Review: Probability Theory

SDS 391P.6, Spring 2026

Pratik Patil

These notes are a work in progress and are provided as-is for instructional purposes only. They are not (yet) at the level of a scholarly document. In particular, the notes draw from various sources and do not (yet) have sufficient references to the original sources. Additionally, almost surely the notes have errors and they are only probably approximately correct. The notes will be updated regularly as the course progresses. Last updated: 2026-02-02.

1 Random variables and their properties

1.1 Basic notation

A random variable X is a real-valued function on an underlying probability space. We write $\mathbb{P}(E)$ for the probability of an event E, and $\mathbb{E}[X]$ for the expectation of X. Linearity of expectation gives:

$$\mathbb{E}\left[\sum_{i=1}^{m} a_i X_i\right] = \sum_{i=1}^{m} a_i \mathbb{E}[X_i] \qquad \text{for any random variables } X_i \text{ and scalars } a_i.$$

The variance and standard deviation of X are:

$$\mathrm{Var}(X) := \mathbb{E}\big[(X - \mathbb{E}X)^2\big] = \mathbb{E}[X^2] - (\mathbb{E}X)^2, \qquad \sigma(X) := \sqrt{\mathrm{Var}(X)}.$$

If $X_1, \dots, X_m$ are independent (or just uncorrelated), then variance is additive:

$$\mathrm{Var}\left(\sum_{i=1}^{m} a_i X_i\right) = \sum_{i=1}^{m} a_i^2 \,\mathrm{Var}(X_i).$$

For two random variables X, Y, the covariance is

$$\mathrm{Cov}(X, Y) := \mathbb{E}\big[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\big] = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y.$$

Covariance is bilinear in each argument (after centering), and symmetric: $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$. If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$. The converse is false in general: $\mathrm{Cov}(X, Y) = 0$ does *not* imply independence.

For any scalars a, b,

$$\mathrm{Var}(aX + bY) = a^2 \,\mathrm{Var}(X) + b^2 \,\mathrm{Var}(Y) + 2ab \,\mathrm{Cov}(X, Y).$$

A distribution can be described by its cumulative distribution function (CDF) $F_X(t) = \mathbb{P}\{X \leqslant t\}$. In many arguments it is more convenient to work with the tail $\mathbb{P}\{X > t\} = 1 - F_X(t)$.

1.2 Some canonical distributions

A large fraction of statistical models are built from a small number of "atomic" distributions.

- Some discrete distributions:
  - Bernoulli $X \sim \mathrm{Ber}(p)$: $X \in \{0, 1\}$ with $\mathbb{P}\{X = 1\} = p$, $\mathbb{P}\{X = 0\} = 1 - p$. Then $\mathbb{E}X = p$ and $\mathrm{Var}(X) = p(1 - p)$. (Models binary labels; logistic regression and classification.)
  - Binomial $X \sim \mathrm{Binom}(m, p)$: $X = \sum_{i=1}^{m} X_i$ with $X_i \sim \mathrm{Ber}(p)$ i.i.d. Then $X \in \{0, \ldots, m\}$ and
    $$\mathbb{P}\{X = k\} = \binom{m}{k} p^k (1 - p)^{m-k}.$$
    (Models counts out of m trials; proportions.)
  - Poisson $X \sim \mathrm{Pois}(\lambda)$: $X \in \{0, 1, 2, \ldots\}$ with
    $$\mathbb{P}\{X = k\} = e^{-\lambda} \frac{\lambda^k}{k!}.$$
    (Models count data; Poisson regression; "rare event" limits; see also Section 5.2)
- Some continuous distributions:
  - Normal/Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$: density
    $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right).$$
    (Noise model; least squares; CLT; Gaussian priors and random features.)
  - Laplace $X \sim \mathrm{Laplace}(\mu, b)$: density
    $$f(x) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right).$$
    (Heavier tails than Gaussian; robust noise; ℓ_1 sparsity connections.)
  - Chi-square $X \sim \chi_k^2$: distribution of $\sum_{i=1}^{k} Z_i^2$ with $Z_i \sim \mathcal{N}(0, 1)$ i.i.d. (Shows up in Gaussian norms, variance estimation, and quadratic forms.)

