

ECE 534: Random Processes

Lecture Notes: Tail Events and Zero–One Laws, the Central Limit Theorem,
the Law of the Iterated Logarithm, and Probabilistic Inequalities

University of Illinois Urbana–Champaign

Fall 2026

Standing notation. $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, $\mathcal{B}(\mathbb{R})$ is the Borel σ -algebra on $\mathbb{R}$, and all random variables are real valued unless stated otherwise. For a sequence $X_1, X_2, \dots$ we write $S_n = X_1 + \dots + X_n$. Φ and ϕ denote the standard normal distribution function and density, and $\varphi_X(t) = \mathbb{E}[e^{itX}]$ denotes the characteristic function of X .¹ “a.s.” means almost surely, “i.o.” means infinitely often, and $\log$ is the natural logarithm.

Contents

1	σ-algebras generated by random variables	2
2	The tail σ-algebra	3
2.1	Examples of tail events	3
3	Kolmogorov’s zero–one law	4
4	S_n/n converges or diverges with probability one	5
5	The central limit theorem	6
5.1	Characteristic functions	6
5.2	Statement and proof	7
5.3	Rate of convergence: the Berry–Esseen theorem	8
5.4	Higher-order corrections: the Edgeworth expansion	9
6	The law of the iterated logarithm and other scaling laws	10
6.1	What the SLLN and CLT leave open	10
6.2	The law of the iterated logarithm	10
6.3	Other scaling laws	12
7	Probabilistic inequalities	13
7.1	Markov’s inequality	13
7.2	Chebyshev’s inequality	13
7.3	Jensen’s inequality	15
7.4	Hölder, Cauchy–Schwarz and Minkowski (statements)	15
	Further reading	16

¹Engineering texts often write j for $\sqrt{-1}$ and $\Phi_X(u) = \mathbb{E}[e^{juX}]$ for the characteristic function. We use i and φ_X so that Φ can be reserved for the normal distribution function.

1 σ -algebras generated by random variables

A random variable X is a function on Ω , and observing X tells us only part of what there is to know about the outcome ω . The σ -algebra generated by X is the precise description of that part.

Definition 1.1 (σ -algebra generated by a random variable). Let $X : \Omega \rightarrow \mathbb{R}$ be a random variable. The σ -algebra generated by X is

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

Because preimages commute with complements and countable unions, $\sigma(X)$ is a σ -algebra, and $\sigma(X) \subseteq \mathcal{F}$ because X is measurable. It is the *smallest* σ -algebra on Ω with respect to which X is measurable: any σ -algebra that makes X measurable must contain every set $X^{-1}(B)$.

Definition 1.2 (σ -algebra generated by a family). Let $\{X_i : i \in I\}$ be a family of random variables (the index set I may be finite, countable, or uncountable). Then

$$\sigma(X_i : i \in I) = \sigma\left(\bigcup_{i \in I} \sigma(X_i)\right)$$

is the smallest σ -algebra with respect to which every X_i is measurable.

The outer $\sigma(\cdot)$ is needed: a union of σ -algebras is usually not a σ -algebra. For finitely many variables there is an equivalent description, $\sigma(X_1, \dots, X_n) = \{\{(X_1, \dots, X_n) \in B\} : B \in \mathcal{B}(\mathbb{R}^n)\}$.

Interpretation as information. $\sigma(X)$ is the collection of events whose occurrence can be decided by looking at the value of X alone. The following standard fact makes this precise.

Proposition 1.3 (Doob–Dynkin lemma). A random variable Y is $\sigma(X_1, \dots, X_n)$ -measurable if and only if $Y = g(X_1, \dots, X_n)$ for some Borel function $g : \mathbb{R}^n \rightarrow \mathbb{R}$.

So “ Y is $\sigma(X)$ -measurable” and “ Y is a function of X ” mean the same thing. In this language, random variables X and Y are independent exactly when $\sigma(X)$ and $\sigma(Y)$ are independent σ -algebras, that is, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for all $A \in \sigma(X)$, $B \in \sigma(Y)$. An infinite family is independent when every finite subfamily is.

Example 1.4 (Indicator). If $X = \mathbf{1}_A$ for an event A , then $\sigma(X) = \{\emptyset, A, A^c, \Omega\}$.

Example 1.5 (Discrete random variable). If X takes the distinct values $x_1, x_2, \dots$, then the events $A_k = \{X = x_k\}$ partition Ω and $\sigma(X)$ consists of all unions $\bigcup_{k \in K} A_k$, $K \subseteq \mathbb{N}$. Knowing X means knowing which cell of the partition contains ω , and nothing more.

Example 1.6 (Binary digits). Let $\Omega = [0, 1)$ with Lebesgue measure and let $X_n(\omega)$ be the n th binary digit of ω ; the X_n are i.i.d. Bernoulli(1/2). Then $\sigma(X_1) = \{\emptyset, [0, \frac{1}{2}), [\frac{1}{2}, 1), \Omega\}$, and $\sigma(X_1, \dots, X_n)$ consists of all unions of the 2^n dyadic intervals $[k2^{-n}, (k+1)2^{-n})$, so it has 2^{2^n} elements. The σ -algebras $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$ increase with n : more digits, finer resolution.

Remark 1.7 (Generating π -systems). A π -system is a collection of sets closed under finite intersections. The “cylinder” events $\{X_1 \in B_1, \dots, X_n \in B_n\}$, $B_i \in \mathcal{B}(\mathbb{R})$, form a π -system that generates $\sigma(X_1, \dots, X_n)$. Working with such generators is how one proves statements about σ -algebras whose elements cannot be listed explicitly; see Section 3.

2 The tail σ -algebra

Definition 2.1 (Tail σ -algebra). Let $X_1, X_2, \dots$ be a sequence of random variables. For $n \geq 1$ let

$$\mathcal{T}_n = \sigma(X_n, X_{n+1}, X_{n+2}, \dots)$$

be the σ -algebra describing the “future from time n on.” The *tail σ -algebra* of the sequence is

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \mathcal{T}_n.$$

Events in $\mathcal{T}$ are *tail events*, and $\mathcal{T}$ -measurable (extended real valued) random variables are *tail random variables*.

The $\mathcal{T}_n$ decrease in n , and an intersection of σ -algebras is a σ -algebra, so $\mathcal{T}$ is well defined. An event A is a tail event when, for *every* n , its occurrence is determined by $X_n, X_{n+1}, \dots$. Informally:

a tail event is one whose occurrence is not affected by changing the values of any finite number of the X_i .

This heuristic is the practical test. To verify it rigorously, show that $A \in \mathcal{T}_m$ for each fixed m .

2.1 Examples of tail events

Throughout, $B_n, B \in \mathcal{B}(\mathbb{R})$ and $c \in \mathbb{R}$.

Example 2.2 (Infinitely often and eventually).

$$\{X_n \in B_n \text{ i.o.}\} = \bigcap_{m \geq 1} \bigcup_{n \geq m} \{X_n \in B_n\}$$

is a tail event: for any m_0 the intersection may be started at $m = m_0$ without changing the set, and then every term lies in $\mathcal{T}_{m_0}$. Taking complements, $\{X_n \in B_n \text{ for all large } n\}$ is a tail event too. For instance, if X_n indicates that a channel is in a deep fade in slot n , “the channel fades in infinitely many slots” is a tail event.

