LIMIT Category A and B Talk

Aritrabha Majumdar

January 2026

We have a couple of topics to select from, feel free to discuss among yourselves and let me know what you prefer.

1 Topics

1.1 Central Limit Theorem

A random variable is a rule that assigns a numerical value to the outcome of a random experiment. For example, it could count the number of heads in several coin tosses, record the waiting time for a bus, or measure the score on a quiz. The central limit theorem explains why sums of many small random effects often look approximately Gaussian, or "bell-shaped." In informal terms, if we add up lots of small random contributions and scale them correctly, the result starts behaving like one special random variable: the normal distribution. This theorem helps explain why bell curves appear so often in science, polling, measurement, and data analysis.

1.2 Erdős–Kac Theorem

The Erdős–Kac theorem is a striking probabilistic statement in number theory. For a positive integer n, let $\omega(n)$ denote the number of distinct prime factors of n. The theorem says that for a "typical" large integer, the quantity $\omega(n)$ is usually close to $\log\log n$, and after centering and scaling it behaves like a normal random variable. What makes this exciting is that prime numbers seem very arithmetic and exact, yet the theorem says that if we look at many integers at once, the number of prime factors behaves in a surprisingly random, predictable way.

1.3 Self-Avoiding Walks

A self-avoiding walk is a path on a grid or lattice that never visits the same point twice. You can think of it as trying to walk through a city laid out like graph paper while refusing to step on any intersection you have already used. These walks arise in combinatorics and in models of long polymer chains, where self-intersection is physically disallowed. Basic questions include how many self-avoiding walks of length n exist, how far they typically travel from the origin, and what their large-scale shape looks like. Even though the rules are simple, the mathematics becomes surprisingly deep very quickly.

1.4 Percolation

Percolation studies the emergence of large connected clusters in random media. A good picture to keep in mind is water trying to move through a porous material, or a rumor trying to spread through a social network. In the standard Bernoulli bond or site percolation model, each edge or vertex of a graph is declared open with probability p and closed otherwise, independently of the others. One then asks whether a giant connected cluster appears, how cluster sizes behave near the critical threshold p_c, and how the system suddenly changes when p passes that threshold. Percolation is one of the simplest mathematical models that shows a phase transition.

1.5 Information Theory

Information theory gives a quantitative language for uncertainty, communication, and compression. One of its big questions is: how many yes-or-no questions do we need, on average, to determine the value of a random variable completely? Its basic quantity is entropy, which measures the amount of information carried by a random variable. For a discrete random variable X with distribution (p_x), the entropy is

$$H(X) = -\sum_x p_x \log p_x.$$

Roughly speaking, larger entropy means more uncertainty, so more information is needed to figure out the outcome. Other key ideas include mutual information, which measures how much knowing one variable tells us about another, and channel capacity, which describes the maximal reliable communication rate through noisy systems. The subject connects probability with coding, statistics, machine learning, and theoretical computer science.

2 CLT and Erdős–Kac Theorem

This section develops the normal distribution from a natural model, then uses characteristic functions to prove the central limit theorem and explain the Erdős–Kac theorem.

2.1 A dartboard model and the Gaussian shape

Imagine throwing darts at a target. Let X be the horizontal error and Y the vertical error. A natural symmetry assumption is that the distribution of the total error only depends on the distance from the center, not on the direction. If the horizontal and vertical errors are independent and follow the same law, then there should be a function f such that the joint density has the form

$$p(x, y) = f\big(\sqrt{x^2 + y^2}\big) = f(x)f(y)$$

for all $x, y \in \mathbb{R}$. The first expression says "only the distance matters," while the second says "horizontal and vertical errors are independent."

Let $g(t) = f(\sqrt{t})$ for $t \geq 0$. Then the functional equation becomes

$$g(u + v) = g(u)g(v), \qquad u, v \geq 0.$$

Assume f is bounded, smooth, and positive near 0. Since $g(0) = f(0)$ and $g(0) = g(0)^2$, we get $g(0) = 1$. Taking logarithms, $h(t) = \log g(t)$ satisfies

$$h(u + v) = h(u) + h(v).$$

Because h is smooth, the only solutions are linear: $h(t) = -ct$ for some constant c. Therefore

$$g(t) = e^{-ct} \qquad \text{and hence} \qquad f(r) = e^{-cr^2}.$$

So the only smooth bounded radial law compatible with independence has Gaussian shape.

2.2 Probability density functions

A probability density function (pdf) on $\mathbb{R}$ is a nonnegative function p such that

$$\int_{-\infty}^{\infty} p(x)\,dx = 1.$$

If X has pdf p, then probabilities are found by integrating:

$$\mathbb{P}(a \le X \le b) = \int_a^b p(x)\,dx.$$

An important example from physics comes from quantum mechanics. If a particle has wave function $\psi(x)$, then the quantity

$$|\psi(x)|^2$$

describes the probability density for the particle's position, provided

$$\int_{-\infty}^{\infty} |\psi(x)|^2\,dx = 1.$$

So in that setting, the wave function itself is not the pdf, but its square magnitude is.

