

Lectures on Mixing Times

A Crossroad between Probability, Analysis and Geometry

Nathanél Berestycki
Cambridge University
N.Berestycki@statslab.cam.ac.uk

November 18, 2011

These lecture notes correspond to a graduate course I taught at Cambridge University in 2008. I would like to thank in particular Ismael Bailleul, Richard Pymar, and Perla Sousi for their contributions. The notes were extensively revised and extended to include representation theory in 2011 when I taught part of this material again at Ecole Polytechnique in Paris.

Apart from the first two lectures which are needed throughout the rest of the course, each lecture is designed to illustrate a particular technique, and may be read independently from the rest.

I would gratefully receive feedback, comments, errors and typos. These should be sent to the above email address.

Contents

| | | |
|----------|--|-----------|
| 1 | Coupling | 4 |
| 1.1 | Total Variation distance | 4 |
| 1.2 | Coupling | 5 |
| 1.3 | Example: Random to top shuffling. | 7 |
| 1.4 | Example: random walk on a hypercube. | 9 |
| 2 | Spectral Gap | 11 |
| 2.1 | Eigenvalue decomposition | 11 |
| 2.2 | Example: Random walk on the circle. | 14 |
| 3 | Dirichlet forms and Poincaré inequalities | 16 |
| 3.1 | Dirichlet forms | 16 |
| 3.2 | Variational characterization | 17 |
| 3.3 | Poincaré inequality and the path method | 18 |
| 3.4 | Example: random walk on the n -dog | 19 |
| 4 | Comparison techniques | 23 |
| 4.1 | Random walks on groups | 23 |
| 4.2 | Heat kernel, ℓ^2 distance and eigenvalues | 23 |
| 4.3 | Comparison techniques | 24 |

| | | |
|-----------|--|-----------|
| 4.4 | Example 1: random transpositions on a graph | 26 |
| 4.5 | Example 2: a random walk on S_n | 27 |
| 5 | Wilson's method | 29 |
| 5.1 | Statement of the result | 29 |
| 5.2 | Example: Random walk on a hypercube | 31 |
| 5.3 | Example: adjacent transpositions. | 32 |
| 6 | Nash inequalities. | 35 |
| 6.1 | Abstract result | 35 |
| 6.2 | Example | 37 |
| 7 | Evolving sets and martingales. | 39 |
| 7.1 | Definition and properties | 39 |
| 7.2 | Evolving sets as a randomized isoperimetric profile | 43 |
| 7.3 | The conductance profile | 46 |
| 8 | Coupling from the past: exact sampling. | 49 |
| 8.1 | The Ising model and the Glauber dynamics | 49 |
| 8.2 | Coupling from the past. | 51 |
| 9 | Representation Theory | 54 |
| 9.1 | Basic definitions and results | 54 |
| 9.2 | Characters | 55 |
| 9.3 | Fourier inversion | 57 |
| 9.4 | Applications to card shuffling: the Diaconis-Shahshahani theorem | 59 |
| 10 | Riffle shuffle | 60 |
| 10.1 | Lower bounds | 63 |
| 10.2 | Guessing the true upper-bound | 65 |
| 10.3 | Seven shuffles are enough: the Bayer-Diaconis result | 66 |

Introduction

Take a deck of $n = 52$ cards and shuffle it. It is intuitive that if you shuffle your deck sufficiently many times, the deck will be in an approximately random order. But how many is sufficiently many ?

In these notes we take a look at some of the mathematical aspects of card shuffling and, more generally, of the mixing times of Markov chains. We pay particular attention to the *cutoff phenomenon*, which says that convergence to the equilibrium distribution of a Markov chain tends to occur abruptly asymptotically as some parameter $n \rightarrow \infty$ (usually the size of the state space of the chain, or the number of cards if talking about a card shuffling problem). If this occurs, the time at which this convergence takes place is called the *mixing time*. proving or disproving the cutoff phenomenon is a major area of modern probability, and despite remarkable progress over the last 25 years since this phenomenon was discovered by Diaconis and Shahshahani and by Aldous, there are still only very few examples which are completely understood.

The techniques which have proved useful so far involve a number of traditionally disjoint areas of mathematics: these can be probabilistic techniques such as coupling or martingales and “evolving sets”, the study of eigenvalues and eigenfunctions, functional and geometric inequalities (Cheeger’s inequality, Poincaré and Nash inequalities), or even representation theory – sadly not in these notes at this point. Yet another set of connections is provided by the fact that many of the Markov chains for which we desire to estimate the mixing time are often of a statistical mechanics nature, such as the Glauber dynamics for the Ising model. For these, there is a spectacular method available devised by Propp and Wilson, called “coupling from the past”, which we will talk about at the end.

1 Coupling

1.1 Total Variation distance

To start with, we need a way to measure how far away from stationarity we are. This is done by introducing the total variation between two probability measure on the state-space of the chain. Let μ and ν be two probability measures on a space S . (For us, S is the state space of a given Markov chain).

Definition 1.1. *The total variation distance between μ and ν is*

$$d_{TV}(\mu, \nu) = \|\mu - \nu\| = \sup_{A \subset S} |\mu(A) - \nu(A)| \quad (1)$$

The total variation distance thus measures the maximal error made when approximating μ by ν to predict the probability of an event. An easy exercise shows that if S is discrete:

Lemma 1.1. *We have the following identities:*

$$\begin{aligned} \|\mu - \nu\| &= \frac{1}{2} \sum_{s \in S} (\mu(s) - \nu(s))^+ \\ &= \sum_{s \in S} |\mu(s) - \nu(s)| \end{aligned}$$

As a consequence of these identities, note that we have:

$$0 \leq d_{TV}(\mu, \nu) \leq 1.$$

That is, the maximal value that the total variation distance can take is 1.

The relevant definitions for a Markov chain on a finite state space are the following:

Definition 1.2. *Let P be a transition probability matrix on a finite state space S , and assume that the Markov chain associated with P is irreducible and aperiodic. Let $\pi(x)$ denote the stationary distribution of X , defined by*

$$\sum_x \pi(x)P(x, y) = \pi(y).$$

Define the distance function for all $t \geq 0$ by:

$$d(t) = \max_{x \in S} \|p^t(x, \cdot) - \pi(\cdot)\|$$

The classical theory of Markov chains tells us that $d(t) \rightarrow 0$ as $t \rightarrow \infty$. In fact, the Perron-Frobenius theorem tells us that, asymptotically as $t \rightarrow \infty$, the distance $d(t)$ decays exponentially fast, with a rate of decay control by the spectral gap of the chain:

Proposition 1.1. *Let λ be the maximal eigenvalue of P which is strictly smaller than 1. Then there exists a constant C such that*

$$d(t) \sim C\lambda^t, \quad t \rightarrow \infty.$$

A priori, this result seems to tell us everything we want about mixing times. Indeed, to make λ^t small it suffices to take t larger than $-1/\log \lambda$. In most chains where the state space is large, the value of λ is close to 1, i.e, $\lambda = 1 - s$ where s is the spectral gap. This tells us that we need to take $t > t_{\text{rel}} := 1/s$ to make $d(t)$ small. As we will see in further details later, this is indeed necessary in general, but far from sufficient - the reason being that C is unknown and depends generally on n , and that this asymptotic decay says nothing about the actual behaviour of the chain at time t_{rel} , only something about extremely large times.

The formal definition of a cutoff phenomenon is the following:

Definition 1.3. *Let X^n be a family of Markov chains. We say that there is (asymptotically) a cutoff phenomenon at τ_n if for every $\varepsilon > 0$,*

$$d((1 - \varepsilon)\tau_n) \rightarrow 1$$

but

$$d((1 + \varepsilon)\tau_n) \rightarrow 0.$$

τ_n is then called the mixing time.

Since $d(t)$ converges to 0 as $t \rightarrow \infty$, it always makes sense to define, for $0 < x < 1$:

$$\tau_n(x) = \inf\{t \geq 0 : d(t) \leq x\}$$

Then cutoff is equivalent to saying that $\tau_n(x) \sim \tau_n(y)$ for all $0 < x < y < 1$. In particular, one may always define

$$t_{\text{mix}} = \inf\{t \geq 0 : d(t) \leq 1/e\},$$

where the constant $1/e$ is arbitrary.

1.2 Coupling

The technique of coupling is one of the most powerful probabilistic tools to obtain quantitative estimates about mixing times. The basic observation is the following. Let μ, ν be two measures on a set S .

Definition 1.4. *A coupling of μ and ν is the realisation of a pair of random variables (X, Y) on the same probability space such that $X \sim \mu$ and $Y \sim \nu$.*

Theorem 1.1. *For all couplings (X, Y) of μ and ν , we have:*

$$\|\mu - \nu\| \leq \mathbb{P}(X \neq Y). \tag{2}$$

Furthermore, there always is a coupling (X, Y) which achieves equality in (2).

Proof. We content ourselves with verifying the inequality (2), which is practice all we are going to use anyway. (But it is reassuring to know that this is a sharp inequality!) Let Q denote the law of a coupling (X, Y) , which is a probability measure on $S \times S$. Let Δ be the diagonal:

$$\Delta = \{(s, s) : s \in S\}.$$

If A is any subset of S , then we have:

$$\begin{aligned}
|\mu(A) - \nu(A)| &= |Q(A \times S) - Q(S \times A)| \\
&= |Q(A \times S \cap \Delta) + Q(A \times S \cap \Delta^c) \\
&\quad - Q(S \times A \cap \Delta) - Q(S \times A \cap \Delta^c)| \\
&= |Q(A \times S \cap \Delta^c) - Q(S \times A \cap \Delta^c)| \\
&\leq Q(\Delta^c) = \mathbb{P}(X \neq Y).
\end{aligned}$$

In the third equality we have cancelled the first and third term because there is equality on the diagonal. \square

This proof is thus simple enough but we will see how powerful it is in a moment. First, a few consequences:

Proposition 1.2. *We have the following facts.*

1. $d(t)$ is non-increasing with time.
2. Let ρ be defined by:

$$\rho(t) = \max_{x, y \in S} \|p^t(x, \cdot) - p^t(y, \cdot)\|.$$

Then

$$d(t) \leq \rho(t) \leq 2d(t).$$

3. ρ is submultiplicative: for all $s, t \geq 0$:

$$\rho(t + s) \leq \rho(t)\rho(s).$$

Proof. We will prove points 2. and 3. The right-hand side of point 2. is simply the triangular inequality. For the left-hand side, observe that by stationarity, if $A \subset S$,

$$\pi(A) = \sum_{y \in S} \pi(y)P^t(y, A)$$

Therefore, by the triangular inequality:

$$\begin{aligned}
\|\pi - p^t(x, \cdot)\| &= \max_{A \subset S} |P^t(x, A) - \pi(A)| \\
&= \max_{A \subset S} \left| \sum_{y \in S} \pi(y)[P^t(x, A) - P^t(y, A)] \right| \\
&\leq \max_{A \subset S} \sum_{y \in S} \pi(y) |P^t(x, A) - P^t(y, A)| \\
&\leq \rho(t) \sum_{y \in S} \pi(y) = \rho(t).
\end{aligned}$$

\square

For point 3, we may use our coupling argument: let (X_s, Y_s) be the optimal coupling of $p^s(x, \cdot)$ with $p^s(y, \cdot)$. Thus

$$\|p^s(x, \cdot) - p^s(y, \cdot)\| = P(X_s \neq Y_s).$$

From X_s and Y_s we construct X_{s+t} and Y_{s+t} in such a way that they form a particular coupling of $p^{s+t}(x, \cdot)$ with $p^{s+t}(y, \cdot)$, as follows. There are two possibilities to consider: either $X_s = Y_s$, or not. Conditionally on the event $A_z = \{X_s = Y_s = z\}$ we take

$$X_{s+t} = Y_{s+t} \sim p^t(z, \cdot)$$

while conditionally on the event $A_{z,z'} = \{X_s = z, Y_s = z'\}$, with $z \neq z'$, then we choose

$$X_{s+t} \sim p^{s+t}(z, \cdot) \text{ and } Y_{s+t} \sim p^{s+t}(z', \cdot)$$

with a choice of X_{s+t} and Y_{s+t} such that X_{s+t} and Y_{s+t} form an optimal coupling of $p^t(z, \cdot)$ with $p^t(z', \cdot)$. Thus

$$\mathbb{P}(X_{s+t} = Y_{s+t} | A_{z,z'}) \leq \rho(t).$$

With these definitions,

$$\begin{aligned} \rho(s+t) &\leq \mathbb{P}(X_{s+t} \neq Y_{s+t}) \\ &= \mathbb{P}(X_s \neq Y_s) \mathbb{P}(X_{s+t} \neq Y_{s+t} | X_s \neq Y_s) \\ &= \rho(s) \mathbb{P}(X_{s+t} \neq Y_{s+t} | X_s \neq Y_s). \end{aligned}$$

Let $\mu(x, y)$ denote the law of (X_s, Y_s) given $X_s \neq Y_s$. By the Markov property at time s , we have:

$$\begin{aligned} \rho(s+t) &\leq \rho(s) \sum_{z \neq z'} \mu(z, z') \mathbb{P}(X_{s+t} \neq Y_{s+t} | A_{z,z'}) \\ &\leq \rho(s) \rho(t) \sum_{z \neq z'} \mu(z, z') \\ &= \rho(s) \rho(t), \end{aligned}$$

as desired.

1.3 Example: Random to top shuffling.

To illustrate the power of the method of coupling, nothing better than to view it on action on the following simple method of shuffling card: at each step, take a randomly chosen card in the deck and insert it at the top of the deck. Mathematically, the state space is \mathcal{S}_n , the permutation group on $\{1, \dots, n\}$ (with $n = 52$ for a real deck of cards). Our convention is that cards are labelled $1, \dots, n$ and that $\sigma(i)$ gives the value of the card in position i of the deck. Equivalently, $\sigma^{-1}(i)$ gives the position of card number i .

The operation of taking the top card and inserting in position i of the deck is thus equivalent to multiplying on the right the current permutation σ by the cycle $(1 \ 2 \dots \ i)$, that is, the permutation which maps $1 \rightarrow 2, 2 \rightarrow 3, \dots, i \rightarrow 1$. That is,

$$\sigma' = \sigma \cdot (i \ i-1 \ \dots \ 2 \ 1)$$

where \cdot is the composition of permutations. [As a check, we note that the card now on top is $\sigma'(1) = \sigma(i)$, was in position i before the shuffle. Taking the alternative convention that $\sigma(i)$ denotes the position of card number i , we are of course led to $\sigma' = (1\ 2\ \dots\ i)\sigma$.]

It is easy to check that the uniform distribution is invariant for this shuffling method and that the chain is aperiodic. The result concerning the mixing time of this chain is as follows:

Theorem 1.2. *The random-to-top chain exhibits cutoff at time $t_{\text{mix}} = n \log n$.*

Proof. Consider two decks X_t and Y_t such that X_0 is initially in an arbitrary order (say the identity Id : by symmetry it does not matter), and Y_0 is a permutation which is chosen according to the stationary measure π . We construct a coupling of X_t and Y_t as follows: at each step, we draw $1 \leq i \leq n$ uniformly at random. In both decks, we take card number i and put it at the top. Note that both decks are evolving according to the transition probabilities of the random-to-top chain, so in particular the right-hand deck Y_t always has the uniform distribution.

A further property of that coupling is that once a card i has been selected, its position in both decks will be identical for all subsequent times, as it will first be on top of the deck and then will move down by one unit each time another card is selected. If it is selected again, it will move to the top of the deck again in both decks and this keeps on going forever. In particular, if

$$\tau = \inf\{t \geq 0 : \text{all cards have been selected once}\}$$

then $X_\tau = Y_\tau$. Hence for all $t \geq 0$:

$$d(t) \leq \mathbb{P}(\tau > t).$$

But it is easy to compute asymptotics for τ , as this problem is the classical *coupon collector* problem: when exactly $0 \leq j \leq n-1$ cards have been touched, the time until a new card will be touched is geometric with parameter $1 - j/n$. Hence:

$$\tau = X_0 + \dots + X_{n-1}$$

where X_j are independent geometric variables with parameter j/n . Thus

$$\begin{aligned} \mathbb{E}(\tau) &= \frac{n}{1} + \frac{n}{2} + \dots + \frac{n}{n} \\ &= n \sum_{j=1}^n \frac{1}{j} \\ &\sim n \log n. \end{aligned}$$

Note that $\text{var}(X_j) = (1 - (j/n))/(j/n)^2 \leq n^2/j^2$ so that

$$\text{var}(\tau) \leq Cn^2.$$

Hence by Chebyshev's inequality there is concentration: $t = (1 + \varepsilon)\mathbb{E}\tau$,

$$d(t) \leq \mathbb{P}(\tau > t) \rightarrow 0.$$

This proves the upper-bound on the mixing time. For the lower-bound, we note the following: let A_j be the event that the j bottom cards of the deck are in their original relative order

(that is, if these cards are in order from the bottom i_1, \dots, i_j then we have $i_1 > i_2 > \dots > i_j$. Naturally, for a uniform permutation,

$$\pi(A_j) = \frac{1}{j!}$$

as any arrangement of the j bottom cards is equally likely. Thus if j is reasonably large, this has very small probability for a uniform permutation. However, if we are significantly before τ then the event A_j has a pretty large chance to hold for some high value of j . Indeed, if $t \leq (1 - \varepsilon)\tau$, then many cards have not been touched. All these cards must be at the bottom of the deck and conserve their initial order.

Thus fix j arbitrarily large. The probability that A_j holds for X_t at time $t = (1 - \varepsilon)n \log n$ is at least the probability that j cards have not been touched. The following lemma will give us the estimate we are looking for:

Lemma 1.2. *Let $b < 1$. For any $b < b' < 1$, if $t = bn \log n$, at least $n^{1-b'}$ cards have not been touched by time t , with probability tending to 1 as $n \rightarrow \infty$.*

The proof of this lemma is exactly the same as before. To conclude the proof of the theorem, fix $\varepsilon > 0$ and let $t = (1 - 2\varepsilon)n \log n$. Let A be the event that n^ε cards have not been touched. Then, by the above lemma, $P(X_t \in A) = 1 - o(1)$. On the other hand for a uniform random permutation, $\pi(A) = 1/(n^\varepsilon!)$ since the order of the n^ε cards at the bottom of the deck is itself uniform random.

$$\begin{aligned} d(t) &\geq \|P(X_t \in A) - \pi(A)\| \\ &\geq \left| 1 - o(1) - \frac{1}{n^\varepsilon!} \right| \rightarrow 1 \end{aligned}$$

Therefore,

$$\lim_{n \rightarrow \infty} d(t) \geq 1.$$

This proves the lower-bound of the cutoff phenomenon. □

1.4 Example: random walk on a hypercube.

Let $H_n = \{0, 1\}^n$ be the n -dimensional hypercube. A random walk on the hypercube (in continuous time) is the following process: at rate 1, we choose a coordinate $1 \leq i \leq n$ uniformly at random, and flip it: that is, if it was 0 we change it to a 1, and if it was a 1 we flip it to a zero.

Again, the uniform measure on H_n (i.e., independent uniform bits 0 or 1) is the stationary measure for this chain. By coupling we obtain the following estimate:

Theorem 1.3. *Let $t = (1/2 + \varepsilon)n \log n$. Then $d(t) \rightarrow 0$.*

Proof. Again, the idea is to couple the process with a different random walk ($Y_t, t \geq 0$) which starts directly from the stationary measure. Because of periodicity issues, we need to assume that X and Y differ at an even number of coordinates. (If not, let Y evolve for one step and do nothing to X first: this problem is then resolved). At each step we pick a i at random in $\{1, \dots, n\}$. If $X_i(t) = Y_i(t)$ then we flip both coordinates X_i and Y_i . If $X_i \neq Y_i$, we find j such that $X_j \neq Y_j$. (For instance, we take the smallest $j > i$ which satisfies this property,

where smallest is interpreted cyclically mod n). We then flip bit i for X and bit j for Y . This has the effect of making two new coordinates agree for X and Y . Naturally, once all coordinates have been touched they stay the same forever after, and thus the mixing time is dominated by the stopping time τ such that all coordinates have been touched. At every step we touch two different coordinates so playing with the coupon collector problem gives us

$$\mathbb{P}(\tau > (1/2 + \varepsilon)n \log n) \rightarrow 0.$$

This proves the result. □

Remark 1.1. *This coupling is not sharp. It is easy to get an exact formula for the distribution of X : each coordinate evolves as an independent bit-flipping Markov chain, changing at rate $1/n$. For this, the probability to be at 0 is*

$$\mathbb{P}(N_{t/n} = 1 \pmod{2}) = \frac{1}{2}(1 - e^{-2t/n}).$$

where N_t is a Poisson process with rate 1. From this exact expression one obtains:

$$d(t) = 2^{-n-1} \sum_{L=0}^n \binom{n}{L} \left| (1 + e^{-2t/n})^{n-L} (1 - e^{-2t/n})^L - 1 \right|$$

so cutoff occurs precisely at time $t_{\text{mix}} = (1/4)n \log n$.

2 Spectral Gap

2.1 Eigenvalue decomposition

Our presentation here follows quite closely Chapter 12 of [7].

Proposition 2.1. *Let P be a transition matrix.*

1. *If λ is a (possibly complex) eigenvalue, then $|\lambda| \leq 1$*
2. *if P is irreducible then eigenspace associated with $\lambda = 1$ is one-dimensional and is generated by $(1, 1, \dots, 1)$.*
3. *If P is irreducible and aperiodic then -1 is not an eigenvalue.*

Let π be the stationary measure of the chain associated with P , and define an inner product on real-valued functions on S , $\langle \cdot, \cdot \rangle_\pi$ by:

$$\langle f, g \rangle_\pi = \sum_{x \in S} f(x)g(x)\pi(x).$$

Equipped with this scalar product the space of real valued functions may be viewed as $\ell^2(\pi)$. One of the traditional techniques for studying Markov chains is to diagonalize them. It is then particularly useful to take a set of eigenfunctions orthonormal with respect to $\langle \cdot, \cdot \rangle_\pi$.