1.3 L_p norms

Two families of summary quantities appear throughout high-dimensional statistics:

- The moment generating function (MGF): $M_X(t) := \mathbb{E}e^{tX}$, when finite, encodes moments and tails. Some comments:
  - For heavy-tailed variables, $M_X(t)$ may be $+\infty$ for all $t > 0$.
  - If $M_X(t)$ is finite on an open interval around 0 and is differentiable there, then
    $$M_X^{(k)}(0) = \mathbb{E}[X^k].$$
    So an MGF packages all moments into a single function.

– If X and Y are independent, then

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

Equivalently, the *log MGF* (also known as the cumulant generating function)

$$\Lambda_X(t) := \log \mathbb{E}e^{tX}$$

satisfies $\Lambda_{X+Y}(t) = \Lambda_X(t) + \Lambda_Y(t)$ for independent X, Y. This additivity is one of the main reasons MGFs are so useful for concentration bounds.

- The L_p norms of X, defined for $p > 0$ by

$$\|X\|_{L_p} := \left(\mathbb{E}|X|^p\right)^{1/p}, \qquad \|X\|_{L_\infty} := \operatorname{ess\,sup} |X|.$$

  Some comments:

  – For $p < 1$, the triangle inequality fails, so $\|\cdot\|_{L_p}$ is not a norm.

  – For $p \geqslant 1$, $\|\cdot\|_{L_p}$ is a norm on the space of random variables with finite L_p norm.

  – The exponent $p = 2$ is special: L_2 is a Hilbert space with inner product $\langle X, Y\rangle_{L_2} := \mathbb{E}[XY]$ and norm $\|X\|_{L_2} = (\mathbb{E}|X|^2)^{1/2}$. With this viewpoint, $\operatorname{Cov}(X, Y)$ is just the L_2 inner product of the centered variables $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$, so covariance measures geometric alignment in L_2.

Just as we recalled for ℓ_p spaces in the last lecture, the following key inequalities appear naturally as soon as we introduce L_p norm:

Cauchy–Schwarz, Hölder, and Minkowski. For $X, Y \in L_2$, $|\mathbb{E}[XY]| \leqslant \|X\|_{L_2}\|Y\|_{L_2}$. More generally, if $p, p' \in [1, \infty]$ are conjugate exponents $\frac{1}{p} + \frac{1}{p'} = 1$, then

$$|\mathbb{E}[XY]| \leqslant \|X\|_{L_p}\|Y\|_{L_{p'}}.$$

Minkowski's inequality is the triangle inequality in L_p: for $p \geqslant 1$, $\|X + Y\|_{L_p} \leqslant \|X\|_{L_p} + \|Y\|_{L_p}$.

Jensen. If $\phi : \mathbb{R} \to \mathbb{R}$ is convex and X is integrable, then

$$\phi(\mathbb{E}X) \leqslant \mathbb{E}\phi(X).$$

Two quick consequences that are often used implicitly:

- If $\|\cdot\|$ is any norm on $\mathbb{R}^d$, then $\|\mathbb{E}X\| \leqslant \mathbb{E}\|X\|$ (since norms are convex).
- The L_p norms of a random variable are *increasing* in p: if $0 < p \leqslant q \leqslant \infty$, then $\|X\|_{L_p} \leqslant \|X\|_{L_q}$.