Example 2.3 (Limits superior and inferior). $\limsup_n X_n$ and $\liminf_n X_n$ are tail random variables, since $\limsup_n X_n = \inf_{m \geq m_0} \sup_{n \geq m} X_n$ for every m_0 . Hence $\{\limsup_n X_n \leq c\}$ and

$$\left\{ \lim_{n \rightarrow \infty} X_n \text{ exists} \right\} = \left\{ \limsup_n X_n = \liminf_n X_n \in \mathbb{R} \right\}$$

are tail events.

Example 2.4 (Convergence of a series). $\{\sum_{n=1}^{\infty} X_n \text{ converges}\}$ is a tail event, because $\sum_{n \geq 1} X_n$ converges if and only if $\sum_{n \geq m} X_n$ converges. The *value* of the sum is not a tail random variable: it changes when X_1 changes.

Example 2.5 (Normalized sums). Let $a_n \rightarrow \infty$. For fixed m and $n \geq m$,

$$\frac{S_n}{a_n} = \frac{S_{m-1}}{a_n} + \frac{X_m + \dots + X_n}{a_n}, \quad \frac{S_{m-1}(\omega)}{a_n} \xrightarrow{n \rightarrow \infty} 0 \text{ for every } \omega.$$

So $\limsup_n S_n/a_n$ and $\liminf_n S_n/a_n$ depend only on $X_m, X_{m+1}, \dots$, for every m ; they are tail random variables. Consequently

$$\left\{ \lim_n \frac{S_n}{a_n} \text{ exists} \right\}, \quad \left\{ \lim_n \frac{S_n}{a_n} = c \right\}, \quad \left\{ \limsup_n \frac{S_n}{a_n} > c \right\}, \quad \left\{ \limsup_n \frac{S_n}{\sqrt{2n \log \log n}} = c \right\}$$

are all tail events. This example drives Sections 4 and 6.

Example 2.6 (Radius of convergence). The radius of convergence of the random power series $\sum_n X_n z^n$ is $R = 1/\limsup_n |X_n|^{1/n}$, a tail random variable by Example 2.3.

Example 2.7 (Events that are *not* tail events). (a) $\{X_1 > 0\}$, and more generally any nontrivial event determined by finitely many coordinates.

(b) $\{\sum_n X_n \geq 1\}$ (on the event that the series converges): the value of the sum depends on every term.

(c) $\{\limsup_n S_n > 0\}$: with no normalization, changing X_1 shifts every S_n by the same amount. In contrast $\{\limsup_n S_n = \infty\}$ is a tail event.

(d) $\{S_n = 0 \text{ i.o.}\}$ for a ± 1 random walk. Along the sequence $+1, -1, +1, -1, \dots$ the walk returns to 0 infinitely often; flipping only X_1 gives a walk that never does. (For i.i.d. sequences, events such as (c) and (d) are still trivial, but by a different result, the Hewitt–Savage zero–one law for permutation-invariant events.)

3 Kolmogorov’s zero–one law

Theorem 3.1 (Kolmogorov’s zero–one law). *Let $X_1, X_2, \dots$ be independent random variables with tail σ -algebra $\mathcal{T}$. Then every tail event is trivial:*

$$A \in \mathcal{T} \implies \mathbb{P}(A) \in \{0, 1\}.$$

The idea of the proof is that a tail event A is independent of $(X_1, \dots, X_n)$ for every n , hence independent of the whole sequence; but A is itself determined by the sequence, so A is independent of *itself*. The only technical point is passing from finitely many coordinates to infinitely many, which is done with the π – λ theorem.

Tools. A collection $\mathcal{L}$ of subsets of Ω is a λ -system if (i) $\Omega \in \mathcal{L}$; (ii) $A, B \in \mathcal{L}$ and $A \subseteq B$ imply $B \setminus A \in \mathcal{L}$; (iii) $A_1 \subseteq A_2 \subseteq \dots$ in $\mathcal{L}$ implies $\bigcup_k A_k \in \mathcal{L}$.

Theorem 3.2 (Dynkin’s π – λ theorem). *If $\mathcal{P}$ is a π -system, $\mathcal{L}$ is a λ -system and $\mathcal{P} \subseteq \mathcal{L}$, then $\sigma(\mathcal{P}) \subseteq \mathcal{L}$.*

Lemma 3.3 (Independence extends from generators). *Let $\mathcal{P}$ and $\mathcal{Q}$ be π -systems of events such that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for all $A \in \mathcal{P}$, $B \in \mathcal{Q}$. Then $\sigma(\mathcal{P})$ and $\sigma(\mathcal{Q})$ are independent.*

Proof. Fix $B \in \mathcal{Q}$ and let $\mathcal{L}_B = \{A \in \mathcal{F} : \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)\}$. Then $\Omega \in \mathcal{L}_B$; if $A \subseteq A'$ are in $\mathcal{L}_B$ then $\mathbb{P}((A' \setminus A) \cap B) = \mathbb{P}(A' \cap B) - \mathbb{P}(A \cap B) = \mathbb{P}(A' \setminus A)\mathbb{P}(B)$; and $\mathcal{L}_B$ is closed under increasing unions by continuity of $\mathbb{P}$ from below. So $\mathcal{L}_B$ is a λ -system containing $\mathcal{P}$, and therefore $\sigma(\mathcal{P}) \subseteq \mathcal{L}_B$. This shows $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for all $A \in \sigma(\mathcal{P})$ and $B \in \mathcal{Q}$. Now fix $A \in \sigma(\mathcal{P})$ and repeat the argument with $\mathcal{L}'_A = \{B \in \mathcal{F} : \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)\} \supseteq \mathcal{Q}$ to get $\sigma(\mathcal{Q}) \subseteq \mathcal{L}'_A$. $\square$

Proof of Theorem 3.1. Let $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$ and $\mathcal{F}_\infty = \sigma(X_1, X_2, \dots)$.

Step 1: $\mathcal{F}_n$ and $\mathcal{T}_{n+1}$ are independent. Let $\mathcal{P}_n$ be the collection of events $\{X_1 \in B_1, \dots, X_n \in B_n\}$ and let $\mathcal{Q}_n$ be the collection of events $\{X_{n+1} \in B_{n+1}, \dots, X_{n+k} \in B_{n+k}\}$, over all $k \geq 1$ and all Borel sets B_i . Both are π -systems (to intersect two events in $\mathcal{Q}_n$ with different k , pad the shorter one with $B_i = \mathbb{R}$), with $\sigma(\mathcal{P}_n) = \mathcal{F}_n$ and $\sigma(\mathcal{Q}_n) = \mathcal{T}_{n+1}$. For $A \in \mathcal{P}_n$ and $B \in \mathcal{Q}_n$, independence of $X_1, \dots, X_{n+k}$ gives

$$\mathbb{P}(A \cap B) = \prod_{i=1}^{n+k} \mathbb{P}(X_i \in B_i) = \mathbb{P}(A) \mathbb{P}(B).$$

Lemma 3.3 now gives the claim.

