MATH 235: Modern Markov Chains

Lecturer: Professor Persi Diaconis

Notes by: Andrew Lin

Spring 2026

1 March 30, 2026

The subject of Markov chains continues to explode enormously, but we still have to start from the basics (unless the whole class tells Professor Diaconis to speed up after a week or two). Office hours will be weekly in room 383D Wednesdays after lectures.

In this first lecture, we'll start with a general motivating overview. Let $\mathfrak{X}$ be a measurable space (but most things in the class will just be finite). Suppose $\pi(x)$ is a probability measure on $\mathfrak{X}$, and $f \in L^1(\pi)$ is an integrable function. The goal is to compute or approximate the expectation $\int f\,d\pi$. In many high-dimensional problems, we can't do this exactly (we can't even write down π or we don't know the normalizing measure), but there are methods of still approximating it by using a Markov kernel $K(x, A)$ (meaning that $x \mapsto K(x, A)$ is measurable and $K(x, \cdot)$ is a probability measure).

The idea is that we can construct a Markov chain K with stationary measure π, meaning that for any A,

$$\int_{\mathfrak{X}} K(x, A)\pi(dx) = \pi(A).$$

Then iterating K, meaning that

$$K^2(x, A) = \int K(y, A)K(x, dy), \quad K^\ell(x, A) = \int K^{\ell-1}(y, A)K(x, dy),$$

yields (under mild conditions) that $K^\ell(x, A) \to \pi(A)$ for any starting state x and any set A. Thus to compute the integral $\int f\,d\pi$ of interest, we generate $X_0 = x, X_1, X_2, \cdots$ a run of the Markov chain. We can then estimate

$$I_\pi(f) = \int f\,d\pi \quad \approx \quad \hat{I}_n(f) = \frac{1}{n+1}\sum_{i=0}^{n} f(X_i).$$

One key question of the course is thus the following:

Problem 1

Given $\varepsilon, \varepsilon'$, how large do we need to make n so that

$$\mathbb{P}_x(|\hat{I}_n(f) - I_\pi(f)| > \varepsilon) < \varepsilon',$$

and does the answer depend on the starting state x?

At a more "down-to-earth level," we can also phrase the problem more directly (though the answer to this one can be different from the one before):

Problem 2

The **total variation distance** (between π and the distribution after ℓ steps when started at x) is defined by

$$||K_x^\ell - \pi|| = \max_A |K^\ell(x, A) - \pi(A)|.$$

How large does ℓ need to be so that $||K_x^\ell - \pi|| < \varepsilon$? How does this depend on x?

This latter formulation is often the way people think about questions like "how many shuffles does it take for a deck of cards to get random." For an example of what these results look like, here's a concrete description:

Example 3

Let our state space $\mathfrak{X} = \{0, 1, \cdots, n\} \times (0, 1)$ be a product of a discrete and a continuous part, and consider the **beta-binomial** stationary distribution

$$\pi(j, \theta) = \binom{n}{j} \theta^j (1-\theta)^{n-j} \quad \text{for } 0 \le j \le n,\ 0 < \theta < 1.$$

This is a joint probability distribution in j and θ, since summing over j yields 1 (binomial distribution) and then integrating that over θ again yields 1.

The story of this setting is the following: 10 years ago, the final project for this course was to read a paper called "Explaining the Gibbs Sampler" by Casella and George. The idea of the Gibbs sampler is that we have some probability measure $f(x_1, \cdots x_k)$ on a product space $\mathfrak{X} = X_1 \times \cdots \times X_k$, and we want to be able to sample from it even though (for example) we're only given f up to a normalizing constant. To get around this, we can start at some state $x^0 = (x_1^0, \cdots, x_k^0)$, **resample the first coordinate** to get $(x_1^1, x_2^0, \cdots, x_k^0)$, then **resample the second coordinate** to get $(x_1^1, x_2^1, x_3^0, \cdots, x_k^0)$, and so on until we've refreshed all k coordinates; the end result is our new state x^1. And the way we resample at each ith step is that we resample from the conditional distribution of f given all other coordinates except the ith one. (For example in statistical physics models, this can be quite easy to do because the conditional density often only depends on some local interactions – it reduces a higher-dimensional sampling problem to a low-dimensional one.)

Under mild conditions, such a Gibbs sampler chain will have f as its stationary distribution. And the paper by Casella and George was explaining this to the statistics community with various examples, and the project was to pick any example from the paper and prove anything about it. The first example in the paper was this beta-binomial distribution, and four students (independently) tried to prove something about it. Well, the Gibbs sampler for this can be described explicitly:

- Start at some state (j, θ).
- Now change the first coordinate to (j', θ), which means $j' \sim \text{Binomial}(n, \theta)$.
- Finally, change the second coordinate to (j', θ'), which means $\theta' \sim \text{Beta}(j+1, n-j+1)$.

A concrete question we could then ask is something like the following: if $n = 100$ and we start from $(j_0, \theta_0) = (100, \frac{1}{2})$, can we find some ℓ such that $||K_{100,1/2}^\ell - \pi|| < \frac{1}{100}$? And using the best technology at the time (Harris recurrence), the students in the course managed to prove $\ell \le 10^{33}$. (The community couldn't really do much better than that either.) But for some common sense, we can run some simulations and keep track of the position of the j-coordinate after various numbers of steps; we can see that the histogram is not very uniform after 50 steps but looks

pretty uniform after 200 steps. So Professor Diaconis had to come in and try to solve the problem "properly," and so here's an example of a theorem that we're aiming to get to:

Theorem 4

Let $\beta_1 = 1 - \frac{2}{n+2}$. Then for **any** starting state (j, θ),

$$\frac{1}{2}\beta_1^\ell \leq ||K_{j,\theta}^\ell - \pi|| \leq \frac{\beta_1^{\ell-1/2}}{1-\beta_1^{2\ell-2}}.$$

Notice that this result doesn't have any unspecified constants, and really what this is saying if we plug things in is that order n steps suffice for mixing of the chain: for $n = 100$ and $\ell = 200$ we have $||K_{100,1/2}^\ell - \pi|| \leq 0.0192$, while for $\ell = 50$ we have $||K_{100,1/2}^\ell - \pi|| \geq 0.1825$. For comparison, what Harris recurrence really gets us when applied to this problem is $||K_{j,1/2}^\ell - \pi|| \leq A\gamma^\ell$, without being told what A and γ are (and that's not really useful for anything other than knowing that the chain actually converges).

For more commentary on this problem, we can see Professor Diaconis' paper with Khare and Saloff-Coste "Gibbs Sampling, Exponential Families and Orthogonal Polynomials." The main thing that was useful was that we could actually diagonalize the chain and get an explicit description of all of the eigenvalues and eigenvectors.

With that example in mind, here's what our course outline will look like:

- The first three weeks will be the geometric theory of Markov chains (spectral theory with eigenvalues and eigenvectors, volume growth, comparison theory).
- The next two weeks will be about "how to construct your own Markov chain," using things like the Metropolis algorithm or Gibbs sampler, auxiliary variables, Swendsen-Weng, and so on.
- The two weeks after that will be certain common stochastic techniques (stationary times, coupling, path-coupling).
- And finally, the three weeks at the end will discuss the cutoff phenomenon, in particular the breakthrough work of Justin Salez. The idea is that many Markov chains have total variation going from almost 1 to almost 0 in a very short window, and there's some emerging theory involving a notion of (Ricci) curvature for Markov chains, something called varentropy, and the idea of log-Sobolev inequalities.(**Edit**: it turns out that we didn't get to this part.)

The course will have about three short homework assignments and the aforementioned project (the list of topics will be given to us). We won't follow any textbook, but if we want some additional reading, it's useful to read Levin and Peres' book "Markov chains and mixing times."

With that, we'll "start the course" – we can do a lot of the things in this class with measure theory, but often the ideas are clearest expressed in finite state spaces without any need to reference more complicated notation.

Let $\mathfrak{X}$ be a finite set, and let K be a **Markov transition matrix** (meaning that $K(x, y) \geq 0$ for all $x, y \in \mathfrak{X}$, and $\sum_y K(x, y) = 1$). Throughout this course, we'll generally assume (unless otherwise stated) that the chain is connected and doesn't have any periodicity issues, meaning that there is some ℓ such that $K^\ell(x, y) > 0$ for all x, y simultaneously. (This is often called **ergodicity**.) This implies that there exists a unique stationary distribution $\pi(x)$ on $\mathfrak{X}$, such that

$$\pi(x) > 0, \quad \sum_x \pi(x) = 1, \quad \sum_x \pi(x)K(x, y) = \pi(y).$$

In probability language, if we pick x from π and then take a step from the chain, then we're still distributed according to π; equivalently, the row vector π is a (left) eigenvector for the matrix K with eigenvalue 1. We write

$$K^\ell(x,y) = \sum_z K^{\ell-1}(x,z)\pi(z,y)$$

(we can just think of this as raising the matrix K to the ℓth power), and the fundamental theorem of Markov chains says that under these circumstances, we have $K^\ell(x,y) \to \pi(y)$ as $\ell \to \infty$ (so the rows of the matrix powers will get very close to each other and all look like the row vector π). We'll measure this convergence in various senses:

Definition 5

A Markov chain K with stationary distribution π is **reversible** if for all $x, y \in \mathfrak{X}$,

$$\pi(x)K(x,y) = \pi(y)K(y,x).$$

This also goes under the name "detailed balance" in some fields of science; it basically says that in stationarity, the chain looks the same when run forwards or backwards.

Example 6

Let $G = (\mathfrak{X}, E)$ be an undirected graph (with loops allowed), where E is some set of undirected edges between the vertices $|\mathfrak{X}|$. We can define $K(x,y)$ to be nearest-neighbor random walk on the graph G, meaning that

$$K(x,y) = \begin{cases} \frac{1}{|N_x|} & \text{if } x \sim y, \\ 0 & \text{otherwise,} \end{cases}$$

where $N_x = \{z \in \mathfrak{X} : (x,z) \in E\}$. In this example, we can compute that the stationary distribution is proportional to the number of adjacent edges, so $\pi(x) = \frac{|N_x|}{2|E|}$, and that K is reversible; indeed,

$$\pi(x)K(x,y) = \frac{1}{2|E|} = \pi(y)K(y,x)$$

if x, y are connected and 0 otherwise.

The same argument also works if we put weights on the edges and choose an edge in our random walk with weight proportional to $w(e)$; the same argument shows that

$$\pi(x) = \frac{W(x)}{Z}, \quad W(x) = \sum_{x \in e} w(e).$$

for Z the normalizing constant. And on the other hand, "any reversible Markov chain is a weighted random walk on a graph."

Most things that are run in practice are reversible, so it's important to be able to prove things about such chains. Some tools are particularly powerful in this setting:

Lemma 7

Let K be reversible with stationary distribution π on some finite state space $\mathfrak{X}$. Define $\ell^2(\pi)$ to be the function space $\{f : \mathfrak{X} \to \mathbb{R}\}$ with inner product $\langle f, g\rangle = \sum_{x\in\mathfrak{X}} f(x)g(x)\pi(x)$. Then for any $f, g \in \ell^2$, we have

$$\langle Kf, g\rangle = \langle f, Kg\rangle$$

(that is, K is a self-adjoint operator on ℓ^2), and in fact K is reversible if and only if it is self-adjoint on the corresponding space.

Proof. Form a basis of the function space with the functions $\delta_a(x) = 1\{x = a\}$. Taking $f = \delta_a, g = \delta_b$, we have (the following is how K acts on functions by thinking of them as column vectors)

$$Kf(x) = \sum_y K(x, y)f(y) = K(x, a),$$

so

$$\langle Kf, g\rangle = \sum_z K(z, a)\delta_b(z)\pi(z) = K(b, a)\pi(b),$$

and similarly

$$\langle f, Kg\rangle = K(a, b)\pi(a).$$

Thus the equality of these two expressions is exactly the same as reversibility checked at a and b, and then we can take linear combinations. □

Self-adjointness means that we have the spectral theorem (and the "best book" if we want to see the matrix analysis perspective on all of this is Horn and Johnson):

Proposition 8

For a reversible Markov chain (K, π), there exist real eigenvalues $\beta_0 \geq \beta_1 \geq \cdots \geq \beta_{|\mathfrak{X}|-1}$ and corresponding eigenvectors $\phi_i(x)$ such that $K\phi_i(x) = \beta_i\phi_i(x)$, such that the ϕ_is are orthonormal in $\ell^2(\pi)$.

This next fact is basically the main classical tool we have for bounding rates of convergence of Markov chains:

Proposition 9

Let $\mathfrak{X}$ be a finite set and (K, π) be a reversible Markov chain on $\mathfrak{X}$. Then we have the following:

1. We have $1 = \beta_0 \geq \beta_1 \geq \cdots \geq \beta_{|\mathfrak{X}|-1} \geq -1$ (with $\phi_0(x) = 1$ identically).
2. If the Markov chain is connected (meaning that for each x, y there is some $\ell = \ell(x, y)$ such that $K^\ell(x, y) > 0$), then $\beta_1 < 1$ (meaning that the largest eigenvalue is unique).
3. If the chain is connected and the graph corresponding to the Markov chain (where we connect x, y if $K(x, y) > 0$) is not bipartite, then $\beta_{|\mathfrak{X}|-1} > -1$,
4. For any $x \in \mathfrak{X}$, we have $\dfrac{1}{\pi(x)} = \displaystyle\sum_{i=0}^{|\mathfrak{X}|-1} \phi_i(x)^2$,
5. (Basic bound) The chi-square distance to stationarity is given by

$$\begin{aligned}
\chi_x(\ell) &= \sum_y \frac{(K^\ell(x, y) - \pi(y))^2}{\pi(y)} \\
&= \left|\left| \frac{K_x^\ell}{\pi} - 1 \right|\right|_2^2 \\
&= \sum_{i=1}^{|\mathfrak{X}|-1} \phi_i^2(x) \beta_i^{2\ell} \\
&\leq \frac{1 - \pi(x)}{\pi(x)} \beta_*^{2\ell} \\
&\leq \frac{\beta_*^{2\ell}}{\pi(x)},
\end{aligned}$$

where $\beta_* = \max(\beta_1, |\beta_{|\mathfrak{X}|-1}|)$ is the second absolute eigenvalue, and where $\frac{K_x^\ell}{\pi}$ is the vector of values $\frac{K^\ell(x,y)}{\pi(y)}$.

So in particular a chain which is connected and not bipartite has all nontrivial eigenvalues strictly within $(-1, 1)$, and so this chi-square distance goes down at least exponentially.

Proof. For (1), say that β, ϕ are an eigenvalue-eigenvector pair for the chain. Choose some state x^* such that $|\phi(x^*)|$ is maximized among all states (in particular it is nonzero). Then

$$\begin{aligned}
|\beta\phi(x^*)| &= |K\phi(x^*)| \\
&= \left| \sum_y K(x^*, y)\phi(y) \right| \\
&\leq \sum_y K(x^*, y)|\phi(y)| \\
&\leq |\phi(x^*)| \sum_y K(x^*, y) \\
&= |\phi(x^*)|,
\end{aligned}$$

so $|\beta| \leq 1$. Of course, $\beta_0 = 1$ and $\phi_0(x) \geq 1$ indeed form an eigenpair.

For (2), notice that if we follow the chain of inequalities and actually have equality for eigenvalue 1 and some eigenvector ϕ, then $\phi(y) = \phi(x^*)$ for all y which are one step away from x^*. But then we can repeat the argument with y in place of x^* and so on, and we find that ϕ is constant over all states by connectedness. Similarly for (3),

follow the chain of inequalities and suppose we have equality for eigenvalue -1 and some eigenvector ϕ. Then the argument above again has to yield equality, which means that $\phi(y)$ must be the negative of $\phi(x^*)$ for every y adjacent to x^*; repeating this yields a bipartition. (Equivalently, we use that a graph is not bipartite if and only if there is some cycle of odd length.)

Next for (4), let $d_x(y)$ be the function which is $\frac{1}{\pi(x)}$ if $x = y$ and 0 otherwise. Then for any eigenvector ϕ_i we have

$$\langle d_x, \phi_i \rangle = \sum d_x(y)\phi_i(y)\pi(y) = \phi_i(x),$$

and so we can expand out d_x in the eigenvector basis:

$$\begin{aligned} d_x(y) &= \sum_i \langle d_x, \phi_i \rangle \phi_i(y) \\ \overset{x=y}{\Longrightarrow} \quad \frac{1}{\pi(x)} &= \sum_{i=0}^{|\mathfrak{X}|-1} \phi_i(x)^2. \end{aligned}$$

And finally for (5), we have

$$\chi_x(\ell) = \sum_y \frac{(K^\ell(x,y) - \pi(y))^2}{\pi(y)} = \sum_y \left(\frac{K^\ell(x,y) - \pi(y)}{\pi(y)} \right)^2 \pi(y) = \left\| \frac{K_x^\ell}{\pi} - 1 \right\|_2^2,$$

and now we can compute

$$\left\langle \frac{K_x^\ell}{\pi}, \phi_i \right\rangle = \sum_y \frac{K^\ell(x,y)}{\pi(y)} \phi_i(y)\pi(y) = \sum_y K^\ell(x,y)\phi_i(y) = \beta_i^\ell \phi_i(x)$$

by the eigenvector equation. Thus we write out "by the Pythagorean theorem"

$$\begin{aligned} \left\| \frac{K_x^\ell}{\pi} - 1 \right\|_2^2 &= \sum_{i=0}^{|\mathfrak{X}|-1} \left\langle \frac{K_x^\ell}{\pi} - 1, \phi_i \right\rangle^2 \\ &= \sum_{i=1}^{|\mathfrak{X}|-1} (\beta_i^\ell \phi_i(x))^2 \\ &= \sum_{i=1}^{|\mathfrak{X}|-1} \beta_i^{2\ell} \phi_i^2(x), \end{aligned}$$

where in the second-to-last step we use that $\langle 1, \phi_i \rangle = 0$ for any eigenvector ϕ_i except the trivial one, and also that the $i = 0$ term is just zero because $\beta_0 = 1$ and $\phi_0 = 1$ identically. The remaining inequalities just come from bounding all of the different β terms by β^* and using our formula from part (4). □

We'll do a serious example where we can write everything down and see how the bounds go next time!

2 April 1, 2026

Last time, we described how to make use of eigenvalues to study mixing time for ergodic reversible Markov chains. Specifically, reversible Markov matrices can be thought of as self-adjoint operators on the function space $\ell^2(\pi)$ via $Kf(x) = \sum_y K(x,y)f(y)$, and its nontrivial eigenvalues lie in $(-1, 1)$. Thus, we can get an expression for the

chi-square distance to stationarity, and the important inequality that relates this to total variation is that

$$\begin{aligned} 4||K_x^\ell - \pi||_{\text{TV}}^2 &\le \chi_x^2(\ell) \\ &= \sum_y \frac{(K^\ell(x,y) - \pi(y))^2}{\pi(y)} \\ &= \sum_{i=1}^{|\mathfrak{X}|-1} \phi_i^2(x)\beta_i^{2\ell} \\ &\le \frac{\beta_*^{2\ell}}{\pi(x)}. \end{aligned}$$

The only inequality we didn't prove last time was the first one, but that's just the Cauchy-Schwarz inequality since

$$\begin{aligned} 4||K_x^\ell - \pi||_{\text{TV}}^2 &= \left(\sum_y |K^\ell(x,y) - \pi(y)|\right)^2 \\ &= \left(\sum_y \frac{|K^\ell(x,y) - \pi(y)|}{\sqrt{\pi(y)}}\sqrt{\pi(y)}\right)^2 \\ &\le \sum_y \frac{(K^\ell(x,y) - \pi(y))^2}{\pi(y)} \sum_y \pi(y) \\ &= \chi_x(\ell) \cdot 1. \end{aligned}$$

So the point is that having the eigenvalues gives us upper bounds on total variation for any ℓ. (And it's useful to keep track of the starting state, since it really can matter.) The reason Cauchy-Schwarz isn't actually so bad to use here is that the equality case comes when the two functions are proportional, and we make use of this inequality when ℓ is large enough so that the chain is close to stationarity and thus it's sharp enough to give us things like cutoff. The other inequality $\beta_i^{2\ell} \le \beta_*^{2\ell}$ can be a lot worse in general, but often it's all we really have unless we can calculate all of the eigenvalues and eigenvectors.

We'll now do an example which goes by the name of the **Ehrenfest urn**, and there's an important historical story behind this. Around 1890, there was a "rebellion" in chemistry and physics against statistical mechanics (as developed by Maxwell, Boltzmann, and many others); people didn't actually know that there were atoms and molecules at the time, and so the calculations were assuming we have something like 10^{23} unverifiable degrees of freedom. Simultaneously, there was the "Zermelo paradox," stated as follows: say we have n gas particles in the left half of a box, and there is a porous membrane which particles can pass through. It seems that "after a long time things are random," and Boltzmann had "proved" the ergodic hypothesis that entropy increases. But Poincaré had also proved the recurrence theorem, which stated that in any dynamical system like this, eventually it must be back where it starts – in particular, we should get back to the starting state, which is not maximum entropy. And all of this isn't "just talk:" statistical mechanics was actually banned from talks at conferences for a while and Boltzmann took his own life. Well, in 1906, Einstein showed that Brownian motion of water particles could be explained by atoms, and if we want more about this, we can read the book "Boltzmann's Atom."

In 1910, the Ehrenfests introduced a simple conceptual model which could explain this:

Example 10 (Ehrenfest urn)

Suppose we have n balls in two urns (which we can think of as the two halves of the box), and at each step, pick a ball at random and move it to the other urn. Then the number of balls X in the left urn, evolves as a Markov chain on $\{0, 1, \cdots, n\}$:

$$K(i, i-1) = \frac{i}{n}, \quad K(i, i+1) = 1 - \frac{i}{n}.$$

Since each ball is equally likely to be in the left or right urn after a long time, it must be that the stationary distribution is Binomial$(n, \frac{1}{2})$, and that's easy to verify: the chain turns out to be reversible with $\pi(j) = \frac{1}{2^n}\binom{n}{j}$. So we can ask a question like "take $n = 10$; we know that after some amount of time all of the balls will be back in the left urn again," and recurrence is not a surprise here. And on the other hand, let

$$P^\ell(j) = K^\ell(n, j)$$

be the distribution of the number of balls in the left urn after ℓ steps. Then the entropy $-\sum_j P^\ell(j) \log P^\ell(j)$ can be checked to be monotone increasing, so there's nothing particularly contradictory about those two facts both being true.

We'll analyze this Markov chain today using the techniques we introduced last time. The first key idea here is that even though we can describe the chain just in terms of the number of balls in the left urn, we can also code up the chain by keeping track of all n balls separately; that is, instead use the state space

$$\mathfrak{X} = C_2^n = \{(x_1, \cdots, x_n) : x_i \in \{0, 1\}\},$$

where $x_i = 0$ if the ball is in the left urn and 1 if it is in the right urn. But then the Ehrenfest urn is actually just **nearest-neighbor random walk on the hypercube** C_2^n, since the dynamics of the Ehrenfest urn pick a ball at random and swap that one coordinate. And since the hypercube is an abelian group, "lifting the Markov chain" turns out to be a good idea (and the question of when we can lift Markov chains is interesting but not something we'll get into here).

The next idea is that actually the chain we started with has a parity problem: after an even (resp. odd) number of steps, the parity of the number of balls in the left urn is the same (resp. different) as the starting state. One way this can be fixed is to make the chain have a small probability of "doing nothing," which we can do by putting a holding probability of $\frac{1}{n+1}$ at each step. So the chain we will analyze has the following dynamics on C_2^n:

$$K(x, y) = \begin{cases} \frac{1}{n+1} & \text{if } y = x + e_i \text{ for } 1 \le i \le n, \\ \frac{1}{n+1} & \text{if } y = x, \\ 0 & \text{otherwise,} \end{cases}$$

where e_i is the ith basis vector $(0, 0, \cdots, 1, \cdots, 0, 0)$. This chain is ergodic and has uniform stationary distribution $\pi(x) = \frac{1}{2^n}$ (because it's symmetric).

Proposition 11

The Markov chain K above has an orthonormal basis of eigenvectors

$$\{\phi_z(x) = (-1)^{x \cdot z} : z \in C_2^n\},$$

where $x \cdot z$ is the usual dot product. The eigenvalue of ϕ_z is $\beta_z = 1 - \frac{2|z|}{n+1}$, where $|z|$ is the number of ones in z.

Of course, this means the eigenvalues have high multiplicity: β_z has multiplicity $\binom{n}{|z|}$.

Proof. We must show that $K\phi_z(x) = \beta_z\phi_z(x)$, so we will compute directly. We have

$$\begin{aligned} K\phi_z(x) &= \sum_{y \in C_2^n} K(x,y)\phi_z(y) \\ &= \frac{1}{n+1}\left(\phi_z(x) + \sum_{i=1}^{n} \phi_z(x+e_i)\right) \\ &= \frac{1}{n+1}\left((-1)^{z\cdot x} + (-1)^{z\cdot(x+e_1)} + \cdots + (-1)^{z\cdot(x+e_n)}\right) \end{aligned}$$

since those are the only values for which $K(x,y)$ is nonzero. And now we can factor things out to get

$$K\phi_z(x) = \frac{(-1)^{z\cdot x}}{n+1}\left(1 + (-1)^{z_1} + \cdots + (-1)^{z_n}\right)$$

where z_i is the value in the ith coordinate of z; thus exactly $|z|$ of these terms in the sum are -1 and the other $n+1-|z|$ are $+1$, leading to

$$K\phi_z(x) = \phi_z(x) \cdot \frac{n+1-2|z|}{n+1} = \beta_z\phi_z(x).$$

And the orthogonality of these functions can be explicitly checked by writing out the sum. □

This is exactly Fourier analysis on the hypercube." The first person to diagonalize the chain was Mark Kac (in a rather complicated way on the original chain), and then later it was seen that lifting to the hypercube gives a simpler proof. And so the point is that we can now apply those eigenvalue bounds that we proved previously. We get that the chi-square distance to stationarity after ℓ steps is

$$\begin{aligned} \chi_x^2(\ell) &= \sum_{z \neq 0} \phi_z^2(x)\left(1 - \frac{2|z|}{n+1}\right)^{2\ell} \\ &= \sum_{z \neq 0}\left(1 - \frac{2|z|}{n+1}\right)^{2\ell} \end{aligned}$$

(so things don't depend on where we start, and that's a feature of having a random walk on the group), and now we have a lot of multiplicity so we can group up the sum:

$$\chi_x^2(\ell) = \sum_{i=1}^{n} \binom{n}{i}\left(1 - \frac{2i}{n+1}\right)^{2\ell}.$$

We can now use the elementary inequalities $\binom{n}{i} \leq \frac{n^i}{n!}$ and $1 - x \leq e^{-x}$ to simplify this further. We can "fold the sum in half" so we only have to go up to $\frac{n}{2}$, and we get

$$\begin{aligned} \chi_x^2(\ell) &= 2\sum_{i=1}^{n/2} \binom{n}{i}\left(1 - \frac{2i}{n+1}\right)^{2\ell} \\ &\leq 2\sum_{i=1}^{n/2} \frac{n^i}{i!} e^{-4i\ell/(n+1)} \\ &\leq 2\sum_{i=1}^{\infty} \frac{1}{i!} e^{i\log n - 4i\ell/(n+1)}. \end{aligned}$$

But now we can choose an "intelligent" ℓ: if we set $\ell = \frac{n+1}{4}(\log n + c)$ (we can put in "integer parts" but we won't be

careful about that here), most of the factors in the exponential cancel out and we get that **for this specific** ℓ,

$$\begin{aligned}\chi_x^2(\ell) &\leq 2\sum_{i=1}^{\infty}\frac{1}{i!}e^{-ic}\\ &= 2\sum_{i=1}^{\infty}\frac{1}{i!}(e^{-c})^i\\ &= 2\left(e^{e^{-c}}-1\right).\end{aligned}$$

So we see that if c is a little bit large (like 10), this right-hand side is quite small because $e^{e^{-c}}-1\approx e^{-c}$. Applying that $4||K_x^\ell-\pi||_{\mathrm{TV}}^2\leq\chi_x^2(\ell)$, we've thus proved the following:

Theorem 12

Suppose that $\ell=\frac{n+1}{4}(\log n+c)$. Then we have the total variation bound to stationarity

$$4||K_x^\ell-\pi||_{\mathrm{TV}}^2\sim 2e^{-c};$$

that is, the chain converges exponentially fast in a linear number of steps after $\frac{1}{4}n\log n$ steps.

Of course, we used a lot of inequalities and threw away terms, but it turns out this is tight.

Remark 13. *Suppose we used the "bad bound"* $4||K_x^\ell-\pi||_{TV}^2\leq\frac{\beta_*^{2\ell}}{\pi(x)}$ *where we bounded all eigenvalues by the worst one. Plugging in what we know, this yields*

$$\begin{aligned}4||K_x^\ell-\pi||_{TV}^2 &\leq 2^n\left(1-\frac{2}{n+1}\right)^{2\ell}\\ &\sim e^{n\log 2-4\ell/(n+1)}.\end{aligned}$$

And this is much worse: it tells us that we need to take $\ell=\frac{n+1}{4}(n+c)$ *of order* n^2 *instead of* $n\log n$ *to guarantee a good upper bound, and then we don't get the correct answer. And the difference between* n^2 *and* $n\log n$ *actually does matter in practice – it's the whole point of the fast Fourier transform – but if we didn't know all of this stuff it's not so bad. And in many of the Markov chains we're stuck on, getting a spectral gap alone would already be a big step!*

Turning now to lower bounds, we would often say "a similar argument shows the lower bound matches." But we'll actually do it out carefully today here for chi-square distance. We've just proven that x$\ell=\frac{n+1}{4}(\log n+c)$ gives us chi-square at most $2(e^{e^{-c}}-1)$, and for a matching bound we can just keep some of the terms (since everything in sight is nonnegative). Only keeping the ones where $|z|=1$, we get

$$\begin{aligned}\chi_x^2(\ell) &= \sum_{z\neq 0}\phi_z^2(x)\left(1-\frac{2|z|}{n+1}\right)^{2\ell}\\ &\geq n\left(1-\frac{2}{n+1}\right)^{2\ell}\\ &\sim e^{\log n-4\ell/(n+1)},\end{aligned}$$

and thus if we take $\ell=\frac{n+1}{4}(\log n+c)$ we get e^{-c}. So things match up to a factor of 2, and thus at least in ℓ^2 we haven't been sloppy. Thus we can say that we have a **cutoff in** ℓ^2: unlike total variation, chi-square can go to infinity, but what ℓ^2 cutoff means is that we go from "going to infinity" to "going to zero exponentially fast" in a smaller window than the mixing time.

But this still doesn't get us a lower bound in ℓ^1, since we still make use of Cauchy-Schwarz. In our homework, we prove that the total variation can also be expressed alternatively as

$$\begin{aligned}||K_x^\ell - \pi||_{\text{TV}} &= \max_A |K^\ell(x, A) - \pi(A)| \\ &= \frac{1}{2}\sum_y |K^\ell(x, y) - \pi(y)|.\end{aligned}$$

Usually we're using the latter form, but the first version says that we can get lower bounds by computing $K^\ell(x, A)$ and $\pi(A)$ for any specified set A. And this idea is due to Professor Diaconis and goes under the name "second moment method:" define the function

$$Z(x) = \sum_{i=1}^n (-1)^{x_i}$$

which is the sum of all of the largest nontrivial (β_1) eigenvectors. This is orthogonal to the trivial eigenvector, and thus $\mathbb{E}_\pi[Z] = 0$. Furthermore, the coordinates of this function are independent, so $\text{Var}(Z) = n$. Thus under π, the function is centered at zero and has standard deviation $\sqrt{n}$.

But under K^ℓ, because we have an eigenvector, we must have $\mathbb{E}_\ell[Z] = \left(1 - \frac{2}{n+1}\right)^\ell \cdot n$ if we start from the all-0s state. And if we want the variance, we have to square Z and that gives us $(-1)^{x_i + x_j}$ functions, which are also okay because they are also eigenfunctions. We end up finding that $\mathbb{E}[Z^2] = n + n(n-1)\left(1 - \frac{4}{n+1}\right)^\ell$, and after some busy work we see that if $\ell = \frac{n+1}{4}(\log n + c)$ (for c positive or negative), $\mathbb{E}_\ell[Z] = \sqrt{n}e^{-c/2}\left(1 + O(\frac{\log n}{n})\right)$ and $\text{Var}_\ell(Z) = n + O(e^{-c}\log n)$ (and these big-O estimates are actually uniform in c up to $\log n$). The point is that for c a little bit small, $\mathbb{E}_\ell[Z]$ is big: the mean is something like $10\sqrt{n}$ and the standard deviation is just $\sqrt{n}$, so under this measure Z looks very different from the measure we would have under π. We can now phrase this in terms of that total variation set A:

Proposition 14

Fix some $\beta > 0$ and let $A = \{x : |Z(x)| \le \beta\sqrt{n}\}$. By Chebyshev's inequality, we have

$$\pi(A) \ge 1 - \frac{\text{const}}{\beta^2}, \quad K_0^\ell(A) \le \frac{2}{(e^{-c/2} - \beta)^2}.$$

Thus choosing c negative and β appropriately shows the total variation distance must be large.

With that, we've proven an actual cutoff in total variation for the Ehrenfest urn: we go from TV near 1 to TV near 0 in a short interval (of length n) around $\frac{n+1}{4}\log n$. And we can get a feeling for the quality of the bounds and what they're good for.

It's natural to ask whether we can say something about what the total variation distance looks like as a function of c, and Professor Diaconis showed in a paper "Asymptotic analysis of a random walk on a hypercube with many dimensions" with Graham and Morrison that if $\ell = \frac{n+1}{4}(\log n + c)$ for $c \in \mathbb{R}$ fixed,

$$||K_x^\ell - \pi||_{\text{TV}} \sim \text{Erf}(e^{-2c}/\sqrt{8})$$

for $\text{Erf} = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ the error function for the normal distribution. So we're actually doubly exponentially close to 1 if c is negative and exponentially close to 0 if c is positive, meaning that the profile is quite sharp.

