables)**.** *Two random variables $X, Y : \Omega \to \mathbb{R}$ are called equivalent if*

$$P(X = Y) = 1 \quad \textit{equivalently,} \quad P(X \neq Y) = 0.$$

In this case, X and Y are indistinguishable under P.

Definition 2.4 (Distribution of a Random Variable)**.** *Any measurable random variable $X : (\Omega, \mathcal{F}, P) \to \mathbb{R}$ induces a probability measure μ_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by*

$$\mu_X(A) = P(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\}), \quad A \in \mathcal{B}(\mathbb{R}).$$

This μ_X is called the distribution of X.

Definition 2.5 (Equal in Distribution)**.** *Two random variables X and Y (possibly on different probability spaces) are said to be equal in distribution, written $X \overset{d}{=} Y$, if they induce the same law:*

$$\mu_X = \mu_Y.$$

Example. Let $\Omega = (0, 1]$ with Lebesgue measure. Define

$$X = 1_{(0, \frac{1}{2}]}, \qquad Y = 1_{(\frac{1}{2}, 1]}.$$

Then X and Y do not satisfy $P(X = Y) = 1$ (in fact, $P(X = Y) = 0$). However, both are Bernoulli(1/2) random variables, hence $X \overset{d}{=} Y$.

Theorem 2.1 (Function Composition)**.** *If $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is measurable and $g : \mathbb{R} \to \mathbb{R}$ is Borel-measurable, then the composition*

$$g \circ X : \Omega \to \mathbb{R}$$

is a random variable.

Proof. Let $A \in \mathcal{B}(\mathbb{R})$. Then

$$\{\omega : g(X(\omega)) \in A\} = X^{-1}(g^{-1}(A)) \in \mathcal{F},$$

since $g^{-1}(A) \in \mathcal{B}(\mathbb{R})$ and X is measurable. Hence $g(X)$ is measurable. □

Remark. This shows that applying measurable transformations (e.g. sin, exp, log, polynomials) to random variables always yields another random variable.

Proposition 2.3 (Algebra of Measurable Functions)**.** *Let X, Y be random variables. Then:*

- *$X+Y$, $X-Y$, and cX (for $c \in \mathbb{R}$) are random variables.*
- *XY is a random variable.*
- *$\max\{X,Y\}$ and $\min\{X,Y\}$ are random variables.*

Sketch. For sums: $\{X+Y<t\}$ can be expressed as a countable union of intersections involving $\{X<q\}, \{Y<t-q\}$ with $q \in \mathbb{Q}$, hence measurable. For products: express $\{XY<t\}$ via rational bounds depending on the signs of X, Y. For max/min: note that

$$\max\{X,Y\} = \tfrac{1}{2}(X+Y+|X-Y|), \quad \min\{X,Y\} = \tfrac{1}{2}(X+Y-|X-Y|).$$

All are measurable as they are compositions of measurable maps. □

Theorem 2.2 (Sup/Inf of Random Variables)**.** *Let $\{X_n\}_{n\geq 1}$ be random variables. Then*

$$\sup_n X_n, \quad \inf_n X_n, \quad \limsup_{n\to\infty} X_n, \quad \liminf_{n\to\infty} X_n$$

are all random variables.

Proof. For $\sup_n X_n$:

$$\{\sup_n X_n \leq t\} = \bigcap_{n=1}^{\infty} \{X_n \leq t\},$$

which is measurable as a countable intersection of measurable sets. Similarly for $\inf_n$. For lim sup:

$$\{\limsup_{n\to\infty} X_n \leq t\} = \bigcap_{m=1}^{\infty} \bigcup_{n\geq m} \{X_n \leq t\},$$

which is measurable. Analogous for lim inf. □

Remark. The constructions of sup, inf, lim sup, and lim inf are all "pointwise" operations, hence measurability follows from closure of $\mathcal{F}$ under countable unions/intersections. This highlights an important principle: *measurability is stable under pointwise limits and algebraic operations*.

Limsup, Liminf, and Measurability of Limits

Definition 2.6 (Limsup and Liminf of Numerical Sequences)**.** *Let $\{a_n\}_{n\in\mathbb{N}}$ be a sequence of real numbers. We define*

$$\limsup_{n\to\infty} a_n = \inf_{m\geq 1} \sup_{n\geq m} a_n, \qquad \liminf_{n\to\infty} a_n = \sup_{m\geq 1} \inf_{n\geq m} a_n.$$

Remark. Intuitively:

- lim sup tracks the "eventual upper envelope" of the sequence — the smallest ceiling that still contains infinitely many terms.
- lim inf tracks the "eventual lower envelope" — the largest floor that still contains infinitely many terms.

It follows that $\liminf a_n \leq \limsup a_n$ always, and if they coincide, the common value is the limit of the sequence.

Theorem 2.3 (Measurability of lim sup and lim inf)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of measurable random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Then*

$$\limsup_{n\to\infty} X_n(\omega), \qquad \liminf_{n\to\infty} X_n(\omega)$$

are measurable random variables.

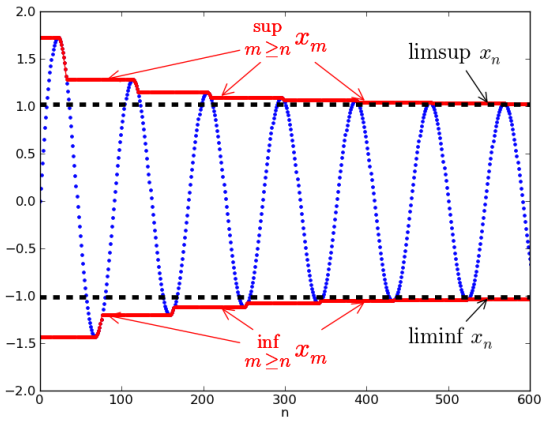

Figure 1: An illustration of limit superior and limit inferior.

Proof. Fix $\omega \in \Omega$. Define

$$Y_m(\omega) = \sup_{n \geq m} X_n(\omega).$$

Each Y_m is measurable because measurability is preserved under sup of a countable family. Then

$$\limsup_{n \to \infty} X_n(\omega) = \inf_{m \geq 1} Y_m(\omega),$$

which is measurable as an infimum of measurable functions. The case of $\liminf$ follows similarly. □

Remark (Alternative Characterization)**.**A useful identity is:

$$\{\limsup_{n \to \infty} X_n \leq x\} = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} \{X_n \leq x\}.$$

This set belongs to $\mathcal{F}$ because it is built from countable unions and intersections of measurable sets. This alternative description is often more practical when proving measurability.

Theorem 2.4 (Measurability of Limits When They Exist)**.** *Let $\{X_n\}$ be measurable random variables. Suppose*

$$\lim_{n \to \infty} X_n(\omega)$$

exists for all ω in a set of probability one (i.e., $\limsup X_n = \liminf X_n$ almost surely). Then

$$\lim_{n \to \infty} X_n$$

is a measurable random variable.

Remark (Spirit of the Result)**.**This theorem emphasizes an important principle: the pointwise limit of measurable random variables is again measurable (at least almost surely). This is essential for probability theory, since random variables are often defined as limits of simpler approximations (e.g., in the construction of expectations, martingales, or stochastic processes).

The deeper idea is that measurability is robust under limits — which is why probability theory works so well with infinite sequences and limiting arguments.

Sigma-Algebra Generated by a Random Variable

Definition 2.7 (Sigma-Algebra Generated by a Random Variable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a measurable random variable. The sigma-algebra generated by X, denoted $\sigma(X)$, is defined by*

$$\sigma(X) = \{X^{-1}(A) : A \in \mathcal{B}(\mathbb{R})\},$$

where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-algebra on $\mathbb{R}$. Equivalently, $\sigma(X)$ is the smallest σ-algebra with respect to which X is measurable.

Example (Flipping Two Coins)**.**Consider $\Omega = \{(0,0),(0,1),(1,0),(1,1)\}$ with $\mathcal{F} = 2^{\Omega}$ and $X(\omega)$ defined as the total number of heads:

$$X(0,0) = 0, \quad X(0,1) = X(1,0) = 1, \quad X(1,1) = 2.$$

Then

$$\sigma(X) = \big\{\emptyset,\ \Omega,\ \{(0,0)\},\ \{(1,1)\},\ \{(0,1),(1,0)\},\ \{(0,0),(0,1),(1,0)\},\ \{(0,1),(1,0),(1,1)\}\big\}.$$

Remark (Interpretation)**.**Two sample points $\omega_1, \omega_2 \in \Omega$ are *indistinguishable with respect to X* if $X(\omega_1) = X(\omega_2)$. In that case, for every $A \in \sigma(X)$, either both ω_1, ω_2 belong to A or neither does. This captures the idea that $\sigma(X)$ contains *exactly the information that can be revealed by knowing the value of X*. It "forgets" everything else about the underlying ω.

Distribution and Distribution Functions

Definition 2.8 (Distribution of a Random Variable)**.** *The distribution of a random variable $X : \Omega \to \mathbb{R}$ is the probability measure μ_X on $\mathcal{B}(\mathbb{R})$ defined by*

$$\mu_X(A) = \mathbb{P}(X \in A), \qquad \forall A \in \mathcal{B}(\mathbb{R}).$$

Definition 2.9 (Distribution Function)**.** *The distribution function $F_X : \mathbb{R} \to [0,1]$ of X is given by*

$$F_X(x) = \mu_X((-\infty, x]) = \mathbb{P}(X \le x).$$

Remark (Properties of Distribution Functions)**.**The distribution function F_X satisfies:

- F_X is non-decreasing.
- F_X is right-continuous: $\lim_{\varepsilon \downarrow 0} F_X(x+\varepsilon) = F_X(x)$.
- $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to +\infty} F_X(x) = 1$.
- The jump of F_X at x equals the probability mass at x:

$$\mathbb{P}(X = x) = F_X(x) - \lim_{\varepsilon \downarrow 0} F_X(x - \varepsilon).$$

Remark (Interpretation)**.**The distribution μ_X "pushes forward" the probability measure $\mathbb{P}$ from Ω to $\mathbb{R}$. In this sense, μ_X tells us *everything probabilistic about X*, independently of the underlying sample space.