Step 2: $\mathcal{F}_n$ and $\mathcal{T}$ are independent, for every n . This is immediate from Step 1 because $\mathcal{T} \subseteq \mathcal{T}_{n+1}$.

Step 3: $\mathcal{F}_\infty$ and $\mathcal{T}$ are independent. Since $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots$, the union $\mathcal{P} = \bigcup_n \mathcal{F}_n$ is closed under finite intersections (two of its sets lie in a common $\mathcal{F}_n$), so it is a π -system, and $\sigma(\mathcal{P}) = \mathcal{F}_\infty$. By Step 2, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for all $A \in \mathcal{P}$ and $B \in \mathcal{T}$. Apply Lemma 3.3 with $\mathcal{Q} = \mathcal{T}$.

Step 4: conclusion. Let $A \in \mathcal{T}$. Since $\mathcal{T} \subseteq \mathcal{T}_1 = \mathcal{F}_\infty$, we have $A \in \mathcal{F}_\infty$ and $A \in \mathcal{T}$, so by Step 3 the event A is independent of itself:

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2.$$

Hence $\mathbb{P}(A) \in \{0, 1\}$. □

Corollary 3.4 (Tail random variables are constant). *If $X_1, X_2, \dots$ are independent and Y is a tail random variable with values in $[-\infty, \infty]$, then there is a constant $c \in [-\infty, \infty]$ with $\mathbb{P}(Y = c) = 1$.*

Proof. For each $y \in \mathbb{R}$ the event $\{Y \leq y\}$ is in $\mathcal{T}$, so $F(y) = \mathbb{P}(Y \leq y) \in \{0, 1\}$. Since F is nondecreasing, let $c = \inf\{y : F(y) = 1\}$ (with $\inf \emptyset = +\infty$). Then $\mathbb{P}(Y \leq y) = 1$ for $y > c$ and $\mathbb{P}(Y \leq y) = 0$ for $y < c$, so $\mathbb{P}(Y = c) = 1$. □

Remark 3.5 (What the law does and does not say). (i) The theorem says the probability is 0 or 1 but not *which*. Deciding usually takes a separate computation. For events of the form $\{X_n \in B_n \text{ i.o.}\}$ with independent X_n , the Borel–Cantelli lemmas settle it: the probability is 0 if $\sum_n \mathbb{P}(X_n \in B_n) < \infty$ and 1 if the sum diverges.

(ii) Independence is essential. If $X_n = X_1$ for all n , then $\mathcal{T} = \sigma(X_1)$, which is nontrivial whenever X_1 is not constant.

(iii) No assumption of identical distributions or of finite moments is made.

Example 3.6 (Random signs). Let $\varepsilon_1, \varepsilon_2, \dots$ be i.i.d. with $\mathbb{P}(\varepsilon_n = \pm 1) = \frac{1}{2}$ and let a_n be real constants. By Example 2.4 and Theorem 3.1, the series $\sum_n \varepsilon_n a_n$ converges with probability 0 or 1. Kolmogorov’s three-series theorem identifies the case: it converges a.s. if $\sum_n a_n^2 < \infty$ and diverges a.s. otherwise. So the random harmonic series $\sum_n \pm \frac{1}{n}$ converges a.s., while $\sum_n \pm \frac{1}{\sqrt{n}}$ diverges a.s.

4 S_n/n converges or diverges with probability one

We now apply the zero–one law to sample means. There is no moment assumption in the next result, and the X_i need not be identically distributed.

Proposition 4.1 (Dichotomy for sample means). *Let $X_1, X_2, \dots$ be independent and $S_n = X_1 + \dots + X_n$. Then*

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} \frac{S_n}{n} \text{ exists and is finite}\right) \in \{0, 1\}.$$

Moreover there are deterministic constants $-\infty \leq \underline{c} \leq \bar{c} \leq \infty$ such that

$$\liminf_{n \rightarrow \infty} \frac{S_n}{n} = \underline{c} \quad \text{and} \quad \limsup_{n \rightarrow \infty} \frac{S_n}{n} = \bar{c} \quad \text{a.s.}$$

In particular, if S_n/n converges a.s., its limit is a.s. a constant.

Proof. By Example 2.5 with $a_n = n$, the random variables $\liminf_n S_n/n$ and $\limsup_n S_n/n$ are tail random variables: for each fixed m , the contribution S_{m-1}/n of the first $m-1$ terms vanishes as $n \rightarrow \infty$. By Corollary 3.4 each is a.s. equal to a constant, $\underline{c}$ and $\bar{c}$. The event that the limit exists and is finite is $\{\liminf_n S_n/n = \limsup_n S_n/n \in \mathbb{R}\}$, a tail event, so it has probability 0 or 1 by Theorem 3.1; it has probability 1 exactly when $\underline{c} = \bar{c} \in \mathbb{R}$, and then the limit equals that common constant. $\square$

So for independent summands there is no middle ground: it cannot happen that S_n/n converges on a set of probability 0.3, nor that it converges to a genuinely random limit. The same argument works for S_n/a_n with any deterministic $a_n \rightarrow \infty$.

Which case occurs? The zero-one law does not say. For i.i.d. sequences the answer is the strong law of large numbers and its converse.

Theorem 4.2 (SLLN and its converse). *Let $X_1, X_2, \dots$ be i.i.d.*

(a) *If $\mathbb{E}|X_1| < \infty$, then $S_n/n \rightarrow \mathbb{E}X_1$ a.s.*

(b) *If $\mathbb{E}|X_1| = \infty$, then S_n/n diverges a.s.*

Proof of (b). Suppose S_n/n converged to a finite limit with positive probability. By Proposition 4.1 it would converge a.s. to a constant c , and then

$$\frac{X_n}{n} = \frac{S_n}{n} - \frac{n-1}{n} \cdot \frac{S_{n-1}}{n-1} \longrightarrow c - c = 0 \quad \text{a.s.,}$$

so $\mathbb{P}(|X_n| \geq n \text{ i.o.}) = 0$. The events $\{|X_n| \geq n\}$ are independent, so the second Borel–Cantelli lemma forces $\sum_n \mathbb{P}(|X_n| \geq n) = \sum_n \mathbb{P}(|X_1| \geq n) < \infty$, which is equivalent to $\mathbb{E}|X_1| < \infty$. This contradicts the hypothesis. $\square$

Example 4.3 (Cauchy samples). If the X_i are i.i.d. standard Cauchy, then S_n/n is again standard Cauchy for every n , and $\mathbb{E}|X_1| = \infty$. By Theorem 4.2(b), S_n/n diverges a.s.; in fact $\limsup_n S_n/n = +\infty$ and $\liminf_n S_n/n = -\infty$ a.s. Averaging does not help at all.

Example 4.4 (Dependence can produce a random limit). Let Θ be uniform on $[0, 1]$ and, given Θ , let $X_1, X_2, \dots$ be i.i.d. Bernoulli(Θ). Then $S_n/n \rightarrow \Theta$ a.s., a nondegenerate random limit. This does not contradict Proposition 4.1: the X_i are exchangeable but not independent, and $\{\lim S_n/n \leq \frac{1}{2}\}$ is a tail event with probability $\frac{1}{2}$.