So "a miracle really did occur" here: we were able to diagonalize the Markov chain here exactly, but to think about what really happens imagine erasing a single edge of the hypercube. It's quite hard to prove that nothing really changes in that setting, and so what's really important for "being able to do all this" is that we had a random walk on an abelian group C_2^n where we can do Fourier analysis. And we can do similar kinds of things on nonabelian groups as well (it's a different subject, but there are known formulas and lots of techniques). If we want to learn more about this, we can

check out Professor Diaconis' book "Group representations in probability and statistics."

3 April 6, 2026

We're in the middle of developing analytic (geometric) theory for Markov chains now, and we'll "get someplace" today. Our setting so far has been that $K(x,y)$ is the transition matrix of some finite-state Markov chain, reversible with respect to some stationary distribution $\pi(x)$ (for example, random walk on a graph with edge weights). We then study such a chain by looking at the function space $\ell^2(\pi)$ and having K act via $Kf(x) = \sum_y K(x,y)f(y)$ (multiplication on column vectors); this setup then guarantees that we have real eigenvectors and eigenvalues for the self-adjoint operator K, which we can use to bound the chi-square distance to stationarity after any number of steps.

Last time, we did this by explicitly computing the eigenvalues in a special case. More generally, we typically get our hands on eigenvalues by thinking about quadratic forms and taking minima and maxima; we'll write this down in a unified language now.

Definition 15

The **Laplacian operator** on $\ell^2(\pi)$ is given by $L = I - K$, meaning that L acts on functions via

$$Lf(x) = f(x) - Kf(x).$$

This is also self-adjoint and has eigenvalues $1 - \beta_i$, and this is the convention that analysts use ("eigenvalues start at 0 and go up," rather than "eigenvalues start at 1 and go down"). But the interpretation probabilistically for L is that Laplacians "replace functions by second difference operators" if we're doing random walk on $\mathbb{Z}^d$, for example.

Definition 16

The **Dirichlet form** $\mathcal{E}(f,g)$ is defined by

$$\mathcal{E}(f,g) = \langle (I-K)f, g\rangle.$$

Writing out the definition of the inner product and then using reversibility yields the following:

Proposition 17

We have

$$\mathcal{E}(f,g) = \frac{1}{2}\sum_{x,y}(f(x)-f(y))(g(x)-g(y))\pi(x)K(x,y).$$

In particular, since $\pi(x)K(x,y) = \pi(y)K(y,x)$, this is a symmetric form.

(This is something we have to check for ourselves to believe, so we won't write it out carefully here.) In the special case where $f = g$, the Dirichlet form

$$\mathcal{E}(f,f) = \frac{1}{2}\sum_{x,y}(f(x)-f(y))^2\pi(x)K(x,y)$$

is always positive semidefinite (and in fact positive definite). We have the minimax characterization of eigenvalues, and we can say it in the following way

Proposition 18

With the notation above, we have the expression for the spectral gaps

$$1-\beta_1 = \min_{f \text{ non-constant}} \frac{\mathcal{E}(f,f)}{\mathrm{Var}(f)}, \quad 1-\beta_{|\mathfrak{X}|-1} = \max_{f \text{ non-constant}} \frac{\mathcal{E}(f,f)}{\mathrm{Var}(f)},$$

where the variance of f (under π) is given by

$$\mathrm{Var}(f) = \sum_x (f(x)-\overline{f})^2 \pi(x) = \frac{1}{2}\sum_{x,y}(f(x)-f(y))^2\pi(x)\pi(y)$$

for $\overline{f} = \sum_x f(x)\pi(x)$ the mean of f. (These last two expressions are equal by expanding out $f(x)-f(y) = (f(x)-\overline{f})-(f(y)-\overline{f})$.)

Proof. Let $W = \left\{f \in \ell^2(\pi) : \langle f, 1\rangle = 0\right\}$ be the set of functions orthogonal to the constants (so our functions will all have mean zero). It's okay to only look at these functions because subtracting off a constant keeps both the numerator and denominator the same.

For any $f \in W$ we thus have $\mathrm{Var}(f) = ||f||_2^2$. The right-hand side is also homogeneous in f, so we can assume $||f||_2^2 = 1$. Expanding out in the eigenbasis, we have

$$f = \sum_{i=1}^{|\mathfrak{X}|-1} \phi_i(x)\langle \phi_i, f\rangle$$

(the $i=0$ term disappears because f is orthogonal to the β_0-eigenfunction, which is a constant). Thus $I-K$ acts on each term of this sum as a constant, meaning

$$(I-K)f = \sum_{i=1}^{|\mathfrak{X}|-1}(1-\beta_i)\phi_i(x)\langle\phi_i, f\rangle.$$

Therefore by orthonormality of the eigenfunctions,

$$\begin{aligned}\mathcal{E}(f,f) &= \langle (I-K)f, f\rangle \\ &= \sum_{i=1}^{|\mathfrak{X}|-1}(1-\beta_i)\langle\phi_i,f\rangle^2 \\ &\geq (1-\beta_1)\sum_{i=1}^{|\mathfrak{X}|-1}\langle\phi_i,f\rangle^2 \\ &= 1-\beta_1\end{aligned}$$

and taking the minimum over all $f \in W$ gets us one side of the inequality; equality occurs if we take $f = \phi_1$.

The other bound is the same but with the inequality $1-\beta_i \leq 1-\beta_{|\mathfrak{X}|-1}$ instead (and equality is achieved at $f = \phi_{|\mathfrak{X}|-1}$). □

We'll now use this to get eigenvalue bounds, specifically **Poincaré inequalities**. The idea is that when we run into a Markov chain, the first thing to try is the following. We **choose some collection of paths** $\{\gamma_{xy} : x, y \in \mathfrak{X}\}$, where γ_{xy} is a connected path from x to y (meaning that we have $[x_0, x_1, x_2, \cdots, x_\ell]$ with $x_0 = x, x_\ell = y$, and $K(x_i, x_{i+1}) > 0$ for all i). (It's common to choose some geodesic (shortest path) if one is clear, and we also often choose the reverse path for γ_{yx}, but we don't necessarily have to.)

Theorem 19

Given the paths $\{\gamma_{xy}\}$ and the Markov chain K as above, we have $\beta_1 \le 1 - \frac{1}{A}$, where

$$A = \max_{\text{edges } e} \frac{1}{Q(e)} \sum_{\text{paths } \gamma_{xy} \ni e} \pi(x)\pi(y)|\gamma_{xy}|,$$

where $|\gamma_{xy}|$ is the length ℓ of the path (the number of edges), and for any edge $e = (z, w)$ we have $Q(e) = \pi(z)K(z, w)$.

If we haven't seen this before, we might be a bit confused how we might actually think about such a quantity A, but the point is that in many real examples we can choose paths and actually get good bounds. Let's first prove it:

Proof. For any (non-constant) function f, we have

$$\mathrm{Var}(f) = \frac{1}{2}\sum_{x,y}(f(x) - f(y))^2\pi(x)\pi(y),$$

and now replace $f(x) - f(y)$ with a telescoping sum along the path γ_{xy}. We find that (letting e be directed edges from e^- to e^+)

$$\begin{aligned}
\mathrm{Var}(f) &= \frac{1}{2}\sum_{x,y}\Big((f(x_0) - f(x_1)) + (f(x_1) - f(x_2)) + \cdots + (f(x_{\ell-1}) - f(x_\ell))\Big)^2\pi(x)\pi(y) \\
&= \frac{1}{2}\sum_{x,y}\Big(\sum_{e\in\gamma_{xy}} f(e^-) - f(e^+)\Big)^2\pi(x)\pi(y) \\
&\le \frac{1}{2}\sum_{x,y}|\gamma_{xy}|\sum_{e\in\gamma_{xy}}(f(e^-) - f(e^+))^2\pi(x)\pi(y)
\end{aligned}$$

by using Cauchy-Schwarz on the blue part. So now swapping the order of summation yields

$$\begin{aligned}
\mathrm{Var}(f) &= \sum_e (f(e^-) - f(e^+))^2 \sum_{\gamma_{xy}\ni e} |\gamma_{xy}|\pi(x)\pi(y) \\
&= \sum_e (f(e^-) - f(e^+))^2 \frac{Q(e)}{Q(e)} \sum_{\gamma_{xy}\ni e} |\gamma_{xy}|\pi(x)\pi(y) \\
&\le A\mathcal{E}(f, f),
\end{aligned}$$

because the red part is exactly the Dirichlet form and the rest is exactly the definition of A. Thus rearranging yields that $1 - \beta_1 \ge \frac{1}{A}$ and thus $\beta_1 = 1 - \frac{1}{A}$, as desired. □

This result is the start of what we call the "geometry of Markov chains."

Example 20

Let's clean this up a bit and specialize to nearest-neighbor random walk on a (simple undirected) graph, meaning that from a vertex x we go to one of its $d(x)$ neighbors with equal probability (so that we have reversibility with respect to the stationary distribution $\pi(x) = \frac{d(x)}{2|E|}$).

In this case, we then have $Q(e) = \frac{d(x)}{2|E|} \cdot \frac{1}{d(x)} = \frac{1}{2|E|}$ for any edge, so

$$\begin{aligned} A &= \max_e 2|E| \sum_{\gamma_{xy} \ni e} \pi(x)\pi(y)|\gamma_{xy}| \\ &\leq 2|E| \left(\frac{d_*}{2|E|}\right)^2 \gamma_* \cdot \max_e \sum_{\gamma_{xy} \ni e} 1 \end{aligned}$$

for d_* the max degree and γ_* the max length of any path (this is a sloppy bound by just pointwise bounding every term). Calling this maximum β_* (that is, β_* is the maximum number of paths using a common edge – eventually we'll learn to think about this as curvature), we find that

$$A \leq \frac{d_*^2 \gamma_* \beta_*}{2|E|} \implies \beta_1 \leq 1 - \frac{2|E|}{d_*^2 \gamma_* \beta_*}.$$

So if we can choose paths between pairs of points which don't use too many common overlapping edges, that keeps β_* small and that gets us a good bound. But if we have a graph which looks something like a complete graph K_n and another complete graph K_n linked only by a single path, then any edge in that middle path requires a lot of common paths (and geometers know that this is indeed some kind of measure of curvature).

Of course, we need to see examples to be convinced and understand how to use this. As sloppy as we've been, it turns out to not be so bad and people have done it in hundreds of cases:

Example 21

Consider a graph where we have a path from vertex 1 to n and a loop only at the endpoints 1 and n. That is, $K(i, i+1) = K(i, i-1) = \frac{1}{2}$ and $K(1,1) = K(n,n) = \frac{1}{2}$, and these self-loops mean that the chain has uniform stationary distribution.

There's only one way to choose paths on this graph if we're not being silly – we just make γ_{ij} the shortest path from i to j, and then we have

$$A = \max_e \frac{1}{Q(e)} \sum_{\gamma_{xy} \ni e} \pi(x)\pi(y)|\gamma_{xy}|.$$

Edges from a vertex to themselves don't actually come into the picture here (because of how we set up the proof), and for any edge $e = (i, i+1)$ we have $Q(e) = \frac{1}{2n}$. If we just bound $|\gamma_{ij}| \leq n$ for all i, j (we'll be a little more careful later), we find that

$$A \leq \frac{2n}{n^2} \cdot n \cdot \max_i \#\{\text{paths passing through } (i, i+1)\}.$$

But the paths using this edge are all connecting something on the left and something on the right, and so in particular the covering number is at most n^2. Plugging this in yields $A \leq 2n^2$ so $\boxed{\beta_1 \leq 1 - \frac{1}{2n^2}}$.

Well, in this problem, we know exactly what the eigenvalues are; they are

$$\beta_j = \cos\left(\frac{\pi j}{n}\right) = 1 - \frac{1}{2}\left(\frac{\pi j}{n}\right)^2 + O\left(\left(\frac{j}{n}\right)^4\right).$$

In particular this means β_1 is $1 - \frac{\pi^2}{2n^2} + O(\frac{1}{n^4})$, and so we're only off by a factor of π^2 here even though we've been a slob. If we do the paths bound more carefully (the worst-case edge is the middle one which has $\frac{n^2}{2}$ paths through it, and also we gain a factor of 2 from paths typically being half of the maximum length), we get $1 - \frac{2}{n^2}$ instead. So really it's not so bad.

Fact 22

The "negative eigenvalues" is a different idea which requires a different lecture; what we need to do is require paths of odd length (see Proposition 24 below). It can then be shown that $\beta_{|\mathfrak{X}|-1} \geq -1 + \frac{4}{n^2}$, so the absolute spectral gap $\beta_* = \max\left(1-\beta_1, 1+\beta_{|\mathfrak{X}|-1}\right)$ is at most $\frac{2}{n^2}$. So using our chi-square bounds to stationarity, we see that

$$4||K_x^\ell - \pi||_{\mathrm{TV}}^2 \leq n\left(1 - \frac{2}{n^2}\right)^{2\ell}$$

and the right-hand side is small after order $n^2 \log n$ steps. As we go on, in fact the answer is just n^2 (we don't need the extra log), but we're pretty close to the right answer.

This example we just did generalizes to a higher-dimensional grid (perhaps with loops on the edges). But then when we have any two points, we have a choice of which path connects them; it turns out we can make the choice randomly or deterministically and lots of different things give us pretty good answers. It turns out for any low-dimensional grid the right answer is n^2, and this gives us $n^2 \log n$ in some generality.

Example 23

Next, let's return to the Ehrenfest urn (since this starts to be more like a high-dimensional problem). As we previously discussed, this is nearest-neighbor random walk on the hypercube C_2^d where we have probability $\frac{1}{d+1}$ of staying and probability $\frac{1}{d+1}$ of adding any standard basis vector.

We have to choose paths, and in most problems we try to do "it doesn't really matter which choice we make;" one thing we can do is say that if $x = (x_1, \cdots, x_d)$ and $y = (y_1, \cdots, y_d)$, then we'll work left-to-right and fix coordinates one by one (so if $x_i = y_i$ try the next coordinate instead, but otherwise make the next edge of the path "add e_i"). The maximum path length is just d, and $\pi(x) = \frac{1}{2^d}$ for any x, and the degree of any vertex is $d+1$ and we have $d2^{d-1}$ edges in the hypercube, so

$$\beta_1 \leq 1 - \frac{2d2^{d-1}}{d(d+1)^2\beta_*}$$

and all that's left is to bound the covering number β_*. Well, any edge is between two d-tuples which are the same except differing at one coordinate, and any path can start anywhere to the left of that (2^{i-1}) and go anywhere to the right ($2^{d-(i-1)}$); that is, it's just 2^d everywhere. So putting everything together yields

$$\beta_1 \leq 1 - \frac{1}{(d+1)^2},$$

which is a nice expression but unfortunately wrong (remember that the right answer is $1 - \frac{2}{d+1}$). And Professor Diaconis tried for a long time to fix this in the original paper ("Geometric Bounds for Eigenvalues of Markov Chains" with Daniel Stroock) and wasn't able to do so. But **on the other hand**, if we use paths directly on the "lumped" Ehrenfest urn chain on the state space $\{0, 1, \cdots, n\}$, there's only one possible choice of paths and now it gives us the right answer. So somehow the symmetry that we use when we lift loses a factor of d.

We'll do one of the "real examples" from computer science theory next time (random walk on the space of matchings) again using paths.

4 April 8, 2026

The TA for the course, Nathan Tung, will hold office hours on Fridays from 3-5pm in the stats department library (downstairs).

We've been doing path arguments (Poincaré and eigenvalue type argments), and to familiarize ourselves more with this we'll now talk about **lower bounds on negative eigenvalues**. As a sidenote, next Monday's lecture will be by Professor Sourav Chatterjee, and it will be about how to do similar things on non-symmetric chains.

As usual, suppose $\mathfrak{X}$ is a finite set and we have a reversible, ergodic Markov chain (K, π) on $\mathfrak{X}$. We've previously been bounding things on the side of $1 = \beta_0 > \beta_1$, and now we'll think about how to bound the other side $\beta_{|\mathfrak{X}|-1} > -1$. The statement will look somewhat similar to Theorem 19:

Proposition 24

In the setting above, we have $\beta_{|\mathfrak{X}|-1} \geq -1 + \frac{2}{B}$, where

$$B = \max_{\text{edges } e} \frac{1}{Q(e)} \sum_{\text{paths } \sigma_x \ni e} |\sigma_x| \pi(x),$$

where $Q(e) = \pi(z)K(z, w)$ as before, $\{\sigma_x\}_{x \in \mathfrak{X}}$ is a collection of paths σ_x of odd length from x to x, and where the maximum includes self-loop edges as well.

Proof. Instead of the quadratic form used last time, we'll consider the quadratic form

$$\begin{aligned}\mathcal{F}(f, g) &= \langle (I + K)f, g \rangle \\ &= \frac{1}{2} \sum_{x,y \in \mathfrak{X}} (f(x) + f(y))(g(x) + g(y))\pi(x)K(x, y).\end{aligned}$$

Since the eigenvalues of $I + K$ are 1 plus the β_is, a similar variational characterization yields that

$$1 + \beta_{|\mathfrak{X}|-1} = \min_{f \text{ nonzero}} \frac{\mathcal{F}(f, f)}{||f||_2^2}$$

(note "nonzero" instead of "nonconstant" here, and $||f||_2^2$ instead of $\text{Var}(f)$). Thus, we just need a bound of the form $||f||_2^2 \leq \frac{B}{2}\mathcal{F}(f, f)$, which will imply that $\frac{2}{B} \leq \frac{\mathcal{F}(f,f)}{||f||_2^2}$, so $\frac{2}{B} \leq 1 + \beta_{|\mathfrak{X}|-1}$.

To get such a bound, the trick in the proof is that if $\sigma_x = [x, y, z, x]$ is a path of length 3, we can write

$$f(x) = \frac{1}{2}\Big((f(x) + f(y)) - (f(y) + f(z)) + (f(z) - f(x)\Big)$$

(this cancels out everything except the initial and final $f(x)$), and then use Cauchy-Schwarz. Indeed,

$$\begin{aligned}||f||_2^2 &= \sum_x f(x)^2 \pi(x) \\ &= \sum_x \Big(\frac{1}{2} \sum_{e \in \sigma_x} (-1)^{p(e)} (f(e^-) + f(e^+))\Big)^2 \pi(x)\end{aligned}$$

where $p(e)$ is the "parity of the edge" which alternates between $+1$ and -1. By Cauchy-Schwarz on the inner sum we thus have

$$||f||_2^2 \leq \sum_x \frac{1}{4} |\sigma_x| \sum_{e \in \sigma_x} \big(f(e^-) + f(e^+)\big)^2 \pi(x),$$

so now swapping the order of summation this yields

$$||f||_2^2 \leq \sum_e \left(f(e^+) + f(e^-)\right)^2 \frac{Q(e)}{Q(e)} \sum_{\sigma_x \ni e} |\sigma_x| \pi(x).$$

Now the blue part can be upper bounded by B and so this whole thing is at most $\frac{1}{2}\mathcal{F}(f,f) \cdot B$, which is exactly what we wanted to show. □

So in our two arguments, we add along the path in one case and subtract along it in another, and we might ask what else is possible. We'll talk about that a bit more later, but first we'll do some examples:

Example 25

Suppose the chain has holding, meaning that there is some $\varepsilon > 0$ such that $K(x,x) \geq \varepsilon$ for all x. This lets us choose σ_x to be just the loop e_x at x, and for each such loop we have $Q(e_x) = \pi(x)K(x,x) \geq \varepsilon Q(e_x)$. Thus the terms in the maximum for B are all of the form $\frac{1}{Q(e_x)}\pi(x) = \frac{1}{K(x,x)}$, and so we can take $B = \frac{1}{\varepsilon}$. That is, ε-holding means the smallest eigenvalue is at least $-1 + 2\varepsilon$.

Here are two concrete examples:

- Consider the Ehrenfest urn on the hypercube that we've been studying in the last few lectures (Example 23). We know that random walk on C_2^d includes a holding term $K(x,x) = \frac{1}{d+1}$, so this tells us that $\beta_{|\mathfrak{X}|-1} \geq -1 + \frac{2}{d+1}$ and this is exactly sharp (that is exactly the correct value of the smallest eigenvalue).
- Next, consider the random walk on a path with holding at the two endpoints (Example 21). Then our odd-length paths σ_x have to go to one of the endpoints, take the loop, and then return back. In that case we have $K(x,y) = \frac{1}{2}$ for any existing edge and $\pi(x) = \frac{1}{n}$ for all x, so $Q(e) = \frac{1}{2n}$. Also, we can bound the longest path length $|\sigma_x|$ by n. Thus we can take

  $$B = \max_e \frac{1}{Q(e)} \sum_{\sigma_x \ni e} |\sigma_x| \pi(x) \leq 2n \sum_{\sigma_x \ni e} 1.$$

  The worst-case edges are the loops on the endpoints, one of which will need to account for $\frac{n}{2}$ of the vertices, and thus we can take $B = n^2$. Thus we have indeed shown that $\beta_{n-1} \geq -1 + \frac{2}{n^2}$ for this chain, as promised (and maybe we can get a factor of 2 better, but it's the right order of magnitude because we know the asymptotics of cosine).

Note that if $K(x,x) \geq \frac{1}{2}$ for all x, all eigenvalues are nonnegative and we don't have to worry about negative eigenvalues at all. "Most people" (especially in computer science) use lazy chains, meaning that they use $\tilde{K} = \frac{I+K}{2}$ instead of K (so with probability $\frac{1}{2}$ do nothing, and with probability $\frac{1}{2}$ run the chain K). That's fine, but if we want to get exactly the right answer and prove cutoff we probably don't want to do this since it's a different chain. And in practice if we wanted to run a chain, we wouldn't want to make it lazy and take twice as many steps anyway.

Another way to get away from negative eigenvalues is to work in continuous time (so that instead of making steps at time $1, 2, 3, \cdots$, we have a Poisson process of rate 1 and make steps when the clock rings). And we could also analyze the chain after an even number of steps and an odd number of steps separately if we really had to – the rates can be different.

All of the examples so far have been "easy examples" in hindsight, so we'll now talk about a real one where the machinery of path arguments was helpful:

Example 26

Suppose we want to approximate the **permanent** of a matrix, which is famous in computer science theory for being #P-complete. For an $n \times n$ matrix A, define

$$\text{per}(A) = \sum_{\sigma \in S_n} \prod_{i=1}^{n} A_{i\sigma_i}$$

(so this is like the determinant but without the alternating sign factor).

If we haven't seen this quantity before, we might ask why people care about it – one motivation is that for a bipartite graph with n boys and n girls where some of the boy-girl pairs like each other, we can define an incidence matrix A where $A_{ij} = 1$ if the ith boy and the jth girl like each other and 0 otherwise. Then $\text{per}(A)$ is the total number of possible **perfect matchings** into n pairs, and matching problems are very important in lots of real applications (medical school matchings, rideshare programs, and so on) – we can see Lovśz and Plummer's book "Matching Theory" for more.

In this setting, it is easy to tell whether or not such a graph has a perfect matching (there's an $O(n^{2.5})$ algorithm that's usable and an $O(n^2)$ algorithm which isn't actually usable in practice), but Leslie Valiant showed in 1979 that counting the number of such matchings is difficult, in fact #P-complete (there's no polynomial-time algorithm in practice). Of course, there are some special cases where we can evaluate the permanent, and that can be important.

For another reason for caring about this quantity, we can consider **Mallows ℓ^p measures** on the symmetric group S_n. We often pick permutations at random, and one example of a model of random permutations which are "clustered around some given permutation" assigns a mass

$$\pi_\theta(\sigma) = \frac{1}{Z} \exp\left(\theta \sum_{i=1}^{n} |\sigma(i) - i|^p\right)$$

for some parameters $\theta, p \geq 0$. (So this centers permutations around the identity – for example, this happens when psychologists play seven tones and ask people to rank them from highest to lowest, and the answers given usually cluster around the right answer with some randomness.) Then the normalizing constant for this measure is

$$Z = \sum_{\sigma \in S_n} \exp\left(\theta \sum_{i=1}^{n} |\sigma(i) - i|^p\right) = \sum_{\sigma \in S_n} \prod_{i=1}^{n} A_{i,\sigma(i)}$$

for $A_{ij} = e^{\theta |i-j|^p}$, and indeed this seems to be a difficult quantity to compute. And there are many other motivations too which we won't talk about here.

Example 27

In celebrated (Turing award-winning) work, Broder, Jerrum, Valiant, and Vazirani introduced a Markov chain Monte Carlo algorithm for approximating permanents. We'll do the special case where we have a bipartite graph on $n + n$ vertices and know that there is a perfect matching; our goal is to count the total number of them. This is a special case of the following more general problem: we have a finite set $\mathfrak{X}$ and we can sample from it uniformly (that's what the algorithm will allow us to do, to good approximation). We then want to use the sample to estimate the total size of the set.

In words, suppose we have a big set and someone lets us pick 1000 things from it. If we don't know anything else about the set, all we can really do is wait for repeats and then use the birthday problem to estimate (since it takes

$1.2\sqrt{n}$ samples to have a reasonable chance of a repeat, and thus we can invert this to make an estimate). But if the set had structure, we could do much better; for example if we knew the set had numbers $\{1, 2, \cdots, n\}$ we could take the largest thing in our sample and stretch it out a bit, and with any unbounded sample size this gets us a good answer.

The idea for how to do the actual estimation in general is the following: suppose we have a nested decreasing sequence of sets $\mathfrak{X} = X_0 \supset X_1 \supset \cdots \supset X_K =$ singleton, where the ratio $\frac{|X_{i+1}|}{|X_i|}$ is "not too small" and K is not too large. Then do the following: sample at random from X_0 and see what proportion land in X_1, then sample at random from X_1 and see what proportion land in X_2, and so on. This gives us estimates for $\frac{|X_i|}{|X_{i+1}|}$ for all i, and then we can multiply them together to get an estimate for $|\mathfrak{X}|$ by telescoping. Large deviations for the binomial distribution say that we get exponentially close for each term. Thus the overall estimate is not so bad and doesn't take too long (polynomial time), as long as the adjacent ratios are in control, there aren't so many of them to telescope, and we have an efficient sampling algorithm for each stage.

Jerrum and Sinclair did this for the special case of the permanent. First, we need to generate a random perfect matching. What we do is we take an existing perfect matching, look at two pairs, and then swap the matching of those pairs. To prove something about this, we have to **enlarge the state space**, and that's an important idea to consider. We work with

$$\mathcal{M}_n(G) \cup \mathcal{M}_{n-1}(G),$$

which is the set of all perfect matchings along with the set of all matchings where we have $(n-1)$ pairs matched. The algorithm is the following, started from some matching or almost-matching x:

1. Pick an edge (u, v) from the graph uniformly.
2. If x is a perfect matching and (u, v) is part of it it, then remove it from the matching.
3. If x is an almost-matching and (u, v) are unmatched, then add it to the matching.
4. On the other hand, if (u, w) are matched and v is unmatched, then replace (u, w) with (u, v). Similarly if (w, v) is matched and u us unmatched, replace (w, v) with (u, v).
5. Otherwise, do nothing.

It can be shown that this Markov chain is connected on this extended state space under some assumptions, and the $\mathcal{M}_n$ part and $\mathcal{M}_{n-1}$ part have comparable sizes. Furthermore, this algorithm is an ergodic symmetric chain. So if we can sample on the whole extended state space, that is good enough for getting us samples from $\mathcal{M}_n(G)$. Jerrum and Sinclair then use paths to say that given two matchings or almost-matchings, we have some way of getting from one to another, and we can get reasonable covering number bounds using some interesting ideas.

Originally this was done using "Cheeger methods," but Professor Diaconis and Stroock took those paths and used Poincaré methods instead and got n^7 instead of n^{14}:

Theorem 28

Let G be a bipartite graph such that each vertex is of degree at least $\frac{n}{2}$. Then $\beta_1 \leq 1 - \frac{1}{6n^7}$.

The same strategy got that quadratic improvement in many other computer science problems as well; to see this all in action, we can see either Mark Jerrum or Alistair Sinclair's books or the original paper.

5 April 13, 2026

Today's lecture is being given by Professor Sourav Chatterjee. We'll discuss one of his recent papers about **mixing in nonreversible chains**, called "Spectral gap of nonreversible Markov chains."

Example 29

Our setting will be the following (we'll use different notation from the rest of the lectures for consistency with the paper). Suppose $X_0, X_1, \cdots$ is a (time-homogeneous) stationary Markov chain on a finite state space which we'll call $\mathcal{S}$, and let P be its transition matrix and $L = I - P$ be the generator. Also suppose μ is an invariant (stationary) probability measure with $\mu(x) > 0$ for all $x \in \mathcal{S}$ (we aren't assuming it's unique, but in all applications it will be). As before, such a measure defines an inner product on the function space $\mathbb{C}^{\mathcal{S}}$ given by $\langle f, g\rangle = \sum_{x\in\mathcal{S}} f(x)\overline{g(x)}\mu(x)$, and it also specifies a norm $||f|| = \sqrt{\langle f, f\rangle}$.

So everything is the same, except we aren't assuming reversibility; in particular P may not be self-adjoint.

The difference now is that understanding L and understanding P will no longer be the same thing, since we define the spectral gap differently from the reversible case:

Definition 30

With the notation above, the **spectral gap** of the chain is the second smallest singular value γ of L with respect to the inner product $\langle\cdot,\cdot\rangle$ above. We then define $\tau = \frac{1}{\gamma}$ to be the **relaxation time** of the chain.

The theory of spectral gap and relaxation time has been attempted for nonreversible chains many times, but somehow this happened to be the right one for various reasons (and hadn't been studied before).

Fact 31

As a reminder, the singular values of L are the square roots of the eigenvalues of L^*L. So in order to define the adjoint L^*, we have to specify an inner product.

For reversible chains the gap is 1 minus the second smallest eigenvalue, so this is exactly coinciding with our definition in the reversbiele case because $L^* = L$ with respect to μ and so the eigenvalues of L^*L are the squares of the eigenvalues of L.

To state the main theorem, we'll set up a bit of notation. For a function $g : \mathcal{S} \to \mathbb{R}$, write $\mu g = \sum_x \mu(x)g(x)$ for the average and $\mu_n g = \frac{1}{n}\sum_{i=1}^N g(X_i)$ for the empirical average. We are curious how long the random quantity $\mu_n g$ converges to its limit μg. Incidentally this is somehow more important than convergence in other notions like mixing time, since this is actually how people actually use Markov chains for empirical estimation. We'll write for any real-valued random variable Z the norm

$$||Z||_{L^2} = \sqrt{\mathbb{E}[Z^2]},$$

and the quantity we will be curious about is

$$\Delta_n = \sup_{g:||g-\mu g||=1} ||\mu_n g - \mu g||_{L^2}$$

(here $g - \mu g$ is deterministic so the norm in the subscript is just coming from the ordinary μ-norm). Note that Δ_n is not monotone in n; it is possible for it to increase as n increases.

Theorem 32

Recall that our Markov chain must be started from stationarity, meaning that $X_0 \sim \mu$. We have the following results:

1. For all $n \geq 1$, we have $\Delta_n \leq \sqrt{\frac{4\tau}{n}}$.
2. Conversely, for all $n \leq \frac{\tau}{3}$, we have $\Delta_n \geq \frac{1}{132}$.
3. Finally, for all $n \geq 1$, we have $\max_{n \leq k \leq 2n} \Delta_k \geq \frac{\tau}{2n + 3\tau}$.

In other words, we need order τ steps to be close to 0, but before time $\frac{\tau}{3}$ we must be big. And while it's possible for a specific value like Δ_τ to be zero, if we look over a big window from (for example) τ to 2τ, there will be some time where Δ_n is big. The weakness in the result though is that it doesn't tell us about fixed initial states, only "μ-averaged" results.

