Mixing Times of Markov Chains: Techniques and Examples

A Crossroad between Probability, Analysis and Geometry

Nathanaël Berestycki
University of Cambridge
N.Berestycki@statslab.cam.ac.uk

November 22, 2016

The purpose of these notes is to showcase various methods which have been developed over the last 30 years to study mixing times of Markov chains and in particular the *cutoff phenomenon*. This refers to the surprising behaviour that many natural Markov chains reach their equilibrium distributions in a very concentrated window of time.

The emphasis of these notes is on techniques explained through examples, and I did not try to make the results as general as possible. Instead, the ideas are exposed in the simplest form that I know of, and a couple of examples are selected to illustrate the idea concretely. The book by Levin, Peres and Wilmer [17] is a key reference that contains many more examples and more precise results; we refer to it often and follow its exposition in a number of cases. However we depart from it in the organisation of the material and several proofs; some topics which are not covered in that book but are included here include the connection to representation theory and the Diaconis–Shahshahani upper bound lemma, as well as analytic methods based on functional inequalities such as Nash and Log-Sobolev inequalities. On the other hand, I didn't include things like the path coupling method, electrical network theory, and important examples such as lamplighters and spin systems such as Glauber dynamics for Ising model and q-colourings.

The notes evolved from a graduate course I taught at Cambridge on a number of occasions, first as a graduate reading course in 2008. 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, and further after a mini-course in Edinburgh at the occasion of a workshop on probability and representation theory. I wish to express my thanks to the organisers for the opportunity to present this material. They were revised and considerably extended again (and exercises were added) in 2016, as I was preparing to teach it as a Part III course in Cambridge.

Apart from the first two chapters 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

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 $n \to \infty$ (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 The coupling method

1.1 Prerequisites on Markov chains

Let S be a finite state space. Let P be a **transition matrix** on S, i.e., the matrix $P = (P(x,y))_{x,y\in S}$ satisfies $P(x,y) \geq 0$ and $\sum_y P(x,y) = 1$ for every $x \in S$. The matrix P defines a **Markov chain** on S: given a state $x \in S$, we consider a process $X = (X_t, t = 0, 1, \ldots)$ such that $X_0 = x$ and for every $t \geq 0$, given $(X_0, \ldots, X_t)$,

$$\mathbb{P}(X_{t+1} = y | X_0, \ldots, X_t) = P(X_t, y).$$

More generally, we could consider the process started not from a fixed $x \in S$, but rather from a certain starting distribution μ on S. When needed, we will usually refer to the law of this process by $\mathbb{P}_\mu$.

Note that for every $t \geq 0$,

$$\begin{aligned} \mathbb{P}(X_t = y | X_0 = x) &= \sum_{x_1, \ldots x_{t-1} \in S} \mathbb{P}(X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t = y | X_0 = x) \\ &= \sum_{x_1, \ldots, x_{t-1} \in S} P(x, x_1) P(x_1, x_2) \ldots P(x_{t-1}, y) \\ &= P^t(x, y), \end{aligned}$$

where P^t is the t-th power of the matrix P. In other words, the t-step transition probabilities of the Markov chain are given by the t-th power of the transition matrix.

A transition matrix P(or a Markov chain) is called **irreducible** if all states can be reached from any starting point: for every $x, y \in S$, there exists $x = x_0, x_1, \ldots, x_{n-1}, x_n = y$ such that $P(x_{i-1}, x_i) > 0$ for $1 \leq i \leq n$. P is called **aperiodic** if $\gcd\{n \geq 1 : P^n(x,x) > 0\} = 1$ for all states x, and is called periodic otherwise. An example of a periodic Markov chain is simple random walk on the relative integers $\mathbb{Z}$, defined by $P(i, i \pm 1) = 1/2$ and $P(i,j) = 0$ otherwise.

Let $(\pi(x), x \in S)$ be a collection of real numbers indexed by the states in S. We say that π defines an **invariant measure** if for all $y \in S$,

$$\sum_{x \in S} \pi(x) P(x,y) = \pi(y), \tag{1}$$

or in matrix notations, $\pi P = \pi$. If the chain is irreducible then solutions to (1) are unique up to a multiplicative constant. Furthermore, if the chain is recurrent then (1) has a solution such that $\pi(x) \geq 0$. If the chain is positive recurrent, in particular if the state space is finite, then π can be normalised to be a probability distribution. In that case, we call $\pi(x)$ the (unique) **stationary distribution** of the chain, or equivalently its **equilibrium distribution**. Then(1) says that choosing $X_0 \sim \pi$, the Markov chain with starting state X_0 is stationary: for every fixed $n \geq 0$, $X_n \sim \pi$.

The **ergodic theorem** for Markov chain states that if P is irreducible, aperiodic and positive recurrent (in particular, if P is irreducible and aperiodic, and the state space S is finite), then for all starting distribution μ on S, then the Markov chain X started from μ converges to the unique stationary distribution π in the long run: that is, $X_t \to \pi$ as $t \to \infty$ in distribution, or equivalently, for all $y \in S$,

$$\mathbb{P}_\mu(X_t = y) \to \pi(y), t \to \infty. \tag{2}$$

A measure $(\pi(x), x \in S)$ is called **reversible** if for all $x, y \in S$:

$$\pi(x)P(x,y) = \pi(y)P(y,x). \tag{3}$$

It is straightforward to check that in that case, π satisfies (1) and is thus the unique invariant measure of the chain up to a constant.

Example 1.1. ***Random walk on a graph.***

A crucial example is that of a finite graph $G = (V, E)$. Here the Markov chain is defined by taking the state space to be the set of vertices, and setting $P(x, y) = 1/\deg(x)$ if y is a neighbour of x (where $\deg(x)$ is the degree of the vertex x, or the number of neighbours of x), and $P(x, y) = 0$ otherwise. In other words, the random walk on G hops from a vertex x to a randomly chosen neighbour y of x. Note that if $\mu(x) = \deg(x)$, then

$$\mu(x)P(x,y) = \deg(x) \times \frac{1}{\deg(x)} = 1 = \mu(y)P(y,x).$$

Consequently, $\pi(x) = \deg(x)/\sum_y \deg(y)$ is the reversible stationary distribution of a random walk on G. Note that the normalising constant $C = \sum_y \deg(y)$ can be expressed also as $C = 2|E|$, since every edge is counted twice in the sum (once for each endpoint).

1.2 Total variation distance

To discuss the notion of mixing time, the first thing we need is a way to measure how far away from stationarity we are. Given two probability measures μ, ν on a state space S (where μ might be the distribution of the chain after a given number of steps, and ν might be the stationary distribution of the chain), there are various notions of distance between μ and ν one can use. The most natural and simplest notion is that of total variation distance, which is defined as follows.

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

$$\|\mu - \nu\|_{\text{tv}} = \sup_{A \subset S} |\mu(A) - \nu(A)| \tag{4}$$

The total variation distance thus measures the maximal error made when approximating μ by ν to predict the probability of an event. Thus μ and ν are close if they are statistically indistinguishable.

Note that $\|\mu - \nu\|_{\text{tv}}$ in any case. That is, the maximal value that the total variation distance can take is 1. We record here some basic properties of total variation distance:

Lemma 1.1. *We have the following identities:*

$$\|\mu - \nu\|_{\text{tv}} = \sum_{s \in S} (\mu(s) - \nu(s))^+ = \frac{1}{2} \sum_{s \in S} |\mu(s) - \nu(s)|.$$

Proof. Let $B = \{x \in S : \mu(x) \geq \nu(x)\}$ and let $A \subset S$ be any event. Then

$$\mu(A) - \nu(A) = [\mu(A \cap B) - \nu(A \cap B)] + [\mu(A \cap B^c) - \nu(A \cap B^c)].$$

The second term is negative by definition of B hence

$$\begin{aligned}\mu(A) - \nu(A) \le (\mu - \nu)(A \cap B) \le (\mu - \nu)(B) \\ = \mu(B) - \nu(B).\end{aligned} \tag{5}$$

since $\mu - \nu$ is nonnegative on B. Likewise, by symmetry,

$$\nu(A) - \mu(A) \le \nu(B^c) - \mu(B^c). \tag{6}$$

We claim however that the right hand sides of (5) and (6) are equal. Indeed, $\mu(B) + \mu(B^c) = 1 = \nu(B) + \nu(B^c)$. Therefore,

$$|\mu(A) - \nu(A)| \le \mu(B) - \nu(B) = \nu(B^c) - \mu(B^c).$$

Since A was arbitrary, taking the sup gives

$$\|\mu - \nu\|_{\text{tv}} \le \mu(B) - \nu(B) = \sum_x (\mu(x) - \nu(x))_+,$$

which is one of the desired conclusions. To get the second one, just observe that since $\mu(B) - \nu(B) = \nu(B^c) - \mu(B^c)$, we deduce

$$\|\mu - \nu\|_{\text{tv}} \le \frac{1}{2}([\mu(B) - \nu(B)] + [\nu(B^c) - \mu(B^c)]) = \frac{1}{2}\sum_x |\mu(x) - \nu(x)|,$$

as desired. □

1.3 Mixing times and the cutoff phenomenon

Definition 1.2. *Let P be an irreducible, aperiodic transition matrix on a finite state space S, and let $\pi(x)$ denote its stationary distribution, defined by (1). Define the* ***distance function*** *for all $t = 0, 1, \ldots$ by:*

$$d(t) = \max_{x \in S} \|P^t(x, \cdot) - \pi(\cdot)\|_{\text{tv}}. \tag{7}$$

We also extend the definition of $d(t)$ to all $[0, \infty)$ by setting $d(t) = d(\lfloor t \rfloor)$.

$d(t)$ is the total variation distance between the distribution of the Markov chain at time t and its equilibrium, started from the *worst* possible starting point x, so that if $d(t)$ is small we know that the chain is close to equilibrium no matter what was its starting point. The ergodic theorem (2) implies that $d(t) \to 0$ as $t \to \infty$. In fact, elementary linear algebra tells us that, asymptotically as $t \to \infty$, the distance $d(t)$ decays exponentially fast, with a rate of decay control by the *spectral gap* of the chain (this will be defined and justified in the next chapter).

Proposition 1.1. *Suppose (P, π) is reversible. Let λ be the eigenvalue of P which is of maximal modulus strictly smaller than 1. Then there exists a constant C such that*

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

The proof is a simple consequence of theorem 2.1. A priori, Proposition 1.1 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, $\lambda = 1 - \gamma$ where γ is the (absolute) spectral gap. This tells us that we need to take $t > t_{\text{rel}} := 1/s$ to make $d(t)$ small. As we will se 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 (i.e., for each $n \geq 1$, $X^n = (X^n(t), t = 0, 1,)$ is a family on some state space S_n). Write $d_n(t)$ for the corresponding distance function defined in (7). We say that there is (asymptotically) a cutoff phenomenon at t_n if for every $\varepsilon > 0$,*

$$d_n((1-\varepsilon)t_n) \to 1$$

but

$$d_n((1+\varepsilon)t_n) \to 0.$$

Now fix a Markov chain X on a state space S. Since $d(t)$ converges to 0 as $t \to \infty$, it always makes sense to define, for $0 < \varepsilon < 1$:

$$t_{\text{mix}}(\varepsilon) = \inf\{t \geq 0 : d(t) \leq \varepsilon\}$$

$t_{\text{mix}}(\varepsilon)$ is called the mixing time at level ε. By convention, if we don't specify a level $\varepsilon \in (0,1)$, the **mixing time of the chain** is

$$t_{\text{mix}} = t_{\text{mix}}(1/4) = \inf\{t \geq 0 : d(t) \leq 1/4\},$$

Note that with these definitions, if X^n is a family of chains as above, and if $t^n_{\text{mix}}(\varepsilon)$ is the mixing time at level ε of X^n, then the cutoff phenomenon occurs if and only if

$$t^n_{\text{mix}}(1-\varepsilon) \sim t^n_{\text{mix}}(\varepsilon)$$

for all $0 < \varepsilon < 1$, where $a_n \sim b_n$ means $a_n/b_n \to 1$ as $n \to \infty$. In this case the cutoff phenomenon occurs at $t_n = t^n_{\text{mix}}$.

1.4 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 $Y \sim \nu$.*

So to construct a coupling we seek two random variables which have the correct distributions μ and ν respectively, but there is complete freedom over how they are correlated. Two extreme examples are as follows:

Example 1.2. *Suppose $\mu = \nu$. Then one possible coupling is to take X a random variable with law μ and take $Y = X$. Another coupling is to take Y to be an independent copy of X (i.e., X and Y are i.i.d. with law μ).*

In general, if μ and ν are distinct, we cannot necessarily find a coupling such that $X = Y$. In fact, whether or not $X = Y$ will occur, in the best case scenario, with a probability which is high only if the total variation distance between μ and ν is small, as the following fundamental result shows.

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

$$\|\mu - \nu\|_{\text{tv}} \le \mathbb{P}(X \neq Y). \tag{8}$$

Furthermore, there always is a coupling (X, Y) which achieves equality in (8).

Proof. We start with the proof of the inequality (8), which is practice the only thing we use. (But it is reassuring to know that this is a sharp inequality!) Let (X, Y) denote a coupling of μ and ν.

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

$$\begin{aligned} &|\mu(A) - \nu(A)| = |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)| \\ \le &|\mathbb{P}(X \in A, X = Y) - \mathbb{P}(Y \in A, X = Y)| + |\mathbb{P}(X \in A, X \neq Y) - \mathbb{P}(Y \in A, X \neq Y)| \end{aligned}$$

Note that the first term is in fact equal to zero, while the second is less or equal to $\mathbb{P}(X \neq Y)$, as desired.

We now construct a coupling which achieves equality. Let

$$p = \sum_{x \in S} \mu(x) \wedge \nu(x),$$

and note that $p \in [0, 1]$. (Here $a \wedge b$ means by definition $\min(a, b)$.) Then we note that

$$\|\mu - \nu\|_{\text{tv}} = \sum_x (\mu(x) - \nu(x))_+ = \sum_x \mu(x) - \mu(x) \wedge \nu(x) = 1 - p.$$

Define a probability distribution $\lambda(x) = (\mu(x) \wedge \nu(x))/p$. We let Z be a random variable with distribution λ, and we let Z_1, Z_2 have distribution respectively $(\mu(x) - \nu(x))_+/\|\mu - \nu\|_{\text{tv}}$ and $(\mu(x) - \nu(x)_+/\|\mu - \nu\|$. Then define a pair of random variables (X, Y) as follows. First toss a coin which comes up heads with probability p. If it comes head, take $X = Y = Z$. Else, if it comes tails, take $X = Z_1$ and $Y = Z_2$. Then we claim that (X, Y) defines a coupling of μ and ν. Indeed,

$$\mathbb{P}(X = x) = p\lambda(x) + (1 - p)\frac{(\mu(x) - \nu(x))_+}{\|\mu - \nu\|_{\text{tv}}} = \mu(x)$$

so $X \sim \mu$ as desired. Likewise, $Y \sim \nu$. Moreover, $X = Y$ if and only if the coin comes up heads. Hence

$$\mathbb{P}(X \neq Y) = 1 - p = \|\mu - \nu\|_{\text{tv}},$$

so this coupling achieves equality as desired. □

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:*
$$\rho(t) = \max_{x,y \in S} \|P^t(x,\cdot) - P^t(y,\cdot)\|_{\text{tv}}.$$
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:

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

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

$$\|P^s(x,\cdot) - P^s(y,\cdot)\|_{\text{tv}} = \mathbb{P}(X_s \neq Y_s).$$

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,

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

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:

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

as desired. □

1.5 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 $\{1,\ldots,n\}$ (with $n = 52$ for a real deck of cards). Our convention is that cards are labelled $1,\ldots,n$ 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 card in position i of the deck and moving it to the top the deck can be described mathematically as multiplying on the right the current permutation σ by the cycle $(1\ 2 \ldots\ i)$, that is, the permutation which maps $1 \to 2$, $2 \to 3, \ldots, i \to 1$. That is,

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

where $\cdot$ is the composition of permutations, σ' denotes the new deck of card and σ the old one. [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 $\sigma' = (1\ 2\ \ldots i)\sigma$.]

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. The proof has two parts: an upper bound (showing that $d((1+\varepsilon)n\log n) \to 0$, and hence $t_{\text{mix}} \le (1+\varepsilon)n\log n)$ and a lower bound (showing that $d((1-\varepsilon)n\log n) \to 1$ and hence $t_{\text{mix}}(1-\varepsilon)n\log n)$.

Step 1. Upper bound. We will use a coupling argument, showing that if X_t denotes the state of the deck after t moves, then X_t can be successfully coupled with a uniform deck with high probability (i.e., with probability tending to 1 as $n \to \infty$.)

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 $1 \le i \le n$ 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 this is a valid coupling). Note in particular that the right-hand deck Y_t always has the uniform distribution for every fixed $t \ge 0$.

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

$$\tau_n = \inf\{t \ge 0 : \text{ all cards have been selected at least once}\}$$

then $X_t = Y_t$ for all $t \ge \tau_n$. Hence for all $t \ge 0$:

$$d(t) \le \mathbb{P}(X_t \ne Y_t) \le \mathbb{P}(\tau_n > t). \tag{9}$$

Now, τ_n can be estimated through the coupon-collector problem.

Lemma 1.2 (Coupon-collector problem). *We have that $\tau_n / n \log n \to 1$ in probability. That is,*

$$\mathbb{P}((1-\varepsilon) n \log n \le \tau_n \le (1+\varepsilon) n \log n) \to 1$$

as $n \to \infty$, for all $\varepsilon > 0$.