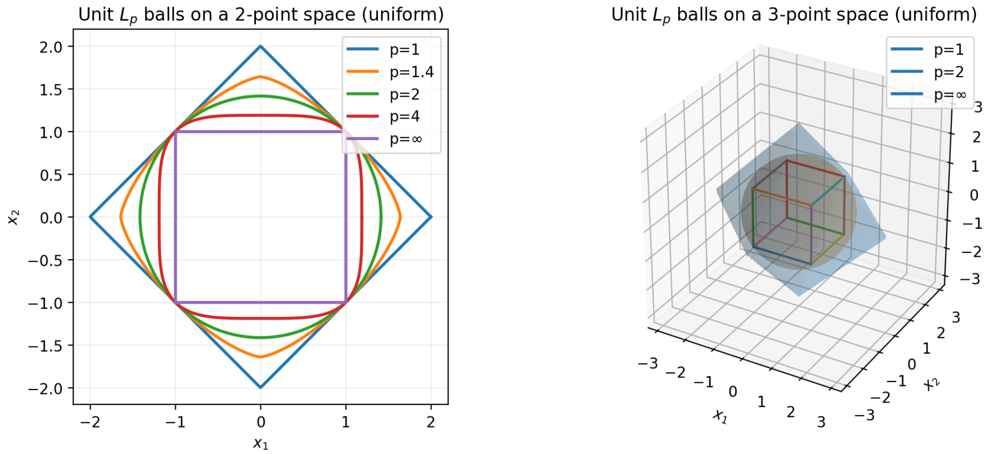

Figure 1: Unit L_p balls on a finite probability space with uniform measure (2-point case on the left, 3-point case on the right). As p increases, the balls become rounder but *shrink* (since $\|X\|_{L_p}$ increases with p).

1.4 Contrasting ℓ_p and L_p norms

The ℓ_p norms of a fixed vector in $\mathbb{R}^n$ are *decreasing* in p, while the L_p norms of a fixed random variable are *increasing* in p. It may seem contradictory, but there is no contradiction: the two are normalized differently.

Indeed, let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and let X be a random variable that takes values $x_1, \ldots, x_n$ each with probability $1/n$. Then

$$\|x\|_{\ell_p} = \Big(\sum_{i=1}^n |x_i|^p\Big)^{1/p}, \qquad \|X\|_{L_p} = \Big(\frac{1}{n}\sum_{i=1}^n |x_i|^p\Big)^{1/p} = n^{-1/p}\|x\|_{\ell_p}.$$

So the L_p unit ball on an n-point uniform space is just a scaled ℓ_p ball:

$$\|X\|_{L_p} \leqslant 1 \quad \iff \quad \|x\|_{\ell_p} \leqslant n^{1/p}.$$

The factor $n^{-1/p}$ flips the monotonicity direction when you compare across p.

On a finite probability space with n atoms (in particular, in the discrete uniform case above), all L_p norms are equivalent up to constants depending on n. On a general probability space, there is *no* dimension parameter to save you: for $p < q$ it is possible to have $X \in L_p$ but $X \notin L_q$ (heavy tails).

1.5 Generalizing L_p norms: Orlicz norms

In high-dimensional probability and random matrix theory, L_p norms are often too crude: we want *uniform* tail control across all p, which leads to Orlicz norms.

An Orlicz function is a convex, increasing function $\psi : [0, \infty) \to [0, \infty)$ with $\psi(0) = 0$ and $\psi(x) \to \infty$ as $x \to \infty$. Given ψ, define the Orlicz norm

$$\|X\|_\psi := \inf\Big\{t > 0 : \mathbb{E}\psi(|X|/t) \leqslant 1\Big\}.$$

The Orlicz space L_ψ is the set of X with $\|X\|_\psi < \infty$.

Some examples:

- If $\psi(x) = x^p$ with $p \geqslant 1$, then $\|\cdot\|_\psi$ is exactly $\|\cdot\|_{L_p}$.
- If $\psi_2(x) = e^{x^2} - 1$, then $\|\cdot\|_{\psi_2}$ is the *sub-Gaussian norm*.
- If $\psi_1(x) = e^x - 1$, then $\|\cdot\|_{\psi_1}$ is the *sub-exponential norm*.

1.6 All norms at once

One can locate sub-Gaussian and sub-exponential variables between L_∞ and all L_p spaces:

$$L_\infty \subset L_{\psi_2} \subset L_{\psi_1} \subset \bigcap_{p<\infty} L_p \subset L_2 \subset L_1.$$

Quantitatively (for $p \geqslant 2$), one often uses the schematic chain

$$\|X\|_{L_1} \leqslant \|X\|_{L_2} \leqslant \|X\|_{L_p} \;\lesssim\; \|X\|_{\psi_1} \;\lesssim\; \|X\|_{\psi_2} \;\lesssim\; \|X\|_{L_\infty},$$

where the hidden factor in one step typically grows like $O(p)$.