Proof of the upper bound (1). Assume that $\gamma > 0$ to avoid trivialities. This means that 0 is a singular value of L of multiplicity 1, so $\dim(\ker L) = 1$ and thus $\dim(\text{range } L) = |\mathcal{S}|_1$. But for any $f \in \text{range } L$, if we write $f = Lg$, we have

$$\langle \mathbf{1}, f \rangle = \sum_x \mu(x)(g(x) - Pg(x)) = 0$$

since μ is an invariant measure for P. Thus range(L) is contained in the orthogonal subspace to span($\mathbf{1}$) (under the μ-inner product); in particular those must be equal by dimension counts. Furthermore for any function $g : \mathcal{S} \to \mathbb{R}$, $g - \mu g$ is in that orthogonal subspace. Therefore **for any function** g we have a solution of the Poisson equation

$$Lf = g - \mu g,$$

and furthermore we can always find an f of this form which is orthogonal to $\mathbf{1}$ by subtracting off a constant. (We may have seen solutions of this type using an infinite series, but our purpose is to avoid that.) Fix such a pair of functions f, g. Since γ is the second-smallest singular value of L, we must have

$$||Lf|| \geq \gamma ||f|| \quad \text{if } f \in \text{span}(\mathbf{1})^\perp;$$

in particular this means that

$$||g - \mu g|| \geq \gamma ||f|| \implies ||f|| \leq \tau ||g - \mu g||.$$

Lemma 33

For all n, we have the inequality

$$\sum_x \mu(x) \left(\mathbb{E}_x \left[\mu_n g - \mu g\right]\right)^2 \leq \frac{4\tau^2 ||g - \mu g||^2}{n^2}.$$

Proof of lemma. We have

$$\begin{aligned}\mathbb{E}_x[g(X_n) - \mu g] &= \mathbb{E}_x[Lf(X_n)] \\ &= \mathbb{E}_x[f(X_n) - Pf(X_n)] \\ &= \mathbb{E}_x\left[f(X_n)\right] - \mathbb{E}_x\left[\mathbb{E}[f(X_{n+1})|X_n]\right] \\ &= \mathbb{E}_x\left[f(X_n) - f(X_{n+1})\right]\end{aligned}$$

by the tower property. So therefore we can write out the empirical average and get

$$\begin{aligned}\mathbb{E}_x[\mu_n g - \mu g] &= \frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}_x[g(X_k) - \mu g] \\ &= \frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}_x[f(X_k) - f(X_{k+1})] \\ &= \frac{1}{n}(f(x) - \mathbb{E}_x[f(X_n)])\end{aligned}$$

by a telescoping sum. So we can plug this into the parenthetical part of our left-hand side:

$$\begin{aligned}\sum_x \mu(x)\left(\mathbb{E}_x\left[\mu_n g - \mu g\right]\right)^2 &= \frac{1}{n^2}\sum_x \mu(x)\left(\mathbb{E}_x[f(X_n) - f(x)]\right)^2 \\ &\leq \frac{1}{n^2}\sum_x \mu(x)\mathbb{E}_x[(f(X_n) - f(x))^2] \\ &\leq \frac{1}{n^2}\sum_x \mu(x)\left(2\mathbb{E}_x[f(X_n)^2] + 2f(x)^2\right)\end{aligned}$$

by Cauchy-Schwarz (or Jensen's inequality) and then using $(a+b)^2 \leq 2a^2 + 2b^2$. But this simplifies to $\frac{2}{n^2}\mathbb{E}[f(X_n)^2] + \frac{2}{n^2}||f||^2$, which is just $\frac{4}{n^2}||f||^2$ by stationarity. Plugging in our bound $||f|| \leq \tau||g - \mu g||$ yields the result. □

So now returning to the proof, take any $g : \mathcal{S} \to \mathbb{R}$; without loss of generality we can take $\mu g = 0$ and $||g|| = 1$. Then we can write explicitly

$$\begin{aligned}||\mu_n g - \mu g||_{L^2}^2 &= ||\mu_n g||_{L^2}^2 \\ &= \sum_x \mu(x)\mathbb{E}_x\left[\left(\frac{1}{n}\sum_{i=1}^{n-1} g(X_i)\right)^2\right].\end{aligned}$$

Expanding this square and then rewriting in a different order yields

$$\begin{aligned}&\frac{1}{n^2}\sum_x \sum_{i,j=0}^{n-1}\mu(x)\mathbb{E}_x\left[g(X_i)g(X_j)\right] \\ &= \frac{2}{n^2}\sum_x\sum_{i=0}^{n-1}\sum_{j=i}^{n-1}\mu(x)\mathbb{E}_x\left[g(X_i)g(X_j)\right] - \frac{1}{n^2}\sum_x\sum_{i=0}^{n-1}\mathbb{E}_x[g(X_i)^2].\end{aligned}$$

We'll ignore the latter term because it's always subtracting off a nonnegative thing, so we have that

$$||\mu_n g - \mu g||_{L^2}^2 \leq \frac{2}{n^2} \sum_x \sum_{i=0}^{n-1} \sum_{j=i}^{n-1} \mu(x) \mathbb{E}_x \left[g(X_i) g(X_j) \right]$$
$$= \frac{2}{n^2} \sum_x \sum_{i=0}^{n-1} \mu(x) \mathbb{E}_x \left[g(X_i) \sum_{j=i}^{n-1} g(X_j) \right].$$

Now for any fixed i, if we define the function $h(x) = \mathbb{E}[\sum_{j=i}^{n-1} g(X_j) | X_i = x]$, we have by time-homogeneity that this is also $\mathbb{E}[\sum_{j=0}^{n-i-1} g(X_j) | X_0 = x] = (n-i)\mathbb{E}_x[\mu_{n-i} g]$ by definition. So in our expression above, the part for any fixed i simplifies by Cauchy-Schwarz to

$$\frac{2}{n^2} \sum_x \mathbb{E}_x \left[g(X_i) h(X_i) \right] \leq \sqrt{\mathbb{E}_x[g(X_i)^2] \mathbb{E}_x[h(X_i)^2]}$$
$$\leq \frac{2}{n^2} \sqrt{\sum_x \mu(x) \mathbb{E}_x[g(X_i)^2]} \sqrt{\sum_x \mu(x) \mathbb{E}_x[h(X_i)^2]}$$
$$\leq \frac{2}{n^2} ||g|| \, ||h||.$$

By our lemma applied to h, we can simplify

$$||h||^2 = (n-i)^2 \sum_x \mu(x) \mathbb{E}_x[\mu_{n-i} g]^2 \leq 4\tau^2 ||g||^2;$$

in particular the factors of n cancel out. So plugging this back in gives us the upper bound of $\frac{2}{n^2}||g||(2\tau||g||) = \frac{4\tau}{n^2}||g|| = \frac{4\tau}{n^2}$; summing over all n values of i yields the result. □

Notice that the challenge here is that we don't have eigenvalues and eigenvectors but still have to do the bounds overall, so we end up having to work with more global quantities.

Proof of the lower bounds (2) and (3). Consider any function $f \in \text{span}(\mathbf{1})^\perp$ with $||f|| = 1$ which minimizes $||Lf||$. (This will not be exactly what witnesses the anomaly in Δ_n, though.) Let $g = Lf$. Then $\gamma = ||Lf||$, and we can define the three functions

$$u_n(x) = f(x) - \frac{1}{n} \sum_{k=n}^{2n-1} \mathbb{E}_x \left[f(X_k) \right],$$

$$v_n(x) = \frac{1}{n} \sum_{k=n}^{2n-1} \mathbb{E}_x \left[f(X_k) \right],$$

$$w_n(x) = \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{E}_x \left[g(X_k) \right] = \mathbb{E}_x[\mu_n g].$$

What we will show is that non-convergence is witnessed either by f or by g. For any n we have the identity $f = u_n + v_n$, so $1 \leq ||u_n|| + ||v_n||$; in particular at least one of u_n and v_n is large. Recalling that $Lf = g$ and $\mu g = 0$, we use the "telescoping" internal step from the lemma

$$\mathbb{E}_x[\mu_n g] = \frac{1}{n}(f(x) - \mathbb{E}_x[f(X_n)]).$$

But we can rewrite

$$\begin{aligned} u_n(x) &= \frac{1}{n}\sum_{k=n}^{2n-1}(f(x) - \mathbb{E}_x[f(X_k)]) \\ &= \frac{1}{n}\sum_{k=n}^{2n-1} k w_k(x). \end{aligned}$$

Now we have

$$\begin{aligned} ||w_n||^2 &= \sum_x \mu(x) w_n(x)^2 \\ &= \sum_x \mu(x)\mathbb{E}_x[\mu_n g]^2 \\ &\leq \sum_x \mu(x)\mathbb{E}_x[(\mu_n g)^2] \\ &= ||\mu_n g||_{L^2}^2 \\ &\leq ||g||^2 \Delta_n^2 \\ &\leq \gamma^2 \Delta_n^2. \end{aligned}$$

by our definition of relaxation time. Thus by the triangle inequality we have

$$||u_n(x)|| \leq \frac{1}{n}\sum_{k=n}^{2n-1} k||w_k(x)|| \leq \frac{1}{n}\sum_{k=n}^{2n-1} k\gamma\Delta_k \leq 2n\gamma \max_{n\leq k\leq 2n-1}\Delta_k.$$

And we can also similarly get a bound on the norm of v_n: we have

$$\begin{aligned} v_n(x) &= \frac{1}{n}\sum_{k=n}^{2n-1}\mathbb{E}_x[f(X_k)] \\ &= \mathbb{E}_x\left[\frac{1}{n}\sum_{k=0}^{2n-1} f(X_k) - \frac{1}{n}\sum_{k=0}^{n-1} f(X_k)\right] \\ &= \mathbb{E}_x\left[2\mu_{2n}f - \mu_n f\right], \end{aligned}$$

meaning that $||v_n|| \leq 2||\mu_{2n}f|| + ||\mu_n f|| \leq 2\Delta_{2n} + \Delta_n$ because f has norm 1 and mean zero. Putting together our bounds yields

$$1 \leq ||u_n|| + ||v_n|| \leq (2n\gamma + 3)\max_{n\leq k\leq 2n}\Delta_k \implies \max_{n\leq k\leq 2n}\Delta_k \geq \frac{1}{2n\gamma+3}.$$

So in particular if n does not exceed the relaxation time by a lot, Δ_k will be large for something in $[n, 2n]$, proving (3). And for the pointwise bound (2), we use some tricks to refine this last estimate. □

Remark 34. *Consider the chain on a circle of length n which always goes to the right by one step. Then after n steps we know that the empirical average will always exactly converge, even though γ is nonzero, so we can't have hopes of a much better lower bound on Δ_n.*

Example 35

For an explicit computation, let's now think about a Markov chain on $\mathcal{S} = \mathbb{Z}/N\mathbb{Z}$ defined by specifying some given probabilities $p_1, \cdots, p_k$ with $\sum_i p_i = 1$ and points $a_1, \cdots, a_k \in \mathbb{Z}/N\mathbb{Z}$ distinct, and at each step the random walk takes jump a_i with probability p_i. (So for simple symmetric random walk we would have $a_1 = +1$ and $a_2 = -1$ and $p_1 = p_2 = \frac{1}{2}$.

Clearly this is a nonreversible chain unless things are symmetric under negating all a_is. For this chain, we can write down the exact formula for the relaxation time:

Theorem 36

We have

$$\tau = \max_{1 \le j \le N-1} \left| 1 - \sum_{r=1}^{k} p_r e^{2\pi i j a_r / N} \right|^{-1}$$

(where the i in the exponent is the imaginary unit). The idea is that the nonzero singular values are the values for various js.

The stationary distribution of such a chain is always uniform, so our inner product is the ordinary inner product. And since P is a circulant matrix, it is always normal (that is, $PP^* = P^*P$) and we can write $P = UDU^*$ and get a diagonalization) and U is of a very specific form. Then we can read off the singular values by just writing out L in terms of the decomposition.

This may look relatively simple, but it's still quite complicated to understand for individual ps and as. One thing we can do is fix k and the probabilities p_i and pick a_is at random (for example we specify two jumps, but we randomly choose the size of the increments); it's a bit hard to guess what the relaxation time will be:

Theorem 37

Fix $p_1, \cdots, p_k$ and choose $a_1, \cdots, a_k$ uniformly at random. Then $\tau = O_p(N^{2/(k+1)})$; more precisely, for all $L > 0$, we have

$$\mathbb{P}(\tau > L^{2/(k+1)}) \le \frac{C}{L^{(k+1)/2}}$$

for some constant C depending only on our choices of $k, p_1, \cdots, p_k$.

So with two points it'll typically take $N^{2/3}$ time for empirical averages to converge. In contrast, what was previously known about this is that the mixing time is of order $N^{2/(K-1)}$ (with both upper and lower bounds), which is much larger! So empirical averages will converge much faster than the chain itself does, and this happens in many examples. It would be nice to have a general theory of why this happens.

For example, on the circle which either stays where it is or moves to the right with probability $1/2$, after k steps it has moved forward by k steps with fluctuation $\sqrt{k}$. So we need N^2 steps to mix because we're concentrated in a small window, but we only need N steps to be sufficient for empirical averages to be close.

In words, the proof idea is to write out

$$\left| 1 - \sum_{r=1}^{k} p_r e^{2\pi i j a_r / N} \right|^2 = \left(\sum_{r=1}^{k} p_r \left(1 - \cos \frac{2\pi j a_r}{N} \right) \right)^2 + \left(\sum_{r=1}^{k} p_r \sin \left(\frac{2\pi j a_r}{N} \right) \right)^2$$

We can reparametrize with variables X_r so that $j a_r \equiv X_r \bmod N$ with X_r chosen to be between $-\frac{N-1}{2}$ and $\frac{N-1}{2}$; the X_rs are iid because the a_rs are. But now if this right-hand side were to be very small, then both terms must be quite

small; if the first term is small then X_rs are close to zero, and therefore in the second term we can replace $\sin x \approx x$. Thus we get that if the whole expression is at most δ, then for all r $|X_r| \leq C\sqrt{\delta}N$ and $|\sum p_r X_r| \leq C\delta N$; this allows us to get an upper bound for an appropriately chosen δ. Then we just sum over all j and union bound.

6 April 15, 2026

We'll return to reversible chains today, looking more at path technologies. There's a version of this for non-reversible chains too, but we won't touch on that in this lecture.

Example 38

We're returning to our usual setting: let $\mathfrak{X}$ be a finite set and (K, π) be a reversible Markov chain on $\mathfrak{X}$. Last time, we showed that $\beta_1 \leq 1 - \frac{1}{A}$ with $A = \max_e \frac{1}{Q(e)} \sum_{\gamma_{xy} \ni e} |\gamma_{xy}| \pi(x)\pi(y)$ with a specific use of Cauchy-Schwarz. We'll sharpen this in two different ways.

First of all, there are often many possible choices of paths (even if many of them end up giving the same answer, or if the simplest way gives us a pretty good answer). So one thing we might consider is to use random paths, and a second is to use Cauchy-Schwarz with a different weight from just 1 when we add and subtract along the path $f(x) - f(y)$ in the variance. In various examples, it's possible that doing these changes buys us a factor of n (or better), and we can even optimize over these choices.

Theorem 39

Let A^* be the optimized bound of A over all choices of random paths and weights in Cauchy-Schwarz. Then

$$1 - \frac{(\log |\mathfrak{X}|)^2}{A^*} \leq \beta_1 \leq 1 - \frac{1}{A^*}.$$

In particular, we get reasonably sharp bounds on the second eigenvalue if we add in this extra freedom, and let's talk about the details of these techniques now.

First, for random paths, we can define a probability measure $\mu_{xy}(\gamma)$ over the set of all paths γ from x to y (for example, we can have it be a specific deterministic path, or we could choose uniformly at random). Then we get a variant of the bound we had before:

Theorem 40

With random paths, we have $\beta_1 \leq 1 - \frac{1}{B}$, where

$$B = \max_e \frac{1}{Q(e)} \sum_{x,y} \sum_{\text{paths } \gamma_{xy} \ni e} \mu_{xy}(\gamma_{xy}) |\gamma_{xy}| \pi(x)\pi(y)$$

(where in this inner summation we sum over all paths γ_{xy} containing a given edge, not just specific chosen ones).

Proof. Much like before, we know that

$$1 - \beta_1 = \max_f \frac{\mathcal{E}(f, f)}{\text{Var}(f)},$$

and we write

$$\begin{aligned}\mathrm{Var}(f) &= \sum_{x,y}(f(x)-f(y))^2\pi(x)\pi(y)\\ &= \frac{1}{2}\sum_{x,y}\left(\sum_{\gamma_{xy}}\mu_{xy}(\gamma_{xy})\right)(f(x)-f(y))^2\pi(x)\pi(y)\\ &= \frac{1}{2}\sum_{x,y}\sum_{\gamma_{xy}}\mu_{xy}(\gamma_{xy})\left(\sum_{e\in\gamma_{xy}}\left(f(e^+)-f(e^-)\right)\right)^2\pi(x)\pi(y).\end{aligned}$$

Now again using Cauchy-Schwarz with weight 1 simplifies this to

$$\begin{aligned}\mathrm{Var}(f) &\le \frac{1}{2}\sum_{x,y}\sum_{\gamma_{xy}}\mu(\gamma_{xy})|\gamma_{xy}|\sum_{e\in\gamma_{xy}}(f(e^+)-f(e^-))^2\pi(x)\pi(y)\\ &= \frac{1}{2}\sum_{e\in\gamma_{xy}}(f(e^+)-f(e^-))^2\frac{Q(e)}{Q(e)}\sum_{\gamma_{xy}\ni e}\mu(\gamma_{xy})|\gamma_{xy}|\pi(x)\pi(y)\end{aligned}$$

by the same swap of summation, and now the blue part is bounded from above by B so $\mathrm{Var}(f)\le B\mathcal{E}(f,f)$ as desired. □

Here's an example where it really does help to have random paths over deterministic ones:

Example 41

Consider the complete bipartite graph $K_{n,n}$, and write the vertices on one side as i and the vertices on the other side as i' for $1\le i\le n$. Our Markov chain will be simple random walk, so that $K(i,j')=K(j',i)=\frac{1}{n}$ for all i,j'. Of course, this chain is reversible with $\pi(i)=\pi(i')=\frac{1}{2n}$, so $Q(i,j')=\frac{1}{2n^2}$ for all edges.

This walk has a parity problem, so it has -1 as an eigenvalue, but what we care about is bounding the second eigenvalue. First of all, we could consider naive deterministic paths – if we're going from i to j' or vice versa we can just take that edge, and then we have to also worry about paths between vertices on the same side. So we can for example choose γ_{ij} to be the path $i\to i'\to j$, and similarly on the other side. Then

$$\begin{aligned}A &= \max_e\frac{1}{Q(e)}\sum_{\gamma_{xy}\ni e}|\gamma_{xy}|\pi(x)\pi(y)\\ &\le 2n^2\cdot 2\cdot\frac{1}{(2n)^2}\max_e\sum_{\gamma_{xy}\ni e}1.\end{aligned}$$

But now if we look at the edge from i to i', that ends up getting used a total of $2n-1$ times, and thus we get $A\le n\implies\beta_1\le 1-\frac{1}{2n-1}$. But this is a bad bound – we actually know that for this chain, the eigenvalues are 1 (with multiplicity 1), -1 (with multiplicity 1), and 0 (with multiplicity $2n-2$). So we're getting the wrong spectral gap here.

Well, random paths helps us out here to avoid the issue of overusing paths. For $\gamma_{ij'}$ we can still just use the edge from i to j', but for γ_{ij} among one side we can now randomly pick a random path of the form $i\to k'\to j$, where k' is chosen uniformly at random among all n vertices on the other side. Path lengths are again bounded by 2, and this

time we get that

$$B = \max_e \frac{1}{Q(e)} \sum_{\gamma_{xy} \ni e} \mu_{xy}(\gamma_{xy}) |\gamma_{xy}| \pi(x)\pi(y)$$
$$\leq \frac{2 \cdot 2n^2}{(2n)^2} \max_e \sum_{\gamma_{xy} \ni e} \mu_{xy}(\gamma_{xy}),$$

and now for any edge (i, j') we get a contribution of 1 from $\gamma_{ij'}$, as well as a contribution of $\frac{1}{n}$ from each of the $(n-1)$ paths γ_{ij}, each of the $(n-1)$ paths γ_{ji}, and also the same on the other side; the total contribution is bounded by $1 + 4 = 5$. So we get that $\beta_1 \leq 1 - \frac{1}{5}$, which is at least of the right order (and saves the order n factor).

Remark 42. *Of course, we could have also tried some "intelligent strategy" with clever or pseudo-random paths. But in a similar setting, consider random walk on the hypercube but now allow double the number of moves, so now we have $2n$ choices instead of just the usual n. If in addition to the usual basis e_i we choose another random n things, it speeds things up from $n \log n$ to $n^{3/4}$, but no one knows a way of choosing $2n$ things which achieves that. For a reference, we can see a paper by Wilson.*

Fact 43

For an example where random paths made a huge difference for real examples, we can see Ben Morris and Alistair Sinclair's paper "Random Walks on Truncated Cubes and Sampling 0-1 Knapsack Solutions." The idea is that we take our random walk on the hypercube C_2^n, but now we consider some hyperplane which cuts it, and we look at all points in the hypercube above that hyperplane. The resulting graph is still connected and now think about nearest-neighbor random walk on that subset which gives us a uniform stationary distribution (by sampling a neighbor at random and not moving below the hypercube if that's what we get to).

Fact 44

For yet another example, we can see Section 5 of Professor Diaconis' paper with Laurent Saloff-Coste "Nash inequalities for finite Markov chains." There are lots of examples where choosing random paths buys us a factor of n, and group theorists use this particular argument.

For our other strategy, we can think about optimizing via Cauchy-Schwarz weights. Let's do a (somewhat) real example:

Example 45

Consider the Metropolis algorithm on the path $\{0, 1, \cdots, n-1\}$ with stationary distribution $\pi(j) = \frac{a^{h(j)}}{Z}$, where $h(j)$ is some increasing function like i^b and $0 < a < 1$; further assume that $h(i+1) - h(i) \geq c \geq 1$. (The idea is that the stationary distribution has exponential fall-off.)

We haven't described the Metropolis algorithm yet, and we'll do that in more detail later and provide lots of motivation, but for now we'll just write down the chain. The idea is that we can use the Metropolis argument to transform a Markov chain to another one with a different stationary distribution by flipping coins, and in this case we can "Metropolize" starting from the base chain which is nearest-neighbor random walk with $\frac{1}{2}$-holding at the two endpoints (Example 21). Our new chain has, for all $1 \leq i \leq n-2$,

$$M(i,i) = \frac{1}{2} - \frac{a^{h(i+1)-h(i)}}{2}, \quad M(i,i+1) = \frac{a^{h(i+1)-h(i)}}{2}, \quad M(i,i-1) = \frac{1}{2},$$

and for the corners we have

$$M(0,0) = 1 - \frac{a^{h(1)-h(0)}}{2}, \quad M(0,1) = \frac{a^{h(1)-h(0)}}{2}, \quad M(n-1,n-2) = M(n-1,n-1) = \frac{1}{2}.$$

So basically, we propose a step from the original chain, and then if the stationary distribution increases we take the step, and otherwise we flip a coin to see whether we take it or not (appropriately chosen to get us the right stationary distribution). What we'll show is the following:

Proposition 46

In the setting above, we have

$$\beta_1 \leq 1 - \frac{(1-a^{c/2})^2}{2}.$$

In particular, this doesn't depend on n.

Obviously this chain will mix pretty quickly because it has most of its mass around 0, and we generally drift towards the left at each step. But this is a quantitative way of getting that.

Proof. We'll bound the variance in terms of the Dirichlet form in the usual way. First of all, write

$$\begin{aligned} 2\mathrm{Var}(f) &= \sum_{x,y} (f(x) - f(y))^2 \pi(x)\pi(y) \\ &= \sum_{x,y} \left(\sum_{e \in \gamma_{xy}} \left(f(e^+) - f(e^-)\right) \right)^2 \pi(x)\pi(y) \end{aligned}$$

(there's only one naive way to choose paths in this case, so we don't need to worry about weights). Normally what we do here is do Cauchy-Schwarz on the inside sum, but this time we can write

$$2\mathrm{Var}(f) = \sum_{x,y} \left(\sum_{e \in \gamma_{xy}} \left(f(e^+) - f(e^-)\right) \frac{Q(e)^\theta}{Q(e)^\theta} \right)^2 \pi(x)\pi(y)$$

where it turns out we'll want to take $\theta = \frac{1}{4}$, but we don't know that yet (we'll see soon why we should take $\theta \in (0, \frac{1}{2})$, though). Define the modified path length $|\gamma_{xy}|_\theta = \sum_{e \in \gamma_{xy}} \frac{1}{Q(e)^{2\theta}}$. Applying Cauchy-Schwarz now gives us

$$\begin{aligned} 2\mathrm{Var}(f) &\leq \sum_{x,y} |\gamma_{xy}|_\theta \left(\sum_{e \in \gamma_{xy}} Q(e)^{2\theta} (f(e^+) - f(e^-))^2 \right) \pi(x)\pi(y) \\ &= \sum_e (f(e^+) - f(e^-)) Q(e) Q(e)^{2\theta-1} \sum_{\gamma_{xy} \ni e} |\gamma_{xy}|_\theta \pi(x)\pi(y) \end{aligned}$$

by swapping order of summation, and therefore we have

$$2\mathrm{Var}(f) \leq 2A\mathcal{E}(f,f) \quad \text{for } A = \max_e Q(e)^{2\theta-1} \sum_{\gamma_{xy} \ni e} |\gamma_{xy}|_\theta \pi(x)\pi(y).$$

We now have to bound this quantity A. The good news is that explicit calculation shows for edges in the middle that $Q(i,i+1) = Q(i,i-1) = \frac{\pi(i+1)}{2}$, and for any $x < y$ the dominant term in $|\gamma_{xy}|_\theta$ is the rightmost one because $\pi(i)$ decreases as i increases, and the sum decays geometrically by assumption of the falloff of π (going to the right by 1

decreases π by a factor of a^c. Bounding by the infinite geometric series then yields

$$|\gamma_{xy}|_\theta \le \frac{\pi(y)^{-2\theta}}{1-a^{2c\theta}}.$$

So plugging into A, we see that for any edge $(i, i+1)$ we need to bound the quantity

$$2^{2\theta}(1-a^{2c\theta})^{-1}Q(e)^{2\theta-1}\sum_{\substack{0\le j\le i\\ i+1\le k\le n-1}}\pi(j)\pi(k)^{1-2\theta}$$

(we're summing over all paths that start to the left of i and end at the right of $i+1$). If we now sum over k we get another geometric series dominated by the term with $k = i+1$, so that can be replaced with $\frac{(\pi(i+1))^{1-2\theta}}{1-a^{c(1-2\theta)}}$; similarly we can just bound the sum over j by 1 since it's a probability measure. Putting things together, our final expression that we want to bound for this edge is

$$A \le \frac{2}{(1-a^{c\cdot 2\theta})(1-a^{c\cdot(1-2\theta)})},$$

and choosing $\theta = \frac{1}{4}$ gets us the best bound because it makes the two exponents equal, yielding the result. □

Everything we've just said can be referenced from the paper "What do we know about the Metropolis algorithm?" by Professor Diaconis and Laurent Saloff-Coste. At the millenium (in 2000), there was a list of "ten great algorithms in scientific computing" which were used thousands or millions of times every day (like the singular value decomposition and quicksort), and the first one on the list was Metropolis. Professor Diaconis noticed that no one really ever proved anything about it, so this paper gave some cases where something could be done. This argument works for grids in low dimensions with probability measures with some kind of unique peak and exponential falloff in the same way; we don't strictly need the assumption that $h(i+1) - h(i) \ge c$ either (polynomial falloff instead of exponential still gives very good bounds). But for any real application of the Metropolis algorithm, things are really still open.

In the case of ordinary exponential falloff where $h(i) = i$ on the path, we know all the eigenvalues and things are nice enough that we can write them down: in fact we have

$$\beta_1 = \frac{(1-a)}{2} + \sqrt{a}.$$

The argument with weighted Cauchy-Schwarz gets $\beta_1 = 1 - \frac{(1-a^{1/2})^2}{2}$, and this is actually exactly the right answer! Of course, having bounds on all of the eigenvalues would get us a sharper answer and so on.

Fact 47

We might ask if we can be more scientific about how to choose these weights, and there's a very nice paper by Nabil Kahale "A semidefinite bound for Mixing Rates of Markov Chains." What the paper manages to do is optimize over weights in Cauchy-Schwarz and measures on paths μ_{xy}, and it shows that the best answer we get is the bound from before. And furthermore, there's a useful way of actually using it to find the answer.

The idea is to let $\ell(e)$ be any arbitrary length function on edges e; this function (putting $\sqrt{\ell(e)}$ everywhere) gives us, for any fixed paths,

$$\beta_1 \le 1 - \frac{1}{A_\ell}, \quad A_\ell = \max_e \frac{1}{\ell(e)Q(e)}\sum_{\gamma_{xy}\ni e}|\gamma_{xy}|_e\pi(x)\pi(y)$$

where $|\gamma_{xy}|_\ell$ is the sum of the $\ell(e)$s along the path. We now want to optimize this over ℓ, and the idea is that the function which takes in a vector of path lengths $\ell(e)$ (indexed by edges) and outputs a vector of values $\frac{1}{Q(e)}\sum_{\gamma_{xy}\ni e}|\gamma_{xy}|_\ell\pi(x)\pi(y)$ (also indexed by edges) is linear; call the matrix M.

Theorem 48

With these choices above, the matrix M has nonnegative entries (so Perron-Frobenius applies), and it has all real eigenvalues and eigenvectors. Letting μ_1 be the largest eigenvalue of M, we have $\mu_1 \leq \min_\ell A_\ell$ and equality is achieved if ℓ is the Perron eigenvector (that is, the eigenvector with all nonnegative entries). In particular, this means that $\beta_1 \leq 1 - \frac{1}{\mu_1}$.

What's important is that this is practical for actual usage: we can start with any initial length vector $\ell(e)$ (like a constant). Then we can repeatedly iterate $\ell' = M\ell$, and we will always have $A_{\ell'} \leq A_\ell$. Thus we get better bounds and we'll converge to the Perron eigenvector and the optimal value of μ_1. In Kahale's paper, using the bounds from Professor Diaconis' paper with Stroock (where they weren't tight) and then iterating with one iteration of M already gets the right answer up to constants.

The part where we optimize similarly over measures on paths involves some result from multi-commodity flows, which is not so bad either but it's less standard than this linear algebra. But again, it's good to look at the paper and read it ourselves for the details. Next time, we'll move to **comparison theory**, which will give us a way of getting bounds on all eigenvalues (not just the first one).

7 April 20, 2026

We're starting a new topic today (in a certain sense), and it's "probably the most useful thing Professor Diaconis has discovered" in Markov chain mixing. The topic is **comparison theory**, and it roughly says that if we have a Markov chain we're interested in, and we can make another Markov chain where we know everything (eigenvalues, geometry, mixing time), then we can prove theorems about the difficult chain, and sometimes we get more or less the right answer. It's easiest to explain for random walks on groups, and many of the useful applications come from that setting, so this is what we'll discuss.

Example 49

Let G be a finite group, and let $\mu(s)$ be a probability measure on G. Consider the random walk on G, started at (for example) $x_0 = \text{id}$ and defined by

$$X_1 = s_1, \quad X_2 = s_2 s_1, \quad \cdots, \quad X_n = s_n X_{n-1}$$

for s_i iid from μ. (In words, we pick an element and multiply on the left.)

All of these walks will have the uniform distribution $U(s) = \frac{1}{|G|}$ as the stationary measure. There are of course many examples of this, but here are three that are good to keep in mind:

- For $G = C_n = \{0, \cdots, n-1\}$ the group of n points on a circle, we can define $\mu(1) = \mu(-1) = \frac{1}{2}$ and get simple random walk on the circle. Taking n odd, it's obvious that after running the walk a long time, we'll get to a uniformly random element.
- For $G = C_2^n$ the usual hypercube in n dimensions, we've already studied nearest-neighbor random walk previously when lifting the Ehrenfest urn, and the corresponding measure is $\mu(0) = \mu(e_i) = \frac{1}{n+1}$ for all $1 \leq i \leq n$.
- For $G = S_n$, let μ be essentially uniform on the set of all transpositions with some holding:

$$\mu(\text{id}) = \frac{1}{n}, \quad \mu(i,j) = \frac{2}{n^2} \text{ for } i \neq j.$$

In words, we have n cards on a table, and we pick two random cards at a time and swap them.

There are many stories and open problems about even these simple examples – there's many more things to say, but we'll do so after developing some more theory. Our goal is to study convergence to the uniform distribution U, and we can do so by thinking about the repeated convolution of our measure μ (the distribution after two or more steps)

$$\mu * \mu(s) = \sum_t \mu(t)\mu(st^{-1}), \quad \mu^{*k}(s) = \sum_t \mu(t)\mu^{*(k-1)}(st^{-1}).$$

Under mild conditions – specifically, the support of μ should not be contained in a coset of a subgroup (intuitively μ doesn't only live on the even number or the odd numbers) – we will indeed have $\mu^{*k}(s) \to \frac{1}{|G|}$ as $k \to \infty$. So our job will be to study how large k must be so that $||\mu^{*k}(s) - U||_{\text{TV}} < \varepsilon$.

For random walk on C_n, it turns out that order n^2 steps are necessary and sufficient, with no cutoff. In contrast, the walk on the hypercube C_2^n takes $\frac{1}{4}(n \log n + c)$ steps, and the random transpositions walk S_n takes $\frac{1}{2}n(\log n + c)$ steps, both with cutoff. There are hundreds of people who continue to work on various chains of this sort, but we'll try not to dive into all of the details of that.

Comparison theory lives within the L^2 theory, so that's what we will study. We'll assume that $\mu(s) = \mu(s^{-1})$ for any $s \in G$ (which is true in the examples above and true a lot of the time), so that the Markov chain $K(s,t) = \mu(st^{-1})$ is reversible (it's symmetric with uniform stationary distribution). Thus we can consider the function space

$$L^2(G) = \{f : G \to \mathbb{R}\}, \quad \langle f, g\rangle = \frac{1}{|G|}\sum_{s\in G} f(s)g(s),$$

and our goal will be to bound eigenvalues with the minimax principle as usual. The two forms $\mathcal{E}$ and $\mathcal{F}$ now take the form

$$\begin{aligned}\mathcal{E}_\mu(f,f) &= \langle (I-K)f, f\rangle \\ &= \frac{1}{2|G|}\sum_{s,t\in G}(f(s)-f(st))^2\mu(t)\end{aligned}$$

and

$$\begin{aligned}\mathcal{F}_\mu(f,f) &= \langle (I+K)f, f\rangle \\ &= \frac{1}{2|G|}\sum_{s,t\in G}(f(s)+f(st))^2\mu(t).\end{aligned}$$

We're going to make use of all of the eigenvalues this time, not just the second one, and here's the form of the minimax characterization we'll use:

Theorem 50 (Minimax characterization)

Let V be a real inner product space, and let $S : V \to V$ be a symmetric operator. For W a subspace of V, define

$$m(W) = \min_{\substack{f \text{ nonzero} \\ f\in W}} \frac{\langle Sf, f\rangle}{\langle f, f\rangle}, \quad M(W) = \max_{\substack{f \text{ nonzero} \\ f\in W}} \frac{\langle Sf, f\rangle}{\langle f, f\rangle}.$$

Let S have eigenvalues $q_0 \le q_1 \le \cdots$. Then

$$q_i = \max_{\substack{W \text{ subspace} \\ \dim(W^\perp)=i}} m(W) = \min_{\substack{W \text{ subspace} \\ \dim(W)=i+1}} M(W).$$

As usual, Horn and Johnson's book (page 176) is a good reference if we haven't seen this before or want the proof

in detail. And now we're ready to describe our setting – the following theorem comes directly from that minimax characterization:

Theorem 51

Let $\mu, \tilde{\mu}$ be two symmetric probability measures on G, where we are interested in understanding μ and we know everything about $\tilde{\mu}$. Let $\mathcal{E}$ and $\tilde{\mathcal{E}}$ be the corresponding quadratic forms, and let β_i and $\tilde{\beta}_i$ be the corresponding ordered eigenvalues.