Proof. Let Z_i denote the time to collect an ith new card after $i-1$ distinct cards have already been collected, $1 \le i \le n$. Then

$$\tau = Z_1 + \ldots + Z_n$$

where the Z_j are independent Geometric random variables, with success probability $p_j = (n-j+1)/n$.

$$\mathbb{E}(\tau_n) = \sum_{j=1}^n \frac{1}{p_j} = \frac{n}{n} + \ldots + \frac{n}{1} = n \sum_{j=1}^n \frac{1}{j} \sim n \log n,$$

as $n \to \infty$. (Here and in the rest of the text, $a_n \sim b_n$ means that $a_n / b_n \to 1$.) Moreover, note that $\mathrm{var}(Z_j) = (1-p_j)/p_j^2 \le 1/p_j^2$. Therefore,

$$\mathrm{var}(\tau_n) = \mathrm{var}(Z_1) + \ldots + \mathrm{var}(Z_n) \le \sum_{i=1}^n \left(\frac{n}{i}\right)^2$$

and thus

$$\mathrm{var}(\tau_n) \le Cn^2,$$

where $C = \pi^2/6$. Hence, since $\mathrm{var}(\tau_n) = o(\mathbb{E}(\tau_n)^2)$, by Chebyshev's inequality τ_n is concentrated near its expectation:

$$\mathbb{P}(\tau_n > (1+\varepsilon)\mathbb{E}(\tau_n)) \le \frac{\mathrm{var}(\tau_n)}{\varepsilon^2 \mathbb{E}(\tau_n)^2} \to 0$$

and similarly for the lower bound. Hence the lemma is proved. □

Recalling (9) we conclude that for $t = (1+\varepsilon)n\log n$,

$$d(t) \leq \mathbb{P}(\tau_n > t) \to 0,$$

which proves the upper-bound on the mixing time.

Step 2: lower bound. 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 $i_1, \ldots, i_j$ then we have $i_1 > i_2 > \ldots > i_j$. 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 1.3. *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 $n \to \infty$.*

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.

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

Therefore,

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

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

1.6 Example: random walk on a hypercube.

Let $H_n = \{0,1\}^n$ be the n-dimensional hypercube. A random walk $(X_t, t = 0, 1, \ldots)$ on the hypercube is the following process: at each step, choose $1 \leq i \leq n$. Then X_{t+1} is obtained from X_t by flipping the coordinate X_t^i. That is, if $X_t^i = 0$ we put $X_{t+1}^i = 1$, and vice-versa. This defines a Markov chain which is irreducible but not aperiodic (e.g., starting from the origin, we can only return to it at even times). To discuss the mixing behaviour of the chain, we have two options. Option 1 is to consider the **lazy chain**: at each step we flip a fair coin, and only perform the above if the coin is heads. Option 2 is to consider the **continuous time chain**: instead of doing such a move at each time step, we wait an exponential random variable with mean 1 to perform it. In other words, we perform a move at rate 1 in continuous time. Both options get around the periodicity issue and hence both chain converge to an equilibrium measure π which it is easy to check is simply the uniform distribution on H_n. Note however that essentially the lazy chain moves half as fast as the continuous one, and thus will basically take twice as long to mix.

Theorem 1.3. *Let $d_L(t)$ denote the distance to stationarity (in total variation) of the lazy chain, and let $d_C(t)$ denote the same for the continuous time chain. Then for any $\varepsilon > 0$, $d_L((1+\varepsilon)n\log n) \to 0$, while $d_C((1/2)(1+\varepsilon)n\log n) \to 0$, as $n \to \infty$.*

Proof. We explain the idea for the lazy chain. Again, we use a coupling argument. Essentially the idea is the following. At every step, we choose a coordinate and flip it with probability 1/2. It is easy to see that the effect on that coordinate is to *randomise* it. The bound in the theorem is therefore just the coupon collector bound.

Formally, we let Y_0 have the uniform distribution on H_n. We couple X_t and Y_t as follows. At each time step, pick $1 \le i \le n$ at random and flip a coin. *Case 1.* Suppose first $X_t^i = Y_t^i$. In that case, we flip X_t^i and Y_t^i simultaneously, if the coin comes up heads. *Case 2.* Now suppose that $X_t^i \neq Y_t^i$. Then flip X_t^i but not Y_t^i if the coin comes up heads, and the other way around if the coin comes up tails. Note that once a coordinate i has been selected, we have $X_t^i = Y_t^i$ forever after. Hence if τ_n is the first time all coordinates have been chosen, we have $X_t = Y_t$ for all $t \ge \tau_n$. The theorem in the lazy case hence follows from the coupon-collector bound, lemma 1.2.

We also sketch the proof of the continuous time case, where we also do a coupling as follows. Let Y_0 have the uniform distribution on H_n. We can assume that X_0 and Y_0 differ at an even number of coordinates. (If not, let X evolve for one step and do nothing to Y first: this problem is then resolved). At each step we pick a i at random in $\{1,\ldots,n\}$. 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_t and bit j for Y_t. 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 bounded by the stopping time τ_n such that all coordinates have been touched. At every step we touch two different coordinates so playing with the coupon collector problem gives us

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

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 in the continuous time case: indeed, each coordinate evolves as an independent bit-flipping Markov chain, changing at rate $1/n$. For this, the probability to be at 0 is*

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

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

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

From this formula it is not hard to see that cutoff occurs precisely at time $t_{\text{mix}} = (1/4)n\log n$ in continuous time.

Intuitively, this is because it it is not necessary to randomise every coordinate: in fact, randomising all but $\sqrt{n}$ gives a sample which is very close to the stationary distribution. (Essentially, by the CLT, if one takes a set of n i.i.d. Bernoulli random variables, and set to zero a random sample of size $m = o(\sqrt{n})$ coordinates, then the resulting sequence would be statistically indistinguishable from the original one). This idea is explored in detail in one of the exercises to show an alternative proof of cutoff.

1.7 Example: couplings for random transpositions*

Random transpositions is the Markov chains on the symmetric group S_n where at each step, two cards are selected by the left and right hand respectively, independently and uniformly at random, and the two cards are swapped (in particular, the two cards can be the same, in which case nothing happens). Mathematically, we view the deck of cards as a permutation $\sigma \in S_n$ where $\sigma(i)$ indicates the position of card with label i.

Then the process above can be written as the Markov chain where the transition probability satisfies

$$P(\sigma, \sigma') = \begin{cases} 1/n & \text{if } \sigma = \sigma' \\ 2/n^2 & \text{if } \sigma' = \sigma \cdot (i,j) \text{ for some } 1 \le i \neq j \le n \\ 0 & \text{else} \end{cases} \tag{10}$$

This is perhaps the simplest shuffle one can think about (from a mathematical point of view, not from a practical one!) Historically this was the first to be analysed. in the landmark paper by Diaconis and Shahshahani [10]. Diaconis recounts that this started with a zealous engineer at Bell Labs who proposed generating permutations by composing $n-1$ random uniform transpositions (since after all, by a result of Cayley, *any* permutation can be written as a product of $n-1$ transpositions). This prompted Diaconis and Shahshahani to investigate how many transpositions were required.

Note that when a card is touched (either with the left or the right hand) then it is moved to a location selected by the other hand, which is independent and uniform. It is therefore natural to believe that when *all* cards have been touched at least once, then the resulting permutation is random. Since essentially two cards are touched at every step (except for the relatively rare case when the two hands chose the same card) by the coupon collector problem, one expects that

$$t_{\text{mix}} = (1 + o(1))\frac{1}{2} n \log n.$$

Note in particular that any card that had never been touched remains in its initial position and so (if the chain was started from the identity, which we can assume without loss of generality) must be a fixed point of the permutation. Using this observation it is easy to see that

$$t_{\text{mix}} \ge (1 + o(1))\frac{1}{2} n \log n. \tag{11}$$

To obtain upper bounds on t_{mix}, we will make use of some marking schemes (i.e., we will mark certain cards as time goes on) which will have the property that conditionally on the set of marked cards, their relative ordering in the deck is uniform at any given time. We will call such a marking scheme *admissible*. Here our presentation is similar to Diaconis [9, Chapter 4B].

Naïve marking scheme. Let L_t and R_t be the label of the cards selected by the left and right hands. Then mark R_t if both L_t and R_t are unmarked.

Proposition 1.3. *The naïve marking scheme is admissible. Furthermore, if τ is the first time only one card remains unmarked, then $\tau \leq (\pi^2/6)n^2$ with probability tending to 1. Hence $t_{\text{mix}} \leq (2\pi^2/3)n^2$.*

Proof. We start with the proof that the scheme is admissible. Informally, one way to build a uniform deck of cards is to add cards one at a time, each time inserting a new card at a uniformly chosen slot. We will see that the marked cards are exactly performing this construction.

To be precise, let $K_t = k$ denote the number of marked cards. Let $\mathcal{M}_t = \{M_1, \dots, M_k\}$ denote the labels of the marked cards, and let $\mathcal{P}_t = \{P_1, \dots, P_k\}$ denote the set of positions in the deck where there is a marked card. Let $\Pi_t : \{M_1, \dots, M_k\} \to \{P_1, \dots P_k\}$ describe how these cards are arranged within the deck. We will check by induction on t that

(a) Given $K = k$: $\mathcal{M}_t, \mathcal{P}_t$ are independent uniform subsets of size k,

(b) Given $K = k$, $\mathcal{M}_t$, and $\mathcal{P}_t$; the bijection $\Pi_t : \mathcal{M}_t \to \mathcal{P}_t$ is uniform.

The fact that (a) holds comes from the fact that when a new card R_{t+1} is marked, it is put in a new position (defined by L_{t+1}) which is independent from the label, and uniform. Hence $\mathcal{M}_{t+1} = \mathcal{M}_t \cup \{R_{t+1}\}$, while $\mathcal{P}_{t+1} = \mathcal{P}_t \cup \{\sigma_t(L_{t+1})\}$ (recall that $\sigma_t(i)$ gives the position at time t of the card with label i). The condition that both cards L_{t+1}, R_{t+1} are unmarked at time t is the condition that $R_{t+1} \notin \mathcal{M}_t$ and that $\sigma_t(L_{t+1}) \notin \mathcal{P}_t$. This proves that (a) holds when a new card is marked, and it is trivial to check it also when no new card is marked at $t + 1$.

To check (b), we proceed a bit more carefully, and consider separately the two cases where a new card is marked at time $t + 1$ and no new card is marked at that time.

- Suppose we are marking a new card at time $t + 1$. By (a), we see that given $\mathcal{M}_{t+1}$ and $\mathcal{P}_{t+1}$, the label m and position p of the newly marked card are uniformly distributed among these two subsets and are independent of each other. Moreover, conditionally on $\mathcal{M}_{t+1}$ and $\mathcal{P}_{t+1}$, and conditionally on $R_{t+1} = m \in \mathcal{M}_{t+1}$ being marked at time $t+1$, and the new position $p \in \mathcal{P}_{t+1}$, being added, we determine uniquely Π_{t+1} by specifying $\Pi_t : \mathcal{M}_t \to \mathcal{P}_t$ and the card L_t for which (under these restrictions) there are exactly $k + 1$ choices. Hence all choices for Π_{t+1} are equally likely.
- Suppose no new card is being added at time $t + 1$. This can happen for two reasons:

  – both cards were already marked. In this case we are applying to Π_t a uniformly chosen transposition, which does not change uniformity.

  – one of the two cards (say L_t) was marked, but not the other. Then the bijection Π_{t+1} is not affected, except for the obvious relabelling of the element of $\mathcal{P}_t$ corresponding to the marked card L_t at time t, whose position is changed from $\sigma_t(L_t)$ to $\sigma_{t+1}(L_t)$. Once again this does not affect uniformity of Π_{t+1}.

Either way, both (a) and (b) hold, so the scheme is admissible.

We now estimate τ, and note that when there are i marked cards, the probability to mark a new one is $((n-i)/n)^2$ at each step. Hence $\tau = \sum_{i=0}^{n-2} N_i$, where N_i is geometric with success probability $((n - i)/n)^2$. It is easy to deduce that $\mathbb{E}(\tau) \leq (\pi^2/6)n^2$ and hence by Markov's inequaility, $t_{\text{mix}} = t_{\text{mix}}(1/4) \leq (2\pi^2/3)n^2$. □

Broder's marking scheme. We now describe a better marking scheme, which is better in the sense that it achieves the right order of magnitude $O(n \log n)$ (but not with the right constant (1/2) in front, so that this is still not sufficient to prove cutoff). This scheme is due to Andre Broder, and goes as follows. At time t, mark R_t if R_t is unmarked and:

- Either L_t marked,
- Or $L_t = R_t$.

Proposition 1.4. *The Broder scheme is admissible. Moreover, if τ is the time at which all cards are marked, then $\tau \le 2(1+o(1))n \log n$ with high probability, hence $t_{\text{mix}} \le 2n \log n$.*

Proof. The proof for this is very similar to the naïve scheme, and we keep the same notations. More generally, by induction on t it is easy to check that, given $K_t = k$ (the number of marked cards), given $\mathcal{M}_t$ (the labels of the marked cards) and given $\mathcal{P}_t$ (the positions of the marked cards), we have Π_t uniform among the bijections $\mathcal{M}_t \to \mathcal{P}_t$.We have two cases to consider. Either a new card is marked, or not.

- Suppose a new card is marked at time t+1. Then, conditionally on $K_{t+1} = k, \mathcal{M}_{t+1}, \mathcal{P}_{t+1}$, then L_{t+1} is equally likely to be any card of $\mathcal{M}_{t+1} = \mathcal{M}_t \cup \{R_t\}$, and $\mathcal{P}_{t+1} = \mathcal{P}_t \cup \{p\}$ where $p = \sigma_t(R_{t+1})$ is the position at time t of the newly marked card. Then conditionally on $\mathcal{M}_{t+1}, \mathcal{P}_{t+1}, R_{t+1}$, we determine uniquely Π_{t+1} by specifying Π_t and the card L_{t+1} (which is equally likely to be in any of the $k+1$ possibilities allowed by these restrictions). Hence every bijection $\Pi_{t+1} : \mathcal{M}_{t+1} \to \mathcal{P}_{t+1}$ is, conditionally on this information, equally likely. Hence this remains true after we remove the conditioning on R_{t+1} and the new position p.
- Suppose no new card is marked at time $t+1$. There are three ways in which no new cards are marked:

  – L_{t+1} and R_{t+1} are different and both unmarked: nothing happens to Π_t.

  – Both L_{t+1} and R_{t+1} were marked at or before time t. In this case we are perform a uniform transposition within the set of marked cards and this does not change the uniformity of Π_{t+1}.

  – Otherwise, L_{t+1} was unmarked and $R_t \neq L_t$ was marked at or before time t. Then the map Π_{t+1} is unchanged, except for updating the position of the marked card R_{t+1} from its old position at time t to its new position, corresponding to $\sigma_t(L_{t+1})$.

Either way, we have proved that the scheme is admissible.

To finish the proposition, we simply estimate τ. Let τ_k be the time for the kth new card to be marked. Then τ_k are independent geometric random variables, with success probability $p_k = (k+1)(n-k)/n^2$. Hence

$$\mathbb{E}(\tau) = \sum_{k=0}^{n-1} \frac{n^2}{(k+1)(n-k)} = \frac{n^2}{n+1} \sum_{k=0}^{n-1} (\frac{1}{k+1} + \frac{1}{n-k}) \sim 2n \log n. \tag{12}$$

Moreover, $\text{var}(\tau_k) = (1-p_k)/p_k^2 \le 1/p_k^2$. Hence, using the fact that $(a+b)^2 \le 2(a^2+b^2)$,

$$\text{var}(\tau) \le \sum_{k=0}^{n-1} \frac{n^4}{(k+1)^2(n-k)^2} \le n^2 \times 4 \sum_{k=0}^{n-1} \frac{1}{k^2} = O(n^2).$$

Hence the result follows from Chebyshev's inequality. □

Matthews' marking scheme. The marking scheme devised by Matthews is an improvement on the Broder scheme, but also incorporates feature of the naïve marking scheme. Essentially, the idea is that the Broder scheme is too slow because we only mark cards with our right hand, and not with our left hand. If we were able to mark with both hands we might hope for a speed up by a factor of two. (As we will see, it is not completely trivial to devise a marking scheme where cards are marked with either hand). That would bring the bound on t_{mix} to $(1+o(1))n\log n$. Even this is not enough to prove cutoff, which we expect to be at $(1/2)n\log n$ (recall our lower bound (11)).

Here the observation is that Broder's scheme is also quite slow at the beginning: for instance in (12), there is an $n\log n$ term which comes from marking the first k cards with $k = \lfloor n/2 \rfloor$. Hence Matthew's idea is to combine the above modification of Broder's scheme with the naïve marking scheme for the first $\lfloor n/2 \rfloor$ cards. This leads to the following scheme: let L_t, R_t be the cards picked by the left and right hand respectively, and let $K_t = k$ be the number of marked cards. Then:

- if $k \le \lfloor n/2 \rfloor$, mark R_t if both L_t, R_t are unmarked.
- if $k > \lfloor n/2 \rfloor$, we mark as follows.

  (a) if one of the two cards is unmarked and not the other, mark the unmarked one.

  (b) Also mark an unmarked card i if $L_t = R_t = i$.

  (c) Also mark an unmarked card i if $(L_t, R_t) = \phi(i)$, where ϕ is an injection from the unmarked cards $[n] \setminus \mathcal{M}_{t-1}$ into the pairs of marked cards $\mathcal{M}_{t-1}^2$.

Theorem 1.4. *Matthews' marking scheme above is admissible. Furthermore, if τ is the first time all cards have been marked, then*

$$\tau \le (1+o(1))(1/2)n\log n$$

with probability tending to 1 as $n \to \infty$. In particular,

$$t_{\text{mix}} \le (1/2)n\log n$$

and the cutoff phenomenon takes place for random transpositions.

Proof. Given the work we have already done in Propositions 1.3 and 1.4 (whose notations we will keep using for simplicity), the proof will be relatively simple. We prove by induction on t that given $\mathcal{M}_t, \mathcal{P}_t$, the map $\Pi_t : \mathcal{M}_t \to \mathcal{P}_t$ describing the arrangements of marked cards into positions is uniform. Of course when $K_t = k \le \lfloor n/2 \rfloor$ this is already known by Proposition 1.3.