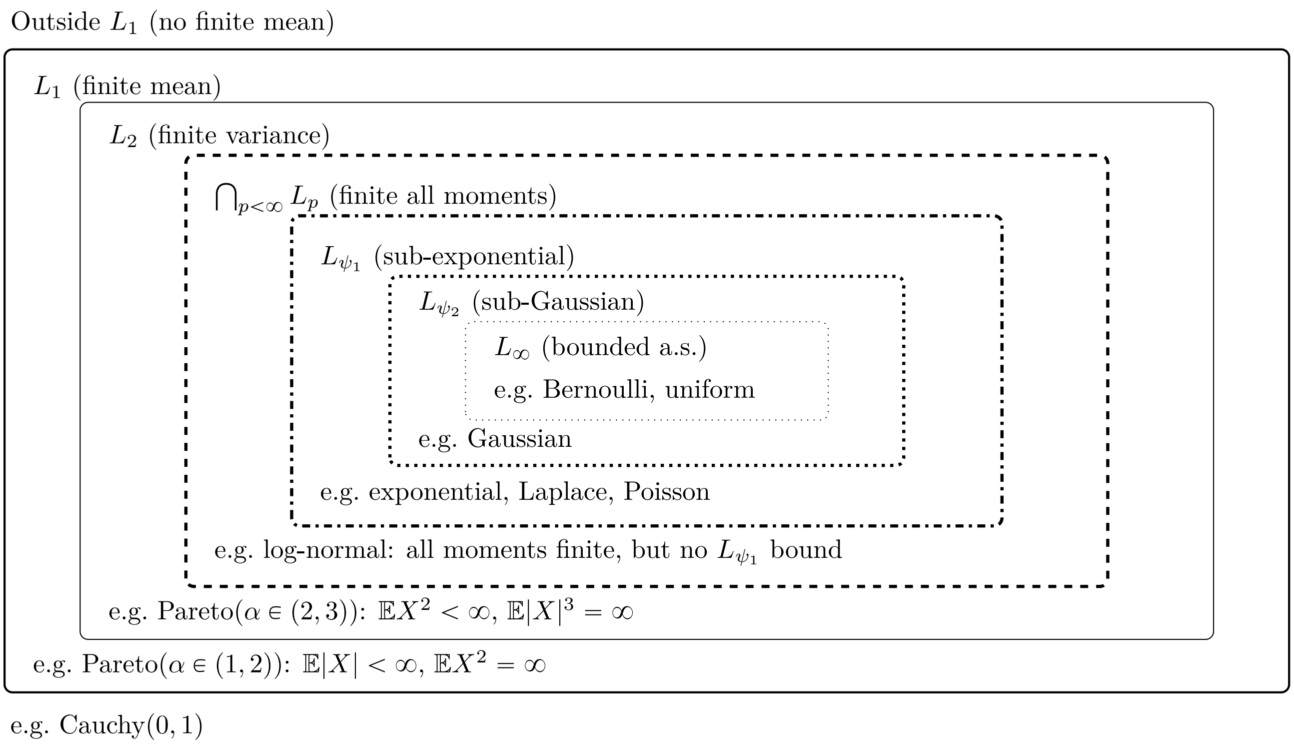

Figure 2: A useful hierarchy of integrability/tail classes with canonical examples. Inclusions (stronger tails $\Rightarrow$ smaller class): $L_\infty \subset L_{\psi_2} \subset L_{\psi_1} \subset \bigcap_{p<\infty} L_p \subset L_2 \subset L_1$. A Cauchy random variable lies outside L_1 (it is finite a.s. but has $\mathbb{E}|X| = \infty$).

Up to absolute constants, one has the moment growth heuristics

$$\|X\|_{L_p} \lesssim \|X\|_{\psi_2}\sqrt{p} \quad \text{(sub-Gaussian)}, \qquad \|X\|_{L_p} \lesssim \|X\|_{\psi_1} p \quad \text{(sub-exponential)}.$$

Thus ψ_2 controls all moments with $\sqrt{p}$ growth, while ψ_1 controls all moments with linear-in-p growth.