Let $|S| = n$, and assume that that π is reversible with respect to P : $\pi(x)P(x, y) = \pi(y)P(y, x) \forall x, y \in S$. Then all eigenvalues are real and we can order them in decreasing order from 1 to -1 :

$$\lambda_1 = 1 > \lambda_2 \geq \dots \geq \lambda_n \geq -1.$$

Theorem 2.1. *Assume that π is reversible with respect to P . Then:*

1. *There exists a set of eigenfunctions f_1, \dots, f_n which are orthonormal for $\langle \cdot, \cdot \rangle_\pi$ and f_1 is the constant vector $(1, \dots, 1)$.*
2. *P^t can be decomposed as:*

$$\frac{P^t(x, y)}{\pi(y)} = \sum_{j=1}^n f_j(x)f_j(y)\lambda_j^t.$$

Proof. This is essentially the classical spectral theorem. if

$$A(x, y) = \sqrt{\frac{\pi(x)}{\pi(y)}} P(x, y)$$

then reversibility of P implies that A is symmetric. Hence by the spectral theorem there exists a set of eigenfunction ϕ_j which diagonalize A and are orthonormal with respect to the Euclidean product $\langle \cdot, \cdot \rangle$. If D_π is the diagonal matrix with entries $\pi(x), x \in S$, then

$$f_j = D_\pi^{-1/2} \phi_j$$

defines the desired eigenfunctions of P as can be readily checked.

For the decomposition in the theorem note that if $s \in S \mapsto \delta_y(s)$ is the function equal to 1 if $s = y$ and 0 otherwise, we can expand this function on the orthonormal basis:

$$\begin{aligned}\delta_y &= \sum_{j=1}^n \langle \delta_y, f_j \rangle_{\pi} f_j \\ &= \sum_{j=1}^n f_j(y) \pi(y) f_j.\end{aligned}$$

Hence, since $P^t(x, y)$ is nothing else but $(P^t \delta_y)(x)$ and λ_j^t is an eigenvalue of P^t we get:

$$P^t(x, y) = \sum_{j=1}^n f_j(y) \pi(y) \lambda_j^t f_j(x)$$

as required. \square

Definition 2.1. Let $\lambda_* = \max\{|\lambda| : \lambda \text{ eigenvalue} \neq 1\}$. $\gamma^* = 1 - \lambda_*$ is called the absolute spectral gap, and $\gamma = 1 - \lambda_2$ is called the spectral gap of P . The relaxation time t_{rel} is defined by

$$t_{\text{rel}} = \frac{1}{\gamma^*}.$$

The negative eigenvalues are in general not so relevant. One way to see this is to define a lazy chain \tilde{P} by saying that, with probability 1/2, the lazy chain does nothing, and with probability 1/2, it takes a step according to P . Thus:

$$\tilde{P} = \frac{1}{2}(I + P)$$

where I is the $|S|$ -dimensional identity matrix. Then by point (i) in Proposition 2.1, we see that all eigenvalues are nonnegative, and hence $\gamma^* = \gamma$. On the other hand, the mixing time of \tilde{P} is essentially twice that of P .

Here is how we can say something about the mixing times using the spectral gap. In practice this is often one of the first things to look at. Let $\pi_{\min} := \min_{x \in S} \pi(x)$ (note that if P is a random walk on a d -regular graph, then $\pi(x) \equiv 1/|S|$ so $\pi_{\min} = 1/|S|$).

Theorem 2.2. Fix $0 < \varepsilon < 1$ arbitrary. Assume that P is aperiodic, irreducible and reversible with respect to π . Then

$$(t_{\text{rel}} - 1) \log \left(\frac{1}{2\varepsilon} \right) \leq t_{\text{mix}}(\varepsilon) \leq \log \left(\frac{1}{2\varepsilon \sqrt{\pi_{\min}}} \right) t_{\text{rel}}$$

Proof. Recall that one of the basic identities for the definition of the total variation distance is

$$\begin{aligned}\|P^t(x, \cdot) - \pi\| &= \frac{1}{2} \sum_y |P^t(x, y) - \pi(y)| \\ &= \sum_y \pi(y) \left| \frac{P^t(x, y)}{\pi(y)} - 1 \right| \\ &= \left\| \frac{P^t(x, \cdot)}{\pi(\cdot)} - 1 \right\|_1\end{aligned}$$

where $\|\cdot\|_1$ refers to the $\ell^1(\pi)$ norm. Taking the square and using Jensen's inequality, we get

$$4\|P^t(x, \cdot) - \pi\|^2 \leq \left\| \frac{P^t(x, \cdot)}{\pi(\cdot)} - 1 \right\|_2^2$$

Expanding the function on the eigenfunction basis f_j using Theorem 2.1, we get

$$\begin{aligned} \left\| \frac{P^t(x, \cdot)}{\pi(\cdot)} - 1 \right\|_2^2 &= \left\| \sum_{j=2}^n f_j(x) f_j(\cdot) \lambda_j^t \right\|_2^2 \\ &= \sum_{j=2}^n \lambda_j^{2t} f_j(x)^2 \leq \lambda_*^{2t} \sum_{j \geq 2} f_j(x)^2. \end{aligned} \quad (3)$$

Now, we claim that $\sum_{j=1}^n f_j(x)^2 = \pi(x)^{-1}$. Indeed, by decomposition:

$$\pi(x) = \langle \delta_x, \delta_x \rangle_\pi = \sum_{j=1}^n f_j(x)^2 \pi(x)^2.$$

Hence

$$\begin{aligned} 4\|P^t(x, \cdot) - \pi\|^2 &\leq \lambda_*^{2t} \pi(x)^{-1} \leq \lambda_*^{2t} \pi_{\min}^{-1} \\ &\leq (1 - \gamma_*)^{2t} \pi_{\min} \leq e^{-2\gamma_* t} \pi_{\min}^{-1}. \end{aligned}$$

Maximising over x and taking the square root, we get

$$d(t) \leq \frac{1}{2} e^{-\gamma_* t} \sqrt{\pi_{\min}^{-1}}. \quad (4)$$

Solving for the right-hand side equal to ε gives us $d(t) \leq \varepsilon$ as soon as $t \geq \frac{1}{\gamma_*} \log\left(\frac{1}{2\varepsilon\sqrt{\pi_{\min}}}\right)$.

For the lower-bound, let $f = f_j$ for some $j \geq 2$, and let $\lambda \neq 1$ be the eigenvalue. Since the eigenfunctions are orthonormal, we get $\langle f, f_1 \rangle_\pi = 0 = \mathbb{E}_\pi(f)$, and hence:

$$\begin{aligned} |\lambda^t f(x)| &= |P^t f(x)| = \left| \sum_{y \in S} [P^t(x, y) f(y) - \pi(y) f(y)] \right| \\ &\leq 2\|f\|_\infty d(t) \end{aligned}$$

Taking x to be the point such that $f(x) = \|f\|_\infty$, we obtain

$$|\lambda|^t \leq 2d(t) \quad (5)$$

Taking $|\lambda| = \lambda_*$ gives the lower-bound: indeed, evaluating at $t = t_{\text{mix}}(\varepsilon)$ gives us

$$\lambda_*^t \leq 2\varepsilon$$

and hence

$$\frac{1}{2\varepsilon} \leq \frac{1}{\lambda_*^t}.$$

Taking the logarithm and using $\log(x) \leq x - 1$ for $x > 1$, we get

$$\log\left(\frac{1}{2\varepsilon}\right) \leq t_{\text{mix}}(\varepsilon) \frac{1}{|\lambda_* - 1|} = t_{\text{mix}}(\varepsilon) t_{\text{rel}}.$$

□

Remark 2.1. *When the chain is transitive, one can obtain a slightly different estimate which is a bit better in some examples: Recall that we have*

$$4\|P_t(x, \cdot) - \pi(\cdot)\|^2 \leq \sum_{j=2}^n \lambda_j^{2t} f_j(x)^2.$$

The left hand side does not depend on x by transitivity, and is thus equal to $4d(t)^2$ for each $x \in S$. We are thus allowed to sum this inequality over $x \in S$ and divide $n = |S|$. We obtain:

$$4d(t)^2 \leq \sum_{j=2}^n \lambda_j^{2t} \sum_{x \in S} \frac{1}{n} f_j(x)^2.$$

Since $\pi(x) = 1/n$, we recognize $\sum_{x \in S} \frac{1}{n} f_j(x)^2 = \|f_j\|_\pi^2 = 1$. Thus

$$4d(t)^2 \leq \sum_{j=2}^n \lambda_j^{2t}. \tag{6}$$

2.2 Example: Random walk on the circle.

We will see what we can get from Theorem 2.2 on a concrete example of a simple random walk on a large cycle $\mathbb{Z}/n\mathbb{Z}$. We view this as a subset of the complex plane $W_n = \{1, \omega, \omega^2, \dots, \omega^{n-1}\}$ with $\omega = e^{2i\pi/n}$.

Let P be the matrix of this walk. To be an eigenfunction f with eigenvalue λ for P means that

$$\lambda f(\omega^k) = Pf(\omega^k) = \frac{1}{2}(f(\omega^{k+1}) + f(\omega^{k-1}))$$

for all $1 \leq k \leq n$. We claim that the functions ϕ_j = “take the j^{th} power” give n eigenvalues. This can be seen geometrically: see Figure 12.1 in [7]. More formally, note that

$$\begin{aligned} \frac{\phi_j(\omega^{k+1}) + \phi_j(\omega^{k-1})}{2} &= \frac{\omega^{jk} \omega^k + \omega^{-k}}{2} \\ &= \phi_j(\omega^k) \operatorname{Re}(\omega^k) \\ &= \phi_j(\omega^k) \cos\left(\frac{2\pi j}{n}\right). \end{aligned}$$

Thus ϕ_j is an eigenfunction with eigenvalue $\cos(2\pi j/n)$.

If n is even, the chain is periodic and the absolute spectral gap is 1. If n is odd, the chain is aperiodic and both the spectral gap and the absolute spectral gap are equal to

$$1 - \cos\left(\frac{2\pi}{n}\right) \sim \frac{2\pi^2}{n^2}$$

as $n \rightarrow \infty$. Thus

$$t_{\text{rel}} \sim \frac{n^2}{2\pi^2}.$$

It makes a lot of intuitive sense that n^2 is the correct order of magnitude. However, since $|S| = n$, the lower and upper bound in Theorem 2.2 don't match. We can get around this fact using slightly improved estimates in the upper-bound.

A better statement is as follows:

Theorem 2.3. Assume that $n \geq 7$ is odd. If $t \geq n^2$,

$$d(t) \leq \exp\left(-\alpha \frac{t}{n^2}\right)$$

where $\alpha = \pi^2/2$. Conversely, for any $t \geq 0$,

$$d(t) \geq \frac{1}{2} \exp\left(-\alpha \frac{t}{n^2} - \beta \frac{t}{n^4}\right)$$

where $\beta = \pi^4/11$.

Proof. We use the sharpened version of Theorem 2.2 (i.e., (6)) to prove this result. We have:

$$\begin{aligned} d(t)^2 &\leq \frac{1}{4} \sum_{j=1}^{n-1} \cos\left(\frac{2\pi j}{n}\right)^{2n} \\ &= \frac{1}{2} \sum_{j=1}^{(n-1)/2} \cos\left(\frac{\pi j}{n}\right)^{2n}. \end{aligned}$$

Since $\cos(x) \leq e^{-x^2/2}$ on $[0, \pi/2]$ (a consequence of concavity of the cosine function over that interval) we see that

$$\begin{aligned} d(t)^2 &\leq \frac{1}{2} \sum_{j=1}^{(n-1)/2} \exp\left(-\frac{\pi^2 j^2 t}{n^2}\right) \\ &\leq \frac{1}{2} \exp\left(-\frac{\pi^2 t}{n^2}\right) \sum_{j=1}^{\infty} \exp\left(-\frac{\pi^2 (j^2 - 1)t}{n^2}\right) \\ &\leq \frac{1}{2} \exp\left(-\frac{\pi^2 t}{n^2}\right) \sum_{j=1}^{\infty} \exp\left(-\frac{3\pi^2 t}{n^2}\right) \\ &= \frac{1}{2} \frac{\exp\left(-\frac{\pi^2 t}{n^2}\right)}{1 - \exp\left(-\frac{3\pi^2 t}{n^2}\right)} \end{aligned}$$

from which the upper-bound follows. For the lower-bound, note that we have a general lower bound using the second eigenvalue: by (5), we have

$$\lambda_2^t \leq 2d(t)$$

and thus here

$$d(t) \geq \frac{1}{2} \cos\left(\frac{\pi j}{n}\right)^t$$

Since $\cos x \geq \exp\left(-\frac{x^2}{2} - \frac{x^4}{11}\right)$ for all $0 \leq x \leq 1/2$, we get the desired lower-bound. \square

3 Dirichlet forms and Poincaré inequalities

We describe a few basic tools coming from the study of Dirichlet forms, which are energy functionals associated with a Markov chain. These are useful to establish so-called Poincaré inequalities, which in turn give us estimates about the spectral gap. Mostly, this will be extremely useful in the next lecture when we introduce the comparison techniques of Diaconis and Saloff-Coste.

3.1 Dirichlet forms

We start with the following basic facts. Let A, B be two normed vector spaces. If $K : A \rightarrow B$ is a linear operator, denote

$$\|K\|_{A \rightarrow B} = \sup_{f \in A, \|f\|_A=1} \|Kf\|_B.$$

Suppose S is a finite state space and P is the transition matrix of an irreducible positive recurrent Markov chain with stationary distribution π . Then P may be viewed as an operator from $\ell^p(\pi)$ to $\ell^q(\pi)$, via the following:

$$(Pf)(x) = \sum_y P(x, y)f(y),$$

for any $f \in \ell^p(\pi)$ (which is simply any function $f : S \rightarrow \mathbb{R}$). Then let

$$\|P\|_{p \rightarrow q} = \|P\|_{\ell^p(\pi) \rightarrow \ell^q(\pi)}$$

Definition 3.1. Let $f, g : S \rightarrow \mathbb{R}$. The Dirichlet form associated with P is defined by

$$\mathcal{E}(f, g) = \langle (I - P)f, g \rangle_\pi.$$

Hence

$$\begin{aligned} \mathcal{E}(f, g) &= \sum_x \pi(x)[f(x) - (Pf)(x)]g(x) \\ &= \sum_x \pi(x) \left[\sum_y P(x, y)(f(x) - f(y)) \right] g(x) \\ &= \sum_{x, y} \pi(x)P(x, y)g(x)(f(x) - f(y)). \end{aligned}$$

When P is reversible with respect to π , another expression for the right hand side is $\sum_{x, y} \pi(y)P(y, x)g(x)(f(x) - f(y))$. Interverting the role of x and y , and summing these two expressions, we get

$$\mathcal{E}(f, g) = \frac{1}{2} \sum_{x, y} \pi(x)P(x, y)[f(y) - f(x)][g(y) - g(x)], \quad (7)$$

a much more useful expression.

Now, it is natural to define, for the edge $e = (x, y)$.

$$Q(e) = \frac{1}{2}[P(x, y)\pi(x) + P(y, x)\pi(y)]$$

which is (up to the factor 1/2) the flow through the edge e at equilibrium. We also call $\nabla f(e) = f(y) - f(x)$ the discrete derivative. Then with these notations, (7) becomes

$$\mathcal{E}(f, g) = \sum_e Q(e) \nabla f(e) \nabla g(e).$$

Hence the Dirichlet energy we have just defined is the analogue of the classical Dirichlet energy from mechanics on a domain $D \subset \mathbb{R}^d$ for f, g smooth real functions on D , their energy is defined to be:

$$\mathcal{E}(f, g) = \int_D \nabla f \cdot \nabla g.$$

3.2 Variational characterization

The following “variational characterization” (or minimax characterization) of the spectral gap in terms of the Dirichlet form:

Theorem 3.1. *Let γ be the spectral gap. Then*

$$\gamma = \min_{\substack{f: S \rightarrow \mathbb{R} \\ f \perp 1, \|f\|_2=1}} \mathcal{E}(f, f) = \min_{\substack{f: S \rightarrow \mathbb{R} \\ f \perp 1}} \frac{\mathcal{E}(f, f)}{\|f\|_2^2}.$$

Equality is attained for $f = f_2$.

Proof. We only do the case $k = 2$, the rest is left as an exercise. By scaling it suffices to prove the first equality. Now note that if f_j is the orthonormal set of functions with $f_1 \equiv 1$ the constant vector, then for any function f with $\|f\| = 1$ with $f \perp 1$, we have

$$f = \sum_{j=1}^n \langle f, f_j \rangle_{\pi} f_j = \sum_{j=2}^n \langle f, f_j \rangle_{\pi} f_j$$

since $\langle f, f_1 \rangle_{\pi} = 0$ by assumption. On the other hand, using the orthonormality of the eigenfunctions, using the order on eigenvalues and the fact that $\|f\|^2 = 1$:

$$\begin{aligned} \mathcal{E}(f, f) &= \langle (I - K)f, f \rangle_{\pi} \\ &= \sum_{j=1}^n |\langle f, f_j \rangle_{\pi}|^2 (1 - \lambda_j) \\ &\geq (1 - \lambda_2) \sum_{j=2}^n \langle f, f_j \rangle_{\pi}^2 = 1 - \lambda_2 \gamma. \end{aligned}$$

On the other hand there is clearly equality for $f = f_2$. Note that the calculation holds even if P is not assumed aperiodic. \square

The other “minimax” characterization is also well-known and will be useful in the lecture:

Theorem 3.2. *For W a subspace of $\ell^2(\pi)$, let*

$$m(W) = \min\{\mathcal{E}(f, f), f \in \ell^2(\pi), \|f\|_2 = 1\},$$

and let

$$M(W) = \max\{\mathcal{E}(f, f), f \in \ell^2(\pi), \|f\|_2 = 1\}.$$

Then

$$1 - \lambda_j = \max\{m(W) : \dim(W^{\perp}) = j - 1\} = \min\{M(W) : \dim(W) = j\}$$

3.3 Poincaré inequality and the path method

Note that if $f : S \rightarrow \mathbb{R}$ and $f \perp 1$, then this means $\mathbb{E}_\pi(f) = 0$. Thus the inequality above says that

$$\gamma \|f\|^2 \leq \mathcal{E}(f, f)$$

and hence, in probabilistic terms,

$$\text{var}_\pi(f) \leq \frac{1}{\gamma} \mathcal{E}(f, f).$$

This motivates the following definition:

Definition 3.2. We say that P satisfies a Poincaré inequality with constant C if, for all functions $f : S \rightarrow \mathbb{R}$,

$$\text{var}_\pi(f) \leq C \mathcal{E}(f, f). \quad (8)$$

As just discussed, a Poincaré inequality *always* holds with $C = 1/\gamma$ (which is the best choice possible), hence a (finite) Poincaré inequality is equivalent to a (bounded) spectral gap. Thus Poincaré inequalities can be used to obtain lower-bounds on the spectral gap, and hence (roughly speaking) upper-bounds on mixing times – more about this later.

For now, let us see a fairly general example of how one would derive a Poincaré inequality in the context of a Markov chain on S . It turns out that there are easy ways to establish Poincaré inequalities by doing some elementary geometric analysis in some concrete examples. Here is how this roughly works. We associate to S an oriented graph where there is an edge from x to y if $P(x, y) > 0$ or conversely. We say that a collection of edges $\mathcal{A} \subset S \times S$ is called adapted if $(x, y) \in \mathcal{A}$ implies $P(x, y) + P(y, x) > 0$.

If \mathcal{A} is adapted, define for all $x, y \in S$, $\Gamma(x, y) = \{\text{paths from } x \text{ to } y \text{ using only edges from } \mathcal{A} \text{ and at most once}\}$. That is, this is the family of paths from x to y such that each path is *edge-self avoiding*.

For any $x, y \in S$, suppose that we fix, once and for all, a certain path $\ell(x, y) \in \Gamma(x, y)$. Let $|\ell(x, y)|$ denote the length of this path, i.e., its number of edges. Then we have the following result:

Theorem 3.3. *The Poincaré inequality (8) holds with*

$$C = \max_{e \in \mathcal{A}} \left\{ \frac{1}{Q(e)} \sum_{x, y: e \in \ell_{x, y}} |\ell(x, y)| \pi(x) \pi(y) \right\}.$$

In particular, $\gamma \geq 1/C$.

Note that the smaller \mathcal{A} can be chosen, the better this bound gets. The number C may be thought of as a *congestion ratio*.

Proof. This is fairly straightforward: if f is a function on the state space,

$$\begin{aligned}
\text{var}_\pi(f) &= \sum_x \left| f(x) - \sum_y \pi(y) f(y) \right|^2 \pi(x) \\
&\leq \sum_{x,y} |f(x) - f(y)|^2 \pi(x) \pi(y) \quad (\text{by Cauchy-Schwarz}) \\
&\leq \sum_{x,y} |\ell(x,y)| \sum_{e \in \ell(x,y)} |\nabla f(e)|^2 \pi(x) \pi(y) \quad (\text{by Cauchy-Schwarz again}) \\
&\leq \sum_e \left(\frac{1}{Q(e)} \sum_{x,y: e \in \ell(x,y)} |\ell(x,y)| \pi(x) \pi(y) \right) \nabla f(e)^2 Q(e) \\
&\leq C \mathcal{E}(f, f).
\end{aligned}$$

This finishes the proof. \square

If S is the vertex set of a finite graph $G = (S, E)$ and X is the simple random walk on G (i.e., the process jumps to a neighbour chosen uniformly at random at each step) then we can obtain the following bound:

Corollary 3.1. *A Poincaré inequality holds with*

$$C = \frac{d_*^2 \Delta_* \eta_*}{2|E|}$$

where d_* is the maximal degree of the graph, $\Delta_* = \max_{x,y} |\ell(x,y)|$ is the diameter of the graph (in the sense of ℓ), and $\eta_* = \max_{e \in E} \#\{(x,y) : e \in \ell(x,y)\}$.