So consider the case $K_t = k > \lfloor n/2 \rfloor$. The point is that case (c) has been added so that conditionally on marking a new card (say i) at time $t+1$, then i has the correct probability of not changing position at the time of being marked, so that altogether its position within $\mathcal{P}_{t+1}$. To see that this is the case, note that the probability to add the card i to $\mathcal{M}_t$ at time $t+1$ is precisely

$$\begin{aligned}\mathbb{P}(\mathcal{M}_{t+1} = \mathcal{M}_t \cup \{i\}) &= \mathbb{P}(L_{t+1} = i, R_{t+1} \in \mathcal{M}_t) + \mathbb{P}(R_{t+1} = i, L_{t+1} \in \mathcal{M}_t) \\ &\quad + \mathbb{P}(L_{t+1} = R_{t+1} = i) + \mathbb{P}((L_{t+1}, R_{t+1}) = \phi(i)) \\ &= (1/n)(k/n) + (1/n)(k/n) + 1/n^2 + 1/n^2 \\ &= 2(k+1)/n^2. \end{aligned} \tag{13}$$

and the probability of i being a fixed point (i.e., to not move position) is precisely coming from the last two events, which contribute $2/n^2$. Hence the conditional probability is $1/(k+1)$, as desired. It is easy to deduce that the position of the newly marked card within $\mathcal{P}_{t+1}$ is therefore uniform. From this the same arguments as in Proposition 1.4 finish the proof of the admissibility of Matthews' marking scheme.

To finish the proof, note that the time N to conclude the first (naïve) phase has expectation

$$\mathbb{E}(N) = \sum_{i=1}^{n/2} \frac{n^2}{(n-i)^2} \leq 2n$$

, hence $N \leq n \log \log n$ with high probability by Markov's inequality.

Also, the time M to conclude the second phase (once the first phase has ended), can be written as

$$M = \sum_{k=\lfloor n/2 \rfloor}^{n-1} \tau_k$$

where τ_k is Geometric with success probability $p_k = 2(k+1)(n-k)/n^2$, from (13). Hence

$$\mathbb{E}(M) \leq \sum_{k=n/2}^{n} 1/p_k \leq \sum_{k=n/2}^{n} \frac{n^2}{2(k+1)(n-k)}.$$

The same proof as in Proposition 1.4 then shows that $M \leq (1+o(1))(1/2)n \log n$ with high probability as $n \to \infty$. Since $\tau = N + M$ the result is proved. □

2 Spectral methods and relaxation time

2.1 Eigenvalue decomposition

Our presentation here follows quite closely Chapter 12 of [17].

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

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

Suppose P is irreducible and aperiodic. Rather than viewing P as a matrix (i.e., a linear operator on vectors), it will be more useful to view P as an operator on functions $f : S \to \mathbb{R}$ by setting:

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

This is nothing but the standard action on column vectors. Likewise, the matrix P^t can be interpreted as an operator on functions in the same way, and recall also that by the Chapman–Kolmogorov equation, $P^t(x,y)$ is nothing but the transition probability of the chain from x to y in t steps: thus $P^t(x,y) = \mathbb{P}(X_t = y | X_0 = x)$.

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 $\ell^2(\pi)$. We will write the corresponding norm by $\|f\|_2^2 = \langle f, f \rangle_\pi$. More generally, we can define

$$\|f\|_p = \left(\sum_x |f(x)|^p \pi(x) \right)^{1/p}.$$

One of the traditional techniques for studying Markov chains is to diagonalise 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: that is,

$$\pi(x)P(x,y) = \pi(y)P(y,x) \forall x, y \in S.$$

Then all eigenvalues are real and we can order them in decreasing order from 1 to -1:

$$\lambda_1 = 1 > \lambda_2 \ge \ldots \ge \lambda_n \ge -1.$$

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 $(1, \dots, 1)^T$.*

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 eignevalue 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. □

2.2 The spectral gap and the relaxation time

Definition 2.1. *Supose P irreducible and aperiodic. 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*

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

Now suppose that P is also reversible. Note that the negative eigenvalues are in general not so relevant. One way to see this is to consider the *lazy chain* which, we recall, is defined by saying that with probability 1/2, the lazy chain does nothing, and with probability 1/2, it takes a step according to P. In particular, its transition matrix $\tilde{P}$ satisfies

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

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*

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

The basic idea for the proof of this theorem is that the total variation distance is an ℓ^1 distance which can be dominated by an ℓ^2 distance thanks to Cauchy–Schwarz. We then observe that the eigenvalues essentially measure convergence in ℓ^2. We start with by giving the definition of the ℓ^2 distance.

Definition 2.2. *Let $d_2(t) = \sup_{x \in S} \left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - 1 \right\|_2$. We call $d_2(t)$, the ℓ_2 distance to stationarity.*

Lemma 2.1. *Assume that P is irreducible, aperiodic (but not necessarily reversible). Then we have $d(t) \leq (1/2)d_2(t)$.*

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

$$4\|P^t(x,\cdot) - \pi\|^2 \leq \left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - 1 \right\|_2^2 \leq d_2(t)^2.$$

Taking the maximum over x gives the result. □

We are now ready to prove the theorem.

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

$$\begin{aligned} \left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - 1 \right\|_2^2 &= \left\| \sum_{j=2}^n f_j(x) f_j(\cdot) \lambda_j^t \right\|_2^2 \\ &= \sum_{j=2}^n \lambda_j^{2t} f_j(x)^2 \leq \lambda_*^{2t} \sum_{j \geq 2} f_j(x)^2. \end{aligned} \tag{14}$$

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

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

Maximising over x and taking the square root, we get

$$d(t) \le \frac{1}{2} e^{-\gamma_* t} \sqrt{\pi_{\min}^{-1}}. \tag{15}$$

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

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

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

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

$$|\lambda|^t \le 2d(t) \tag{16}$$

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

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

and hence

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

Taking the logarithm. we get

$$\log\left(\frac{1}{2\varepsilon}\right) \le t_{\text{mix}}(\varepsilon) \log(1-\gamma_*).$$

Using $x \ge (1-x)\log(1-x)$ for all $x \in [0,1]$, applied to $x = \gamma_*$, we deduce

$$t_{\text{mix}}(\varepsilon) \ge \log(\frac{1}{2\varepsilon})(t_{\text{rel}} - 1)$$

as desired. □

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*

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

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

$$4d(t)^2 \leq \sum_{j=2}^{n} \lambda_j^{2t}. \tag{17}$$

2.3 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 $W_n = \{1, \omega, \omega^2, \ldots, \omega^{n-1}\}$ 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 $\phi_j(z) = z^j$, $1 \leq j \leq n$, give us n eigenvalues. This can be seen geometrically: see Figure 12.1 in [17]. More formally, note that

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

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 0. 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 $n \to \infty$. 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 detailed knowledge of all the eigenvalues. (However, we already note that another method will be provided with Nash inequalities. Finally, we note that we will get an order of magnitude on the spectral gap in the next lectures using Poincaré inequalities and the path method for Dirichlet energies.)

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., (17)) to prove this result. We have:

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

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 (16), 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. □

2.4 Example: random walk on the hypercube

Consider the lazy random walk on the hypercube of section 1.6. The hypercube can be thought of as the group $(\mathbb{Z}/2\mathbb{Z})^n$ and it is therefore not surprising that the diagonalisation can be done explicitly. Set $H_n = \{-1, 1\}^n$ and for $J \subset \{1, \ldots, n\}$ let

$$f_J(x) = \prod_{i \in J} x_i; \qquad x \in H_n.$$

It is straightforward to check that f_I are orthonormal with respect to $\ell^2(\pi)$ and hence (since there are 2^n such functions) form an orthonormal basis of $\ell^2(\pi)$. Furthermore, if P is the transition matrix of the lazy random walk on H_n, then

$$Pf_J(x) = (1 - |J|/n)f_J(x)$$

since with probability $1 - |J|/n$, none of the coordinates in J are changed (in which case f_J remains unchanged) and otherwise one of the coordinates in J is randomised, making f_J equal to zero on average.

Hence each $f_J, J \subset \{1, \ldots, n\}$ is an eigenfunction and because $|H_n| = 2^n$ we have found them all. In particular, the spectral gap is

$$\gamma = 1/n. \tag{18}$$

Note that the relation between spectral gap and mixing only implies $t_{\text{mix}} \leq Cn^2$, whereas in reality $t_{\text{mix}} = (1/2 + o(1))n \log n$ (for the lazy walk).

3 Geometric methods

The previous chapter related mixing to spectral notions such as the spectral gap. In many cases, it is difficult to compute the spectral gap explicitly, and instead this has to be estimated. In this chapter we gather several geometric methods for estimating the spectral gap: the first is the method of canonical paths of Diaconis and Saloff-Coste, which gives a Poincaré inequality (and thus, as we will see below, an estimate of the spectral gap) by a path counting argument. The second is Cheeger's inequality, which relates the spectral gap to bottleneck ratios.

We start with Dirichlet forms, which are energy functionals associated with a Markov chain, and the variational characterisation of the spectral gap. These naturally lead to Poincaré inequalities, which are the basis of both methods.

3.1 Dirichlet forms and variational characterisation of spectral gap

We start with the following basic facts. Let S be a finite state space with $|S| = n$. Let P be a transition matrix on S. Suppose P has an invariant distribution π, and recall our notation $\langle f, g\rangle_\pi = \sum_x f(x)g(x)\pi(x)$. The following is an important notion in the theory of Markov chains.

Definition 3.1. *Let $f, g : S \to \mathbb{R}$. 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

$$\mathcal{E}(f, g) = \frac{1}{2}\sum_{x,y} \pi(x)P(x, y)[f(y) - f(x)][g(y) - g(x)], \tag{19}$$

a much more useful expression.

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

$$Q(e) = \pi(x)P(x, y)$$

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, (19) becomes

$$\mathcal{E}(f, g) = \frac{1}{2}\sum_e Q(e)\nabla f(e)\nabla g(e).$$

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:

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

Thus when $f = g$, $\mathcal{E}(f,f)$ measures how rough or how smooth the function f is.

The following "variational characterisation" (or minimax characterization) of the spectral gap in terms of the Dirichlet form is very useful in practice:

Theorem 3.1. *Assume (P, π) is reversible, let γ be the spectral gap. Then*

$$\gamma = \min_{\substack{f:S\to\mathbb{R} \\ \mathbb{E}_\pi(f)=0, \|f\|_2=1}} \mathcal{E}(f,f) = \min_{\substack{f:S\to\mathbb{R} \\ \mathbb{E}_\pi(f)=0}} \frac{\mathcal{E}(f,f)}{\|f\|_2^2}.$$

Equality is attained for $f = f_2$.

Remark 3.1. *The variational problem in the above theorem is standard in mathematical physics. It has a nice geometric interpretation: the aim is to find the smoothest function on S with zero mean and L^2 norm equal to one. It is instructive to think about what this optimal function would look like in the case of a domain $D \subset \mathbb{R}^d$.*

Proof. By scaling it suffices to prove the first equality. Now note that $\mathbb{E}_\pi(f) = 0 = \langle f, 1\rangle_\pi$ so the condition $\mathbb{E}_\pi(f) = 0$ means that $f \perp 1 \equiv f_1$. Thus consider 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. Using orthonormality of the eigenfunctions, and the fact that $\|f\|^2 = 1$:

$$\begin{aligned}
\mathcal{E}(f,f) &= \langle (I-P)f, f\rangle_\pi \\
&= \sum_{j=1}^n |\langle f, f_j\rangle_\pi|^2 (1-\lambda_j) \\
&\geq (1-\lambda_2) \sum_{j=2}^n \langle f, f_j\rangle^2 = (1-\lambda_2)\|f\|_2^2 = \gamma.
\end{aligned}$$

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

3.2 Poincaré inequalities and the canonical paths method

Note that if $f : S \to \mathbb{R}$ 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,

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

Note that in this form, this inequality is true for all functions $f : S \to \mathbb{R}$, not just those with mean zero. This motivates the following definition:

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

$$\mathrm{var}_\pi(f) \le C\mathcal{E}(f,f). \tag{20}$$

As just discussed, a Poincaré inequality *always* holds with $C = 1/\gamma$. Conversely, if a Poincaré inequality holds with constant C, then $\gamma \ge 1/C$. Hence if one establishes the inequality (20) for some constant C this shows $t_{\mathrm{rel}} \le C$ at least if the chain is lazy.

We will now see a general and very useful method, called the **canonical paths method** of Diaconis and Saloff-Coste, which shows how Poincaré inequalities can be established from geometric consideration. In turn, this provides estimates on the spectral gap by theorem 3.1 and hence also on the mixing time by theorem 2.2. (While in this form the result was stated by Diaconis and Saloff–Coste [11], various precursor results were proved earlier, notably by Jerrum and Sinclair [15].) For any $x, y \in S$, suppose that we fix, once and for all, a certain path $\gamma_{x,y}$, i.e., a collection of states $x_0, \ldots, x_k$ such that $x_0 = x$ and $x_k = y$, and $P(x_i, x_{i+1}) > 0$ for all $1 \le i \le k-1$. Let $|\gamma_{x,y}| = k$ denote the length of this path. Then we have the following result:

Theorem 3.2. *The Poincaré inequality (20) holds with*

$$C = \max_e \left\{ \frac{2}{Q(e)} \sum_{x,y:e\in\gamma_{x,y}} |\gamma_{xy}|\pi(x)\pi(y) \right\},$$

where the maximum is over all $e = (u,v) \in S \times S$ such that $P(u,v) > 0$. In particular, $\gamma \ge 1/C$.

The number C may be thought of as a *congestion ratio*: it is large if there is an edge e such that e is on γ_{xy} for "many" choices of x and y.

Proof. The proof is a straightforward application of Cauchy–Schwarz: if f is a function on the state space,

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

This finishes the proof. □

3.3 Some simple examples

As a direct application of the path method, we show how to get bounds on the spectral gap of random walk on some natural examples of graphs, without explicitly computing the eigenvalues.

3.3.1 Example: random walk on a box.

Consider the box $[n]^d = \{1, \dots, n\}^d$ with the restriction of the edges of $\mathbb{Z}^d$ to these vertices. Then there exists C which depends on d but not on n such that $\gamma \geq C/n^2$.

Proof. We apply the path method: for $x, y \in V$, we choose the path $\gamma_{x,y}$ defined as follows. We first try to match the first coordinate of x and y, then the second coordinate, and so on until the last coordinate. Each time, the changes in coordinate is monotone. As an example if $d = 2$ and $x = (x_1, x_2)$ and $y = (y_1, y_2)$, then define $z = (y_1, x_2)$. Then γ_{xy} is the reunion of two paths, γ_1 and γ_2, such that γ_1 goes in horizontal line from x to z and γ_2 in vertical line from z to y.

Then it is straightforward that $\pi(x) \leq C/n^d$ for all $x \in V$, while $Q(e) \geq c/n^d$. Also, the maximal length of a path $\gamma_{x,y}$ is clearly no more than Cn for some $C > 0$. Hence

$$\begin{aligned} C &= \max_e \left\{ \frac{2}{Q(e)} \sum_{x,y: e \in \gamma_{x,y}} |\gamma_{xy}| \pi(x) \pi(y) \right\} \\ &\leq \max_e \left\{ \frac{O(1)}{n^{d-1}} \#\{x, y : e \in \gamma_{x,y}\} \right\}. \end{aligned}$$

Now we claim that the number above is at most n^{d+1}. This is obtained when e is roughly at the center of the box. In this case it is easy to see that for each choice of y in the right-half of the box there are approximately n choices of x for which the path γ_{xy} uses he edge e. □

3.3.2 Example: random walk on a tree.

Consider a finite tree T, with n vertices and maximal degree d, and maximal height H. Then for some absolute constant c,

$$\gamma \geq \frac{c}{d^2 n H}.$$

Hence in particular, if T is a uniform random tree with n vertices,

$$\gamma \geq c(\log n)^{-2-\varepsilon} n^{-3/2},$$

with high probability as $n \to \infty$ for any $\varepsilon > 0$. This is known to be optimal except for the log prefactors. (However, on the regular binary tree this bound is far from optimal, as the spectral gap is known to be a constant).

For a uniform random tree, it is known that $H \leq \sqrt{n}(\log n)^\varepsilon$ with high probability for any $\varepsilon > 0$, from which the claim follows.

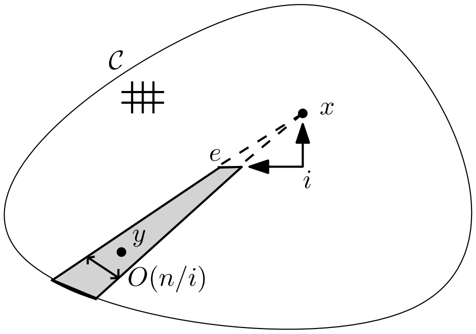

Figure 1: Canonical paths method for random walk on a convex set. For a given edge e and a starting point x at distance i from e, the edge e will be in γ_{xy} if y is in the shaded area, which forms a cone. A slice of that cone will have diameter $O(n/i)$ and hence size $O((n/i)^{d-1})$. Summing over all such slices, we get that the size of the shaded area is $O(n^d/i^{d-1})$.

3.3.3 Example: random walk on a convex set.

Let $\mathcal{C}$ be a bounded convex set in $\mathbb{R}^d$. Consider lazy simple random walk on $G_n = (1/n)\mathbb{Z}^d \cap \mathcal{C}$. Then we claim that

$$t_{\text{rel}} \leq O(n^2) \tag{21}$$

Hence $t_{\text{mix}} = O(n^2 \log n)$. Again, this is sharp except for the logarithmic factor. (More delicate arguments based on Nash inequalities are required to get the sharp bound $t_{\text{mix}} = O(n^2)$.)

Proof. We apply the method of canonical paths, choosing for γ_{xy} a lattice path staying as close as possible from the straight Euclidean segment $[x, y]$. (Note that since $\mathcal{C}$ is assumed to be convex, this stays entirely within $\mathcal{C}$). Again, we find by Theorem 3.2, that a Poincaré inequality holds with constant

$$C = O(1)\frac{1}{n^{d-1}} \max_e \#\{x, y : e \in \gamma_{xy}\}.$$

We need to estimate $\#\{x, y : e \in \gamma_{xy}\}$ for a given edge e. Hence fix an edge e and a point x at (lattice) distance i from e, where $1 \leq i \leq O(n)$ (we write dist for the lattice distance, so $\text{dist}(e, x) = i$. Then the numbers of y such that $e \in \gamma_{x,y}$ is at most $O(n^d/i^{d-1})$ (see figure: this is the size of the grey area). Consequently,

$$\#\{x, y : e \in \gamma_{xy}; \text{dist}(x, e) = i\} \leq O(i^{d-1})O(\frac{n^d}{i^{d-1}}) = O(n^d).$$

Summing over $1 \leq i \leq O(n)$ gives us,

$$\#\{x, y : e \in \gamma_{xy}\} \leq O(n^{d+1}).$$

Hence a Poincaré inequality holds with

$$C \leq O(n^2),$$

giving $t_{\text{rel}} = O(n^2)$ as desired for the lazy random walk. □

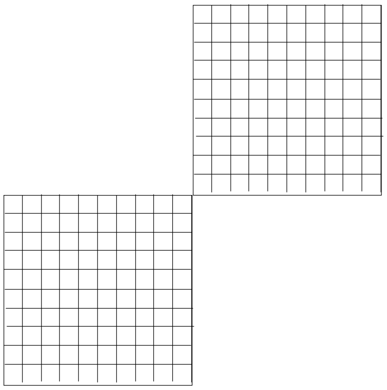

Figure 2: The n-dog graph D_n.