The standard normal density is

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

It is clearly nonnegative, so we only need to check that it integrates to 1. Let

$$I = \int_{-\infty}^{\infty} e^{-x^2/2}\,dx.$$

Then

$$I^2 = \int_{\mathbb{R}^2} e^{-(x^2+y^2)/2}\,dx\,dy.$$

Switch to polar coordinates:

$$I^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r\,dr\,d\theta = 2\pi \int_0^{\infty} e^{-r^2/2} r\,dr = 2\pi.$$

Thus $I = \sqrt{2\pi}$, and therefore

$$\int_{-\infty}^{\infty} \varphi(x)\,dx = 1.$$

So φ really is a pdf.

2.3 Moments, skewness, and kurtosis

If X is a random variable, its kth moment is usually $\mathbb{E}[X^k]$, provided the expectation exists. The first few moments are especially important:

- The first moment $\mathbb{E}[X]$ is the mean, or average value.
- The second moment $\mathbb{E}[X^2]$ measures size; from it we get the variance
$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$
- The third centered moment $\mathbb{E}[(X - \mathbb{E}[X])^3]$ measures asymmetry.
- The normalized third centered moment,
$$\frac{\mathbb{E}[(X - \mathbb{E}[X])^3]}{\mathrm{Var}(X)^{3/2}},$$
is called the skewness.
- The fourth centered moment $\mathbb{E}[(X - \mathbb{E}[X])^4]$ measures how heavy the tails are.
- The normalized fourth centered moment,
$$\frac{\mathbb{E}[(X - \mathbb{E}[X])^4]}{\mathrm{Var}(X)^2},$$
is called the kurtosis.

For the standard normal distribution, the mean is 0, the variance is 1, the skewness is 0, and the kurtosis is 3.

2.4 Characteristic functions

The characteristic function of a random variable X is

$$\phi_X(t) = \mathbb{E}[e^{itX}], \qquad t \in \mathbb{R}.$$

This always exists because $|e^{itX}| = 1$. Characteristic functions are useful because they turn sums into products: if X and Y are independent, then

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t).$$

Also, the derivatives of ϕ_X at 0 encode moments whenever the moments exist.

Theorem 2.1. *If two random variables have the same characteristic function, then they have the same distribution.*

Proof. If X has pdf p_X, then its characteristic function is the Fourier transform

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} p_X(x)\, dx.$$

Under mild regularity assumptions, we can recover p_X from ϕ_X by the inverse Fourier transform:

$$p_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_X(t)\, dt.$$

So if two random variables have the same characteristic function, then this inversion formula gives the same density, and hence the same distribution.

More generally, even when a density is not available, one can recover the distribution function from the characteristic function using the Lévy inversion formula:

$$\frac{F(b) + F(b^-)}{2} - \frac{F(a) + F(a^-)}{2} = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \phi_X(t)\, dt.$$

for continuity points $a < b$ of the distribution function F. This also shows that the characteristic function uniquely determines the distribution. □

2.5 Proof of the central limit theorem

We now prove the classical iid version of the central limit theorem.

Theorem 2.2. *Let $X_1, X_2, \dots$ be independent and identically distributed random variables with*

$$\mathbb{E}[X_1] = 0, \qquad \mathrm{Var}(X_1) = 1.$$

Then

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}}$$

converges in distribution to the standard normal random variable $\mathcal{N}(0,1)$.

Proof. Let $\phi(t) = \mathbb{E}[e^{itX_1}]$. Since $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[X_1^2] = 1$, the Taylor expansion near 0 gives

$$\phi(u) = 1 - \frac{u^2}{2} + o(u^2) \qquad \text{as } u \to 0.$$

Let

$$S_n = \frac{X_1 + \cdots + X_n}{\sqrt{n}}.$$

By independence,

$$\phi_{S_n}(t) = \Big(\phi\Big(\frac{t}{\sqrt{n}}\Big)\Big)^n.$$

Using the expansion with $u = t/\sqrt{n}$,

$$\phi\Big(\frac{t}{\sqrt{n}}\Big) = 1 - \frac{t^2}{2n} + o\Big(\frac{1}{n}\Big).$$

Therefore

$$\phi_{S_n}(t) = \left(1 - \frac{t^2}{2n} + o\Big(\frac{1}{n}\Big)\right)^n \longrightarrow e^{-t^2/2}.$$

Now let $Z \sim \mathcal{N}(0,1)$, with density

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Its characteristic function is

$$\phi_Z(t) = \mathbb{E}[e^{itZ}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2}\, dx.$$