A hierarchy of tail/moment assumptions. In decreasing order of strength, one often encounters:

1. surely bounded: $|X| \leqslant M$ surely;
2. almost surely bounded: $|X| \leqslant M$ a.s.;
3. sub-Gaussian tail: $\mathbb{P}(|X| \geqslant t) \leqslant Ce^{-ct^2}$;
4. sub-exponential tail (more generally): $\mathbb{P}(|X| \geqslant t) \leqslant Ce^{-ct^a}$ for some $a > 0$;
5. finite k-th moment: $\mathbb{E}|X|^k < \infty$ for some $k \geqslant 1$;
6. integrable: $\mathbb{E}|X| < \infty$;
7. finite a.s.: $|X| < \infty$ a.s.

For example:

- If X is bounded (e.g. Bernoulli, or uniform on a bounded set), then $X \in L_\infty$, hence subgaussian and subexponential.
- Gaussians are subgaussian.
- Exponential, Poisson, and geometric random variables are canonical subexponential (but not subgaussian) examples.
- Cauchy is a canonical heavy-tailed example: it is almost surely finite, but does not have finite first moment.

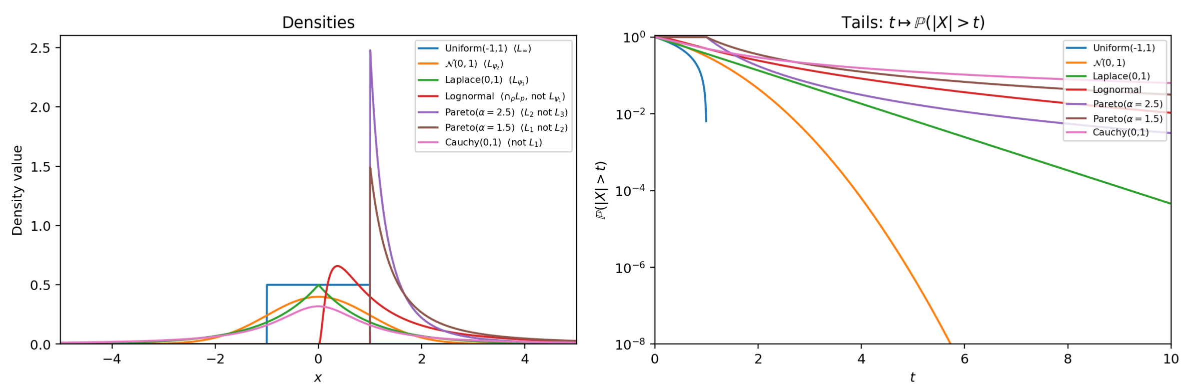

Figure 3: Representative distributions across the integrability/tail hierarchy. Left: densities. Right: tail probabilities $t \mapsto \mathbb{P}(|X| > t)$ (with a log scale on the vertical axis). Examples shown: Uniform$(-1, 1)$ (bounded, hence L_∞), Gaussian $\mathcal{N}(0, 1)$ (sub-Gaussian, L_{ψ_2}), Laplace$(0, 1)$ (sub-exponential, L_{ψ_1} but not L_{ψ_2}), Lognormal $\exp(\mathcal{N}(0, 1))$ (all moments finite but not L_{ψ_1}), Pareto$(\alpha = 2.5)$ (in L_2 but not L_3), Pareto$(\alpha = 1.5)$ (in L_1 but not L_2), and Cauchy$(0, 1)$ (not in L_1). For nonnegative distributions (lognormal/Pareto), $\mathbb{P}(|X| > t) = \mathbb{P}(X > t)$.

2 Tails, L_p norms, and basic concentration tools

A recurring theme is that moments (L_p norms) and tails ($\mathbb{P}(|X| > t)$) control each other.

Tails ↔ moments. For a nonnegative random variable X,

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}\{X > t\}\, dt.$$

This "integrated tail" identity will be a core identity for us. If you can bound tails, you can integrate to get expectations More generally, for $p > 0$ one has (under finiteness)

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

Markov and Chebyshev. Markov's inequality bounds tails using only an expectation: for $X \geqslant 0$ and $t > 0$,

$$\mathbb{P}\{X \geqslant t\} \leqslant \frac{\mathbb{E}X}{t}.$$

Chebyshev's inequality applies Markov to $(X - \mathbb{E}X)^2$: for any $t > 0$,

$$\mathbb{P}\big\{|X - \mathbb{E}X| \geqslant t\big\} \leqslant \frac{\mathrm{Var}(X)}{t^2}.$$

In high dimensions, Chebyshev is often too crude, but it is still a useful baseline and is frequently used inside more sophisticated arguments.