The function F_X is simply the cumulative form of this measure. Right-continuity arises because $(-\infty, x]$ are the basic generators of the Borel σ-algebra, and monotonicity comes from set inclusion. The jumps of F_X correspond to atoms of the distribution — this explains why discrete distributions appear as step functions.

Types of Random Variables

Definition 2.10 (Simple Random Variable)**.** $X(\omega) = \sum_{k=1}^{n} x_k 1_{A_k}(\omega)$ *with $x_k \in \mathbb{R}$, $A_k \in \mathcal{F}$, $\{A_k\}$ a measurable partition of Ω.*

Definition 2.11 (Equivalent Random Variables)**.** *X and Y are equivalent if $P(X = Y) = 1$, i.e. they differ only on a null set. Notation: $X \overset{a.s.}{=} Y$.*

Lemma 2.1 (Approximation Lemma)**.** *Every nonnegative random variable X can be approximated by an increasing sequence of simple random variables $X_n \uparrow X$ pointwise.*

Distribution of a Random Variable

Definition 2.12 (Induced Measure)**.** *Each random variable $X : \Omega \to \mathbb{R}$ induces a probability measure μ_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by*

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

Definition 2.13 (Distribution Function)**.** *The cumulative distribution function (CDF) of X is*

$$F_X(x) = P(X \le x).$$

Proposition 2.4 (Properties of CDF)**.** *1. F_X is non-decreasing.*

2. *F_X is right-continuous.*
3. $\lim_{x\to-\infty} F_X(x) = 0, \quad \lim_{x\to\infty} F_X(x) = 1.$
4. $P(a < X \le b) = F_X(b) - F_X(a).$
5. $P(X = x) = F_X(x) - \lim_{y\uparrow x} F_X(y).$

Remark. Every probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ corresponds uniquely to a distribution function, and vice versa.

Definition 2.14 (Equality in Distribution)**.** *$X \overset{d}{=} Y$ if $F_X = F_Y$ (equivalently $\mu_X = \mu_Y$), even if X, Y live on different spaces.*

4 Types of Distributions

- **Discrete:** $P(X = x_k) > 0$ on a countable set $\{x_k\}$.
- **Continuous:** $P(X = x) = 0$ for all x.
- **Absolutely continuous:** $F_X(x) = \int_{-\infty}^{x} f(t)\,dt$ for a density f.
- **Singular continuous:** F continuous, $F'(x) = 0$ a.e., e.g. Cantor distribution.

Cantor Function as Counterexample

Example (Cantor Function)**.**The Cantor function $C : [0,1] \to [0,1]$ is continuous, non-decreasing, and satisfies

$$C(0) = 0, \qquad C(1) = 1.$$

It has the following properties:

- $C'(x) = 0$ for almost every $x \in [0,1]$ (with respect to Lebesgue measure).
- C is not constant; in fact, it strictly increases along the Cantor set.
- C is not absolutely continuous, since

$$C(x) - C(0) \neq \int_0^x C'(t)\,dt,$$

the right-hand side being identically zero.

Therefore the Cantor function illustrates that *continuity and differentiability a.e. are not enough* to guarantee absolute continuity. Its associated measure is the **Cantor distribution**, which is a prime example of a *singular continuous measure*.

Absolute Continuity and The Fundamental Theorem of Calculus

Theorem 2.5 (Fundamental Theorem of Calculus for Lebesgue Integrals [Folland Thm.3.35)**.** *Let $-\infty < a < b < \infty$ and $F : [a,b] \to \mathbb{C}$. The following are equivalent:*

(a) F is absolutely continuous on $[a,b]$.

(b) There exists $f \in L^1([a,b])$ such that

$$F(x) - F(a) = \int_a^x f(t)\,dt, \qquad \forall x \in [a,b].$$

(c) F is differentiable almost everywhere on $[a,b]$, with $F' \in L^1([a,b])$, and

$$F(x) - F(a) = \int_a^x F'(t)\,dt, \qquad \forall x \in [a,b].$$

Remark (Interpretation)**.**This theorem characterizes absolute continuity in three equivalent ways:

- **Definition-based:** F is absolutely continuous if small total length of intervals implies small total variation of F.
- **Integral form:** Absolutely continuous functions are precisely those which can be written as indefinite Lebesgue integrals of some L^1 function.
- **Derivative form:** They are differentiable almost everywhere, their derivative belongs to L^1, and the classical fundamental theorem of calculus holds in the Lebesgue sense.

Thus, absolute continuity identifies the "good" class of functions where differentiation and integration are perfectly compatible.

Common Distributions

Example (Discrete)**.**

- Bernoulli(p): $P(X = 1) = p$, $P(X = 0) = 1 - p$.
- Binomial(n, p): $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.
- Geometric(p): $P(X = k) = (1-p)^{k-1} p$, $k \geq 1$.
- Poisson(λ): $P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$.

Example (Absolutely Continuous)**.**

- Uniform(a, b): $f(x) = \frac{1}{b-a}$ on (a, b).
- Exponential(λ): $f(x) = \lambda e^{-\lambda x}$ on $[0, \infty)$.
- Normal(μ, σ^2): $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$.

Remark (Memoryless Property)**.**Among all discrete distributions, the *Geometric* distribution is the only one with the memoryless property:

$$P(X > m + n \mid X > m) = P(X > n).$$

Among continuous distributions, the unique memoryless law is the *Exponential* distribution:

$$P(X > t + s \mid X > t) = P(X > s).$$

Remark. Convolutions describe sums of independent random variables:

$$F_{X+Y} = F_X * F_Y, \qquad f_{X+Y} = f_X * f_Y.$$

If one variable has a density, then the sum does as well.

Joint Distributions and Random Vectors

Definition 2.15 (Random Vector)**.** *An n-dimensional random vector is*

$$X = (X_1, X_2, \ldots, X_n)^\top,$$

where each X_k is a random variable.

Definition 2.16 (Joint Distribution Function)**.** *For $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, the joint distribution of X is*

$$F_X(x) = P(X \leq x) = P(X_1 \leq x_1, \ldots, X_n \leq x_n).$$

- **Discrete case:** Characterized by the joint probability function

$$p_X(x) = P(X = x).$$

- **Absolutely continuous case:** Characterized by the joint density

$$f_X(x) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F_X(x).$$

Definition 2.17 (Marginal Distribution)**.** *If $X = (X, Y)$ with joint density $f_{X,Y}$, the marginal density of X is*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy.$$

Remark. Marginals describe lower-dimensional components of a joint distribution, but they do not in general determine the joint distribution uniquely.

3 Expectation: Definition, Properties, and Convergence

Definition of Expectation

Definition 3.1 (Expectation of a Random Variable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a measurable random variable. The expectation (or mean) of X, denoted $\mathbb{E}[X]$, is the Lebesgue integral of X with respect to $\mathbb{P}$:*

$$\mathbb{E}[X] := \int_\Omega X(\omega)\, d\mathbb{P}(\omega),$$

whenever this integral exists (possibly $= \pm\infty$).

Remark (Interpretation)**.**Expectation measures the "center of mass" of a random variable under the probability law $\mathbb{P}$. It depends only on the distribution of X:

$$\mathbb{E}[X] = \int_{\mathbb{R}} x\, \mu_X(dx), \quad \text{where } \mu_X(A) = \mathbb{P}(X \in A)$$

is the pushforward (distribution) of X.

Expectation is the measure-theoretic counterpart of integration. It provides the bridge between probability and analysis, generalizing the concept of an "average" to random quantities. We construct it systematically, starting from the simplest measurable functions and extending to all integrable random variables.

Construction from Simple to General

Step 1: Simple random variables.

Definition 3.2 (Simple function)**.** *Let $\{A_i\}_{i=1}^n$ be disjoint measurable sets and $c_i \in \mathbb{R}$. A simple random variable is a function*

$$X(\omega) = \sum_{i=1}^n c_i \mathbf{1}_{A_i}(\omega), \qquad A_i \cap A_j = \emptyset \ (i \neq j).$$

The expectation is defined as

$$\mathbb{E}[X] = \sum_{i=1}^n c_i\, \mathbb{P}(A_i).$$

- Simple functions are the most elementary measurable objects: finite-valued, piecewise constant.
- They serve as the building blocks of the Lebesgue integral, much like step functions in the Riemann integral.
- The definition $\mathbb{E}[X] = \sum c_i \mathbb{P}(A_i)$ makes expectation a probability-weighted average of the coefficients c_i.

Step 2: Nonnegative random variables.

Definition 3.3 (Nonnegative random variables)**.** *If $X(\omega) \geq 0$, then*

$$\mathbb{E}[X] = \sup\{\mathbb{E}[Y] : 0 \leq Y \leq X,\ Y \text{ simple}\}.$$

Here, $\mathbb{E}[X]$ may equal $+\infty$.

Approximation trick: construct X_n by dyadic approximation:

$$X_n(\omega) = \begin{cases} \frac{k-1}{2^n}, & \frac{k-1}{2^n} \leq X(\omega) < \frac{k}{2^n},\ k = 1, \ldots, n2^n, \\ n, & X(\omega) \geq n. \end{cases}$$

Then $X_n \uparrow X$ as $n \to \infty$, and

$$\mathbb{E}[X] = \lim_{n\to\infty} \sum_{k=1}^{n2^n} \frac{k-1}{2^n}\, \mathbb{P}\big(\tfrac{k-1}{2^n} \leq X < \tfrac{k}{2^n}\big)\,.$$

- **Density principle:** Simple functions are dense in $L^1(\mathbb{P})$ and more generally in $L^p(\mathbb{P})$ $(1 \leq p < \infty)$.

- This ensures that defining $\mathbb{E}[X]$ for simple functions uniquely determines $\mathbb{E}[X]$ for all nonnegative measurable X.
- Conceptually, this means: starting from indicators $\mathbf{1}_A$, extending to finite sums $\sum c_i \mathbf{1}_{A_i}$, and finally taking monotone limits recovers all nonnegative random variables.

Step 3: General random variables.

Definition 3.4. *Let $X : \Omega \to \mathbb{R}$ be measurable. Define the positive and negative parts:*

$$X^+ := \max(X, 0), \qquad X^- := \max(-X, 0).$$

Then $X = X^+ - X^-$ and $|X| = X^+ + X^-$.