5 The central limit theorem

The law of large numbers says $S_n - n\mu = o(n)$. The CLT identifies the actual size of the fluctuations, $\sqrt{n}$, and their limiting shape.

5.1 Characteristic functions

The *characteristic function* of a random variable X is

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \mathbb{E}[\cos tX] + i \mathbb{E}[\sin tX], \quad t \in \mathbb{R}.$$

It always exists, because $|e^{itX}| = 1$. We need the following facts.

(C1) $\varphi_X(0) = 1$, $|\varphi_X(t)| \leq 1$, and φ_X is uniformly continuous.

(C2) $\varphi_{aX+b}(t) = e^{itb}\varphi_X(at)$.

(C3) If X and Y are independent, then $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$.

(C4) (Uniqueness) $\varphi_X = \varphi_Y$ implies that X and Y have the same distribution.

(C5) If $Z \sim N(0, 1)$ then $\varphi_Z(t) = e^{-t^2/2}$.

(C6) (**Lévy's continuity theorem**) $X_n \xrightarrow{d} X$ if and only if $\varphi_{X_n}(t) \rightarrow \varphi_X(t)$ for every $t \in \mathbb{R}$.

Property (C3) is why characteristic functions suit sums of independent variables, and (C6) converts a statement about distributions into pointwise convergence of functions.

Lemma 5.1 (Taylor estimate). *For all real x ,*

$$\left| e^{ix} - \left(1 + ix - \frac{x^2}{2} \right) \right| \leq \min\left(\frac{|x|^3}{6}, x^2\right).$$

Consequently, if $\mathbb{E}X = 0$ and $\mathbb{E}X^2 = \sigma^2 < \infty$, then for all t

$$\left| \varphi_X(t) - \left(1 - \frac{\sigma^2 t^2}{2} \right) \right| \leq \mathbb{E}\left[\min\left(\frac{|t|^3 |X|^3}{6}, t^2 X^2\right)\right]. \quad (1)$$

Proof. Integrating $\frac{d}{ds}e^{is} = ie^{is}$ repeatedly from 0 to x gives, in turn, $|e^{ix} - 1| \leq |x|$, $|e^{ix} - 1 - ix| \leq x^2/2$, and $|e^{ix} - 1 - ix + x^2/2| \leq |x|^3/6$. The second of these and the triangle inequality give $|e^{ix} - 1 - ix + x^2/2| \leq |e^{ix} - 1 - ix| + x^2/2 \leq x^2$. For (1), put $x = tX$, take expectations, and use $|\mathbb{E}W| \leq \mathbb{E}|W|$ together with $\mathbb{E}X = 0$, $\mathbb{E}X^2 = \sigma^2$. $\square$

The point of the minimum: the cubic bound is good for small $|tX|$ but needs a third moment; the quadratic bound needs only a second moment. Together they show the right side of (1) is $o(t^2)$ as $t \rightarrow 0$ under a finite variance alone.

Lemma 5.2 (Comparing products). *If $z_1, \dots, z_n$ and $w_1, \dots, w_n$ are complex numbers of modulus at most 1, then*

$$\left| \prod_{k=1}^n z_k - \prod_{k=1}^n w_k \right| \leq \sum_{k=1}^n |z_k - w_k|.$$

Proof. For $n = 2$: $z_1 z_2 - w_1 w_2 = (z_1 - w_1)z_2 + w_1(z_2 - w_2)$. Induct on n . $\square$

5.2 Statement and proof

Theorem 5.3 (Central limit theorem, Lindeberg–Lévy). *Let $X_1, X_2, \dots$ be i.i.d. with mean μ and variance $\sigma^2 \in (0, \infty)$. Then*

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1), \quad \text{that is,} \quad \lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq x) = \Phi(x) \quad \text{for every } x \in \mathbb{R}.$$

Proof. Replacing X_k by $(X_k - \mu)/\sigma$, we may assume $\mu = 0$ and $\sigma = 1$, so $Z_n = S_n/\sqrt{n}$. Write $\varphi = \varphi_{X_1}$. By (C2) and (C3),

$$\varphi_{Z_n}(t) = \left[\varphi\left(\frac{t}{\sqrt{n}}\right) \right]^n.$$

Fix $t \in \mathbb{R}$. By (C5) and (C6) it suffices to show $\varphi_{Z_n}(t) \rightarrow e^{-t^2/2}$.

Take $n \geq t^2$, so that $0 \leq 1 - \frac{t^2}{2n} \leq 1$. Apply Lemma 5.2 with $z_k = \varphi(t/\sqrt{n})$ and $w_k = 1 - \frac{t^2}{2n}$ for all k , and then (1) with t replaced by $t/\sqrt{n}$:

$$\left| \varphi\left(\frac{t}{\sqrt{n}}\right)^n - \left(1 - \frac{t^2}{2n}\right)^n \right| \leq n \left| \varphi\left(\frac{t}{\sqrt{n}}\right) - \left(1 - \frac{t^2}{2n}\right) \right| \leq \mathbb{E} \left[\min\left(\frac{|t|^3 |X_1|^3}{6\sqrt{n}}, t^2 X_1^2\right) \right].$$

The integrand on the right tends to 0 pointwise as $n \rightarrow \infty$ and is dominated by $t^2 X_1^2$, which is integrable. By dominated convergence the right side tends to 0. Finally $(1 - \frac{t^2}{2n})^n \rightarrow e^{-t^2/2}$ by elementary calculus. Therefore $\varphi_{Z_n}(t) \rightarrow e^{-t^2/2}$ for every t , and Lévy's continuity theorem gives $Z_n \xrightarrow{d} N(0, 1)$. $\square$

Remark 5.4 (Reading the theorem). (i) *Universality*. The limit depends on the distribution of X_1 only through μ and σ^2 . In the proof this appears as $\log \varphi(t) = -t^2/2 + o(t^2)$: after the scaling $t \mapsto t/\sqrt{n}$ and raising to the n th power, only the quadratic term survives.

(ii) *Why $\sqrt{n}$* . $\text{Var}(S_n) = n\sigma^2$, so $\sqrt{n}$ is the only normalization giving a nondegenerate limit with finite variance.

(iii) *Uniformity*. Because Φ is continuous, the convergence is uniform in x (Pólya's theorem): $\sup_x |\mathbb{P}(Z_n \leq x) - \Phi(x)| \rightarrow 0$. The CLT gives no rate; that is the Berry–Esseen theorem below.

(iv) *Mode of convergence*. The convergence is in distribution only. Z_n does *not* converge in probability, let alone almost surely. Indeed (with $\mu = 0, \sigma = 1$) $Z_{2n} - Z_n = (\frac{1}{\sqrt{2}} - 1)Z_n + (S_{2n} - S_n)/\sqrt{2n}$ is a sum of two independent terms and converges in distribution to $N(0, 2 - \sqrt{2})$, so (Z_n) is not Cauchy in probability. Section 6 describes the almost-sure behavior.

Remark 5.5 (Beyond i.i.d.: Lindeberg–Feller). The same proof strategy handles independent, non-identically distributed summands. Let $X_{n,1}, \dots, X_{n,k_n}$ be independent with mean zero and $\sum_k \mathbb{E} X_{n,k}^2 = 1$ for each n . If $\sum_k \mathbb{E}[X_{n,k}^2 \mathbf{1}\{|X_{n,k}| > \varepsilon\}] \rightarrow 0$ for every $\varepsilon > 0$ (the *Lindeberg condition*), then $\sum_k X_{n,k} \xrightarrow{d} N(0, 1)$. The condition says that no single summand contributes a non-negligible share of the total variance.