MGF trick. Applying Markov to $e^{\lambda X}$ yields, for any $\lambda > 0$,

$$\mathbb{P}\{X \geqslant t\} = \mathbb{P}\{e^{\lambda X} \geqslant e^{\lambda t}\} \leqslant e^{-\lambda t}\mathbb{E}e^{\lambda X}.$$

Optimizing over λ is the starting point for Chernoff/Hoeffding/Bernstein-type concentration bounds.

Paley–Zygmund. If $X \geqslant 0$ has finite second moment, then for any $\lambda \in (0, 1)$,

$$\mathbb{P}\big(X \geqslant \lambda \mathbb{E}X\big) \geqslant (1 - \lambda)^2 \frac{(\mathbb{E}X)^2}{\mathbb{E}X^2}.$$

Union bound. For any events $E_1, \ldots, E_m$,

$$\mathbb{P}\Big(\bigcup_{i=1}^m E_i\Big) \leqslant \sum_{i=1}^m \mathbb{P}(E_i).$$

This is one of the main ways we convert pointwise/high-probability statements into uniform ones. Typical example: Let $Z_1, \ldots, Z_p$ be random variables. Then, we have $\mathbb{P}\{\max_j Z_j \geqslant t\} \leqslant \sum_{j=1}^p \mathbb{P}\{Z_j \geqslant t\}$.

3 Random vectors and covariance matrices

A random vector $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ has mean $\mathbb{E}X = (\mathbb{E}X_1, \ldots, \mathbb{E}X_d) \in \mathbb{R}^d$ defined coordinatewise. There are two closely related notions of "variance" in multiple dimensions:

- The scalar variance $\mathbb{E}\|X - \mathbb{E}X\|_2^2$, which measures average squared Euclidean deviation.

- The covariance matrix

$$\operatorname{Cov}(X) := \mathbb{E}\big[(X - \mathbb{E}X)(X - \mathbb{E}X)^\top\big] = \mathbb{E}[XX^\top] - \mathbb{E}X(\mathbb{E}X)^\top \in \mathbb{R}^{d\times d},$$

  whose (i, j) entry is $\operatorname{Cov}(X_i, X_j)$.

By construction, $\operatorname{Cov}(X)$ is symmetric positive semidefinite, and it satisfies $v^\top \operatorname{Cov}(X)v = \operatorname{Var}(v^\top X) \geqslant 0$ for any direction $v \in \mathbb{R}^d$.
Also, note that these two notions are connected via:

$$\operatorname{tr}(\operatorname{Cov}(X)) = \sum_{j=1}^{d} \operatorname{Var}(X_j) = \mathbb{E}\|X - \mathbb{E}X\|_2^2.$$

If $A \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m$, then

$$\operatorname{Cov}(AX + b) = A \operatorname{Cov}(X)A^\top.$$

4 Conditioning

Conditional probability is $\mathbb{P}(E \mid F) = \mathbb{P}(E \cap F)/\mathbb{P}(F)$ (when $\mathbb{P}(F) > 0$). Conditional expectation $\mathbb{E}[X \mid Y]$ is a random variable (a function of Y).
The most used identity is the law of total expectation:

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

Applied to an indicator 1_E, it gives the law of total probability in a convenient form:

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

To compute or bound $\mathbb{E}X$ or $\mathbb{P}(E)$, it often helps to *first* analyze the quantity with some part of the randomness held fixed (conditioning), and *then* average over what remains.

5 Stochastic convergences

5.1 Basic modes of convergences

Let X_n and X be random variables.

- Almost sure (a.s.) convergence. We write $X_n \to X$ almost surely if

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

  This is the strongest of the common modes below: it says the sample paths converge except on a null event.

- Convergence in probability. We write $X_n \to X$ in probability if for every $\varepsilon > 0$,

$$\mathbb{P}\{|X_n - X| > \varepsilon\} \to 0.$$

  This is the natural notion for consistency of estimators.

- L_p convergence. For $p \geqslant 1$, we write $X_n \to X$ in L_p if

$$\|X_n - X\|_{L_p} = \left(\mathbb{E}|X_n - X|^p\right)^{1/p} \to 0.$$

  By Markov's inequality, L_p convergence implies convergence in probability.