This inequality isn't very sharp but it can easily be modified to obtain something slightly more precise. The modification consists in introducing another parameter to optimize on, which is a weight function $w : E \rightarrow \mathbb{R}_{>0}$ on the edges. We then define the weight of a path $\ell = \ell(x,y)$ to be:

$$|\ell|_w = \sum_{e \in \ell} \frac{1}{w(e)}.$$

Corollary 3.2. *A Poincaré inequality holds with*

$$C = \max_{e \in \mathcal{A}} \left\{ \frac{w(e)}{Q(e)} \sum_{x,y: e \in \ell(x,y)} |\ell(x,y)|_w \pi(x) \pi(y) \right\}.$$

The proof is a simple adaptation of Theorem 3.3. We call this the path method.

3.4 Example: random walk on the n -dog

As an example of application of this technique, we study the random walk on the so-called n -dog D_n . This is the subgraph of \mathbb{Z}^2 which consists of joining two copies of the square of size n by a corner, say the North-East corner of the first square with the South-West corner of the second one. See Figure 3.4 for a picture.

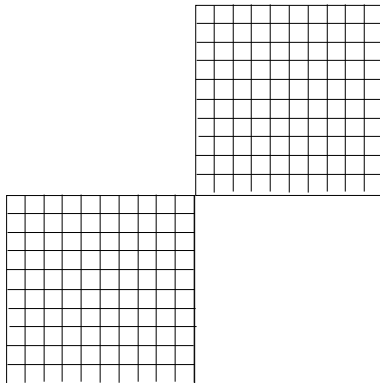


Figure 1: The n -dog graph D_n .

Heuristics. It takes approximately n^2 units of time for a random walk to mix on a single square. However, the presence of the small bottleneck significantly slows down mixing. Indeed, if the walk starts from the centre of one of the two squares say, then mixing can't occur if the bottleneck site hasn't been visited with high probability. The time needed to visit this site is comparable to the *cover time* of the square, i.e., the time needed to visit every vertex at least once. It is not hard to convince yourself that the cover time is approximately $n^2 \log n$ in two dimension, and hence mixing should take approximately $n^2 \log n$ units of time as well. (Once the bottleneck has been visited, it takes another n^2 units of time to mix in the other square, which is negligible compared to $n^2 \log n$). To see that the cover time $n^2 \log n$, note that in two dimensions the range R_t of the walk at time t is approximately $t/\log t$. Indeed, $\mathbb{E}(R_t) = \sum_{i=1}^t \mathbb{P}(E_i)$ where E_i is the event that the walk visits a new site at time i . By reversibility of the simple random walk, this is the same as the event that a random walk run for time i does not come back to the origin. This has probability $1/\log i$ approximately, so

$$\mathbb{E}(R_t) = \sum_{i=1}^t \mathbb{P}(E_i) \approx \sum_{i=1}^t \frac{1}{\log i} \approx \frac{t}{\log t}.$$

Thus while $t \ll n^2 \log n$, then R_t is much smaller than n^2 and so the cover time much be greater. This explains why the cover time is of order $n^2 \log n$, and hence explains the heuristics. \square

What we see next is a way of making this heuristics precise. To ease computations, assume that each vertex on the border of the square is equipped with a self-loop, and a vertex at a corner is equipped with two self-loops. Thus every vertex in D_n has degree 4 exactly, and hence the stationary distribution is the uniform measure on D_n .

Theorem 3.4. *Let γ be the spectral gap of the random walk on D_n and let $t_{\text{rel}} = \gamma^{-1}$ be the relaxation time. Then for all $n \geq 1$,*

$$t_{\text{rel}} \leq 64(n+1)^2 \log(2n+1),$$

while for n large enough:

$$t_{\text{rel}} \geq 2n^2 \log n.$$

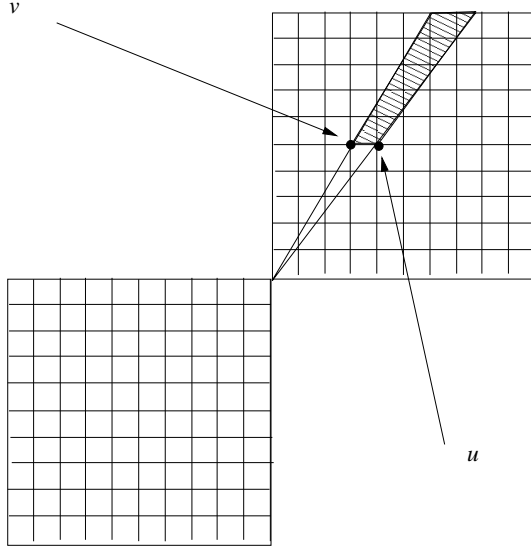


Figure 2: The shaded region corresponds to all the points x, y such that $e = (u, v) \in \ell(x, y)$.

Proof. We prove the upper-bound. Note that $|S| = 2(n+1)^2 - 1$ since each square has $(n+1)^2$ vertices (but the centre shouldn't be counted twice). π is the uniform measure, so $\pi \equiv 1/|S|$. The equilibrium flow $Q(e)$ satisfies $Q(e) = (4|S|)^{-1}$ for all edges e . We wish to apply Corollary 3.2, and for this we need to choose two things. The first one is the set of paths $\ell(x, y)$ for $x, y \in D_n$, and the second is the weight function $w : E \rightarrow \mathbb{R}$. For the special paths $\ell(x, y)$ we only define $\ell(x, 0)$ (where 0 is the junction between the two squares) and we define $\ell(x, y)$ to be the concatenation of the two paths $\ell(x, 0)$ with $\ell(0, y)$: that is, $\ell(x, y) = \ell(x, 0) \cup \ell(0, y)$. If $x \in D_n$, we define $\ell(x, 0)$ to be the lattice path which stays closest to the Euclidean segment $[x, 0]$ between x and 0. This is the path that stays at distance $\leq \sqrt{2}$ from $[x, 0]$.

Before choosing the weights, a word of motivation. If $e = (u, v)$ is an edge with $d(u, 0) = i+1$ and $d(v, 0) = i$ (here d is the graph distance, so this simply means v is closest to 0), then:

$$\#\{x \in S : e \in \ell(x, 0)\} \leq \frac{4(n+1)^2}{i+1}.$$

(This is the cardinality of the shaded region in figure 3.4.) Thus

$$\#\{(x, y) \in S^2 : e \in \ell(x, y)\} \leq \frac{4(n+1)^2}{i+1} \leq \frac{8(n+1)^2|S|}{i+1}.$$

This is because for e to be on $\ell(x, y)$, e has to be on $\ell(x, 0)$ and y could be anywhere in S , or conversely.

This motivates the following choice: if $d(0, e) = i$ then take $w(e) = i+1$. Thus for $x, y \in S$,

$$|\ell(x, y)|_w \leq 2 \sum_{i=0}^{2n-1} \frac{1}{i+1} \leq 2 \log(2n+1).$$

Thus by Corollary 3.2, there is a Poincaré inequality with

$$\begin{aligned} C &= \max_{e \in \mathcal{A}} \left\{ \frac{w(e)}{Q(e)} \sum_{x, y: e \in \ell(x, y)} |\ell(x, y)|_w \pi(x) \pi(y) \right\} \\ &\leq \max_{0 \leq i \leq n} \left\{ (i+1) 4|S| 2 \log(2n+1) \frac{\#\{x, y \in S : e \in \ell(x, y)\}}{|S|^2} \right\} \\ &\leq 64 \log(2n+1) n^2. \end{aligned}$$

This gives the upper-bound. For the other direction, take $f(x) = \text{sgn}(x) \log(1 + d(0, x))$, where the function $\text{sgn}(x)$ is $-1, 0$, or 1 depending on whether x is in the left square, the right square or the centre. Then $\mathbb{E}_\pi(f) = 0$ by symmetry.

Moreover, since there are $i+2$ points at distance i from 0 in one square for $i \leq n$,

$$\begin{aligned} \text{var}_\pi f &= \mathbb{E}_\pi(f^2) \geq \frac{1}{|S|} \sum_{i=0}^n (i+2) \log(i+1)^2 \\ &\geq \frac{n^2 (\log n)^2}{2|S|} = (\log n)^2 4n^2 \end{aligned}$$

for n large enough. On the other hand, it is not hard to see that

$$\begin{aligned} \mathcal{E}(f, f) &= \frac{1}{4|S|} \sum_{i=0}^{2n-1} [(i+1) \wedge (2n-i+1)] (\log(i+2) - \log(i+1))^2 \\ &\leq \frac{1}{4|S|} \sum_{i=1}^{2n-1} \frac{1}{i+1} \\ &\leq \frac{\log(2n+1)}{4|S|}. \end{aligned}$$

Thus

$$\gamma \leq \frac{\mathcal{E}(f, f)}{\text{var}_\pi(f)} \leq \frac{1}{2n^2 \log n}$$

for n large enough. □

4 Comparison techniques

We describe here a wonderful set of tools, due to Diaconis and Saloff-Coste in a series of papers dating from around 1993, which show to get estimates on mixing times (specifically, spectral gaps) for Markov chains which may be compared to other Markov chains where answers are known. In fact, the two Markov chains can be quite different, which yields some spectacular applications and examples.

4.1 Random walks on groups

Let G be a group of cardinality $n = |G|$, and let S be a generating set of G such that $G^{-1} = G$. Let $p(\cdot)$ be a probability measure on S such that $p(x) > 0$ for all $x \in S$. The random walk on G based on p is the Markov chain whose transition probabilities are given by the following:

$$P(x, y) = p(yx^{-1}).$$

Random walks on groups have many symmetries that make it possible to use varied techniques. If p is a symmetric kernel ($p(x) = p(x^{-1})$ for all $x \in S$), and thus in particular if p is the uniform distribution on S , then the uniform measure π on G is clearly reversible for the chain. Also, by symmetry, $\|p^n(x, \cdot) - \pi\|_{TV}$ does not depend on the starting point $x \in G$. To simplify the discussion we will focus on this setup in this lecture. We will then call $P^n(x) = P^n(o, x)$ where o is the identity element of G .

4.2 Heat kernel, ℓ^2 distance and eigenvalues

Let G be a finite state space and P the transition matrix of a symmetric random walk on G . For what follows it will be convenient to work in continuous time and hence we define the heat kernel, which is the continuous-time version of the transition probabilities.

Let $(X_t, t \geq 0)$ the continuous-time version of the chain P , i.e., the random process which waits an exponential amount of time with parameter 1 and jumps to a location chosen according to $P(x, \cdot)$. By definition, The *heat kernel* H_t denotes the law of the process at time t started from x . That is,

$$\begin{aligned} H_t(x, y) &= \mathbb{P}_x(X_t = y) \\ &= \sum_{k=0}^{\infty} e^{-t} \frac{t^k}{k!} P^k(x, y) \\ &= \exp\{-t(I - P)\} \end{aligned}$$

where the exponential of a matrix is defined by the usual expansion $\exp M = \sum_{k=0}^{\infty} M^k/k!$. Recall that if P is irreducible, then (since S is also finite) there exists a unique π such that $\pi H_t = \pi$, and moreover $\max_x \|H_t^x - \pi\|_{TV} \rightarrow 0$ as $t \rightarrow \infty$.

Our analysis below will make use of an ℓ^2 distance: let

$$d_2(t) := \sqrt{n \sum_{y \in G} |H_t(x, y) - \pi(y)|^2} \tag{9}$$

$$= \left\| \frac{H_t(x, \cdot)}{\pi(\cdot)} - 1 \right\|_2. \tag{10}$$

Observe that, by the Cauchy-Schwarz inequality, we always have

$$d(t) = \|H_t(x, \cdot) - \pi(\cdot)\|_{TV} \leq \frac{1}{2}d_2(t). \quad (11)$$

On the other hand, the distance $d_2(t)$ is easily evaluated in terms of the eigenvalues: we have in fact already done this computation in (6). We recall it here: by the spectral theorem,

$$\frac{P_k(x, \cdot)}{\pi(\cdot)} - 1 = \sum_{j=2}^n c_j \lambda_j^k f_j$$

with $c_j = f_j(x)$. Thus by conditioning on the number of jumps of the random walk up to time t , we deduce a similar formula for $H_t(x, y)$:

$$\begin{aligned} \frac{H_t(x, \cdot)}{\pi(\cdot)} - 1 &= \sum_{k=0}^{\infty} e^{-t} \frac{t^k}{k!} \left(\frac{P_k(x, y)}{\pi(y)} - 1 \right) \\ &= \sum_{j=2}^n c_j f_j \sum_{k=0}^{\infty} e^{-t} \frac{t^k \lambda_j^k}{k!} \\ &= \sum_{j=2}^n c_j f_j e^{-t\mu_j} \end{aligned}$$

with $\mu_j = 1 - \lambda_j$.

Hence, for all $x \in G$, by orthonormality of the f_j ,

$$\left\| \frac{H_t(x, \cdot)}{\pi(\cdot)} - 1 \right\|_2^2 = \sum_{j=2}^n e^{-2t\mu_j} f_j(x)^2.$$

Since the left hand side does not depend on x by symmetry, we can average over $x \in G$ and get the following identity for $d_2(t)$:

$$\begin{aligned} d_2(t)^2 &= \sum_{j=2}^n e^{-2t\mu_j} \frac{1}{n} \sum_{x \in G} f_j(x)^2 \\ &= \sum_{j=2}^n e^{-2t\mu_j}, \end{aligned} \quad (12)$$

since for all $1 \leq j \leq n$, $\frac{1}{n} \sum_{x \in G} f_j(x)^2 = \|f_j\|_2^2 = 1$.

To summarize,

$$\boxed{d(t) \leq \frac{1}{2}d_2(t) = \frac{1}{2} \sqrt{\sum_{j=2}^n e^{-2t\mu_j}}.} \quad (13)$$

4.3 Comparison techniques

Consider two random walks P, \tilde{P} on a group G . We will have in mind a situation where the mixing behaviour of the random walk \tilde{P} is completely understood – it will be important that this understanding extends to $d_2(t)$ distance – and we wish to deduce an understanding of the mixing behaviour of P .

As it turns out, the correct way to compare two Markov chains is to compare their Dirichlet energy:

Lemma 4.1. *Let \mathcal{E} and $\tilde{\mathcal{E}}$ be the Dirichlet forms of P, \tilde{P} and denote respectively by $\lambda_1 \geq \dots \geq \lambda_n$ and $\tilde{\lambda}_1 \geq \dots \tilde{\lambda}_n$ their eigenvalues. Assume that for some constant $A > 0$, for all $f : G \rightarrow \mathbb{R}$,*

$$\tilde{\mathcal{E}}(f, f) \leq A\mathcal{E}(f, f). \quad (14)$$

Then $\tilde{\mu}_j \leq A\mu_j$, for all $1 \leq j \leq n$.

Proof. This is a straightforward consequence of the minimax principle (Theorem 3.2). Since $\tilde{\mathcal{E}}(f, f) \leq A\mathcal{E}(f, f)$, we directly have, for an arbitrary subspace W of codimension $j - 1$:

$$\min\{\tilde{\mathcal{E}}(f, f) : f \in W, \|f\|_2 = 1\} \leq A \min\{\mathcal{E}(f, f) : f \in W, \|f\|_2 = 1\}$$

i.e., $\tilde{m}(W) \leq Am(W)$. Hence, maximizing over subspaces W of codimension $j - 1$:

$$\tilde{\mu}_j = \max\{\tilde{m}(W), \dim(W^\perp) = j - 1\} \leq A \max\{m(W) : \dim(W^\perp) = j - 1\} = A\mu_j.$$

This concludes the proof. Here note that we have used that the scalar product $\ell^2(\pi)$ is the same “for both chains” because it really only depends on the stationary measure π , which is necessarily the uniform distribution in our setup. \square

As a consequence of this lemma and (13), we obtain the trivial but crucial relation:

Corollary 4.1. *Assume that $\tilde{\mathcal{E}}(f, f) \leq A\mathcal{E}(f, f)$ for all functions $f : G \rightarrow \mathbb{R}$. Then*

$$d_2(t) \leq \tilde{d}_2(t/A).$$

Hence if the ℓ_2 mixing behaviour of \tilde{P} is understood, a comparison inequality such as (14) gives us immediately a bound on the ℓ^2 mixing time of P (and hence a total variation bound as well).

As it turns out, a comparison can be set up in a way very similar to Theorem 3.3. Fix E a generating set for G (which could in principle be distinct from either S, \tilde{S} , but will in general be S). For each $y \in G$, fix a geodesic path from o to y using only edges from S , the generating set for P , i.e. a path of the form $y = z_1 \cdot \dots \cdot z_k$, where $z_i \in S$ and $k = |y|_S$ is minimal. For each generator $z \in S$, let $N(z, y)$ denote the number of times the edge z is used on this path.

Theorem 4.1. *Let P, \tilde{P} be two symmetric random walks on G . Assume that the support of p contains E . Then $\tilde{\mathcal{E}}(f, f) \leq A\mathcal{E}(f, f)$ where*

$$A = \max_{z \in S} \frac{1}{p(z)} \sum_{y \in G} |y| N(z, y) \tilde{p}(y). \quad (15)$$

The proof is a straightforward adaptation of the proof of Theorem 3.3, so we leave it to the reader. When $E = S$, the assumption is trivial.

4.4 Example 1: random transpositions on a graph

We now illustrate the power of comparison techniques with some examples coming from random walks on permutation groups such as S_n . The random walk which very well understood, and which will serve as a comparison benchmark for other random walks, is the random transposition process, where for each transposition $\tau = (i, j)$, $\tilde{p}(\tau) = 1/n^2$, $\tilde{p}(id) = 1/n$ (to avoid periodicity issues).

Using sophisticated techniques from representation theory, Diaconis and Shahshahani (1981) were able to completely analyse this chain (in ℓ^2 sense) and establish the cutoff phenomenon for $\tilde{d}_2(t)$ at time

$$\widetilde{t}_{\text{mix}} = \frac{1}{2}n \log n.$$

More precisely, they showed:

Theorem 4.2. (Diaconis-Shahshahani [5]) *Let $c > 0$. Then there exists a universal $\alpha > 0$ such that $\tilde{d}_2(t) \leq \alpha e^{-c}$ whenever $t \geq (1/2)n(\log + c)$.*

Consider now a fix graph $\mathcal{G} = (V, E)$. One can define a random walk on the permutations of V by imagining that there is a card on each vertex of V and at each step, we exchange two neighbouring cards at random, or do nothing with probability $1/n$ (where $n = |V|$). The case of random transpositions corresponds then corresponds to $\mathcal{G} = K_n$, the complete graph. Interesting examples include the case of *adjacent transpositions* where cards are arranged on a one-dimensional array, and the *star transpositions* where \mathcal{G} is the star. This can also be seen as the “top with random” transposition scheme.

For each vertex $x, y \in V$ fix γ_{xy} a path from x to y , and set

$$\Delta = \text{length of longest path ,}$$

(where length is measured in the number of edges),

$$K = \max_{e \in E} |\{(x, y) \in V^2 : e \in \gamma_{x,y}\}|.$$

Then we have the following theorem:

Theorem 4.3. *The comparison $\tilde{\mathcal{E}} \leq A\mathcal{E}$ holds with*

$$A = \frac{8|E|\Delta K}{n(n-1)}.$$

As a result, if $t = n(A+1)(\log n + c)$,

$$d(t) \leq \alpha e^{-c}$$

for some universal constant $\alpha > 0$.

Proof. We apply Theorem 4.1. Here, if e is an edge of the graph (identified with an abuse of notation with a transposition in $\mathcal{S}(V)$), then $p(e) = (n-1)/(n|E|)$. If (x, y) are arbitrary vertices, the transposition can be constructed by first transposing successively all the edges in

γ_{ij} and then reversing these transpositions except for the last one. Thus $|(x, y)| \leq 2|\gamma_{xy}| \leq 2\Delta$. Therefore, by Theorem 4.1, we may take

$$\begin{aligned} A &= \max_{e \in E} \frac{1}{p(e)} \sum_{y \in S_n} |y| N(e, y) \tilde{p}(y) \\ &\leq \max_{e \in E} \frac{n|E|}{n-1} \sum_{y \in \tilde{S}} 2\Delta \frac{2}{n^2} N(e, y) \\ &\leq \frac{8\Delta|E|K}{n(n-1)} \end{aligned}$$

as $\max_{e \in E} \sum_{y \in \tilde{S}} \mathbf{1}_{\{e \in \gamma_y\}} = K$, by definition, and each fixed edge e appears at most twice in the path to the transposition $y \in \tilde{S}$. Applying the result of Diaconis and Shahshahani (Theorem 4.2) on random transpositions finishes the proof. \square

First application: Suppose \mathcal{G} is a segment of length n , so that the random walk P is the adjacent transposition process. Then \mathcal{G} is a tree so paths γ_{xy} are forced. Clearly $\Delta = n - 1$, $|E| = n - 1$, $K \leq 2(n/2)^2$. Thus Theorem 4.3 shows that if

$$t = 4n^3(\log n + c)$$

then $d(t) \leq \alpha e^{-c}$.