5.3 Rate of convergence: the Berry–Esseen theorem

Theorem 5.6 (Berry–Esseen). *Let $X_1, X_2, \dots$ be i.i.d. with mean μ , variance $\sigma^2 > 0$ and finite third absolute central moment $\rho = \mathbb{E}|X_1 - \mu|^3$. Let F_n be the distribution function of $Z_n = (S_n - n\mu)/(\sigma\sqrt{n})$. Then for all $n \geq 1$,*

$$\sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{C \rho}{\sigma^3 \sqrt{n}},$$

where C is an absolute constant.

The best possible constant is not known. The inequality holds with $C = 0.4748$ (Shevtsova, 2011), and Esseen showed that it cannot hold with any $C < \frac{\sqrt{10+3}}{6\sqrt{2\pi}} \approx 0.4097$.

Remark 5.7. (i) The bound is *non-asymptotic*: it holds for every n , with an explicit constant.

(ii) The rate $n^{-1/2}$ cannot be improved in general. For fair ± 1 steps and even n , $\mathbb{P}(S_n = 0) \sim \sqrt{2/(\pi n)}$, so F_n has a jump of that size at 0 while Φ is continuous; hence $\sup_x |F_n(x) - \Phi(x)| \gtrsim 1/\sqrt{2\pi n}$.

- (iii) For independent, non-identically distributed, mean-zero summands with variances σ_k^2 and third absolute moments ρ_k , the distribution function G_n of $S_n/(\sum_k \sigma_k^2)^{1/2}$ satisfies

$$\sup_{x \in \mathbb{R}} |G_n(x) - \Phi(x)| \leq \frac{C_0 \sum_{k=1}^n \rho_k}{(\sum_{k=1}^n \sigma_k^2)^{3/2}},$$

where one may take $C_0 = 0.5600$.

- (iv) The bound controls the *absolute* error. In the tails, where $1 - \Phi(x) \ll n^{-1/2}$, it says nothing useful about the *relative* error. Tail probabilities at that depth are the subject of large deviations theory (see the last row of Table 1).

5.4 Higher-order corrections: the Edgeworth expansion

Berry–Esseen bounds the error of the normal approximation. The Edgeworth expansion describes the error: it is a series in powers of $n^{-1/2}$ whose coefficients involve the higher cumulants of X_1 . Let

$$\kappa_3 = \frac{\mathbb{E}(X_1 - \mu)^3}{\sigma^3} \quad (\text{skewness}), \quad \kappa_4 = \frac{\mathbb{E}(X_1 - \mu)^4}{\sigma^4} - 3 \quad (\text{excess kurtosis}),$$

and let He_k be the (probabilists') Hermite polynomials, defined by $\phi^{(k)}(x) = (-1)^k He_k(x)\phi(x)$:

$$He_2(x) = x^2 - 1, \quad He_3(x) = x^3 - 3x, \quad He_5(x) = x^5 - 10x^3 + 15x.$$

Theorem 5.8 (Edgeworth expansion). *Let $X_1, X_2, \dots$ be i.i.d. with variance $\sigma^2 > 0$, and let F_n be as in Theorem 5.6.*

- (a) *If $\mathbb{E}|X_1|^3 < \infty$ and the distribution of X_1 is not a lattice distribution, then, uniformly in x ,*

$$F_n(x) = \Phi(x) - \phi(x) \frac{\kappa_3 He_2(x)}{6\sqrt{n}} + o(n^{-1/2}).$$

- (b) *If $\mathbb{E}X_1^4 < \infty$ and Cramér's condition $\limsup_{|t| \rightarrow \infty} |\varphi_{X_1}(t)| < 1$ holds (it holds, for example, whenever X_1 has a density), then, uniformly in x ,*

$$F_n(x) = \Phi(x) - \phi(x) \left[\frac{\kappa_3 He_2(x)}{6\sqrt{n}} + \frac{1}{n} \left(\frac{\kappa_4 He_3(x)}{24} + \frac{\kappa_3^2 He_5(x)}{72} \right) \right] + o(n^{-1}).$$

Where the terms come from. With $\mu = 0$, $\sigma = 1$, expand the logarithm of the characteristic function one step further than in the CLT proof:

$$\log \varphi_{Z_n}(t) = n \log \varphi\left(\frac{t}{\sqrt{n}}\right) = -\frac{t^2}{2} + \frac{\kappa_3(it)^3}{6\sqrt{n}} + \frac{\kappa_4(it)^4}{24n} + \dots$$

Exponentiating,

$$\varphi_{Z_n}(t) = e^{-t^2/2} \left[1 + \frac{\kappa_3(it)^3}{6\sqrt{n}} + \frac{1}{n} \left(\frac{\kappa_4(it)^4}{24} + \frac{\kappa_3^2(it)^6}{72} \right) + \dots \right].$$

Since $(it)^k e^{-t^2/2}$ is the Fourier transform of $(-1)^k \phi^{(k)}(x) = He_k(x)\phi(x)$, inverting term by term gives the density expansion

$$f_n(x) \approx \phi(x) \left[1 + \frac{\kappa_3 He_3(x)}{6\sqrt{n}} + \frac{1}{n} \left(\frac{\kappa_4 He_4(x)}{24} + \frac{\kappa_3^2 He_6(x)}{72} \right) \right],$$

and integrating with $\int_{-\infty}^x He_k(y)\phi(y) dy = -He_{k-1}(x)\phi(x)$ gives Theorem 5.8. Making the term-by-term inversion rigorous is what requires the smoothness (non-lattice or Cramér) conditions.

- Remark 5.9.* (i) The leading correction is a *skewness* correction of order $n^{-1/2}$; *kurtosis* enters at order n^{-1} . If the distribution is symmetric ($\kappa_3 = 0$) and satisfies the conditions of (b), the normal approximation is accurate to $O(1/n)$ rather than $O(1/\sqrt{n})$.
- (ii) For lattice distributions the expansion in (a) fails: there is an additional discontinuous term of order $n^{-1/2}$ coming from the jumps of F_n . This is the same phenomenon as in the remark after Theorem 5.6.
- (iii) The expansion is asymptotic in n , not a convergent series, and the truncated “density” can be negative in the tails. It is most accurate near the center of the distribution.

Example 5.10 (Sum of exponentials). Let X_i be i.i.d. $\text{Exp}(1)$, so $\mu = \sigma = 1$ and $\kappa_3 = 2$, and take $n = 10$. Then $S_{10} \sim \text{Gamma}(10, 1)$ and the exact value is $\mathbb{P}(S_{10} \leq 10) = 0.5421$. The CLT gives $\Phi(0) = 0.5$. The one-term Edgeworth correction gives

$$\Phi(0) - \phi(0) \frac{\kappa_3 \text{He}_2(0)}{6\sqrt{10}} = 0.5 + \frac{0.3989 \times 2}{6\sqrt{10}} = 0.5421,$$

which agrees with the exact value to four decimals.