1. Suppose there exists some $A > 0$ such that $\tilde{E}(f,f) \leq A\mathcal{E}(f,f)$ for all f. Then for all i, we have $\beta_i \leq 1 - \frac{1-\tilde{\beta}_i}{A}$.
2. Similarly, if $\tilde{F}(f,f) \leq B\mathcal{F}(f,f)$, then $\beta_i \geq -1 + \frac{1+\tilde{\beta}_i}{B}$.

To really make use of this, we can use the following:

Theorem 52

With notation as above, suppose that $\tilde{\mathcal{E}}(f,f) \leq A\mathcal{E}(f,f)$. Then the chi-square distance to stationarity satisfies

$$|G| \cdot ||\mu^{*n} - U||_2^2 \leq \beta_{|G|-1}^{2n} + e^{-n/A} + |G| \cdot \left|\left|\tilde{\mu}^{*\lfloor \frac{n}{2A} \rfloor} - U\right|\right|_2^2.$$

Similarly, suppose we have both $\tilde{\mathcal{E}}(f,f) \leq A\mathcal{E}(f,f)$ and $\tilde{\mathcal{F}}(f,f) \leq A\mathcal{F}(f,f)$. Then we can remove a term from above:

$$|G| \cdot ||\mu^{*n} - U||_2^2 \leq e^{-n/A} + |G| \cdot \left|\left|\tilde{\mu}^{*\lfloor \frac{n}{2A} \rfloor} - U\right|\right|_2^2.$$

This is one of those results where we "have to see examples," but the point is that if it takes k steps to make the μ^* chain close to uniform, then $2Ak$ steps makes the last term on the right-hand side small (and A can depend on n).

Remark 53. *Comparison theory can be used even for rather large perturbations to our original chain. For example, a small generating set of S_n would be a single transposition and an n-cycle (this is basically bubble sort), which is completely different from random transpositions. But comparison theory actually gets the right answer. And it even works on very different state spaces as well with some extra machinery.*

Let's first prove this, especially since the proof itself doesn't depend so much on groups:

Proof. Everything follows from using the appropriate bounds $1 - x \leq e^{-x}$ for all x and $1 - x \geq e^{-2x}$ for $0 < x < \frac{1}{2}$. (If we aren't used to having these kinds of inequalities on hand, we should take a look at Laszlo Kozma's "inequalities cheat sheet.") Write down the left-hand side as

$$\begin{aligned} |G| \cdot ||\mu^{*n} - U||_2^2 &= \sum_s \left(\mu^{*n}(s) - \frac{1}{|G|}\right)^2 \\ &= -\frac{1}{|G|} + \sum_s (\mu^{*n}(s))^2 \\ &= \mu^{*(2n)}(\mathrm{id}) - \frac{1}{|G|} \\ &= \frac{1}{|G|} \sum_{i=1}^{|G|-1} \beta_i^{2n}. \end{aligned}$$

Here the second line follows by expanding out the square and using that $\sum_s \mu^{*n}(s) = 1$, and then the third line follows by noticing that after $2n$ steps we're back at the identity if we multiply by s after n steps and then by s^{-1}

after n steps – this uses symmetry of the chain. And the last line uses that $\mu^{*(2n)}(\mathrm{id}) = K^{2n}(s,s)$ for any s, and $\sum_{i=0}^{|G|-1} \beta_i^{2n} = \mathrm{tr}(K^{2n})$ and all diagonal terms are equal, so we divide through by a factor of $|G|$. And now we can use our inequalities: observe that

$$\frac{1}{|G|}\sum_{i=1}^{|G|-1}\beta_i^{2n} \leq \beta_{|G|-1}^{2n} + \frac{1}{|G|}\sum_{i:\beta_i>0}\beta_i^{2n},$$

since all of the negative eigenvalues can be bounded below by the smallest one. Then we know by assumption and the useful inequality that

$$0 \leq \beta_i \leq 1 - \frac{1-\tilde{\beta}_i}{A} \leq e^{-(1-\tilde{\beta}_i)/A},$$

so plugging back in yields

$$\begin{aligned}\frac{1}{|G|}\sum_{i=1}^{|G|-1}\beta_i^{2n} &\leq \beta_{|G|-1}^{2n} + \frac{1}{|G|}\sum_{i:\tilde{\beta}_i>0} e^{-2n(1-\tilde{\beta}_i)/A} \\ &\leq \beta_{|G|-1}^{2n} + e^{-n/A} + |G|\cdot||\tilde{\mu}^{*\lfloor\frac{n}{2A}\rfloor} - U||_2^2,\end{aligned}$$

where we still have to justify the last inequality. We're using that $1-x \geq e^{-2x}$ for $0 \leq x \leq \frac{1}{2}$ by breaking up the positive $\tilde{\beta}_i$s into two sets, the ones above and below $\frac{1}{2}$:

- When $0 \leq 1-\tilde{\beta}_i \leq \frac{1}{2}$ we have that $e^{-2(1-\tilde{\beta}_i)} \leq \tilde{\beta}_i$, so raising to the appropriate power yields $e^{-2n/(1-\tilde{\beta}_i)/A} \leq \tilde{\beta}_i^{2\lfloor\frac{n}{2A}\rfloor}$. This is some part of the sum of the chi-square distance for $\tilde{\mu}$.
- For the other set where $1-\tilde{\beta}_i > \frac{1}{2}$, we have $-2(1-\tilde{\beta}_i) < -1$ so $e^{-2n(1-\tilde{\beta}_i)/A} \leq e^{-n/A}$. So all of the contribution from these added up (and divided by $|G|$) add up to at most $e^{-n/A}$.

The other inequality is proved similarly; we can see Professor Diaconis' paper with Laurent Saloff-Coste "Comparison Techniques for Random Walk on Finite Groups" for the details. □

A real question at this point is "how do we compare forms?," and the answer is a slightly fancy version of the path arguments we've previously developed. Before that, let's look back at our three examples from before and state some questions:

Example 54

For the simple random walk on C_n for n odd (or alternatively we could take $\mu(-1) = \mu(0) = \mu(1) = \frac{1}{3}$ and not worry about parity), we know (by elementary Fourier analysis on the circle) that n^2 steps is necessary and sufficient.

Even for this, there are some open questions and interesting comments. First of all, if $\mu(a) = \mu(b) = \frac{1}{2}$ for some other a, b, we still get the same answer, so changing the support to two points doesn't matter. But now (suppose n is prime to avoid issues) allow three points with $\mu(a) = \mu(b) = \mu(c) = \frac{1}{3}$, and things change:

Theorem 55

In the setting above, for almost all triples (a,b,c), order n steps are necessary and sufficient.

It was open for a few years to find any such actual example, and now we know (thanks to number theory). And more generally, if μ is supported on k points, for almost all sets of points it is necessary and sufficient to take order $n^{2/(k-1)}$ steps (for k bounded and n large).

For fixed k, none of these walks have cutoff, and so it was a question of whether there was any probability distribution on C_n (with support growing) where cutoff does occur. It turns out that (due to Robert Hough) if we take $\mu(\pm 1) = \mu(\pm 2) = \cdots = \mu(\pm 2^k) = \frac{1}{2k}$, then it's like "sticking the hypercube in the cycle" and we do in fact get cutoff. But for most sets of $\log k$ points, probably cutoff still does not occur, though that is still open.

Related to this, a very nice problem to consider is related to the Peres conjecture:

Conjecture 56 (Peres "conjecture")
Call the **mixing time** of a chain the smallest time k such that $||\mu^{*k} - U||_{\text{TV}} \leq \frac{1}{e}$, and define the spectral gap in the usual way. Then a Markov chain has cutoff if and only if the product of the gap and mixing time diverges to ∞.

This turns out to be true in basically all practical cases, but annoyingly not always true (there are some constructed counterexamples). It's open to resolve whether this is true even for the cycle, and if we want to look more into this, we can (wait for the last third of the course or) look at Justin Salez's recent Saint-Flour lecture notes. Most of what is discussed generalizes to abelian groups and "low-class nilpotent groups," and in particular the Heisenberg group of matrices $\left\{ \begin{bmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{bmatrix} : x, y, z \in C_n \right\}$ with walk "take a row at random and add or subtract it from the row above it."

Example 57
Jumping now to random transpositions, this will be the Markov chain we use to do lots of comparison theory next time, and we'll walk through what's known about it now.

The analysis of this Markov chain was done by Professor Diaconis and Mehrdad Shahshahani in 1981, using Fourier analysis on S_n. In short, this makes use of representation theory: a representation of S_n is a map $\rho : S_n \to \text{GL}_d(V)$ taking group elements to matrices, such that $\rho(st) = \rho(s)\rho(t)$. We then have a Fourier transform

$$\hat{\mu}(\rho) = \sum_s \rho(s)\mu(s)$$

which takes convolutions into products, meaning that $\widehat{\mu * \mu}(\rho) = \hat{\mu}(\rho)\hat{\mu}(\rho)$ and thus we just have to study the eigenvalues of the matrices $\hat{\mu}$. These matrices can be quite big – the biggest ones for S_n are of size $\sqrt{n!}$ – but Schur's lemma says that if the representation is irreducible and the measure μ is constant on conjugacy classes (meaning that $\mu(s) = \mu(t^{-1}st)$, then the Fourier transform $\hat{\mu}(\rho)$ is just some multiple of the identity $c_\mu I$, where

$$c_\mu = \frac{1}{d_\rho} \sum_s \chi_\rho(s)\mu(s)$$

where $\chi_\rho(s) = \text{tr}(\rho(s))$ is the associated character. So we just have to raise numbers to high powers instead of matrices.

Well, our random-transposition measure is constant on conjugacy classes, and Frobenius describes the characters for some representations: in this case we have for any partition $\mu = (\lambda_1, \lambda_2, \cdots)$ (partitions index the irreducible representations of S_n) that

$$c_\mu = \frac{1}{n(n-1)} \sum_{i=1}^{\ell} \lambda_i(\lambda_i - 1) - i(i-1).$$

So we "know all the eigenvalues" of our chain here! The paper in 1981 basically carries this out to get appropriate

bounds, and it does enough to show that the total variation distance satisfies

$$\left|\left|\mu^{*k} - U\right|\right|_{\text{TV}} \leq 2e^{-c} \text{ for } k = \frac{1}{2}n(\log n + c),$$

as well as a matching lower bound to get cutoff. So this is the measure that we know everything about, and the content of this part of the subject now is that this lets us do just about any measure on the symmetric group even if it's wildly different from transpositions!

Remark 58. *There are lots of finite groups, and we can try to do this story on any of them. Professor Diaconis "tricks the group theorists" into doing the work of getting analogous bounds for other groups, and thus now we just need to take a walk we're interested in and figure out how to do the combinatorics of comparison theory (which we'll do next time). So there's a lot of opportunities for applying this!*

Fact 59

One wonderful problem is the following: we have sharp results for random transvections on $\text{SL}_n(\mathbb{F}_q)$, in which we repeatedly multiply the matrix by a reflection of the form $I - 2uu^T$ (so we reflect about the orthogonal hyperplane to u). These are generators, and if we pick u at random and use that to get a random walk, we get a chain that ends up being studied for cryptographic purposes. This is a constant-on-conjugacy-classes walk, and Hildebrand showed that n steps are necessary and sufficient (and even that there is a cutoff). But this is really the only other example where results are known, so we should be encouraged to try the same on other groups.

8 April 22, 2026

This lecture is the "central part of the course" – we'll continue studying comparison theory via some illustrative techniques and examples. We mentioned last time that if we have bounds on the Dirichlet form $\tilde{\mathcal{E}}(f,f) \leq A\mathcal{E}(f,f)$, then we get associated bounds on the eigenvalues $\beta_i(\mu) \leq 1 - \frac{1-\tilde{\beta}_i(\tilde{\mu})}{A}$, and then this also gets us bounds on the chi-square distance (Theorem 52).

But the question we now have to answer is how we actually prove comparison of forms (and whether it ends up being useful in practice). That's what we'll answer today.

Definition 60

Let $S \subseteq G$ be a symmetric set of generators of the group (meaning that if $s \in G$, then $s^{-1} \in G$ as well, and any element of G can be written as a product of elements in S). Then for each $y \in G$ we can write $y = z_1 \cdots z_k$ for $z_i \in S$, and we let the **length** of y, denoted $|y|$ be the smallest k over all ways of expressing y this way. We will then let $N(z,y)$ be the number of times z appears in that expression for y.

Theorem 61

Let $\mu, \tilde{\mu}$ be symmetric measures on G, and suppose the support of μ contains a symmetric generating set S. Then we have $\tilde{\mathcal{E}}(f,f) \leq A\mathcal{E}(f,f)$ for

$$A \leq \max_{z \in S} \frac{1}{\mu(z)} \sum_{y \in G} |y| N(z,y) \tilde{\mu}(y).$$

The proof is really the same argument as what we've been doing:

Proof. We'll use the expressions for y in terms of S-elements as our paths. First write, for any $x, y \in G$,

$$f(x) - f(xy) = (f(x) - f(xz_1)) + (f(xz_1) - f(xz_1z_2)) + \cdots + (f(xz_1 \cdots z_{k-1}) - f(xy)),$$

and now do the same trick as usual of squaring both sides and using Cauchy-Schwarz to get

$$(f(x) - f(xy))^2 \leq |y|\Big((f(x) - f(xz_1))^2 + \cdots + (f(xz_1 \cdots z_{k-1}) - f(xy))^2\Big).$$

Now summing both sides in x yields

$$\sum_x (f(x) - f(xy))^2 \leq |y| \sum_{z \in S} \sum_{x \in G} (f(x) - f(xz))^2 N(z, y)$$

because each term on the right-hand side is some x and we can group by how many times each z element is used. If we now multiply both sides by $\frac{\tilde{\mu}(y)}{2|G|}$ and sum over y, we get that

$$\begin{aligned}\mathcal{E}_{\tilde{\mu}}(f, f) = \frac{1}{2|G|} \sum_{x,y \in G} (f(x) - f(xy))^2 \tilde{\mu}(y) &\leq \frac{1}{2|G|} \sum_{x,y \in G} \sum_{z \in S} (f(x) - f(xz))^2 |y| N(z, y) \tilde{\mu}(y) \\ &= \frac{1}{2|G|} \sum_{x \in G} \sum_{z \in S} (f(x) - f(xz))^2 \frac{\mu(z)}{\mu(z)} \sum_{y \in G} |y| N(z, y) \tilde{\mu}(y),\end{aligned}$$

so we can now bound the right-hand side by $A\mathcal{E}_\mu(f, f)$ because the blue part is bounded by A, and then we can sum over all $z \in G$ instead of all $z \in S$ because all terms are nonnegative. □

Example 62

As a basic starting case, let μ be any symmetric measure on any group G, and let $S \subseteq \text{supp}(\mu)$. Take $\tilde{\mu}(s) = \frac{1}{|G|}$ to be uniform, and observe that $\tilde{\mathcal{E}}(f, f) = \frac{1}{|G|^2} \frac{1}{2} \sum_{x,y} (f(x) - f(xy))^2$ is just the variance of f.

So comparing μ with this uniform measure yields the following result:

Proposition 63

Let $\eta = \min_{s \in S} \mu(s)$, and let γ be the diameter of G in the generators S. Then $\beta_1 \leq 1 - \frac{\eta}{\gamma^2}$.

Proof. By applying our previous result, we know that $\text{Var}(f) \leq A\mathcal{E}(f, f)$, where

$$\begin{aligned}A &= \max_{z \in S} \frac{1}{\mu(z)} \sum_{y \in G} |y| N(z, y) \tilde{\mu}(y) \\ &\leq \frac{1}{\eta} \sum_{y \in G} \gamma^2 \tilde{\mu}(y) \\ &= \frac{\gamma^2}{\eta},\end{aligned}$$

and so by our variational characterization (from when we first talked about Dirichlet forms) we must have $\beta_1 \leq 1 - \frac{1}{A} = 1 - \frac{\eta}{\gamma^2}$. (Another way of thinking about this is that for the uniform distribution, we have an eigenvalue of 1 and all other eigenvalues 0, so the gap for $\tilde{\mu}$ is 1.) □

This works for any group and any measure, and so we might ask if it's any good. The following example is rather instructive and also gives an interesting result:

Example 64

Let $G = S_n$, and let our generating set S be $\{\mathrm{id}, (1,2), (1,2,\cdots,n), (n,n-1,\cdots,1)\}$ in cycle notation (so in a deck of cards, this contains the identity, flipping the top two cards, bottom-to-top, and top-to-bottom). Let $\mu(y)$ be $\frac{1}{4}$ for any $y \in S$ and 0 otherwise. We wish to understand the resulting "shuffling mechanism" under μ.

But applying Proposition 63 is quite easy in this case: we clearly have $\eta = \frac{1}{4}$, and we claim that $\gamma = 3\binom{n}{2} \leq \frac{3n^2}{2}$. Indeed, write $t = (1,2)$ and $c = (1,\cdots,n)$. Now for any $y \in S_n$ we can write it in terms of t and c by "working from bottom to top:"

- First, bring card n to the bottom by repeated applications of c. This requires at most n moves.
- Now, suppose the bottom cards $i+1$ through n have been moved to the bottom. Now we have to bring card i to be adjacent to $i+1$, and we do so by bringing i to the top using c, then transposing the top two with t, then putting the new top card on the bottom with c, then transposing again with t, until i is next to $i+1$. Then finally apply c or c^{-1} repeatedly until cards i through n are at the bottom.

We can of course do a little better and improve the constant, but the overall idea is that it takes order n^2 steps and this is basically bubble sort. So plugging in our values of η and γ tells us that $\beta_1 \leq 1 - \frac{1}{9n^4}$.

In many problems like this, we would be thrilled to have such a bound, but we can actually improve this to $\beta_1 \leq 1 - \frac{C}{n^3}$ by comparison with random transpositions instead of the uniform distribution:

Example 65

Recall from last lecture that if we take $\tilde{\mu}(\mathrm{id}) = \frac{1}{n}$ and $\tilde{\mu}((i,j)) = \frac{2}{n^2}$ (and $\tilde{\mu}$ zero otherwise), we know the eigenvalues of this (random-transpositions) chain and in fact $\tilde{\beta}_1 = 1 - \frac{2}{n}$ from character theory. We can perform comparison of μ with this, using the same four-element set S as before.

The result will then be that $\tilde{\mathcal{E}}(f,f) \leq A\mathcal{E}(f,f)$, where

$$\begin{aligned} A &= \max_{z \in S} \frac{1}{\mu(z)} \sum_y |y| N(z,y) \tilde{\mu}(y) \\ &\leq 4(3n)^2, \end{aligned}$$

since the only ys that appear in the sum are transpositions (each with $\tilde{\mu}(y) = \frac{2}{n^2}$, and we can write (i,j) in terms of c and t by moving card i to the top, then transpose-cut-transpose-cut until i and j are adjacent, then do all of those steps in reverse so that j and i's roles are now swapped. So Theorem 51 tells us that

$$\beta_1 \leq 1 - \frac{2/n}{36n^2} = 1 - \frac{1}{18n^3},$$

which is indeed the factor-of-n saving that we promised. And that's the right answer up to a constant, so in fact comparison theory "really can be useful."

On the other side of the spectrum, we also know that $\beta_{|G|-1}(\mu) \geq -1 + 2 \cdot \frac{1}{4} = -\frac{1}{2}$, since our μ-chain has $\frac{1}{4}$-holding. So in this chain the second absolute eigenvalue is $\beta_*(\mu) \leq 1 - \frac{1}{18n^3}$, and so the "easy bound" using chi-square distance says that

$$\begin{aligned} 4||\mu^{*k} - \mu||_{\mathrm{TV}}^2 &\leq \frac{1}{\pi_{\min}} \beta_*^{2k} \\ &= n! \left(1 - \frac{1}{18n^3}\right)^{2k}. \end{aligned}$$

since π is constant on the whole symmetric group and is equal to $\frac{1}{n!}$. We can rewrite this expression using Stirling's approximation and bound

$$4||\mu^{*k} - U||_{\text{TV}}^2 \leq \exp\left(-\frac{2k}{18n^3} + n\log n\right),$$

so if $k = 9n^3(n\log n + c)$ then the right-hand side is at most e^{-c}. This actually turns out to still be off, and we can improve this by using all of the eigenvalues:

Example 66

We can further improve the mixing time calculation to $n^3 \log n$ with a more careful analysis, since we know sharp mixing rates. Saloff-Coste and Zuniga found that for random transpositions, we have

$$n!||\tilde{\mu}^{*k} - U||_2^2 \leq 16e^{-2c} \quad \text{if } k = \frac{1}{2}n(\log n + c).$$

Well, using the same paths and the same $A = 36n^2$ bound, we know by Theorem 52 that

$$4||\mu^{*k} - U||_{\text{TV}}^2 \leq n!||\mu^{*k} - U||_2^2 \leq \left(\frac{1}{2}\right)^{2k} + e^{-k/(36n^2)} + n!\left|\left|\tilde{\mu}^{\left\lfloor \frac{k}{2(36n^2)}\right\rfloor} - U\right|\right|_2^2.$$

Thus we need $\frac{k}{72n^2} = \frac{1}{2}n(\log n + c) \implies k = 36n^3(\log n + c)$ to make this last term smaller than $16e^{-2c}$; at that order the first two terms are exponentially small so are negligible. So this tells us that in fact order $n^3 \log n$ steps are sufficient for chi-square distance to be small, and then taking square roots and then dividing by 2 tells us that $||\mu^{*k} - U||_{\text{TV}} \leq 2e^{-c} + o(1)$ for some explicitly known (super-exponentially small) $o(1)$ after this many steps.

Remark 67. *The paper that Professor Diaconis wrote with Saloff-Coste has many more examples and stories, many of which are using random transpositions to get good answers. In particular, Borel wrote a book called "The Mathematical Theory of Bridge," and in the middle it says that "you can make math out of shuffling" and suggested a bunch of shuffles. These shuffles included things like "top-to-random bottom-to-random" (meaning we take the top card and put it somewhere, then take the bottom card and put it somewhere), or "random-to-random," or "Borel shuffles" where we cut into three piles uniformly and then order them 3-2-1 instead of 1-2-3. Using comparison with random transpositions, the paper does all of these and gets the right order of mixing (with matching lower bounds).*

We might notice that top-to-random bottom-to-random is not reversible, but there's a parallel theory by considering the reversibilization and going back to the original chain; we'll talk about later on in the course.

All of these chains are conjectured to have cutoff – specifically, Professor Diaconis conjectured that any generating set of S_n has cutoff, and furthermore the slowest such chain is to take a transposition and an n-cycle so it's always $n^3 \log n$ at worst. (Well, specifically, there's a parity problem, so we can either put the card on the bottom or second from the bottom to fix it.)

There's been some developments on this: it's a fact that if we pick σ, τ uniformly in S_n and look at the group $\langle \sigma, \tau \rangle$ that they generate, then $G = S_n$ with probability $\frac{3}{4} + o(\frac{1}{n})$ and A_n with probability $\frac{1}{4} + o(\frac{1}{n})$. (Something like this is true for more general groups too.) So we might ask what the mixing time is for these generators, and Helfgott, Seress, and Zuk showed that if μ is $\frac{1}{4}$ on each of $\{\sigma, \tau, \sigma^{-1}, \tau^{-1}\}$, then the chain mixes in order $n^3(\log n)^8$ steps, and they do it using comparison with random transpositions. So we can try the same question ourselves with some other group if we'd like, such as $\text{SL}_n(\mathbb{F}_q)$ and comparing with random transvections!

We've been using the extra structure of random walks on groups to keep things simpler, and next time we'll see how this goes for general reversible chains and do some literature review.

9 April 27, 2026

In the last two lectures, we've done comparison theory for groups. We could spend another lecture or two on examples of that type (and that's what was done last time this course was offered), but what we'll do instead is consider comparison theory for general (reversible) Markov chains.

Example 68

Our setting will be the following: $\mathfrak{X}$ is a finite set, and $K(x,y)$ is an ergodic reversible Markov chain with respect to the stationary distribution $\pi(x)$. We then have another comparison chain $(\tilde{K},\tilde{\pi})$ where we know everything (because it involves orthogonal polynomials, or we were lucky in some way).

To do comparison, we need paths, and the way we set that up is that for every x,y with $\tilde{K}(x,y)>0$, we pick some path γ_{xy} from x to y **using steps possible for the chain that we care about**. That is, we have

$$\gamma_{xy}=[x_0=x,x_1,\cdots,x_\ell=y],\quad \text{where } K(x_{i-1},x_i)>0 \text{ for all } i.$$

Theorem 69

With the notation above, let $\mathcal{E},\tilde{\mathcal{E}}$ be the Dirichlet forms for the two chains. Then $\tilde{\mathcal{E}}(f,f)\le A\mathcal{E}(f,f)$ for

$$A=\max_{\substack{e=(z,w):\\ K(z,w)>0}}\frac{1}{\pi(z)K(z,w)}\sum_{\text{paths }\gamma_{xy}\ni e}|\gamma_{xy}|\tilde{\pi}(x)\tilde{K}(x,y).$$

Proof. This is the same argument we've been using for the last few results. We write down

$$\begin{aligned}\tilde{\mathcal{E}}(f,f)&=\frac{1}{2}\sum_{x,y}(f(x)-f(y))^2\tilde{\pi}(x)\tilde{K}(x,y)\\&=\frac{1}{2}\sum_{x,y}\left(\sum_{e\in\gamma_{xy}}f(e^-)-f(e^+)\right)^2\tilde{\pi}(x)\tilde{K}(x,y)\\&\le\frac{1}{2}\sum_{x,y}|\gamma_{xy}|\sum_{e\in\gamma_{xy}}(f(e^-)-f(e^+))^2\tilde{\pi}(x)\tilde{K}(x,y)\end{aligned}$$

by Cauchy-Schwarz with weight 1. And now swapping the order of summation yields

$$\begin{aligned}\tilde{\mathcal{E}}(f,f)&\le\frac{1}{2}\sum_e(f(e^-)-f(e^+))^2\frac{\pi(e^+)K(e^+,e^-)}{\pi(e^+)K(e^+,e^-)}\sum_{\gamma_{xy}\ni e}|\gamma_{xy}|\tilde{\pi}(x)\tilde{K}(x,y)\\&\le A\mathcal{E}(f,f).\end{aligned}$$

□

Fact 70

A similar result holds for the "negative eigenvalue" Dirichlet forms too: if we pick paths γ^*_{xy} of **odd length**, we can write the same kind of alternating telescoping series $f(x) - f(y) = (f(x) + f(x_1)) - (f(x_1) + f(x_2)) + \cdots$ and use Cauchy-Schwarz with this instead. The result is that $\tilde{\mathcal{F}}(f,f) \le B\mathcal{F}(f,f)$ for

$$B = \max_{\substack{e=(z,w):\\ K(z,w)>0}} \frac{1}{\pi(z)K(z,w)} \sum_{\gamma^*_{xy} \ni e} |\gamma_{xy}| \tilde{\pi}(x)\tilde{K}(x,y).$$

So interpreting these results, suppose $\tilde{\mathcal{E}}(f,f) \le A\mathcal{E}(f,f)$ **and furthermore we have** $\tilde{\pi} = \pi$. Then we have $\beta_i \le 1 - \frac{1-\tilde{\beta}_i}{A}$. Similarly, if $\tilde{\mathcal{F}}(f,f) \le B\mathcal{F}(f,f)$ and the stationary distributions agree, then we have $\beta_i \ge -1 + \frac{1+\tilde{\beta}_i}{B}$. For random walks on groups we didn't have to worry about π because it was always uniform; in general we have to actually bound the stationary distribution above and below, and we'll see examples of that very soon.

The point is that the story works in general and these bounds can be interesting. But now we'll see why we did the groups case first: if we want to bound total variation distance, we usually do it through eigenvalues and the series of calculations

$$4||K^\ell_x - \pi||^2_{\mathrm{TV}} \le \chi^2_x(\ell) = \sum_y \frac{(K^\ell(x,y) - \pi(y))^2}{\pi(y)} = \sum_{i=1}^{|\mathfrak{X}|-1} f_i(x)^2 \beta_i^{2\ell}$$

for $\{f_i\}$ an orthonormal basis of eigenfunctions for K. The problem is we don't really know the f_is for our chain well enough to evaluate them all at x; **in the group case that factor didn't appear** because the mixing time is the same from any x by symmetry.

One way to get around this is to consider the **average chi-square distance**

$$\chi^2_{\mathrm{avg}}(\ell) = \sum_x \pi(x)\chi^2_x(\ell).$$

The practical implication of this is that if we had a Markov chain that we were running in practice, we might run things for a long time so we're roughly distributed as π. Then we might ask the question of how many more steps we need until the state is independent from the previous measurement, and that can be quantified by how long it takes $\chi^2_{\mathrm{avg}}(\ell)$ to decay. But it's also very useful for computation, because

$$\chi^2_{\mathrm{avg}}(\ell) = \sum_x \pi(x)\chi^2_x(\ell) = \sum_x \sum_{i=1}^{|\mathfrak{X}|-1} f_i(x)^2 \pi(x) \beta_i^{2\ell},$$

and now the sum over x is just 1 because of orthonormality. Thus we always have

$$\chi^2_{\mathrm{avg}}(\ell) = \sum_{i=1}^{|\mathfrak{X}|-1} \beta_i^{2\ell}$$

without needing to know the eigenfunctions at all.

Example 71

Here's an example where we wouldn't really know how to study the chain of interest at all directly. Consider an $n \times n$ square grid on $\mathbb{Z}^2$; we know how to bound simple random walk on this because we have a product chain, but now suppose we remove a few edges while keeping the chain connected. Specifically, suppose we do so in a way so that at most one edge per square is removed.

The nearest-neighbor random walk on this new graph G now has stationary distribution $\pi(x) = \frac{\deg(x)}{2|E|}$ for E the total number of edges, and $\pi(x)K(x,y) = \frac{1}{2|E|}$ if x, y are connected and 0 otherwise.

Comparison turns out to be very useful in this setting if we compare with the product of two "random walks on a path with holding at the endpoints" from Example 21. The resulting product chain is nearest-neighbor random walk on a square grid, except with holding proportional to 2 on the corners and proportional to 1 on the edges: specifically $\tilde{K}(x,y) = \frac{1}{4}$ if x, y are adjacent, $\tilde{K}(x,x) = \frac{1}{2}$ for the four corners, and $\tilde{K}(x,x) = \frac{1}{4}$ for all other vertices on the edges. And $\tilde{\pi}(x) = \frac{1}{n^2}$ and we know all of the eigenvalues of the path chain, hence the product chain as well: it turns out that

$$\tilde{K} \text{ has eigenvalues } \frac{1}{2}\left(\cos\left(\frac{\pi j}{n}\right) + \cos\left(\frac{\pi k}{n}\right)\right) \text{ for } 0 \le j, k \le n-1,$$

and so in particular the biggest nontrivial eigenvalues on either side are

$$\tilde{\beta}_1 = \frac{1}{2}\left(1 + \cos\frac{\pi}{n}\right), \quad \tilde{\beta}_{n^2-1} = -\cos\left(\frac{\pi}{n}\right).$$

Now these two chains have different stationary distributions, but they aren't so badly different: letting E denote the number of edges in the modified graph, we have

$$\frac{n^2}{|E|}\tilde{\pi}(x) \le \pi(x) \le \frac{4n^2}{|E|}\tilde{\pi}(x) \text{ for all } x.$$

For paths, we must connect things adjacent in the $\tilde{K}$ chain using edges from the K chain (since only the terms where $\tilde{K}(x,y) > 0$ contribute to the sum). But the point of our assumption is that **we can always replace a missing edge by going around the other three edges in one of its adjacent squares**, so the path length will always be at most 3. And to count the number of paths that use any given edge to bound the covering number, we can check that if we pick Some convention for how to loop around, at most 7 different adjacent pairs use a given edge in the deleted graph. Putting this together, we get $\tilde{\mathcal{E}} \le \frac{7\cdot 3\cdot |E|}{n^2}\mathcal{E}$, so

$$\beta_i \le 1 - \frac{n^2}{21|E|}(1 - \tilde{\beta}_i).$$

Of course, this is a toy example, but it's one where we wouldn't know what to do otherwise. And again, all of this can be found in the paper "Comparison Theorems for Reversible Markov Chains" with Saloff-Coste.

Example 72

A more substantial application is the following: let G be a connected simple graphs on n vertices, for example the n-cycle, and put down k particles for some $0 < k < n$ so that at most one of them lies at every site. In the **symmetric exclusion process**, we pick a particle at random, pick one of its neighboring sites, and then move if it's not occupied (and stay if it is). So here we have a Markov chain on k-element subsets of the n vertices.

Because of the holding and reversibility, the stationary distribution is uniform over all configurations, meaning $\pi(x) = \frac{1}{\binom{n}{k}}$ for all sets x. The existing bounds at the time were very bad for mixing time on this, but actually we can do **comparison with random transpositions** and get sharp bounds on β_1 and $\beta_{\min}$. Here when we say "random transpositions," consider the chain where we pick something in the k-set and something outside the k-set and switch them; this is essentially the **Bernoulli–Laplace urn model** and it is well understood.

Example 73

In the actual Bernoulli–Laplace urn, we have k red balls and $n-k$ black balls in two urns, and each time we pick a ball in the left and right urn and switch them. So if we label the balls it's the same thing. This chain was really "the first Markov chain" back from 1780, and it was used as a model of how gas goes through a porous membrane; they got pretty good estimates for how long it would take via a continuous model. And Professor Diaconis and Shahshahani found all of the eigenvalues and eigenvectors of $\hat{K}(x,y)$ so that it could indeed be used for comparison.

Of course, there is some combinatorics we have to do to relate the two chains and deal with the restrictions on transpositions, since $\tilde{K}$ has nothing to do with the geometry of the graph but K relies on it. So we need to choose paths, and we can do so in some complicated way.

Again, dozens of other examples can be found from the citations of the comparison theory paper (in various fields like physics and computer science). But the point is just to see that it can be useful, and also that having nice chains where we know everything is very useful even if things are very special.

Fact 74

There's been an amazing development in this area recently (though it's rather technical) in the paper "Sharp Poincaré and log-Sobolev inequalities for the switch chain on regular bipartite graphs" of Tikhomirov and Youssef. This paper has a systematic machine for comparing Markov chains on different state spaces, for example if we removed a few vertices instead of just edges from our square grid in Example 71. And with their machinery, it's even possible to compare stationary distributions that are wildly different and still get good bounds on the eigenvalues.