3.3.4 An example with a bottleneck: random walk on the n-dog*

Sometimes, to take more into account the specific geometry of a graph it will be useful to weigh differently certain edges. Thus we introduce a weight function $w : E \to \mathbb{R}_{>0}$ on the edges. We then define the weight of a path $\gamma = \gamma_{xy}$ to be:

$$|\gamma|_w = \sum_{e \in \gamma} w(e).$$

Proposition 3.1. *A Poincaré inequality holds with*

$$C = \max_{e \in \mathcal{A}} \left\{ \frac{2}{w(e)Q(e)} \sum_{x,y: e \in \gamma_{xy})} |\gamma_{xy}|_w \pi(x)\pi(y) \right\}.$$

In particular, $t_{\text{rel}} \le C$ for the corresponding lazy random walk.

The proof is a simple adaptation of Theorem 3.2. We call this the **weighted path method**. In applications we will give a heavy weight to congested edges and less important weight to edges that are relatively free.

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 3.3.4 for a picture.

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. □

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 3.3. *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:

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

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 3.1, and for this we need to choose two things. The first one is the set of paths $\gamma(x,y)$ for $x, y \in D_n$, and the second is the weight function $w : E \to \mathbb{R}$. For the special paths $\gamma(x,y)$ we only define $\gamma(x,0)$ (where 0 is the junction between the two squares) and we define $\gamma(x,y)$ to be the concatenation of the two paths $\gamma_{x,0}$ with $\gamma_{0,y}$: that is, $\gamma_{x,y} = \gamma_{x,0} \cup \gamma_{0,y}$. If $x \in D_n$, we define $\gamma_{x,0}$ 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:

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

(This is the cardinality of the shaded region in figure 3.3.4.) Thus

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

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

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

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

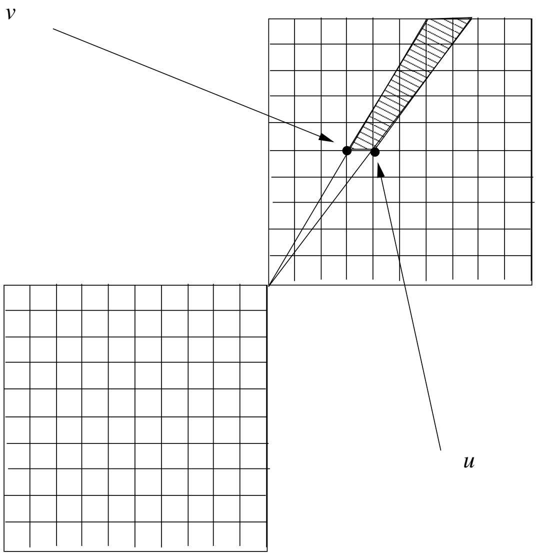

Figure 3: The shaded region corresponds to all the points x, y such that $e = (u, v) \in \gamma_{x,y}$.

Thus by Corollary 3.1, there is a Poincaré inequality with

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

This gives the upper-bound. For the other direction, take $f(x) = \operatorname{sgn}(x)\log(1 + d(0,x))$, where the function $\operatorname{sgn}(x)$ 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 $i \leq n$,

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

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

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

for n large enough. □

3.4 Cheeger's inequality

The method of canonical paths already introduces the idea that if many paths between pairs of vertices go through the same edge, then this slows down the mixing. This idea is taken more systematically with Cheeger's inequality, which relates the *bottleneck ratio* or *isoperimetric number* of a graph to its spectral gap.

Recall our notation $Q(e) = \pi(x)P(x,y)$ for the equilibrium flow through the edge $e = (x,y)$. Let $Q(A,B) = \sum_{x\in A, y \in B} Q(x,y)$ and define the **bottleneck ratio** of a set A to be

$$\Phi(A) = \frac{Q(A,A^c)}{\pi(A)}$$

Essentially this is a measure of the size of the boundary relative to the total size of the set A.

Definition 3.3. *The* ***bottleneck ratio*** *of the Markov chain is defined by*

$$\Phi_* = \min_{A:\pi(A)\le 1/2} \Phi(A).$$

We now state Cheeger's inequality:

Theorem 3.4. *Suppose P is reversible and let $\gamma = 1 - \lambda_2$ be the spectral gap. Then*

$$\frac{\Phi_*^2}{2} \le \gamma \le 2\Phi_*.$$

This was originally proved in the context of Markov chains by Jerrum and Sinclair [15], and is the direct analogue of the inequality proved by Cheeger for Riemannian manifolds (hence the name Cheeger inequality). While the left and right hand side are in general quite different from one another, either bound can be sharp in some examples (see e.g. examples 3.1 and 3.2), so it is remarkable that such an inequality holds in this degree of generality.

A slightly more precise (but equally general) result will be proved later on in Theorem 9.2, which takes into account the whole isoperimetric profile, i.e. the function of $r \le 1/2$ obtained by minimising the bottleneck ratio over sets of mass less than r.

Proof of upper bound of Theorem 3.4. We start with the proof that $\gamma \le 2\Phi^*$ which is easier. By the variational characterisation of the spectral gap,

$$\gamma = \min_{f\neq 0; \mathbb{E}_\pi(f)=0} \frac{\mathcal{E}(f,f)}{\mathrm{var}_\pi(f)}$$

Define f to be the function which is constant on A and A^c by $f(x) = -\pi(A^c)$ if $x \in A$, $f(x) = \pi(A)$ if $x \in A^c$. Then note that $\mathbb{E}_\pi(f) = 0$. Moreover,

$$\mathcal{E}(f,f) = \frac{1}{2}\sum_{x,y} \pi(x)P(x,y)(f(x)-f(y))^2 = \frac{1}{2}[Q(A,A^c) + Q(A^c,A)] = Q(A,A^c).$$

On the other hand,

$$\begin{aligned}\mathrm{var}_\pi(f) = \mathbb{E}_\pi(f^2) &= \sum_{x\in A} \pi(x)\pi(A^c)^2 + \sum_{x \in A^c} \pi(x)\pi(A^2)\\ &= \pi(A)\pi(A^c)^2 + \pi(A^c)\pi(A)^2\\ &= \pi(A)\pi(A^c) \geq \pi(A)/2.\end{aligned}$$

Consequently,

$$\gamma \leq \frac{Q(A,A^c)}{\pi(A)/2}.$$

Taking the minimum over all sets A such that $\pi(A) \leq 1/2$ gives $\gamma \leq 2\Phi_*$, as desired. □

Before the proof of the lower bound (which is harder) we state and prove the following lemma.

Lemma 3.1. *Let $f : S \to [0,\infty)$ be a nonnegative function such that $\pi(f > 0) \leq 1/2$. Order S so that f is non-increasing.*

$$\mathbb{E}_\pi(f) \leq \Phi_*^{-1} \sum_{x<y} Q(x,y)(f(x) - f(y)).$$

Proof. Fix $t > 0$ and let $A = \{x : f(x) > t\}$. Since Φ_* minimises the bottleneck ratio,

$$\Phi_* \leq \frac{Q(A;A^c)}{\pi(A)} = \frac{\sum_{x,y:f(x)>t\geq f(y)} Q(x,y)}{\pi(f>t)}$$

Hence (since we can restrict over x, y such that $x < y$ without loss of generality),

$$\pi(f > t) \leq \frac{1}{\Phi_*} \sum_{x<y} Q(x,y)\mathbf{1}_{\{f(x)>t\geq f(y)\}}.$$

Integrating over t, the left hand side is $\mathbb{E}_\pi(f)$ by Fubini's theorem, so we find

$$\mathbb{E}_\pi(f) \leq \Phi_*^{-1} \sum_{x<y} Q(x,y)(f(x) - f(y))$$

as desired. □

Proof of lower bound of Theorem 3.4. Let f_2 be an eigenfunction of P with eigenvalue $\lambda_2 = 1-\gamma$. On a graph where there is a bottleneck, we can expect the function f_2 to be positive on one side of the bottleneck and negative on the other, since f_2 minimises the Dirichlet energy.

It is therefore natural to assume that $\pi(f_2 > 0) \leq 1/2$ (if not one can use $-f_2$) and let $f = \max(0, f_2)$. It is clear that this reduces the Dirichlet energy, and a moment of thought shows that intuitively, this even reduces the ratio $\mathcal{E}(f,f)/\mathbb{E}_\pi(f^2)$ (note that γ minimises this ratio but only over functions over zero mean, so there is no contradiction).

Lemma 3.2. *We have $\mathcal{E}(f,f) \leq \gamma \mathbb{E}_\pi(f^2)$.*

Proof. Note that $f \geq f_2$ and thus $Pf \geq Pf_2 = \lambda f_2$. Suppose first that $f(x) > 0$, so $f(x) = f_2(x)$. In this case, $(I - P)f(x) = f_2(x) - Pf(x) \leq (1 - \lambda_2)f_2(x) = \gamma f(x)$. Now suppose that $f(x) = 0$. Then $(I - P)f(x) \leq 0 = \gamma f(x)$. Either way,

$$(I - P)f(x) \leq \gamma f(x).$$

Hence taking the inner product with f,

$$\langle (I - P)f, f\rangle_\pi \leq \gamma \langle f, f\rangle_\pi$$

as desired. □

Now, $\mathbb{E}_\pi(f^2)$ can be upper-bounded by Lemma 3.1. We get:

$$\mathbb{E}_\pi(f^2) \leq \Phi_*^{-1} \sum_{x<y} Q(x,y)(f^2(x) - f^2(y)).$$

We do the factorisation $(f^2(x) - f^2(y)) = (f(x) - f(y))(f(x) + f(y))$ and wish to apply the Cauchy–Schwarz inequality with respect to the measure $\mathbf{1}_{\{x<y\}}Q(x,y)$. Note that this doesn't have total mass equal to 1, but the inequality can still be applied. Hence we get:

$$\mathbb{E}_\pi(f^2) \leq \Phi_*^{-1}\big(\sum_{x<y} Q(x,y)(f(x) - f(y))^2\big)^{1/2}\big(\sum_{x<y} Q(x,y)(f(x) + f(y))^2\big)^{1/2}.$$

Using $(a + b)^2 \leq 2(a^2 + b^2)$,

$$\begin{aligned}\mathbb{E}_\pi(f^2) &\leq \Phi_*^{-1}\big(\sum_{x<y} Q(x,y)(f(x) - f(y))^2\big)^{1/2}\big(\sum_{x<y} Q(x,y)(f(x) + f(y))^2\big)^{1/2}\\ &\leq \Phi_*^{-1}\mathcal{E}(f,f)^{1/2}\big(2\sum_{x<y} Q(x,y)(f(x)^2 + f(y)^2)\big)^{1/2}\\ &\leq \Phi_*^{-1}\mathcal{E}(f,f)^{1/2}\big(2\mathbb{E}_\pi(f^2)\big)^{1/2},\end{aligned}$$

where we have used the fact that $\sum_{x<y} Q(x,y)f(x)^2 = \sum_{x<y} Q(x,y)f(y)^2 \leq (1/2)\mathbb{E}_\pi(f^2)$. Squaring and making the cancellation,

$$\mathbb{E}_\pi(f^2) \leq 2\mathcal{E}(f,f)\Phi_*^{-2}.$$

Plugging into Lemma 3.2, we get $\Phi_*^2 \leq 2\mathcal{E}(f,f)/\mathbb{E}_\pi(f^2) \leq 2\gamma$ by Lemma 3.2. This concludes the proof. □

Example 3.1. *Consider random walk on the n-cycle.*

In that case it is not hard to see that $\Phi_* \asymp 1/n$, while the spectral gap is by our previous analysis $\gamma \asymp 1/n^2$. Hence in this case $\gamma \asymp \Phi_*^2$.

On the other hand, here is an example where the other side of Cheeger's inequality is sharp.

Example 3.2. *Consider the lazy random walk on the hypercube $H_n = \{0,1\}^n$.*

We estimate the bottleneck ratio as follows. Set A to be the set

$$A = \{x \in H_n : x_1 = 0\}.$$

Note that $\pi(A) = 1/2$. On the other hand,

$$\begin{aligned} Q(A, A^c) &= \sum_{x\in A, y\in A^c} \pi(x)P(x,y) \\ &= \sum_{x\in A} \frac{1}{2^n} \times \frac{1}{2n} \end{aligned}$$

since from a given $x \in A$ there is only one vertex y such that $P(x,y) > 0$ (obtained by flipping the first coordinate of x), and in that case $P(x,y) = 1/(2n)$. Consequently,

$$\Phi_* \le \frac{Q(A, A^c)}{\pi(A)} \le \frac{1}{2n}.$$

But we know by Cheeger's inequality that $\Phi^* \ge (\gamma/2)$ and we already computed γ in this case in (18), where we found $\gamma = 1/n$. Thus

$$\Phi_* = (\gamma/2) = 1/(2n).$$

In particular in this example, the other side of Cheeger's inequality is sharp.

Remark 3.2. *As mentioned earlier, Cheeger's inequality says in particular that if there is a bottleneck in the graph (namely, if Φ_* is small) then γ is small and hence mixing takes long. In fact, we remark that this remains true even if the chain is not reversible: Theorem 7.3 in [17] states that $t_{\text{mix}} \ge 1/(4\Phi_*)$ even if the chain is nonreversible.*

3.5 Expander graphs*

Cheeger's inequality tells us that when a graph has no bottleneck then the walk must mix fast, which is intuitive. The "best" graphs from that point of view are those for which the Cheeger constant is bounded below.

Definition 3.4. *A family of graphs $\{G_n\}$ is called an **expander family** if the Cheeger constant satisfies $\Phi_* \ge \alpha$ for some $\alpha > 0$ and for all $n \ge 1$.*

Sometimes G_n is also required to be d-regular, meaning that all vertices are of degree d. With this extra requirement, it is not *a priori* clear whether such graphs exist. However, it turns out that *most* d-regular graphs are actually expanders, as shown by the usual proposition.

Theorem 3.5. *Let G_n be a graph uniformly chosen among all d-regular graphs on n vertices. Then there exists $\alpha > 0$ sufficiently small that with probability tending to 1 as $n \to \infty$, G_n is an α-expander.*

Proof. Though the proof falls somewhat outside the scope of these notes, we will include a sketch of proof here (and it is clear that the same proof for random graphs models with given degree distribution, provided that the minimal degree is 3). Essentially the method is to use the construction of a random regular graph from the *configuration model* due to Bollobàs and de la Vega [8]. This gives an explicit way of checking the neighbourhood structure of a given subset of vertices and in particular of estimating the probability that the boundary is small. Summing over all subsets and using Stirling's formula essentially concludes the proof. The proof which follows is adapted from Lemma 3.5 in [12].

The *configuration model* for a random d-regular graph ($d \geq 3$) is as follows: each of the n vertices is given d stubs or half-edges. Then these dn half edges are paired uniformly at random, giving rise to a graph which may or may not contain loops and multi-edges.

Lemma 3.3. *Let E be the event that the graph G generated by the configuration model contains no self-loops or multi-edges. Then $\mathbb{P}(E)$ is bounded away from zero. Moreover, conditionally on E, G is a uniformly chosen random d-regular graph.*

See [8] for a proof. Hence it suffices to prove the theorem for the configuration model G.

Let $D = dn$ be the total degree and let $|E| = D/2$ be the totall number of edges. Let S be a subset of vertices with total degree $d(S) = s$ and suppose that $s \leq (1/2)|E| = D/4$. Consider the event that $|E(S, S^c)| \leq \alpha s$. Then there must be $k \leq s$ points outside of S which are matched to points in S, and $s - k$ points of S which are matched to each other. Note that if we have a set of size m even, then the number of perfect matchings of this set is $(m-1)!! = (m-1)(m-3)\dots 1$. Hence

$$\begin{aligned}\mathbb{P}(|E(S,S^c)| \leq \alpha s) &\leq \sum_{k=0}^{\alpha s} \mathbb{P}(E(S,S^c) = k)\mathbf{1}_{\{s \equiv k \mod 2\}} \\ &\leq \frac{1}{(D-1)!!}\sum_{k=0}^{\alpha s}\binom{s}{k}(s-k-1)!!\binom{D-s}{k}k!(D-s-k-1)!! \\ &\lesssim \frac{1}{D!!}\sum_{k=0}^{\alpha s}\binom{s}{k}(s-k)!!\binom{D-s}{k}k!(D-k-s)!!\end{aligned}$$

In the last inequality we have used the fact that by Stirling's formula, $m!! \asymp (\sqrt{m!})m^{1/4}$, hence $m!! \asymp m^{1/2}(m-1)!!$, as well as the fact that $\sqrt{D} \lesssim \sqrt{s-k}\sqrt{D-s-k}$ (because $k \leq \alpha s \leq D/4$). Hence

$$\begin{aligned}\mathbb{P}(|E(S,S^c)| \leq \alpha s) &\lesssim \sum_{k=0}^{\alpha s}\frac{(D-s-k)^{1/4}(s-k)^{1/4}}{D^{1/4}}\sqrt{\frac{(s-k)!(D-s-k)!}{D!}}\binom{s}{k}\binom{D-s}{k} \\ &\leq \sum_{k=0}^{\alpha s} s^{1/4}\left(\frac{\binom{s}{k}\binom{D-s}{k}}{\binom{D}{s}}\right)^{1/2}\end{aligned}$$

after making cancellations. Now, by Stirling's formula,

$$\binom{n}{k} \asymp \sqrt{\frac{n}{k(n-k)}}\exp(nH(k/n)); \text{ where } \quad H(x) = -x\log x - (1-x)\log(1-x).$$

Hence

$$\begin{aligned}\mathbb{P}(|E(S,S^c)| \leq \alpha s) &\lesssim \sum_{k=0}^{\alpha s} s^{1/2}\exp(\frac{1}{2}[H(\frac{k}{s})s + H(\frac{k}{D-s}) - H(\frac{s}{D})D]) \\ &\leq s^{3/2}\exp(\frac{1}{2}[H(\alpha)s + H(\frac{2\alpha s}{D}) - H(\frac{s}{D})D])\end{aligned}$$

where we have used the fact that $H(x)$ is monotone increasing over $x \in [0,1/2]$ and $k/s \le \alpha \le 1/4$, as well as $k/(D-s) \le \alpha s/(D/2)$.