6 The law of the iterated logarithm and other scaling laws

6.1 What the SLLN and CLT leave open

Let the X_i be i.i.d. with mean 0 and variance $\sigma^2 \in (0, \infty)$. The SLLN says $S_n/n \rightarrow 0$ a.s., so $|S_n|$ grows more slowly than n . The CLT says $S_n/\sqrt{n}$ is of order one *in distribution*, for each fixed large n . Neither tells us how large S_n gets *along a single sample path*. A first answer comes from the zero-one law.

Proposition 6.1. *Under the assumptions above, $\limsup_n S_n/\sqrt{n} = +\infty$ and $\liminf_n S_n/\sqrt{n} = -\infty$ a.s. In particular $S_n/\sqrt{n}$ does not converge almost surely.*

Proof. Fix $M > 0$ and let $A_n = \{S_n/\sqrt{n} > M\}$. Then $\mathbb{P}(A_n \text{ i.o.}) \geq \limsup_n \mathbb{P}(A_n) = 1 - \Phi(M/\sigma) > 0$ by the CLT. Since $\{A_n \text{ i.o.}\} \subseteq \{\limsup_n S_n/\sqrt{n} \geq M\}$ and the latter is a tail event (Example 2.5) of positive probability, it has probability 1 by Theorem 3.1. Let $M \rightarrow \infty$ along the integers. The $\liminf$ statement follows by applying this to $-X_i$. $\square$

So $\sqrt{n}$ is too small a normalization for almost sure boundedness, and n is too large. The LIL finds the exact boundary.

6.2 The law of the iterated logarithm

Theorem 6.2 (Hartman–Wintner law of the iterated logarithm). *Let $X_1, X_2, \dots$ be i.i.d. with mean 0 and variance $\sigma^2 \in (0, \infty)$. Then, almost surely,*

$$\limsup_{n \rightarrow \infty} \frac{S_n}{\sqrt{2n \log \log n}} = \sigma \quad \text{and} \quad \liminf_{n \rightarrow \infty} \frac{S_n}{\sqrt{2n \log \log n}} = -\sigma.$$

Moreover, almost surely, the set of limit points of the sequence $S_n/\sqrt{2n \log \log n}$ is the whole interval $[-\sigma, \sigma]$.

For variables with mean $\mu \neq 0$, apply the theorem to $X_i - \mu$.

Meaning. Write $h(n) = \sigma\sqrt{2n \log \log n}$. The theorem is equivalent to the following two statements, valid for every $\varepsilon > 0$ with probability one:

- (*upper envelope*) $S_n \leq (1 + \varepsilon)h(n)$ for all sufficiently large n ;
- (*attained*) $S_n \geq (1 - \varepsilon)h(n)$ for infinitely many n .

The same holds for $-S_n$. So the curves $\pm h(n)$ are the sharp almost-sure envelope of the random walk: slightly wider curves are eventually never crossed, slightly narrower curves are crossed infinitely often. The clause on limit points says that the normalized walk does not settle near $\pm\sigma$ either; it keeps sweeping across the entire band.

Relation to the zero-one law. By Example 2.5 and Corollary 3.4, $\limsup_n S_n/\sqrt{2n \log \log n}$ must be an a.s. constant. The content of the LIL is that this constant is finite, nonzero, and equal to σ .

Relation to the CLT. For fixed n , the CLT says S_n is typically of size $\sigma\sqrt{n}$, so

$$\frac{S_n}{\sqrt{2n \log \log n}} \rightarrow 0 \quad \text{in probability,} \quad \text{while} \quad \limsup_{n \rightarrow \infty} \frac{S_n}{\sqrt{2n \log \log n}} = \sigma \quad \text{a.s.}$$

There is no contradiction. At any given large time the walk is very likely to be far inside the envelope; excursions out to the envelope are rare, but they recur forever. This is a natural example of a sequence that converges in probability but not almost surely. The CLT describes a snapshot at one time; the LIL describes the whole path.

Intuition: where $\log \log n$ comes from. The values S_n and S_{n+1} are almost identical, so along a path there are far fewer than n “independent looks” at the walk by time n . The walk decorrelates only when time is multiplied by a constant factor. Sample it along a geometric sequence $n_k = \theta^k$, $\theta > 1$: the values $S_{n_k}/\sqrt{n_k}$ are approximately independent and approximately $N(0, \sigma^2)$, and up to time n there are only about $\log n$ of them. Using the Gaussian tail $\mathbb{P}(N(0, 1) > x) \approx e^{-x^2/2}$,

$$\mathbb{P}\left(S_{n_k} > \sigma\sqrt{2\lambda n_k \log \log n_k}\right) \approx e^{-\lambda \log \log n_k} = (k \log \theta)^{-\lambda} \asymp k^{-\lambda}.$$

The series $\sum_k k^{-\lambda}$ converges if $\lambda > 1$ and diverges if $\lambda \leq 1$. By the first Borel–Cantelli lemma, levels with $\lambda > 1$ are exceeded only finitely often; by the second (applied to the nearly independent increments $S_{n_k} - S_{n_{k-1}}$), levels with $\lambda < 1$ are exceeded infinitely often. The threshold $\lambda = 1$ gives the constant $\sqrt{2}$. Equivalently: the maximum of m independent $N(0, 1)$ variables is about $\sqrt{2 \log m}$, and here $m \approx \log n$. The rigorous proof follows this outline, with a maximal inequality to control the walk between the times n_k .

How slow is $\log \log$? The envelope is only barely wider than $\sqrt{n}$: $\sqrt{2 \log \log n} \approx 2.29$ for $n = 10^6$ and ≈ 3.30 for $n = 10^{100}$. Figure 1 shows one simulated path.

Remark 6.3. The finite-variance assumption is sharp: if $X_1, X_2, \dots$ are i.i.d. and

$$\limsup_{n \rightarrow \infty} \frac{|S_n|}{\sqrt{2n \log \log n}} < \infty \quad \text{with positive probability,}$$

then $\mathbb{E}X_1^2 < \infty$ and $\mathbb{E}X_1 = 0$ (Strassen).

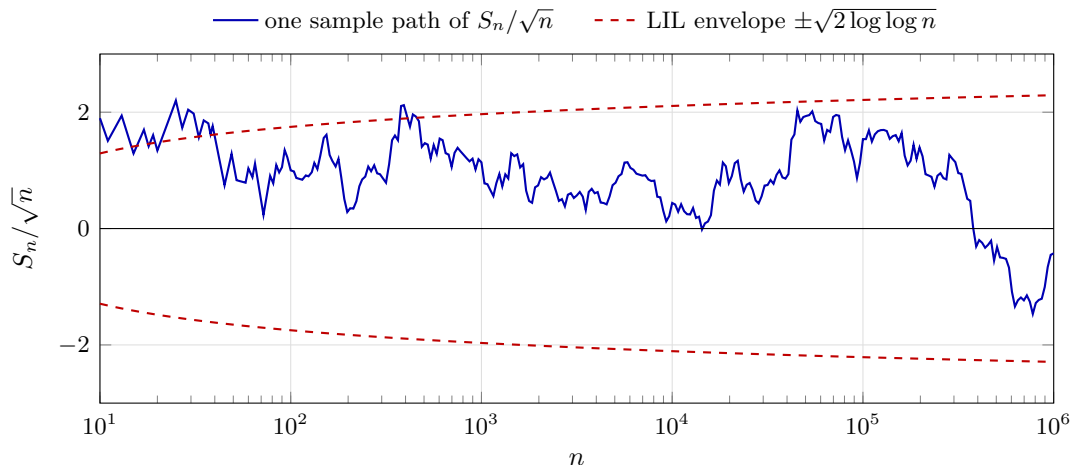


Figure 1: A simulated fair ± 1 random walk ($\sigma = 1$), plotted as $S_n/\sqrt{n}$ against n on a logarithmic time axis, with the LIL envelope. At a typical time the path is well inside the envelope (CLT); the LIL says that it comes arbitrarily close to both dashed curves infinitely often, and exceeds any wider multiple of them only finitely often. The statement is asymptotic, so brief crossings at small n are not excluded.