This example is known as the *random adjacent transpositions* process. It is easy to check that n^3 is necessary and guess that $n^3 \log n$ is the right order of magnitude. This example will be discussed further in the next lecture, devoted to Wilson's algorithm, where :

- A conjecturally sharp lower bound of $1/\pi^2 n^3 \log n$ is obtained (as an illustration of Wilson's method)
- An upper-bound using a coupling argument is shown. The upper-bound is twice the above, ie $(2/\pi^2)n^3 \log n$.

Second application: Suppose \mathcal{G} is the star graph. Then here again, \mathcal{G} is a tree so paths are forced. We clearly have $\Delta = 2$, $|E| = n - 1$, $K = 2(n - 1)$, hence $A \leq 32$. Thus for $t = 33n(\log n + c)$, we have $d(t) \leq \alpha e^{-c}$.

4.5 Example 2: a random walk on S_n

Consider the random walk generated by the identity, $(1, 2)$ and the n -cycle $(1, 2, \dots, n)$ as well as its inverse. Thus at each step, with probability $1/4$ we:

- do nothing
- exchange top two cards
- put the top card at bottom or the other way round.

Theorem 4.4. *If $t = 64n^3(\log n + c)$, then $d(t) \leq \alpha e^{-c}$.*

Proof. To see this, note that any transposition τ may be obtained by performing at most $4n$ moves from the generator of the walk P . Indeed, say we wish to build the transposition (i, j) with $i < j$. Moving $i - 1$ cards from top to bottom, the i th card sits at the top of the deck. Then transpose the two top cards and shift repeatedly until the i th card is just next to the j th card (at the top), and transpose them. Moving backward brings us back to the original deck but where cards i and j have been exchanged. At most $2n + 2n$ moves have been performed. Thus for y a transposition, $|y| \leq 4n$. Since $N(z, y) \leq |y|$ clearly, we deduce using (15) that

$$A \leq 4 \sum_y |y|^2 \tilde{p}(y) \leq 64n^2.$$

Using the Diaconis and Shahshahani result (Theorem 4.2) finishes the proof. \square

5 Wilson's method

5.1 Statement of the result

David Wilson devised a general method for proving lower bounds on the mixing times. As we will see, this can provide very sharp estimates in some examples. The idea is to produce a general function which will serve as a distinguishing statistics. The following lemma is elementary but rather tedious. Its proof can be found in Proposition 7.7 of [7] (note the typo in the statement, however).

Lemma 5.1. *Let μ, ν be two probability measures on a finite set S . Let f be a function on S and let $r \geq 0$ be such that*

$$|\mathbb{E}_\mu(f) - \mathbb{E}_\nu(f)| \geq r\sigma$$

where $\sigma^2 = \frac{1}{2}(\text{var}_\mu(f) + \text{var}_\nu(f))$. Then the total variation distance between μ and ν satisfy:

$$\|\mu - \nu\| \geq 1 - \frac{4}{4 + r^2}.$$

This is useful when r is large, in which case it is natural to expect that the TV distance between μ and μ is also large. We now state the result which is the basis of Wilson's method (see Theorem 13.5 in [7]).

Theorem 5.1. *Let X_t be an irreducible aperiodic Markov chain. Let Φ be an eigenfunction with eigenvalue $1/2 < \lambda < 1$. Fix $0 < \epsilon < 1$ and let $R > 0$ satisfy:*

$$\mathbb{E}_x(|\Phi(X_1) - \Phi(x)|^2) \leq R$$

for all $x \in S$. Then

$$t_{\text{mix}}(\epsilon) \geq \frac{1}{2 \log(1/\lambda)} \left[\log \left(\frac{(1-\lambda)\Phi(x)^2}{2R} \right) + \log \left(\frac{1-\epsilon}{\epsilon} \right) \right].$$

Proof. The proof is directly taken from [7]. Since Φ is an eigenfunction of P , we immediately get

$$\mathbb{E}_x(\Phi(X_t)) = \lambda^t \Phi(x). \tag{16}$$

Let $D_t = \Phi(X_{t+1}) - \Phi(X_t)$ be the difference process. Then we know

$$\mathbb{E}_x(D_t | X_t = z) = (\lambda - 1)\Phi(z)$$

and

$$\mathbb{E}_x(D_t^2 | X_t = z) \leq R$$

Therefore,

$$\begin{aligned} \mathbb{E}_x(\Phi(X_{t+1})^2) &= \mathbb{E}((\Phi(X_t) + D_t)^2) \\ &= \Phi(z)^2 + 2\mathbb{E}_x(D_t \Phi(z) | X_t = z) + \mathbb{E}_x(D_t^2 | X_t = z) \\ &\leq \Phi(z)^2(2\lambda - 1) + R \end{aligned}$$

so that taking the expectations, we find:

$$\mathbb{E}_x(\Phi(X_{t+1})^2) \leq (2\lambda - 1)\mathbb{E}_x(\Phi(X_t)^2) + R.$$

This is an inequality which may apply iteratively. This leads us to summing a certain geometric series. Or, more clearly (and equivalently), we may subtract $R/(2(1-\lambda))$ from both sides and get, noting $Z_t = \Phi(X_t)^2 - R/(2(1-\lambda))$,

$$\mathbb{E}_x(Z_{t+1}) \leq (2\lambda - 1)\mathbb{E}_x(Z_t).$$

Hence if $t \geq 0$,

$$\mathbb{E}_x(Z_t) \leq (2\lambda - 1)^t \left(\Phi(x)^2 - \frac{R}{2(1-\lambda)} \right),$$

and thus

$$\mathbb{E}_x(\Phi(X_t)^2) \leq (2\lambda - 1)^t (\Phi(x)^2) + \frac{R}{2(1-\lambda)}.$$

Using (16), this gives us:

$$\begin{aligned} \text{var}_x(\Phi(X_t)^2) &\leq [(2\lambda - 1)^t - \lambda^{2t}] \Phi(x)^2 + \frac{R}{2(1-\lambda)} \\ &\leq \frac{R}{2(1-\lambda)} \end{aligned}$$

since $2\lambda - 1 < \lambda$. (This may look crude, but what we are losing here is in practice very small). Note that as $t \rightarrow \infty$, we also get that

$$\text{var}_\pi(\Phi) \leq \frac{R}{2(1-\lambda)}.$$

We now wish to apply Lemma 5.1, with $\mu = P(x, \cdot)$ and $\nu = \pi$. Note then that

$$\mathbb{E}_\mu(\Phi) = P^t \Phi(x) = \lambda^t \Phi(x)$$

and that by orthogonality of eigenfunctions

$$\mathbb{E}_\nu(\Phi) = \sum_x \pi(x) \Phi(x) = 0$$

since Φ is an eigenfunction associated with $\lambda < 1$. Thus we may write

$$|\mathbb{E}_\mu(\Phi) - \mathbb{E}_\nu(\Phi)| \geq r\sigma$$

where r^2 is defined by

$$\begin{aligned} r^2 &= \frac{|\mathbb{E}_\mu(\Phi) - \mathbb{E}_\nu(\Phi)|^2}{\frac{1}{2} \text{var}_\mu f + \frac{1}{2} \text{var}_\nu f} \\ &\geq \frac{\lambda^{2t} \Phi(x)^2 2(1-\lambda)}{R} \end{aligned}$$

Thus by Lemma 5.1, we find:

$$\begin{aligned} \|P^t(x, \cdot) - \pi\| &\geq 1 - \frac{4}{4 + r^2} \\ &= \frac{(1-\lambda)\lambda^{2t}\Phi(x)^2}{2R + (1-\lambda)\lambda^{2t}\Phi(x)^2}. \end{aligned}$$

Thus if $t \geq \frac{1}{2\log(1/\lambda)} \left[\log \left(\frac{(1-\lambda)\Phi(x)^2}{2R} \right) + \log \left(\frac{1-\varepsilon}{\varepsilon} \right) \right]$, then

$$(1-\lambda)\lambda^{2t}\Phi(x)^2 \geq \frac{\varepsilon}{1-\varepsilon} 2R$$

and hence the total variation at this time must be greater than ε . □

5.2 Example: Random walk on a hypercube

We have already studied a random walk on the hypercube $\{0, 1\}^n$ by means of coupling (and have also computed the exact distribution of the walk at time t) but we return to it to illustrate Wilson's method on a first concrete example.

To find out the eigenfunctions and eigenvalues of the chain, it is easier to think of the walk as taking its values on $\{-1, 1\}^n$. Then the chain in continuous time can be thought of as choosing a random coordinate at rate 1 and putting a random sign. Since this is a "product" chain (i.e., n independent chains run in parallel), some fairly simple general theory tells us that any eigenfunction can be written as product of eigenfunctions over a given subset of coordinates. That is, must be of the following form: if J is a subset of $\{1, \dots, n\}$,

$$f_J(x) = \prod_{j \in J} f_j(x_j)$$

where f_j is an eigenfunction for the j^{th} chain. Here, the one-dimensional chains are trivial, with the function $f(x) = x$ being the only nontrivial eigenfunction. (This is where it is useful to view the state of a coordinate as a sign). Hence for $J \subset \{1, \dots, n\}$, the function

$$f_J(x) = \prod_{j \in J} x_j$$

is an eigenfunction of the *lazy* version of the chain, and the associated eigenvalue is

$$\lambda_J = \frac{\sum_{j=1}^n 1 - \mathbf{1}_{\{j \in J\}}}{n} = \frac{n - |J|}{n}.$$

This gives us all the eigenfunctions and hence

$$\gamma = \frac{1}{n} \text{ and hence } t_{\text{rel}} = n.$$

This is far from sharp, as we know from previous work that in fact $t_{\text{mix}} = \frac{1}{2}n \log n$ (because this is the lazy version of the chain).

Now, consider Wilson's method. We wish to take Φ an eigenfunction associated with the second largest eigenvalue. The associated eigenspace has dimension n (i.e., the number of choices of J such that $|J| = n - 1$). But a convenient representative is

$$\Phi(x) = W(x) - \frac{n}{2}$$

where $W(x)$ is the number of 1's in the string x . (You may easily check that this is an eigenfunction associated with $\lambda = 1 - 1/n$. Then

$$\mathbb{E}_x((\Phi(X_1) - \Phi(x))^2) = \frac{1}{2}$$

since Φ changes by exactly ± 1 whenever the chain actually moves (i.e., with probability $1/2$). Hence if we take $R = 1/2$ and the initial state to be the all 1's vector, then we find:

$$\begin{aligned} t_{\text{mix}}(\varepsilon) &\geq \frac{1}{-2 \log(1 - n^{-1})} \left[\log\{n^{-1}(\frac{n}{2})^2\} + \log\{(1 - \varepsilon)/\varepsilon\} \right] \\ &= \frac{1}{2}n \log n + O(n). \end{aligned}$$

This is, as explained before, indeed sharp.

5.3 Example: adjacent transpositions.

Random adjacent transpositions is the random walk on \mathcal{S}_n which results when the shuffling method consists in selecting a position $1 \leq i \leq n - 1$ at random and exchanging the two neighbouring cards at position i and $i + 1$. (Note that this is not done cyclically). To avoid problems we consider the lazy version of this chain as usual.

Heuristics. If you follows the trajectory of a single card, this like a delayed random walk on the segment $\{1, \dots, n\}$ with reflection at the boundaries. The card moves only with probability $1/(2n)$ if it is not at the boundary, so since it takes approximately n^2 units of time for a reflecting random walk to mix on $\{1, \dots, n\}$ we can expect a single card to mix in about $O(n^3)$. Maximizing over all n possible cards, we guess

$$t_{\text{mix}}(1/e) \approx n^3 \log n.$$

However, it seems that the cutoff phenomenon is not proved in this example. It seems that all that is known is the following:

Theorem 5.2. *Let*

$$t = (1 - \varepsilon) \frac{1}{\pi^2} n^3 \log n.$$

Then $d(t) \rightarrow 1$. On the other hand, if

$$t \geq (1 + \varepsilon) \frac{2}{\pi} n^3 \log n,$$

then $d(t) \rightarrow 0$.

The lower-bound is conjectured to be the correct mixing time. It is obtained through an application of Wilson's method which we now describe. For this we need to find a good eigenfunction for our distinguishing statistics as well as a good initial state.

Lemma 5.2. *Let ϕ be an eigenfunction for the "single card" chain. Fix $1 \leq k \leq n$ and let $\hat{\phi}(\sigma) = \phi(\sigma(k))$. Then $\hat{\phi}$ is an eigenfunction of the original random walk.*

This is trivial to prove but tells us that we can start looking for an eigenfunction for the single card chain (which is basically delayed reflecting random walk) and lift it to an eigenfunction on the symmetric group.

Now, reflecting random walk on the interval is easy to analyse. Indeed its eigenfunction can be obtained from those of the random walk on the one-dimensional torus simply by observing that the projection of random walk on the torus onto the x -coordinate forms such a reflecting walk. Thus, let M be the transition probability of random walk on the n -path with holding probability $1/2$ at the endpoints. Let P' be the transition matrix of the single card chain: thus

$$P' = \frac{1}{n-1} M + \frac{n-2}{n-1} I$$

Then

$$\varphi(k) = \cos\left(\frac{(2k-1)\pi}{2n}\right)$$

is an eigenfunction of M and thus of P' , with eigenvalue:

$$\lambda = \frac{1}{n-1} \cos\left(\frac{\pi}{n}\right) + \frac{n-2}{n-1} = 1 - \frac{\pi^2}{2n^3} + O(n^{-3}).$$

Thus $\sigma \in S_n \mapsto \varphi(\sigma(k))$ is an eigenfunction of the adjacent transposition walk for all $1 \leq k \leq n$. Since these eigenfunctions lie in the same eigenspace, we may define:

$$\Phi(\sigma) = \sum_{1 \leq k \leq n} \varphi(k)\varphi(\sigma(k)) \quad (17)$$

which is also an eigenfunction of the chain with eigenvalue λ . When $\sigma = \text{id}$ is the identity permutation, then it can be shown that

$$\Phi(\sigma) = \sum_{k=1}^n \cos\left(\frac{(2k-1)\pi}{2n}\right)^2 = \frac{n}{2}$$

(it can be shown that functions of the form (17) are necessarily maximised at $\sigma = \text{id}$). This is why we choose this specific Φ and this specific starting point: when Φ is small, we know we are far away from the identity.

Now, let us see what is the value of R in Theorem 5.1. For this we need to compute the effect of one adjacent transposition $(k-1, k)$ onto $\Phi(\sigma)$. Note that only two terms in (17) change. Thus

$$\begin{aligned} |\Phi(\sigma(k-1, k)) - \Phi(\sigma)| &= |\varphi(k)\varphi(\sigma(k-1)) + \varphi(k-1)\varphi(\sigma(k)) \\ &\quad - \varphi(k)\varphi(\sigma(k)) - \varphi(k-1)\varphi(\sigma(k-1))| \\ &= |\varphi(k-1) - \varphi(k)| |\varphi(\sigma(k)) - \varphi(\sigma(k-1))|. \end{aligned}$$

Now note that $|\varphi'(x)| \leq \pi/n$ so the first term is smaller than π/n , and that since $|\varphi(x)| \leq 1$ the second term is smaller than 2. Therefore,

$$|\Phi(\sigma(k-1, k)) - \Phi(\sigma)| \leq \sqrt{R} := \frac{2\pi}{n}.$$

To compute the lower-bound given by Theorem 5.1, note that

$$\begin{aligned} t_{\text{mix}}(\varepsilon) &\geq \frac{1}{-2\log(\lambda)} \left[\log\left(\frac{(1-\lambda)\Phi(x)^2}{2R}\right) + C_\varepsilon \right] \\ &= \frac{n^3}{\pi^2} \left[\log\left(\frac{\frac{\pi^2}{2n^3}(n/2)^2}{2(2\pi^2/n)}\right) + C_\varepsilon \right] \\ &= \frac{n^3}{\pi^2} \log n + C'_\varepsilon \end{aligned}$$

as claimed for the lower-bound.

Upper-bound by coupling. The following coupling was introduced by Aldous, but we follow the presentation in [7], 16.1.2. It is based on the single card chain as well. While this is not sharp (and not the sharpest known either), it still gives the correct order of magnitude for the mixing time. We prove that

$$\text{if } t = 2n^3 \log_2 n \text{ then } d(t) \rightarrow 0. \quad (18)$$

Assume that we have two decks σ_t and σ'_t (we think of left and right decks) and that a is a fixed card $1 \leq a \leq n$. We wish to put card a at the same position in both decks. (We will later maximise over all possible $1 \leq a \leq n$.) The coupling is the following. Choose a position $1 \leq i \leq n-1$ at random in the deck: we are considering whether to perform the transposition $(i, i+1)$ on each deck. (This must be done with probability $1/2$ for each deck.)

- If $\sigma_t(i) = \sigma'_t(i+1)$ or $\sigma_t(i+1) = \sigma'_t(i)$ then perform the opposite things on the left and right deck: transpose on the right if the left stays still, and vice versa.
- Otherwise, perform the same action on both decks.

Let D_t denote the distance between the positions of the cards in both decks, and observe that once $D_t = 0$, then this stays true forever i.e. the cards are matched. The point is that D_t is approximately a Markov chain, where D_t can change with probability $1/(n-1) + 1/(n-1) = 2/(n-1)$ (the first term is the probability that the left card moves, the right is the probability that the right card moves) if both cards are at the interior and at distance $D_t > 1$. When D_t moves, it is equally likely to move up or down. However if one of the two cards is at the top or at the bottom then the distance may not increase. Thus in general,

$$\mathbb{P}(D_t = d + 1 | \sigma_t, \sigma'_t, D_t = d) \leq M(d, d + 1)$$

and

$$\mathbb{P}(D_t = d - 1 | \sigma_t, \sigma'_t, D_t = d) = M(d, d - 1)$$

where M is the transition matrix described above. Even though D_t is not a Markov chain, it is stochastically bounded above by the random walk Y_t with transition matrix M . It is not hard to prove that if τ is the first time that $Y = 0$, then we have:

$$\mathbb{E}_k(\tau) \leq \frac{(n-1)n^2}{2}$$

no matter what the starting point of Y is. Thus if τ_a is the first time $D_t = 0$, we have $\mathbb{E}(\tau_a) \leq (n-1)n^2/2$ as well. Therefore, by Markov's inequality:

$$\mathbb{P}(\tau_a > n^3) \leq \frac{1}{2}. \tag{19}$$

Suppose we run the chain for blocks of time of duration n^3 each, and we run $2 \log_2 n$ such blocks. Since (19) is independent of the starting point, the probability that $\tau_a > 2 \log_2 n n^3$ is smaller than the probability that it didn't couple in any of these runs, and hence:

$$\mathbb{P}(\tau_a > 2n^3 \log_2 n) \leq \left(\frac{1}{2}\right)^{2 \log_2 n} = \frac{1}{n^2}.$$

Now, maximising over all possible choices of a ,

$$\mathbb{P}\left(\max_{1 \leq a \leq n} \tau_a > 2n^3 \log_2 n\right) \leq n \frac{1}{n^2} = \frac{1}{n}.$$

But note that if $t \geq \max_{1 \leq a \leq n} \tau_a$, the decks are identical, and hence

$$\text{if } t = 2n^3 \log_2 n \text{ then } d(t) \leq 1/n \rightarrow 0$$

as claimed.

6 Nash inequalities.

6.1 Abstract result

We have already seen that a Poincaré inequality (i.e., a control on the spectral gap) was not sharp up to logarithms when we use the standard relation between spectral gap and mixing times (Theorem 2.2). In previous cases, such as the random walk on the circle, we overcame this difficulty by using an explicit control on *all eigenvalues* and symmetry properties of the graph (essentially, vertex-transitivity).

The following result is what people usually refer to as Nash’s theorem, although this isn’t the language in which it was stated, and uses a slightly sharpened version of the Poincaré inequality, which doesn’t lead to a $\log n$ loss when translating to mixing times. We start with defining what is a Nash inequality.

Definition 6.1. *Assume that (K, π) is irreducible and reversible. We say that it satisfies a Nash inequality if, for all $g \in \ell^2(\pi)$,*

$$\mathrm{var}_\pi(g)^{1+2/d} \leq C\mathcal{E}(g, g)\|g\|_1^{4/d}. \quad (20)$$

We will see that in practice, d often represents the “true dimension” of the ambient space and that C is a constant of the order of the relaxation time.

Theorem 6.1. *Then for all $t > 0$,*

$$\|h_t^x - 1\|_2 \leq \left(\frac{dC}{4t}\right)^{d/4}.$$

and for all $t > 0$,

$$|h_t(x, y) - 1| \leq \left(\frac{dC}{2t}\right)^{d/2}.$$

Proof. Fix f a function such that $\|f\|_1 = 1$, and set

$$u(t) = \|H_t(f - \pi f)\|_2^2 = \mathrm{var}_\pi(H_t f).$$

By our assumption (20), we have

$$u(t)^{1+2/d} \leq C\mathcal{E}(H_t f)\|H_t f\|_1^{4/d}$$

and we have already seen that $\mathcal{E}(H_t f) = -\frac{1}{2}u'(t)$. (Here $\mathcal{E}(h) := \mathcal{E}(h, h)$).