- *If $\mathbb{E}[X^+] < \infty$ and $\mathbb{E}[X^-] < \infty$, then*

$$\mathbb{E}[X] := \mathbb{E}[X^+] - \mathbb{E}[X^-],$$

  and X is called integrable.
- *If exactly one of $\mathbb{E}[X^+], \mathbb{E}[X^-]$ is finite, we allow $\mathbb{E}[X] = \pm\infty$.*
- *If both are infinite, then $\mathbb{E}[X]$ is undefined.*

Remark. The philosophy is universal: definitions begin with simple objects (indicators, step functions, simple processes), and the density of these objects in the relevant function spaces guarantees the extension to the full generality. This density principle underlies Lebesgue integration, L^p theory, and stochastic integration alike.

Connection to Stochastic Integration.

- In stochastic calculus, the Itô integral $\int_0^T H_t \, dB_t$ is *first defined* for simple predictable processes H (piecewise constant, adapted).
- These simple processes are dense in $L^2(\Omega \times [0, T])$, so the Itô integral extends uniquely to all square-integrable predictable H.
- This mirrors the construction of expectation: start from indicators, extend by linearity, then extend by density.

Properties of Expectation

Basic Properties

Let X, Y be random variables and $c \in \mathbb{R}$. Then:

- If $\mathbb{P}(X = 0) = 1$, then $\mathbb{E}[X] = 0$.
- If $\mathbb{P}(X \geq 0) = 1$, then $\mathbb{E}[X] \geq 0$.
- If $\mathbb{E}[X] = 0$ and $X \geq 0$, then $\mathbb{P}(X = 0) = 1$.
- Linearity: $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, and $\mathbb{E}[cX] = c\mathbb{E}[X]$.
- Monotonicity: If $X \leq Y$ almost surely, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.
- If $\mathbb{P}(X = Y) = 1$, then $\mathbb{E}[X] = \mathbb{E}[Y]$.
- If $\mathbb{E}[|X - Y|] = 0$, then $\mathbb{P}(X = Y) = 1$.

Theorem 3.1 (Law of the Unconscious Statistician (LOTUS))**.** *If X has distribution μ_X and $g : \mathbb{R} \to \mathbb{R}$ is measurable, then*

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) \, \mu_X(dx).$$

In particular,

- *If X has density f_X: $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \, dx$.*
- *If X is discrete: $\mathbb{E}[g(X)] = \sum_x g(x) P(X = x)$.*

Example: Compute the Expectation of Independent Random Variables

A fundamental property is that if X and Y are independent, then

$$\mathbb{E}[XY] \;=\; \mathbb{E}[X]\cdot\mathbb{E}[Y].$$

Why is this true? Independence intuitively means that the "joint behavior" of X and Y factors into their separate distributions. Thus, the expected value of the product should factor as the product of expectations. To prove this rigorously, we must handle general measurable random variables step by step, using approximation.

We have already introduced the notion of independence for events in Unit 1, and for random variables and σ-algebras generated by it in Unit 2. Now, we introduce

Definition 3.5 (Independence of Random Variables)**.** *Given random variables $X_1,\dots,X_n$, recall that each random variable X_i generates a σ-algebra*

$$\sigma(X_i) := \{X_i^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\},$$

the collection of all measurable events determined by X_i.

We say that the random variables $X_1,\dots,X_n$ are independent if the σ-algebras $\sigma(X_1),\dots,\sigma(X_n)$ are independent. Equivalently, $X_1,\dots,X_n$ are independent if for all Borel sets $A_1,\dots,A_n \subseteq \mathbb{R}$,

$$\mathbb{P}(X_1 \in A_1,\dots,X_n \in A_n) \;=\; \prod_{i=1}^{n}\mathbb{P}(X_i \in A_i).$$

Proof of $\mathbb{E}[XY] \;=\; \mathbb{E}[X]\cdot\mathbb{E}[Y]$

Step 1. Indicator random variables. If $X = \mathbf{1}_A$ and $Y = \mathbf{1}_B$ with $A, B \in \mathcal{F}$, then by independence,

$$\mathbb{E}[XY] = \mathbb{P}(A\cap B) = \mathbb{P}(A)\mathbb{P}(B) = \mathbb{E}[X]\mathbb{E}[Y].$$

Step 2. Simple random variables. If $X = \sum_{i=1}^{n} a_i\mathbf{1}_{A_i}$ and $Y = \sum_{j=1}^{m} b_j\mathbf{1}_{B_j}$ are simple r.v.s with disjoint partitions, then

$$\mathbb{E}[XY] = \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} a_ib_j\,\mathbf{1}_{A_i\cap B_j}\right] = \sum_{i=1}^{n}\sum_{j=1}^{m} a_ib_j\,\mathbb{P}(A_i\cap B_j).$$

Independence implies $\mathbb{P}(A_i\cap B_j) = \mathbb{P}(A_i)\mathbb{P}(B_j)$. Thus

$$\mathbb{E}[XY] = \Big(\sum_{i=1}^{n} a_i\mathbb{P}(A_i)\Big)\Big(\sum_{j=1}^{m} b_j\mathbb{P}(B_j)\Big) = \mathbb{E}[X]\mathbb{E}[Y].$$

Step 3. Nonnegative random variables. For general nonnegative X, Y, choose simple approximations $X_n \uparrow X$ and $Y_n \uparrow Y$. Then by monotone convergence,

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

Step 4. General integrable random variables. For arbitrary X, Y, decompose into positive and negative parts:

$$X = X^+ - X^-, \qquad Y = Y^+ - Y^-.$$

Expanding XY gives a combination of products of nonnegative variables (e.g. X^+Y^+, X^+Y^-, etc.), each of which factors by Step 3. Recombining terms yields the general result.

Remark. The approximation strategy is essential:

- Start from the "atoms" of measurable functions (indicators).
- Extend linearly to simple functions.
- Pass to limits via monotone convergence.

- Finally handle signed functions by decomposition.

This pattern underlies much of measure-theoretic probability.

Convergence of Expectation

Convergence of Expectation

A natural question: suppose $X_n \to X$ almost surely, i.e.

$$\mathbb{P}\Big(\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\Big) = 1.$$

Do we necessarily have

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

Answer: No, not in general.

Example (Counterexample: "Skinny Tall Rectangles")**.**Let $\Omega = (0,1)$, $\mathcal{F}$ the Borel σ-algebra, and $\mathbb{P}$ the Lebesgue measure. Define

$$X_n(\omega) = \begin{cases} n, & \omega \in (0, \frac{1}{n}], \\ 0, & \omega \in (\frac{1}{n}, 1]. \end{cases}$$

Then for each n,

$$\mathbb{E}[X_n] = n \cdot \tfrac{1}{n} + 0 \cdot \Big(1 - \tfrac{1}{n}\Big) = 1.$$

On the other hand, $X_n(\omega) \to 0$ for almost every $\omega \in (0,1)$, so $X = 0$ a.s. and hence $\mathbb{E}[X] = 0$.

Thus

$$\lim_{n\to\infty} \mathbb{E}[X_n] = 1 \quad \neq \quad \mathbb{E}[X] = 0.$$

Remark. This shows that almost sure convergence does not in general imply convergence of expectations. Additional conditions (such as *dominated convergence theorem* or *monotone convergence theorem*) are required to interchange limits and expectation.

Convergence Theorems

Fatou's Lemma with a Nontrivial Proof

Theorem 3.2 (Fatou's Lemma)**.** *Let $\{X_n\}_{n\geq 1}$ be a sequence of nonnegative random variables. Then*

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

Proof. The key idea is to compare $\liminf X_n$ with a sequence of simple random variables that approximate it from below, so that we can apply bounded convergence.

Step 1. For any bounded simple random variable Y such that

$$0 \leq Y(\omega) \leq \liminf_{n\to\infty} X_n(\omega),$$

we want to show

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

Step 2. Define for each n,

$$Y_n(\omega) := \min\{\, Y(\omega), X_n(\omega) \,\}.$$

This is the "best of both worlds" construction: it guarantees

$$0 \leq Y_n(\omega) \leq Y(\omega) \quad \text{and} \quad Y_n(\omega) \leq X_n(\omega).$$

Thus, Y_n is bounded by Y, but also never exceeds X_n.

Step 3. Now define

$$Z_n(\omega) := \inf_{m\geq n} Y_m(\omega).$$

Then (Z_n) is an increasing sequence (since the index set shrinks as n grows), and by construction,

$$\lim_{n\to\infty} Z_n(\omega) = \liminf_{n\to\infty} Y_n(\omega) \;\le\; \liminf_{n\to\infty} X_n(\omega).$$

Step 4. By the monotone convergence theorem (applied to the nonnegative increasing sequence Z_n),

$$\mathbb{E}\Big[\lim_{n\to\infty} Z_n\Big] = \lim_{n\to\infty} \mathbb{E}[Z_n].$$

But since $Z_n \le Y_n \le X_n$, we get

$$\mathbb{E}[Z_n] \le \mathbb{E}[Y_n] \le \mathbb{E}[X_n].$$

Step 5. Combining these,

$$\mathbb{E}[Y] \;\le\; \lim_{n\to\infty} \mathbb{E}[Z_n] \;\le\; \liminf_{n\to\infty} \mathbb{E}[X_n].$$

Step 6. Finally, take the supremum over all such bounded simple Y with $Y \le \liminf X_n$. By the definition of the Lebesgue integral, this supremum equals $\mathbb{E}[\liminf X_n]$. Hence,

$$\mathbb{E}\Big[\liminf_{n\to\infty} X_n\Big] \;\le\; \liminf_{n\to\infty} \mathbb{E}[X_n].$$

□

Remark. The clever step is defining $Y_n = \min\{Y, X_n\}$. This ensures two things simultaneously:

- $Y_n \le Y$, so the sequence is uniformly bounded (good for convergence theorems).
- $Y_n \le X_n$, so we can compare its expectation directly with $\mathbb{E}[X_n]$.

This "best of both worlds" trick is typical in measure-theoretic proofs: construct an auxiliary sequence that inherits the best properties of both sides of the inequality.

Bounded Convergence Theorem (BCT)

Theorem 3.3 (Bounded Convergence Theorem)**.** *Suppose $X_n \to X$ almost surely, and there exists a constant $M > 0$ such that*

$$|X_n(\omega)| \le M \quad \textit{for all } n \textit{ and all } \omega.$$

Then

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

Proof sketch. We want to show $\mathbb{E}[X_n - X] \to 0$.