Complete the square in the exponent:

$$itx - \frac{x^2}{2} = -\frac{1}{2}(x - it)^2 - \frac{t^2}{2}.$$

Therefore

$$\phi_Z(t) = e^{-t^2/2} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-it)^2/2}\, dx = e^{-t^2/2}.$$

So $e^{-t^2/2}$ is the characteristic function of $\mathcal{N}(0,1)$. By uniqueness of characteristic functions, S_n converges in distribution to $\mathcal{N}(0,1)$. □

If X_1 has mean μ and variance σ^2, we apply the theorem to

$$Y_i = \frac{X_i - \mu}{\sigma}$$

and obtain the usual statement

$$\frac{(X_1 + \cdots + X_n) - n\mu}{\sigma\sqrt{n}} \Rightarrow \mathcal{N}(0,1).$$

2.6 The $\log\log x$ form of Mertens' theorem

The Erdős–Kac theorem relies on the asymptotic formula

$$\sum_{p \le x} \frac{1}{p} \sim \log\log x.$$

Here is a short proof using the prime number theorem.

Theorem 2.3. *As $x \to \infty$,*

$$\sum_{p \le x} \frac{1}{p} = \log\log x + O(1).$$

In particular,

$$\sum_{p \le x} \frac{1}{p} \sim \log\log x.$$

Proof. Let

$$A(x) = \sum_{p \le x} 1 = \pi(x).$$

By partial summation,

$$\sum_{p \le x} \frac{1}{p} = \frac{\pi(x)}{x} + \int_2^x \frac{\pi(t)}{t^2}\, dt.$$

Now the prime number theorem says that

$$\pi(t) \sim \frac{t}{\log t}.$$

Therefore

$$\frac{\pi(x)}{x} \sim \frac{1}{\log x} \to 0,$$

and

$$\int_2^x \frac{\pi(t)}{t^2}\,dt = \int_2^x \frac{1}{t\log t}\,dt + o\left(\int_2^x \frac{dt}{t\log t}\right).$$

But

$$\int_2^x \frac{dt}{t\log t} = \log\log x - \log\log 2.$$

So we obtain

$$\sum_{p\le x}\frac{1}{p} = \log\log x + O(1).$$

Since the right-hand side tends to infinity, dividing by $\log\log x$ shows that

$$\sum_{p\le x}\frac{1}{p} \sim \log\log x.$$

□

2.7 Proof idea for the Erdős–Kac theorem

The proof is due to *Turán*, and we have

$$\sum_{n\le x}\nu(n) = \sum_{n\le x}\sum_{p\mid n} 1 = \sum_{p\le x}\sum_{n\le x\, p\mid n} 1 = \sum_{p\le x}\lfloor\frac{x}{p}\rfloor = \sum_{p\le x}\left(\frac{x}{p} + O(1)\right)$$

Now

$$\sum_{p\le x}\frac{1}{p} = \log\log x + c + O\left(\frac{1}{\log n}\right)$$

So

$$\sum_{n\le x}\nu(n) = x\log\log x + O(x) \implies \frac{1}{x}\sum_{n\le x}\nu(n) = \log\log x + O(1)$$

We fix $M = n^{1/3}$. Now we define

$$Z_n = \sum_{p\le M}\mathbf{1}_{(p\mid X_n)}$$

Evidently,

$$\nu(n) - 3 \le Z_n \le \nu(n).$$

$$\mathbb{E}(Z_n) = \sum_{p\le M}\mathbb{P}(p \mid X_n) = \sum_{p\le M}\frac{1}{n}\left\lfloor\frac{n}{p}\right\rfloor = \sum_{p\le M}\left(\frac{1}{p} + O\left(\frac{1}{n}\right)\right) \sim \log\log x + O(1)$$

$$\mathrm{Var}(Z_n) = \sum_{p\le M}\mathrm{Var}\big(\mathbf{1}_{(p\mid X_n)}\big) + 2\sum_{p<q\le M}\mathrm{Cov}\big(\mathbf{1}_{(p\mid X_n)}, \mathbf{1}_{(q\mid X_n)}\big).$$

Now,

$$\mathrm{Var}\big(\mathbf{1}_{(p|X_n)}\big) = \mathbb{E}\big(\mathbf{1}_{(p|X_n)}\big) - \Big(\mathbb{E}\mathbf{1}_{(p|X_n)}\Big)^2$$

$$= \frac{1}{p}\left(1 - \frac{1}{p}\right) + O\left(\frac{1}{n}\right).$$

Hence,

$$\sum_{p\le M} \mathrm{Var}\big(\mathbf{1}_{(p|X_n)}\big) = \sum_{p\le M} \frac{1}{p}\left(1 - \frac{1}{p}\right) + O\left(\frac{1}{n}\right) = \log\log n + O(1).$$