Note also that since $\|f\|_1 \leq 1$, and using reversibility:

$$\begin{aligned} \|H_t f\|_1 &= \sum_x |H_t f(x)|\pi(x) \\ &= \sum_x \left| \sum_y H_t(x, y)f(y) \right| \pi(x) \\ &\leq \sum_{x, y} H_t(x, y)|f(y)|\pi(x) \\ &= \sum_y \pi(y)|f(y)| \sum_x H_t(y, x) \\ &= \|f\|_1 \leq 1. \end{aligned}$$

So we conclude that:

$$u(t)^{1+2/d} \leq -\frac{C}{2}u'(t).$$

Thus if we set $v(t) = (dC/4)u(t)^{-2/d}$, we have: $v'(t) \geq 1$ for all t and hence since $v(0) \geq 0$, this implies $v(t) \geq t$ for all t . From this it follows that

$$\|H_t(f - \pi f)\|_2 \leq \left(\frac{dC}{4t}\right)^{d/4}$$

for all f such that $\|f\|_1 = 1$. But note that this inequality is scale invariant, so it must hold for all f . Furthermore, specializing to $f = f_x(y) = \frac{\mathbf{1}_{\{y=x\}}}{\pi(y)}$, note that $H_t f(y) = h_t(x, y)$ by reversibility and that $\mathbb{E}_\pi(f) = 1$, so we obtain the first part of the conclusion, which is:

$$\|h_t^x - 1\|_2 \leq \left(\frac{dC}{4t}\right)^{d/4}.$$

Using Lemma ??, together with Cauchy-Schwarz's inequality, this immediately entails the second part of the conclusion. \square

Remark. The constant C in (20) must satisfy:

$$C \geq 1/\gamma. \tag{21}$$

Indeed, if g is such that $\mathbb{E}_\pi(g) = 0$ then by (20) we have:

$$\text{var}_\pi(g)^{1+2/d} \leq C\mathcal{E}(g)\|g\|_1^{4/d}$$

so by Jensen's inequality (i.e., Cauchy-Schwartz):

$$\text{var}_\pi(g)^{1+2/d} \leq C\mathcal{E}(g)\|g\|_2^{4/d}$$

But note that $\|g\|_2^2 = \text{var}_\pi g$ so this actually means:

$$\text{var}_\pi(g) \leq C\mathcal{E}(g)$$

which stays true even if $\mathbb{E}_\pi(g) \neq 0$ by adding a suitable constant to g . Since on the other hand by the variational formulation we know

$$\gamma = \min_{\text{var}_\pi g \neq 0} \frac{\mathcal{E}(g)}{\text{var}_\pi(g)}$$

this implies $\gamma \geq 1/C$ which is (21).

The conclusion of Theorem 6.1 is usually strong when t is not too large, otherwise Theorems ?? and ?? typically take over. Taking this into account leads to the slightly refined estimate:

Corollary 6.1. *If Nash's inequality (20) is satisfied, then*

$$\|h_t^x - 1\|_2 \leq \min \left\{ \left(\frac{dC}{4t}\right)^{d/4}, e^{-\gamma(t-dC/4)} \right\}.$$

See Corollary 2.3.2 in [10] for a proof.

6.2 Example

Basically, Nash inequalities are excellent tools to deal with situation where the geometry is subtle. However at this stage we haven't yet developed any of the corresponding useful geometric tools, so our example is fairly basic. Consider the interval $S = \{-n, \dots, n\}$, and consider simple random walk on S with holding probability $1/2$ at both hands. (This walk is always aperiodic). The uniform measure $\pi(x) \equiv 1/(2n+1)$ is stationary and even reversible. If $f : S \rightarrow \mathbb{R}$, the Dirichlet form is:

$$\mathcal{E}(f) = \frac{1}{2n+1} \sum_{i=-n}^{n-1} |f(i+1) - f(i)|^2.$$

Now, it is obvious that

$$|\max f - \min f| \leq \sum_{i=-n}^{n-1} |f(i+1) - f(i)|$$

so if f is not of constant, in particular

$$\|f\|_\infty \leq \sum_{i=-n}^{n-1} |f(i+1) - f(i)|.$$

Let g be such that $\mathbb{E}_\pi(g) = 0$, and define a function $f = \text{sgn}(g)g^2$. Then writing $\Delta f(i)$ for the increment $f(i+1) - f(i)$, it follows

$$|\Delta f(i)| \leq |\Delta g(i)|(|g(i+1)| + |g(i)|)$$

so that by Cauchy-Schwartz:

$$\|f\|_\infty \leq \left(\sum_i |\Delta g(i)|^2 \right)^{1/2} \left(\sum_i (|g(i+1)| + |g(i)|)^2 \right)^{1/2}$$

The first term is nothing but $\mathcal{E}(g)^{1/2}(2n+1)^{1/2}$, while the second is smaller than $2\|g\|_2$ by Cauchy-Schwartz's inequality, so we obtain:

$$\|f\|_\infty = \|g\|_\infty^2 \leq 2(2n+1)\mathcal{E}(g)^{1/2}\|g\|_2$$

Using Hölder's inequality:

$$\begin{aligned} \|g\|_2^4 &\leq \|g\|_\infty^2 \|g\|_1^2 \\ &\leq 2(2n+1)\mathcal{E}(g)^{1/2} \|g\|_2 \|g\|_1^2 \end{aligned}$$

and thus, dividing by $\|g\|_2$ we get:

$$\|g\|_2^3 \leq 2(2n+1)\mathcal{E}(g)^{1/2} \|g\|_1^2$$

Since g has mean 0, this is the same as

$$(\text{var}_\pi g)^3 \leq 4(2n+1)^2 \mathcal{E}(g) \|g\|_1^4$$

after squaring. Changing g into $\tilde{g} + m$ with $m = \mathbb{E}(g)$ and $\mathbb{E}(\tilde{g}) = 0$, we see that $\|\tilde{g}\|_1 \leq 2\|g\|_1$ so the following holds for all g :

$$(\text{var}_\pi g)^3 \leq 64(2n+1)^2 \mathcal{E}(g) \|g\|_1^4.$$

This is Nash's inequality (20) with $d = 1$ (the dimension of the space) and $C = 64(2n+1)^2$. Thus, by (21),

$$\gamma \geq \frac{1}{64(2n+1)^2}.$$

Nash's theorem tells us:

$$\|h_t^x - 1\|_2 \leq \left(\frac{64(2n+1)^2}{2t} \right)^{1/4}$$

which is the right order of magnitude, while the spectral gap estimate (??) only gives:

$$\|h_t^x - 1\|_2 \leq \sqrt{2n+1} e^{-t/(64(2n+1)^2)}$$

which is off because of the square root in front (it shows that roughly $n^2 \log n$ units of time are enough to mix, which is more than necessary).

7 Evolving sets and martingales.

Evolving sets is an auxiliary process with values in the subsets of the state space V , which was introduced by Morris and Peres in 2005. They can be used to prove some remarkable general results about mixing times, which we now describe.

The setup is as follows: we have a countable state space V with irreducible aperiodic transition probability $p(x, y)$ and stationary distribution $\pi(x)$. We define the equilibrium flow from x to y as

$$Q(x, y) = \pi(x)p(x, y)$$

which is a slight change compared to our previous notion of flow in the previous chapters. (We used to take $Q(e) = \frac{1}{2}(\pi(x)K(x, y) + \pi(y)K(y, x))$. Thus the two definitions coincide when the chain is reversible). If $S \subset V$, we further define $Q(S, y) = \sum_{x \in S} Q(x, y)$.

7.1 Definition and properties

Definition 7.1. *The evolving set process is a set-valued Markov chain $(S_n, n \geq 0)$, whose transition probabilities are as follow. Given $S_n = S \subset V$, pick U a uniform random variable on $(0, 1)$. Then $S_{n+1} = \tilde{S}$ where*

$$\tilde{S} = \{y \in V : Q(S, y) \geq U\pi(y)\}.$$

Note that an immediate consequence of this definition is that if $y \in V$, then

$$\mathbb{P}(y \in S_{n+1} | S_n = S) = \mathbb{P}(Q(S, y) \geq U\pi(y)) = \frac{Q(S, y)}{\pi(y)}.$$

To get a feel for how this chain works, consider the example where V is given by the $n \times n$ torus in 2 dimensions, and X is the lazy chain: that is, it stays wherever it is with probability $1/2$ and move to a randomly chosen neighbour with probability $1/2$. (Thus the chain is irreducible and aperiodic). The stationary distribution π is then uniform. Thus a given point y belongs to \tilde{S} if and only if $\sum_{x \in S} p(x, y) > U$. Now, if x is a neighbour from y , then $p(x, y) = 1/8$, while if $x = y$, $p(x, y) = 1/2$. Thus if $U < 1/2$, the set will grow. If $1/8 < U < 2/8$ in the example below, any point on the boundary of S is added provided that it has at least two neighbours. If on the other hand, $6/8 < U < 7/8$ then only points in S with at least three neighbours in S will be kept next round. This is illustrated in the picture below.

We now state some of the properties of the evolving set process. The first is martingale property which shall be very useful in the following.

Lemma 7.1. *The sequence $\{\pi(S_n)\}_{n \geq 0}$ is a martingale.*

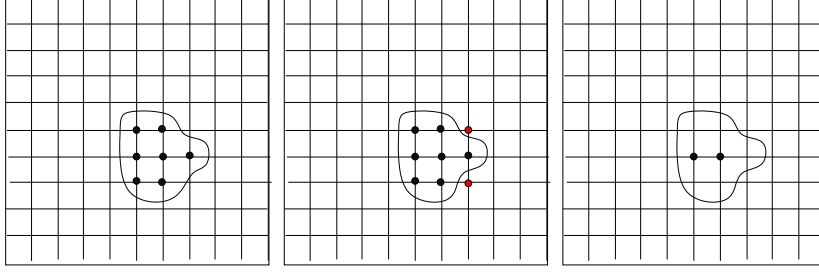


Figure 3: The initial state of the evolving set and two possible transitions: (a) $1/8 < U < 2/8$ and (b) $6/8 < U < 7/8$.

Proof.

$$\begin{aligned}
\mathbb{E}(\pi(S_{n+1})|S_n) &= \sum_{y \in V} \pi(y) \mathbb{P}(y \in S_{n+1}|S_n) \\
&= \sum_{y \in V} \pi(y) \frac{Q(S_n, y)}{\pi(y)} \\
&= \sum_{y \in V} \pi(S_n) p(S_n, y) \\
&= \pi(S_n) \sum_{y \in V} p(S_n, y) = \pi(S_n).
\end{aligned}$$

□

The next lemma relates the evolving set to the transition probabilities of the Markov chain:

Lemma 7.2. *For all $n \geq 0$, we have:*

$$p^n(x, y) = \frac{\pi(y)}{\pi(x)} \mathbb{P}_x(y \in S_n).$$

Here \mathbb{P}_x means that the evolving set starts at $S_0 = \{x\}$.

Proof. The proof proceeds by induction. The case $n = 0$ is trivial so assume that $n \geq 1$ and

that the result is true for $n - 1$. Then by decomposing on the state of the chain at time $n - 1$,

$$\begin{aligned}
p^n(x, y) &= \sum_z p^{n-1}(x, z)p(z, y) \\
&= \sum_z \frac{\pi(z)}{\pi(x)} \mathbb{P}_x(z \in S_{n-1})p(z, y) \\
&= \frac{\pi(y)}{\pi(x)} \sum_z \underbrace{\pi(z)p(z, y)}_{=Q(z, y)} \frac{1}{\pi(y)} \mathbb{P}_x(z \in S_{n-1}) \\
&= \frac{\pi(y)}{\pi(x)} \mathbb{E}_x \left(\frac{1}{\pi(y)} Q(S_{n-1}, y) \right) \\
&= \frac{\pi(y)}{\pi(x)} \mathbb{E}_x (\mathbb{P}_x(y \in S_n | S_{n-1})) \\
&= \frac{\pi(y)}{\pi(x)} \mathbb{P}_x(y \in S_n).
\end{aligned}$$

□

The next lemma states a duality property between S_n and S_n^c :

Lemma 7.3. *The complement S_n^c of the evolving set is also an evolving set process with the same transition probabilities.*

Proof. Fix $n \geq 0$. Note that $Q(S_n, y) + Q(S_n^c, y) = Q(V, y) = \pi(y)$ by stationarity. Therefore, $Q(S_n^c, y) = \pi(y) - Q(S_n, y)$. It follows that if U is the random variable used for the construction of S_{n+1} given S_n ,

$$\begin{aligned}
S_{n+1}^c &= \{y \in V : Q(S_n, y) < U\pi(y)\} \\
&= \{y \in V : \pi(y) - Q(S_n^c, y) < U\pi(y)\} \\
&= \{y \in V : Q(S_n^c, y) > (1 - U)\pi(y)\}
\end{aligned}$$

Since $U \stackrel{d}{=} 1 - U$, S_{n+1}^c has the same transition probabilities as the original evolving set. □

We may now start to describe the relationship between evolving sets and mixing. We start by defining the ℓ^2 -distance between μ and π , where π is a distribution on V , $\chi(\mu, \pi)$:

$$\chi(\mu, \pi) = \left(\sum_{y \in V} \pi(y) \left[\frac{\mu(y)}{\pi(y)} - 1 \right]^2 \right)^{1/2}$$

To make sense of this definition, note that $\chi(\mu, \pi)^2$ is the second moment (with respect to π) of the Radom-Nikodyn derivative of μ with respect to π , $\mu(y)/\pi(y)$, minus 1. This derivative would be exactly 1 if $\mu \equiv \pi$ so $\chi(\pi, \pi) = 0$. It turns out that χ is a distance, and is a stronger way to measure distance to stationarity than the total variation distance, as is shown by the

following computation:

$$\begin{aligned}
\|\mu - \pi\| &= \frac{1}{2} \sum_{y \in V} |\mu(y) - \pi(y)| \\
&= \frac{1}{2} \sum_{y \in V} \pi(y) \left| \frac{\mu(y)}{\pi(y)} - 1 \right| \\
&\leq \frac{1}{2} \chi(\mu, \pi)
\end{aligned}$$

by Cauchy-Schwarz's inequality. Thus if $\chi(\mu, \pi)$ is small, then so is $\|\mu - \pi\|$. Note furthermore that by expanding the square in the definition of $\chi(\mu, \pi)$, we have

$$\chi(\mu, \pi)^2 = \sum_y \frac{\mu^2(y)}{\pi(y)} - 1.$$

We introduce the following notation:

$$S^\# = \begin{cases} S & \text{if } \pi(S) \leq 1/2 \\ S^c & \text{otherwise} \end{cases} \quad (22)$$

Lemma 7.4. *Let $\mu_n = p^n(x, \cdot)$ be the distribution of the Markov chain after n steps started from x . Then*

$$\chi(\mu_n, \pi) \leq \frac{1}{\pi(x)} \mathbb{E}_{\{x\}} \left(\sqrt{\pi(S_n^\#)} \right). \quad (23)$$

Proof. The idea is to introduce two replicas (independent copies) of the evolving set process S_n and Λ_n . Then note that

$$\begin{aligned}
\chi(\mu_n, \pi)^2 &= \left(\sum_y \frac{\mu_n(y)^2}{\pi(y)} \right) - 1 \\
&= \left(\sum_y \left[\frac{\pi(y)}{\pi(x)} \mathbb{P}_x(y \in S_n) \right]^2 \frac{1}{\pi(y)} \right) - 1 \\
&= \sum_y \frac{\pi(y) \mathbb{P}_x(y \in S_n)^2}{\pi(x)^2} - 1 \\
&= \frac{1}{\pi(x)^2} \left(\sum_y \pi(y) \mathbb{P}_x(y \in S_n)^2 - \pi(x)^2 \right) \\
&= \frac{1}{\pi(x)^2} \left(\sum_y \pi(y) \mathbb{P}_x(y \in S_n; y \in \Lambda_n) - \pi(x)^2 \right)
\end{aligned}$$

Now, recall that by the martingale property, $\pi(x) = \mathbb{E}_x(\pi(S_n))$, so that by independence between S_n and Λ_n , the above may be written as

$$\chi(\mu_n, \pi)^2 = \frac{1}{\pi(x)^2} \mathbb{E}_x \left(\pi(S_n \cap \Lambda_n) - \pi(S_n)\pi(\Lambda_n) \right)$$

On the other hand, for any sets $\Lambda, S \subset V$ we always have

$$\pi(\Lambda) = \pi(\Lambda)\pi(S) + \pi(\Lambda)\pi(S^c)$$

and

$$\pi(\Lambda) = \pi(\Lambda; S) + \pi(\Lambda; S^c)$$

so that

$$|\pi(S \cap \Lambda) - \pi(S)\pi(\Lambda)| = |\pi(S^c)\pi(\Lambda) - \pi(S^c \cap \Lambda)|$$

But note that the expression in the right-hand side is invariant if one replaces Λ by Λ^c or S by S^c . Therefore,

$$|\pi(S \cap \Lambda) - \pi(S)\pi(\Lambda)| \leq |\pi(S^\sharp)\pi(\Lambda^\sharp) - \pi(S^\sharp \cap \Lambda^\sharp)|$$

Letting $p = \pi(S^\sharp) \wedge \pi(\Lambda^\sharp)$, this means

$$\begin{aligned} |\pi(S \cap \Lambda) - \pi(S)\pi(\Lambda)| &\leq |\pi(S^\sharp)\pi(\Lambda^\sharp) - \pi(S^\sharp \cap \Lambda^\sharp)| \\ &\leq |p - p^2| \leq p \leq \sqrt{\pi(S^\sharp)\pi(\Lambda^\sharp)} \end{aligned}$$

whence

$$\chi(\mu_n, \pi)^2 = \frac{1}{\pi(x)^2} \mathbb{E}_x \left(\sqrt{\pi(S_n^\sharp)\pi(\Lambda_n^\sharp)} \right)$$

and therefore, by independence:

$$\chi(\mu_n, \pi) = \frac{1}{\pi(x)^2} \mathbb{E}_x \left(\sqrt{\pi(S_n^\sharp)} \right)$$

which ends the proof. \square

It is interesting to think about the last result in the case where V is say finite. The evolving set process is a Markov chain where the only two absorbing states are the empty set and states otherwise communicate. Hence S_n eventually gets absorbed in one of those two states. When this happens, then $S_n^\sharp = \emptyset$, so (23) suggests that the distance is then close to 0. This idea can be carried further to construct what is known as a *strong stationary time*, i.e., a random time T such that $X_T \stackrel{d}{=} \pi$ exactly, and moreover T is independent of X_T . See section 17.7 in [7] for more information about this.

7.2 Evolving sets as a randomized isoperimetric profile

For a set $S \subset V$, let \tilde{S} denote a step of the evolving set process started from S . Define the *boundary gauge*:

$$\Psi(S) = 1 - \mathbb{E}_S \left[\sqrt{\frac{\pi(\tilde{S})}{\pi(S)}} \right]$$

and let

$$\psi(r) = \begin{cases} \inf\{\Psi(S) : \pi(S) \leq r\} & \text{if } r \in [\pi^*, \frac{1}{2}] \\ \psi(1/2) & \text{otherwise.} \end{cases}$$

Here, π^* denotes as usual the minimum value of the stationary distribution $\pi^* = \inf_{x \in V} \pi(x)$.

Note that $\psi(r)$ is non-increasing on $r \geq \pi^*$. The definition of Ψ and ψ is reminiscent of the definition of the isoperimetric constant I in Lecture 6. In fact, intuitively speaking $\psi(r)$ is essentially a “randomized isoperimetric constant” among all sets of mass smaller than r (where mass is measured in terms of the stationary distribution). It is randomized in the sense that we don’t simply measure boundary over volume $Q(S, S^c)/\pi(X)$ as we do for the isoperimetric constant, but we compare the masses of \tilde{S} with S , where \tilde{S} is chosen according to the evolving set rules. $\psi(r)$ can thus be thought of as a *randomized isoperimetric profile*. We will see more about this line of thought in the next subsection.

The following result gives us an explicit upper-bound for the mixing times of the chain in terms of this function ψ .

Theorem 7.1. *Let $x \in V$ and let $\mu_n = p^n(x, \cdot)$. Then for all $\varepsilon > 0$,*

$$\chi(\mu_n, \pi)^2 \leq \varepsilon \text{ for all } n \geq \int_{4\pi(x)}^{4/\varepsilon} \frac{du}{u\psi(u)}.$$

In particular

$$t_{\text{mix}}(\sqrt{\varepsilon}) \leq \int_{4\pi^*}^{4/\varepsilon} \frac{du}{u\psi(u)}.$$

Proof. Let $K(S, A)$ denote the transition kernel of the evolving set process. We define the transformed kernel

$$\hat{K}(S, A) = \frac{\pi(A)}{\pi(S)} K(S, A),$$

for which it is easy to check that this is also a transition kernel. To explain the definition, we note that starting from a state S , the probability that S_n will get absorbed by V rather than by \emptyset is, by the optional stopping theorem, $\pi(S)$, since $\pi(S_n)$ is a martingale. Thus the transition of \hat{K} are those of K weighted by the probability that, starting from the new state A , the chain will eventually get absorbed by V rather than by \emptyset . Doob’s theory of h -transforms tells us that this is indeed the transition probabilities of the Markov chain S_n conditioned on eventual absorption by V .

Moreover, by induction on n

$$\hat{K}^n(S, A) = \frac{\pi(A)}{\pi(S)} K^n(S, A)$$

and thus for any nonnegative function f :

$$\hat{E}_S(f(S_n)) = \mathbb{E}_s \left(\frac{\pi(S_n)}{\pi(S)} f(S_n) \right)$$

by the monotone class theorem. Let $Z_n = \sqrt{\pi(S_n^\#)/\pi(S_n)}$. Then if $\pi(S_n) \leq 1/2$, we have $Z_n = \sqrt{1/\pi(S_n)}$, i.e., $\pi(S_n) = Z_n^{-2}$ for $\pi(S_n) < 1/2$.

Using (23), we get:

$$\begin{aligned} \chi(\mu_n, \pi) &\leq \frac{1}{\pi(x)} \mathbb{E}_x \sqrt{\pi(S_n^\#)} = \mathbb{E}_x \left(\frac{\pi(S_n)}{\pi(x)} \frac{\sqrt{\pi(S_n^\#)}}{\pi(S_n)} \right) \\ &\leq \hat{\mathbb{E}}_x \left(\frac{\sqrt{\pi(S_n^\#)}}{\pi(S_n)} \right) = \hat{\mathbb{E}}_x(Z_n). \end{aligned}$$

Thus to control the ℓ^2 distance it suffices to have good bounds on $\hat{E}_x(Z_n)$. However, we have the following lemma.