We now sum over all possible sets S such that $d(S) = s$, where $s \le D/4$. The number of such subsets is of course $\binom{n}{s/d} \lesssim (1/\sqrt{s}) \exp(H(s/D)D/d)$. Hence summing over all $s \le D/4$, we get

$$\begin{aligned}\mathbb{P}(\exists S \subset [n], d(S) = s, |E(S;S^c)| \le \alpha s) &\lesssim s^2 \exp(D[\frac{1}{d}H(\frac{s}{D}) + \frac{1}{2}H(\alpha)\frac{s}{D} + \frac{1}{2}H(\frac{2\alpha s}{D}) - \frac{1}{2}H(\frac{s}{D})]) \\ &\le s^2 \exp(D[-\frac{1}{6}H(\frac{s}{D}) + \frac{s}{2D}H(\alpha) + \frac{1}{2}H(\frac{2\alpha s}{D})]),\end{aligned}$$

where we have used $d \ge 3$. Notice that as $x \to 0$, $H(\alpha x)/H(x) \to \alpha$. Hence we can choose $\alpha > 0$ sufficiently small that $H(2\alpha x) \le (1/100)H(x)$ for all $x \le [0,1/4]$.

Hence the right hand side less than

$$\lesssim s^2 \exp(-\frac{D}{12}H(\frac{s}{D})) \lesssim s^2 \exp(-s\log(D/s)).$$

Summing over all $s \le D/4$, we see that

$$\mathbb{P}(\exists S \subset [n], d(S) \le D/4, |E(S;S^c)| \le \alpha d(S)) \to 0$$

for this value of α. This concludes the proof. □

Note that in particular, $t_{\text{mix}} \le C \log n$. It is natural to ask whether cutoff occurs for a given family of expanders. Proving a conjecture of Durrett [14] (which was implicit already in Berestycki and Durrett [6]), Lubetzky and Sly [19] showed:

Theorem 3.6. *Cutoff occurs on a random d-regular graph at time $t_{\text{mix}} = (d/(d-2)) \log_{d-1} n$, with high probability.*

This was later extended by Berestycki, Lubetzky, Peres and Sly [7] to more general random graphs models, including in particular the case of the giant component of an Erdős–Renyi random graph.

Peres conjectured in 2004 that cutoff occurs on *every* expander which is vertex-transitive (i.e., the graph looks the same from every vertex – more formally for every $x, y \in V$ there is a graph isomorphism taking x to y.) For many years there was not a *single example* where this was known to be true (and naturally no counterexample either). Recently, Lubetzky and Peres [18] have shown that cutoff takes place on every Ramanujan graph, which are expander graphs where the spectral gap is as big as it can possibly be (i.e., the best possible expanders). Since Ramaujan graphs are expanders, this provided the first example where this conjecture was verified. Still, the general problem remains, even on any fixed deterministic vertex-transitive expander graph which is not Ramanujan.

xxxx Write about Kesten's theorem.

4 Comparison methods

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.

zzzzz Discussion about how comparison in TV seems very difficult in general.

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 $S^{-1} = S$. Let $p(\cdot)$ be a probability measure on S such that $p(x) > 0$ for all $x \in S$. (We call p the **kernel** of the random walk.) The **(right) random walk on G based on p** is the Markov chain whose transition probabilities are given by the following:

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

In other words, at each step if the walk is at $x \in G$, we choose $s \in S$ with probability $p(s)$, and jump from x to xs. (The corresponding left random walk would jump from x to sx, and note that when p is symmetric, the left and right random walks have exactly the same distribution).

Random walks on groups have 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. We will then call $P^n(x) = P^n(o, x)$ where o is the identity element of G. If p is uniform on S then the Markov chain X can be identified with a random walk on the **Cayley graph** of G induced by S: this is the graph where the vertices are the elements of the group, and there is an edge between x and xy if $y = xs$ for some $s \in S$.

Note that, by symmetry, $\|p^n(x, \cdot) - u\|_{TV}$ does not depend on the starting point $x \in G$; more generally a random walk on a group is an example of a **transitive Markov chain** (and the associated Cayley graph is a transitive graph). In particular, for a random walk on a group we have (see remark 2.1)

$$d_2(t)^2 = \sum_{j=1}^{n} \lambda_j^{2t} \tag{22}$$

where (λ_j) are the eigenvalues of the chain.

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 mount 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,

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

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 $\max_x \|H_t^x - \pi\|_{TV} \to 0$ as $t \to \infty$. In fact π is also the stationary distribution of the discrete time chain associated with P.

We write down the analogue of (22) for the chain in continuous time:

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

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,

$$d_2(t)^2 = \left\| \frac{H_t(x,\cdot)}{\pi(\cdot)} - 1 \right\|_2^2 = \sum_{j=2}^{n} e^{-2t\mu_j} f_j(x)^2.$$

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 $d_2(t)$:

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

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

We deduce

$$4d(t)^2 \le d_2(t)^2 = \sum_{j=2}^{n} e^{-2t\mu_j}. \tag{23}$$

Example. Random transpositions: let $G = S_n$ be the symmetric group with n elements. Consider the generating set S to be the set of transpositions $\tau = (i,j)$ for some $i \neq j$ and the identity, and define the kernel p on S as follows:

$$p(s) = \begin{cases} 1/n^2 & \text{if } s = (i,j) \text{ and } i < j \\ 1/n & \text{if } s = id \text{ is the identity element.} \end{cases} \tag{24}$$

This definition avoids periodicity issues (though in continuous time this isn't relevant); and note that $\sum_{s \in S} p(s) = 1$.

Let $(X_t, t \geq 0)$ be the continuous-time version of this chain. Using techniques from representation theory, Diaconis and Shahshahani (1981) were able to completely analyse this chain and establish the cutoff phenomenon for $d_2(t)$ at time

$$t_{\text{mix}} = \frac{1}{2} n \log n.$$

It is easy to show the lower bound for t_{mix} (before $(1/2)n \log n$ there are too many fixed points.) The difficiult part is the upper bound, and it turns out that after $(1/2)n \log n$ even the ℓ^2 distance is small. More precisely, they showed:

Theorem 4.1 (Diaconis–Shahshahani 1981 [10])**.** *Let $c > 0$. Then there exists a universal $\alpha > 0$ such that $d_2(t) \leq \alpha e^{-c}$ whenever $t \geq (1/2)n(\log n + c)$.*

We will discuss a proof of this result in Section 6.6.

4.3 Comparison techniques

Consider two random walks $P, \tilde{P}$ on a group G. Note that since the chains are translation invariant, both have the uniform distribution π as their invariant measure. We will have in mind a situation where the mixing behaviour of the random walk $\tilde{P}$ is completely understood. It will be important here 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. We first need the following minimiax characterisation of eigenvalues.

Theorem 4.2. *Suppose P is irreducible, aperiodic, and reversible, let λ_j be the eigenvalues ordered in nonincreasing order, so $1 = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Then for $1 \leq j \leq n$*

$$1 - \lambda_j = \max_{\phi_1, \ldots, \phi_{j-1}} \min\{\mathcal{E}(f,f) : \|f\|_2 = 1, f \perp \phi_1, \ldots, \phi_{j-1}\}.$$

Proof. Fix some arbitrary functions $\phi_1, \ldots, \phi_{j-1}$, and let $W = \text{Span}(\phi_1, \ldots, \phi_{j-1})^\perp$. Note that $\dim(W) \geq n - j + 1$. Hence $W \cap \text{Span}(f_1, \ldots, f_j) \neq \emptyset$. Let $g = \sum_{i=1}^j \alpha_i f_i$ be in this intersection, and assume without loss of generality that $\|g\|^2 = \sum_{i=1}^j \alpha_i^2 = 1$. Then

$$\begin{aligned} \mathcal{E}(g,g) &= \langle (I - P)g, g \rangle_\pi \\ &= \langle \sum_{i=1}^j \alpha_i (1 - \lambda_i) f_i, \sum_{i=1}^j \alpha_i f_i \rangle_\pi \\ &= \sum_{i=1}^j \alpha_i^2 (1 - \lambda_i) \leq 1 - \lambda_j. \end{aligned}$$

Therefore, the minimum over *all* functions f in W of the Dirichlet energy must be even smaller:

$$\min\{\mathcal{E}(f,f) : f \in W, \|f\|_2 = 1\} \leq \mathcal{E}(g,g) \leq 1 - \lambda_j.$$

Now taking the maximum over $\phi_1, \ldots, \phi_{j-1}$, we deduce that

$$\max_{\phi_1,\ldots,\phi_{j-1}} \min\{\mathcal{E}(f,f) : \|f\|_2 = 1, f \perp \phi_1, \ldots, \phi_{j-1}\} \leq 1 - \lambda_j.$$

We get equality by setting $\phi_i = f_i$ for $1 \leq i \leq j-1$. □

Corollary 4.1. *Let $P, \tilde{P}$ be two reversible, irreducible, aperiodic Markov chains on a set S with common invraiant distribution π. Let $\mathcal{E}$ and $\tilde{\mathcal{E}}$ be their respective Dirichlet forms and denote by $\lambda_1 \geq \ldots \geq \lambda_n$ and $\tilde{\lambda}_1 \geq \ldots \tilde{\lambda}_n$ their respective eigenvalues. Let $\mu_j = 1 - \lambda_j$ (resp. $\tilde{\mu}_j = 1 - \tilde{\lambda}_j$) be the corresponding gaps. Assume that for some constant $A > 0$, for all $f : G \to \mathbb{R}$,*

$$\tilde{\mathcal{E}}(f,f) \leq A\mathcal{E}(f,f). \tag{25}$$

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 4.2). The point is that the notion of orthogonality depends only on π and not on the chains themselves. Hence for a fixed $j \geq 1$, and for fixed functions $\phi_1, \ldots, \phi_{j-1}$, if $W = \mathrm{Span}(\phi_1, \ldots, \phi_{j-1})$,

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

Taking the maximum over $\phi_1, \ldots, \phi_{j-1}$, we get

$$1 - \lambda_j \leq 1 - \tilde{\lambda}_j,$$

as desired. □

As a consequence of this lemma and (23), we obtain the trivial but crucial relation:

Theorem 4.3. *Assume that $P, \tilde{P}$ are two transitive Markov chains on a set S, and suppose that $\tilde{\mathcal{E}}(f,f) \leq A\mathcal{E}(f,f)$ for all functions $f : S \to \mathbb{R}$. Then if $d_2(t), \tilde{d}_2(t)$ denote the respective ℓ^2 distance to stationarity of their continuous-time versions,*

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

Hence if the ℓ^2 mixing behaviour of $\tilde{P}$ is understood, a comparison inequality such as (25) 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, when we specialise to random walks on a group with two distinct symmetric generating sets S and $\tilde{S}$ such that $S = S^{-1}$ and $\tilde{S} = \tilde{S}^{-1}$, a general comparison can be set up in a way very similar to Theorem 3.2. Fix E a generating set for G (which could in principle be distinct from either set S but we will assume that $E \subset S$). For each $y \in \tilde{S}$, fix a path from o to y using only edges from E, i.e. a path of the form $y = z_1 \cdot \ldots \cdot z_k$, where $z_i \in E$ and let $|y| = k$ be the length of this path. For each generator $z \in E$, let $N(z,y)$ denote the number of times the edge z is used on this path.

Theorem 4.4. *Let $P, \tilde{P}$ be two symmetric random walks on a finite group G with respective symmetric kernels p and $\tilde{p}$ on the sets S and $\tilde{S}$. Let $E \subset S$. Then $\tilde{\mathcal{E}}(f,f) \leq A\mathcal{E}(f,f)$ where*

$$A = \max_{z \in E} \frac{1}{p(z)} \sum_{y \in \tilde{S}} |y| N(z,y) \tilde{p}(y). \tag{26}$$

The proof is similar to the proof of Theorem 3.2, but we include it because it can be confusing nevertheless.

Proof. Note that the Dirichlet form can be expressed as

$$\tilde{\mathcal{E}}(f,f) = \sum_{x \in G, y \in \tilde{S}} \pi(x)\tilde{p}(y)(f(xy) - f(x))^2.$$

Suppose $x \in G, y \in \tilde{S}$, and write $y = z_1 \dots z_k$ where $z_i \in E$. Then

$$f(xy) - f(x) = \sum_{i=1}^{k} f(xz_1 \dots z_i) - f(xz_1 \dots z_{i-1})$$

so that by Cauchy–Schwarz,

$$(f(xy) - f(x))^2 \leq |y| \sum_{i=1}^{k} [f(xz_1 \dots z_i) - f(xz_1 \dots z_{i-1})]^2.$$

Summing in $x \in G$ gives

$$\sum_{x \in G} [f(xy) - f(x)]^2 \leq |y| \sum_{i=1}^{k} \sum_{g \in G} [f(gz_i) - f(g)]^2 \leq |y| \sum_{z \in E} \sum_{g \in G} [f(gz) - f(g)]^2 N(z,y).$$

We multiply by $1/n$ and by $\tilde{p}(y)$, and sum over $y \in \tilde{S}$ to find

$$\begin{aligned}
\tilde{\mathcal{E}}(f,f) &\leq \sum_{z \in E} \sum_{g \in G} \frac{1}{n} [f(gz) - f(g)]^2 \sum_{y \in \tilde{S}} |y| \tilde{p}(y) N(z,y) \\
&\leq \sum_{z \in E} \sum_{g \in G} \frac{1}{n} [f(gz) - f(g)]^2 p(z) \left[\frac{1}{p(z)} \sum_{y \in \tilde{S}} |y| \tilde{p}(y) N(z,y) \right] \\
&\leq A\mathcal{E}(f,f)
\end{aligned}$$

as desired. □

4.4 Example: 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. This is a random walk on S_n where the kernel is uniform on the above set of four generators. 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 4.5. *If $t = 72n^3(\log n + c)$, then for the continuous-time chain, $d_2(t) \le \alpha e^{-c}$.*

Proof. We use a comparison argument, taking our benchmark $\tilde{P}$ to be the random walk on S_n induced by random transpositions, defined by (24). Take $E = S$, and let $y \in \tilde{S}$ be a fixed transposition (say $y = (i, j)$ with $i < j$). We must build y from elementary moves allowed by S.

We first put card with label i on top of the deck by putting the top card at the bottom repeatedly. We then reduce the gap between the cards i and j by repeatedly swapping the top two cards and putting the top card at bottom of deck, until i and j are next to one another (at most $3n$ moves so far). We then switch i and j and reverse all the previous moves, which gets us back to the same deck as the original deck but with i and j swapped. Hence $|y| \le 6n$. Since $N(z, y) \le |y|$ clearly, we deduce using theorem 4.4 that $\tilde{\mathcal{E}}(f, f) \le A\mathcal{E}(f, f)$ with

$$A \le 4 \sum_y |y|^2 \tilde{p}(y) \le 144n^2.$$

Using the Diaconis and Shahshahani result (Theorem 6.9) together with theorem 4.3 finishes the proof. □

4.5 Example: the interchange process

Consider now a fix connected graph $\mathcal{G} = (V, E)$ on n vertices labelled $\{v_1, \dots, v_n\}$. 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 $n = |V|$). In other words the group is $G = \mathcal{S}(V) \simeq S_n$ and the set S of generators consists of the identity along with the transpositions (v_i, v_j) for every pair of neighbouring vertices v_i, v_j. The kernel p is defined to be $p(id) = 1/n$, and $p((v_i, v_j)) = (1 - 1/n)(1/|E|)$.

Note that this set of generators is symmetric and the kernel is symmetric. Moreover since $\mathcal{G}$ is connected every transposition (v, w) can be built from neighbouring transpositions, and hence the set S generates all of $\mathcal{S}(V)$.

The case of random transpositions corresponds then corresponds to $\mathcal{G} = K_n$, the complete graph on n vertices. 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 on $\mathcal{G}$ from x to y, and set

$$\Delta = \max_{x,y} |\gamma_{xy}|,$$

(where $|\gamma_{xy}|$ is the length of the path γ_{xy}, 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 4.6. *The comparison $\tilde{\mathcal{E}} \leq A\mathcal{E}$ holds with*

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

As a result, if $t = (A/2)n(\log n + c)$, $d_2(t) \leq \alpha e^{-c}$ for some universal constant $\alpha > 0$.

Proof. We apply Theorem 4.4 again, with $\tilde{p}$ being the kernel of random transpositions. 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 $x, y \in V$ are arbitrary vertices (not necessarily neighbours), the transposition $\tau = (x, y)$ can be constructed by first transposing successively all the edges in γ_{xy} (here and below we identify an edge $e = (u, v)$ with the transposition $(u, v) \in \mathcal{S}(V)$) and then reversing these transpositions except for the last one. This gives path $\tilde{\gamma}(\tau)$ on $G = \mathcal{S}(V)$ of length $|\tau| = |\tilde{\gamma}(\tau)| \leq 2|\gamma_{xy}| \leq 2\Delta$. Moreover no transposition is used more than twice, hence for any $\tau \in \tilde{S}$, and any edge e, $N(e, \tau) \leq 2 \cdot \mathbf{1}_{\{e \in \tilde{\gamma}(\tau)\}}$. Therefore, by Theorem 4.4, we may take

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

as $\sum_{\tau \in \tilde{S}} \mathbf{1}_{\{e \in \tilde{\gamma}(\tau)\}} \leq K$, by definition of K. Applying the result of Diaconis and Shahshahani (Theorem 6.9) on random transpositions finishes the proof. □

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 forces. Clearly $\Delta = n - 1$, $|E| = n - 1$, $K \leq 2(n/2)^2$. Thus Theorem 4.6 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 $\Delta = 2$, $|E| = n - 1$, $K = n - 1$, hence $A \leq 16$. Thus for $t = 8n(\log n + c)$, we have $d(t) \leq \alpha e^{-c}$.