6.3 Other scaling laws

Table 1 places the LIL among the main limit theorems for $S_n - n\mu$. Each row answers the question “what happens to $S_n - n\mu$ at scale a_n ?” for a different a_n and a different mode of convergence.

Scale a_n	Name	Statement (i.i.d., $\mathbb{E}X_1 = \mu$)	Mode
n	Strong law	$(S_n - n\mu)/n \rightarrow 0$	a.s.
$n^{1/p}$, $1 \leq p < 2$	Marcinkiewicz–Zygmund law	$(S_n - n\mu)/n^{1/p} \rightarrow 0$ iff $\mathbb{E} X_1 ^p < \infty$	a.s.
$\sqrt{2n \log \log n}$	Law of the iterated logarithm	$\limsup = \sigma$, $\liminf = -\sigma$	a.s.
$\sqrt{n}$	Central limit theorem	$(S_n - n\mu)/(\sigma\sqrt{n}) \xrightarrow{d} N(0, 1)$	in dist.
$\sqrt{n} \ll a_n \ll n$	Moderate deviations	$\frac{n}{a_n^2} \log \mathbb{P}(S_n - n\mu \geq a_n x) \rightarrow -\frac{x^2}{2\sigma^2}$	tail prob.
n	Large deviations (Cramér)	$\frac{1}{n} \log \mathbb{P}(S_n \geq nx) \rightarrow -I(x)$, $x > \mu$	tail prob.

Table 1: Scaling laws for sums of i.i.d. random variables. Rows 3–4 assume $\sigma^2 < \infty$; rows 5–6 assume a finite moment generating function near 0, and $x > 0$ in row 5. In row 6, $I(x) = \sup_{\theta} \{\theta x - \log \mathbb{E} e^{\theta X_1}\}$ is the rate function.

Some further scaling laws worth knowing:

- **Heavy tails and stable laws.** If the variance is infinite, the $\sqrt{n}$ scaling fails. If X_1 is symmetric with $\mathbb{P}(|X_1| > x) \sim cx^{-\alpha}$ for some $0 < \alpha < 2$, then $S_n/n^{1/\alpha}$ converges in distribution to a symmetric α -stable law, which is not Gaussian. The sum is then dominated by its few largest terms. The Cauchy case is $\alpha = 1$, with scaling n .

- **Chung’s “other” LIL.** The LIL describes how *large* the walk gets; Chung’s law describes how *small* its running maximum can stay:

$$\liminf_{n \rightarrow \infty} \sqrt{\frac{\log \log n}{n}} \max_{1 \leq k \leq n} |S_k| = \frac{\pi \sigma}{\sqrt{8}} \quad \text{a.s.}$$

- **Functional versions.** Donsker’s theorem upgrades the CLT to paths: $t \mapsto S_{\lfloor nt \rfloor} / (\sigma \sqrt{n})$ converges in distribution to standard Brownian motion B on $[0, 1]$. Brownian motion obeys its own LIL, both at infinity and at zero: almost surely,

$$\limsup_{t \rightarrow \infty} \frac{B_t}{\sqrt{2t \log \log t}} = 1 \quad \text{and} \quad \limsup_{t \downarrow 0} \frac{B_t}{\sqrt{2t \log \log(1/t)}} = 1.$$

- **Extremes.** For comparison, let M_n be the maximum of n i.i.d. standard normal variables. Then $M_n / \sqrt{2 \log n} \rightarrow 1$ a.s. The LIL has the same form, with n replaced by the effective number $\log n$ of independent looks.

7 Probabilistic inequalities

Limit theorems say what happens as $n \rightarrow \infty$. Inequalities give guarantees that hold as stated, for every n and under minimal assumptions.

7.1 Markov’s inequality

Theorem 7.1 (Markov). *If $X \geq 0$ and $a > 0$, then*

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

Proof. Pointwise, $a \mathbf{1}\{X \geq a\} \leq X$. Take expectations. □

A useful generalization: if $g \geq 0$ is nondecreasing on the range of X and $g(a) > 0$, then $\mathbb{P}(X \geq a) \leq \mathbb{P}(g(X) \geq g(a)) \leq \mathbb{E}g(X)/g(a)$. The choice of g is where all the flexibility lies: $g(x) = x^2$ gives Chebyshev, $g(x) = |x|^p$ gives moment bounds, and $g(x) = e^{\theta x}$ gives the Chernoff bound (Example 7.6).

Example 7.2 (Markov is tight). Fix $a > m > 0$ and let $X = a$ with probability m/a and $X = 0$ otherwise. Then $\mathbb{E}X = m$ and $\mathbb{P}(X \geq a) = m/a$, with equality. Markov’s inequality cannot be improved using the mean alone.

7.2 Chebyshev’s inequality

Theorem 7.3 (Chebyshev). *If X has mean μ and variance $\sigma^2 < \infty$, then for $a > 0$*

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Proof. Apply Markov’s inequality to $(X - \mu)^2$ at level a^2 . □

Equivalently, $\mathbb{P}(|X - \mu| \geq k\sigma) \leq 1/k^2$: any random variable with finite variance is within k standard deviations of its mean with probability at least $1 - 1/k^2$. There is also a one-sided version (*Cantelli's inequality*):

$$\mathbb{P}(X - \mu \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2}, \quad a > 0,$$

obtained by applying Markov to $(X - \mu + u)^2$ and optimizing over $u \geq 0$ (the optimum is $u = \sigma^2/a$).

Example 7.4 (Weak law of large numbers). If $X_1, \dots, X_n$ are uncorrelated with common mean μ and variance σ^2 , then $\text{Var}(S_n/n) = \sigma^2/n$ and Chebyshev gives

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0.$$

Example 7.5 (How many samples? Polling). We estimate an unknown proportion p by the sample mean $\hat{p} = S_n/n$ of n i.i.d. Bernoulli(p) observations, and want $\mathbb{P}(|\hat{p} - p| \geq \varepsilon) \leq \delta$ with $\varepsilon = \delta = 0.05$. Since $\text{Var}(\hat{p}) = p(1-p)/n \leq 1/(4n)$:

- *Chebyshev*: $\frac{1}{4n\varepsilon^2} \leq \delta$ requires $n \geq \frac{1}{4\varepsilon^2\delta} = 2000$.
- *Hoeffding's inequality* (stated here without proof) says that $\mathbb{P}(|\hat{p} - p| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$ for independent $[0, 1]$ -valued observations, so $2e^{-2n\varepsilon^2} \leq \delta$ requires $n \geq \frac{\log(2/\delta)}{2\varepsilon^2} \approx 738$.
- *CLT approximation*: $n \approx (1.96)^2/(4\varepsilon^2) \approx 385$.