Lemma 7.5. *Let $f : [0, \infty) \rightarrow [0, 1]$ be a nondecreasing function. Suppose that Z_n is a sequence of random variables such that $Z_0 = \mathbb{E}(Z_0) = L_0$ (say), and for all $n \geq 0$:*

$$\mathbb{E}(Z_{n+1}|Z_n) \leq Z_n(1 - f(Z_n)).$$

Then for every $n \geq \int_{\delta}^{L_0} \frac{2dz}{zf(z/2)}$ we have $\mathbb{E}(Z_n) \leq \delta$.

Proof. The proof is split into two steps. The first step is to show that if $L_n = \mathbb{E}(Z_n)$, then

$$L_{n+1} \leq L_n(1 - g(L_n)) \tag{24}$$

where $g(u) = \frac{1}{2}f(u/2)$. Indeed, if $A = \{Z_n > \mathbb{E}(Z_n)/2\}$, then

$$\mathbb{E}(Z_n \mathbf{1}_{\{A^c\}}) \leq \frac{1}{2}\mathbb{E}(Z_n)$$

so

$$\mathbb{E}(Z \mathbf{1}_{\{A\}}) \geq \frac{1}{2}\mathbb{E}(Z).$$

Thus since g is nondecreasing:

$$\mathbb{E}(Z_n g(2Z_n)) \geq \mathbb{E}(Z_n \mathbf{1}_{\{A\}} g(L_n)) = \frac{1}{2}L_n g(L_n).$$

On the other hand,

$$\mathbb{E}(Z_{n+1} - Z_n) \leq -\mathbb{E}(Z_n f(Z_n)) = -2\mathbb{E}(Z_n g(2Z_n)) \leq -L_n g(L_n)$$

which proves the claim.

The second step is as follows. Note that it suffices to prove that

$$\int_{L_n}^{L_0} \frac{dz}{zf(z)} \geq n.$$

However,

$$L_{n+1} \leq L_n(1 - g(L_n)) \leq L_n e^{-g(L_n)},$$

so

$$\int_{L_{k+1}}^{L_k} \frac{dz}{zf(z)} \geq \frac{1}{f(L_k)} \int_{L_{k+1}}^{L_k} \frac{dz}{z} = \frac{1}{f(L_k)} \log \frac{L_k}{L_{k+1}} \geq 1.$$

Summing up over $k \in \{0, \dots, n-1\}$ gives the result. □

End of the proof of Theorem 7.1. Let us compute $\hat{\mathbb{E}}_x(Z_{n+1}/Z_n|S_n)$.

$$\begin{aligned} \hat{\mathbb{E}}_x \left(\frac{Z_{n+1}}{Z_n} \middle| S_n \right) &= \mathbb{E}_x \left(\frac{\pi(S_{n+1})}{\pi(S_n)} \frac{Z_{n+1}}{Z_n} \middle| S_n \right) \\ &= \mathbb{E}_x \left(\frac{\sqrt{\pi(S_{n+1}^\sharp)}}{\sqrt{\pi(S_n^\sharp)}} \middle| S_n \right) \\ &= 1 - \Psi(S_n^\sharp). \end{aligned}$$

Note that $\Psi(S_n^\sharp) \geq \psi(\pi(S_n^\sharp))$

$$\hat{\mathbb{E}}_x \left(\frac{Z_{n+1}}{Z_n} \middle| S_n \right) \leq 1 - \psi(\pi(S_n^\sharp))$$

Now, ψ is non-increasing so $1 - \psi(\cdot)$ is nondecreasing. On the other hand we note that it is always true that $\pi(S_n^\sharp) \leq Z_n^{-2}$. (It is an equality if $\pi(S_n) \leq 1/2$.) Indeed, this is equivalent to saying

$$\pi(S_n^\sharp) \leq \frac{\pi(S_n)^2}{\pi(S_n^\sharp)}$$

or equivalently, $\pi(S_n^\sharp) \leq \pi(S_n)$, which is obviously true. Thus by monotonicity we get

$$\hat{\mathbb{E}}_x \left(\frac{Z_{n+1}}{Z_n} \middle| S_n \right) \leq 1 - \psi(Z_n^{-2})$$

and note that if $f(z) = \psi(1/z^2)$, which is nondecreasing, then Lemma 7.5 tells us that if $L_0 = Z_0 = \pi(x)^{-1/2}$, then $\hat{E}(Z_n) \leq \delta$ for all

$$n \geq \int_\delta^{\pi(x)^{-1/2}} \frac{2dz}{z\psi(4/z^2)},$$

or, after making the change of variable $u = 4/z^2$,

$$n \geq \int_{\pi(x)}^{4/\delta^2} \frac{du}{u\psi(u)}.$$

Thus since $\hat{E}(Z_n) = \chi(\mu_n, \pi)$, taking $\delta = \sqrt{\varepsilon}$, we get $\chi(\mu_n, \pi)^2 \leq \varepsilon$ for all

$$n \geq \int_{\pi(x)}^{4/\varepsilon} \frac{du}{u\psi(u)}.$$

This finishes the proof of Theorem 7.1. □

7.3 The conductance profile

Theorem 7.1 can be used to prove a bound on mixing times which is slightly more intuitive than the above, and which is given in terms of the *conductance profile* of the chain. Let us briefly discuss these notions. We have seen in Lecture 6 on geometric tools II how the isoperimetric constant

$$I = \min_{S \subset V, \pi(S) \leq 1/2} \frac{Q(S, S^c)}{\pi(S)}$$

can be used to bound the spectral gap: we have $\gamma \geq I^2/8$ and thus

$$t_{\text{mix}}(\varepsilon) \leq t_{\text{rel}} \log((\pi_* \varepsilon)^{-1}) \leq 8I^{-2} \left(\log \frac{1}{\pi_*} + \log \frac{1}{\varepsilon} \right). \quad (25)$$

The quantity $\Phi_S := Q(S, S^c)/\pi(S)$ is called the conductance of a set S . One idea that emerged in the late 90's is that generally speaking, sets which are “small” (in the sense of

stationary distribution, say) tend to have a higher conductance, so it is very pessimistic to always bound it below by I . Instead, it was suggested to consider the *isoperimetric profile* or *conductance profile*

$$\Phi(r) = \inf\{\Phi_S : \pi(S) \leq r\}.$$

It should thus be possible to prove a bound which use the decreasing function $\Phi(r)$ rather than the constant function $I = \Phi(1/2)$. Morris and Peres were able to use evolving sets to prove the following result. Recall the first that the separation distance $s(\mu, \pi) = \max_y(1 - \mu(y)/\pi(y))$ is such that $\|\mu - \pi\| \leq s(\mu, \pi)$.

Theorem 7.2. *Assume that the chain is irreducible and that $p(x, x) \geq 1/2$ for all $x \in V$ (in particular, it is aperiodic). Then for all n such that*

$$n \geq 1 + \int_{4\pi^*}^{4/\varepsilon} \frac{4du}{u\Phi^2(u)}$$

then

$$\left| \frac{p^n(x, y) - \pi(y)}{\pi(y)} \right| \leq \varepsilon.$$

In particular, $s(\mu_n, \pi) \leq \varepsilon$ and thus $\|\mu_n - \pi\| \leq \varepsilon$.

Note that, using the monotonicity of $\Phi(u)$ (which is weakly decreasing with u) we have $\Phi(u) \geq I$ for all $u \leq 1/2$, so we find better bounds than (25).

Proof. The proof is essentially a consequence of Theorem 7.1, and of the following lemma which relates the conductance Φ_S to the boundary gauge $\Psi(S)$ used in the previous theorem:

Lemma 7.6. *Let $S \subset V$ be nonempty, and assume that $p(x, x) \geq 1/2$. Then*

$$\Psi(S) = 1 = \mathbb{E}_S \sqrt{\frac{\pi(\tilde{S})}{\pi(S)}} \geq \frac{1}{2} \Phi_S^2.$$

In particular, $\Phi(r)^2 \leq 2\psi(r)$ for all $r \in [\pi^*, 1/2]$.

See section 4 of the original paper of Morris and Peres for a proof of this result.

Let us now turn to the proof of Theorem 7.2. First recall the *time-reversal* $q(\cdot, \cdot)$ which is a different transition matrix on $V \times V$, which satisfies

$$\pi(y)p(y, z) = \pi(z)q(z, y), \quad y, z \in V.$$

There is a similar formula for the m -step transition probabilities of q , which is given by

$$\pi(y)p^m(y, z) = \pi(z)q^m(z, y), \quad y, z \in V, m \geq 1.$$

by summing up over intermediary states and induction on $m \geq 1$. Now, note that

$$p^{n+m}(x, z) - \pi(z) = \sum_{y \in V} \left(p^n(x, y) - \pi(y) \right) \left(p^m(y, z) - \pi(z) \right)$$

and therefore:

$$\begin{aligned}
\left| \frac{p^{n+m}(x, z) - \pi(z)}{\pi(z)} \right| &= \left| \sum_{y \in V} (p^n(x, y) - \pi(y)) \left(\frac{p^m(y, z) - \pi(z)}{\pi(z)} \right) \right| \\
&= \left| \sum_{y \in V} \pi(y) \left(\frac{p^n(x, y)}{\pi(y)} - 1 \right) \left(\frac{q^m(z, y)}{\pi(y)} - 1 \right) \right| \\
&\leq \chi(p^n(x, \cdot), \pi) \chi(q^m(z, \cdot), \pi)
\end{aligned}$$

by Cauchy-Schwarz's inequality. But now, observe that $Q(S, S^c)$ is the asymptotic fraction of transitions of the chain from a state in S to a state in S^c at equilibrium. However, every such transition must be followed by a transition from a state in S^c to a state in S , and therefore, the asymptotic frequency of these transitions must be equal. It follows that $Q(S, S^c) = Q(S^c, S)$, and as a consequence the conductance profile of the chain q is identical to the conductance profile of the chain p . It follows that if

$$m, \ell \geq \int_{4\pi^*}^{4/\varepsilon} \frac{2du}{u\Phi(u)^2},$$

then

$$\chi(p^m(x, \cdot), \pi) \leq \sqrt{\varepsilon}, \chi(q^\ell(x, \cdot), \pi) \leq \sqrt{\varepsilon}$$

and therefore,

$$\left| \frac{p^{m+\ell}(x, z) - \pi(z)}{\pi(z)} \right| \leq \varepsilon.$$

This finishes the proof of Theorem 7.2. □

8 Coupling from the past: exact sampling.

Around 1996, Propp and Wilson came up with a brilliant algorithm to generate an *exact* sample from the equilibrium distribution of a wide class of Markov chains – and this algorithm also decides how long to run the chain for. This algorithm is known as *coupling from the past*, for reasons which will become clear. The setup in which this algorithm is simplest is when the Markov chain’s state space has a natural notion of partial order. To keep things simple, we first introduce a prototypical example of the class of Markov chain to which coupling from the past applies, and then describe how this works.

8.1 The Ising model and the Glauber dynamics

The Ising model is one of the most basic models of statistical physics, which is a probability measure on spin configurations over a given graph G . Suppose $G = (V, E)$ is a finite graph, such as the $n \times n$ torus, and let $\sigma \in S := \{-1, 1\}^V$, i.e., σ is a function from the vertices of G into $\{-1, 1\}$ (which is the value of the spin at every vertex). Define the *Hamiltonian* of the system by:

$$H(\sigma) = - \sum_{i \sim j} \sigma_i \sigma_j - \sum_{i \in V} B_i \sigma_i$$

where $(B_i, i \in V)$ are given numbers called the external field of the system. We define a probability measure on spin configurations $\sigma \in S$ by:

$$\mu_\beta(\sigma) := Z^{-1} \exp(-\beta H(\sigma))$$

where $\beta > 0$ and Z^{-1} is a normalizing constant which makes the probabilities add up to 1. μ_β is called the Gibbs distribution of the Ising (ferromagnetic) model with inverse temperature β . Thus μ_β favors configurations on which neighbouring spins agree, and the greater β , the greater this tendency.

To digress a little bit from the main topic, looking at simulations of this model, one guesses the following phenomenon: there is a phase transition as β increases from 0 to ∞ during which the following occurs: for small $\beta > 0$, the connected clusters of identical spins are small and widespread “random” (whatever this means), while for large β they are organized: for instance if $B_i > 0$ then spins are overwhelmingly negative.

To make simulations, one needs an efficient algorithm for sampling and this is usually done with the help of the following Markov chain called the Glauber dynamics: this is a Markov chain which updates the spin values of one site at a time, and does so as follows. We select a site random, $i \in V$ say, and let σ_i be the value of the spin at i . The update of the site is essentially the following: we pretend the neighbours $j \sim i$ are already at equilibrium, and choose the new value of σ_i according to the conditional equilibrium distribution of σ_i given the values of σ_j . In practice, this means the following: let U be a uniform random variable, let $p = \mu_\beta(\sigma_i = +1 | \sigma_j, j \neq i)$ and let $q = 1 - p$. Then update $\sigma_i = 1$ if and only if $U < p$, or in other words,

$$U < \frac{1}{1 + q/p},$$

and put $\sigma_i = -1$ otherwise. Now, observe that q/p may be written as

$$\frac{q}{p} = \exp(-\beta \Delta H), \text{ where } \Delta H = 2 \sum_{j \sim i} \sigma_j + 2B_i.$$

Thus note that we don't even need to estimate the normalizing constant Z to do this!

Theorem 8.1. *The Gibbs distribution μ_β is the unique invariant distribution for the Glauber dynamics.*

Proof. It suffices to prove that the detailed balance condition

$$\mu_\beta(\sigma)P(\sigma, \sigma') = \mu_\beta(\sigma')P(\sigma', \sigma)$$

where P denotes the transition kernel for the Glauber dynamics.

To check it, it suffices to consider σ, σ' which differ at exactly one vertex $i \in V$. Assume for instance $\sigma_i = -1$ while $\sigma'_i = +1$. In $\mu_\beta(\sigma)$, we eliminate the dependence on things other than σ_i by writing

$$\mu_\beta(\sigma) = C \exp \left(-\beta \sum_{j \sim i} \sigma_j - \beta B_i \right)$$

and

$$\mu_\beta(\sigma') = C \exp \left(\beta \sum_{j \sim i} \sigma_j + \beta B_i \right).$$

Thus it suffices to check

$$C \exp \left(-\beta \frac{\Delta H}{2} \right) \frac{1}{1 + \exp(-\beta \Delta H)} = C \exp \left(\beta \frac{\Delta H}{2} \right) \left(1 - \frac{1}{1 + \exp(-\beta \Delta H)} \right)$$

or equivalently, after cancellation of $C/(1 + \exp(\beta \Delta H))$:

$$\exp \left(-\beta \frac{\Delta H}{2} \right) = \exp \left(\beta \frac{\Delta H}{2} \right) \exp(-\beta \Delta H)$$

which is obvious. □

Monotonicity. There is a natural order relation on spin configurations σ , which is to say $\sigma \preceq \sigma'$ if $\sigma_i \leq \sigma'_i$ for all $i \in V$. Note that the Glauber dynamics *respects* this order relation: that is, if $\sigma_1 \preceq \sigma_2$, then their respective updates σ'_1 and σ'_2 will also satisfy the same relations. This is an immediate consequence of the fact that

$$\Delta H = 2 \sum_i B_i + 2 \sum_i \sigma_i \quad \text{monotone increasing in every } \sigma_i$$

There is one maximal state $\widehat{1}$ which is the spin configuration where all spins are pointing up, while there is a minimal configuration $\widehat{-1}$ such that all spins are pointing down.

This monotonicity (and the existence of a minimal and maximal states) are the properties we are looking for. Rather than state precise conditions, we now describe the method of coupling from the past for the Glauber dynamics. It will be clear from the example how this works in general.

8.2 Coupling from the past.

The algorithm, and the proof that it works, are both deceptively simple: but change one ingredient and the whole thing collapses. The initial idea is the following. Instead of running the Markov chain starts at time 0 and we need to run it for a long time, we imagine instead it has run forever, and we need to choose the starting point far enough into the past (and the starting states suitably) so that the sample from the Markov chain at time 0 can be guaranteed to be exactly in equilibrium. To do that, we use the monotonicity of the Glauber dynamics as follows. Assume that some independent uniform random variables U_{-1}, U_{-2}, \dots have been fixed once and for all. Let $T > 0$ and consider the Glauber chain runs between times $-T$ and 0 using these same random variables for the updates, and suppose also that T is large enough that if we had started the chain at time $-T$ from the maximal $\widehat{1}$, then the state at time 0 would be identical to the state which we would have obtained if we had started from the minimal state $\widehat{-1}$. In that case, note that any other starting state is always such that the chain run from that state using the updates U_{-T}, \dots, U_0 is always sandwiched between the chain started from the extremal states. We say that the chain has *coalesced*.

If the chain has not coalesced during $[-T, 0]$, we start again from $-2T$ and keep running the chain using the *same* updates $U_{-2T}, \dots, U_{-T}, \dots, U_{-1}$, and start again checking whether the chain has coalesced during $[-2T, 0]$. So long as the chain hasn't coalesced then we keep multiplying T by 2 and checking if the two extremal states coalesce starting from time $-T$ before time 0. If that is the case, we define our sample X to be the value of the chain at time 0. This is the *coupling from the past*.

Theorem 8.2. *This algorithm terminates almost surely in finite time. Moreover, $X \stackrel{d}{=} \pi$.*

Proof. The update rule of configuration σ given the randomness U may be written as a map $\sigma' = \phi(\sigma, U)$. For $s < t \in \mathbb{Z}$, let

$$f_t : S \rightarrow S \text{ defined by } f_t(\sigma) = \phi(\sigma, U_t)$$

and let $F_s^t = f_{t-1} \circ f_{t-2} \circ \dots \circ f_s$. Note that the maps f_t are i.i.d. Since the chain is ergodic, there is an L such that $P_{\widehat{-1}}^{\widehat{1}}(X_t = \widehat{1}) = \varepsilon > 0$. By monotonicity, this implies

$$\mathbb{P}(F_{-(i+1)L}^{-iL} \text{ is constant}) \geq \varepsilon > 0, \text{ for all } i \geq 0.$$

Since these events are independent, by the Borel-Cantelli lemma, a.s. there is some $i \geq 0$ such that $F_{-(i+1)L}^{-iL}$ is constant. In this case it follows that $F_{-(i+1)L}^0$ is also constant, and thus F_{-T}^0 is almost surely constant for T sufficiently large. Call $F_{-\infty}^0$ this value, which is the value returned by the algorithm. It remains to check that $F_{-\infty}^0$ is distributed according to π . But note that

$$F_{-\infty}^0 =_d F_{-\infty}^1$$

and on the other hand, $F_{-\infty}^1$ is obtained from $F_{-\infty}^0$ by performing a step of the Markov chain. Thus the distribution of $F_{-\infty}^0$ is invariant, and is hence equal to π . \square

Remark. One could imagine lots of variations to this idea, but it is important to realise that most will fail: for instance, if you try to coalesce in the future and consider the first time after time 0 that the coupled chains started from the top and the bottom agree, the resulting state need not be a sample of the equilibrium distribution. (This is because this time T^* is

random and is not independent from the chain.) Similarly, it is essential to use the same fixed randomness $U_{-1}, \dots, U_{-T}, \dots$ at every step of the algorithm. For instance, if coupling fails and we need to look $2T$ backwards in time, we cannot refresh the variables U_{-1}, \dots to generate the chain again.

Let T_\star be the running time of the algorithm, i.e., the first T such that F_{-T}^0 is constant. Along with a statement that coalescence eventually occurs, Propp and Wilson show that actually the distribution of the coalescence time T_\star is not much greater than the mixing time, in the following sense. Let

$$t_{\text{mix}} = t_{\text{mix}}(1/e),$$

and let H denote the length of the longest totally ordered chain between the minimal and maximal elements $\widehat{-1}$ and $\widehat{1}$.

Theorem 8.3.

$$\mathbb{E}(T_\star) \leq 2 t_{\text{mix}}(1 + \log H).$$

In particular, this means that coupling from the past is very efficient, since of course T_\star cannot be smaller than t_{mix} . For instance, in the case of Glauber dynamics on a subgraph G of \mathbb{Z}^d , H is only of the order of the volume of the subgraph, which means $\log H$ is of order $\log n$ if G has diameter of order n .

Proof. To prove the result, we note that T_\star has the same distribution as T^\star , where T^\star is the time of coalescence forward in time from time 0, i.e., T^\star is the first T such that F_0^T is constant. Thus we only prove the result about T^\star , which is conceptually much simpler than T_\star .