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 [17] (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 $\sigma^2 = \frac{1}{2}(\mathrm{var}_\mu(f) + \mathrm{var}_\nu(f))$. 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 ν is also large. We now state the result which is the basis of Wilson's method (see Theorem 13.5 in [17]).

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

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

Proof. The proof is directly taken from [17]. Since Φ is an eigenfunction of P, we immediately get

$$\mathbb{E}_x(\Phi(X_t)) = \lambda^t \Phi(x). \tag{27}$$

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 substract $R/(2(1-\lambda))$ 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

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

Using (27), this gives us:

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

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

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

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

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

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 greater that 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.

We already know the eigenvalues of the chain: if J is a subset of $\{1,\dots,n\}$, and we write the hypercube H_n as $H_n = \{-1,+1\}^n$, then

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

is an eigenfunction of the lazy random walk on H_n, 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

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

Now, consider Wilson's method. We wish to take Φ an eigenfunction associated with the second largest eigenvalue, i.e., with the eigenvalue associated with the spectral gap. 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 $d(t) \to 1$. On the other hand, if

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

then $d(t) \to 0$.

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:

$$\Phi(\sigma) = \sum_{1 \leq k \leq n} \varphi(k)\varphi(\sigma(k)) \tag{28}$$

which is also an eigenfunction of the chain with eigenvalue λ. When $\sigma = \mathrm{id}$ 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 (28) are necessarily maximised at $\sigma = \mathrm{id}$). 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 (28) change. Thus

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

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

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

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

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

as claimed for the lower-bound.

Upper-bound by coupling. The following coupling was introduced by Aldous, but we follow the presentation in [17], 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

$$\text{if } t = 2n^3 \log_2 n \text{ then } d(t) \to 0. \tag{29}$$

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 $1 \le a \le n$. We wish to put card a at the same position in both decks. (We will later maximise over all possible $1 \le a \le n$.) The coupling is the following. Choose a position $1 \le i \le n-1$ 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 $\sigma_t(i+1) = \sigma(t)$ 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:

$$\mathbb{P}(\tau_a > n^3) \leq \frac{1}{2}. \tag{30}$$

Suppose we run the chain for blocks of time of duration n^3 each, and we run $2\log_2 n$ such blocks. Since (30) is independent of the starting point, the probability that $\tau_a > 2\log_2 nn^3$ 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

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

as claimed.

6 Representation theoretic methods

6.1 Basic definitions and results of representation theory

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

The purpose of representation theory is to describe a given group G by asking the reverse question: what structures does G act on? 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 6.1. *A group representation of G is a map $\rho : G \to GL(V)$, 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, $\rho(e) = Id$.

In short, a representation is simply an embedding of the group into some linear group $GL(V)$, where the group multiplication and inverse correspond to the matrix multiplication and inverse. The dimension of V is called d_ρ, the dimension of ρ.

Example 6.1. *A trivial (one-dimensional) representation is one for which $\rho(s)v = v$ for all $v \in V$ and all $s \in G$. When $G = S_n$ is the symmetric group, another less trivial (but still one-dimensional) representation is the sign representation: $\rho(s)v = \mathrm{sgn}(s)v$.*

Example 6.2. *When $G = S_n$ an interesting representation is the* permutation *representation. This is a representation ρ in an vector space V of dimension n defined as follows. For $s \in S_n$, the linear map $\rho(s)$ is defined by setting $\rho(s)(e_i) = e_{s(i)}$, where $(e_1, \ldots, e_n)$ is a basis of V.*

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 6.1. *Let $\rho : G \to GL(V)$ 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 $(,)$ 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 $\langle , \rangle$ 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. □

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

Theorem 6.1. *Every representation ρ is the finite sum of irreducible representations.*

Example 6.3. *The permutations representation above is not irreducible. Indeed, observe that the vector $v = (1, \dots, 1)$ is invariant under $\rho(s)$ for any $s \in S_n$. Hence if $W_1 = Span(v)$, W_1 is stable. It follows that there is a complement W_2 such that W_2 is also invariant. This can be described as the subspace $\{v \in V : \sum_i v_i = 0\}$. The induced representation ρ_1 is the trivial one-dimensional representation. The induced representation ρ_2 on W_2 is called the $(n-1)$-dimensional representation. It will be shown in Example 6.5 that ρ_2 is in fact irreducible; this will play an important role in the theory of random transpositions.*

Remark 6.1. *Since the scalar product $\langle , \rangle$ 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.*

6.2 Characters

Ultimately, our main goal will be to use representations of G to do some Fourier analysis. To do this, our starting point is to find a natural collection of elementary functions $(e_i)_{i \in I}$ such that any function $f : G \to \mathbb{R}$ or $f : G \to \mathbb{C}$ (in fact, any function f subject to a certain symmetry condition, see Theorem 6.7) can be represented as a linear combination $\sum_i \alpha_i e_i$, where the coefficients α_i can be computed simply. The elementary functions e_i are called characters and they are defined in terms of representations.

Definition 6.2. *Let ρ be a representation of G. The character associated to ρ is the function $\chi_\rho : G \to \mathbb{R}$ defined by $\chi_\rho(s) = \mathrm{Tr}(\rho(s))$.*

Example 6.4. *In the permutation representation ρ of Example 6.3, the character χ_ρ is the function giving the number of fixed points of a permutation $s \in S_n$.*

The following properties are trivial but worth keeping in mind:

Proposition 6.2. *(i) $\chi_\rho(e) = d_\rho$, (ii) $\chi_\rho(s^{-1}) = \overline{\chi_\rho(s)}$, and (iii) the characters are invariant by conjugacy: $\chi_\rho(t^{-1}st) = \chi_\rho(s)$.*

Proof. (i) just follows from the fact $\rho(e) = Id$ always,

(ii) In the basis where $\rho(s)$ is unitary, we have that $\rho(s^{-1}) = \rho(s)^{-1} = \rho(s)^*$ thus, taking the trace, $\chi_\rho(s^{-1}) = \overline{\chi_\rho(s)}$.

(iii) just comes from the fact that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$. □

The decomposition of any function as a linear combination of characters will come from the fundamental observation (below) that these form an orthonormal basis for the natural scalar product. Recall the usual scalar product on functions from $G \to \mathbb{C}$:

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

Theorem 6.2. *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 below 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 $f : V \to V'$ (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 6.1. *Let ρ, ρ' be two irreducible representations into V, V' respectively. Let $f : V \to V'$ 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 $f = \lambda I$ for some $\lambda \in \mathbb{C}$.

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 $\rho(s) \in \ker f$ as well. Likewise, the image of f, Imf 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 Im$f = V'$, so f is an isomorphism. For (b), let λ be an eigenvalue of f. Then the map $\tilde{f} = f - \lambda I$ has a nontrivial kernel and satisfies $\tilde{f}\rho = \rho'\tilde{f}$. Thus by the above $\tilde{f} = 0$ i.e. $f = \lambda I$. □

It is the following corollary which we use here:

Corollary 6.1. *Let $h : V \to V'$ be linear. Define*

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

Then

(a) if ρ, ρ' are not equivalent then $\tilde{h} = 0$.

(b) If $V = V'$ and $\rho = \rho'$ then we have $\tilde{h} = \lambda I$ with $\lambda = \mathrm{Tr}(h)/d_\rho$.

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. □

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 6.3. *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 $V = W_1 \oplus \ldots \oplus W_m$. Then we have $\chi_\rho = \sum_i \chi_i$ where χ_i is the character or W_i. Hence $(\chi_\rho|\chi_\psi) = \sum_{i=1}^m (\chi_i|\chi_\psi)$. 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. □

Corollary 6.2. *Two representations are equivalent if and only they have the same character.*

Corollary 6.3. *Let ρ be a representation. Then $(\chi_\rho|\chi_\rho)$ is a positive integer, equal to 1 if and only if ρ is irreducible.*

Example 6.5. *Let us check that the $(n-1)$ dimensional representation of S_n, introduced in Example 6.3 is irreducible. Recall the permutation representation ρ. We have found two induced representation ρ_1 and ρ_2. It therefore suffices to show that $(\chi_\rho|\chi_\rho) = 2$. Recall that $\chi_\rho(s)$ is just the number of fixed points of $s \in S_n$, hence $(\chi_\rho|\chi_\rho) = \mathbb{E}(X^2)$, where X is the number of fixed points of a randomly chosen permutation $\sigma \in S_n$. The result then follows from the fact that*

$$\begin{aligned}\mathbb{E}(X^2) &= \sum_{i=1}^n \mathbb{P}(\sigma_i = i) + \sum_{i \neq j} \mathbb{P}(\sigma_i = i; \sigma_j = j) \\ &= n \times \frac{1}{n} + n(n-1) \times \frac{1}{n} \times \frac{1}{n-1} \\ &= 2.\end{aligned}$$

Indeed σ_i is uniformly distributed on $\{1, \ldots n\}$ and given $\sigma_i = i$, $|sigma_j$ is uniformly on $\{1, \ldots, n\} \setminus \{i\}$. Thus ρ contains two irreducible representations, and hence ρ_1 and ρ_2 are both 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 6.4. *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 $(\chi_\rho|\chi_\psi) = \frac{1}{|G|}\sum_s \chi_\psi(s)\bar{\chi}(s) = \chi_\psi(e) = d_\psi$. □

Corollary 6.4. *We have $\sum_\rho d_\rho^2 = |G|$, where $\sum_\rho$ 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 $\chi(s) = 0$.

Remark 6.2. *The identity $\sum d_\rho^2 = |G|$ suggests that in general there is a natural probability distribution on irreducible representations of a group G, given by $p(\rho) = d_\rho^2/|G|$. When $G = S_n$, this is known as the Plancherel measure, a subject of great interest in random matrix theory, completely integrable systems etc.*

6.3 Fourier inversion

Let $f : G \to \mathbb{C}$ 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). To summarise, and to make the link with classical Fourier analysis, the particular way to embed G into a linear group $GL(V)$ (an irreducible representation ρ) is a "frequency". Then given a function f on G, the "amplitude" of that frequency $\hat{f}(\rho)$ is the sum (in $GL(V)$) of all the group elements, weighted by the function f. (However one has to bear in mind that this "amplitude" is a matrix. In fact it would be more appropriate to say that the amplitude is $d_\rho \hat{f}(\rho)$.)

The Fourier inversion theorem says that it is possible to reconstruct entirely any function f from its Fourier transform as follows:

Theorem 6.5. *We have the following identity: for all $s \in G$,*

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

where the sum is over all irreducible representations.

This is the analogue of the classical Fourier inversion theorem – which, in its discrete form, is just a particular case of this result with $G = \mathbb{Z}/n\mathbb{Z}$. Indeed in this case, the representations are all one-dimensional (as in any Abelian group) and so each representation ρ_j is determined by a complex number $\rho_j(x)$ for each $x \in \mathbb{Z}/n\mathbb{Z}$ which respect the group structure of G; hence for a given representation ρ one has $\rho(x+y) = \rho(x)\rho(y)$. Hence $\rho = \rho_j$ for some frequency $1 \le j \le n$, where $\rho_j(x) = e^{2\mathbf{i}\pi jx/n}$. The Fourier transform of a function $f : \mathbb{Z}/n\mathbb{Z} \to \mathbb{C}$, is computed as $\hat{f}(j) = \sum_x f(x)e^{2\mathbf{i}\pi jx/n}$. The Fourier inversion theorem states that

$$f(x) = \frac{1}{n} \sum_{j=0}^{n-1} \hat{f}(j)e^{-2\mathbf{i}\pi jx/n},$$

which is indeed the statement we are familiar with in classical Fourier analysis.

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 6.4. □

Theorem 6.6. *Let $f, g : G \to \mathbb{C}$ 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. □

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

$$\sum_s f(s)h(s) = \frac{1}{|G|}\sum_\rho \operatorname{Tr}(\hat{f}(\rho)\hat{g}(\rho)^*). \tag{31}$$

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 $s \in G$.

6.4 Class functions

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 $c_1 \ldots \cdot c_m$, and σ is a permutation, then $\sigma\pi\sigma^{-1}$ is the permutation having cycle distribution equal to $\sigma(c_1)\ldots\sigma(c_m)$, where if $c = (x_1, \ldots, x_k)$ we denote by $\sigma(c)$ the cycle $(\sigma(x_1), \ldots, \sigma(x_k))$. 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 6.2. *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,

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

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

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

Theorem 6.7. *The characters form an orthonormal basis of the space of class functions.*

Proof. Note that the characters are themselves class functions since $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$. 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 $(f|\bar{\chi}_\rho) = 0$ 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 $f = 0$. □

6.5 Diaconis–Shahshahani lemma

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

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

Then it is straightforward that

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

Thus 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.

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

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

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 (31) to this we get

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

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. [change argument] □

The following corollary makes explicit what we learn for the total variation distance in the case of a random walk on G, and is known in the literature as the Diaconis–Shahshahani upper bound lemma.

Corollary 6.5. *Suppose that P be a random walk kernel which is a class function. Then for all $t \ge 1$, we have, if $d(t) := \|P^{*t} - \pi\|_{TV}$, we have:*

$$d(t)^2 \le \frac{1}{4} d_2(t)^2 = \frac{1}{4} \sum_* d_\rho^2 |\lambda_\rho|^{2t}$$

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

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

Note that when P is symmetric, i.e., $P(s) = P(s^{-1})$, then λ_ρ is real, as $\chi_\rho(s) = \overline{\chi_\rho(s^{-1})}$.

Originally this was discovered by the authors in the context of their investigation of random transpositions, but has since been used in a very wide variety of examples.

6.6 Example: random transpositions

We will now discuss and sketch a proof of the following fundamental theorem, due to Diaconis and Shahshahani.

Theorem 6.9. *(Diaconis-Shahshahani [10]) Let $c > 0$. Then there exists a universal $\alpha > 0$ such that $\tilde{d}_2(t) \le \alpha e^{-c}$ whenever $t \ge (1/2)(n \log + cn)$.*

Conversely, for all $\varepsilon > 0$ there is a $c > 0$ such that $d(t) \ge 1 - \varepsilon$ for $t \le (1/2)(n \log n - cn)$.

Sketch of proof. We start by the *upper bound* (which is harder than the lower bound).

We apply Corollary 6.5. Note that $\lambda_\rho = \frac{1}{n} + \frac{n-1}{n} r(\rho)$, 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 (it does not matter which since characters are class functions). Hence

$$d(t)^2 \le \frac{1}{4} \sum_* d_\rho^2 \left(\frac{1}{n} + \frac{n-1}{n} r(\rho) \right)^{2t}. \tag{32}$$

We start by considering the $(n-1)$-dimensional representation of Example 6.3. For this we have $d_\rho = n - 1$ and it can be seen that $\chi_\rho(\tau) = n - 3$. Indeed for the permutation representation the character is the number of fixed points, which is $n - 2$. Since this is the sum of the characters of the trivial and the $(n-1)$-dimensional representation, we deduce $\chi_\rho(\tau) = n - 3$ as claimed. Thus $r(\rho) = (n-3)/(n-1)$ and hence the contribution of this representation to the sum in (32) is

$$(n-1)^2 \left(\frac{n-2}{n} \right)^{2t} \le \exp(2 \log n - 4t/n).$$

For $t = (1/2)(n \log n + cn)$ we see that this is less than e^{-c}. The bulk of the proof of the upper bound consists in showing that for this time t, the sum of contributions for *all other* irreducible representations in (32) is negligible. The main reason why this holds is that for "most" representations, the character ratio $r(\rho)$ is bounded away from 1. Suppose for instance it was the case that $r(\rho) \le r < 1/e$. Then

$$\sum_{*} d_\rho^2 \left(\frac{1}{n} + \frac{n-1}{n} r(\rho) \right)^{2t} \le r^{2t} \sum_{*} d_\rho^2 = r^{2t} n!$$

by Corollary 6.4. Now, observe that for $t \ge (1/2) n \log n$ we have

$$\begin{aligned} r^{2t} n! &\le \exp(2t \log r + n \log n) \\ &\le \exp(n \log n (1 + \log r)) \to 0. \end{aligned}$$

This is a gross simplification of the argument, but helps to explain the gist. To make things rigorous requires an exact formula for the character of a given irreducible representation ρ. The irreducible representations ρ of S_n can be indexed by Young diagrams $\lambda = (\lambda_1, \ldots, \lambda_k)$, that is partitions of n (hence $\lambda_1 + \ldots + \lambda_k = n$, and we write them in decreasing order). A formula due to Frobenius then gives, for the corresponding irreducible representation ρ,

$$r(\rho) = \frac{1}{n(n-1)} \sum_j \lambda_j^2 - (2j-1)\lambda_j. \tag{33}$$

Likewise it is well known in representation theory how to compute the dimension d_ρ of the associated representation. If we think of the Young diagram as a stack of boxes on top of one another, then d_ρ is the number of ways to fill the boxes with labels $1, \ldots, n$ in such a way that the labels are always increasing from left to right and from top to bottom. Hence $d_\rho \le \binom{n}{\lambda_1} d_{(\lambda_2, \ldots, \lambda_n)}$. The desired upper bound follows, with plenty of careful estimates. See [10, Chapter 3D] for an eminently readable account of the proof.

Lower bound. We now check that for $t \le (1/2)(n \log n - cn)$, $d(t) \ge 1 - \varepsilon$. We are able to find an explicit event such that $\mathbb{P}(X_t \in A) \ge 1 - \varepsilon$ but $\mathbb{P}(\sigma \in A) \le \varepsilon$ for a uniform random permutation σ. This event A is given by

$$A = \{s \in S_n : s \text{ has more than } K \text{ fixed points}\}$$

where K is arbitrarily large. Observe that if a card i has never been touched up to time t then it is a fixed point of the permutation X_t. Since we are collecting two cards at a time, the coupon collector problem tells us that for $t = (1/2)(n \log n - cn)$, $\mathbb{P}(A) \ge 1 - \varepsilon$ by choosing c sufficiently large. But of course $\mathbb{P}(\sigma \in A) \le \varepsilon$ if K is large enough.