Also,

$$\mathrm{Cov}\big(\mathbf{1}_{(p|X_n)}, \mathbf{1}_{(q|X_n)}\big) = \mathbb{E}\big(\mathbf{1}_{(pq|X_n)}\big) - \mathbb{E}\big(\mathbf{1}_{(p|X_n)}\big)\mathbb{E}\big(\mathbf{1}_{(q|X_n)}\big)$$

$$= \frac{1}{n}\left\lfloor \frac{n}{pq} \right\rfloor - \frac{1}{n^2}\left\lfloor \frac{n}{p} \right\rfloor \left\lfloor \frac{n}{q} \right\rfloor$$

$$\le \frac{1}{pq} - \left(\frac{1}{p} - \frac{1}{n}\right)\left(\frac{1}{q} - \frac{1}{n}\right) \le \frac{1}{n}\left(\frac{1}{p} + \frac{1}{q}\right).$$

Therefore,

$$\sum_{p<q\le M} \mathrm{Cov}\big(\mathbf{1}_{(p|X_n)}, \mathbf{1}_{(q|X_n)}\big) \le \frac{1}{n} \sum_{p<q\le M} \left(\frac{1}{p} + \frac{1}{q}\right)$$

$$= \frac{1}{n} M \log\log n + \frac{1}{n} M \log\log n + O\left(\frac{M}{n}\right).$$

So $Var(Z_n) = \log\log n + O(1)$, and by CLT

$$\frac{\nu(n) - \log\log n}{\sqrt{\log\log n}} \implies \mathcal{N}(0,1)$$

3 Percolation and Self-Avoiding Walks

This section gives a more mathematical description of bond percolation and self-avoiding walks, but keeps the main ideas as simple as possible.

3.1 The Bond Percolation Model

For $d \in \mathbb{N}$, we think of the lattice $\mathbb{Z}^d$ as sitting inside $\mathbb{R}^d$. Two vertices $x, y \in \mathbb{Z}^d$ are joined by an edge when their ℓ^1 distance is 1, that is,

$$\|x - y\|_1 = 1.$$

The resulting graph is the usual nearest-neighbour lattice, which we denote by $\mathbb{L}^d = (\mathbb{Z}^d, E^d)$.

Now fix a parameter $p \in [0,1]$. Each edge is declared *open* with probability p and *closed* with probability $1-p$, independently of all the other edges. A configuration is therefore a function

$$\omega : E^d \to \{0,1\},$$

where $\omega(e) = 1$ means that e is open and $\omega(e) = 0$ means that e is closed. The sample space is

$$\Omega = \prod_{e \in E^d} \{0,1\},$$

equipped with the product probability measure $\mathbb{P}_p$.

If two sets of vertices A and B can be connected by a path using only open edges, we write

$$A \leftrightarrow B.$$

One of the most important questions is whether there exists an infinite open cluster. This leads to the *critical probability*

$$p_c(d) = \inf\bigl\{p \in [0,1] : \mathbb{P}_p(0 \leftrightarrow \infty) > 0\bigr\}.$$

Roughly speaking, when $p < p_c(d)$, open clusters are typically small, while for $p > p_c(d)$ there is a positive chance that the origin lies in an infinite cluster.

3.2 Defining the Self-Avoiding Walk

A self-avoiding walk is a path on the lattice that never visits the same vertex twice. For each $n \in \mathbb{N}_0$, we define

$$\mathcal{W}_n = \Bigl\{w = (w_0, w_1, \ldots, w_n) \in (\mathbb{Z}^d)^{n+1} : w_0 = 0,\ \|w_{i+1} - w_i\|_1 = 1,\ \text{and } w_i \neq w_j \text{ for } i \neq j\Bigr\}.$$

So $\mathcal{W}_n$ is the set of all self-avoiding walks of length n starting at the origin.

More generally, one can put a Gibbs measure on $\mathcal{W}_n$. If $H_n(w)$ is an energy function, or Hamiltonian, and $\beta \in \mathbb{R}$ is the inverse temperature, then the Gibbs measure is defined by

$$P_n^\beta(w) = \frac{1}{Z_n(\beta)} e^{-\beta H_n(w)}, \qquad w \in \mathcal{W}_n,$$

where

$$Z_n(\beta) = \sum_{w \in \mathcal{W}_n} e^{-\beta H_n(w)}$$

is the partition function that normalizes the probabilities.

Let

$$c_n = |\mathcal{W}_n|$$

be the number of such walks. If we choose one uniformly from $\mathcal{W}_n$, then each walk has probability

$$P_n(w) = \frac{1}{c_n}.$$

This is the simplest model, and it corresponds to taking no extra energy term in the Gibbs weight, for example $H_n \equiv 0$. In that case, the partition function is just $Z_n = c_n$.

An important quantity is the *connective constant*

$$\mu(d) = \lim_{n \to \infty} c_n^{1/n},$$

which measures the exponential growth rate of the number of self-avoiding walks.