Fix $\delta > 0$. Note that for each ω,

$$-\delta\,\mathbf{1}_{\{|X_n - X| < \delta\}} - 2M\,\mathbf{1}_{\{|X_n - X| \ge \delta\}} \;\le\; X_n - X \;\le\; \delta\,\mathbf{1}_{\{|X_n - X| < \delta\}} + 2M\,\mathbf{1}_{\{|X_n - X| \ge \delta\}}.$$

Taking expectations and applying monotonicity,

$$-2M\,\mathbb{P}(|X_n - X| \ge \delta) - \delta\mathbb{P}(|X_n - X| < \delta) \;\le\; \mathbb{E}[X_n - X] \;\le\; 2M\,\mathbb{P}(|X_n - X| \ge \delta) + \delta.$$

Since $X_n \to X$ almost surely, we have $\mathbb{P}(|X_n - X| \ge \delta) \to 0$. Thus letting $n \to \infty$,

$$-\delta \;\le\; \liminf_{n\to\infty} \mathbb{E}[X_n - X] \;\le\; \limsup_{n\to\infty} \mathbb{E}[X_n - X] \;\le\; \delta.$$

Because $\delta > 0$ was arbitrary, $\mathbb{E}[X_n - X] \to 0$, i.e. $\mathbb{E}[X_n] \to \mathbb{E}[X]$. □

Remark (Trick:)**.**The key idea is to control $X_n - X$ using two bounds simultaneously:

- A *local bound* ($|X_n - X| < \delta$) gives a small error δ.
- A *global bound* ($|X_n - X| \ge \delta$) uses the uniform bound M and the fact that $\mathbb{P}(|X_n - X| \ge \delta) \to 0$.

By combining these via indicator functions and taking expectations, we get a bound of the form

$$|\mathbb{E}[X_n - X]| \leq 2M\,\mathbb{P}(|X_n - X| \geq \delta) + \delta,$$

which vanishes in the limit. This is a recurring theme in measure-theoretic convergence proofs.

Monotone Convergence Theorem (MCT)

Remark. It can happen that $\liminf_{n\to\infty} \mathbb{E}[X_n] = +\infty$. In this case, the conclusion of Fatou's lemma is still valid, but it provides no nontrivial information. The Monotone Convergence Theorem can be viewed as the case where the inequality of Fatou's lemma strengthens to an equality.

Theorem 3.4 (Monotone Convergence Theorem)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of nonnegative random variables with $X_n(\omega) \leq X_{n+1}(\omega)$ for all ω. If $X_n \uparrow X$ pointwise, then*

$$\lim_{n\to\infty} \mathbb{E}[X_n] \;=\; \mathbb{E}\Big[\lim_{n\to\infty} X_n\Big].$$

Proof. Monotonicity gives $\mathbb{E}[X_n] \leq \mathbb{E}[\lim_{n\to\infty} X_n]$ for each n, so

$$\limsup_{n\to\infty} \mathbb{E}[X_n] \;\leq\; \mathbb{E}\Big[\lim_{n\to\infty} X_n\Big].$$

On the other hand, Fatou's lemma yields

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

Since $\liminf X_n = \lim X_n = X$, this gives

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

Combining both inequalities shows

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

□

Remark. The MCT shows that expectations commute with monotone limits for nonnegative sequences. It is the analytic foundation of defining $\mathbb{E}[X]$ as the supremum of expectations of simple functions bounded by X. This result is so central that nearly every construction in integration theory follows the same scheme: define on simple objects, extend by monotone limits, and guarantee equality of limit and integral via MCT.

Dominated Convergence Theorem (DCT)

Theorem 3.5 (Dominated Convergence Theorem)**.** *Let $(X_n)_{n\geq 1}$ be random variables with $X_n \to X$ **almost surely**. Suppose there exists an integrable Y such that $|X_n| \leq Y$ almost surely for all n (**domination assumption**). Then X is integrable and*

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

Proof. We adapt the proof strategy using Fatou's lemma. Note first that $Y - X_n \geq 0$ and $Y + X_n \geq 0$, which satisfy the non-negative conditions for applying Fatou's lemma.

Moreover, since $X_n \to X$ a.s., we have

$$Y - X_n \;\to\; Y - X, \qquad Y + X_n \;\to\; Y + X \quad \text{almost surely.}$$

Applying Fatou's lemma to each nonnegative sequence gives

$$\mathbb{E}[Y - X] \;\leq\; \liminf_{n\to\infty} \mathbb{E}[Y - X_n], \qquad \mathbb{E}[Y + X] \;\leq\; \liminf_{n\to\infty} \mathbb{E}[Y + X_n].$$

Now use linearity of expectation. For the **first** inequality,

$$\mathbb{E}[Y - X] = \mathbb{E}[Y] - \mathbb{E}[X], \qquad \mathbb{E}[Y - X_n] = \mathbb{E}[Y] - \mathbb{E}[X_n].$$

Also recalling that $\liminf(-a_n) = -\limsup a_n$, when a_n is sequence of real numbers, thus

$$\mathbb{E}[Y] - \mathbb{E}[X] \;\le\; \liminf_{n\to\infty} \big(\mathbb{E}[Y] - \mathbb{E}[X_n]\big) = \mathbb{E}[Y] - \limsup_{n\to\infty} \mathbb{E}[X_n],$$

Cancelling $\mathbb{E}[Y]$ from both sides and rearrange

$$\mathbb{E}[X] \;\ge\; \limsup_{n\to\infty} \mathbb{E}[X_n].$$

For the second inequality, similarly it yields

$$\mathbb{E}[X] \;\le\; \liminf_{n\to\infty} \mathbb{E}[X_n].$$

Together, the two inequalities give

$$\mathbb{E}[X] \;\le\; \liminf_{n\to\infty} \mathbb{E}[X_n] \le \limsup_{n\to\infty} \mathbb{E}[X_n] \;\le\; \mathbb{E}[X].$$

Since lim sup and lim inf coincide, $\lim \mathbb{E}[X_n] = \mathbb{E}[X]$. □

Remark. The domination assumption is the critical control: it *rules out "skinny tall rectangles"—rare but arbitrarily large spikes that can prevent convergence of expectations*. The proof cleverly applies Fatou's lemma not directly to X_n but to the *modified sequences* $Y \pm X_n$, which are nonnegative and converge to $Y \pm X$. This double application yields both upper and lower bounds, forcing equality. The argument showcases how domination turns a one-sided inequality (Fatou) into a two-sided equality.

A Useful Remark on Series of Random Variables

Remark. If $\sum_{n=1}^{\infty} \mathbb{E}[|X_n|] < \infty$, then $\sum_{n=1}^{\infty} X_n$ converges absolutely in L^1, and moreover

$$\mathbb{E}\left[\sum_{n=1}^{\infty} X_n\right] = \sum_{n=1}^{\infty} \mathbb{E}[X_n].$$

Indeed, since $|X_n| \ge 0$, we have

$$\mathbb{E}\left[\sum_{n=1}^{N} |X_n|\right] = \sum_{n=1}^{N} \mathbb{E}[|X_n|] < \infty,$$

so $\sum X_n$ converges in L^1. The equality follows from the linearity of expectation and bounded convergence. This is the probabilistic analogue of absolute convergence of series in analysis.

4 Conditional Distribution and Expectation

Discrete Case

If X, Y are discrete random variables, then

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x,\, Y = y)}{\mathbb{P}(Y = y)}.$$

Sometimes we write

$$P_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y).$$

Continuous Case

If X and Y are continuous, then

$$\mathbb{P}(X \in A \mid Y = y) = \frac{\mathbb{P}(X \in A,\, Y = y)}{\mathbb{P}(Y = y)} = \frac{0}{0},$$

which is ill-defined. Instead, define the conditional probability through a limiting argument:

$$\mathbb{P}(X \in A \mid Y = y) = \lim_{\delta \downarrow 0} \frac{\mathbb{P}(X \in A,\, Y \in [y, y+\delta])}{\mathbb{P}(Y \in [y, y+\delta])} = \lim_{\delta \downarrow 0} \frac{\int_A \int_y^{y+\delta} f_{XY}(x,z)\, dz\, dx}{\int_y^{y+\delta} f_Y(z)\, dz}.$$

Using the approximation $f_{XY}(x,z) \approx f_{XY}(x,y)$ and $f_Y(z) \approx f_Y(y)$ for small δ,

$$\mathbb{P}(X \in A \mid Y = y) = \int_A \frac{f_{XY}(x,y)}{f_Y(y)}\, dx.$$

Example (Normal Example)**.**Let $X_1, X_2 \sim N(0,1)$ be independent and define $S_1 = X_1$, $S_2 = X_1 + X_2$. Then

$$\mathbb{P}(S_2 > 8 \mid S_1 = 7) = \mathbb{P}(X_2 > 1),$$

since $S_2 - S_1 = X_2$ is independent of S_1.