To describe the strategies in more detail, we need to understand log-Sobolev inequalities, which we won't do know. But the point is that comparison theory is alive and well.

What we'll do now is state three tractable open problems for simple random walk which we can do with our techniques, and Professor Diaconis thinks they should be doable. All three are on the symmetric group:

Problem 75

We previously considered the (symmetrization of) the chain "transpose cards 1 and 2, or put the top card on the bottom." Now consider the transposition $(1,m)$ and the n-cycle $(1,2,\cdots,n)$; these generate S_n if and only if $m-1$ and n are relatively prime. Thus we can define the symmetrized chain with $Q(s)=\frac{1}{4}$ for $s=\mathrm{id},(1,m),(1,\cdots,n),(n,\cdots,1)$ and $Q(s)=0$ otherwise; describe the rate of convergence of this random walk (specifically how it depends on n).

Problem 76

Suppose $3|n$. In 1901, Miller produced generators s_1, s_2 for S_n such that $s_1^2 = s_2^3 = \mathrm{id}$. More concretely, let

$$s_2 = (1,2,3)(4,5,6)\cdots(n-2,n-1,n)$$

Now pick a prime p such that $\frac{n}{2} < p < n-2$, define $a = n-p$, and define, if a is even, the product of disjoint transpositions

$$s_1 = (3,4)(6,7)(9,10)\cdots(n-3,n-2)\left(1,\frac{3a+5}{2}\right),(n-4,n-1),$$

and the same but with $\left(2,\frac{3a+4}{2}\right)$ if a is odd. These are pretty "rigorous" group multiplications, so intuitively it should mix quite fast. But it's not known what the diameter of the group is or how to get mixing times.

Problem 77

For a reference on this one, see "Generating Symmetric Groups" by Isaacs and Zieschang. For every $n \neq 4$, for all non-identity $x \in S_n$, there exists some $y \in S_n$ such that x, y together generate S_n. In particular, if $x = (1,2,\cdots,m)$ for any m, n relatively prime (take card m out and put it on top), we can always take $y = (1,2,\cdots,n)$. Bound the mixing time of the random walk for these two generators and see how it depends on n.

The point is that even if we know "almost all generators give at most $n^3 \log n$ mixing time," there are famous or interesting or basic specific parametrized families of problems where we don't know the answer yet!

Remark 78. *A good reference on the more recent developments is Dyer, Goldberg, Jerrum, and Martin's expository paper "Markov chain comparison." In particular, it has some new theorems and techniques for non-reversible chains.*

Next time, we'll discuss non-reversible chains in a different way from what we saw in Professor Chatterjee's lecture; there will be some new ideas there. Then we'll move on to the second part of the course, where we discuss "how to make Markov chains that do what we want in real scientific settings."

10 April 29, 2026

We'll consider non-reversible chains today, still sticking with the case where we have finite state space $\mathfrak{X}$ and an ergodic Markov chain $K(x,y)$ with stationary distribution $\pi(x)$, but now **without the assumption of reversibility**. Many, if not most, Markov chains run in practice for applications (and proven with theory) are reversible, but many are not:

Example 79

In the top-to-random shuffle, we take a deck of cards and repeatedly take the top card and put it somewhere at random. Similarly, in riffle shuffling, we split the deck in half and then "riffle" them together in one of various ways. we Both of these examples on S_n are of interest but are not reversible.

Example 80

For some examples on C_n, the integers mod n, we can consider simple random walk with drift (so probability p of going to the left and q of going to the right). And we can also consider the chain

$$X_{n+1} = 2X_n + \varepsilon_{n+1}, \quad \varepsilon_i \in \{-1, 0, 1\} \text{ with probability } \frac{1}{3} \text{ each;}$$

this is known as the **Chung–Diaconis–Graham** and has been studied extensively.

Recently in the applied literature, it's been shown that nonreversible chains mix faster in some sense; for example, it's been shown that the spectral gap can be made bigger while keeping the stationary distribution the same if we go from reversible to nonreversible. For more, we can see various papers by Radford Neal or the paper by Professor Diaconis and Laurent Miclo "On the spectral analysis of second-order Markov chains."

Fact 81

For reversible chains for (weighted) random walks on graphs, it's very easy to write down the stationary distribution. But if we do the same for a nonreversible chain, it can be quite a nightmare to understand. For example, consider the chain on C_p for p prime given by

$$X_{k+1} = X_k^2 + \varepsilon_{k+1}$$

for ε_{k+1} the same as before. Then the stationary distribution is wildly non-uniform, and no one really understands how to describe it.

Getting to business now, write $K_x(y)$ for $K(x, y)$ and let K act on functions as

$$Kf(x) = \sum_y K(x, y) f(y).$$

We'll be working in $\ell^2(\pi)$, which is the set of all functions $f : \mathfrak{X} \to \mathbb{R}$ (note that often we would use $\mathbb{C}$, but we're not doing that here) with inner product as usual,

$$\langle f_1, f_2 \rangle = \sum_{x \in \mathfrak{X}} f_1(x) f_2(x) \pi(x).$$

Since we're working in $\ell^2(\pi)$, we integrate with respect to π and so the "density of the measure K_x with respect to π" is $\frac{K_x}{\pi}(y) = \frac{K(x,y)}{\pi(y)}$. (As we might be able to tell, everything we talk about here generalizes to more abstract spaces, and the language we use here works very generally.)

As usual, we're curious about the total variation distance to stationarity

$$||K_x^n - \pi||_{\text{TV}} = \frac{1}{2} \sum_y |K^n(x, y) - \pi(y)| = \max_{A \subseteq \mathfrak{X}} |K^n(x, A) - \pi(A)|,$$

and we still have, by Cauchy-Schwarz,

$$2||K_x^n - \pi||_{\text{TV}} \leq \left|\left| \frac{K_x^n}{\pi} - 1 \right|\right|_2.$$

The new idea now is to **bound total variation by using reversibilizations of** K. There are lots of ways we can do this, but the main one is the following:

Definition 82

For a Markov kernel K, let $K^*(x, y) = \frac{\pi(y)K(y,x)}{\pi(x)}$ be its **adjoint** (also sometimes called the **time-reversal**).

Observe that this is nonnegative and when summed over y we get $\frac{\pi(x)}{\pi(x)} = 1$ by the definition of stationarity, so K^* is also a Markov kernel with π as stationary distribution because

$$\sum_x \pi(x)K^*(x,y) = \sum_x \pi(y)K(y,x) = \pi(y).$$

This adjoint is also the adjoint in the linear algebra sense; that is, $\langle Kf_1, f_2\rangle = \langle f_1, K^*f_2\rangle$ for all f_1, f_2. This K^* is not reversible either, but it lets us define the **multiplicative symmetrization** K^*K and check that it is reversible; similarly we can define the **additive symmetrization** $\frac{1}{2}(K + K^*)$. Thus we can use eigenvalues and paths and spectral theory and everything else, and (this is the burden of this lecture) then use them to get bounds on the original chain.

The idea is to break the ℓ^2 norm up into two parts: for any nonnegative n_1, n_2 with $n_1 + n_2 = n$, we can write

$$\boxed{\max_x \left|\left| \frac{K_x^n}{\pi} - 1 \right|\right|_2 = ||K^n - \Pi||_{2\to\infty} \le ||K^{n_1}||_{2\to\infty} ||K^{n_2} - \Pi||_{2\to 2}},$$

where Π is the matrix whose rows are all π. To explain this notation, suppose A is an $m \times n$ matrix and we define $||\cdot||_\alpha, ||\cdot||_\beta$ to be vector norms on $\mathbb{R}^n$, $\mathbb{R}^m$ respectively (we'll just be using $\ell^1, \ell^2, \ell^\infty$). Then we write down the operator norm

$$||A||_{\alpha\to\beta} = \max_{x\in\mathbb{R}^n : ||x||_\alpha \le 1} ||Ax||_\beta.$$

There's some trickery here, and for more details we can again check Horn and Johnson's book. In particular, we'll find the following "little theorem:" if $m = n$, we have

$$||A||_{2\to\infty} = \max_{1\le i\le n} ||A_{i:}||_2$$

for $A_{i:}$ the ith row of the matrix A, and this explains the first equality in the boxed identity above. And the inequality afterward uses the fact that $||AB|| \le ||A||\,||B||$ for any matrices A, B, where we're writing

$$K^n - \Pi = K^{n_2}K^{n_1} - \Pi K^{n_1} = (K^{n_2} - \Pi)K^{n_1}.$$

The two norms on our right-hand side should both be familiar (or workable) for us: we will denote

$$\mu(n) = ||K^n - \pi||_{2\to 2} = \sup_{f\in\ell_2(\pi) : ||f||_2 = 1} ||(K^n - \pi)f||_2$$

and similarly (this is sometimes called the **decay rate**)

$$D(n) = ||K^n||_{2\to\infty} = \sup_x \left|\left| \frac{K_x^n}{\Pi} \right|\right|_2.$$

The strategy now is that $\mu(n)$ can be estimated via eigenvalues of the multiplicative reversibilization K^*K, and $D(n)$ can be estimated via Nash inequalities (which is some refinement of Poincaré). We won't go into the latter in this course in too much detail, but we can find more details (and everything else in this lecture) in the paper "Nash inequalities for finite Markov chains" by Professor Diaconis and Saloff-Coste.

Lemma 83

Let $\pi_* = \min \pi(x)$. Then

$$2||K_x^n - \pi||_{\mathrm{TV}} \le \left|\left| \frac{K_x^n}{\pi} - 1 \right|\right|_2 \le \min_{\substack{n_1,n_2\ge 0: \\ n_1+n_2=n}} D(n_1)\mu(n_2) \le \pi_*^{-1/2}\mu(1)^n.$$

Proof. Everything except the last inequality was already in the boxed identity above. For the last one, we just have to

check that we can use $n_1 = 0$ and $n_2 = n$ and compute $D(0) = \pi_*^{-1/2}$ explicitly. □

So now we need to be able to say something about $\mu(1) = ||K^n - \pi||_{2\to 2}$ using eigenvalues. Note that K^*K is reversible with respect to π, so it has some eigenvalues $\beta_0 = 1 \geq \beta_1 \geq \cdots \geq \beta_{|\mathfrak{X}|-1} \geq 0$ (the lower bound is because the minimax characterization of eigenvalues uses $\langle K^*Kf, f\rangle = \langle Kf, Kf\rangle \geq 0$). We can then define the top (nontrivial) singular value

$$\mu(K) = \sqrt{\beta_1(K^*K)}.$$

The minimax characterization then tells us that

$$1 - \mu^2 = \min_{\substack{f:\mathfrak{X}\to\mathbb{R}:\\ \pi(f)=0\\ ||f||_2=1}} \mathcal{E}_*(f, f),$$

where the Dirichlet form of the symmetrized chain is

$$\mathcal{E}_*(f, f) = \langle (I - K^*K)f, f\rangle = \frac{1}{2}\sum_{x,y}(f(x) - f(y))^2\pi(x)K^*K(x, y).$$

But in this case because we can move the adjoint over, we also have

$$\mathcal{E}_*(f, f) = ||f||_2^2 - \langle Kf, Kf\rangle = ||f||_2^2 - ||Kf||_2^2.$$

Therefore, we have the following (remembering that we want to consider f with $||f|| = 1$):

Lemma 84

Let (K, π) be a Markov chain on $\mathfrak{X}$. Then the $2 \to 2$ norm is exactly $||K - \Pi||_{2\to 2} = \mu$, and therefore

$$||K^n - \Pi||_{2\to 2} \leq \mu^n.$$

Proof. Observe that $(K - \Pi)f = Kf - \Pi f$, and $||f - \Pi f||_2^2 = ||f||_2^2 - ||\Pi f||_2^2$. Thus

$$||K - \Pi||_{2\to 2} = \max_{\substack{f:\mathfrak{X}\to\mathbb{R}:\\ \Pi f=0,\\ ||f||_2=1}} ||Kf||_2 = \mu,$$

since we've subtracted off the stationary distribution to get rid of the top singular value. Then we get the other statement by multiplicativity. □

Laurent Saloff-Coste's lectures on finite Markov chains (in the Springer lecture notes) also has lots of this content in more detail. In particular, it covers Nash and log-Sobolev inequalities, but it begins with all of this symmetrization.

Example 85

Let $\mathfrak{X} = S_n$ and define a probability measure (in cycle notation)

$$Q(\mathrm{id}) = Q((1, 2)) = Q((1, 2, \cdots, n)) = \frac{1}{3}.$$

(So either swap card 1 with 2 or move it to the bottom.) We then get the associated Markov chain $K(\sigma, \eta) = Q(\eta\sigma^{-1})$, and as with any other random walk on a group the stationary distribution is uniform over all $n!$ permutations.

The "reversed" random walk is then $Q^*(\mathrm{id}) = Q^*((1,2)) = Q^*((n, n-1, \cdots, 1)) = \frac{1}{3}$, and therefore we can compute K^*K by taking a step from Q and then a step from Q^*. It turns out that

$$Q^*Q(\mathrm{id}) = \frac{1}{3}, \quad Q^*Q((1,2)) = \frac{2}{9}, \quad Q^*Q((1,2,\cdots,n)) = Q^*Q((n,n-1,\cdots,1)) = \frac{1}{9},$$

$$Q^*Q((1,2)(1,2,\cdots,n)) = Q^*Q((n,n-1,\cdots,1)(1,2)) = \frac{1}{9}.$$

So this is a very specific reversible Markov chain, supported at six points, and thus it's very easy to do this by paths via comparison with random transpositions. We find that

$$\mu^2(Q) = \beta_1(Q^*Q) \leq 1 - \frac{1}{41n^3}.$$

So this gives us some rate: we have that

$$2||K^\ell - \pi||_{\mathrm{TV}} \leq \sqrt{n!}\left(1 - \frac{1}{41n^3}\right)^{\ell/2}.$$

So we have an $e^{n \log n}$ prefactor and thus something like $n^4 \log n$ steps is sufficient to mix with this calculation. (The right answer turns out to be $n^3 \log n$, but that's another story.)

Remark 86. *As a word of caution, using K^*K can destroy connectivity of the chain. For example, if we instead take $Q((1,2)) = Q((1,\cdots,n)) = \frac{1}{2}$ in the above example, we get $Q^*Q(id) = \frac{1}{2}$ and $Q^*((1,2)(1,\cdots,n)) = Q^*((n,\cdots,1)(1,2)) = \frac{1}{4}$, and so Q^*Q is concentrated only on the cyclic subgroup generated by the element $(1,2)(1,\cdots,n)$, which means our eigenvalue bounds won't quite work.*

Luckily, this can be fixed by symmetrizing $Q * Q$ instead (we want to bound powers of Q anyway, so bounding powers of Q^2 is also okay), and then it generates everything. In fact, all kinds of polynomials in Q and Q^* (which are symmetric in the right way) end up getting used, and it's worth knowing that they can all work.

Fact 87

Consider the Chung–Diaconis–Graham walk on the cyclic group C_p from above, given by $X_{n+1} = 2X_n + \varepsilon_{n+1}$. It was proven (in a paper by those authors) that order $\log p \log\log p$ steps suffice, and also that this many steps are needed for infinitely many p.

If we try to do this via symmetrization and paths, the multiplicative symmetrization gives $\mu^2 = 1 - \frac{c}{p^2} + O(\frac{1}{p^4})$, which says that order p^2 steps suffice. So we're unlucky in that it gives us an exponentially wrong result in this case. But lots of other examples do give us useful (and more correct) bounds.

Remark 88. *Consider the random walk on all nonnegative integers $\{0,1,2,\cdots\}$ which goes left with probability $\frac{2}{3}$ and right with probability $\frac{1}{3}$, and if we're at 0 we stay with probability $\frac{1}{2}$ and go right with probability $\frac{1}{2}$. (We can cut it off at finite N if we want.) The stationary distribution of this walk is geometric with parameter $\frac{1}{2}$, and the spectral gap is of constant size. So if we start at some large n and use the bound $\frac{1}{\sqrt{\pi(n)}}\mu^\ell$, it tells us that ℓ needs to be of order n until it's accurate. And similarly, if we start near 0, some techniques tell us that ℓ can be much smaller and we're still close to being mixed. So this does kind of give us the right behavior in that way, and we get other tight results using more techniques of Nash inequalities (which don't require reversibility at all).*

11 May 4, 2026

We're starting a "new subject" today. So far, we've been mentioning some techniques for bounding rates of convergence of Markov chains, and while we could do that forever, we'll switch now to actually **constructing Markov chains**. If we look in the literature, there are many ideas but really there are just two or three central ones, and we'll try to make sense of them here so that we can go in our own directions.

As a quick outline, we'll begin with the **Metropolis algorithm** and understanding how it can be thought of as an L^1 projection. We'll then discuss the other extremely widely-used **Gibbs sampler** (also **Glauber dynamics**) and the connection to von Neumann's alternating projections. Both of these are widely used in various areas and applications. There's also something called the **exchange algorithm** for doubly intractable Markov chains. And there's a whole class of algorithms sometimes called **hit and run** or **auxiliary variables** or **data augmentation**, which are all the same idea restated in different notation coming from different fields.

Finally, the two families of chains most often used these days are **Hamiltonian** (also **hybrid**) **Monte Carlo** and the **no-U-turn sampler**. In nice algebraic cases, we can prove some things, but for any actual application (even if we make up our own versions) nothing can be proved.

We'll begin with the Metropolis algorithm, which was listed in 2000 as the first of "ten great algorithms." Everything will be said for general state spaces eventually, but for now let's first consider the finite case.

Example 89

Our setting will be that $\mathfrak{X}$ is a finite set and $\pi(x)$ is some distribution on it, and our goal is to sample from π. We are given some way of moving around on $\mathfrak{X}$ via a Markov chain $K(x,y)$ that we can run, except that **the stationary distribution of K does not have to be π** (and the chain doesn't have to be reversible). The algorithm will essentially "run K with some additional coin-flipping" so that π is the stationary distribution for the new chain $M(x,y)$.

Definition 90

In the setting above, define the **acceptance ratio**

$$A(x,y) = \frac{\pi(y)K(y,x)}{\pi(x)K(x,y)}.$$

The **Metropolis algorithm** is the following: from x, choose y from $K(x,y)$. Then flip a coin with heads-probability $\min(A(x,y),1)$; if the coin comes up heads then move to y, and otherwise stay at x.

Thus, our Metropolis algorithm has

$$M(x,y) = K(x,y)\min(A(x,y),1) \quad \text{if } x \neq y$$

and therefore (either proposing $x \mapsto x$, or adding up the tail probabilities to adjacent states)

$$M(x,x) = K(x,x) + \sum_{z \neq x : A(x,z) < 1} (1 - A(x,z)).$$

Lemma 91

With the notation above, the chain $M(x,y)$ is reversible with respect to the (stationary) distribution π.

Proof. Consider any $x, y \in \mathfrak{X}$. First suppose $A(x,y) < 1$. Then

$$\begin{aligned}\pi(x)M(x,y) &= \pi(x)K(x,y)A(x,y)\\ &= \pi(y)K(y,x)\\ &= \pi(y)M(y,x)\end{aligned}$$

since $A(y,x) = \frac{1}{A(x,y)} \geq 1$. Similarly if $A(x,y) > 1$, we just run the chain of equalities in reverse with x and y swapped. Finally if $A(x,y) = 1$, then $\pi(x)K(x,y) = \pi(y)K(y,x)$ and there are no coin flips so the same is true for M. □

So this takes any chain and turns it into a π-reversible chain, and the key point is that we can run this Metropolis chain without knowing the normalizing constant Z in situations where we know that $\pi(x) = Z^{-1}e^{\beta H(x)}$ for some energy function $H(x)$ (and where Z is typically not known in lots of models). Instead, the normalizing constant cancels out in the expression for $A(x,y)$ and so everything is somehow local. Our job in this lecture will be to understand "how anyone thought of this" (where it arose), and then use that to understand some useful variations and applications.

To begin, the idea (derivation) is that we have π given to us up to normalization, and we have some way of moving around our state space (which is our given $K(x,y)$ – think of this as a geometry of some sort). Our idea would then perhaps be to take a step from K and either go there (move to y) or don't (stay at x). Letting the probability of moving by $\theta(x,y)$, we now want to figure out what θ should be to make π stationary. There could be many ways of doing this, but the easiest is to make the combined chain $M(x,y)$ π-reversible, meaning that

$$\pi(x)K(x,y)\theta(x,y) = \pi(y)K(y,x)\theta(y,x) \implies \theta(x,y) = \frac{\pi(y)K(y,x)}{\pi(x)K(x,y)}\theta(y,x) = A(x,y)\theta(y,x).$$

In particular, this expression must be at most $A(x,y)$ since $\theta(y,x)$ is a probability, and for the same reason $\theta(x,y)$ needs to also be at most 1:

Lemma 92

Any choice of $\theta(x,y)$ such that $\theta(x,y) \leq \min(A(x,y),1)$ and $\theta(y,x) = \frac{\theta(x,y)}{A(x,y)}$ will also yield a reversible Markov chain with stationary distribution π.

So the Metropolis algorithm choice of $\theta(x,y) = \min(A(x,y),1)$ is exactly the largest possible coin flip, meaning that we have the fastest possible chain in terms of moving away from x. But for example chemists use the **Barker dynamics**

$$\theta(x,y) = \frac{\pi(x)K(x,y)}{\pi(x)K(x,y)+\pi(y)K(y,x)} = \frac{A(x,y)}{1+A(x,y)}.$$

which also makes sense. (And for another variation of this algorithm, we could have also picked two different potential y_1, y_2 from $K(x,\cdot)$ and then flip coins to decide between the different options for which to move to. This goes under the name of "multiple-try Metropolis.")

Remark 93. *In many contexts, our starting chain K is symmetric. Then $A(x,y) = \min(\frac{\pi(y)}{\pi(x)},1)$, and so we can describe our chain in words as "always move from x to y if $\pi(y) > \pi(x)$, and otherwise only go there with probability $\frac{\pi(y)}{\pi(x)}$."*

Example 94

Abstracting this to more general state spaces now, let $(\mathfrak{X},\mu)$ be a σ-finite measure space, and let $\pi(dx)$ be a probability measure on $\mathfrak{X}$ such that $\pi \ll \mu$. We'll write $\pi(x)$ for the density of π with respect to μ, and let $K(x,dy)$ be any Markov kernel (way of moving around) on $\mathfrak{X}$.

In order to define Metropolis, we need, for each x, that $K(x,\cdot) \ll \mu$ except for a potential atom $K(x,\{x\}) \geq 0$. Letting its density be written $k(x,y)$, we get the same recipe as before for Metropolis:

$$M(x,dy) = k(x,y)\min(1, A(x,y))\,\mu(dy) + a(x)\delta_x(dy), \quad A(x,y) = \frac{\pi(y)k(y,x)}{\pi(x)k(x,y)},$$

where

$$a(x) = K(x,\{x\}) + \int (1 - A(x,z))_+ k(x,z)\mu(dz)$$

is the total probability of staying where we are. And the same results (reversible with stationary π) still hold; remember that we do still have to check connectedness and aperiodicity, but that isn't usually a problem.

Example 95

We're now ready to show what this chain "does for a living." Going back to a finite state space $\mathfrak{X}$ now, let $\mathcal{S}(\mathfrak{X})$ be the set of all stochastic (row-sums 1) $|\mathfrak{X}| \times |\mathfrak{X}|$ matrices, or in other words the set of all Markov chains on $\mathfrak{X}$. This is a convex subset of $\mathbb{R}^{|\mathfrak{X}|^2}$ of dimension $|\mathfrak{X}| \cdot (|\mathfrak{X}| - 1)$. Let $\mathcal{R}_\pi(\mathfrak{X})$ be the set of all π-reversible such elements of $\mathcal{S}(\mathfrak{X})$; this is some convex subset of $\mathfrak{X}$ of dimension $\frac{|\mathfrak{X}|\cdot(|\mathfrak{X}|-1)}{2}$. Then the Metropolis algorithm is a map $\mathcal{M}$ from $\mathcal{S}(\mathfrak{X})$ into $\mathcal{R}_\pi(\mathfrak{X})$ which takes some K into $M(K)$ via the above construction.

We'll introduce a metric d_π on $\mathcal{S}(\mathfrak{X})$ which measures how far apart two Markov chains are:

$$d_\pi(K, K') = \sum_x \pi(x) \sum_{y \neq x} |K(x,y) - K'(x,y)|.$$

So this is like the total variation distance except omitting the diagonal (which is necessary, as we will soon see). This is symmetric and satisfies the triangle inequality, and if $d_\pi(K,K') = 0$ then $K(x,y) = K'(x,y)$ for all $x \neq y$ and thus $K(x,x) = K'(x,x)$ as well because the matrices are stochastic.

So the picture is that $\mathcal{S}(\mathfrak{X})$ is some big convex set and $\mathcal{R}_\pi(\mathfrak{X})$ is some convex subset of it.

Theorem 96

The Metropolis algorithm, thought of as a map $\mathcal{M} : \mathcal{S}(\mathfrak{X}) \mapsto \mathcal{R}_\pi(\mathfrak{X})$, minimizes the distance $d_\pi(K,\cdot)$ when taken into $\mathcal{R}_\pi$. That is, $\mathcal{M}$ maps to the closest point in the convex set $\mathcal{R}_\pi$ in this funny metric d_π, and in fact $\mathcal{M}(K)$ is the unique closest element of $\mathcal{R}_\pi$ which is coordinate-wise smaller than $K(x,y)$ on the off-diagonal entries.

Remark 97. *To elaborate a bit more, for $x \neq y$ we have*

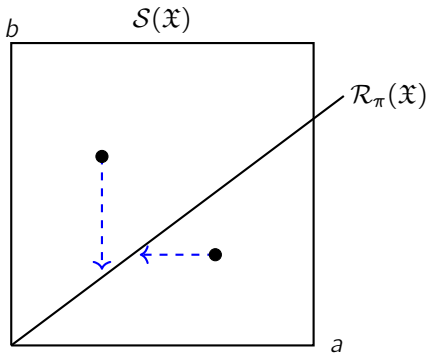
$$\mathcal{M}(K)(x,y) = K(x,y)\min\left(\frac{\pi(y)K(y,x)}{\pi(x)K(x,y)}, 1\right) = \min\left(\frac{\pi(y)K(y,x)}{\pi(x)}, K(x,y)\right) \leq K(x,y),$$

so the Metropolized $\mathcal{M}(K)$ is "always less" than K. And this is important because combined with the minimax characterization, it implies $\beta_i(M) \leq \beta_i(\tilde{K})$ whenever $\tilde{K}$ is in this "coin-flipping class" where we follow K but additionally combine it with coin flips. So in that sense it is also the fastest mixing such chain under these restrictions.

Example 98

For an instructive example, suppose we're on the two-state chain $\mathfrak{X} = \{0,1\}$. Then $\mathcal{S}(\mathfrak{X})$ is the set of matrices $\begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$ for $0 \leq a, b \leq 1$, and $\mathcal{R}_\pi$ is the subset of it such that $\pi(0)a = \pi(1)b$

If $\pi(1) > \pi(0)$ the diagram looks something like this (the square is $\mathcal{S}$, the line is $\mathcal{R}_\pi$, and the map $\mathcal{M}$ maps (a, b) to $(\frac{\pi(1)}{\pi(0)}b, b)$ when it's below the line and (a, b) to $(a, \frac{\pi(0)}{\pi(1)}a)$ when above it – it's piecewise coordinate-wise linear in the two sliced-up parts.



Fact 99

The same theorem holds in the abstract case: if we define

$$d_\pi(K, K') = \int_{\mathfrak{X}\times\mathfrak{X}\setminus\Delta} \pi(x)|k(x, y) - k'(x, y)|\mu(dx)\mu(dy),$$

then all of the results still hold.

We won't go through the proof of this result in class because it's rather technical (and so it's hard to get something out of it live). But a good reference is the paper "A Geometric Interpretation of the Metropolis–Hastings Algorithm" by Billera and Professor Diaconis.

Example 100

The original use for the Metropolis algorithm was the **hard disks in a box** problem – we want to have non-overlapping disks, so the configuration space is specified by the n disk centers. This inherits a uniform distribution from the Lebesgue measure, and the task is to sample uniformly from this set.

Three big classes of algorithms – Metropolis, Glauber, and molecular dynamics – were all invented to solve this. And Metropolis basically comes down to "pick a disk, try to move it within ϵ of the original location, and then perform it if allowed."

Fact 101

Metropolis itself isn't used so often these days, but what takes its place (and uses Metropolis under the hood) is hybrid Monte Carlo. In a few sentences, we run a deterministic diffusion via some differential equation, but we get roundoff error in high dimensions even with many decimal places of accuracy. So we use that as a proposal and then perform Metropolis to see whether to accept a run or not. (And similarly, Metropolis works well if we want to discretize a continuous density and still be exact, for example by shifting where we do the discretization.)

Similarly, the Gibbs sampler takes a probability distribution in many coordinates and resamples each one conditionally on the others. If we actually want to do this and pick from the one-dimensional conditional distribution, we have some density and again often we use Metropolis to make each step work if we can't do it in closed form.

Many other examples (for example in understanding the mathematics of gerrymandering and how far off a proposed division is from the "expected distribution"). And next time we'll talk a bit about whether we can prove anything about Metropolis.

12 May 6, 2026

The goal of today is to understand whether we can prove anything about the Metropolis algorithm. The answer is, as in many questions of this type, "yes and no." There's been lots of papers which have proved things (many by Professor Diaconis), but mostly the places where we can prove things are in "nice" toy problems with some algebra which allows the use of group theory or something similar, or where the setting is in low dimensions. But for real problems really we can't prove anything useful.

Fact 102

It's worth keeping in mind that Metropolis "doesn't always work well." For example, if our distribution is the union of two spaced-out Gaussian peaks (say, discretized), and our starting walk $K(x, y)$ is nearest-neighbor random walk, then it's exponentially unlikely for us to cross the gap between the two peaks (since we have to keep flipping heads to go to things with smaller and smaller $\pi(x)$). So in real problems with multi-modal distributions, we do end up getting stuck in one of the local groups a lot of the time.

But Metropolis does turn out to be okay in low-dimensional "unimodal-type" examples; we'll basically climb to the top of the single peak and stay around that. And remember that we did prove an example of this with paths earlier on in the course (see Example 45) in a case where the stationary distribution falls off exponentially from its peak, and we managed to get a constant spectral gap. In any low-dimensional example (a grid in two or three dimensions) with π of this form, we'll see the same kind of story where it takes order n steps to get to the peak and then is basically random.

Fact 103

For another situation, suppose π is proportional to a low-dimensional polynomial, say $\pi(j) = \frac{aj^3+bj^2+cj+d}{Z}$, even if we can have "polynomial wiggles." Then order n^2 steps turn out to be sufficient – we can get across polynomial-sized peaks, and we can prove this using the path techniques from earlier in the course. (The constants depend on the degree of the polynomial.) The paper "What do we know about the Metroplis algorithm?" by Professor Diaconis and Saloff–Coste goes through this in detail.

Fact 104

Besides these cases, "sometimes a miracle happens" and we can explicitly diagonalize the Metropolis algorithm using algebra. For example, suppose we have the random walk on the hypercube C_2^n, cycle C_n, or symmetric group S_n. Then we can Metropolize in interesting ways that are still tractable.

For an explicit example, we'll consider the following setup:

Example 105

Consider the hypercube C_2^d and let K be our nearest-neighbor random walk with no holding (since Metropolis will take care of the holding for us), meaning that $K(x,y) = \frac{1}{d}$ if $d_{\text{Ham}}(x,y) = 1$ and 0 otherwise. We'll then Metropolize to the distribution

$$\pi_\theta(x) = \frac{\theta^{H(x)}}{(1+\theta)^d}, \quad 0 < \theta \leq 1,$$

where $H(x) = d_{\text{Ham}}(0,x)$ is the number of ones in x.

So if we want to change to the new chain, we have that

$$M(x,y) = \begin{cases} \frac{1}{d} & \text{if } d_{\text{Ham}}(x,y) = 1 \text{ and } H(y) < H(x), \\ \frac{\theta}{d} & \text{if } d_{\text{Ham}}(x,y) = 1 \text{ and } H(y) > H(x), \\ (1-\theta)\left(1 - \frac{H(x)}{d}\right) & \text{if } x = y. \end{cases}$$

For example, the matrix for $d = 2$ looks like (rows and columns indexed by $00, 01, 10, 11$ in that order)

$$\begin{bmatrix} 1-\theta & \frac{\theta}{2} & \frac{\theta}{2} & 0 \\ \frac{1}{2} & \frac{1-\theta}{2} & 0 & \frac{\theta}{2} \\ \frac{1}{2} & 0 & \frac{1-\theta}{2} & \frac{\theta}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}.$$

Notice that everything in sight here is invariant under permuting the coordinates – in particular, S_d acts on C_2^d and we have $M(x,y) = M(x^\sigma, y^\sigma)$ for any $\sigma \in S_d$. Thus we can reduce to the orbit chain where we only report the number of ones $H(x)$. The resulting chain on $\{0, 1, \cdots, d\}$ is then

$$\tilde{M}(i,j) = \begin{cases} \frac{i}{d} & \text{if } j = i-1, \\ \left(1 - \frac{i}{d}\right)\theta & \text{if } j = i+1, \\ \left(1 - \frac{i}{d}\right)(1-\theta) & \text{if } j = i. \end{cases}$$

(The orbit chain for $d = 2$ on $\{0,1,2\}$ would then be $\begin{bmatrix} 1-\theta & \theta & 0 \\ \frac{1}{2} & \frac{1-\theta}{2} & \frac{\theta}{2} \\ 0 & 1 & 0 \end{bmatrix}$.) The stationary distribution of this lumped chain is binomial; that is,

$$\tilde{\pi}(i) = \binom{d}{i}\theta^i(1+\theta)^{-d}.$$

The point is that we do get a miracle in this special case:

Theorem 106

The lumped chain $\tilde{M}(i,j)$ has eigenvalues

$$\beta_i = 1 - \frac{i}{d}(1+\theta), \quad 0 \le i \le d,$$

and the corresponding eigenfunctions are the **Krawtchouk polynomials** $P_i(j)$, which are the orthogonal polynomials for the binomial distribution: P_i is the degree-i polynomial

$$P_i(j) = \left(\theta^i \binom{d}{i}\right)^{-1/2} \sum_{k=0}^{i} (-1)^k \binom{j}{k}\binom{d-j}{i-k}\theta^{i-k}.$$

We know everything about these polynomials – in particular, they look like Hermite polynomials for large d. So we have an expression for bounding ℓ^1 by ℓ^2 and it works great here:

Theorem 107

For $0 < \theta < 1$ fixed, start the Metropolized $M(x,y)$ chain at $x = 0$. Then there is some function $f(\theta, c)$ such that for $k = \frac{d}{2(1+\theta)}\log(\theta d) + c$,

$$||M_0^k - \pi_\theta||_{\mathrm{TV}} \le f(\theta, c),$$

where f decreases to 0 as $c \to \infty$ and increases to 1 as $c \to -\infty$. And there turns out to be a matching lower bound so that we actually get cutoff.