Theorem 3.1. *For every $d \geq 1$, the connective constant satisfies*

$$\mu(d) \leq 2d - 1.$$

Proof. The first step of a self-avoiding walk has at most $2d$ choices. After that, at each new step the walk cannot immediately return to the vertex it just came from, so there are at most $2d-1$ choices. Therefore

$$c_n \leq 2d(2d-1)^{n-1}.$$

Taking nth roots and letting $n \to \infty$ gives

$$\mu(d) = \lim_{n\to\infty} c_n^{1/n} \leq 2d - 1.$$

□

3.3 Free Energy and Fekete's Lemma

In physics and chemistry, *free energy* is the amount of energy available to do useful work. Two important examples are the Helmholtz free energy,

$$F = U - TS,$$

and the Gibbs free energy,

$$G = H - TS,$$

where U is internal energy, H is enthalpy, T is temperature, and S is entropy. In chemistry, the Gibbs free energy is especially important because it helps predict whether a reaction is thermodynamically favorable at constant pressure and temperature.

In lattice models and statistical mechanics, the free energy is the large-scale exponential growth rate of the partition function. For the walk model above, we define

$$f(\beta) = \lim_{n\to\infty} \frac{1}{n} \log Z_n(\beta),$$

provided the limit exists.

When $H_n \equiv 0$, we have $Z_n = c_n$, so the free energy becomes

$$f = \lim_{n\to\infty} \frac{1}{n} \log c_n = \log \mu(d).$$

To justify the existence of limits of this kind, one often uses Fekete's lemma.

Lemma 3.2 (Fekete's lemma)**.** *If $(a_n)_{n\geq 1}$ is a subadditive sequence, meaning that*

$$a_{m+n} \leq a_m + a_n \qquad \text{for all } m, n \geq 1,$$

then the limit

$$\lim_{n\to\infty} \frac{a_n}{n}$$

exists in $[-\infty, \infty)$ and equals

$$\inf_{n\geq 1} \frac{a_n}{n}.$$

Proof. Let

$$\alpha = \inf_{n\geq 1} \frac{a_n}{n}.$$

Then certainly

$$\liminf_{n\to\infty} \frac{a_n}{n} \geq \alpha.$$

So it remains to show that

$$\limsup_{n\to\infty} \frac{a_n}{n} \leq \alpha.$$

Fix $k \geq 1$. For any n, write

$$n = qk + r \qquad \text{with } 0 \leq r < k.$$

By subadditivity,

$$a_n = a_{qk+r} \leq q a_k + a_r.$$

Dividing by n gives

$$\frac{a_n}{n} \leq \frac{qk}{n} \cdot \frac{a_k}{k} + \frac{a_r}{n}.$$

As $n \to \infty$ with k fixed, we have $qk/n \to 1$ and $a_r/n \to 0$ because $r < k$ takes only finitely many values. Hence

$$\limsup_{n\to\infty} \frac{a_n}{n} \leq \frac{a_k}{k}.$$

Since this holds for every k, we obtain

$$\limsup_{n\to\infty} \frac{a_n}{n} \leq \inf_{k\geq 1} \frac{a_k}{k} = \alpha.$$

Combining this with the lower bound for the liminf shows that

$$\lim_{n\to\infty} \frac{a_n}{n} = \alpha.$$

□

For the partition function, subadditivity often appears in the form

$$\log Z_{m+n}(\beta) \leq \log Z_m(\beta) + \log Z_n(\beta).$$

When this holds, Fekete's lemma implies that

$$f(\beta) = \lim_{n\to\infty} \frac{1}{n} \log Z_n(\beta) = \inf_{n\geq 1} \frac{1}{n} \log Z_n(\beta).$$

There is also a useful convexity property. If the Hamiltonian depends linearly on the parameter β, then Hölder's inequality gives

$$Z_n\left(\frac{\beta_1 + \beta_2}{2}\right) \leq \sqrt{Z_n(\beta_1) Z_n(\beta_2)}.$$

Taking logarithms and dividing by n shows that

$$f_n\left(\frac{\beta_1 + \beta_2}{2}\right) \leq \frac{f_n(\beta_1) + f_n(\beta_2)}{2}, \qquad f_n(\beta) = \frac{1}{n} \log Z_n(\beta).$$

So each f_n is convex, and the limiting free energy is convex whenever the limit exists. Convexity is important because non-differentiability of free energy is often interpreted as a sign of a phase transition.

As an additional piece of information, there are strong upper bounds on the number of self-avoiding walks. A classical Hammersley–Welsh bound says that

$$c_n \leq \mu(d)^n \exp\bigl(O(\sqrt{n})\bigr).$$

More recently, Hugo Hutchcroft gave a simpler proof of a bound of this form and refined it further. The main message is that c_n grows essentially like $\mu(d)^n$, up to a much smaller stretched-exponential correction.