Conditional Expectation

For discrete X,

$$\mathbb{E}(X \mid Y = y) = \sum_x x\, \mathbb{P}(X = x \mid Y = y).$$

For continuous X,

$$\mathbb{E}(X \mid Y = y) = \int_{-\infty}^{\infty} x\, \frac{f_{XY}(x,y)}{f_Y(y)}\, dx.$$

Example (Random Walk)**.**Let $\{X_k\}_{k=1}^{\infty}$ satisfy $\mathbb{P}(X_k = 1) = p$, $\mathbb{P}(X_k = -1) = 1 - p$, and define $S_n = \sum_{k=1}^{n} X_k$.

$$\begin{aligned}
\mathbb{P}(S_4 = z \mid S_2 = z) &= \mathbb{P}(X_3 + X_4 = 0 \mid S_2 = z) \\
&= \mathbb{P}(X_3 + X_4 = 0) && \text{(independence)} \\
&= \mathbb{P}(X_3 = 1, X_4 = -1) + \mathbb{P}(X_3 = -1, X_4 = 1) \\
&= 2p(1-p).
\end{aligned}$$

Then

$$\mathbb{E}S_n = \sum_{k=1}^{n} \mathbb{E}X_k = n(2p-1), \qquad \mathbb{E}\left[\frac{S_n}{n}\right] = 2p - 1,$$

and

$$\mathbb{E}(S_{323} \mid S_{322} = 30) = 30 + \mathbb{E}X_{323} = 30 + (2p-1).$$

General Conditional Expectation

$\mathbb{E}(X \mid Y)$ is a random variable measurable w.r.t. $\sigma(Y)$ such that

$$\mathbb{E}(X \mid Y)(\omega) = \mathbb{E}(X \mid Y = y) \quad \text{whenever } Y(\omega) = y.$$

In the discrete case,

$$\mathbb{E}[S_n \mid S_k] = S_k + \mathbb{E}[S_n - S_k \mid S_k], \qquad k < n.$$

Law of Total Expectation

$$\mathbb{E}[\mathbb{E}(X \mid Y)] = \mathbb{E}X.$$

Sketch of Proof (Discrete Case).

$$\mathbb{E}[\mathbb{E}(X \mid Y)] = \sum_y \mathbb{E}(X \mid Y = y)\,\mathbb{P}(Y = y) = \sum_y \sum_x x\,\mathbb{P}(X = x \mid Y = y)\,\mathbb{P}(Y = y) = \mathbb{E}X.$$

For continuous Y,

$$\mathbb{E}[\mathbb{E}(X \mid Y)] = \int_{-\infty}^{\infty} \mathbb{E}(X \mid Y = y) f_Y(y)\,dy.$$

□

Distributions with Random Parameter

Example (Poisson–Binomial Mixture)**.**Suppose $N \sim (\lambda)$ and, given N, $X \mid N \sim (N, p)$. Then

$$\mathbb{E}X = \mathbb{E}[\mathbb{E}(X \mid N)] = \mathbb{E}[Np] = p\lambda.$$

Moreover,

$$\mathbb{P}(X = k) = \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} e^{-\lambda} \frac{\lambda^n}{n!} = e^{-p\lambda} \frac{(p\lambda)^k}{k!},$$

so $X \sim (p\lambda)$.

Example (Dice and Coins)**.**Roll a fair die $X \sim \text{Uniform}\{1, \ldots, 6\}$, then flip X fair coins: $Y \mid X \sim (X, 1/2)$. Then

$$\mathbb{E}Y = \mathbb{E}[\mathbb{E}(Y \mid X)] = \sum_{k=1}^{6} \frac{k}{2} \cdot \frac{1}{6} = \frac{7}{4}.$$

Random Sums of Random Variables

Let $\{X_k\}_{k \geq 1}$ be real-valued r.v.'s and N integer-valued. Define $S_N = \sum_{k=1}^{N} X_k$. Then

$$\mathbb{P}(S_N \leq x) = \sum_{n=0}^{\infty} \mathbb{P}(S_n \leq x)\,\mathbb{P}(N = n),$$

and

$$\mathbb{E}S_N = \sum_{n=0}^{\infty} \mathbb{E}S_n\,\mathbb{P}(N = n) = \mathbb{E}X_1 \cdot \mathbb{E}N.$$

Also,

$$\mathbb{E}[S_N^2] = \sum_{n=0}^{\infty} \mathbb{E}[S_n^2]\,\mathbb{P}(N = n) = \mathbb{V}\text{ar}(X_1)\mathbb{E}N + (\mathbb{E}X_1)^2 \mathbb{E}[N^2],$$

so that

$$\mathbb{V}\text{ar}(S_N) = \mathbb{E}[S_N^2] - (\mathbb{E}S_N)^2.$$

Galton–Watson Branching Process

Model. We begin with a single ancestor:

$$X_0 = 1,$$

and define recursively

$$X_{n+1} = \sum_{k=1}^{X_n} \xi_k^{(n+1)},$$

where $\xi_k^{(i)}$ are i.i.d. nonnegative integer–valued random variables, each representing the number of offspring produced by one individual in generation i.

Thus, X_n denotes the population size at generation n.

Expectation Dynamics. By the law of total expectation and independence of offspring counts,

$$\mathbb{E}[X_{n+1} \mid X_n] = X_n \, \mathbb{E}[\xi], \quad \text{so} \quad \mathbb{E}X_{n+1} = \mathbb{E}X_n \cdot \mathbb{E}\xi.$$

Iterating gives

$$\mathbb{E}X_n = (\mathbb{E}\xi)^n.$$

$$\boxed{\mathbb{E}\xi < 1 \Rightarrow \mathbb{E}X_n \to 0, \quad \mathbb{E}\xi = 1 \Rightarrow \mathbb{E}X_n = 1, \quad \mathbb{E}\xi > 1 \Rightarrow \mathbb{E}X_n \to \infty.}$$

This gives the *expected population growth law*, but says nothing yet about *extinction vs. survival* on a given sample path.

Extinction Probability. Define the extinction event

$$E := \{\exists n : X_n = 0\} = \{\lim_{n\to\infty} X_n = 0\}.$$

Let $q := \mathbb{P}(E)$.

Since $X_{n+1} = 0$ iff all X_n individuals produce no offspring, we have

$$q = \mathbb{P}(\text{eventual extinction}) = \sum_{k=0}^{\infty} \mathbb{P}(X_1 = k)\, q^k = \mathbb{E}[q^{\xi}].$$

Thus q satisfies the fixed–point equation

$$\boxed{q = G(q), \qquad \text{where } G(s) = \mathbb{E}[s^{\xi}] \text{ is the generating function of } \xi.}$$

Qualitative Behavior. Since G is convex, increasing, with $G(1) = 1$:

$$q = \begin{cases} 1, & \mathbb{E}[\xi] \le 1, \\ \text{the unique root in } (0,1), & \mathbb{E}[\xi] > 1. \end{cases}$$

Hence:

$$\mathbb{E}[\xi] \le 1 \implies \mathbb{P}(\lim_{n\to\infty} X_n = 0) = 1,$$

while

$$\mathbb{E}[\xi] > 1 \text{ and } \mathbb{P}(\xi = 0) > 0 \implies 0 < \mathbb{P}(\lim_{n\to\infty} X_n = 0) < 1.$$

Interpretation. Why does extinction still have positive probability when $\mathbb{E}[\xi] > 1$? Because although the average number of offspring exceeds one, there remains a nonzero chance that early generations suffer bad luck — for instance, the first ancestor might produce no children ($\mathbb{P}(\xi = 0) > 0$), or its offspring might all fail to reproduce. Since reproduction events are independent, the process can get trapped in the absorbing state $X_n = 0$ forever once it hits it.

Heuristically: "High fertility does not guarantee survival."

The branching process can *explode exponentially or die out entirely*; which happens is random, and the extinction probability q quantifies this uncertainty.

Critical and Supercritical Regimes.

- **Subcritical ($\mathbb{E}[\xi] < 1$):** population declines in expectation and almost surely becomes extinct.
- **Critical ($\mathbb{E}[\xi] = 1$):** still extinct a.s., though extinction takes longer (heavy–tailed survival time).
- **Supercritical ($\mathbb{E}[\xi] > 1$):** explosion possible, but extinction remains a nonzero event whenever $\mathbb{P}(\xi = 0) > 0$.

Comment. This dichotomy between expectation growth and extinction probability is one of the most profound insights in stochastic population models: *mean growth does not imply survival.* The Galton–Watson process thus became a paradigm for randomness in multiplicative systems—from nuclear chain reactions to epidemic spread.

5 Borel–Cantelli Lemmas

Idea. These lemmas are extremely useful for determining whether a sequence of random events happens infinitely often (abbreviated i.o.).

$$A_n \text{ i.o. } := \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

First Borel–Cantelli Lemma

Theorem 5.1 (First Borel–Cantelli)**.** *If*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty,$$

then

$$\mathbb{P}(A_n \text{ i.o.}) = 0.$$

Proof. By continuity of probability:

$$\mathbb{P}(A_n \text{ i.o.}) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \right) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{k=n}^{\infty} A_k \right).$$

By subadditivity:

$$\mathbb{P}\left(\bigcup_{k=n}^{\infty} A_k \right) \le \sum_{k=n}^{\infty} \mathbb{P}(A_k) \to 0,$$

so the probability of infinitely many occurrences is 0. □

Example (Finite-Sum Case)**.**Let X_n satisfy

$$\mathbb{P}(X_n = 0) = 1 - \frac{1}{n^2}, \qquad \mathbb{P}(X_n = n^2) = \frac{1}{n^2}.$$

Then $\mathbb{E}X_n = 1$, yet

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n \neq 0) = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty.$$

Hence, by Borel–Cantelli, $X_n = 0$ eventually a.s. So although $\mathbb{E}X_n = 1$ for each n, we have $\mathbb{P}(\lim_{n\to\infty} X_n = 0) = 1$.

Second Borel–Cantelli Lemma

Theorem 5.2 (Second Borel–Cantelli)**.** *If the A_n are independent and $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then*

$$\mathbb{P}(A_n \text{ i.o.}) = 1.$$

Proof Sketch. We compute the probability that only finitely many A_n occur:

$$\mathbb{P}(\text{finitely many } A_n) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c \right) = \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{k=n}^{\infty} A_k^c \right).$$

By independence,

$$\mathbb{P}\left(\bigcap_{k=n}^{\infty} A_k^c \right) = \prod_{k=n}^{\infty} \mathbb{P}(A_k^c) = \exp\left(\sum_{k=n}^{\infty} \log(1 - \mathbb{P}(A_k)) \right).$$

Since $\log(1-x) \sim -x$ for small x and $\sum \mathbb{P}(A_k) = \infty$, the sum diverges to $-\infty$. Hence the product $\to 0$, giving $\mathbb{P}(A_n \text{ i.o.}) = 1$. □

Example (Infinite-Sum Case)**.**Let X_k be independent with

$$\mathbb{P}(X_k = k) = \frac{1}{k}, \quad \mathbb{P}(X_k = 0) = 1 - \frac{1}{k}.$$

Then $\mathbb{E}X_k = 1$, but

$$\sum_{k=1}^{\infty} \mathbb{P}(X_k \neq 0) = \sum_{k=1}^{\infty} \frac{1}{k} = \infty.$$

Thus, by the second Borel–Cantelli lemma,

$$\mathbb{P}(X_k \neq 0 \text{ infinitely often}) = 1.$$

Remark. To show almost sure convergence $X_n \to X$, it suffices to verify that for every $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) < \infty.$$

Then, by the first Borel–Cantelli lemma, only finitely many large deviations occur.