- Convergence in distribution. We write $X_n \Rightarrow X$ if the CDFs converge at continuity points of F_X:

$$F_{X_n}(t) \to F_X(t) \quad \text{for all continuity points } t.$$

  This is the weakest of the standard modes, but it is the one used in the CLT.

Some implications:

$$X_n \to X \text{ a.s.} \implies X_n \to X \text{ in probability} \implies X_n \Rightarrow X.$$

In general, none of these implications can be reversed without extra assumptions.

5.2 Limit theorems

Limit theorems formalize what happens when we average independent samples. Let $X_1, X_2, \dots$ be i.i.d. with mean μ and variance $\sigma^2 < \infty$, and set $S_N = \sum_{i=1}^N X_i$.

- Laws of large numbers (LLN):
  - Weak law (WLLN): $\frac{S_N}{N} \to \mu$ in probability. A one-line intuition comes from Chebyshev: $\mathrm{Var}(S_N/N) = \sigma^2/N$, so $\mathbb{P}\{|\frac{S_N}{N} - \mu| \geqslant \varepsilon\} \leqslant \sigma^2/(N\varepsilon^2) \to 0$.
  - Strong law (SLLN): $\frac{S_N}{N} \to \mu$ almost surely. This is strictly stronger than the WLLN.

- Central limit theorem (CLT): After centering and scaling, sums become approximately Gaussian:

$$\frac{S_N - \mathbb{E}S_N}{\sqrt{\mathrm{Var}(S_N)}} = \frac{1}{\sigma\sqrt{N}} \sum_{i=1}^N (X_i - \mu) \Rightarrow \mathcal{N}(0, 1).$$

- Poisson limit theorem (PLT): If $X_{N,i} \sim \mathrm{Ber}(p_{N,i})$ are independent, $\max_i p_{N,i} \to 0$, and $\sum_{i=1}^N p_{N,i} \to \lambda$, then $S_N = \sum_{i=1}^N X_{N,i} \Rightarrow \mathrm{Pois}(\lambda)$. This is the correct limit when we sum many *rare* events and the expected total count stays $O(1)$.

Source material

Parts of this lecture are based on references: Vershynin (2018), in addition to the author's accumulated experience working on related topics.

References

Vershynin, R. (2018). *High-dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

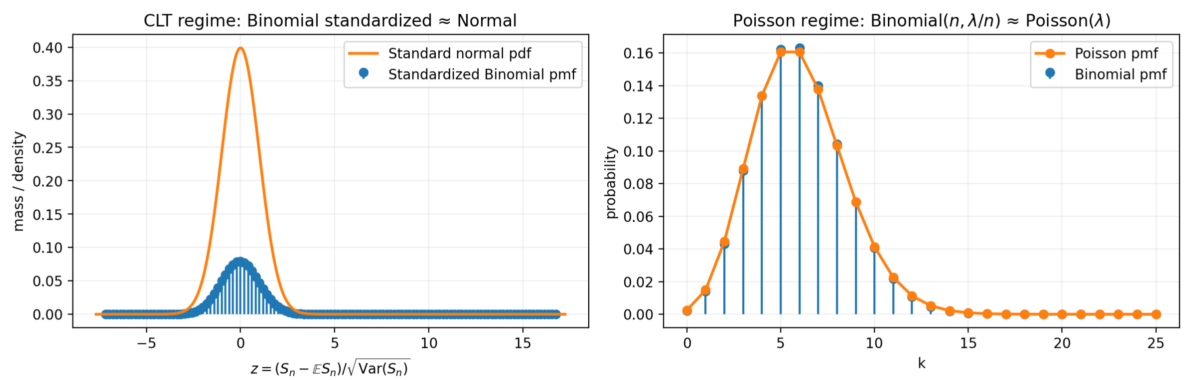

Figure 4: CLT vs Poisson limit. Left: in the classical regime (e.g. Binomial(n, p) with fixed p), the standardized sum approaches $N(0, 1)$. Right: in the sparse regime (e.g. Binomial$(n, \lambda/n)$), the sum approaches Poisson(λ).