The first two are guarantees valid for every p and n . The third is an approximation; it is the smallest because it uses the exact limiting shape, but it comes with no error bound unless combined with Berry–Esseen.

Example 7.6 (Chernoff bound). For any $\theta > 0$, applying Markov to $e^{\theta X}$ gives $\mathbb{P}(X \geq a) \leq e^{-\theta a} \mathbb{E}e^{\theta X}$, and hence

$$\mathbb{P}(X \geq a) \leq \inf_{\theta > 0} e^{-\theta a} M_X(\theta), \quad M_X(\theta) = \mathbb{E}e^{\theta X}.$$

For $Z \sim N(0, 1)$, $M_Z(\theta) = e^{\theta^2/2}$ and the optimum is $\theta = a$, giving $\mathbb{P}(Z \geq a) \leq e^{-a^2/2}$ for $a > 0$: exponentially better than the $1/a^2$ of Chebyshev.

Example 7.7 (Comparing the bounds). Let S be the number of heads in 100 fair coin tosses, so $\mathbb{E}S = 50$ and $\text{Var}(S) = 25$. Bounds on $\mathbb{P}(S \geq 75)$:

Method	Bound	Value
Markov	50/75	0.67
Chebyshev (two-sided)	25/25 ²	4.0 × 10 ⁻²
Cantelli (one-sided)	25/(25 + 25 ²)	3.8 × 10 ⁻²
Hoeffding	$e^{-2 \cdot 25^2/100} = e^{-12.5}$	3.7 × 10 ⁻⁶
Chernoff (optimized)	$e^{-100 D(0.75 \ 0.5)}$	2.1 × 10 ⁻⁶
Exact	$\sum_{k \geq 75} \binom{100}{k} 2^{-100}$	2.8 × 10 ⁻⁷

Here $D(x \| p) = x \log \frac{x}{p} + (1-x) \log \frac{1-x}{1-p}$, and the Hoeffding row uses the one-sided form $\mathbb{P}(\hat{p} - p \geq \varepsilon) \leq e^{-2n\varepsilon^2}$ of the inequality in Example 7.5 with $n = 100$, $\varepsilon = 0.25$. Using more information about the distribution (mean; variance; boundedness; the full moment generating function) gives progressively better bounds.

7.3 Jensen's inequality

Theorem 7.8 (Jensen). *Let $g : I \rightarrow \mathbb{R}$ be convex on an interval I , and let X be a random variable with values in I and $\mathbb{E}|X| < \infty$. Then*

$$g(\mathbb{E}X) \leq \mathbb{E}g(X).$$

If g is strictly convex, equality holds if and only if X is a.s. constant. For concave g the inequality is reversed.

Proof. Let $m = \mathbb{E}X \in I$. A convex function has a supporting line at every point: there is a slope c with $g(x) \geq g(m) + c(x - m)$ for all $x \in I$. Substitute $x = X$ and take expectations: $\mathbb{E}g(X) \geq g(m) + c(\mathbb{E}X - m) = g(\mathbb{E}X)$. $\square$

The same proof gives the conditional version $g(\mathbb{E}[X | \mathcal{G}]) \leq \mathbb{E}[g(X) | \mathcal{G}]$ a.s.

Example 7.9 (Standard consequences). (a) $g(x) = x^2$: $(\mathbb{E}X)^2 \leq \mathbb{E}X^2$, that is, $\text{Var}(X) \geq 0$. Also $|\mathbb{E}X| \leq \mathbb{E}|X|$.

(b) *Lyapunov's inequality.* For $0 < r < s$, applying Jensen to $g(x) = x^{s/r}$ and the variable $|X|^r$ gives $(\mathbb{E}|X|^r)^{1/r} \leq (\mathbb{E}|X|^s)^{1/s}$. So $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$ is nondecreasing in p , and a finite s th moment implies finite lower moments.

(c) $g(x) = 1/x$ on $(0, \infty)$: $\mathbb{E}[1/X] \geq 1/\mathbb{E}X$ for $X > 0$. For instance, the mean time to travel a fixed distance at a random speed is at least the distance divided by the mean speed.

(d) $g = \log$ (concave): $\mathbb{E} \log X \leq \log \mathbb{E}X$ for $X > 0$. Taking X uniform on $\{x_1, \dots, x_n\}$ gives the arithmetic-geometric mean inequality $(x_1 \cdots x_n)^{1/n} \leq \frac{1}{n} \sum_k x_k$.

(e) *Information inequality.* For probability mass functions p, q on a common alphabet,

$$-D(p||q) = \sum_x p(x) \log \frac{q(x)}{p(x)} \leq \log \sum_x p(x) \frac{q(x)}{p(x)} \leq \log 1 = 0,$$

so the relative entropy satisfies $D(p||q) \geq 0$, with equality iff $p = q$. Taking q uniform on an alphabet $\mathcal{X}$ gives the entropy bound $H(p) \leq \log |\mathcal{X}|$.

7.4 Hölder, Cauchy–Schwarz and Minkowski (statements)

Write $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$ for $p \geq 1$.

Theorem 7.10 (Hölder's inequality). *Let $p, q \in (1, \infty)$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$\mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|Y|^q)^{1/q}, \quad \text{that is,} \quad \|XY\|_1 \leq \|X\|_p \|Y\|_q.$$

Theorem 7.11 (Cauchy–Schwarz inequality).

$$|\mathbb{E}[XY]| \leq \mathbb{E}|XY| \leq \sqrt{\mathbb{E}X^2} \sqrt{\mathbb{E}Y^2}.$$

Theorem 7.12 (Minkowski's inequality). *For $p \geq 1$,*

$$(\mathbb{E}|X + Y|^p)^{1/p} \leq (\mathbb{E}|X|^p)^{1/p} + (\mathbb{E}|Y|^p)^{1/p}, \quad \text{that is,} \quad \|X + Y\|_p \leq \|X\|_p + \|Y\|_p.$$

Cauchy–Schwarz is the case $p = q = 2$ of Hölder; applied to centered variables it gives $|\text{Cov}(X, Y)| \leq \sigma_X \sigma_Y$, so correlation coefficients lie in $[-1, 1]$. Minkowski's inequality is the triangle inequality for $\|\cdot\|_p$, which makes L^p a normed space.

Further reading

- B. Hajek, *Random Processes for Engineers*, Cambridge University Press. (Convergence of random variables, characteristic functions, CLT, and the Chernoff bound.)
- R. Durrett, *Probability: Theory and Examples*, Cambridge University Press. (Zero-one laws, SLLN, CLT, LIL.)
- P. Billingsley, *Probability and Measure*, Wiley. (π - λ theorem, generated σ -algebras, characteristic functions.)
- W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, Wiley. (Berry-Esseen and Edgeworth expansions.)