Example: Records in an i.i.d. Sequence (A Betting View)

Let $X_1, X_2, \ldots$ be i.i.d. continuous random variables. We call X_j a *record* if it exceeds all previous values:

$$X_j > \max\{X_1, \ldots, X_{j-1}\}.$$

Define the *rank* of the j-th observation:

$$R_j = 1 + \#\{\, i < j : X_i \geq X_j \,\}.$$

Thus, $R_j = 1$ means that X_j is the largest so far — a "record winner."

A surprising symmetry. Because the variables are i.i.d. and continuous (so no ties), every ordering of $(X_1, \ldots, X_n)$ is equally likely. Hence each of the n positions is equally likely to contain the overall maximum:

$$\mathbb{P}(R_n = 1) = \frac{1}{n}.$$

This means that even though the sequence length grows, your chance that "the next one beats all the previous ones" is exactly the same as drawing the winning ticket from n equally likely spots.

A fair betting interpretation. Imagine a gambler who, at each time j, places a bet of \$1 that a new record occurs. At time j, the probability of success is $1/j$. So the expected gain from each individual bet is zero — a fair game. But the gambler plays infinitely many rounds.

Now, since $\sum_{j=1}^{\infty} 1/j = \infty$, the **second Borel–Cantelli lemma** tells us:

$$\mathbb{P}(\text{record occurs infinitely often}) = 1.$$

In words: even though each bet becomes harder to win, the gambler will win infinitely many times almost surely. The "record-breaking moments" never stop coming — only their frequency slows down.

Why the ranks are independent. Every permutation of $(R_1, \ldots, R_n)$ is equally likely:

$$\mathbb{P}(R_1 = r_1, \ldots, R_n = r_n) = \frac{1}{n!}.$$

And since $\mathbb{P}(R_j = r_j) = 1/j$ for each j, observe that

$$\frac{1}{n!} = \prod_{j=1}^{n} \frac{1}{j}.$$

Hence, the joint distribution factors as a product — the ranks are *independent*.

In other words, the position of the current maximum is independent of how previous maxima appeared — the past does not influence which index wins next. Each "race" for the maximum resets the odds.

Rare streaks and record clusters. So far, we have seen that single records occur infinitely often, but consecutive ones do not. Let us now extend the reasoning: what if the gambler bets not only on two *back-to-back* records, but on two records that occur *within a short gap*?

For instance, define the event

$$A_{j,m} := \{R_j = 1,\, R_{j+m} = 1\},$$

which means a record occurs at time j and another one m steps later.

Because the ranks are independent,

$$\mathbb{P}(A_{j,m}) = \mathbb{P}(R_j = 1)\mathbb{P}(R_{j+m} = 1) = \frac{1}{j(j+m)}.$$

Now consider the series

$$\sum_{j=1}^{\infty} \mathbb{P}(A_{j,m}) = \sum_{j=1}^{\infty} \frac{1}{j(j+m)} < \infty \quad \text{for every fixed } m \geq 1.$$

Therefore, by the **first Borel–Cantelli lemma**,

$$\mathbb{P}(\text{infinitely many pairs of records separated by } m \text{ steps}) = 0.$$

Interpretation. The gambler who bets on *every single record* wins infinitely many times. But the gambler who bets on *pairs of records*—whether back-to-back or with any fixed gap—will almost surely lose infinitely often. Such "clusters" of records occur only finitely many times with probability 1.

This shows that while records never stop appearing, they become increasingly isolated: the waiting time between records grows without bound.

Reflection. The record process embodies a fundamental probabilistic tension:

(i) Each new observation has a small but universal chance $1/n$ to set a new record.

(ii) Yet as n increases, records drift further apart, making "record clusters" almost impossible.

From a "betting" perspective:

- A gambler who wagers on each round (first Borel–Cantelli's divergent case) wins infinitely often.
- A gambler who requires *two wins close together* (convergent case) almost surely stops winning eventually.

Thus, even though the game of chance continues forever, the intervals between record wins keep lengthening — a perfect probabilistic metaphor for the fading frequency of extraordinary events.

Example: Logarithmic Growth Rate of Exponential Variables

Let $X_n \sim \text{Exp}(1)$ be i.i.d. random variables with density $f(x) = e^{-x}\mathbf{1}_{\{x>0\}}$. We claim that

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\log n} = 1\right) = 1.$$

1. Tail probability computation. For any fixed $x > 0$,

$$\mathbb{P}\left(\frac{X_n}{\log n} > x\right) = \mathbb{P}(X_n > x \log n) = \int_{x\log n}^{\infty} e^{-t}\, dt = e^{-x\log n} = n^{-x}.$$

2. Behavior of the series $\sum_{n=1}^{\infty} n^{-x}$.

$$\sum_{n=1}^{\infty} n^{-x} \begin{cases} < \infty, & \text{if } x > 1, \\ = \infty, & \text{if } 0 < x \leq 1. \end{cases}$$

This follows from the p–series test: $\sum n^{-p}$ converges iff $p > 1$.

3. Apply the Borel–Cantelli lemmas. Define the events $A_n(x) := \{X_n/\log n > x\}$.

- If $x > 1$, then $\sum_n \mathbb{P}(A_n(x)) = \sum n^{-x} < \infty$. By the **first Borel–Cantelli lemma**,

$$\mathbb{P}(A_n(x) \text{ i.o.}) = 0.$$

  Hence, for every $\varepsilon > 0$, only finitely many n satisfy $X_n/\log n > 1 + \varepsilon$. Therefore,

$$\limsup_{n\to\infty} \frac{X_n}{\log n} \leq 1 \quad \text{a.s.}$$

- If $0 < x < 1$, then $\sum_n \mathbb{P}(A_n(x)) = \infty$, and since the X_n are independent, the **second Borel–Cantelli lemma** gives

$$\mathbb{P}(A_n(x) \text{ i.o.}) = 1.$$

  Thus, for every $\varepsilon > 0$, infinitely many n satisfy $X_n/\log n > 1 - \varepsilon$, meaning

$$\limsup_{n\to\infty} \frac{X_n}{\log n} \geq 1 \quad \text{a.s.}$$

4. Combine both bounds. Since

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\log n} \leq 1\right) = 1 \quad \text{and} \quad \mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\log n} \geq 1\right) = 1,$$

we conclude that

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\log n} = 1\right) = 1.$$

□

Remark. This example illustrates a typical "threshold phenomenon": the exponential tail e^{-x} and the logarithmic growth of $\log n$ balance precisely at $x = 1$. Similar limiting thresholds appear in the **Law of the Iterated Logarithm (LIL)** for Brownian motion, where the normalization involves $\sqrt{2t \log\log t}$. Both results capture how random fluctuations behave at extreme scales.

6 Probability Inequalities

Markov Inequality

Lemma 6.1 (Markov Inequality)**.** *Let $X \geq 0$ and $x > 0$. Then*

$$\mathbb{P}(X > x) \leq \frac{\mathbb{E}X}{x}.$$

Proof. Since $\mathbf{1}_{\{X>x\}} \leq X/x$ (equality iff $X = x$ a.s.), we have

$$\mathbb{P}(X > x) = \mathbb{E}[\mathbf{1}_{\{X>x\}}] \leq \mathbb{E}\left[\frac{X}{x}\mathbf{1}_{\{X>x\}}\right] \leq \mathbb{E}\left[\frac{X}{x}\right] = \frac{\mathbb{E}X}{x}.$$

□

Remark (Historical Note)**.**Today the proof of Markov's inequality fits in a single line, but historically it was far from so concise. In fact, the original variational proof by *V. Markov in 1892 ran to more than 110 pages of detailed argumentation*. For a modern perspective, the survey "Twelve Proofs of the Markov Inequality"(Cambridge) collects many different proofs (analytic, probabilistic, variational) and also discusses this historical background.

Lemma 6.2 (Generalized Markov Inequality)**.** *Let $g : \mathbb{R} \to [0, \infty)$ be increasing. Then*

$$\mathbb{P}(X > x) = \mathbb{P}(g(X) > g(x)) \leq \frac{\mathbb{E}[g(X)]}{g(x)}.$$

Special cases.

- For $g(x) = |x|^p$, $p > 0$:
$$\mathbb{P}(|X| > x) \leq \frac{\mathbb{E}|X|^p}{x^p}.$$
- **Chebyshev's inequality:**
$$\mathbb{P}(|X - \mathbb{E}X| > x) \leq \frac{\mathbb{E}[(X - \mathbb{E}X)^2]}{x^2} = \frac{\mathbb{V}\mathrm{ar}X}{x^2}.$$
Equality occurs only for degenerate two-point distributions.

Application: Markov Inequality and Borel–Cantelli Imply Almost Sure Convergence

Assume that for some $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{E}[|X_n|] < \infty.$$

By the **Markov inequality**,

$$\Pr(|X_n| > \varepsilon) \ \leq \ \frac{\mathbb{E}[|X_n|]}{\varepsilon}.$$

Summing over n gives

$$\sum_{n=1}^{\infty} \Pr(|X_n| > \varepsilon) \ \leq \ \frac{1}{\varepsilon} \sum_{n=1}^{\infty} \mathbb{E}[|X_n|] \ < \ \infty.$$

Now apply the **first Borel–Cantelli lemma**. For any sequence of events $\{A_n\}$,

$$\sum_{n=1}^{\infty} \Pr(A_n) < \infty \quad \Longrightarrow \quad \Pr(\limsup_{n\to\infty} A_n) = 0,$$

where

$$\limsup_{n\to\infty} A_n = \bigcap_{m=1}^{\infty} \bigcup_{n\geq m} A_n$$

is the event that infinitely many of the A_n occur.