(We can see the details for this in the report "Eigen Analysis for Some Examples of the Metropolis Algorithm" by Professor Diaconis and Hanlon, and it's worth noting that it's also possible to do the analysis on the full "unlumped" chain.)

We'll now do a more complicated "surprising" example:

Example 108

Consider the **Ewens sampling measure** on S_n, which is a nonuniform distribution on permutations which is used in genetics and biology to predict how far back in the phylogenetic tree things started from. Let $d(\pi, \sigma)$ be the minimum number of transpositions (not necessarily adjacent) required to bring π to σ; this is called **Cayley distance** and it is left-invariant and right-invariant, so that for example if we want to measure the difference between two sets of rankings it doesn't depend on the order that we list the data. Cayley found that in fact

$$d(\pi, \sigma) = n - C(\pi\sigma^{-1}),$$

where $C(\sigma)$ is the number of cycles in the permutation σ. We can thus define the **Ewens measure**

$$\pi_\theta(\sigma) = Z^{-1}(\theta)\theta^{d(\sigma, \mathrm{id})}$$

(we can use a different "central permutation" instead of the identity too), where the normalizing constant is $Z(\sigma) = \prod_{i=1}^{n}(1 + \theta(i-1))$.

We don't actually need Markov chains or the Metropolis algorithm to sample from this measure. But if we change the metric in any other way, we don't actually know how to sample directly without Metropolis, so let's consider this

special case by using K the random transpositions chain. So our "base chain" is

$$K(\sigma,\tau)=\frac{1}{\binom{n}{2}} \quad \text{if } \tau=\sigma\circ(i,j)$$

(again, we don't have to worry about holding), and now our Metropolized version will want to move closer to the identity. So for $n=3$, with the rows and columns indexed by id, (12), (13), (23), (123), (132), our matrix will look like

$$\begin{bmatrix} 1-\theta & \frac{\theta}{3} & \frac{\theta}{3} & \frac{\theta}{3} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3}(1-\theta) & 0 & 0 & \frac{\theta}{3} & \frac{\theta}{3} \\ \frac{1}{3} & 0 & \frac{2}{3}(1-\theta) & 0 & \frac{\theta}{3} & \frac{\theta}{3} \\ \frac{1}{3} & 0 & 0 & \frac{2}{3}(1-\theta) & \frac{\theta}{3} & \frac{\theta}{3} \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \end{bmatrix},$$

with stationary distribution $\pi(\sigma)\sim\begin{bmatrix}1 & \theta & \theta & \theta & \theta^2 & \theta^2\end{bmatrix}$ (we can check it is a left eigenvector of eigenvalue 1). So it's a bit of a mess, but because the chain is invariant on both sides it's invariant under conjugation, and so $M(\sigma,\tau)=M(\sigma^\eta,\tau^\eta)$ where $\sigma^\eta=\eta^{-1}\sigma\eta$. Thus we can lump to conjugacy classes, which gives us a 3×3 chain (rows and columns indexed by cycle types / partitions $1^3, 12, 3$)

$$\begin{bmatrix} 1-\theta & \theta & 0 \\ \frac{1}{3} & \frac{2}{3}(1-\theta) & \frac{2}{3}\theta \\ 0 & 1 & 0 \end{bmatrix}.$$

It turns out we can diagonalize this chain, and to describe the eigenvalues and eigenfunctions we need some symmetric function theory. Consider a polynomial $f(x_1,\cdots,x_k)$ which is symmetric in its k arguments (meaning that $f(x_1,\cdots,x_k)=f(x_{\sigma(1)},\cdots,x_{\sigma(k)})$ for all $\sigma\in S_k$). Examples include the powersums $p_\ell(x_1,\cdots,x_k)=\sum_{i=1}^k x_i^\ell$, and then correspondingly for any partition $\lambda=(\lambda_1,\cdots,\lambda_r)$ the polynomials

$$p_\lambda(x_1,\cdots,x_k)=\prod_{i\geq 1} p_i^{a_i} \text{ if } \lambda \text{ has } a_i \text{ parts of size } i.$$

Then there is a "fundamental theorem of symmetric functions" which states that the powersums $\{p_\lambda : \lambda \text{ a partition of } n\}$ form a basis of the space of degree-n symmetric polynomials.

There are also other families of symmetric polynomials: we have the elementary symmetric functions

$$e_1=\sum_i x_i, \quad e_2=\sum_{i<j} x_i x_j, \quad \cdots$$

and so from the e_ns we can get the e_λs like we got p_λs from the p_ns. There are lots of other bases too and we know how to get between them.

But the basis that we need for our problem is the **Jack symmetric functions** $j_\lambda^\alpha(x_1,\cdots,x_k)$, which are really objects coming from statistics. We won't give the explicit definition of them here, other than to say that $0\leq\alpha\leq 1$ is some fixed parameter. This used to be a standard topic in a "multivariate analysis" course (which no longer exists), and the Wishart distribution (coming from estimated covariance matrices), which is a measure on symmetric positive definite matrices, has orthogonal polynomials which correspond to the $\alpha=\frac{1}{2}$ case.

The point is that we have the Jack polynomials and also the powersums, and we can write down the change-of-basis

$$j_\lambda^\alpha=\sum_{\mu \text{ partition of } n} C(\lambda,\mu)p_\mu(x).$$

For example, we have for $n = 3$ that

$$j_{1^3} = p_1^3 - 3p_{12},$$

$$j_{12} = p_1^3 + (\alpha - 1)p_{12} - \alpha p_3,$$

$$j_3 = p_1^3 - 3\alpha p_{12} + 2\alpha^2 p_3.$$

Theorem 109

For $0 < \theta \leq 1$, the lumped Markov chain $\tilde{M}^\theta(\lambda, \mu)$ on partitions has eigenvalues

$$\beta_\lambda = (1 - \theta) + \frac{\theta(n(\lambda^T) + n(\lambda))}{\binom{n}{2}},$$

where λ^T is the transpose of λ and $n(\lambda) = \sum_{i=1}^n (i-1)\lambda_i$. The corresponding eigenvectors for β_λ is the set of change-of-basis coefficients $C_\lambda(\cdot)$.

We can thus use this to get rates of convergence and cutoff:

Theorem 110

Started at the identity, the Metropolized chain converges to the Ewens distribution after $k = \frac{1}{2}\max(\frac{1}{\theta}, 1)n(\log n + c)$ steps.

For example, $\theta = 1$ is random transpositions and we recover the $\frac{1}{2}n \log n$ fact from before.

Fact 111

The cyclic group C_n has a similar story: we know that C_n is generated by simple random walk with steps $\{\pm 1\}$, and if we instead make $\pi_\theta(j) = \frac{\theta^{\text{dist}(0,\,j)}}{Z}$. The ordinary eigenfunctions are exponentials or cosines, and if we Metropolize we get the Chebyshev polynomials instead.

Everything we've said "should work for complex reflection groups" as well, and it's likely true that if we take the usual minimal generating sets and close those up into conjugacy classes like random transpositions, and then use that to define a length function on the group, a similar story is likely to be true.

Remark 112. *Everything we've done "starts from the peak," and that's really because we've been analyzing the lumped chain and the peak is (in our examples) in its own orbit when lumped – for example, the conjugacy class containing the identity element is just that one element. If we wanted to start from some other state on the unlumped chain, we wouldn't be able to just use the eigenanalysis on the lumped chain, so we would need to be able to diagonalize the full chain instead. This has been done for examples like the hypercube and cycle, but not transpositions.*

Other examples where "a miracle occurs" can be found in papers that Professor Diaconis wrote with Arun Ram (using things like the Iwahori–Hecke algebra); things really are nice in those cases.

Example 113

To conclude, remember that "what we're doing with the Metropolis algorithm" is that we have $\pi(x)$ on some finite state space, we're given K with some stationary distribution $\sigma(x)$, and we modify K to a Markov chain M so that we can estimate, for example,

$$\mu(f) = \sum_x f(x)\pi(x) = \mathbb{E}_\pi[f(x)].$$

Running Metropolis to generate $X_1, X_2, \cdots, X_n$ from π (supposing that we start from stationarity) would then estimate this by

$$\hat{\mu}_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i),$$

and this is an unbiased, consistent, asymptotically normal estimator of $\mu(f)$. But there's another way to do this estimation using Markov chains as well, called **importance sampling**: we run the original chain $K(x, y)$ to generate $Y_1, \cdots, Y_N$, and then do an appropriate weighting to adjust:

$$\hat{\mu}_{n,\text{IMP}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\pi(Y_i)}{\sigma(Y_i)} f(Y_i).$$

And again, if we start from stationarity, then this is also unbiased, consistent, and asymptotically normal.

So we can either estimate via K and then adjust, or we can sample from K with extra coin flips; somehow these are really quite similar. And so by now we have enough examples to try to figure out which is better, and that's done in the paper "Examples comparing importance sampling and the Metropolis algorithm" by Bassetti and Professor Diaconis. It turns out that "nobody won," but most of the time Metropolis was comparable or even better than importance sampling.

Remark 114. *We've said a lot of nice things about the Metropolis algorithm, and it's also "nice math." But if we take the original example of the hard spheres problem (picking a sphere at random and trying to move it a little bit), we can ask whether we can prove anything about its rates of convergence. In papers with Lebeau, Professor Diaconis found that if we have N disks of radius h, and we move disks within h of where they are, we have*

$$||M^k - Unif_{TV}|| \leq Ae^{-kh^2},$$

and unfortunately here $A = d^{d^2}$ for d the dimension of the problem (so $2N$ here). So even using techniques like microlocal analysis, we're not really getting anything particularly close to the truth. The problem is that even if we try doing this in one dimension and with only a single disk, the spectrum is continuous – there are some eigenvalues and eigenvectors and then a continuous spectrum on top of it. So doing any kind of analysis really is quite torturous.

13 May 11, 2026

This course is somehow "about the geometry of Markov chains," and today we'll discuss the **Gibbs sampler** (also called **Glauber dynamics**), which is historically the second most widely-used algorithm in a lot of areas.

Example 115

Our setting will be the following: let $\mathfrak{X}, \mathfrak{Y}$ be two measurable spaces with σ-finite measures μ, ν respectively, so that $\mu \times \nu$ is the product measure on $\mathfrak{X} \times \mathfrak{Y}$ (think of this as Lebesgue measure). Let $f(x, y)$ be a probability density with respect to this product measure on $\mathfrak{X} \times \mathfrak{Y}$; our goal is to sample from f.

Let the marginals of $f(x, y)$ on $\mathfrak{X}$ and $\mathfrak{Y}$ be

$$m_1(x) = \int_{\mathfrak{Y}} f(x, y)\nu(dy), \quad m_2(y) = \int_{\mathfrak{X}} f(x, y)\mu(dx).$$

Suppose for simplicity that m_1, m_2 are positive everywhere, just so we don't have to worry about more notation. Then we have conditional densities

$$f(x|y) = \frac{f(x, y)}{m_2(y)}, \quad f(y|x) = \frac{f(x, y)}{m_1(x)}.$$

Definition 116

The **Gibbs sampler** is a Markov chain which proposes to sample from $f(x, y)$ in the following way. From a state (x, y), draw the second coordinate y' from $f(\cdot|x)$, and then draw the first coordinate x' from $f(\cdot|y')$; our new state is (x', y').

Thus, our Markov chain has density with respect to $\mu \times \nu$ given by

$$K(x, y; x', y') = f(x'|y')f(y'|x),$$

and after n steps the density is given by

$$K^n(x, y; x', y') = \int_{\mathfrak{X}\times\mathfrak{Y}} K^{n-1}(x, y; w, z)K(w, z; x', y')\mu(dw)\nu(dz).$$

Such a chain also gives rise to an operator K acting on the function space $L^2(f) = L^2_f$ (the square-integrable functions with respect to $f(x, y)$).

Example 117

Let $\mathfrak{X} = \{0, 1, 2, \cdots\}$ with the counting measure μ, and let $\mathfrak{Y} = [0, \infty)$ with Lebesgue measure. Recall that the Poisson(y) distribution has probability mass $f(x|y) = \frac{e^{-y}y^x}{x!}$. Letting e^{-y} be the usual density of a standard exponential random variable (this is the **conjugate prior** in statistics language), we get the joint density

$$f(x, y) = \frac{e^{-2y}y^x}{x!}.$$

We can compute the marginals explicitly here: it turns out $m_1(x) = \frac{1}{2^{x+1}}$ is the **geometric** distribution, and we know $m_2(y) = e^{-y}$ is the **exponential**. Similarly, we can compute the conditional distributions: we know $f(x|y) = \frac{e^{-y}y^x}{x!}$ is the **Poisson**, and it turns out $f(y|x) = \frac{2^{x+1}e^{-2y}y^x}{x!}$, which is the **gamma** density on y with shape parameter $x + 1$ and scale paramter $\frac{1}{2}$.

So the Gibbs sampler asks us to go back and forth between sampling a gamma density and a Poisson random variable, and this gets us a Markov chain. And under mild conditions (which we'll talk about), this converges to f.

Remark 118. *There's nothing special about using two coordinates; we can always use n different coordinates and sequentially sample one coordinate at a time conditional on all other $(n-1)$ coordinates.*

Fact 119

This description of the Gibbs sampler always sounded to Professor Diaconis like "iterated projections" in the functional analysis world. For a reminder of that story, let $\mathcal{H}$ be a Hilbert space, and let $\mathcal{M}_1, \mathcal{M}_2$ be closed subspaces of $\mathcal{H}$ and $\mathcal{M}_I = \mathcal{M}_1 \cap \mathcal{M}_2$. It's a fact that there is always a unique vector in a closed subspace which is the orthogonal projection; thus the iterated projections algorithm is to project onto $\mathcal{M}_1$, then project the result onto $\mathcal{M}_2$, and so on, which converges to an element of $\mathcal{M}_I$.

Such an algorithm turns out to always work:

Theorem 120 (von Neumann)

With the notation above, let P_1, P_2 be orthogonal projections onto $\mathcal{M}_1, \mathcal{M}_2$, and let $\mathcal{P}_I$ be the orthogonal projection onto $\mathcal{M}_I$. Then for the linear operator $T = P_2 P_1$, we have weak convergence $T^n \to \mathcal{P}_I$, meaning that for every $h \in \mathcal{H}$ we have $T^n h \to \mathcal{P}_I h$ pointwise.

There is a quantitative version of this as well:

Theorem 121 (Aronszajn)

Again with notation as above, define the "cosine of the angle between the subspaces"

$$c = \sup\left\{\langle v_1, v_2\rangle : v_i \in \mathcal{M}_i \cap (\mathcal{M}_1 \cap \mathcal{M}_2)^\perp, \quad |v_i| = 1\right\}.$$

Then for any vector h, we have exponential convergence:

$$||(P_2 P_1)^n h - \mathcal{P}_I h|| \leq c^{2n-1}||h||.$$

We'll do a similar analysis now in our context, and the next result can be explained by the following demonstration in one dimension. Suppose we have two paperclips on a string, and we do the following thing repeatedly: take the left endpoint of the string and fold it over with the right paperclip, then move the left paperclip tight (as far out as it can go). Then take the right endpoint of the string and fold it over with the left paperclip, then move the right paperclip tight. Intuitively, the paperclips will eventually move to $\frac{1}{3}$ and $\frac{2}{3}$ if we iterate this enough times

Mathematically, suppose the distances are x, y, z. Then folding with the right endpoint turns (x, y, z) into $(x, \frac{y+z}{2}, \frac{y+z}{2})$, and then folding over with the left endpoint further turns it into $\left(\frac{2x+y+z}{4}, \frac{2x+y+z}{4}, \frac{y+z}{2}\right)$. So in $\mathbb{R}^3$, what we're doing in one step is given by the matrices

$$P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \quad P_2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \implies P_1 P_2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

The eigenvalues of this matrix are 1 (for the vector $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$), $\frac{1}{4}$ (for $\begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$) and 0 (for the vector $\begin{bmatrix} 0 & 1 \\ -1 \end{bmatrix}$). So because we can write

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \frac{x+y+z}{3}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \left(\frac{2}{3}x - \frac{y}{3} - \frac{z}{3}\right)\begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} + (y-z)\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix},$$

and applying $(P_2P_1)^n$ would turn this into

$$(P_2P_1)^n \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \frac{x+y+z}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \left(\frac{1}{4}\right)^n \left(\frac{2}{3}x - \frac{y}{3} - \frac{z}{3}\right) \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

and thus we indeed converge exponentially fast to the state $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. The measure-space version of this is a nice result:

Theorem 122 (Burkholder)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $H = L^2(\mathbb{P})$. Consider the σ-algebras $\mathcal{A}_1, \mathcal{A}_2$ "generated by the two projections," so that $\mathcal{M}_1 = L^2(\Omega, \mathcal{A}_1, \mathbb{P})$ and $\mathcal{M}_2 = L^2(\Omega, \mathcal{A}_2, \mathbb{P})$. For any $U \in L^2(\mathbb{P})$, the orthogonal projection of U onto $\mathcal{M}_i$ is then given by the conditional expectation $\mathbb{E}[U|\mathcal{A}_i]$. (This is often how we prove that such a projection even exists.)

Let $U_0 = U$ and define $U_{2i+1} = \mathbb{E}[U_{2i}|\mathcal{A}_1]$ and $U_{2i+2} = \mathbb{E}[U_{2i+2}|\mathcal{A}_2]$. Then we have almost-sure convergence of the random variables $U_n \to \mathbb{E}[U|\overline{\mathcal{A}_1} \cap \overline{\mathcal{A}_2}]$ if and only if $U \in L \log L$ (meaning that $\int |U| \log(1 + |U|) d\mathbb{P} < \infty$).

The converse here is strong in the sense that if we are given a function $U \in L^1$ but not in $L \log L$, then there exist some sub-σ-algebras such that the iterated projection does not converge. Burkholder was interested in whether there was a sufficient statistic for two σ-algebras at the same time – it was only known that conditioning on a single one gives a sufficient statistic (this is Rao–Blackwell). The proof is pretty involved (30 pages in the Annals of Probability), so we won't go through it here.

To unpack what's happening in our "string example," let Ω be $\{1, 2, 3\}$, $\mathcal{F}$ be all subsets, and $\mathbb{P}(j) = \frac{1}{3}$. Then let U be the function which assigns x, y, z to $1, 2, 3$, and define the σ-algebras $\mathcal{A}_1 = \sigma(\{1, 2\}, \{3\})$ and $\mathcal{A}_2 = \sigma(\{1\}, \{2, 3\})$. Then $\mathbb{E}[U|\mathcal{A}_1]$ is a function of j, and it assigns $\frac{x+y}{2}, \frac{x+y}{2}, z$ to $1, 2, 3$ (since it must not distinguish between the first two coordinates). Similarly $\mathbb{E}[U|\mathcal{A}_2]$ is a function of j which assigns x, $\frac{y+z}{2}$, $\frac{y+z}{2}$ to $1, 2, 3$, and the intersection $\mathcal{A}_1 \cap \mathcal{A}_2$ is trivial and the conditional expectation is $\frac{1}{3}$ for all j.

Remark 123. *This may all seem like it's overcomplicating things, but this next comment shows that "it's not quite so obvious." Suppose now that we have three σ-algebras on $(\Omega, \mathcal{F}, \mathbb{P})$, called $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$, and we do the obvious thing where we project sequentially. It was open for 50 years whether updating $1, 2, 3, 1, 2, 3 \cdots$ actually made things converge. It was known that symmetrizing, so $1, 2, 3, 3, 2, 1, \cdots$ was okay, but the sequential one required non-commutative spectral analysis. And of course, the $L \log L$ condition is funny so there's something to understand there.*

Returning to the Gibbs sampler, suppose our probability measure is $\mathbb{P}(dx, dy) = f(x, y)(\mu \times \nu)(dx dy)$. Our Hilbert space now is $\mathcal{H} = L^2(f)$, and we can let our projection subspaces be

$$\mathcal{M}_1 = L^2_f(\sigma(\mathfrak{Y})), \quad \mathcal{M}_2 = L^2_f(\sigma(\mathfrak{X})$$

be the spaces just generated by single coordinates. Letting P_1, P_2 be the corresponding orthogonal projections, we know that our Markov chain is $K = P_1P_2$. We'll now try to get something useful about rates of convergence using Aronszajn's theorem, which requires computing angles between subspaces (via **maximal correlation** and eigenvalues).

Remark 124. *Notice that P_1P_2 on K is not reversible, and one way we can make it reversible is to do "random scan" instead (that is, use $K = \frac{1}{2}P_1 + \frac{1}{2}P_2$ to pick a coordinate at random and resample); another is $P_1P_2P_1$. So the problem is that P_1P_2 itself may not have real eigenvalues.*

From earlier, we have $c = \sup\left\{\langle v_1, v_2\rangle : v_i \in \mathcal{M}_i \cap (\mathcal{M}_1 \cap \mathcal{M}_2)^\perp, \quad |v_i| = 1\right\}$. To understand "what correlation actually is," we know that we can have uncorrelated random variables which are very related (for example Z and Z^2 for Z standard normal), and there's a related notion which may be more useful for situations like this:

Definition 125

The maximal correlation between two σ-algebras $\mathcal{A}_1, \mathcal{A}_2$ is

$$\gamma(\mathcal{A}_1, \mathcal{A}_2) = \sup\left\{\mathbb{E}[X_1 X_2] : X_i \in \mathcal{A}_i, \quad \mathbb{E}[X_i] = 0, \quad \mathbb{E}[X_i^2] = 1\right\}.$$

Similarly, the maximal correlation between two random variables

$$\gamma(X, Y) = \sup_{f,g}\left\{\mathbb{E}[f(X)g(Y)] : \mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0, \quad \mathrm{Var}(f(X)) = \mathrm{Var}(g(Y)) = 0\right\}.$$

(The latter definition turns out to actually be useful in real statistical contexts.)

Next, we can consider $Q = P_1 P_2 P_1$ (which is an operator on L^2_f). This is a self-adjoint contraction, and we can let β_1 be the second-largest eigenvalue (the largest is 1 because constants are preserved); the minimax characterization yields

$$\beta_1 = \sup_Q\left\{\langle Qg, g\rangle : g \in \mathcal{M}_1, \quad \mathbb{E}[g] = 0, \quad \mathrm{Var}(g) = 1.\right\}$$

(Because $P_1 P_2 P_1$ starts by resampling the first coordinate, everything here depends only on y.)

Theorem 126

Assume that we have trivial intersection between the spaces, meaning that $\overline{\mathcal{A}_1} \cap \overline{\mathcal{A}_2} = \{0, \Omega\}$. Then the three notions we've described above are equivalent:

$$c = \gamma(\mathcal{A}_1, \mathcal{A}_2) = \sqrt{\beta_1}.$$

A reference for this whole lecture, but in particular, this theorem, can be found in Diaconis, Khare, and Saloff-Coste's "Stochastic Alternating Projections." The main point to get across here is the geometric picture of the Gibbs sampler: it really is just iterated projections. So Aronszajn's theorem tells us that we have a weak rate of convergence in L^2 with rate given by this constant, which is something.

Remark 127. *We've been mentioning these "closures" $\mathcal{A}_i$ which were not required in von Neumann because everything was already closed. But a good example to keep in mind is the following: suppose Ω is the unit square $[0,1]^2$, $\mathbb{P}$ is uniform on the diagonal $x_2 = x_1$, and $\mathcal{A}_i$ are the σ-algebras generated by the coordinate projections (so $\mathcal{A}_i = \sigma(X_i)$). Then $\mathcal{A}_1 \cap \mathcal{A}_2$ is trivial, but it is not true that repeatedly projecting will converge because you will just stay where you are. Instead, we have to add in the nullsets to get $\overline{\mathcal{A}_1}$ and $\overline{\mathcal{A}_2}$, which adds anything off of the diagonal – it turns out $\mathcal{A}_1 \cap \mathcal{A}_2$ becomes all Borel sets, so the theorem does not apply. That's why for example requiring $f(x, y)$ to be positive everywhere would fix the issue, for example.*

Fact 128

Now that we have Burkholder's theorem and the von Neumann setup, it's worth pointing out that those two settings are very different. In particular, not all closed subspaces are the range of $\mathbb{E}[X|\mathcal{A}_i]$ for some sub-σ-algebra.

For example, let $\mathcal{H} = L^2([-\pi, \pi])$, and consider the subspace of functions f such that $f(\omega) = 0$ for all ω in some positive-measure subset S. This is indeed a linear space, but it doesn't contain the nonzero constant functions, and the range of a conditional expectation should indeed at least contain those.

Similarly, let $\mathcal{M}'$ be the set of all functions f whose Fourier series coefficients $\hat{f}(k)$ are all zero for $k < 0$. But if $X \geq 0$, then $\mathbb{E}[X|\mathcal{M}'] \geq 0$ as well, and functions with nonnegative Fourier coefficients will not satisfy that.

It turns out that a necessary and sufficient condition is that $\mathcal{M} \subseteq L^2(\Omega, \mathcal{A}, \mathbb{P})$ is in the range of some conditional expectation if and only if $\mathcal{M}$ contains the constants, and also $f \in \mathcal{M}$ implies that the positive part $f_+ \in \mathcal{M}$. So Burkholder only applies to a tiny subset of problems, but we do get strong results in those cases!

14 May 13, 2026

Last time, we talked about the Gibbs sampler in a rather abstract manner. Today will be much more practical, but one thing we'll need (and which is worth saying) is the following.

Fact 129

We know that a function of a Markov chain $f(X_n)$ isn't always a Markov chain – for example, if we have simple random walk on the circle and we just report $+$ or $-$ depending on which half of the circle we're on, then the future depends on the past beyond just the present. And in the setting of the Gibbs sampler on $\mathfrak{X}_1 \times \mathfrak{X}_2 \times \mathfrak{X}_3$, where we have some density $f(x_1, x_2, x_3)$ and resample the coordinates one at a time conditional on the others, we get a process (X_1^n, X_2^n, X_3^n). Then X_1^n is not actually Markov.

But if we only have two coordinates and consider the Gibbs sampler on $\mathfrak{X} \times \mathfrak{Y}$, then each coordinate is indeed Markov.

If we write things down for discrete spaces for simplicity, we can directly compute the $\mathfrak{X}$ transition to be

$$K(x, x') = \sum_y f(y|x) f(x'|y).$$

But this is indeed a Markov transition kernel – it says that for any x, we sample y, then sample x', and that will be our new state. And the same is true for the $\mathfrak{Y}$ chain, and what's important is that **if the $\mathfrak{X}$ chain is close to random after ℓ steps, so is the full bivariate chain on $\mathfrak{X} \times \mathfrak{Y}$ after $\ell + 1$ steps**. Indeed, if x was exactly random, then we sample y from the correct conditional distribution and that exactly gets us the desired density:

Theorem 130

We have the bound on chi-square distance to stationarity

$$\chi_x^2(\ell) \leq \chi_{x,y}^2(\ell) \leq \chi_x^2(\ell - 1),$$

where the chi-square in the middle is for the bivariate chain, and the others are for the $\mathfrak{X}$ chain only.

Example 131

Consider, as we did last lecture in Example 117, the Poisson density $\frac{e^{-y}y^x}{x!}$ of parameter y, and suppose y is standard exponential. Thus we want to run the Gibbs sampler on the density $f(x, y) = \frac{e^{-2y}y^x}{x!}$ on $\mathfrak{X} \times \mathfrak{Y}$.

If we just think about what this chain looks like on the $\mathfrak{X}$-marginal, we have $f(x|y) = \frac{e^{-y}y^x}{x!}$ and $f(y|x) = \frac{2^{x+1}e^{-2y}y^x}{x!}$

(just by Bayes' rule). Thus our $\mathfrak{X}$ chain has transition matrix

$$K(x,x') = \int \frac{e^{-y}y^{x'}}{x'!} \cdot \frac{2^{x+1}e^{-2y}y^x}{x!} dy$$
$$= \frac{2^{x+1}}{3^{x+x'+1}} \binom{x+x'}{x},$$

and its stationary distribution is the marginal distribution on $\mathfrak{X}$ (the geometric distribution $m_1(x) = \frac{1}{2^{x+1}}$). Notice in particular that this is reversible, since

$$m_1(x)K(x,x') = \frac{\binom{x+x'}{x}}{3^{x+x'+1}}$$

which is indeed symmetric in x, x'. (There's also a similar $\mathfrak{Y}$ chain which just lives on $(0,\infty)$ instead.)

Theorem 132

We have the following:

1. The $\mathfrak{X}$ chain has eigenvalues $\frac{1}{2^j}$ for all nonnegative integers j.
2. The (right) eigenvectors are the orthogonal polynomials for the geometric distribution, which are called the **Meixner polynomials**.
3. The chi-square distance to stationarity when started from x after ℓ steps satisfies

$$\chi^2_x(\ell) \leq e^{-2c} \quad \text{if} \quad \ell = \log_2(1+x) + c \text{ for } c > 0.$$

In the other direction, we also have

$$\chi^2_x(\ell) > 2^{2c} \quad \text{if} \quad \ell = \log_2(x-1) - c \text{ for } c > 0.$$

So we have very sharp bounds here, and in fact we know how to control distances with explicit constants from any given starting point. So again "a miracle occurred" because we could diagonalize the operator, but it's really not such a miracle because there are many other examples where such a thing also happens – see "Gibbs Sampling, Exponential Families and Orthogonal Polynomials" by Professor Diaconis, Khare, and Saloff-Coste for more exposition and examples. Basically, we can take any standard exponential family (like how we mentioned uniform and binomial on the first day of the course) and a similar diagonalization story happens with the corresponding orthogonal polynomials.

In general, the Gibbs sampler with two coordinates is described as "choose x' from $f(\cdot|y)$, then y' from $f(\cdot|x')$." But there's of course the question of "how do we actually sample from the conditional" if we can't do it from the joint distribution to begin with. Older statisticians would say that we can use Metropolis with some proposal chain, and that combination has its own name in the literature as "Metropolis on Gibbs." But really, we can "do anything" in the following sense:

Theorem 133

Let $K^1_y(x,x')$ be any Markov chain on $\mathfrak{X}$ with stationary distribution $f(\cdot|y)$, and let $K^2_x(y,y')$ similarly be any Markov chain on $\mathfrak{Y}$ with stationary distribution $f(\cdot|x)$. Then perform the following steps: pick x' by running one step of $K^1_y(x,x')$, then y' by running one step of $K^2_x(y,y')$. Such a procedure always has the joint density $f(x,y)$ as its stationary distribution.

Proof. This is a completely abstract general result, but let's do the discrete case. Let $\tilde{K}(x,y;x,y') = K^1_y(x,x')K^2_{x'}(y,y')$

be the kernel of this chain. Then we can bravely check that for any fixed x', y',

$$\begin{aligned}
\sum_{x,y} f(x,y)K(x,y;x',y') &= \sum_{x,y} f(x,y)K_y^1(x,x')K_{x'}^2(y,y') \\
&= \sum_y K_{x'}^2(y,y')m_1(y)\sum_x \frac{f(x,y)}{m_1(y)}K_y^1(x,x') \\
&= \sum_y K_{x'}^2(y,y')m_1(y)f(x'|y) \\
&= \sum_y K_{x'}^2(y,y')f(x',y) \\
&= m_2(x')\sum_y K_{x'}^2(y,y')f(y|x') \\
&= m_2(x')f(y'|x') \\
&= f(x',y'),
\end{aligned}$$

as desired. □

This all works for more than just two coordinates as well, and so we just need any chain at all to run the Gibbs sampler.

Fact 134

Unfortunately, as a "reality check," most Markov chains being run in practice (or which people care about) are quite different from these toy models we've talked about: they're very high-dimensional, and we don't have any way of exactly solving them. And to some extent, we can really only prove things in these relatively toy examples.

However, there are two areas where people have worked hard and gotten useful results for real chains, namely in **statistical mechanics** for (fairly general) Ising and Potts models and also in certain areas of **statistics**. But it would take us multiple lectures to go through any of those examples properly, and the techniques may not apply to other models. We'll talk a bit about one example in each area.

Example 135

In the statistics domain, James Hobert has lots of papers, and one particular recommendation from Professor Diaconis is "Honest Exploration of Intractable Probability Distributions Via Markov Chain Monte Carlo" by Jones and Hobert. This is an expository (and very readable) paper about "Harris recurrence-type techniques."

Today, we'll instead talk in more detail about a recent paper, "Wasserstein-based methods for convergence complexity analysis of MCMC with applications," by Hobert and Qin.

First, we need to set up the statistics problem where these models come from – the topic is **probit analysis**. The setup is that we are given some data $(y_1, x_1), \cdots, (y_n, x_n)$ where each $y_i \in \{0, 1\}$ (we can think of this as a binary result at the end of an experiment) and $x_i \in \mathbb{R}^p$ are some vectors of covariates (for example properties of the test subjects). We wish to use this data to predict the effect of those covariates on the result; that is, the function $\mathbb{P}(Y = 1|X)$.

In **probit analysis**, we assume this function takes the specific form

$$f(x) = \Phi(x \cdot \beta)$$

for some $\beta \in \mathbb{R}^p$, where $\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-t^2/2}dt$ is the (one-dimensional) cdf of the standard normal. So what

people typically do is fit the best possible β they can from the data, and then they use that to see which coefficients of our covariates are important.