3.4 Mean Square Displacement

The mean square displacement measures how far the walk typically gets from the origin after n steps. If the endpoint of a walk is denoted by S_n, then we define

$$\mathbb{E}\bigl(\|S_n\|^2\bigr) = \frac{1}{c_n} \sum_{w \in \mathcal{W}_n} \|S_n(w)\|^2.$$

In words, we average the squared distance from the origin over all self-avoiding walks of length n.

Physicists predicted that

$$\mathbb{E}\bigl(\|S_n\|^2\bigr) = \begin{cases} Dn^{2\nu}, & d \neq 4, \\ Dn(\log n)^{1/4}, & d = 4, \end{cases}$$

for some constant $D > 0$. In two dimensions, the predicted value is

$$\nu = \frac{3}{4}.$$

There is a simple heuristic for where this exponent comes from. Suppose a self-avoiding walk of length n typically fills a region of radius L. Then its density is roughly

$$\frac{n}{L^d}$$

monomers per site. A rough repulsion argument suggests that the total self-avoidance cost is about

$$\frac{n^2}{L^d}.$$

On the other hand, for an ordinary simple random walk, ending up at radius L costs roughly

$$\exp\left(-\frac{L^2}{n}\right).$$

Putting these two effects together suggests that a self-avoiding walk of radius L should have weight roughly

$$\exp\left(-\left\{\frac{n^2}{L^d} + \frac{L^2}{n}\right\}\right).$$

Now set $L = n^\nu$. Then the exponent becomes

$$n^{2-d\nu} + n^{2\nu-1}.$$

To balance the two competing terms, we set the exponents equal:

$$2 - d\nu = 2\nu - 1.$$

Solving gives

$$\nu = \frac{3}{d+2}.$$

In particular, when $d = 2$, this predicts $\nu = 3/4$.

3.5 A Basic Result About Critical Probability

The function

$$\theta(p) = \mathbb{P}_p(0 \leftrightarrow \infty)$$

is nondecreasing in p: if we make edges more likely to be open, then infinite open clusters become more likely.

Theorem 3.3. *For every $d \geq 2$,*

$$0 < p_c(d) < 1.$$

Idea of proof. To show $p_c(d) > 0$, consider the number $N(n)$ of open self-avoiding paths of length n starting at the origin. Each particular path of length n is open with probability p^n, so

$$\mathbb{E}_p[N(n)] = p^n c_n.$$

Using the bound $c_n \leq (\mu(d) + o(1))^n$, we get

$$\theta(p) \leq \mathbb{P}_p(N(n) \geq 1) \leq \mathbb{E}_p[N(n)] \leq (p\mu(d) + o(1))^n.$$

If $p < 1/\mu(d)$, the right-hand side tends to 0, so $\theta(p) = 0$. Hence $p_c(d) \geq 1/\mu(d) > 0$.

To show $p_c(d) < 1$, it is enough to work in two dimensions and use a Peierls argument. When p is close to 1, closed dual circuits become very unlikely. If there is no closed dual circuit surrounding the origin, then the origin has a good chance to connect to infinity through open edges. This shows that for p sufficiently close to 1, we have $\theta(p) > 0$, so $p_c(2) < 1$. Since $p_c(d+1) \leq p_c(d)$, it follows that $p_c(d) < 1$ for every $d \geq 2$. □

In dimension 2, a famous theorem of Kesten shows that

$$p_c(2) = \frac{1}{2}.$$

4 Information Theory

Information theory asks how much uncertainty is present in a random object, and how much information is gained when we observe it. It gives a mathematical language for data compression, communication through noisy channels, counting arguments, and several geometric inequalities.

4.1 Entropy, Conditional Entropy, and Mutual Information

Let X be a discrete random variable taking values in a finite or countable set with probabilities $p(x) = \mathbb{P}(X = x)$. Its *entropy* is

$$H(X) = -\sum_x p(x) \log p(x).$$

This measures the average uncertainty in X. If all outcomes are equally likely, then the entropy is large; if one outcome is overwhelmingly likely, then the entropy is small.

If X and Y are discrete random variables, the *conditional entropy* of X given Y is

$$H(X \mid Y) = \sum_y \mathbb{P}(Y = y) H(X \mid Y = y) = -\sum_{x,y} p(x,y) \log p(x \mid y).$$

This measures the amount of uncertainty left in X after we are told the value of Y.