Take $A_n = \{|X_n| > \varepsilon\}$. Then $\Pr(\limsup_{n\to\infty} A_n) = 0$ means

$$\Pr(|X_n| > \varepsilon \text{ infinitely often}) = 0.$$

In words, for almost every ω, there exists a random index $N(\omega)$ such that

$$|X_n(\omega)| \le \varepsilon, \qquad \forall n \ge N(\omega).$$

This shows that for almost all outcomes ω, the sequence $|X_n(\omega)|$ eventually stays below any fixed $\varepsilon > 0$. Hence

$$X_n(\omega) \to 0 \quad \text{for almost every } \omega,$$

or equivalently,

$$\Pr\Big(\lim_{n\to\infty} X_n = 0\Big) = 1.$$

Comment. The lim sup formulation clarifies the logic: Markov's inequality bounds each tail probability, ensuring that the total tail mass is summable; Borel–Cantelli then removes the possibility of infinitely many large deviations, so X_n must converge to 0 almost surely.

Application: L^r Convergence Implies Convergence in Probability (via Markov Inequality)

A direct and important consequence of the Markov inequality is that

Theorem 6.1. *Convergence in L^r (for any $r > 0$) implies convergence in probability.*

Application. Let $\{X_n\}$ be a sequence of random variables and X another random variable such that

$$\mathbb{E}[|X_n - X|^r] \to 0, \qquad \text{for some } r > 0.$$

Then for any $\varepsilon > 0$, apply Markov's inequality to the nonnegative random variable $|X_n - X|^r$:

$$\Pr(|X_n - X| \ge \varepsilon) \;=\; \Pr(|X_n - X|^r \ge \varepsilon^r) \;\le\; \frac{\mathbb{E}[|X_n - X|^r]}{\varepsilon^r}.$$

As $\mathbb{E}[|X_n - X|^r] \to 0$, the right-hand side tends to 0, hence

$$\Pr(|X_n - X| \ge \varepsilon) \xrightarrow[n\to\infty]{} 0.$$

Therefore, $X_n \to X$ **in probability**. □

Remark. This argument works for any $r > 0$, including $r = 1$. Intuitively, L^r convergence controls the expected size of deviations between X_n and X, and the Markov inequality transforms that control into a probabilistic statement: the probability of large deviations must vanish as n increases.

Kolmogorov Inequality

Let $X_1, \ldots, X_n$ be independent random variables with mean zero, and $\mathbb{V}\mathrm{ar}(X_k) < \infty$ for all k. Then for any $x > 0$ and $S_n = \sum_{k=1}^n X_k$,

$$\mathbb{P}\Big(\max_{1\le k\le n} |S_k| > x\Big) \;\le\; \frac{\mathbb{V}\mathrm{ar}(S_n)}{x^2}.$$

Proof. We wish to control the probability that the partial sums of independent, mean-zero random variables ever exceed a given boundary:

$$\mathbb{P}\Big(\max_{1\le k\le n} |S_k| > x\Big).$$

Chebyshev's inequality gives us control of $\mathbb{P}(|S_n| > x)$, but not of the *maximum* over all k. How can we translate a global "maximum" event into something additive, so that expectations and variances can enter naturally?

Idea. If we knew the *first time* the process crosses the threshold, we could isolate that event and use disjointness. Let us therefore define, for each k,

$$A_k := \Big\{\max_{1\le j<k} |S_j| \le x,\ |S_k| > x\Big\}.$$

The event A_k records that the process remains within $[-x, x]$ up to time $k-1$ but exits at time k. These A_k are disjoint, and their union equals the event $\{\max_{1\le j\le n} |S_j| > x\}$.

$$\mathbb{P}\left(\max_{1\le j\le n} |S_j| > x\right) = \sum_{k=1}^{n} \mathbb{P}(A_k).$$

What can we compare this to? Variance appears naturally in expressions like $\mathbb{E}[S_n^2]$. Could $\mathbb{E}[S_n^2]$ be large enough to "cover" all those disjoint events? Let's try expanding it over the same partition:

$$\mathbb{E}[S_n^2] = \sum_{k=1}^{n} \mathbb{E}[S_n^2\, 1_{A_k}].$$

Now, what is S_n on A_k? Since S_n is built from **independent increments**,

$$S_n = S_k + (S_n - S_k).$$

$$\mathbb{E}[S_n^2\, 1_{A_k}] = \mathbb{E}\Big[(S_k + (S_n - S_k))^2\, 1_{A_k}\Big] = \mathbb{E}[S_k^2\, 1_{A_k}] + \mathbb{E}[(S_n - S_k)^2\, 1_{A_k}] + 2\,\mathbb{E}[S_k(S_n - S_k)\, 1_{A_k}].$$

The cross term vanishes! because S_k and $(S_n - S_k)$ are independent, and $\mathbb{E}[S_n - S_k] = 0$. Therefore

$$\mathbb{E}[S_k(S_n - S_k)\, 1_{A_k}] = \mathbb{E}[S_k\, 1_{A_k}]\,\mathbb{E}[S_n - S_k] = 0.$$

This leaves

$$\mathbb{E}[S_n^2\, 1_{A_k}] = \mathbb{E}[S_k^2\, 1_{A_k}] + \mathbb{E}[(S_n - S_k)^2\, 1_{A_k}].$$

The second term is nonnegative, so

$$\mathbb{E}[S_n^2\, 1_{A_k}] \ge \mathbb{E}[S_k^2\, 1_{A_k}].$$

Summing over k. Add up all $k = 1, \dots, n$:

$$\mathbb{E}[S_n^2] = \sum_{k=1}^{n} \mathbb{E}[S_n^2\, 1_{A_k}] \ge \sum_{k=1}^{n} \mathbb{E}[S_k^2\, 1_{A_k}].$$

But what is S_k^2 on A_k? By construction of A_k, we know $|S_k| > x$, hence

$$S_k^2\, 1_{A_k} \ge x^2\, 1_{A_k}.$$

Taking expectations gives

$$\mathbb{E}[S_k^2\, 1_{A_k}] \ge x^2\, \mathbb{P}(A_k).$$

Summing again:

$$\mathbb{E}[S_n^2] \ge x^2 \sum_{k=1}^{n} \mathbb{P}(A_k) = x^2\, \mathbb{P}\left(\max_{1\le j\le n} |S_j| > x\right).$$

What have we achieved? We converted the variance of the total sum—an additive, computable quantity—into a bound on the probability of ever crossing the boundary. Rearranging yields

$$\mathbb{P}\left(\max_{1\le k\le n} |S_k| > x\right) \le \frac{\mathbb{V}\mathrm{ar}(S_n)}{x^2}.$$

This is Kolmogorov's inequality. It tells us that the entire trajectory of the random walk is controlled by its total variance. A single variance bound protects us from large deviations at *any* time, a remarkable and forward-looking idea that prefigures Doob's martingale inequalities.

□

Remarks and Comments

Remark (On the A_k Trick)**.**The construction

$$A_k = \{\max_{1\le j<k} |S_j| \le x,\ |S_k| > x\}$$

is remarkably elegant. It decomposes the complex global event $\{\max_{1\le j\le n} |S_j| > x\}$ into disjoint "first exit" events, turning a nonlinear condition on the maximum into a linear sum over disjoint probabilities. This transformation is a discrete precursor to the concept of a *stopping time* in martingale theory. It is precisely this structural insight that gives Kolmogorov's inequality its exact constant 1 and reveals the intimate link between variance additivity and first-hitting behavior.

Remark (Refinement over Chebyshev)**.**Kolmogorov's inequality is a refinement of Chebyshev's inequality. Applying Chebyshev directly to the maximum would yield

$$\mathbb{P}\Big(\max_{1\le k\le n} |S_k| > x\Big) \le \frac{\mathbb{E}[(\max_{1\le k\le n} |S_k|)^2]}{x^2},$$

which is weaker because $\mathbb{E}[(\max |S_k|)^2] \ge \mathbb{V}\mathrm{ar}(S_n)$. Kolmogorov's argument exploits the *first hitting time decomposition* (via the sets A_k), replacing $\mathbb{E}[(\max |S_k|)^2]$ by $\mathbb{V}\mathrm{ar}(S_n)$ and yielding the sharp constant 1. This illustrates how a stopping-time viewpoint can produce strictly stronger inequalities than direct moment bounds.

Remark (Connection to Doob's Martingale Inequality)**.**Kolmogorov's inequality is the earliest discrete prototype of **Doob's L^2 maximal inequality** for martingales. Indeed, (S_k) is a martingale with respect to $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ since $\mathbb{E}[S_{k+1} \mid \mathcal{F}_k] = S_k$. Doob's result states that for any square-integrable martingale (M_k),

$$\mathbb{E}\Big[\big(\max_{1\le k\le n} |M_k|\big)^2\Big] \le 4\,\mathbb{E}[M_n^2], \qquad \mathbb{P}\Big(\max_{1\le k\le n} |M_k| > x\Big) \le \frac{4\,\mathbb{E}[M_n^2]}{x^2}.$$

Kolmogorov's bound corresponds to this inequality with constant 1, valid in the special case of independent increments. The underlying proof technique—partitioning according to the *first exit time* from $[-x, x]$—anticipates Doob's martingale framework by *nearly two decades*.

Exponential Bounds (Chernoff Type)

For any $\lambda > 0$,

$$\mathbb{P}(X > x) = \mathbb{P}\big(e^{\lambda X} > e^{\lambda x}\big) \le \frac{\mathbb{E}e^{\lambda X}}{e^{\lambda x}} = \exp\big(-\lambda x + \log \mathbb{E}e^{\lambda X}\big).$$

This is the *Chernoff bound*. The right-hand side depends on λ, so the sharpest inequality comes from

$$\mathbb{P}(X > x) \le \inf_{\lambda>0} \exp\big(-\lambda x + \log \mathbb{E}e^{\lambda X}\big).$$

Connection to Large Deviations. The quantity

$$\Lambda(\lambda) := \log \mathbb{E}e^{\lambda X}$$

is the *cumulant generating function* (cgf) of X. Taking the infimum above is a convex duality step: by definition,

$$I(x) := \sup_{\lambda>0}\{\lambda x - \Lambda(\lambda)\}$$

is the Legendre–Fenchel transform of Λ. Thus, the Chernoff bound can be written succinctly as

$$\mathbb{P}(X > x) \le e^{-I(x)}.$$

This is the same rate function that appears in **Cramér's Theorem** in large deviations theory.