Intringuingly, this simple probabilistic lower bound has a representation theoretic counterpart. Consider the character χ_ρ of the $(n-1)$-dimensional representation – which is the one whose contribution to the sum (32) can be designed as the culprit for cutoff phenomenon. As we have seen earlier, the character of the permutation representation counts the number of fixed points, so χ_ρ counts the number of fixed points minus one (indeed the permutation representation is the direct sum of the trivial representation and the $(n-1)$-dimensional one). Now, if X is uniformly distributed over S_n,

$$\mathbb{E}(\chi_\rho(X)) = (\chi_\rho | 1) = 0; \text{ and } \operatorname{var}(\chi_\rho(X)) = \mathbb{E}(\chi_\rho^2(X)) = (\chi_\rho | \chi_\rho) = 1.$$

(This could also be deduced from the fact that the number of fixed points is, at least for n large, approximately a Poisson random variable with mean 1). But under the random walk measure,

$$\begin{aligned}\mathbb{E}(\chi_\rho(X_t)) &= \sum_g P^t(g)\,\mathrm{Tr}(\rho(g))\\ &= \mathrm{Tr}(\sum_g P^t(g)\rho(g))\\ &= \mathrm{Tr}(\widehat{P^{*t}}(\rho))\\ &= \mathrm{Tr}[(\hat{P}(\rho)^t]\\ &= d_\rho \lambda_\rho^t\end{aligned}$$

where $\lambda_\rho = (1/n + (n-1)/nr(\rho)) = (n-2)/n$, as before. Again, one finds that for $t \leq (1/2)(n\log n - cn)$,

$$\mathbb{E}(\chi_\rho(X_t) \geq K$$

where K can be made arbitrarily large if c is sufficiently large. This is not quite enough to conclude that X_t looks very different from the uniform distribution (in terms of total variation): the second moment method is needed to show that $\chi_\rho(X_t)$ is in fact large with high probability. This can be done by computing $\mathbb{E}(\chi_\rho(X_t)^2)$. To do this, we may observe that χ^2 is also a character: this is the character or the representation $\rho \otimes \rho$ (where $\otimes$ denotes the tensor product). The explicit decomposition of $\rho\otimes\rho$ in terms of irreducible representation is not hard to find, and involves just three irreducible nontrivial representations. Computing $\mathrm{var}(\chi_\rho(X_t))$ can thus be done by following the above steps. □

6.7 A conjecture on cutoff

To some extent, in "finite dimensions" we expect Markov chains typically to have scaling limits. This implies that we should expect (in finite dimensions) that the typical situations is one where the convergence to equilibrium happens gradually, on the time scale of its diffusive scaling limit. This suggests that we should only expect a cutoff phenomenon in "high dimensions". I propose the following conjecture (which is possibly slightly optimistic).

Conjecture 6.1. *Suppose G_n is a sequence of groups, and suppose that we consider a random walk on the Cayley graph of G_n induced by a set of generators C_n which is invariant under conjugacy.*

Let ρ_n be the lowest dimensional representation of G_n which which is nontrivial, and let $d_n = \dim(\rho_n)$. Suppose that $d_n \to \infty$. Then the cutoff phenomenon occurs (no matter what set of generators is chosen, so long as it is conjugacy invariant), and the cutoff time is given by

$$t_{\mathrm{mix}} = \max_\rho \frac{\log d_\rho}{-\log r(\rho)}$$

where $r(\rho) = \chi_\rho(c)/d_\rho$ is the character ratio of a representation ρ evaluated at any element $c \in C_n$ of the conjugacy class generating the random walk.

Note that typically the maximum of this ratio is attained at the lowest nontrivial representation, but this isn't always the case [CITATION NEEDED.]

7 Other notions of mixing

7.1 Strong stationary times and separation distance

In our first example of cutoff, (random to top) we saw an example of a coupling strategy obtained by consider the first time τ that all cards were touched. At this stopping time, the distribution of the deck of cards is not just approximately stationary, but exactly! Moreover the distribution of the deck of cards is independent of the time τ. Such a time is called a **strong stationary time**.

It turns out that many coupling arguments have this property, which was first investigated in a celebrated paper of Aldous and Diaconis [3]. Just the same way as total variation is intricately linked to the notion of coupling, it turns out that such strong stationary times measure convergence in a stronger metric than total variation, which is the *separation distance.* We start by making the following definitions.

Definition 7.1. *A* ***filtration*** $(\mathcal{F}_n)_{n\ge 0}$ *is an increasing sequence of σ-algebras. A Markov chain (X_n) is* ***adapted*** *to the filtration if X_n is $\mathcal{F}_n$-measurable for every $n \ge 0$ (i.e., all the information about $(X_1, \ldots, X_n)$ is contained in $\mathcal{F}_n$. We say that $(X_n, n \ge 0)$ is a Markov chain* ***with respect to*** $\mathcal{F}$ *if $\mathbb{P}(X_{n+1} = y|\mathcal{F}_n) = P(X_n, y)$ almost surely, where P is the transition matrix.*

A (randomised) ***stopping time*** *τ for the Markov chain X is a random variable τ with values in $\{0, 1, \} \cup \{\infty\}$ such that for some filtration $\mathcal{F}$, with respect to which X is a Markov chain, τ is an $\mathcal{F}$-stopping time: $\{\tau \le n\} \in \mathcal{F}_n$.*

Adapted filtrations contain the information generated by the Markov chain, and typically contain a bit more randomness that does not ruin the Markov property: for instance, they might contain the information generated by the chain and some extra independent coin toss.

In other words, a stopping time for the chain X is a random time such that the event $\{\tau \le n\}$ depends only on $(X_1, \ldots, X_n)$ and possibly some independent random variable Y. It is therefore a way to stop the chain at a time which depends on the chain and possibly on some extra independent randomness. Sometimes people write randomised stopping times to emphasise the fact that some extra randomness is allowed.

Definition 7.2. *A strong stationary time for the chain X is a randomised stopping time such that*

$$\mathbb{P}_x(X_\tau = y; \tau = n) = \pi(y)\mathbb{P}(\tau = n).$$

for all $n \ge 0$ and for all $y \in S$, where π is the invariant distribution of the chain. That is, X_τ has the stationary distribution and is independent of τ.

Note that we could also define the notion of a strong stationary time started from x for a fixed starting state $x \in S$, in which case the above is only required to hold for all $n \ge 0$ and for all $y \in S$. Then τ itself may depend on the starting state x.

Example 7.1. *Random to top: show that the time all cards have been touched is a strong stationary time.*

Here are three examples, the proofs of which are left as exercises.

Example 7.2. *Top to random: Let τ be the first time such that the card which was originally at the bottom is at the top. Show that $\tau + 1$ is a strong stationary time. (In fact, these two examples are related: how?)*

Example 7.3. *On the n-cycle. The cover time (the first time that all vertices have been visited) is a strong stationary time.*

After a strong stationary time, the Markov chain retains the stationary distribution:

Lemma 7.1. *If τ is a strong stationary time (starting from $x \in S$), we have*

$$\mathbb{P}_x(X_t = y; \tau \leq t) = \pi(y)\mathbb{P}_x(\tau \leq t).$$

Proof. We can sum over all times $s \leq t$ and positions z of the chain at time s:

$$\begin{aligned}
\mathbb{P}_x(X_t = y; \tau \leq t) &= \sum_{s=0}^{t} \sum_{z \in S} \mathbb{P}_x(X_t = y, X_s = z, \tau = s) \\
&= \sum_{s=0}^{t} \sum_{z \in S} \mathbb{E}_x[\mathbb{P}_x(X_t = y, X_s = z, \tau = s | \mathcal{F}_s)] \\
&= \sum_{s=0}^{t} \mathbb{P}_x(\tau = s, X_s = z) P^{t-s}(z, y) \\
&= \sum_{s=0}^{t} \sum_{z \in S} \pi(z) P^{t-s}(z, y) \mathbb{P}_x(\tau = s) \\
&= \pi(y)\mathbb{P}_x(\tau \leq t)
\end{aligned}$$

as desired, where we have used the invariance of π to say that $\sum_z \pi(z)P^{t-s}(z,y) = \pi(y)$. □

We now define the separation distance.

Definition 7.3. *If μ, ν are two probability measures on S, we let*

$$\operatorname{sep}(\mu, \nu) = \sup_{x \in S}(1 - \frac{\mu(x)}{\nu(x)}).$$

Note the absence of absolute value in the definition of the separation distance, which may be surprising at first. In particular, it is not symmetric and hence not a distance in the usual sense of the word. Suppose ν is the uniform distribution over S (as will often be the case in our applications). Intuitively, the separation distance is small if even the least likely x for μ is pretty close to $\nu(x)$. More precisely, we will have $\operatorname{sep}(\mu, \nu) \leq \varepsilon$ if $\mu(x) \geq (1 - \varepsilon)\nu(x)$ for *every* x, even the least likely x.

In contrast, the total variation distance requires $\mu(x)$ and $\nu(x)$ to be close in an L^1 sense. This is clearly weaker, and we obtain the following simple but important result:

Theorem 7.1. *We have $\|\mu - \nu\| \leq \operatorname{sep}(\mu, \nu)$.*

Proof. Recall that one expression for $\|\mu - \nu\|$ is $\sum_x(\nu(x) - \mu(x))_+$. Let $\delta = \operatorname{sep}(\mu, \nu)$. Then $\mu(x) \geq (1 - \delta)\nu(x)$, hence $(\nu(x) - \mu(x))_+ \leq \delta\nu(x)$. Summing over x gives us the result. □

We now address the relation between strong stationary times and separation distance.

Theorem 7.2. *Let $s(t) = \sup_x \operatorname{sep}(P^t(x, \cdot), \pi(\cdot))$. If τ is any strong stationary time, then $s(t) \leq \mathbb{P}(\tau > t)$. Conversely, there exists a strong stationary time τ such that $s(t) = \mathbb{P}(\tau > t)$.*

Proof. We only prove the upper bound, which is a trivial consequence of Lemma 7.1. Indeed let $\delta = \mathbb{P}(\tau > t)$. Then we have

$$\mathbb{P}_x(X_t = y) \geq \mathbb{P}(\tau \leq t)\pi(y)$$

so that

$$\frac{P^t(x,y)}{\pi(y)} \geq (1-\delta)$$

for all $x, y \in S$, meaning that $\text{sep}(P^t(x,\cdot), \pi(\cdot)) \leq \delta$ for all x. Hence $s(t) \leq \delta$ as desired. □

While the separation distance is stronger than total variation distance, it is not much stronger. In fact the mixing times can differ by a factor of at most two in the reversible case, because of the following result.

Theorem 7.3. *For reversible chains we have* $s(2t) \leq 1 - (1 - 2d(t))^2$.

Proof. By the Markov property and reversibility,

$$\begin{aligned}
\frac{P^{2t}(x,y)}{\pi(y)} &= \sum_z \frac{P^t(x,z)P^t(z,y)}{\pi(y)} \\
&= \sum_z \frac{P^t(x,z)P^t(y,z)}{\pi(z)} = \sum_z \pi(z)\frac{P^t(x,z)P^t(y,z)}{\pi(z)^2} \\
&\geq \left(\sum_z \pi(z)\frac{P^t(x,z)^{1/2}P^t(y,z)^{1/2}}{\pi(z)}\right)^2 \qquad \text{by Jensen, since } \mathbb{E}(Z) \geq \mathbb{E}(Z^{1/2})^2 \\
&\geq \left(\sum_z \min(P^t(x,z), P^t(y,z))\right)^2 \\
&= \left(1 - \|P^t(x,\cdot) - P^t(y,\cdot)\|\right)^2 \qquad \text{using the fact that } \min(a,b) = b - (b-a)_+ \\
&\geq (1 - \bar{d}(t))^2 \leq (1 - 2d(t))^2.
\end{aligned}$$

The result follows immediately. □

7.2 Example: separation cutoff on the hypercube

For lazy random walk on the hypercube H_n we already know that cutoff in the total variation sense takes place at time $t_{\text{mix}} = (1/2)n\log n$. Here we show that in separation distance this takes place at time $t_{\text{sep}} = n \log n$. (Note that this is indeed within a factor of 2 from t_{mix}, as per Theorem 7.3.

First we consider the upper bound. For this we consider the first time τ where all coordinates have been touched. It is easy to check that τ is a strong stationary time. Since $\tau \leq (1+\varepsilon)n\log n$ with high probability, we have by Theorem 7.2 that $\text{sep}((1+\varepsilon)n\log n) \to 0$, which gives the upper bound.

For the lower bound, we point out that at any time t, intuitively the least likely point on the hypercube for X_t is the opposite corner $y = 1 \ldots 1$, since this requires all coordinates to have been touched. (Indeed it is easy to check, by observing that every touched coordinate is random). Hence

$$\text{sep}(t) \geq 1 - \frac{P^t(x,y)}{\pi(y)} = 1 - \frac{\mathbb{P}(\tau \leq t)\pi(y)}{\pi(y)} = \mathbb{P}(\tau > t)$$

Hence if $t = (1-\varepsilon)n\log n$, $\tau > t$ with high probability and $\mathrm{sep}(t) \to 1$, as desired.

7.3 Lovász–Winkler optimality criterion

Here we focus our attention to stationary times (which are not necessarily strong stationary times) and a beautiful optimality criterion due to Lovász and Winkler. Here optimal refer to the fact that that the expected value of τ is minimal among all stopping times such that X_τ has the desired stationary distribution π. For this we need the notion of a halting state.

We will consider **stationary times**, which are stopping times τ having the property that $\mathbb{P}(X_\tau = y) = \pi(y)$ (but note that X_τ is not assumed to be independent of τ, so τ is not necessarily a strong stationary time). We will look to characterise "best" or "optimal" stationary times. First, let us now define precisely the notion of optimality we wish to consider.

Definition 7.4. *For a given starting state x, or more generally for a given starting distribution μ, we say that the stationary time τ is* mean-optimal *or simply* optimal *for the starting distribution μ on S if $X_\tau = \pi$ in distribution and $\mathbb{E}_\mu(\tau)$ is minimal among all stationary times.*

The following beautiful optimality criterion for stationary times is due to Lovász and Winkler, and relies on the notion of a **halting state**.

Definition 7.5. *Let τ be a (randomised) stopping time and suppose $X_0 = x \in S$. We say that τ has a halting state $y \in S$, if $X_t = y$ implies $\tau \le t$. In other words, y is a halting state if $\tau \le T_y$ with probability one, where T_y is the hitting time of y.*

Example 7.4. *Consider the lazy random walk on the hypercube H_n. The first time τ that all coordinates have been touched is a strong stationary time. Moreover, the state $y = 11\dots1$ is a halting state for τ and the starting state $x = 00\dots0$. Indeed, if $X_t = y$ then all coordinates must have been touched so $\tau \le t$.*

Theorem 7.4. *A stationary time τ is optimal for a starting distribution μ if and only if it has a halting state.*

For instance, for the lazy random walk the hypercube, the strong stationary time τ where all coordinates have been touched has a halting state (namely, the opposite corner $1\dots1$) and hence is mean-optimal.

Proof. We start with *if* part (if there is a halting state then the stationary time is optimal). Let V_x denote the number of times the chain leaves vertex $x \in S$ before being stopped at time τ, so $V_x = \sum_{t=0}^{\tau-1} \mathbf{1}_{\{X_t = x\}}$. (Note that since there is a halting state, τ is finite a.s. and in fact $\mathbb{E}(\tau) < \infty$).) Let $v_x = \mathbb{E}(V_x) < \infty$ (sometimes called the exit frequencies). So $\mathbb{E}(\tau) = \sum_x v_x$. We claim that for every $y \in S$,

$$\sum_x p_{xy} v_x = v_y + \pi_y - \mu_y. \tag{34}$$

The reason is that the number of times that the chain leaves y is equal the number of times the chain enters y, except for the initial and final visits (recall that μ is the starting distribution). Now let τ' be another stationary time, and we can assume for the proof of optimality that

$\mathbb{E}(\tau') < \infty$ as well (since at least τ has finite expectation). Let v'_x denote the corresponding exit frequencies. Then from (34) we see that

$$\sum_x p_{xy}(v'_x - v_x) = v'_y - v_y.$$

Since the chain is irreducible, invariant measures are unique up to constants. It follows that for some $\alpha \in \mathbb{R}$,

$$v'_x - v_x = \alpha\pi_x \tag{35}$$

for all $x \in S$. Summing over x, we get $\alpha = \mathbb{E}(\tau') - \mathbb{E}(\tau)$. Now let y be a halting state for τ (so far we have not really used that assumption except to say $\mathbb{E}(\tau) < \infty$). Then $v_y = 0$. Since $v'_y \geq 0$ we deduce from (35) that $\alpha = \mathbb{E}(\tau') - \mathbb{E}(\tau) \geq 0$, so τ is optimal.

We now examine the converse, the *only if* part (if a stationary time is optimal then it must have a halting state). Suppose we can find *some* stationary time τ with a halting state. Let us check that any optimal stationary time τ' must also have a halting state. To see this, note that by the first part we know τ is also optimal so $\mathbb{E}(\tau) = \mathbb{E}(\tau')$. Hence by (35) we see that the exit frequencies v_x and v'_x are equal. In particular since τ has a halting state y say, we see that $v_y = v'_y = 0$. This implies that y is also a halting state for τ' (a state y is halting if and only if $v_y = 0$!) Hence the proof of the theorem is complete if we can prove that there exists a stationary time τ with a halting state.

Lovász and Winkler give several examples of a stationary time with a halting state. Here is one simple one, which they call the **local rule**. Fix a set v_x of numbers satisfying the "conservation equations" (34) subject to the constraint $\min_x v_x = 0$; these are our candidate exit frequencies for the stationary time. It is not hard to see that a solution to these equations exist: we are trying to solve $(I - P)v = \pi - \mu$. The solutions of this equation must of the form $v + t\pi$, where $t \in \mathbb{R}$ and v is any particular solution to this equation, such as the exit frequencies obtained by letting the Markov chain run until it hits an independent sample from the equilibrium distribution. Since v, π both have positive entries it is easy to check that we can find $t \in \mathbb{R}$ such that $\min_x(v_x + t\pi_x) = 0$, e.g. by considering the infimum over all t such that $\min_x(v_x + t\pi_x) \geq 0$, which must be finite by considering the situation as $t \to -\infty$.