The *mutual information* between X and Y is defined by

$$I(X;Y) = H(X) - H(X \mid Y).$$

It measures how much knowing Y tells us about X. By symmetry one also has

$$I(X;Y) = H(Y) - H(Y \mid X).$$

Lemma 4.1 (Chain rule)**.** *For discrete random variables X and Y,*

$$H(X,Y) = H(Y) + H(X \mid Y) = H(X) + H(Y \mid X).$$

Proof. By definition,

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y).$$

Since $p(x,y) = p(y)p(x \mid y)$, we get

$$\log p(x,y) = \log p(y) + \log p(x \mid y).$$

Substituting this into the formula for $H(X,Y)$ gives

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(y) - \sum_{x,y} p(x,y) \log p(x \mid y).$$

The first term is $H(Y)$ and the second is $H(X \mid Y)$. □

Corollary 4.1.1. *For discrete random variables X and Y,*

$$I(X;Y) = H(X) + H(Y) - H(X,Y).$$

Proof. Apply the chain rule:

$$I(X;Y) = H(X) - H(X \mid Y) = H(X) - \big(H(X,Y) - H(Y)\big).$$

□

4.2 Kullback–Leibler Divergence and Basic Inequalities

If P and Q are probability distributions on the same finite or countable set, with masses $p(x)$ and $q(x)$, then the *Kullback–Leibler divergence* is

$$D(P\|Q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}.$$

It is not a true metric, but it is one of the most important ways to measure how different two distributions are.

Theorem 4.2 (Gibbs inequality)**.** *For probability distributions P and Q,*

$$D(P\|Q) \geq 0,$$

with equality if and only if $P = Q$.

Proof. Use the elementary inequality

$$\log u \le u - 1 \qquad \text{for } u > 0,$$

with equality if and only if $u = 1$. Setting $u = q(x)/p(x)$ gives

$$\log \frac{q(x)}{p(x)} \le \frac{q(x)}{p(x)} - 1.$$

Multiply by $p(x)$ and sum over x:

$$\sum_x p(x) \log \frac{q(x)}{p(x)} \le \sum_x q(x) - \sum_x p(x) = 0.$$

Rearranging gives $D(P\|Q) \ge 0$. □

Corollary 4.2.1. *For discrete random variables X and Y,*

$$I(X;Y) = D(P_{X,Y}\|P_X P_Y) \ge 0.$$

In particular,

$$H(X \mid Y) \le H(X).$$

Proof. By direct computation,

$$D(P_{X,Y}\|P_X P_Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = H(X) + H(Y) - H(X,Y) = I(X;Y).$$

Since mutual information is nonnegative, we get

$$H(X \mid Y) = H(X) - I(X;Y) \le H(X).$$

□

Corollary 4.2.2 (Subadditivity)**.** *For discrete random variables X and Y,*

$$H(X,Y) \le H(X) + H(Y),$$

with equality if and only if X and Y are independent.

Proof. From the previous corollary,

$$H(X,Y) = H(X) + H(Y) - I(X;Y) \le H(X) + H(Y).$$

Equality holds exactly when $I(X;Y) = 0$, which is equivalent to independence. □

Lemma 4.3 (Maximum entropy of the uniform law)**.** *If X takes values in a finite set S, then*

$$H(X) \le \log |S|,$$

with equality if and only if X is uniform on S.

Proof. Let U be the uniform distribution on S. Then

$$D(P_X\|U) = \sum_{x \in S} p(x) \log \frac{p(x)}{1/|S|} = \sum_{x \in S} p(x) \log p(x) + \log |S|.$$

By Gibbs inequality, $D(P_X\|U) \ge 0$, so

$$-H(X) + \log |S| \ge 0.$$

Thus $H(X) \le \log |S|$. □

4.3 Differential Entropy and the Normal Distribution

If X is a continuous random variable with density f, its *differential entropy* is

$$h(X) = -\int_{\mathbb{R}^n} f(x)\log f(x)\,dx,$$

provided the integral exists. Differential entropy behaves somewhat differently from discrete entropy, but Gaussian distributions remain central.

Theorem 4.4. *If $X \sim \mathcal{N}(0,\sigma^2)$ on $\mathbb{R}$, then*

$$h(X) = \frac{1}{2}\log(2\pi e\sigma^2).$$

More generally, if $X \sim \mathcal{N}(0,\Sigma)$ in $\mathbb{R}^n$ with positive definite covariance matrix Σ, then

$$h(X) = \frac{1}{2}\log\big((2\pi e)^n \det\Sigma\big).$$

Proof. In one dimension,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/(2\sigma^2)}.$$

Hence

$$-\log f(x) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{x^2}{2\sigma^2}.$$

Taking expectation gives

$$h(X) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathbb{E}[X^2] = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}.$$

Since $\frac{1}{2} = \frac{1}{2}\log e$, this becomes

$$h(X) = \frac{1}{2}\log(2\pi e\sigma^2).$$

The multivariate formula is proved in exactly the same way using the density of $\mathcal{N}(0,\Sigma)$. □

4.4 A Determinant Inequality for Convex Combinations

The function $\log\det$ plays a basic role in information theory because Gaussian entropy depends on $\det\Sigma$.