Example: Gaussian tails. If $X \sim N(0,1)$ then

$$\mathbb{E}[e^{\lambda X}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda y - y^2/2}\, dy = e^{\lambda^2/2}.$$

So $\Lambda(\lambda) = \lambda^2/2$, and

$$\mathbb{P}(X > x) \;\le\; \inf_{\lambda>0} \exp\Big(-\lambda x + \tfrac{\lambda^2}{2}\Big).$$

Minimize the exponent: derivative $\lambda - x = 0 \Rightarrow \lambda = x$. Hence

$$I(x) = \sup_{\lambda>0}(\lambda x - \tfrac{\lambda^2}{2}) = \tfrac{x^2}{2}.$$

Therefore,

$$\mathbb{P}(X > x) \;\le\; e^{-x^2/2}.$$

Some comment on "Tails":

- This bound is not exact (the Gaussian tail behaves like $\frac{1}{x}e^{-x^2/2}$ for large x), but it captures the correct *exponential decay rate*.
- More generally, for any sub-exponential random variable (i.e. with finite mgf in a neighborhood of 0), the Chernoff bound shows that tails are dominated by $e^{-I(x)}$ where I is the Legendre transform of the cgf.
- In large deviation theory, $I(x)$ is the *rate function*: it determines how probabilities of rare events decay exponentially.
- Thus the Chernoff inequality is a one-shot, finite-n precursor to the full **Cramér–Chernoff method** used to prove large deviations for sums of i.i.d. random variables.

Summary. Chernoff bounds are more than just inequalities: they encode the *variational principle of exponential decay*. They show that the tail of X is controlled by the Legendre dual $I(x)$, which later becomes the central object in large deviation theory.

Convergence of Sums via Kolmogorov Inequalities

Example

Assume X_k are i.i.d. with

$$X_k = \begin{cases} 1, & \text{with prob. } \frac{1}{2}, \\ -1, & \text{with prob. } \frac{1}{2}. \end{cases}$$

Let $S_n = \sum_{k=1}^n X_k$.

By Kolmogorov's inequality, for $x > 0$,

$$\mathbb{P}\left(\max_{1\le k\le n} |S_k| > x \right) \;\le\; \frac{\mathbb{V}\mathrm{ar}(S_n)}{x^2}.$$

Since

$$\mathbb{V}\mathrm{ar}(S_n) = \sum_{k=1}^n \mathbb{E}[X_k^2] = \sum_{k=1}^n 1 < \infty,$$

we have

$$\mathbb{P}\left(\max_{1\le k\le n} |S_k| > x \right) \;\le\; \frac{\sum_{k=1}^n 1}{x^2}.$$

Toward proving S_n converges

If we want to prove that S_n converges, we look into increments:

$$\mathbb{P}\left(\max_{j\le k\le n} |S_k - S_j| > x \right) \;\le\; \frac{\sum_{k=j+1}^n \mathbb{V}\mathrm{ar}(X_k)}{x^2}.$$

Hence for every $\varepsilon > 0$, we can choose j large enough so that

$$\mathbb{P}\left(\max_{j\le k\le n} |S_k - S_j| > x \right) < \varepsilon,$$

which is equivalent to

$$\mathbb{P}\left(\exists j \quad \max_{j\le k\le n} |S_k - S_j| \le x\right) \ge 1-\varepsilon, \quad \forall \varepsilon > 0.$$

Thus, with probability 1 there exists some j such that $\max_{j\le k\le n} |S_k - S_j| \le x$, implying

$$\mathbb{P}(S_n \text{ converges as } n \to \infty) = 1.$$

(Here one typically takes $x_m = 2^{-m}$ to force convergence.)

Lemma (Converse)

Let X_n be independent with $\sup |X_n| \le A$ (a.s.) for some deterministic A, and $\mathbb{E}X_n = 0$. Let $S_n = \sum_{k=1}^n X_k$. Then

$$\mathbb{P}\left(\max_{1\le k\le n} |S_k| > x\right) \ge 1 - \frac{(x+A)^2}{\sum_{k=1}^n \mathbb{V}\mathrm{ar}(X_k)}.$$

So if $\sum_{k=1}^\infty \mathbb{V}\mathrm{ar}(X_k) = \infty$, then S_n cannot converge a.s.

Summary

If X_k are independent, uniformly bounded, with mean zero, then

$$\mathbb{P}(S_n \text{ converges as } n\to\infty) = \begin{cases} 1, & \sum_{k=1}^\infty \mathbb{V}\mathrm{ar}(X_k) < \infty, \\ 0, & \sum_{k=1}^\infty \mathbb{V}\mathrm{ar}(X_k) = \infty. \end{cases}$$

Example

Let $\theta > 0$, and define

$$X_k = \begin{cases} 1/k^\theta, & \text{with prob. } \frac{1}{2}, \\ -1/k^\theta, & \text{with prob. } \frac{1}{2}. \end{cases}$$

Then

$$\sum_{k=1}^\infty X_k \quad \text{converges a.s. iff} \quad \sum_{k=1}^\infty \mathbb{V}\mathrm{ar}(X_k) < \infty.$$

But

$$\sum_{k=1}^\infty \mathbb{V}\mathrm{ar}(X_k) = \sum_{k=1}^\infty \frac{1}{k^{2\theta}}.$$

Hence:

$$\sum X_k \text{ converges a.s.} \iff \theta > \tfrac{1}{2}.$$

Remark

If $\mathbb{E}X_k \neq 0$ but $\mathbb{E}X_k < \infty$, the same result holds for

$$S_n = \sum_{k=1}^n (X_k - \mathbb{E}X_k).$$

That is, if X_k are bounded and independent, then S_n converges a.s. iff

$$\sum_{k=1}^\infty \mathbb{V}\mathrm{ar}(X_k) < \infty.$$

Consequence and Converse

If $\sum \mathbb{V}\mathrm{ar}X_k < \infty$, the bound implies S_n converges a.s. Conversely:

Lemma 6.3 (Converse)**.** *If X_k are independent, mean 0, uniformly bounded ($|X_k| \le A$), then*

$$\mathbb{P}\left(\max_{1\le k\le n} |S_k| > x\right) \ge 1 - \frac{(x+A)^2}{\sum_{k=1}^n \mathbb{V}arX_k}.$$

Thus if $\sum \mathbb{V}arX_k = \infty$, S_n cannot converge a.s.

Summary. For independent, bounded, mean-zero variables,

$$S_n \text{ converges a.s.} \quad \Longleftrightarrow \quad \sum_{k=1}^{\infty} \mathbb{V}\mathrm{ar}X_k < \infty.$$

Hölder Inequality

Theorem 6.2 (Hölder). *For $p > 1$, $1/p + 1/q = 1$,*

$$\mathbb{E}|XY| \le (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|Y|^q)^{1/q},$$

with equality iff $|X|^p$ and $|Y|^q$ are proportional a.s.

Proof. For normalized $\mathbb{E}|X|^p = \mathbb{E}|Y|^q = 1$, Young's inequality gives $|XY| \le |X|^p/p + |Y|^q/q$, hence $\mathbb{E}|XY| \le 1$. Scaling X, Y yields the general case. □

Minkowski Inequality

Theorem 6.3 (Minkowski). *For $p \ge 1$,*

$$(\mathbb{E}|X+Y|^p)^{1/p} \le (\mathbb{E}|X|^p)^{1/p} + (\mathbb{E}|Y|^p)^{1/p}.$$

Sketch. For $p = 1$ this is triangle inequality. For $p > 1$, write $q = p/(p-1)$ and apply Hölder to $\mathbb{E}|X+Y|^{p-1}|X|$ and $\mathbb{E}|X+Y|^{p-1}|Y|$, then rearrange:

$$\mathbb{E}|X+Y|^p \le (\mathbb{E}|X+Y|^p)^{1/q} \big[(\mathbb{E}|X|^p)^{1/p} + (\mathbb{E}|Y|^p)^{1/p} \big].$$

Divide both sides by $(\mathbb{E}|X+Y|^p)^{1/q}$ to obtain the result. □

c_r–Inequality

Theorem 6.4 (c_r–Inequality). *For $r > 0$,*

$$\mathbb{E}|X+Y|^r \le c_r (\mathbb{E}|X|^r + \mathbb{E}|Y|^r), \qquad c_r = \begin{cases} 1, & r \le 1, \\ 2^{r-1}, & r \ge 1. \end{cases}$$

Proof. If $0 < r \le 1$, $x \mapsto x^r$ is concave so $(x+y)^r \le x^r + y^r$. If $r \ge 1$, convexity gives $\left(\frac{x+y}{2}\right)^r \le \frac{x^r + y^r}{2} \Rightarrow (x+y)^r \le 2^{r-1}(x^r + y^r)$. Take expectations. □

Correlation and Cauchy–Schwarz

$$\rho(X,Y) = \frac{(X,Y)}{\sqrt{\mathbb{V}\mathrm{ar}X \, \mathbb{V}\mathrm{ar}Y}} \in [-1, 1].$$

By Cauchy–Schwarz ($p = q = 2$):

$$|(X,Y)| \le \sqrt{\mathbb{V}\mathrm{ar}X \, \mathbb{V}\mathrm{ar}Y}.$$

Equality holds iff $Y = a + bX$ a.s. Examples:

- $Y = a + bX$, $b > 0 \Rightarrow \rho = 1$;
- $Y = a + bX$, $b < 0 \Rightarrow \rho = -1$;
- Independence $\Rightarrow \rho = 0$ (but converse false, e.g. $Y = X^2$).

Jensen Inequality

Definition 6.1. *$f : \mathbb{R} \to \mathbb{R}$ is convex if $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$ for all $\lambda \in (0,1)$.*

Theorem 6.5 (Jensen). *If f is convex and $\mathbb{E}|X| < \infty$, then*

$$f(\mathbb{E}X) \le \mathbb{E} f(X),$$

with equality iff X is a.s. constant or f is linear on the support of X.

Proof. Let $x_0 = \mathbb{E}X$. Convexity implies existence of a supporting line $f(x) \geq a(x - x_0) + f(x_0)$ for all x. Taking expectations gives $\mathbb{E}f(X) \geq f(\mathbb{E}X)$. □

Examples. For convex f:

- $f(x) = e^x$: $\mathbb{E}[e^X] \geq e^{\mathbb{E}X}$ (by AM–GM).
- $f(x) = x^2$: $\mathbb{V}\mathrm{ar}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 \geq 0$.
- $f(x) = |x|$: $\mathbb{E}|X| \geq |\mathbb{E}X|$.