In the Bayesian version of this, we begin with some prior on β, and the usual one we use is the Gaussian

$$\omega(\beta) \propto \exp\left(-\frac{1}{2}(\beta-\nu)^T Q(\beta-\nu)\right),$$

where ν is some prior mean and Q is the prior covariance (we have to specify these as inputs). So this specifies a model – we don't know β, but we think this model is what generated the data, and so Bayes' theorem says that our posterior (given the data and the prior) is

$$\pi(\beta|X,Y) \propto \exp\left(-\frac{1}{2}(\beta-\nu)^T Q(\beta-\nu)\right) \prod_{i=1}^{n} \Phi(x_i\cdot\beta)^{y_i}(1-\Phi(x_i\cdot\beta))^{1-y_i},$$

and so **our goal is**, given all of our data, **to sample from this measure** on $\mathbb{R}^p$. (Perhaps we have $n = 500$ data points and $p = 20$ covariates in a typical example.)

The algorithm everyone uses is not quite the Gibbs sampler, but it's very close – it goes by the name of **data augmentation** or **auxiliary variables**, and specifically we'll consider the **Albert–Chib algorithm**. We have an $n \times p$ real matrix of covariates X given (called the **design matrix**, since sometimes in an experiment we get to choose them). Let $\Sigma = X^T X + Q$ (this is symmetric and positive definite if Q is). For $\mu \in \mathbb{R}$, $\tau > 0$, and $a \in \{0,1\}$, the **truncated normal distribution** is

$$\mathrm{TN}(\mu,\tau^2,a) \sim N(\mu,\tau^2) \text{ truncated on } \begin{cases} (-\infty,0) & \text{if } a=0, \\ (0,\infty) & \text{if } a=1. \end{cases}$$

Our Markov chain will thus be the following two-step process, started from some $\beta \in \mathbb{R}^p$:

1. Choose $\{z_i\}_{i=1}^n$ from $\mathrm{TN}(x_i^T\beta, 1, y_i)$; write $Z = (z_1, \cdots, z_n)$.
2. Then draw $\beta' \in \mathbb{R}^p$ from the p-variate normal $N(\Sigma^{-1}(X^T Z + Q\nu), \Sigma^{-1})$.

We'll do auxiliary variables in more generality next week, but this is one such example actually used in practice that we can keep in mind.

Theorem 136

The Markov chain listed above is reversible with $\pi(\cdot|X,Y)$ as its stationary distribution.

Thus we can draw 1000 βs from this distribution using the Markov chain and try to understand what they look like, and perhaps we can see that this is a practical solution to a practical example. Analyzing this Markov chain is what Qin and Hobert set out to do in their paper.

To understand that analysis, we need to explain two other concepts.

Definition 137

The **Wasserstein distance** is a different metric on probability measures which is much weaker than total variation (the latter of which is usually "asking for too much" in a lot of high-dimensional problems). For $(\mathfrak{X}, d)$ a separable metric space, let p, q be Borel probability measures on $\mathfrak{X}$. The **Wasserstein distance** is then

$$d_W(p,q) = \sup_{f \in \mathrm{Lip}_1} \{\mathbb{E}_p[f] - \mathbb{E}_q[f]\},$$

where Lip_1 is the set of Lipschitz functions with respect to the metric (meaning that $d(f(x), f(y)) \le d(x,y)$).

Basically total variation lets us use a function f which is 1 on an arbitrary set and 0 on the rest, but the Lipschitz condition forces a more restrictive class of functions.

Theorem 138

We can also write the Wasserstein distance as the optimal coupling of the two measures

$$d_W(p,q) = \inf_{(X,Y):X\sim p, Y\sim q} \mathbb{E}[d(X,Y)].$$

So finding any function gets us a lower bound on Wasserstein distance, and finding any coupling gets us an upper bound. The idea is that this metrizes weak* convergence, in that

$$\left\{ p_n \overset{\text{weak}^*}{\to} p_\infty \text{ and } \mathbb{E}[d(X_n, x_0)] \to \mathbb{E}[d(X_\infty, x_0)] \right\} \quad \text{if and only if} \quad d_W(p_n, p_\infty) \to 0.$$

For the Albert–Chib algorithm, our distance in our metric space depends on the data we're given; specifically we have the norm $||x|| = x^T \Sigma x$, which then gives us a metric d_Σ. (And in fact, one of the points of the paper is that we can go from Wasserstein distance to total variation in these problems.) For more about this particular topic, we can see Dudley's book "Real Analysis and Probability."

The other aside is **iterated random functions**, and the best place to look for this is Professor Diaconis' paper with Freedman of the same name.

Example 139

Let $(\mathfrak{X}, d)$ be a metric space, and suppose we have a one-parameter family of functions $f_\theta(x) : \mathfrak{X} \to \mathfrak{X}$ for $0 \le \theta \le 1$. From this family, we get a recipe for making a Markov chain: choose $\theta_1, \theta_2, \cdots$ iid uniformly on $[0,1]$, and then start our chain at some x_0 and successively apply f_{θ_1}, f_{θ_2}, f_{θ_3} and so on via $X_n = f_{\theta_n}(X_{n-1})$.

The point is that all Markov chains can be represented in this form if we specify the functions in the right way, since θ is just our source of randomness.

Theorem 140

Suppose our functions f_θ are contractions "on average," meaning that f shrinks distances on the state space $\mathfrak{X}$. (This is actually a bit subtle, so we'll say it more carefully next time.) Then the associated Markov chain represented in this way has a unique stationary distribution π, and we have exponential convergence to π.

So the argument basically is to take the Albert–Chib algorithm, find a representation as a Lipschitz contraction, and then get bounds on the contraction constants. So nothing in that argument uses eigenvalues and it's a different perspective on how to get rates of convergence from what we've seen so far! (But it really is a nightmare, and if we wanted to answer the question of "how long do we need to run the chain for $n = 500$ and $p = 20$ and given data before we have distance-to-stationarity below $\frac{1}{100}$," we probably wouldn't get any answer within our lifetimes.)

15 May 18, 2026

We did a "complicated real example" of Bayesian pro-bit regression to see the Gibbs sampler in a high-dimensional case. Specifically, we explained the back-and-forth algorithm used to sample, and we mentioned that we can get some bound on mixing time which ends up really not being useful in practice. Professor Diaconis actually wrote the authors of the

paper, and they managed to get a fairly reasonable answer (say 8000 steps for $n = 500$, $p = 200$). But before we talk about that more, we'll first discuss **hit-and-run** (also **auxiliary variables**, **data augmentation**, and the **Burnside process**) in a more abstract setting. The idea is that these four ideas look like different algorithms when they really are basically all the same.

Lots of the content in this lecture comes from the paper "Hit and run as a unifying device" by Anderson and Professor Diaconis.

Example 141 (Hit-and-run)
Let $f(x)$ be a probability density on $\mathbb{R}^d$ (possibly unnormalized), which we want to sample from. To do so, run the following algorithm. From $x \in \mathbb{R}^d$, first choose z uniformly on the surface of the unit ball $B_1(x)$. (This is easy to do: let $Z = (Z_1, \cdots, Z_d)$ be iid normal standard normals and define $z = x + \frac{Z}{\sum_{i=1}^n Z_i^2}$.) Letting ℓ_{xz} be the line in $\mathbb{R}^d$ through x and z, we can then restrict the density f to ℓ_{xz} and sample y according to the density. That point y is the result of one step of our Markov chain.

The picture to have is that we repeatedly choose a line through our current point and resample on that line. The tricky step here is to sample from ℓ_{xz}; sometimes we can just do it directly in some closed-form manner, but otherwise we can take one step from Metropolis or discretize the line into small pieces and sample from the resulting discrete distribution.

We can think of this as a "coordinate-free Gibbs sampler," since the Gibbs sampler is exactly this kind of algorithm except that we only consider lines that are parallel to one of the coordinate axes (so we resample from the conditional distribution on one of the d coordinates). But sometimes coordinates are not special and so this makes sense instead.

This algorithm has f as a stationary distribution (under very few regularity conditions). And it's useful because in one step, we can go a long way and so we're not limited in mixing time like we are in local algorithms. The idea is originally due to Turcin in the paper "On the computation of multidimensional integrals by the Monte Carlo Method."

In operations research, this algorithm is used to sample uniformly from a compact convex set. Indeed, any lines through points will intersect the convex set at a line segment, so it's easy to sample from the one-dimensional distributions here. And for convex problems (especially in high dimensions) lots of analysis has been done, and the paper "Hit-and-run is fast and fun" by Lovasz and Vempala covers lots of that.

At this point, there are lots of questions we may want to ask: for example, "why a ball," or "why uniform on the ball," or "why lines," or "what about discrete problems?". And that's the reason for generalizing beyond this particular Euclidean setting.

Example 142 (Abstract hit-and-run)

We'll work in the finite setting, but everything here is completely general. Let $\mathfrak{X}$ be a finite or countable set and $\pi(x)$ a positive probability distribution on $\mathfrak{X}$. We need the following three ingredients:

- Let $\{L_i\}_{i\in I}$ be a finite or countable family of lines, where each L_i is some (arbitrary) subset of $\mathfrak{X}$ and $\bigcup_{i\in I} L_i = \mathfrak{X}$. Let $I(x) = \{i \in I : x \in L_i\}$ be the subset corresponding to lines that contain x.
- For each $x \in \mathfrak{X}$, we specify some probability distribution $\omega_x(i)$ on $I(x)$, and suppose for clarity that $\omega_x(i) > 0$ for all $i \in I(x)$.
- For each $x \in \mathfrak{X}$ and each line L_i containing x, specify a Markov kernel $K_i(x, y)$ on L_i whose stationary distribution proportional to $\pi(y)\omega_y(i)$.

With these three ingredients, we can write out the total probability of going from x to y in one step of hit-and-run: first sample a line from $\omega_x(i)$, then take a step from $K_i(x, y)$, so that

$$K(x, y) = \sum_i \omega_x(i) K_i(x, y).$$

Proposition 143

The chain above has π as its stationary distribution.

Proof. By explicit computation, we have

$$\begin{aligned}\sum_x \pi(x)K(x, y) &= \sum_x \pi(x) \sum_i \omega_x(i)K_i(x, y)\\ &= \sum_i \sum_x (\pi(x)\omega_x(i))\, K_i(x, y)\\ &= \sum_i \pi(y)\omega_y(i)\\ &= \pi(y),\end{aligned}$$

where the third line comes from the definition of K_i's stationary distribution on the line L_i. □

Remark 144. *Notice that for our Markov kernel on the lines, we can always choose it to be $\frac{1}{Z}\pi(y)\omega_y(i)$ (independent of x) if we can directly sample from the correct stationary distribution. But any other Markov chain with that stationary distribution works.*

Remark 145. *It's sometimes true that $|I(x)| = k$ is independent of x, so that we can just choose $\omega_x(i) = \frac{1}{k}$ to be uniform among all lines. Then we're back to the random scan Gibbs sampler setting.*

Fact 146

If all chains $K_i(x, y)$ are reversible, so is K. Also, remember that this framework does not automatically guarantee ergodicity, so we still need to reason through why that holds using some more careful analysis.

For a useful example, Professor Diaconis realized we can really speed things in in the following setting:

Fact 147

Let $\mathfrak{X}$ be the set of $I \times J$ matrices with some fixed (given) row and column sums (these are called **contingency tables**). In statistical work, we often want to sample such tables and compare that to real data to see if the row and column categorizations are independent.

Suppose in particular that we want to sample from the uniform distribution over all such tables. Counting the number of such tables exactly is #P-complete, but we can use the following algorithm to run a Markov chain on the set. Pick i, i' and j, j'; suppose we have row sums r_i and $r_{i'}$ and column sums $c_j, c_{j'}$. Then we can choose a uniform table with a chain, and in one step we resample the (r_i, c_j) entry over all allowed values and then fill in the other three entries $(r_i, c_{j'})$, $(r_{i'}, c_j)$ and $(r_{i'}, c_{j'})$ deterministically given the row and column sums.

What was previously being done was to repeatedly pick rows and columns and do local changes of the form $\begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix}$ (which doesn't change the row and column sums). But of course we can make the steps bigger in hopes that makes things go faster.

We could then ask what happens if we use a three-way table instead of a two-way tables (for example, pick a pair of parallel planes and try to make the same kinds of updates). But using this set of moves, the graph isn't actually connected. The way to fix the chain is its own subject, and this and many examples are found in Professor Diaconis' paper "Algebraic algorithms for sampling from conditional distributions" with Sturmfel.

So abstractly we have a way of moving from a point to a line that contains it, and then we have a way of moving on the line via the Markov chain. Let's think about another perspective, suggested by physicists:

Example 148 (Auxiliary variables)

Again let $\pi(x)$ be a probability distribution on a finite set $\mathfrak{X}$. This time, let I be some (arbitrary) set of auxiliary variables, and suppose $\omega_x(i)$ is a probability measure on I for all $x \in \mathfrak{X}$. These define a joint distribution

$$f(x, i) = \pi(x)\omega_x(i),$$

which thus yields $f(x|i) = \frac{f(x,i)}{M(i)}$ for the normalizing constant $M(i) = \sum_y f(y, i)$. We then also need to specify, for all i and x, a Markov chain $K_i(x, y)$ on $\mathfrak{X}$ whose stationary distribution is $f(x|i)$. Then again we have

$$K(x, y) = \sum_i \omega_x(i) K_i(x, y).$$

We can check directly (by the same argument as before) that K has π as the stationary distribution. This is a useful abstraction of the Swendsen–Wang algorithm, and we'll see a real example of how it can be useful now:

Example 149

Suppose we have the family of probability measures

$$\pi_\beta(x) = Z^{-1}(\beta) \exp\left(\sum_{j=1}^{J} \beta_j T_j(x)\right)$$

for $\beta \in \mathbb{R}^T$ and T_i some statistics (functions). This is sometimes called an **exponential family.** Our goal is, for β fixed, to sample from π_β.

For this, take the auxiliary index set to be $I = (0, \infty)^J$. The idea will be to take i to be uniform over all vectors $(i_1, \cdots, i_J)$ satisfying $i_j \leq e^{\beta_j T_j(x)}$ for all $1 \leq j \leq J$. (This is easy to do, since we can just sample one coordinate at a time from intervals.) Well, that means the joint distribution (x, i) has density uniform on the set

$$\left\{(x, i) : 0 \leq i_j \leq e^{\beta_j T_j(x)} : 1 \leq j \leq J\right\},$$

so indeed this is exactly what we wanted to do to get the factor in $\pi_\beta(x)$ in our marginal over $\mathfrak{X}$! Thus our algorithm for sampling from $\pi_\beta(x)$ is to do the following:

- From x, choose i uniformly on $[0, \exp(\sum_{j=1}^J \beta_j T_j x)]$.
- From i, choose y uniformly on the set

$$\left\{y : 0 \leq i_j \leq e^{\beta_j T_j(y)} \text{ for } 1 \leq j \leq J\right\} = \left\{y : T_j(y) \geq \frac{\log(i_j)}{\beta_j} \text{ for } 1 \leq j \leq J\right\}.$$

  In general, if T_j is some complicated function, then this could be hard to do. But (this is why this discovery is being mentioned) in many models this is actually tractably invertible so that we can efficiently sample y.

Example 150

The Mallows model is a distribution on permutations which is peaked around some particular permutation (for example, if study participants are trying to rank tones or taste test different flavors of something, their answers will cluster near some center). In the Mallows ℓ_2 model, we have the probability distribution

$$\pi_\beta(\sigma) = Z^{-1}(\beta) \exp\left(-\beta \sum_{i=1}^n (\sigma(i) - \sigma_0(i))^2\right),$$

where $\beta \geq 0$ is a parameter and σ_0 is the "central permutation" to cluster around. (This is used for example to predict horse-racing results.)

For $\beta = 0$ this is just uniform, but for $\beta \to \infty$ we penalize permutations with σ and σ_0 far apart. We can ask various enumerative questions about permutations (for example, the number of cycles or the length of the longest increasing subsequence), but to get a more concrete answer it's usually helpful to just sample.

For this, we'll use auxiliary variables in exactly the way we mentioned above. For simplicity, suppose $\sigma_0(i) = i$ is the identity permutation to cut down on notation. Expanding out the square, we get that

$$\pi_\beta(\sigma) \propto \exp\left(2\beta i \sigma(i)\right)$$

since $\sum \sigma(i)^2$ and $\sum i^2$ are both constants and can get absorbed into the normalizing factor. But now that means we can define our statistics to be

$$T_i(\sigma) = 2i\sigma(i),$$

and then we fit into the exponential family framework above! We thus have n auxiliary variables for sampling our permutation, and the way we will do it is the following:

- Started from $\sigma \in S_n$, choose $(u_1, \cdots, u_n)$, where u_j are independent and (continuous) uniform on $[0, e^{2\beta j \sigma(j)}]$.
- Now given $(u_1, \cdots, u_n)$, pick τ such that $u_j \leq e^{2\beta j \tau(j)}$ holds for all j; that is, we require that

$$\tau(j) \geq \frac{\log(u_j)}{2\beta j}$$

for all j. But the point is that this is actually also easy to do: the set of all permutations such that $\tau(j) \geq b_j$ for some given constants b_j are nice combinatorially. What we do is look at all j such that $b_j \leq 1$ and choose a uniform such index in which to put the value 1. Then out of the remaining indices, choose a uniform index with $b_j \leq 2$ and put the value 2 there, and so on. Eventually we will have filled up all n values in all n points in the permutation.

We know in particular that there is some τ that always works, which is the σ we started with. So this algorithm does always give us some permutation and that is the τ we get after one step of the chain, and it even tells us how many such permutations will satisfy our conditions (because the number of total choices we have at each stage is always the same, and we just multiply them together).

So alternating between permutations and $(u_1, \cdots, u_n)$, we get a chain that lets us sample approximately from this Mallows measure. Of course we had to use the special algebraic structure here of "expanding out the squares" to extract our statistics T_j, but small variations of this strategy work if we replace $(\sigma(i) - \sigma_0(i))^2$ with $|\sigma(i) - \sigma_0(i)|^p$ for other p and also for the distribution $\pi_B(\sigma) = \exp(\mathrm{tr}(B\rho(\sigma)))$, where $\rho(\sigma)$ is the permutation matrix for σ and B is some given $n \times n$ matrix of parameters. (This is one of the examples in Michael Howes' thesis.) The point is that we really "can do auxiliary variables," and in various models we can really see that it does work well.

The main example of course is still the Swendsen–Weng model for sampling from Ising or Potts models in statistical physics, but we're not going over it here because it's written down clearly in many places. But if we want to see more examples of auxiliary variables, good references are the papers by Besag and Green and by Higdon (which are cited in the paper with Andersen).

16 May 20, 2026

We'll finish our discussion about auxiliary variables today. The way we've been setting this up is that we have some $\pi(x)$ on $\mathfrak{X}$, and we have some other auxiliary set I, and we go back and forth between the two by specifying probability measures $\omega_x(i)$ for each x and then Markov chains K_i with stationary distributions $f(x|i)$; then one step of the chain is given by

$$K(x, y) = \sum_{i \in I} \omega_x(i) K_i(x, y)$$

and this has π as stationary distribution. (And we saw how to use this to sample from the Mallows ℓ_2 model on permutations in an explicit way, even without knowing the normalizing constant Z – see Example 150.)

We'll talk a bit more about that example now, since there's still more to learn from it.

Fact 151

Constructing "exponential models of this sort" (perhaps using some other metric on the symmetric group) is done all the time, and the models are more sensitive than we might think. Specifically, understanding this measure π_β depends a lot on β; for example if we take $n = 52$ and $\beta = 1$ (which doesn't seem so pathological) and we try to sample from it, it doesn't actually move much.

It turns out that if $\beta \ll \frac{c}{n^2}$, then π_β is approximately uniform in the "permuton" sense, which is some sense of weak convergence. But if $\beta \gg \frac{c}{n^2}$, then π_β is well-approximated by a point mass at the centering permutation σ_0. And in the regime $\beta \sim \frac{c}{n^2}$, there is some nontrivial permuton limit for π_β as $n \to \infty$. (See Sumit Mukherjee's paper "Estimation in exponential families on permutations" for more exposition.) So we have to be careful about using the right scaling, and this kind of thing happens in all kinds of exponential models.

Remark 152. *Suppose now that we have fixed some β and we do want to sample from it without knowing about auxiliary variables. The first thing we might think of is to do Metropolis: if we're at some permutation, try to do a random transposition (or an adjacent transposition) and reject it with some probability if it moves us farther away from σ_0. In practice this is really what people always do, but it's pretty slow (something like n^3).*

On the other hand, auxiliary variables do a good job for our chain, and we can see Chenyang Zhong's paper "Mallows permutation models with L^1 and L^2 distances:"

Theorem 153 (Zhong)
Take $\beta = \frac{c}{n^2}$. Then $\log n$ steps are necessary and sufficient for auxiliary variables.

Remark 154. *On another note, these Mallows measures do turn out to be surprisingly useful. In "Mallows Model with Learned Distance Metrics: Sampling and Maximum Likelihood Estimation" by Alimohammadi and Asgari, the authors found real data with basketball teams or other rankings, and approximating it via Mallows*

$$\pi_{\beta,\sigma_0,p}(\sigma) = Z^{-1} \exp\left(-\beta \sum_{i=1}^{n} |\sigma(i) - \sigma_0(i)|^p\right)$$

can actually be done (while allowing us to tune all three of those parameters). And comparing to a much richer model of permutations, the Luce model, it turns out that Mallows does much better.

A few lectures ago, we also did an example of auxiliary variables in another real problem, specifically the Bayesian probit regression (Example 135). Recall that we wanted to approximate some data $(y_i, x_i) \in \{0,1\} \times \mathbb{R}^p$ via the model

$$\mathbb{P}(Y = 1|x) = \Phi\left(\beta^T x\right),$$

and we did so by starting with a Gaussian prior $\pi(\beta) \sim N(\nu, Q)$ and then updating via Bayes' theorem to get

$$\pi(\beta|X,Y) \propto \pi(\beta) \prod_{i=1}^{N} \Phi(X_i^T\beta)^{y_i} \left(1 - \Phi(X_i^T\beta)\right)^{1-y_i}$$

The Albert–Chib algorithm then has $\mathfrak{X} = \mathbb{R}^p$ and $I = \mathbb{R}^p$, and it's a Markov chain on the space of possible βs: from β, we choose $(Z_1, \cdots, Z_p)$ for Z_i truncated normals $\mathrm{TN}(x_i^T\beta, 1, y_i)$, and then we choose β' to be normal with some mean and covariance $N(\Sigma^{-1}(X^TZ + Q\nu), \Sigma^{-1})$, where $\Sigma = X^TX + Q$. All of this can be derived by just looking at the form of the density and see what auxiliary variables tells us to do.

Theorem 155 (Hobert–Qin)
For some $0 < \rho < 1$, we have the convergence in Wasserstein distance

$$d_W(K^\ell_{\beta_0}, \pi(\beta|X,Y)) \le C\rho^\ell$$

for some explicit $C > 0$ and $0 < \rho < 1$ (but pretty complicated and not very informative in practice).

What we're really curious about is, for example, how large ℓ needs to be to get $d_W \le \frac{1}{100}$. The answer that the authors gave Professor Diaconis was that if you use the proof rather than the result to get bounds, and we take $X_{n\times p}$ to have iid standard normal entries, $\nu = 0$, and $Q = I$, it's good enough to take 8000 steps for $n = 500$ and $p = 20$. And that's great, since we can really do this in practice. (Though in practice for serious simulations where we care about the answers, we might take something like $n = 5000$ and $p = 500$, and in general in high-dimensional problems people do really care about whether we can get real answers.)

Remark 156. *The connection of auxiliary variables to hit-and-run should be pretty clear to us; the lines L_i passing through points in our state space are exactly the auxiliary variables, and the transition matrix takes exactly the same form. So auxiliary variables contains hit-and-run, but also hit-and-run contains auxiliary variables by letting our lines be $L_i = \{x : \omega_x(i) > 0\}$. So the advantage here is really having the right perspective through some instructive examples, and the examples and even proof techniques look different in different applications.*

Example 157

For one more "big topic" in the same world of stories as auxiliary variables, we'll learn about **data augmentation**, which is an area where we have missing data in a dataset and want to fill it in. This is an issue in any real statistics problem (for example if we want to measure records for clinical trials), and we need to be able to do something about it without throwing it away.

We'll just do one simple example, and we'll see that it's the same as hit-and-run. Let $X_1, \cdots, X_n$ be the results of n balls dropped into k boxes via a multinomial distribution, where $\mathbb{P}(X = j) = \theta_j$ for $\theta_j \geq 0$ and $\sum \theta_j = 1$. In a Bayesian perspective, we observe that X_is and want to estimate θ starting from some prior distribution π on the simplex Δ_k. However, suppose that instead of observing the X_is, we have some set-partition σ^i of the cells $\{1, \cdots, k\}$ for each i, and we only get to observe which set-partition X_i lands in (not the actual value).

Let $Y_1, \cdots, Y_n$ be our actual observations (the names of the blocks that the X_is are in), and our goal is to sample from $\pi(\theta|Y_1, \cdots, Y_n)$. This would be a mess if we had to try to write everything down explicitly, but data augmentation saves us here:

- Start at some vector θ^0 (for example, the midpoint of the simplex or $(1, 0, \cdots, 0)$). For all $1 \leq i \leq n$, sample X_i from θ^0, conditioned on whichever block Y_i of σ^i we land in. (So for example if Y_1 is the block $\{2, 5, 9\}$, then $X_1 = 2, 5, 9$ with probabilities $\frac{\theta_2^0}{\theta_2^0+\theta_5^0+\theta_9^0}$, $\frac{\theta_5^0}{\theta_2^0+\theta_5^0+\theta_9^0}$, $\frac{\theta_9^0}{\theta_2^0+\theta_5^0+\theta_9^0}$.
- Now we have complete (augmented) data $(X_1, \cdots, X_n)$, and we can sample $\theta^{(1)}$ from the posterior

$$\pi(\theta(X_1, \cdots, X_n)) \propto \prod_{i=1}^{k} \theta_k^{n_j} \pi(\theta)$$

  for $n_j = \#\{X_i = j\}$, which is usually easy to do if we choose the right prior.

This algorithm converges to $\pi(\theta|Y_1, \cdots, Y_n)$ and is extremely widely used in real data, since it's an easy way to get going and give us useful information. And this is the most simple case of data augmentation; for more (history and examples) we can see the survey paper "The Art of Data Augmentation" by van Dyk and Meng.

Of course, there is a clear link to auxiliary variables here by letting our state space $\mathfrak{X}$ be the simplex Δ_k, the desired distribution be $\pi(\theta|Y_1, \cdots, Y_n)$, and the auxiliary variables be the actual cell values $(X_1, \cdots, X_n)$ (with $X_i \in Y_i$). And remember that we don't even need to know how to sample from the posterior exactly; we can just do Metropolis or any other Markov chain.

But there really aren't papers being written about rates of convergence for data augmentation; we can take some specific case like this one and try to see whether we can prove anything about it. It's really a reasonable subject to work on!

We'll now slow down and "prove some things" about another smaller corner of auxiliary variables, the **Burnside process**. This was invented for computer science theory in a completely different setting, and it has its own stories.

Example 158

Our setting is that $\mathfrak{X}$ is a finite set and G is a finite group acting on it. Then $\mathfrak{X}$ will be split into disjoint orbits $\mathcal{O}_1 \cup \cdots \cup \mathcal{O}_k$, and in combinatorics (and many other fields) we may want to understand the orbits. Specifically, we might want to know how many of them there are, how large they are, "whether they have nice names," and "do they fit together into some kind of moduli space" (is it a partially ordered set, manifold, metric space, variety, etc.).

For example, we know from Cayley's formula that there are n^{n-2} labeled trees on n vertices. For $n = 4$, if we root our trees at 1, we have 6 ways of forming a path, 6 ways of having two descendants and then one more descendant of one of those, 3 ways of having one descendant and then a branch for the other two vertices, and 1 way to have three descendants. But the symmetric group S_{n-1} acts on trees by permuting the labels of all other vertices besides 1, and so the orbits in this case are unlabeled (rooted) trees; these are called **Pólya trees** and so in our case we have four of them for $n = 4$. But unlabeled trees are much harder to study; for example we don't even have a nice formula for how to enumerate them like Cayley's formula for labeled trees.

The subject of enumeration under symmetry is called **Pólya theory**; for example, benzene has six carbon and six hydrogen atoms, but hydrogens can be replaced with chlorines, and we might want to count the number of possible configurations of different molecules if we treat dihedral symmetry as the same. So counting labeled graphs of this sort up to symmetry is of some interest, and there are many more examples coming.

Definition 159

The **Burnside process** is a way of sampling a uniformly random orbit, and once we can do that we can use Monte Carlo methods to answer things about the orbits. This is a Markov chain on $\mathfrak{X}$ defined as follows:

- From $x \in \mathfrak{X}$, choose a uniform group element $s \in G$ which fixes x.
- From $s \in G$, choose a uniform $y \in \mathfrak{X}$ fixed by s.

Thus the chance of going from x to y is given by

$$K(x, y) = \frac{1}{|G_x|} \sum_{s \in G_x \cap G_y} \frac{1}{|\mathfrak{X}_s|}$$

where G_x is the set of group elements fixing x and $\mathfrak{X}_s$ is the set of elements fixed by s.

Theorem 160

The Markov chain $K(x, y)$ is reversible and has stationary distribution $\pi(x) = \frac{Z^{-1}}{|\mathcal{O}_x|}$, where $\mathcal{O}_x$ is the orbit containing x.

Proof. By the orbit-stabilizer theorem we have $|\mathcal{O}_x| = \frac{|G|}{|G_x|}$ (since the group acts transitively on any orbit), so

$$\begin{aligned}\pi(x)K(x, y) &= Z^{-1}\frac{|G_x|}{|G|} \cdot \frac{1}{|G_x|} \sum_{s \in G_x \cap G_y} \frac{1}{|\mathfrak{X}_s|} \\ &= \frac{Z^{-1}}{|G|} \sum_{s \in G_x \cap G_y} \frac{1}{|\mathfrak{X}_s|}\end{aligned}$$

and this expression is symmetric under x, y. □

In particular, this means that any orbit has an equal probability under π (and the normalizing constant Z is exactly the number of orbits). This chain is connected since we can always go to s the identity element, which then allows us to go to any other element in $\mathfrak{X}$. So running the chain and only reporting the orbit we're in will eventually get us something uniform.

Example 161

Take $\mathfrak{X} = C_2^n$ to be the binary n-tuples and $G = S_n$ to act via permutation of coordinates. Then

$$\mathfrak{X} = \mathcal{O}_0 \cup \mathcal{O}_1 \cup \cdots \cup \mathcal{O}_n$$

where $\mathcal{O}_i$ is the set of n-tuples with i ones (since we can always permute the ones to be in any location, but we can't change the count of them).

Thus running this Burnside process gives us a Markov chain which is uniform on $\{0, 1, \cdots, n\}$, which is of course silly because we already know how to sample from distribution. But it's still an interesting object of study, and in particular we can actually run the two steps of the Burnside process in this case:

- From $x \in \mathfrak{X}$, we pick a permutation which fixes it; this must permute among the zeros and the ones separately, so we sample $s \in S_k \times S_{n-k}$ for k the number of ones.
- Then from $s \in S_n$, we pick some y fixed by s. For this, we write s in cycle notation and y has to be constant on each cycle so that the cyclic shifts keep it the same; we then do coin flips on each cycle to install in its coordinates either 0 or 1 with probability $\frac{1}{2}$.

In this example, both steps are easy to do and we can write down an expression for the Markov kernel $K(x, y)$, but it's not always to do so (like in data augmentation or hit-and-run). These Burnside-type problems are nice math problems, but next time we'll mention three real examples where people actually want to use them too.

Remark 162. *The picture to have in mind for these Burnside chains is the following: form a bipartite graph between $\mathfrak{X}$ and G with an undirected edge between x and g if g fixes x. Then doing simple random walk on the graph and keeping track of every other step gives us exactly the Burnside process on $\mathfrak{X}$, and of course it also shows us that the group elements G also serve as the auxiliary variables or "lines."*

17 May 27, 2026

We'll "do some proofs" of a friendly type related to the Burnside process today, specifically three different proofs of the same theorem with potentially useful techniques.

Example 163

The chain we'll be discussing today is the **binary Burnside process** described in Example 161, in which we start with an n-tuple, choose a uniform permutation which fixes it, and then choose a uniform n-tuple fixed by it. (Recall that both of these steps are easy to do by decomposing the objects appropriately; see the end of last lecture.)

This process was introduced first by Mark Jerrum, who showed that $\sqrt{n}$ steps of the chain suffice for mixing (which isn't so bad compared to the 2^n total states in the state space). Then Aldous showed that $\log n$ steps suffice (which is exponentially better in some sense), and in fact we can further show that a bounded number of steps suffice for any

n (so the running time doesn't depend on n). This is maybe not so surprising given how "vigorous" the dynamics are (the moves are not so local), but we'll be much more careful as we go on.

Fact 164

The first observation to make is any of these Burnside-type chains **lump to orbits** $\mathcal{O}_i$, and in our case the orbits are labeled by the number of ones in x. Identifying the orbit $\mathcal{O}_i = \{x \in C_2^n : |x| = i\}$ with the integer i, we get an orbit chain $K(i,j)$ on $\{0, 1, \cdots, n\}$.

(It's not always true that a function of a Markov chain is Markov, but in this particular lumping it always is.) We then get a symmetric matrix and thus a uniform stationary distribution $\pi(i) = \frac{1}{n+1}$, and we'll begin by studying this orbit chain and seeing what it can tell us.

Example 165

Our first method is the **Doeblin technique**. Doeblin was a probabilist from the 1930s who invented coupling and many other things, and he proved the fundamental theorem of Markov chains with this strategy. Let's first state what it says for finite chains.