Lemma 8.1. *For all $k \geq 1$,*

$$\frac{\mathbb{P}(T^\star > k)}{H} \leq d(k) \leq \mathbb{P}(T^\star > k).$$

The second inequality is trivial, since we have a coupling between X_k and π which works with probability at least $\mathbb{P}(T^\star \leq k)$. The first inequality goes as follows. Let $h(x)$ denotes the length of the longest totally ordered chain whose top element is x . Then if $X_-^k = F_0^k(\widehat{-1})$ is different from $X_+^k = F_0^k(\widehat{1})$, it must be the case that

$$h(X_-^k) \leq h(X_+^k) - 1$$

since we know that $X_-^k \preceq X_+^k$. Therefore,

$$\begin{aligned} \mathbb{P}(T^\star > k) &= \mathbb{P}(X_-^k \neq X_+^k) \\ &\leq \mathbb{E}[h(X_+^k) - h(X_-^k)] \\ &\leq \mathbb{E}_{\widehat{1}}[h(X_k)] - \mathbb{E}_{\widehat{-1}}[h(X_k)] \\ &\leq \|p_k(\widehat{1}, \cdot) - p_k(\widehat{-1}, \cdot)\| H \\ &\leq Hd(k) \end{aligned}$$

from which the inequality follows. □

Lemma 8.2. *The quantity $\mathbb{P}(T^* > k)$ is submultiplicative: for $k_1, k_2 \geq 0$:*

$$\mathbb{P}(T^* > k_1 + k_2) \leq \mathbb{P}(T^* > k_1)\mathbb{P}(T^* > k_2).$$

Proof. The event that $F_0^{k_1}$ is a constant map and the event that $F_{k_1}^{k_1+k_2}$ is a constant map are independent, and if either of those occurs then $F_0^{k_1+k_2}$ is also constant. \square

Lemma 8.3. *For all $k \geq 1$,*

$$k\mathbb{P}(T^* > k) \leq \mathbb{E}(T^*) \leq \frac{k}{\mathbb{P}(T^* \leq k)}.$$

Proof. The first inequality is a trivial consequence of Markov's inequality. For the second, let $\varepsilon = \mathbb{P}(T^* > k)$. By submultiplicativity,

$$\mathbb{P}(T^* > ik) \leq \varepsilon^i$$

and thus

$$\begin{aligned} \mathbb{E}(T^*) &= \sum_{j=0}^{\infty} \mathbb{P}(T^* > j) \leq \sum_{i=0}^{\infty} k\mathbb{P}(T^* > ki) \\ &\leq \sum_{i=0}^{\infty} k\varepsilon^i = \frac{k}{1-\varepsilon} = \frac{k}{\mathbb{P}(T^* \leq k)}. \end{aligned}$$

This proves the lemma. \square

Proof. (of Theorem 8.3). By definition of t_{mix} , $d(t_{\text{mix}}) \leq 1/e$. Since d is also submultiplicative, it follows that for $k = t_{\text{mix}}(1 + \log H)$, $d(k) \leq 1/(eH)$. Therefore, by Lemma 8.1,

$$\mathbb{P}(T^* > k) \leq Hd(k) \leq \frac{1}{e}$$

i.e., $\mathbb{P}(T^* \leq k) \geq 1 - 1/e$. Thus by Lemma 8.3

$$\mathbb{E}(T^*) \leq \frac{k}{1-1/e} \leq 2k = 2t_{\text{mix}}(1 + \log H),$$

as claimed. \square

9 Representation Theory

9.1 Basic definitions and results

Given a (complicated) set S , one powerful way to understand it is to find a group G that *acts* on it: i.e., find a homeomorphism $\rho : G \rightarrow S$ (we usually simply note $\rho(g)(x) = g \cdot x$) such that $g \cdot (h \cdot x) = (gh) \cdot x$ and $e \cdot x = x$.

Representation theory asks the reverse question to describe a group G . That is, given a group G , on what structures does it act? This is far too complicated a question, so we restrict ourselves by looking at (finite-dimensional, complex) *linear structures*, and ask moreover that the action respects the inverse.

Definition 9.1. *A group representation of G is a map $\rho : G \rightarrow GL(V)$, where V is a finite-dimensional vector space on \mathbb{C} , which respects the group structure of G . That is, for all $s, t \in G$:*

$$\rho(st) = \rho(s)\rho(t),$$

and $\rho(s^{-1}) = \rho(s)^{-1}$. In particular, $\rho(e) = Id$.

The dimension of V is called d_ρ , the dimension of ρ . If $W \subset V$ is a subspace of V which is stable under G (i.e., $\rho(s)W \subset W$ for all $s \in G$) then the restriction of ρ to W gives us a subrepresentation. If no such space exists, the representation is called *irreducible*.

Our first task is to show that every representation is the finite sum of irreducible representations, where the sum $\sigma = \rho \oplus \rho'$ between two representations ρ and ρ' is defined as one would expect: $\sigma(s)(v + w) = \rho(s)(v) + \rho'(s)(w)$ for $v \in V, w \in W$. This is a representation into $V \oplus W$.

The basic tool for proving this result is the following:

Proposition 9.1. *Let $\rho : G \rightarrow GL(V)$ be a representation of G . Suppose $W \subset V$ is stable. Then there is a complement W' (i.e., a subspace such that $W \cap W' = \{0\}$ and $W + W' = V$) such that W' is also stable.*

Proof. Fix $\langle \cdot, \cdot \rangle$ any scalar product on V . Then we can define a new scalar product on V as follows: $\langle v, w \rangle = \sum_s (\rho(s)v, \rho(s)w)$. Then $\langle \cdot, \cdot \rangle$ is invariant in the sense that $\langle \rho(s)v, \rho(s)w \rangle = \langle v, w \rangle$. Let W' be an orthogonal complement of W . Then W' is a complement of W and moreover W' is stable under ρ : indeed it suffices to check that for all $s \in G$, and all $w' \in W'$, $\rho(s)(w') \in W'$. In other words we need to check that $\langle \rho(s)w', w \rangle = 0$ for all $w \in W$. But by invariance, $\langle \rho(s)w', w \rangle = \langle w', \rho(s^{-1})w \rangle = 0$ since W is stable. \square

By induction on the dimension, we obtain the desired result:

Theorem 9.1. *Every representation ρ is the finite sum of irreducible representations.*

Remark 9.1. *Since the scalar product $\langle \cdot, \cdot \rangle$ defined above is invariant under the action of $\rho(s)$ for any $s \in G$, we deduce that we can choose basis of V such that the matrix representation of $\rho(s)$ in this basis is unitary. In the following we will always make such a choice without saying it.*

9.2 Characters

Ultimately, our main goal will be to use representations of G to do some Fourier analysis. In order to do this we will need basis functions. These are provided by the characters:

Definition 9.2. Let ρ be a representation of G . The character associated to ρ is the function $\chi_\rho : G \rightarrow \mathbb{R}$ defined by $\chi_\rho(s) = \text{Tr}(\rho(s))$.

The following properties are trivial but worth keeping in mind:

Proposition 9.2. (i) $\chi_\rho(e) = d_\rho$, (ii) $\chi_\rho(s^{-1}) = \overline{\chi_\rho(s)}$, and (iii) the characters are invariant by conjugacy: $\chi_\rho(t^{-1}st) = \chi_\rho(s)$.

(i) just follows from the fact $\rho(e) = Id$ always, (ii) from the fact that $\rho(s)^n = \rho(s^n) = Id$ for $n = |G|$, hence all the eigenvalues of $\rho(s)$ are complex roots of the unity. Thus

$$\overline{\chi_\rho(s)} = \overline{\text{Tr}(\rho(s))} = \sum_i \bar{\lambda}_i = \sum_i \frac{1}{\lambda_i} = \text{Tr}(\rho(s^{-1})) = \chi_\rho(s^{-1}).$$

(iii) is the most useful property and just comes from the fact that $\text{Tr}(AB) = \text{Tr}(BA)$.

To check that the characters are the right basis elements we need to check that they are the orthonormal basis elements for some scalar product. We define the usual scalar product on functions from $G \rightarrow \mathbb{C}$:

$$(f|g) = \frac{1}{|G|} \sum_{s \in G} f(s)\bar{g}(s).$$

Theorem 9.2. The characters are orthonormal functions with respect to $(|)$.

Proof. Let χ, χ' be two characters associated with the representations ρ, ρ' . Then $(\chi|\chi') = \frac{1}{|G|} \sum_s \chi(s)\bar{\chi}'(s)$.

The proof relies on a very useful albeit elementary result: Schur's lemma. Two representations ρ, ρ' are called equivalent if there exists an isomorphism of vector spaces $f : V \rightarrow V'$ (linear and one-to-one) such that $f \circ \rho(s) = \rho'(s) \circ f$ for all $s \in G$. Such an f is called a morphism of representations.

Lemma 9.1. Let ρ, ρ' be two irreducible representations into V, V' respectively. Let $f : V \rightarrow V'$ be linear such that

$$f \circ \rho(s) = \rho'(s) \circ f$$

for all $s \in G$. Then

- (a) If ρ, ρ' are not equivalent then $f = 0$.
- (b) If $V = V'$ and $\rho = \rho'$ then f must be the identity.

Proof. Observe that the kernel of f is invariant under ρ . Indeed, if $s \in G$ and $v \in \ker f$, then $f(\rho(s)(v)) = \rho'(s)(f(v)) = \rho'(s)(0) = 0$ so $\rho(s)v \in \ker f$ as well. Likewise, the image of f , $\text{Im} f$ is also invariant for ρ' . Thus (by irreducibility) both kernels and images are either the whole spaces or trivial. Thus for (a), if $f \neq 0$ then $\ker f = \{0\}$ and $\text{Im} f = V'$, so f is an isomorphism. For (b), let λ be an eigenvalue of f . Then the map $\tilde{f} = f - \lambda Id$ has a nontrivial kernel and satisfies $\tilde{f} \circ \rho = \rho' \tilde{f}$. Thus by the above $\tilde{f} = 0$ i.e. $f = \lambda Id$. \square

It is the following corollary which we use here:

Corollary 9.1. *Let $h : V \rightarrow V'$ be linear. Define*

$$\tilde{h} = \frac{1}{|G|} \sum_s \rho'(s^{-1})h\rho(s) : V \rightarrow V'.$$

Then

- (a) *if ρ, ρ' are not equivalent then $h = 0$.*
- (b) *If $V = V'$ and $\rho = \rho'$ then we have $h = \lambda Id$ with $\lambda = \text{Tr}(h)/d_\rho$.*

This follows simply from the observation that for all $t \in G$, $\rho'_{t^{-1}}\tilde{h}\rho_t = \sum_s \rho'_{(st)^{-1}}h\rho_{st} = \tilde{h}$, so Schur's lemma applies.

Returning to the proof of the theorem, fix a basis of V and a basis for V' , and let $r_{ij}(t)$ (resp. $r'_{ij}(t)$) denote the coordinates of the matrix $\rho(t)$ (resp. $\rho'(t)$). Then to show that $(\chi|\chi') = 0$ we must show that $\sum_{i,j} \sum_t \bar{r}_{ii}(t)r'_{jj}(t) = 0$. Fix i, j , and let \tilde{x}_{ij} denote the coordinates of the linear map \tilde{h} for some choice of h . Then

$$\tilde{x}_{ij} = \frac{1}{|G|} \sum_{t \in G} \sum_{k,l} r'_{ik}(t^{-1})x_{kl}r_{lj}(t) = 0$$

by the corollary. Taking $x_{kl} = 0$ unless $k = i, l = j$ where we choose $x_{ij} = 1$ yields $\tilde{x}_{ij} = 0 = \sum_t r'_{ii}(t^{-1})r_{jj}(t)$. Since $\chi'(t^{-1}) = \bar{\chi}'(t)$, the result follows. The calculations for the case $\chi = \chi'$ are identical. \square

The above theorem is very powerful. Here are some illustrations. We start with the following question: let (ρ, V) be a (non-necessarily irreducible) representation. Let (ψ, W) be an irreducible one. Does W appear in the decomposition of V ? If so, how many times?

Theorem 9.3. *The number of times W arises in the decomposition of V is equal to $(\chi_\rho|\chi_\psi)$.*

Proof. Note that it is not even obvious that the right-hand side is an integer! However, write $V = W_1 \oplus \dots \oplus W_m$. Write χ_i for the character of W_i . Then we have $\chi_\rho = \sum_i \chi_i$ where χ_i is the character of W_i . Then $(\chi|\chi_\psi) = \sum_i (\chi_i|\chi_\psi)$. But note that $(\chi_i|\chi_\psi)$ is, by orthonormality, equal to 1 or 0 according to whether $W = W_i$ or not. The result follows. \square

Corollary 9.2. *Two representations are equivalent if and only they have the same character.*

Corollary 9.3. *Let ρ be a representation. Then $(\chi_\rho|\chi_\rho)$ is a positive integer, equal to 1 if and only if ρ is irreducible.*

Consider now the *regular* representation of G : let V be the vector space of functions on G , of dimension $|G|$, and let e_s be the basis element of that space which is the function equal to 1 at s , and 0 elsewhere. Then define

$$\rho(s)(e_t) = e_{st}.$$

Observe that $\chi(e) = |G|$ since $\chi(e)$ is the identity, and if $s \neq e$ then $\rho(s)(e_t) = e_{st} \neq e_t$ so each diagonal coefficient of ρ is zero in this basis. Thus $\chi(s) = 0$.

Theorem 9.4. *Every irreducible representation is contained in the regular representation, with multiplicity equal to its dimension. In particular, there are only a finite number of irreducible representations.*

Proof. Indeed, if ψ is an irreducible representation, its multiplicity in ρ is equal to $(\rho|\psi) = \frac{1}{|G|} \sum_s \chi(s)\chi_\rho(s) = \chi_\rho(e) = d_\rho$. \square

Corollary 9.4. *We have $\sum_\rho d_\rho^2 = |G|$, where \sum_ρ is the sum over all irreducible representations. If $s \neq e$, $\sum_\rho d_\rho \chi_\rho(s) = 0$.*

Indeed, note that, keeping χ for the character of the regular representation, $\chi(s) = \sum_\rho d_\rho \chi_\rho(s)$, by the above. Taking $s = e$ gives the first formula, and the second follows equally since then we know $\chi(s) = 0$.

9.3 Fourier inversion

Let $f : G \rightarrow \mathbb{C}$ be a function. We define its Fourier transform, evaluated at a representation ρ , to be

$$\hat{f}(\rho) = \sum_{s \in G} f(s)\rho(s).$$

Thus $\hat{f}(\rho)$ is a matrix (or a linear map from V to itself).

The *convolution* between two functions f, g is defined as

$$f \star g(s) = \sum_t f(st^{-1})g(t).$$

Then it is straightforward that

$$\widehat{f \star g} = \hat{f}\hat{g}.$$

The the Fourier transform changes a convolution into a product - this will be at the basis of our analysis of a random walk, whose n -step transition probability is precisely an n -fold convolution of the kernel.

Theorem 9.5. *We have the following identity: for all $s \in G$,*

$$f(s) = \frac{1}{|G|} \sum_* d_\rho \operatorname{Tr}(\rho(s^{-1})\hat{f}(\rho)).$$

This is the analogue of the classical Fourier inversion theorem - which, in its discrete form, is just a particular case of this result with $G = \mathbb{Z}/n\mathbb{Z}$.

Proof. Since both sides are linear it suffices to prove the result for $f = e_t$. The $\hat{f}(\rho) = \sum_{z \in G} f(z)\rho(z) = \rho(t)$, so the right-hand side equals $(1/|G|) \sum_\rho d_\rho \chi(s^{-1}t)$, which is nonzero only if $s = t$, in which case it is equal to 1 by Corollary 9.4. \square

Theorem 9.6. *Let $f, g : G \rightarrow \mathbb{C}$ be two functions. Then*

$$\sum_s f(s)g(s^{-1}) = \frac{1}{|G|} \sum_\rho d_\rho \operatorname{Tr}(\hat{f}(\rho)\hat{g}(\rho)).$$

Proof. Taking $f = e_t$ amounts to showing $g(t^{-1}) = \frac{1}{|G|} \sum_\rho d_\rho \operatorname{Tr}(\rho(t)\hat{g}(\rho))$, which is precisely the Fourier inversion theorem. \square

In particular, a way to rephrase this is to say that

$$\sum_s f(s)h(s) = \frac{1}{|G|} \sum_\rho \text{Tr}(\hat{f}(\rho)\hat{g}(\rho)^*). \quad (26)$$

where M^* is the conjugate transpose of a matrix M . This follows from the fact that $\rho(s)$ is unitary and hence $\rho(s^{-1}) = \rho(s)^{-1} = \rho(s)^*$ for any $s \in G$.

We immediately use this result to show a few applications. Let $s, t \in G$. We say s and t are conjugate if there exists $g \in G$ such that $gsg^{-1} = t$. This defines an equivalence relation on G , its equivalence classes are simply called *conjugacy classes*, a notion that is quite important in group theory. A function that is constant on conjugacy classes is called a *class function*.

When $G = S_n$, there is an easy way to find out whether two permutations are conjugate: if π is a permutation having cycle decomposition $c_1 \dots c_m$, and σ is a permutation, then $\sigma\pi\sigma^{-1}$ is the permutation having cycle distribution equal to $\sigma(c_1) \dots \sigma(c_m)$, where if $c = (x_1, \dots, x_k)$ we denote by $\sigma(c)$ the cycle $(\sigma(x_1), \dots, \sigma(x_k))$. It follows that two permutations are conjugate if and only if they have the same *cycle structure*: the same number of cycles of size 1, of size 2, etc. Thus a typical class function would be $f(\sigma) =$ the number of cycles of σ . However, an even more interesting one is $p_n(\sigma) = \mathbb{P}(X_n = \sigma)$, the n -step transition probabilities for the random transpositions process on S_n .

Lemma 9.2. *Let f be a class function on G , and let ρ be an irreducible representation. Then there exists $\lambda \in \mathbb{C}$ such that $\hat{f}(\rho) = \lambda Id$. Moreover,*

$$\lambda = \frac{|G|}{d_\rho} (f|\bar{\chi}_\rho).$$

Proof. Consider the linear application $\rho(s)\hat{f}(\rho)\rho(s)^{-1}$, for any $s \in G$. Then an expression for it is

$$\begin{aligned} \rho(s)\hat{f}(\rho)\rho(s)^{-1} &= \sum_{t \in G} f(t)\rho(s)\rho(t)\rho(s^{-1}) \\ &= \sum_{t \in G} f(t)\rho(sts^{-1}) \\ &= \sum_{t \in G} f(sts^{-1})\rho(sts^{-1}) \\ &= \hat{f}(\rho) \end{aligned}$$

since f is a class function. So by Schur's lemma, $\hat{f}(\rho) = \lambda I$ for some $\lambda \in \mathbb{C}$. Taking the traces, we find,

$$\lambda = \frac{1}{d_\rho} \text{Tr}(\hat{f}(\rho)).$$

By linearity of the trace, $\text{Tr}(\hat{f}(\rho)) = \sum_s f(s) \text{Tr}(\rho(s)) = \sum_s f(s)\chi_\rho(s) = |G|(f|\bar{\chi}_\rho)$. \square

With this theorem, we immediately deduce the following result, of fundamental importance in many studies:

Theorem 9.7. *The characters form an orthonormal basis of the space of class functions.*

Proof. Note that the characters are themselves class functions since $\text{Tr}(AB) = \text{Tr}(BA)$. We already know that they are orthonormal, so it remains to prove that they generate all class functions. To see this it suffices to check that if f is a class function such that $(f|\chi_\rho) = 0$ for all irreducible representations ρ , then f is zero. However, by the above lemma, in this case $\hat{f}(\rho) = 0$ for all irreducible representation ρ and thus by Fourier inversion $f = 0$. \square

9.4 Applications to card shuffling: the Diaconis-Shahshahani theorem

We come to one of the important results in the section, which shows the relationship between mixing times and representation theory. Recall that the trivial representation is the one-dimensional representation such that $\rho(s)x = x$ for all $x \in \mathbb{C}$.

Theorem 9.8. *Let P be a probability distribution on G and let π be the uniform distribution. Then*

$$d_2(P, \pi)^2 := |G| \sum_{s \in G} (P(s) - \pi(s))^2 = \sum_{*} d_\rho \text{Tr}(\hat{P}(\rho) \overline{\hat{P}(\rho)}).$$

where the sum \sum_{*} is over all nontrivial irreducible representations ρ .

Proof. Let $f(s) = P(s) - \pi(s)$ and $g(s) = f(s)$. Applying the Plancherel formula to this we get

$$\begin{aligned} d_2(P, \pi) &= \sum_{s \in G} f(s)^2 \\ &= \sum_{\rho} d_\rho \text{Tr}(\hat{f}(\rho) \hat{f}(\rho)^*). \end{aligned}$$

Note that when ρ is the trivial representation, $\hat{P}(\rho) = 1 = \hat{\pi}(\rho)$ so $\hat{f}(\rho) = 0$. When ρ is nontrivial, we have that $\hat{\pi}(\rho) = 0$, e.g. as a consequence of the orthogonal relations between the characters since the function 1 is the character of the trivial representation. \square

The following corollary makes explicit what we learn when $P = P^*k$ is the k -step transition probability of a random walk on G whose kernel P is a class function. Then:

Corollary 9.5. *For all $k \geq 1$, we have*

$$4d(k)^2 \leq d_2(k)^2 = \sum_{*} d_\rho^2 \lambda_\rho^{2k}$$

where $\lambda_\rho = \frac{1}{d_\rho} \sum_{s \in G} P(s) \chi_\rho(s)$.

Proof. We use the fact that $\widehat{P^{*k}} = \hat{P}^k$ and that since P is a class function, $\hat{P} = \lambda I$ with an explicit λ as in the previous lemma. So $\text{Tr}(\hat{P}^k (\hat{P}^k)^*) = \lambda^{2k} d_\rho$. \square

Thus if $G = S_n$ and P is the kernel of random transpositions, $\lambda_\rho = \frac{1}{n} + \frac{n-1}{n} r(\rho)$, where

$$r(\rho) = \frac{\chi_\rho(\tau)}{d_\rho} = \frac{\chi_\rho(\tau)}{\chi_\rho(1)}$$

is the co-called character ratio. Here $\chi_\rho(\tau)$ denotes the character of ρ evaluated at any transposition (since characters are class functions).

10 Riffle shuffle

What follows is a set of (informal) notes designed to walk you through the mathematics of the riffle shuffle, which is a model for the card shuffling method used by casinos and professional dealers. This was analysed in remarkable detail first by Aldous who found the asymptotic mixing time in [1] and by Bayer and Diaconis who found an exact formula which considerably sharpened Aldous' result.