Theorem 4.5 (Concavity of $\log\det$)**.** *If A and B are positive definite matrices and $0 \le t \le 1$, then*

$$\det\big((1-\lambda)A + \lambda B\big) \ge \det(A)^{1-\lambda}\det(B)^{\lambda}.$$

Equivalently,

$$\log\det\big((1-t)A + tB\big) \ge (1-t)\log\det A + t\log\det B.$$

Proof. Let $X_1 \sim \mathcal{N}(0,A)$ and $X_2 \sim \mathcal{N}(0,B)$ be independent random vectors in $\mathbb{R}^n$. Let X_z be amother random variable where Z is 1 with probability $1-\lambda$ and 2 with probability λ. Now

$$\frac{1}{2}\log(2\pi e)^n\det\big((1-\lambda)A+\lambda B\big) = h(X_Z) \ge h(X_Z|Z) = \frac{1-\lambda}{2}\log(2\pi e)^n\det(A) + \frac{\lambda}{2}\log(2\pi e)^n\det(B)$$

and thus the result follows.

□

4.5 Shearer's Lemma

Let $X = (X_1, \dots, X_n)$ be a discrete random vector. If $A \subseteq \{1, \dots, n\}$, write $X_A = (X_i)_{i \in A}$.

Theorem 4.6 (Shearer's lemma)**.** *Let $\mathcal{A}$ be a family of subsets of $\{1, \dots, n\}$ such that every index i belongs to at least k members of $\mathcal{A}$. Then*

$$H(X_1, \dots, X_n) \le \frac{1}{k} \sum_{A \in \mathcal{A}} H(X_A).$$

Radhakrishnan's proof. Choose a permutation π of $\{1, \dots, n\}$ uniformly at random. By the chain rule,

$$H(X) = \sum_{i=1}^{n} H\big(X_{\pi(i)} \mid X_{\pi(1)}, \dots, X_{\pi(i-1)}\big).$$

Now fix a set $A \in \mathcal{A}$. Applying the chain rule only inside the coordinates of A gives

$$H(X_A) = \sum_{i \in A} H\big(X_i \mid X_j : j \in A,\ \pi(j) < \pi(i)\big).$$

The conditioning here is on fewer variables than all earlier variables in the permutation, so conditioning reduces entropy and gives

$$H\big(X_i \mid X_j : j \in A,\ \pi(j) < \pi(i)\big) \ge H\big(X_i \mid X_j : \pi(j) < \pi(i)\big).$$

Therefore,

$$H(X_A) \ge \sum_{i \in A} H\big(X_i \mid X_j : \pi(j) < \pi(i)\big).$$

Summing over all $A \in \mathcal{A}$, every coordinate i appears at least k times, so

$$\sum_{A \in \mathcal{A}} H(X_A) \ge k \sum_{i=1}^{n} H\big(X_i \mid X_j : \pi(j) < \pi(i)\big) = kH(X).$$

This is exactly the desired inequality. □

4.6 Loomis–Whitney Inequality

The Loomis–Whitney inequality is a beautiful geometric consequence of Shearer's lemma.

Theorem 4.7 (Discrete Loomis–Whitney)**.** *Let $S \subseteq X_1 \times \cdots \times X_d$ be finite, and let $\pi_i(S)$ denote the projection of S onto all coordinates except the ith. Then*

$$|S|^{d-1} \le \prod_{i=1}^{d} |\pi_i(S)|.$$

Proof. Let $X = (X_1, \dots, X_d)$ be uniformly distributed on S. Then

$$H(X) = \log |S|.$$

For each i, let $A_i = \{1, \dots, d\} \setminus \{i\}$. Each coordinate belongs to exactly $d - 1$ of the sets A_i. Applying Shearer's lemma gives

$$H(X) \le \frac{1}{d-1} \sum_{i=1}^{d} H(X_{A_i}).$$

But X_{A_i} takes values in $\pi_i(S)$, so

$$H(X_{A_i}) \le \log |\pi_i(S)|.$$

Hence

$$\log |S| \le \frac{1}{d-1} \sum_{i=1}^{d} \log |\pi_i(S)|.$$

Exponentiating gives

$$|S|^{d-1} \le \prod_{i=1}^{d} |\pi_i(S)|.$$

□

4.7 Shuffling Increases Entropy

Suppose $X = (X_1, \dots, X_n)$ is a random vector and Π is a random permutation of $\{1, \dots, n\}$, chosen independently of X. Define the shuffled vector

$$Y = (X_{\Pi(1)}, \dots, X_{\Pi(n)}).$$

Theorem 4.8. *Shuffling does not decrease entropy:*

$$H(Y) \ge H(X).$$

The same statement holds for differential entropy whenever the relevant densities exist.

Proof. Condition on the value of Π. For each fixed permutation π, the map

$$X \mapsto (X_{\pi(1)}, \dots, X_{\pi(n)})$$

is just a relabeling of coordinates, so it does not change entropy. Therefore

$$H(Y \mid \Pi) = H(X).$$

Now use the identity

$$H(Y) = H(Y \mid \Pi) + I(Y; \Pi).$$

Since mutual information is nonnegative, we get

$$H(Y) \ge H(Y \mid \Pi) = H(X).$$

□