Theorem 166

Suppose that $\mathfrak{X}$ is any finite set and $K(x,y)$ is a Markov chain on $\mathfrak{X}$ with stationary distribution $\pi(x)$, not necessarily reversible. Suppose there is some $0 < c < 1$ such that we always have the **Doeblin condition**

$$K(x,y) \geq c\pi(y).$$

Then

$$||K_x^\ell - \pi||_{\text{TV}} \leq (1-c)^\ell.$$

Proof. This is an "auxiliary randomization" argument. We write K as a linear combination

$$K(x,y) = c\pi(y) + (1-c)\frac{K(x,y) - c\pi(y)}{1-c}.$$

Probabilistically, this means that $K(x,y)$ flips a coin with heads-probability c; if it comes up heads then we take a step from π, and if it comes up heads we take a step from some other Markov chain with stationary distribution π. But then once the coin has come up heads once, we will be at π forever; thus if T is the (random) first time the coin comes up heads, we must have

$$\mathbb{P}(X_k \in A | T \leq k) = \pi(A).$$

Thus by the law of total probability (everything here should have a subscript for "where we start the chain," but we'll omit that),

$$\begin{aligned}\mathbb{P}(X_k \in A) &= \mathbb{P}(X_k \in A | T \leq k)\mathbb{P}(T \leq k) + \mathbb{P}(X_k \in A | T \geq k)\mathbb{P}(T > k) \\ &= \pi(A)(1 - (1-c)^k) + \mathbb{P}(X_k \in A | T \geq k)(1-c)^k,\end{aligned}$$

which means that for any event A we have

$$\mathbb{P}(X_k \in A) - \pi(A) = (1-c)^k \left(\mathbb{P}(X_k \in A | T \geq k) - \pi(A)\right),$$

and the right-hand side is at most $(1-c)^k$. Taking the supremum over all A yields the desired result. □

Remark 167. *The Doeblin condition is actually very strong; for example a lot of random walks take a while to move throughout the state space. But we can instead require that $K^\ell(x,y) \geq c\pi(y)$, and then instead we get that*

$$||K_x^k - \pi||_{TV} \leq (1-c)^{\lfloor k/l \rfloor}.$$

So really this "can be used in theory" for any ergodic Markov chain.

Fact 168

Suppose we're in a general Polish space and $K(x,dy)$ is some Markov kernel with stationary distribution $\pi(dy)$. Then one version of the Doeblin condition says that if $K(x,O) \geq c\pi(O)$ for every open set O, then again we have $||K_x^\ell - \pi||_{\text{TV}} = (1-c)^\ell$. But again keep in mind that this is really a very strong condition.

Professor Diaconis was "brought up to believe" that the Doeblin condition is useless for the kinds of problems that we care about. For example, suppose we have simple random walk on the circle C_n (for n odd), with $K(j, j\pm 1) = \frac{1}{2}$. Then $\pi(j) = \frac{1}{n}$, but in order to use Remark 167 we would need to understand the walk. Indeed, if we try to use $\ell = \lfloor \frac{n}{2} \rfloor$ (the first time we actually have a positive probability of getting everywhere), we'd have to repeatedly go to the left or go to the right a bunch of times in a row, so that

$$K^{\lfloor n/2 \rfloor}\left(0, \left\lfloor \frac{n}{2} \right\rfloor\right) = \frac{2}{2^{\lfloor n/2 \rfloor}}.$$

This is exponentially small, so if we compare to $\frac{c}{n}$ we have to take c so small that the $(1-c)^{\lfloor k/l \rfloor}$ total variation bound is very uninformative! So compared to the right answer of n^2, things don't look good. (Of course, in this example we could have used ℓ a little bigger to avoid making c very small. But then to do that we really have to understand the chain better anyway.)

But really Doeblin techniques aren't completely useless, and we can see one good use of them now. Indeed, these hit-and-run algorithms go a long way in one step, and the first quantitative steps for proving rates of convergence for sampling from a compact convex set C in $\mathbb{R}^d$ uses these types of techniques to get an upper bound of order n^d steps. (Then Lovász and Vempala strengthened this to n^2 steps in any dimension with some log corrections depending on d.)

Example 169

Let's try the Doeblin bound for the binary Burnside chain lumped to $\{0, 1, \cdots, n\}$. The first thing we have to do is actually write down a closed-form expression for $K(i,j)$.

In other words, we take any n-tuple with i ones, and we want to know the probability of ending up at some n-tuple with j ones. It turns out to be a bit of combinatorics:

Fact 170

Write $\alpha_j^n = \frac{\binom{2j}{j}\binom{2(n-j)}{n-j}}{2^{2n}}$ for $0 \leq j \leq n$. (This is the **discrete arcsine distribution** on $\{0, 1, \cdots, n\}$; it takes larger values near 0 and n than in the middle. Chapter 3 of Feller volume 1 has many examples where this exact distribution comes up!) Then

$$K(0,j) = \alpha_j^n, \quad K(j,k) = \sum_\ell \alpha_\ell^j \alpha_{k-\ell}^{n-j}$$

(where the sum runs over the values where α makes sense).

Well, we know what the discrete arcsine looks like, and in particular we know by Stirling's formula and other computations that

$$K\left(0, \left\lfloor \frac{n}{2} \right\rfloor\right) \geq \frac{1}{\pi n}, \quad K(i,j) \geq K\left(0, \left\lfloor \frac{n}{2} \right\rfloor\right).$$

Thus we have the Doeblin condition with $\pi(j) = \frac{1}{n+1}$ and $c = \frac{1}{\pi} + o\left(\frac{1}{n}\right)$. This therefore tells us that

$$||K_x^\ell - \pi||_{\mathrm{TV}} \leq \left(1 - \frac{1}{\pi}\right)^\ell,$$

meaning that a bounded number of steps suffice for this lumped chain! Thinking about how this relates back to the original chain now, this therefore proves that if we start at the all-zeros state,

$$||K_0^\ell - \pi||_{\mathrm{TV}} \leq \left(1 - \frac{1}{\pi}\right)^\ell$$

(now π is the stationary distribution of the lifted chain on C_2^n), since the orbit $\mathcal{O}_0$ only contains the single all-zeros state $\vec{0}$, and then everything is invariant under permutations. Indeed, the probabilities of being at states with a fixed number of ones are all equal – specifically, $K^\ell(\vec{0}, y) = \frac{1}{\binom{n}{|y|}} K^\ell(0, |y|)$ – so

$$\begin{aligned} ||K_{\vec{0}}^\ell - \pi||_{\mathrm{TV}} &= \frac{1}{2} \sum_{y \in C_2^n} |K^\ell(\vec{0}, y) - \pi(y)| \\ &= \frac{1}{2} \sum_{j=0}^{n} \binom{n}{j} |K^\ell(\vec{0}, y_j) - \pi(y_j)| \end{aligned}$$

for y_j the fixed state with j ones followed by $(n-j)$ zeros, and now the summand is exactly $|K^\ell(0,j) - \pi(j)|$ because $\pi(y_j) = \frac{1}{\binom{n}{j}} \frac{1}{n+1}$.

So the point is that (remembering that we get to choose where we start when we run simulations, so we might as well start from the all-zeros state here) we at least get some quantitative guarantee this way, even if the bound doesn't work from all starting states.

Remark 171. *This same argument also "works" if we have the Burnside process for S_n acting on C_k^n for $k \geq 2$ (so n-tuples on a larger alphabet). With this technique, we end up with a bound $(1 - c_k)^\ell$, but unfortunately c_k is of order $\frac{1}{k!}$ and so things break down for example if $k = n$ (and it's the wrong answer).*

Instead, Aldous has an argument which instead yields, for any starting state,

$$||K_x^\ell - \pi||_{\mathrm{TV}} \leq n \left(1 - \frac{1}{k}\right)^\ell.$$

So if $k = 2$, this tells us that $\log n$ steps suffice when started from any state, and even if $k = n$ this answer isn't so bad. We'll do this argument later, since it's a different flavor of proof which is useful too.

Remark 172. *Before that, we can ask the question of "what happened to the Doeblin condition?". If we want to improve this statement, there's a notion of "drift-minorization," also related to "small sets" and "Harris recurrence." The idea is that Doeblin requires $K^\ell(x,y) \geq c\pi(y)$ for all x, and a* **small set** *$\mathcal{C}$ is such that $K^\ell(x, A) \geq \theta\pi(A)$ for all $x \in \mathcal{C}$ (we should think of this as a central part of our state space). Then we can ask for how long it takes to hit the small set, which is some standard probability problem. (If we hit $\mathcal{C}$ many times, then we're in business.) This is the most popular technique among statisticians for proving rates of convergence, and we might ask whether we can do anything like it in this problem.*

Example 173

We're now ready to move on to our second method, which is that **sometimes a miracle happens and we can diagonalize the chain**. A reference for this is Professor Diaconis' paper "Hahn polynomials and the Burnside process" with Chenyang Zhong.

In this case, the mathematics comes out pretty nicely, and we'll just state some of the main results:

Theorem 174

Let $K(i,j)$ be the lumped binary Burnside process on $\{0,1,\cdots,n\}$.

1. The nonzero eigenvalues of this chain are
$$\beta_k = \frac{\binom{2k}{k}^2}{2^{4k}}, \quad 0 \le k \le \left\lfloor \frac{n}{2} \right\rfloor$$
(in particular, these don't depend on n except for how many of them there are). The other half of the eigenvalues are zero.
2. The eigenfunction for β_k is the Chebyshev polynomial T_{2k} on $\{0,1,\cdots,n\}$ of degree $2k$ (the orthogonal polynomials for the uniform distribution, which we can obtain via Gram–Schmidt).

The nonzero eigenvalues start off $\beta_0 = 1, \beta_1 = \frac{1}{4}, \beta_2 = \frac{9}{64}, \cdots$, and we can use asymptotics to get bounds on rates of convergence:

Corollary 175

We have the bounds on the stationary distribution started from 0
$$\frac{1}{4}\left(\frac{1}{4}\right)^\ell \le ||K_0^\ell - \pi||_{\text{TV}} \le 4\left(\frac{1}{4}\right)^\ell$$
for any $n, \ell \ge 3$.

The main ingredient in this theorem is the **Cannings argument**. Basically, if we have a Markov chain $P(i,j)$ on $\{0,1,\cdots,n\}$ and stationary distribution $\pi(j)$, and call a run of the chain $x_0 = x, X_1, X_2, \cdots$. Then the condition we have to check is that for any positive integer m,

$$\mathbb{E}\left[X_1^m | X_0 = x\right] \text{ is a polynomial in } x \text{ of degree at most } m.$$

If this occurs and the leading coefficient is a_m, then we have a Markov operator which takes polynomials to polynomials without increasing the degree. In such a situation, if (P,π) is reversible, then the eigenfunctions are always the orthogonal polynomials for π, and the eigenvalues are exactly a_m.

For example, suppose that $\mathbb{E}[X_1|X_0 = x] = ax + b$. Then subtracting off an appropriate constant c yields $\mathbb{E}[(X_1 - c)|x] = \alpha(x - c)$, and α will then be the degree-1 eigenvalue.

Fact 176

So what we need to do in our example is compute, if we start at a state with j = ones, the moments for the number of ones after one step of the chain. But for example for the first moment we can condition on the permutation σ that we choose, so

$$\mathbb{E}[X_1|X_0 = k] = \mathbb{E}[\mathbb{E}[X|\sigma, X_0 = k]|X_0 = k],$$

and once we pick any permutation the number of ones is always $\frac{n}{2}$ in expectation because each cycle is independently chosen to be 0 or 1. So in this case the leading coefficient is zero and we get the eigenvalue 0 for degree-1. A more complicated calculation would get us the eigenvalue $\frac{1}{4}$ for degree-2 by computing the second moment, and so on.

18 June 1, 2026

We've seen two methods of proof for rates of convergence of the binary Burnside chain now, specifically Doeblin bounds and explicit diagonalization of the lumped chain. These arguments gave pretty sharp bounds on total variation distance, showing that a bounded number of steps suffice when started from the all-zeros state.

Example 177

We might also be curious about how to study the chain $K(x, y)$ for the "unlumped" chain on C_2^n. Often there isn't much of a difference between the answers for the lumped and unlumped chains, but in this case it does make a big difference. It turns out that in chi-square (ℓ^2) distance, order $\frac{n}{\log n}$ steps are necessary and sufficient.

We were previously studying the chain on $\{0, 1, \cdots, n\}$, but now if we consider the unlumped chain we now have to think about the chain on 2^n different states.

Theorem 178

We have the following for the binary Burnside chain:

1. If $x_0 = (0, \cdots, 0, 1, \cdots, 1)$ is the state with $\lfloor \frac{n}{2} \rfloor$ zeros and $\lceil \frac{n}{2} \rceil$ ones and $\ell < 0.1\frac{n}{\log n}$, then $\chi^2_{x_0}(\ell) \to \infty$ as $n \to \infty$.
2. On the other hand, the average chi-square distance $\chi^2_{\text{avg}}(\ell) = \sum_x \pi(x)\chi^2_x(\ell)$ tends to zero if $\ell > 10\frac{n}{\log n}$, and in fact the same is true for any specific state.

In particular, ℓ^1 and ℓ^2 bounds are often comparable for a lot of Markov chains, but in this case they're exponentially different. More results (and sharper bounds) can be found in the papers with Andrew Lin (me) and Arun Ram, "A curiously slowly mixing Markov chain" and "Schur–Weyl duality for diagonalizing a Markov chain on the hypercube." The math turns out to be interesting despite this being a rather simple example; for example, the eigenvalues for this chain are the same eigenvalues that appear in the lumped chain, but with higher multiplicities: $\beta_k = \frac{\binom{2k}{k}^2}{2^{4n}}$ appears with multiplicity $\binom{n}{2k}$ instead of multiplicity 1, and 0 appears with multiplicity 2^{n-1}. This then in particular allows for studying rates of convergence from a state like $(0, 0, \cdots, 0, 1)$.

So the point really is just that "sometimes analyzing a feature of a Markov chain (the orbit) isn't the same as analyzing the whole chain."

Example 179

We'll now turn to the third method, **coupling**, for getting rates of convergence for this chain. This method now works beyond the binary case; let's consider $\mathfrak{X} = C_k^n$ to be the set of n-tuples which take on one of k values, with S_n acting again via permutation of coordinates.

We'll describe the Aldous coupling, which is a nice argument which gets good bounds (and there's a lot of generalization that can be done too). The hope will be that these kinds of arguments will be helpful in other settings, since they haven't been worked out for other problems yet.

Theorem 180

For the Burnside process on C_k^n, from any starting state x, we have $||K_x^\ell - \pi||_{\text{TV}} \leq n\left(1 - \frac{1}{k}\right)^\ell$. In particular, order $k(\log n + c)$ steps are sufficient for mixing.

We'll need a lemma to prove this, which would be very useful if it could be abstracted to a more general setting:

Lemma 181

Let F_1, F_2 be finite sets, not necessarily disjoint. Then there exists a pair of permutations $(\sigma^1, \sigma^2) \in S_{F_1} \times S_{F_2}$ which are marginally uniform, meaning that $\mathbb{P}(\sigma^1 = \eta) = \frac{1}{|F_1|!}$ for all η and similar for σ^2, and a pair of labelings C_j^1, C_j^2 of the cycles of σ^1, σ^2, such that

$$C_j^1 \cap \{F_1 \cap F_2\} = C_j^2 \cap \{F_1 \cap F_2\}.$$

In words, we can take the permutations σ^1, σ^2 and break them up into cycles, and we label each one with numbers. Furthermore, we can make it so that the numbers agree on the intersection $F_1 \cap F_2$.

Proof. The main fact we need is the following: let $T \subseteq S$ be a subset of a finite set, and let σ be uniform over S_S. Now write σ as a product of cycles and cross out everything not in T. The result will then be a uniform permutation in S_T, written in cycle notation. For example, if $S = \{1, 2, \cdots, 9\}$ and $T = \{1, 2, \cdots, 5\}$ and $\sigma = (139)(2468)(75)$, then the permutation we end up with is $(13)(24)(5)$.

Using this fact, we can prove the lemma by using $S = F_1 \cup F_2$. We then choose a uniform permutation $\sigma \in S_S$, and then the restrictions to F_1 and F_2 are our desired σ^1, σ^2. And to get the labelings of the cycles, we just label the cycles in the original σ and project down. □

Proof of Theorem 180. We will use our lemma to build a coupling; that is, a bivariate chain on $(C_k^n)^2$. That is, we'll construct a chain $\overline{K}(x^1, x^2; y^1, y^2)$, where $x^i, y^j \in C_k^n$, such that marginally each coordinate evolves according to K:

$$\sum_{y^2} \overline{K}(x^1, x^2, y^1, y^2) = K(x^1, y^1) \quad \text{for any } x^2 \in C_k^n,$$

and the same for the second coordinate. We could of course just make our chains independent, but we will couple them in a way where they come together.

Recall that the Burnside process on C_k^n samples a permutation which fixes x (meaning it permutes each of the k groups among themselves), then breaks those permutations into disjoint cycles and installs one of the values $\{1, \cdots, k\}$ uniformly and independently on each cycle. Our coupling will be described as follows: for each $a \in C_k$, define the sets of coordinates equal to a

$$F^{1a} = \{i : x_i^1 = a\}, \quad F^{2a} = \{i : x_i^2 = a\}.$$

Now to choose our uniform permutations σ^1, σ^2 fixing x^1, x^2, we must choose σ^{1a} uniformly from F^{1a} and σ^{2a} uniformly from F^{2a}, and we couple them exactly as in the previous lemma for each a, independently for different a. To complete one step of the chain, we must label our cycles uniformly on $\{1, \cdots, k\}$ within each coordinate, but we're again allowed to couple between the two coordinates. So we'll pick iid labels $\alpha_j^a \in \{1, \cdots, k\}$ for all a, j, and then (using the cycle labeling from our lemma) if a cycle in F^{1a} or F^{2a} is labeled j, we install the value α_j^a for it. The result will be our new state $(y_1, y_2) \in (C_k^n)^2$.

The key point is that because the cycles agree on the intersection of the sets $F^{1a} \cap F^{2a}$, if $x_i^1 = x_i^2 = a$ for some coordinate i, then after one step we will still have $y_i^1 = y_i^2$, since they're being labeled by the same cycle label. And if $x_i^1 \neq x_i^2$, then after one step they have a probability $\frac{1}{k}$ of being the same (since they will be labeled by independent uniform cycle labels).

So now we can use the **coupling bound** to conclude: if we evolve $\overline{K}(x^1, x^2, y^1, y^2)$ and let T be the first time that $y^1 = y^2$, then for any starting state $x^1 = x$ (and having x^2 initially distributed as π),

$$||K_x^\ell - \pi||_{\text{TV}} \leq \mathbb{P}(T > \ell)$$

(here K is the original chain, not the bivariate one). And since each of our n coordinates has a probability at most $\left(1 - \frac{1}{k}\right)^\ell$ of not agreeing after ℓ steps, a union bound yields the result. □

Coupling is one of the main ways that modern probabilists analyze rates of convergence, and so it really is worth reading and learning about.

Remark 182. *It "must be possible" to generalize the shape of this argument for any other example, say for wreath products or $GL_n(\mathbb{F}_q)$. For example, Pólya would explain Pólya theory by asking how many ways there are to color the faces of ten dice with two colors, where rotating a die or swapping two of them counts as the same configuration. Thus the symmetry group is the wreath product $S_n \ltimes S_4^n$ (or we could consider other symmetry groups in place of S_4), and maybe a similar story can be told here.*

Fact 183

This "Aldous coupling" is understandable but not quite sharp; in particular if $k = 2$ it gives $\log n$ steps even if the starting state is the all-zeros state. On the other hand, in very recent work by Nestoridi and Sly, it's been shown that there are starting states x_0 where $\log n$ steps are needed in total variation, and in fact we have cutoff.

Remark 184. *It's interesting to compare these coupling arguments to the path arguments from earlier in the course. Eigenvalues and coupling seem different, and a lot of people have wondered if there's more of a connection. For example, Peter Matthews has a way to construct a coupling if we know all of the eigenvalues and eigenvectors, but at that point we could also just use ℓ^2 bounds and comparison (since there's no comparison theory for coupling).*

That concludes the "three proofs" for the binary Burnside process.

Example 185

The last thing we'll do in the course (today and next time) will be to do three serious (real) examples of the Burnside process where the Markov chain can really do something, convergence seems to occur, but we really can't prove anything. Those three examples will be Pólya trees, partitions of n, and matrices.

When Pólya began his work, he would start by explaining the difference between labeled rooted trees (Cayley) and unlabeled rooted trees (Pólya) – see the discussion below Example 158. As we described, S_{n-1} acts on the non-root

labels, and the orbits are tree shapes, and we want to know lots of things about these trees (how many of them are there for a fixed size n, what can we say about their statistics like tree height or width, degree distribution, and so on).

We do know asymptotics for these Pólya trees; while the number of Cayley trees of size n is exactly n^{n-2}, the number of Pólya trees is asymptotically

$$\frac{b\sqrt{e}}{2\sqrt{\pi}} n^{-3/2} \rho^{-n} \left(1 + O\left(\frac{1}{n}\right)\right),$$

where $b = 2.68\cdots$ and $\rho = 0.3383\cdots$ are known to many decimal places. (So we have exponential growth, but much fewer than Cayley trees – for $n = 10$ there are only 719 Pólya trees.) The literature on tree enumeration is quite established, and we can for example see Michael Drmota's book for a good reference.

For other questions about statistics of these trees, some more is known. For example, we might let $w_k(T)$ be the number of vertices of the tree at depth k from the root and study $\max_k w_k(T)$, the width of the tree, or $h(T)$, the maximum depth of any vertex in the tree. For things like this, there are elaborate developments proving various theorems for Cayley trees. But to do the same for Pólya trees, much more complicated theorems are needed, and two questions we might ask are "are the limit theorems that we get out of those results actually correct?" and "are they pretty accurate for $n = 100$ or 10^6; when do asymptotics kick in?".

To do this, we can run our Burnside process to generate random tree shapes. It looks empirically like it takes 20 steps when $n = 10^7$, and that lets us generate histograms for many of these statistics. In Professor Diaconis' paper with Laurent Bartholdi "An algorithm for uniform generation of unlabeled trees (Pólya trees), with an extension of Cayley's formula," lots of this data is shown. But here's an example of a result:

Theorem 186

The height $h(T_n)$ of a uniform Pólya tree T_n of size n (the maximum distance of a leaf in T_n to the root) satisfies

$$\frac{b\sqrt{e}}{2\sqrt{\pi}} \frac{H(T_n)}{\sqrt{n}} \to F(x)$$

for $F(x) = 1 - 2\prod_{k\geq 1}(-1)^{k-1} e^{-k^2x^2}$ (this is the limiting distribution for a Brownian excursion).

Running simulations of 1 million samples for trees of size $n = 1000$, we see that the general shape is correct but there is a systematic affine shift in some of these statistics (so something isn't quite right). But for some other statistics like degree distribution, the simulations do match the predicted values.

Next time, we'll explain "how we actually run such a Burnside process" – in particular, we need to pick an automorphism of the tree and a tree fixed by that automorphism, and there's some interesting mathematics there.

19 June 3, 2026

As promised last time, we'll describe three Burnside process algorithms for real problems where nothing can really be proved at the moment.

Example 187

We started discussing the first example of rooted tree shapes (Pólya trees) last time. In particular, we described the Burnside process on Cayley trees and what its simulations can do for comparing to limit theorems.

We suppressed discussing the actual algorithm last time, but that leads to interesting math as well and that's what we'll discuss now.

What we need to be able to do to run our algorithm is the following: our state space $\mathfrak{X}$ is the set of all n^{n-2} labeled trees, and our group $G = S_{n-1}$ acts on the non-root vertices by permuting labels. So we need to be able to do the following to run our Markov chain:

- From any tree $T \in \mathfrak{X}$, pick a uniform automorphism $G \in S_{n-1}$ which fixes it. (For example if a vertex has two leaves as children, then we can transpose those two labels.)
- From any $\sigma \in S_{n-1}$, pick a uniform tree in $\mathfrak{X}$ fixed by it.

For the first task, we're essentially asking to compute the automorphism group $\mathrm{Aut}(T) = \{\sigma : T^\sigma = T\}$. To do this, first remove the root, which breaks what's under it into a bunch of rooted labeled trees (rooted at the original root's children). Some of those might be isomorphic, so we can permute the isomorphic ones; this gives us a symmetric group factor. Then within each of the isomorphism classes we can repeat the same thing, and so we end up with an inductive wreath product expression for $\mathrm{Aut}(T)$. So then we have to pick an element of the group uniformly, which we do by picking appropriate generators.

Remark 188. *Code for this is available in Professor Diaconis' paper with Laurent Bartholdi. But there's programs which do this kind of thing online, and it does the computation even faster than the code in the paper! So it's always worth looking around for this kind of thing.*

The second step now is that given some permutation $\sigma \in S_{n-1}$, we must choose a uniform tree fixed by it.

To do this, first consider the simplest case $\sigma = \mathrm{id}$, which means we want to pick a uniform rooted labeled tree at random. One way is to use the proof of Cayley's formula via Prüfer codes: trees are in bijection with sequences $(a_1, \cdots, a_{n-2}) \in [n]^{n-2}$, where the procedure is to repeatedly remove the lowest labeled leaf and record down the label of its parent. Then we can recover the tree with an iterative procedure (so this is indeed a bijection). Thus because it's very easy to pick a uniform random sequence of this kind, we can indeed efficiently sample a uniform rooted labeled tree at random.

It turns out a similar story actually works in general:

Theorem 189

Fix any $\sigma \in S_n$ with $\sigma(1) = 1$, and suppose σ has λ_d cycles of degree d (meaning $\sum d\lambda_d = n$). Define

$$\mu_d = \sum_{\substack{e|d \\ e<d}} e\lambda_e.$$

Then the number of Cayley trees fixed by σ is

$$t_{n,\sigma} = \lambda_1^{\lambda_1 - 2} \prod_{k>2} f(\lambda_d, d, \mu_d), \quad f(m, x, y) = (mx + y)^{m-1} y.$$

So this is a refinement of Cayley's theorem; indeed if σ is the identity then $\lambda_1 = n$ and so $t_{n,\mathrm{id}} = n^{n-2}$. So interestingly, just trying to understand how to run the Burnside process often leads to new interesting mathematics!

The proof of this formula is bijective, and it's an analog of the Prüfer code argument (though a bit more complicated). Prüfer codes turn out to be useful for lots of things, and so it's probably true that there are more applications of this bijection too.

Fact 190

For example, if we have a tree, the degrees of the vertices of the tree can be read off from the Prüfer code (the degree of vertex i is exactly 1 plus the number of times i appears in the Prüfer code). So this implies that if T is a uniform Cayley tree, the joint distribution of the vertex degrees is the distribution of $(n_1+1, \cdots, n_n+1)$, where we drop $(n-2)$ balls in n boxes and n_i are the counts of the balls.

We know the answer to essentially any answer about placing balls in boxes, so this gives us lots of useful information about degrees on trees. For example, if $L(T)$ is the number of leaves of T, then we have $\mathbb{E}[\ell(T)] \sim \frac{n}{e}$ and $\mathrm{Var}(T) \sim \frac{N}{e}\left(1-\frac{1}{e}\right)$, and we have the central limit theorem

$$\mathbb{P}\left(\frac{\ell(T)-\frac{n}{e}}{\sqrt{\frac{n}{e}(1-\frac{1}{e})}} \le x\right) \to \Phi(x).$$

Similarly, we know that the maximal box count is around $\log n$, but we can't get a limit law without having oscillations because this is a maximum for discrete distributions. The point, though, is that we still get theorems out of this, and maybe we can do something similar for random trees invariant under σ using the same bijections. (And another project worth doing would be to do the same for phylogenetic trees, or some other family of trees of applied interest.)

Example 191

For our second example (a completely new topic), consider the **commuting graph process** on any finite group G. Here we have $\mathfrak{X} = G$, where the group acts on itself by conjugation (meaning $x^s = s^{-1}xs$). The orbits are then the conjugacy classes of the group, and the Burnside process then lets us choose a uniform conjugacy class of G at random.

In this case, "both halves" of the Burnside process are the same: from $x \in G$, pick some $s \in G$ such that $s^{-1}xs = x \implies xs = sx$. So we can think of having a graph on G where two elements are connected if they commute, and we perform simple random walk on this graph.

There are a lot of different groups for which we can try running this, but we might ask "who cares?". Well, Professor Diaconis was approached by "the spy services" for a way to generate thousands of permutations of size 10^6 efficiently, and he suggested to do this with the commuting graph walk on S_n. Indeed, the conjugacy classes are cycle types and thus indexed by partitions, and it turns out that doing the Burnside process is infinitely faster than the "standard" rejection algorithm which takes hours to generate a single partition.

To run this algorithm, we need the following:

Fact 192

Suppose $\sigma \in S_n$ has a_i cycles of length i, which is often written $\sigma \sim \prod i^{a_i}$. Then the centralizer (the set of all permutations commuting with σ) can be written

$$C_{S_n}(\sigma) = \prod_{i=1}^{n} S_{a_i} \ltimes C_i^{a_i}.$$

For example, if our permutation has 5 3-cycles, then we can permute those cycles among themselves, and then we can also cycle each cycle by any amount we want, and that doesn't change the permutation. So we can do that independently for each cycle type, and in particular it's very easy to choose an element from this centralizer (pick

independent cyclic elements and a permutation of size a_i, and install those things). That means the Burnside process is easy to run in this case.

In the same way as for trees, we can use this algorithm to pick a random partition and compare with data (counting the number of ones, for example, which has a known limit distribution) – see Professor Diaconis' paper with Michael Howes "Random sampling of contingency tables and partitions: Two practical examples of the Burnside process." And in every theorem tested (comparing the histograms to enumerative theory), everything was good after quite a small number of steps.

But there are some practical lessons to learn here as well:

- In this case, it turns out that running the chain directly on the orbit chain (partitions) is much faster than running the chain on the original chain (permutations).
- For any partition, we can take its transpose by swapping the role of rows and columns in its Ferrers diagram. This is a bijection, so it preserves the uniform distribution on partitions; it turns out that **alternating steps** between the Burnside process and transposition speeds up the algorithm amazingly (it empirically really takes 50 instead of 500 steps for something like $n = 10^8$). So deterministic steps of this kind are often a big deal, and that's its own subject.

Example 193

For another group where this commuting graph walk does something useful, consider the group of **unitriangular matrices** $U_n(q)$, which have 1s on the diagonal, elements of $\mathbb{F}_q$ above the diagonal, and 0 below the diagonal.

For example, if $n = 3$, this is the Heisenberg group with entries mod p; this is one of two groups of order p^3. And for general n, $U_n(q)$ is the Sylow p-subgroup for $\mathrm{GL}_n(\mathbb{F}_q)$.

We have $|U_n(q)| = q^{\binom{n}{2}}$, and this is somehow an analog of the symmetric group for p-groups (any finite p-group is a subgroup of this). So we might ask about the character theory for this group, and in particular that requires us to understand the conjugacy classes of $U_n(q)$. It turns out "no one will ever know," in the sense that if we had a good description of this then there's an unsolvable word problem that we're solving (way beyond P = NP). But there's a famous conjecture of Graham Higman that the number of characters (that is, the number of conjugacy classes) $k_n(q)$ is a polynomial in q. For example, for $n = 3$ and $q = p$ a prime, we know that

$$k_3(p) = p^2 + p - 1,$$

We can see John Thompson's paper trying to prove this conjecture, but there's no success yet. What is known is the following:

Theorem 194

For all q, n, we have

$$q^{n^2/12} \leq k_n(q) \leq q^{n^2/4}.$$

So one thing we can try to do is to run the Burnside process on this group; if we can sample randomly, then we can use that to estimate the number of elements in the set. To run the algorithm, we need, from any $M \in U_n(q)$, to choose some M' such that $MM' = M'M$. It turns out this is easy to do in this case – writing $N = M - I$ and $N' = M' - I$, we require that $NN' = N'N$, and it turns out this is a linear problem in N' given a fixed N. So we can use Gaussian elimination to solve the linear system and efficiently pick a uniform solution (the dimension of the solution

space will depend on the current choice of N) via $N' = \sum_i \varepsilon_i V_i$ for V_i a basis and ε_i iid uniform on $\mathbb{F}_q$. That means it is easy to run the Burnside process here.

We can see the details of how to use this to count the number of group orbits in Professor Diaconis' paper with Chenyang Zhong "Counting the number of group orbits by marrying the Burnside process with importance sampling." The group theory community knows the answer for the number of conjugacy classes of $k(U_n(F_2))$ up to $n = 16$, and when the simulations were run using the Burnside process they agreed quite well with the actual values. But this method allows us to extrapolate out beyond $n = 30$, and it suggests that some refinement of $q^{n^2/12}$ is the right answer here! (And a similar story for $p = 3$ also holds.)

Example 195

The last thing that we haven't described yet is how to actually do the counting. Specifically, suppose that $\mathfrak{X}$ is a finite set and we have a way of sampling uniformly from $\mathfrak{X}$, and our goal is to estimate $|\mathfrak{X}|$. There are many different approaches to this, and we can see the paper "Estimating the size of a set using cascading exclusion" (with Susan Holmes and Sourav Chatterjee) for many different cases.

The idea we'll use in this case is the following: suppose $\mathfrak{X} = X_0 \supseteq X_1 \supseteq \cdots \supseteq X_L = \{1\}$ is a nested decreasing sequence of subsets. Then we have

$$|\mathfrak{X}| = \prod_{i=0}^{L-1} \frac{|X_i|}{|X_{i+1}|},$$

and we can estimate each ratio in this expression by sampling uniformly from X_i and seeing what proportion of the points lie in X_{i+1} (and taking the reciprocal of that estimate). For this to work, we need each ratio to be not too large (or else the probability of landing in the set will be too small to estimate with random samples) and L to also be not too large (or else we have very large error terms). But in this case such a strategy did indeed work (using the **pattern subgroups** of $U_n(q)$), and it provably works in the sense that large deviations for the binomial distribution gives us exponentially small error.

Remark 196. *This type of estimation is the basic technique related to estimation of #-P complete problems in theoretical computer science, and Alistar Sinclair's book on this material is a good reference.*