The basic framework is the Gilbert-Shannon-Reeds model for card shuffling, which is defined as follows. We first cut the deck in two piles of size k and $n - k$, where the position k of the cut follows a Binomial $(n, 1/2)$ distribution. Then, if we imagine that we hold the two piles in our left and right hand, drop the next card from the left or right pile with probability proportional to the size of the pile. That is, if there are a cards in the left hand and b cards in the right hand, drop from the left with probability $a/(a + b)$ and from the right with probability $b/(a + b)$. This gives you a new deck which is the result of one shuffle. This shuffle is then repeated many times. We are going to show the proof of the two following results.

Theorem 10.1. (Aldous 1983 [1]) *There is a cutoff phenomenon at time*

$$t_{\text{mix}} = \frac{3}{2} \log_2 n.$$

The following results of Bayer and Diaconis analyze this in an exact and much sharper way. The first amazing result is an exact formula for the probability distribution of the walk after m steps.

Theorem 10.2. (Bayer-Diaconis 1992 [3]) *After m shuffles,*

$$P(X_m = \sigma) = \frac{1}{2^{mn}} \binom{2^m + n - R(\sigma)}{n}$$

where $R(\sigma)$ is the number of rising sequences of σ , defined below.

Using this exact formula, Bayer and Diaconis were able to study in great detail what happens near the cutoff point, after of order $(3/2) \log_2 n$ shuffles have been performed.

Theorem 10.3. (Bayer-Diaconis 1992 [3]) *Let $m = \log_2(n^{3/2}c)$. Then*

$$d(m) = 1 - 2\Phi\left(-\frac{1}{4\sqrt{3}c}\right) + O(n^{-1/4})$$

where $\Phi(x)$ is the cumulative distribution function of a standard normal random variable:

$$\Phi(x) = \int_{-\infty}^x e^{-u^2/2} \frac{du}{\sqrt{2\pi}}$$

We now comment on the numerical values of those constants for $n = 52$. First, note that in this case,

$$(3/2) \log_2 n = 8.55\dots$$

which indicates that of order 8 or 9 shuffles are necessary and sufficient.

However, based on the Bayer-Diaconis formula and an exact expression for the number of permutation with a given number of rising sequences (an *Eulerian number*, discussed later), we obtain the exact value for $d(m)$

| | | | | | |
|--------|------|-------|------|-------|-------|
| m | 5 | 6 | 7 | 8 | 9 |
| $d(m)$ | 0.92 | 0.614 | 0.33 | 0.167 | 0.085 |

As we see from this table, it is clear that convergence to equilibrium occurs after no less than 7 shuffles. The total variation distance decreases by 2 after each successive shuffle following the transition point.

Remark. It is interesting to note that while 7 is a very small number compared to the size of the state-space ($52!$ which has about 60 digits), this is a rather large number in practice. Nobody ever shuffles a deck of card more than 3 or 4 times. It is easy to take advantage of this in magic tricks (and in casinos, apparently). Bayer and Diaconis describe some very pleasant tricks which exploit the non-randomness of the deck at this stage, which are based on the analysis of the riffle shuffle and in particular of the rising sequences. The reading of the original paper [3] is wholeheartedly recommended !

We will present the key ideas that lead to the proof of Aldous' results (Theorem 10.1.) As we will see, many of the ideas that were used by Bayer and Diaconis were already present in that paper, which appeared about 10 years before.

Before we do anything, we need to define what are the rising sequences of a permutation σ , as the analysis essentially concentrates on the description of their evolution under the shuffle.

Definition 10.1. *Let $\sigma \in \mathcal{S}_n$. The rising sequences of the arrangement of cards σ are the maximal subsets of successive card labels such that these cards are in increasing order.*

This definition is a little hard to digest at first but a picture illustrates the idea, which is very simple. For instance, if $n = 13$ and the deck consists of the following arrangement:

1 7 2 8 9 3 10 4 5 11 6 12 13

then there are two rising sequences:

| | | | | | |
|---|---|---|----|----|-------|
| 1 | 2 | 3 | 4 | 5 | 6 |
| 7 | 8 | 9 | 10 | 11 | 12 13 |

The number of rising sequences of σ is denoted by $R(\sigma)$. Note that rising sequences form a partition of the card labels $1, \dots, n$.

The reason why rising sequences are so essential to the analysis is because when we perform a shuffle, we can only double $R(\sigma)$. The above example illustrates well this idea. The two rising sequences identify the two piles that have resulted from cutting the deck and that have been used to generate the permutation σ in one shuffle. This leads to the following equivalent description of the Gilbert-Shannon-Reeds riffle shuffle measure μ .

Description 2. μ is uniform on the set R of permutation with exactly two rising sequences, and gives mass $(n + 1)2^{-n}$ to the identity.

To see this, fix a permutation $\sigma \in R$. The two rising sequences of σ have length L and $n - L$, say. Then as explained above, they identify the cut of the two piles that have resulted

from cutting the deck. The probability of having made exactly this cut is $\binom{n}{L}2^{-n}$. We then need to drop the cards from the two piles in the correct order. This corresponds to the product of terms of the form $a/(a+b)$, where a and b are the packet sizes. If we focus on the denominators first, note that this will always be the number of cards remaining in our hands, hence it will be $n, n-1, \dots, 2, 1$. As for the numerators, cards dropping from the left hand will give us the terms $L, L-1, \dots, 2, 1$ and terms from the right hand will give us $n-L, n-L-1, \dots, 2, 1$. It follows that the probability of obtaining σ

$$\mu(\sigma) = \frac{\binom{n}{L}}{2^n} \frac{1}{n!} L!(n-L)! = 2^{-n}$$

Note that a riffle is entirely specified by saying which card comes from the left pile and which from the right pile. Thus, we associate to each card c a binary digit $D(c) = 0$ or 1 , where 0 indicates left and 1 indicates right. By the above description, the resulting deck can be described by sequence of n bits which is uniformly distributed over all possible sequences of n binary digits. (Check that this works with the identity as well). This leads to the following description. Let $\mu'(\sigma) = \mu(\sigma^{-1})$ be the measure associated with the reverse move.

Description 3. The reverse shuffle (i.e., the shuffle associated with the measure μ'), can be described as assigning i.i.d. 0-1 digits to every card c , with $P(D(c) = 1) = 1/2$ and $P(D(c) = 0) = 1/2$. The set of cards c such that $D(c) = 0$ is then put on top of the set of cards with $D(c) = 1$.

The beautiful idea of Aldous is to notice that this reverse description (the backward shuffle) is a lot easier to analyze. Let $(X'_m, m \geq 0)$ be the random walk associated with the shuffling method μ' . Since

$$X'_m = g'_1 \dots g'_m = g_1^{-1} \dots g_m^{-1} = (g_m \dots g_1)^{-1}$$

we see that

$$X'_m \stackrel{d}{=} X_m^{-1}$$

and it follows easily that the mixing time of the forward shuffle X is the same as the mixing time of the backward shuffle X' . In fact if d' is the total variation function for the walk X' we have

$$d(m) = d'(m)$$

We are thus going to analyze X' and show that it takes exactly $3/2 \log_2 n$ steps to reach equilibrium with this walk.

To describe the state of the deck after m backward shuffles, we successively assign i.i.d. binary digits 0 or 1 to indicate (respectively) top or bottom pile. E.g., after 2 shuffles:

| deck | 1st shuffle | 2d shuffle |
|------|-------------|------------|
| — | 1 | 0 |
| — | 0 | 0 |
| — | 0 | 1 |
| — | 1 | 1 |
| — | 1 | 0 |
| — | 0 | 0 |

Reading right to left, it is easy to see that the deck consists of the cards with labels 00, then 01, then 10, then 11. This generalizes as follows. For any card c , we attach m binary digits 0

and 1 which tell us if the card is going to the top or the bottom pile in m successive backward shuffles. We may interpret this sequence by reading from right to left as the binary expansion of a number $D_m(c)$. Then the fundamental properties of the deck are:

- (a) The deck is ordered by increasing values of $D_m(c)$.
- (b) If two cards c and c' have the same value of $D_m(c) = D_m(c')$ then they retain their initial ordering.

Note that the numbers $(D_m(c), 1 \leq c \leq n)$ are i.i.d. for different cards, with a distribution that uniform on $\{0, \dots, 2^m - 1\}$.

10.1 Lower bounds

A first upper-bound

One immediate consequence of properties (a) and (b) is that if $T =$ the first time at which all labels $D_m(c)$ are distinct, then the deck is exactly uniformly distributed. We use this remark to get a first upper-bound on the time it takes to get close to stationarity.

Lemma 10.1. *If $m \gg 2 \log_2 n$ then with high probability all labels $D_m(c)$ are distinct.*

Proof. The proof is elementary, and is a reformulation of the Birthday problem. We view the $M = 2^m$ possible values of $D_m(c)$ as M urns and we are throwing independently n balls into them at random. The probability that they all fall in distinct urns is

$$\begin{aligned} P(\text{all labels distinct}) &= 1 \left(1 - \frac{1}{M}\right) \left(1 - \frac{2}{M}\right) \dots \left(1 - \frac{n-1}{M}\right) \\ &= \exp\left(\sum_{j=0}^{n-1} \ln\left(1 - \frac{j}{M}\right)\right) \\ &\approx \exp\left(-\sum_{j=0}^{n-1} \frac{j}{M}\right) \approx \exp(-n^2/2M) \end{aligned}$$

It follows that if $M \ll n^2$ then some cards will have the same label, but if $M \gg n^2$ then with high probability all cards will have distinct labels. But $M = n^2$ is equivalent to $m = 2 \log_2 n$. \square

To rigorously use Lemma 1 to conclude that the distance function at time $(1 + \varepsilon) \log_2 n$ is small, we recall that the total variation distance is small if there is a successful coupling with high probability. Since $X'_T =_d U$ is uniform, the above lemma tells us that

$$d(m) \leq P(T > m)$$

and $P(T > m) \rightarrow 0$ if $m = (2 + \varepsilon) \log_2 n$. This is not the $(3/2) \log_2 n$ we were hoping for, but building on these ideas we will do better a bit later.

In the forward shuffle, the essential concept is that of rising sequences. In the backward shuffle, the equivalent notion is that of descents of a permutation. We say that σ has a descent at j (where $1 \leq j \leq n - 1$) if $\sigma(j) > \sigma(j + 1)$. Let

$$\text{Des}(\sigma) = \#\text{descents of } \sigma = \sum_j a_j \tag{27}$$

where a_j is the indicator of the event that σ has a descent at j . It is trivial to observe that

$$R(\sigma) = \text{Des}(\sigma^{-1}) - 1$$

In this lower-bound, we will show that for $m < \log_2 n$, the number of descents of X'_m is not close to the number of descents of a uniform permutation. This will show that the distance is approximately 1.

Lemma 10.2. *Let $\sigma =_d U$. Then*

$$E(\text{Des}(\sigma)) = (n - 1)/2 \text{ and } \text{var } \text{Des}(\sigma) \sim n/12 \quad (28)$$

The expectation is very easy to compute. In a random permutation each j has probability $1/2$ of being a descent. Moreover there is a lot of independence between the a_j , so it is not surprising that the variance is of order n . In fact, as we will mention later, $\text{Des}(\sigma)$ is approximately normally distributed with this mean and variance.

Now, consider our urn representation of the deck X'_m . Each of the 2^m urns corresponds to a possible value of $D_m(c)$, and those cards which fall in the same urn retain their initial order. It is *obvious* that each urn can create at most one descent when we put piles on top of each other (wince within each urn, the order is the same as initially). It follows that

$$\text{Des}(X'_m) \leq 2^m - 1$$

If $m = (1 - \varepsilon) \log_2 n$ then $\text{Des}(X'_m) \leq n^{1-\varepsilon}$ and thus this is incompatible with (28). The two distributions (X'_m and U) concentrate on permutations with very different number of descents, hence the total variation is close to 1.

A true lower-bound

Here we we push a bit further the lower-bound of the previous section. We will show that for $m = \alpha \log_2 n$ and $\alpha < 3/2$, then

$$E(\text{Des}(X'_m)) = \frac{n - 1}{2} - n^\beta \quad (29)$$

with $\beta > 1/2$, while the variance of $\text{Des}X'_m$ stays $O(n)$. This will imply again that the total variation distance is approximately 1 in this regime. Indeed, (28) implies that for a uniform permutation, the number of descents is $n/2 \pm \sqrt{n}$. Here, (29) implies that the number of descents is $n/2 - n^\beta \pm \sqrt{n}$. Since $\beta > 1/2$, this implies that the two distributions concentrate on permutations with a distinct number of descents.

We need the following lemma, which is a simple modification of the Birthday problem.

Lemma 10.3. *Throw n balls in M urns, and suppose $M \sim n^\alpha$. Let*

$$U_n = \#\{j \leq n : \text{ball } j \text{ and ball } i \text{ fall in same urn, for some } i < j\}$$

Then

$$E(U_n) \sim \frac{1}{2}n^{2-\alpha} \text{ and } \text{var}(U_n) \sim \frac{1}{2}n^{2-\alpha} \quad (30)$$

There surely is a central limit theorem, too.

To prove (29), consider the set J_m of positions j in the resulting deck such that the card in position j and in position $j+1$ have the same value of D_m . Then note that this j can not be a

descent for X'_m . On the other hand, note that the random variables a_j are almost iid outside of J_m . More precisely, conditionally on J_m , the random variables $(a_j, j \text{ odd and } j \notin J_m)$ are independent, and each has expectation $1/2$ (and similarly with even values of J). From this we deduce:

$$E(\text{Des}(X'_m)|J_m) = \frac{n-1}{2} - \#J_m$$

(each integer gives us probability $1/2$ of being a descent, except those who are in J_m). Also,

$$\text{var Des}(X'_m) = O(n)$$

Now, to conclude, remark that $\#J_m =_d U_n$ in equation (30) and thus

$$E(\#J_m) \sim \frac{1}{2}n^{2-\alpha}.$$

Since $\beta = 2 - \alpha > 1/2$, the lower-bound is proved.

10.2 Guessing the true upper-bound

We now wish to prove that after $m = (3/2 + \varepsilon) \log_2 n$, the deck is well-mixed. Aldous [1] has a calculation that looks pretty simple but that I haven't managed to clarify completely. Instead I propose the following intuitive explanation.

After $\alpha \log_2 n$ shuffles and $\alpha > 3/2$, the number of descents can still be written as

$$\frac{n-1}{2} - n^{2-\alpha} + \text{standard deviation term}$$

What happens is that $n^{2-\alpha}$ becomes $o(n^{1/2})$ and hence the variance term takes over. It is in fact not hard to believe that at this stage, $\text{Des}X'_m$ is in fact approximately normally distributed with mean $n/2 + o(n^{1/2})$ and variance cn for some $c > 0$. This is almost the same thing as for a uniform permutation, except that the constant for the variance may be different.

Lemma 10.4. *Let X and Y have two normal distribution with mean 0 and variance σ_1^2 and σ_2^2 . Then*

$$d_{TV}(X, Y) = f(\sigma_1/\sigma_2).$$

f satisfies $0 < f(x) < 1$ for all $x \neq 1$

Lemma 10.4 and the above comment thus imply that the total variation distance between the law of $\text{Des}X'_m$ and $\text{Des}\sigma$ (where σ is uniform) is at most a constant < 1 .

While that seems pretty far away from our desired conclusion (the total variation distance between X'_m and σ is also < 1), we can in fact get there by using in anticipation the Bayer-Diaconis formula. That formula shows that the number of rising sequences of X_m is a sufficient statistics for X_m . (Here, sufficient statistics refers to the fact that knowing $R(\sigma)$ is enough to know the chance of σ - the meaning may be different in statistics...). Thus, $\text{Des}(X'_m)$ is a sufficient statistics for X'_m , and it is obviously so for a uniform permutation as well. On the other hand,

Lemma 10.5. *The total variation distance between X and Y is equal to the total variation distance between any two sufficient statistics of X and Y .*

This is a pretty intuitive fact, and from there the upper-bound follows easily.

10.3 Seven shuffles are enough: the Bayer-Diaconis result

All the foundations are now laid down, and the Bayer-Diaconis formula will follow instantly from the following description of the forward riffle shuffle. (It is a consequence of the urns and balls description of Aldous, but can be proved by other elementary means).

Description 4. X_m is uniform over all ways of splitting the deck into 2^m piles and then riffing the piles together.

We now prove the Bayer-Diaconis formula:

$$P(X_m = \sigma) = \frac{1}{2^{mn}} \binom{2^m + n - R(\sigma)}{n}$$

Let $a = 2^m$. There are a^n shuffles in total. Hence it suffices to prove that the number of ways to obtain the permutation σ is $\binom{2^m + n - R(\sigma)}{n}$.

Note that after the a piles are riffled together, the relative order of the cards within a pile remains constant. Hence this gives at most a rising sequences. Let $r = R(\sigma)$, and consider the partition of σ induced by the rising sequences. These r blocks must correspond to r cuts of the deck. The remaining $a - r$ cuts may be placed anywhere in the deck. To count how many ways there are of doing this, we use what Bayer and Diaconis call the ‘‘stars and bars’’ argument. Increase the deck size to $n + a - r$. Now, we must choose $a - r$ positions to put our $a - r$ cuts. There are

$$\binom{n + a - r}{a - r} = \binom{n + a - r}{n}$$

of doing so. Hence the result!

Using the above formula we can be very explicit about the total variation distance function. Note that

$$d(m) = \sum_{\pi \in \mathcal{S}_n} \left(P^m(\pi) - \frac{1}{n!} \right)^+ = \sum_{\pi \in \mathcal{S}_n} \frac{1}{n!} (n! P^m(\pi) - 1)^+ \quad (31)$$

Let $m = \log_2(n^{3/2}c)$.

$$\begin{aligned} n! P^m(\pi) &= n! \frac{1}{2^{mn}} \frac{(2^m + n - r) \dots (2^m + -r + 1)}{n!} \\ &= \frac{2^m + n - r}{2^m} \dots \frac{2^m - r}{2^m} \\ &= \exp \left(\sum_{i=0}^{n-1} \ln \left(1 + \frac{n - r - i}{2^m} \right) \right) \end{aligned}$$

After an exciting expansion of the log up to the 4th order, and replacing $2^m = n^{3/2}c$ and writing $r = n/2 + h$ (where h may range from $-n/2 + 1$ to $n/2$, we get

$$n! P^m(\pi) = f_n(h) := \exp \left(\frac{-h}{c\sqrt{n}} - \frac{1}{24c^2} - \frac{1}{2} \left(\frac{h}{cn} \right)^2 + O(1/n) + O(h/n) \right) \quad (32)$$

Let h^* be defined by

$$h \leq h^* \iff P^m(\pi) \geq 1/n!$$

This h^* tells us what are the nonzero terms in (31). Now, by setting the exponent equal to 0 in (32), we obtain

$$h^* = -\frac{\sqrt{n}}{24c} + \frac{1}{12c^3} + B + O(1/\sqrt{n}) \quad (33)$$

It follows that

$$d(m) = \sum_{-n/2 \leq h \leq h^*} \frac{R_{nh}}{n!} (f_n(h) - 1)$$

where R_{nh} is the number of permutations with $n/2 + h$ rising sequences. This number is well-known to combinatorists. The number of permutations with j rising sequences is called the Eulerian number a_{nj} , see, e.g. Stanley [12]. Tanny and Stanley show the remarkable formula that if X_1, \dots, X_n are i.i.d. uniform on $(0, 1)$

$$\frac{a_{nj}}{n!} = P(j \leq X_1 + \dots + X_n \leq j + 1) \quad (34)$$

This implies in particular the normal approximation for the descents (or the rising sequences) of a uniform random permutation, with variance equal to $n/12$ as claimed in (28).

From then on, it is essentially a game of algebraic manipulations to obtain Theorem 10.3. We refer the interested reader to p. 308 of [3] for details.

References

- [1] D. Aldous (1983). Random walks on finite groups and rapidly mixing Markov chains. *Seminaire de Probabilités XVII. Lecture Notes in Math.* 986, 243–297. Springer, New-York.
- [2] Aldous, D. and J. Fill. 1999. Reversible Markov chains and random walks on graphs, in progress. Manuscript available at www.stat.berkeley.edu/~aldous/RWG/book.html.
- [3] D. Bayer and P. Diaconis (1992). Trailing the dovetail shuffle to its lair. *Ann. Probab.*, 2, 294-313.
- [4] Diaconis, P. 1988. *Group Representations in Probability and Statistics*, Lecture Notes - Monograph Series, vol. 11, Inst. Math. Stat., Hayward, CA.
- [5] P. Diaconis and M. Shahshahani, Generating a random permutation with random transpositions. *Z. Wahrsch. Verw. Gebiete* **57:2** (1981), 159–179.
- [6] P. Diaconis and L. Saloff-Coste. Comparison techniques for random walks on finite groups. *Ann. Probab.*, 21, 2131–2156 (1993).
- [7] D. Levin, Y. Peres and E. Wilmer. 2009. *Markov chains and mixing times*. American Mathematical Society.
- [8] B. Morris and Y. Peres. Evolving sets, mixing and heat kernel bounds. *Probab. Theor. Rel. Fields*, 133: 245–266 (2005).
- [9] J. Propp and D. Wilson (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Rand. Struct. Algor.*, 9, pp. 223–252
- [10] Saloff-Coste, L. 1997. *Lectures on finite Markov chains*, Lectures on Probability Theory and Statistics, Ecole d'Été de Probabilités de Saint-Flour XXVI - 1996, pp. 301-413
- [11] L. Saloff-Coste, 2003. *Random Walks on Finite Groups*. In: H. Kesten, ed. *Probability on Discrete Structures*, Encyclopaedia of Mathematical Sciences (110), Springer.
- [12] R. Stanley (1977). Eulerian partitions of a unit hypercube. In: *Higher Combinatorics* (M. Aigner, ed.) Reidel, Dordrecht.