Given such a vector v_x, the local rule is simply as follows. At node x, stop with probability $\pi_x/(\pi_x + v_x)$. The probability of stopping depends just on the state and not on the time. Moreover clearly if $v_x = 0$ then the rule stops at x with probability 1, so x is a halting state. It remains to check that τ is stationary.

Lemma 7.2. *The local rule induces a stationary time, i.e., $\mathbb{P}(X_\tau = x) = \pi$.*

Proof. Let v'_x denote the exit frequency of τ at x and let π'_x denote the distribution of X_τ. We would like to show that $\pi' = \pi$; note that it suffices to show that $v' = v$ (by (34)). The main claim is that $v'_x/v_x = \pi'_x/\pi_x$. To see this, note that

$$\pi'_x = \mathbb{P}(X_\tau = x) = \sum_{k=0}^{\infty} \mathbb{P}(X_k = x, \tau = k) = \sum_{k=0}^{\infty} \mathbb{P}(X_k = x, \tau \geq k)\frac{\pi_x}{v_x + \pi_x}$$

whereas

$$v'_x = \sum_{k=0}^{\infty} \mathbb{P}(X_k = x; \tau > k) = \sum_{k=0}^{\infty} \mathbb{P}(X_k = x, \tau \geq k)\frac{v_x}{v_x + \pi_x}.$$

Noting that v'_x is zero whenever $v_x = 0$, we deduce

$$\frac{v'_x}{v_x} = \frac{\pi'_x}{\pi_x}.$$

Let $\alpha = \max_x \pi'_x/\pi_x$. Since both vectors sum to 1, it must be that $\alpha \geq 1$. Hence using the conservation law (34) but for v',

$$\sum_x p_{xy} v'_x = v'_y + \pi'_y - \mu_y \tag{36}$$

If y is a maximiser of π'_x/π_x, which is also a maximiser of v'_x/v_x by the above, we see that

$$\sum_x p_{xy} \alpha v_x \geq \sum_x p_{xy} v'_x = \alpha(v_y + \pi_y) - \mu_y \geq \alpha(v_y + \pi_y - \mu_y).$$

But the right hand side is, using the conservation law for v, also equal to $\alpha \sum_x p_{xy} v_x$, which is the left hand side. Hence all inequalities must have been equalities, and we deduce that $v'_x = \alpha v_x$ for every x such that $p_{xy} > 0$.

Applying the same argument to those x and going on inductively, we deduce (since the chain is irreducible we can eventually reach every x in the state space) that $v'_x = \alpha v_x$ for every x. But then v'/v is the constant vector α, and hence so is π'/π. Since $\sum_x \pi'_x = \sum_x \pi_x = 1$ it must be $\alpha = 1$ and we have $\pi'_x = \pi_x$. □

Since the lemma implies the "only if" part of the result, we have finished the proof of Theorem 7.4. □

7.4 Stationary times are comparable to mixing times

We have so far in these notes focused on the notion of mixing in the sense of approximating the invariant distribution for the total variation distance. But another natural definition is to define the mixing time as a mean-optimal stationary time in the above sense of Lovász and Winkler. This leads the following definition.

Definition 7.6. *Let $t_{\text{stat}} = \sup_x \inf_\tau \mathbb{E}_x(\tau)$, where the sup is over all starting point x and the inf over all stationary times from x. We call t_{stat} the stationary mixing time.*

A natural question is to ask whether t_{mix} and t_{stat} are roughly identical. While $t_{\text{stat}} \leq c\, t_{\text{mix}}$ is always true, it turns out that the converse requires nonperiodicity or at least laziness (as the following example shows: consider two vertices joined by an edge plus a self-loop of weight e^{-n}: then t_{mix} is of order e^n, but t_{stat} is of order 1). However, making these assumptions, we have the following result.

Theorem 7.5 (Aldous [1])**.** *There exist two universal constants c_1, c_2 such that if X is a lazy reversible chain then*

$$c_1\, t_{\text{mix}} \leq t_{\text{stat}} \leq c_2\, t_{\text{mix}}\,. \tag{37}$$

Proof. We only show here the "easy direction", which is $t_{\text{stat}} \leq c_2\, t_{\text{mix}}$ for some universal constant c_2 (and which doesn't require any assumptions on the chain). Set $t = t_{\text{mix}}$ and define a stopping time τ as follows. Take a coupling of X_t with π. If the coupling succeeds set $\tau = t$. If not, condition on X_t and consider the chain at time $2t$, try again to couple it with π. If the coupling succeeds, set $\tau = 2t$, and so on and so forth. Then note that X_τ has by definition the distribution of π, and moreover τ has the same distribution as $N\, t_{\text{mix}}$, where N is a Geometric random variable with probability of success $1/4$. Hence $\mathbb{E}(\tau) = 4\, t_{\text{mix}}$ and we get the desired inequality with $c_2 = 4$.

The converse direction was shown by Aldous [1] for a reversible chain in continuous time. An argument for the lazy chain in discrete time can be found in a paper by Peres and Sousi [22] (roughly following the same lines as Aldous' original argument, though a small problem in the argument of [1] has been corrected there). In that paper, it is further shown that even if the walk is not assumed to be lazy, then averaging over two consecutive steps is enough to guarantee the inequality (37). See also [21] for a related independent and simultaneous work. □

7.5 Cover times

We end this section with a brief discussion of cover times. The cover time of a Markov chain is the first time that all states have been visited at least once. This can be thought of as an alternative notion of mixing (and indeed in some cases such as on the n-cycle, the cover time is a strong stationary time).

Everything in this Chapter I have learnt from Yuval Peres in a conversation at dinner in a sushi restaurant in San Jose. I thank him warmly for his enthusiastic account of Matthew's bound, which is close to the one I will give here (and is of course also closely related but perhaps a bit simpler than the one in [17]).

We start with an incredibly general upper bound on the cover time, due to Matthews, which relate cover time to maximal hitting times of states. First, some definitions. We let τ_{cov} is the first time all vertices have been covered and set

$$t_{\text{cov}} = \inf_{x \in S} \mathbb{E}_x(\tau_{\text{cov}}).$$

Example 7.5. *On the complete graph K_n with self-loops, τ_{cov} is exactly the coupon collector time so $t_{\text{cov}} = (1 + \ldots + 1/n)$.*

Example 7.6. *Show that on the n-cycle, $t_{\text{cov}} = n(n-1)/2$.*

We turn to hitting times. For $y \in S$, set $\tau_y = \inf\{n \geq 0 : X_n = y\}$ and set

$$t_{\text{hit}} = \max_{x,y} \mathbb{E}_x(\tau_y).$$

Matthews' bound is as follows.

Theorem 7.6. *If $|S| = n$ we have and X is an irreducible Markov chain on S,*

$$t_{\text{cov}} \leq \left(1 + \ldots + \frac{1}{n}\right) t_{\text{hit}}$$

At a heuristic level, we "collect" every state at rate at least $1/\, t_{\text{hit}}$, so the right hand side is nothing but the coupon collector bound.

Proof. The proof idea is both simple and subtle. The main is the following. Suppose we order the states of S uniformly at random, according to some uniform random permutation $\sigma \in \mathcal{S}_n$.

Let T_k be the time collect (i.e., visit) states $\sigma(1), \ldots, \sigma(k)$. Then the time to visit $\sigma(k+1)$ is at most t_{hit} in expectation.

But there is also a possibility that this time is zero, if for instance we have already visited $\sigma(k+1)$ by time T_k. This will happen whenever $\sigma(k+1)$ is not the last visited sites among $\{\sigma(1), \ldots, \sigma(k+1)\}$ by the chain. However, since σ is independent of the chain, then the latter has (by symmetry) probability $(1 - 1/(k+1))$. Summing gives us the result.

More precisely, let $\mathcal{F}$ be the filtration generated by X and by the permutation σ. Then T_k is an $\mathcal{F}$-stopping time. Furthermore if A_k is the event that $\sigma(k+1)$ is already visited by time T_k then A_k is $\mathcal{F}_{T_k}$-measurable and $\mathbb{P}(A_k) = 1/(k+1)$ by symmetry. Hence

$$\begin{aligned}\mathbb{E}(T_{k+1} - T_k) &= \mathbb{E}(\mathbb{E}(T_{k+1} - T_k | \mathcal{F}_{T_k}))\\ &\le \mathbb{E}(\mathbf{1}_{\{A_k\}}\, t_{\text{hit}})\\ &= \frac{t_{\text{hit}}}{k+1}.\end{aligned}$$

Summing from $k = 0$ to $n - 1$ gives the result. □

Surprisingly, Matthew's bound is sharp in many examples. To get a lower bound, we can proceed in a relatively similar manner by consider the minimal hitting time, restricted to sets which are relatively sparse (so that minimal hitting time can be quite large). In graphs where the walk needs to mix before hitting points, this will give us the right answer.

Thus fix $A \subset S$ and let $t^A_{\min} = \min_{x \in A, y \in A, x \neq y} \mathbb{E}_x(\tau_y)$. The proof of the following result is very similar and is left as an exercise.

Theorem 7.7. *For any $A \subset S$,*

$$t_{\text{cov}} \ge t^A_{\min}(1 + \ldots + \frac{1}{k-1}),$$

where $k = |A|$.

A striking recent development in connection with cover times is the beautiful result of Ding, Lee and Peres [13], which relates the cover time t_{cov} on a graph with vertex se S to the maximum of the Gaussian free field $(h_x, x \in S)$. This is the centered Gaussian random vector such that $h_{v_0} = 0$ for some fixed $v_0 \in S$ and with covariance $\mathbb{E}((h_x - h_y)^2) = R(x, y)$ the effective resistance between x and y. Then the Ding, Lee and Peres theorem [13] is the following:

Theorem 7.8.

$$t_{\text{cov}} \asymp |E| \left(\mathbb{E}(\max_{x \in S} h_x)\right)^2$$

where the implied constants are universal.

This falls outside the scope of these notes, so we will not give a proof of this result.

7.6 Example: cover time of the torus in dimension $d \ge 3$

We explain how to apply Matthew's method to estimate the cover time on $\mathbb{Z}^d_n$, the d-dimensional torus of sidelength as $n \to \infty$, for fixed $d \ge 3$. The key to doing this is an estimate on the hitting time of points. The main estimate is the following:

Proposition 7.1. *There exists c_1, c_2 such that for any $x \neq y \in \mathbb{Z}_n^d$, then*

$$c_1 n^d \le \mathbb{E}_x(\tau_y) \le c_2 n^d.$$

Note in particular that these estimates do not depend on the distance between x and y. The reason is that even when x, y are close, due to transience of $\mathbb{Z}^d$ for $d \ge 3$, there is a positive probability for the walk to escape. Then the walk will mix before hitting y. Thus the hitting time is of order 1 on an event of positive probability (when x, y are very close) but of order n^d on an event of positive probability. So up to constant the expectation will be of order n^d.

First we explain why n^d is indeed the right order of magnitude, observe that in time n^d the walk is highly mixed, because $n^d \gg n^2$ which is the mixing time. Thus it takes time of order n^d before the expected amount of visits to y is of order 1. A second moment argument then shows that at that time, the probability of visiting is indeed positive, since at each visit the total number of visits will be locally bounded (again by transience). The book [17] proves this using some fairly general theory of electrical network theory (based on looking at the effective resistance and commute time), but we give a completely elementary proof here.

Proof. We consider separately the lower and upper bound. In both cases we will need the following lemma.

Lemma 7.3. *If $x \in Z_n^d$ and $s = cn^2$ for some fixed $c > 0$ then $\mathbb{P}_x(\tau_y \ge s) \ge 1/C$ for some constant C. Also, $\mathbb{E}_x L(s, y) \le C$ for some constant $C > 0$ where $L(s, x)$ is the number of visits to x by time s (or local time of the chain at x by time s).*

Proof. Note that in $\mathbb{Z}^d$, by transience, the walk hits the boundary of a box of size $n/2$ around y without hitting y with probability bounded below. Furthermore, if that is the case, there is positive probability staying away from the box of size $n/4$ around y during time cn^2, but stay in the box of size $3n/4$. If that is the case, the walk on $\mathbb{Z}_n^d$ will definitely not hit y during $[0, cn^2]$, as desired.

The second part follows by applying the Markov property at each successive visit to y. □

Given this lemma, let us consider the lower bound. First of all choose c_1 and set $s = 2\,t_{\text{mix}}$, and let $I = [s, c_1 n^d]$. Then during I, the expected number of visits to y is at most $2\pi(y)|I| \le c_1 n^d \times 2/n^d = 2c_1$. Hence with probability at least $1 - 2c_1$ there is no visit during I. We also know from Lemma 7.3 that with positive probability there is no visit during $[0, s]$. Hence by choosing c_1 sufficiently small and a union bound, we see that $\tau_y \ge c_1 n^d$ with positive probability. Hence $\mathbb{E}_x(\tau_y) \ge c_1' n^d$, as desired.

We now turn to the upper bound on τ_y. Set $s = 2\,t_{\text{mix}}$ as above. We claim that the probability of hitting y during $[s, 2s]$ is at least c/n^{d-2}. Indeed let N denote the number of visits to y during $[s, 2s]$. Then by Cauchy–Schwarz,

$$\mathbb{P}(N > 0) \ge \frac{\mathbb{E}(N)^2}{\mathbb{E}(N^2)}.$$

Now, on the numerator we have

$$\mathbb{E}(N) \ge (1/2)\pi(y)s \ge cn^{2-d}.$$

On the other hand, in the denominator, by the Markov property,

$$\begin{aligned}\mathbb{E}(N^2) &\le \sum_{s\le j\le k\le 2s} \mathbb{E}(\mathbf{1}_{\{X_j=y,X_k=y\}})\\ &\le \sum_{s\le j\le 2s} \mathbb{E}(\mathbf{1}_{\{X_j=y\}}\mathbb{E}_y(L(s,y)))\\ &\le C\mathbb{E}(N)\end{aligned}$$

by Lemma 7.3. Consequently,

$$\mathbb{P}(N>0)\ge c\mathbb{E}(N)\ge cn^{2-d}.$$

Iterating this bound,

$$\mathbb{P}_k(\tau_y>2ks)\le(1-cn^{2-d})^k\le\exp(-kcn^{2-d}).$$

Summing over all k, we get

$$\mathbb{E}_x(\tau_y)\le Csn^{d-2}=Cn^d$$

as desired. □

We now estimate the cover times on the torus. By Theorem 7.6 and Proposition 7.1, we see that

$$t_{\text{cov}}\le c_2n^d\log n.$$

On the other hand, for the lower bound we can take A to be $\mathbb{Z}_n^d$ itself, since the minimum hitting time is of the same order of magnitude as the maximum hitting time (even when the points are close, as explained). Hence

$$t_{\text{cov}}\ge c_1n^d\log n.$$

8 Nash inequalities.

`The reader is advised that this chapter requires extensive rewriting.`

8.1 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 8.1. *Assume that (K, π) is irreducible and reversible. We say that it satisfies a* Nash inequality *if, for all $g \in \ell^2(\pi)$,*

$$\mathrm{var}_\pi(g)^{1+2/d} \le C\mathcal{E}(g,g)\|g\|_1^{4/d}. \tag{38}$$

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 8.1. *Then for all $t > 0$,*

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

and for all $t > 0$,

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

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 (38), we have

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

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 $\|f\|_1 \le 1$, and using reversibility:

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

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. □

Remark. The constant C in (38) must satisfy:

$$C \geq 1/\gamma. \tag{39}$$

Indeed, if g is such that $\mathbb{E}_\pi(g) = 0$ then by (38) we have:

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

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

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

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

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

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

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

this implies $\gamma \geq 1/C$ which is (39).

The conclusion of Theorem 8.1 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 8.1. *If Nash's inequality (38) 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 [24] for a proof.

8.2 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 $f : S \to \mathbb{R}$, 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:

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

This is Nash's inequality (38) with $d = 1$ (the dimension of the space) and $C = 64(2n+1)^2$. Thus, by (39),

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

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).

9 Martingale methods and evolving sets

`The reader is advised that this chapter requires extensive rewriting`

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 $Q(S,y) = \sum_{x\in S} Q(x,y)$.

9.1 Definition and properties

Definition 9.1. *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 9.1. *The sequence $\{\pi(S_n)\}_{n\geq 0}$ is a martingale.*

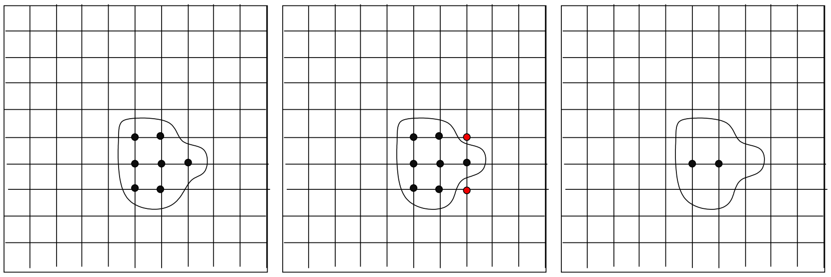

Figure 4: The initial state of the evolving set and two possible transitions: (a) $1/8 < U < 2/8$ and (b) $6/8 < U < 7/8$.

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 9.2. *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 $n-1$,

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

□

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

Lemma 9.3. *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 $U \overset{d}{=} 1-U$, 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:

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

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:

$$S^\sharp = \begin{cases} S & \text{if } \pi(S) \le 1/2 \\ S^c & \text{otherwise} \end{cases} \tag{40}$$

Lemma 9.4. *Let $\mu_n = p^n(x,\cdot)$ be the distribution of the Markov chain after n steps started from x. Then*

$$\chi(\mu_n,\pi) \le \frac{1}{\pi(x)}\mathbb{E}_{\{x\}}\left(\sqrt{\pi(S_n^\sharp)}\right). \tag{41}$$

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

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

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. □

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 (41) 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 $X_T \overset{d}{=} \pi$ exactly, and moreover T is independent of X_T. See section 17.7 in [17] for more information about this.

9.2 Evolving sets as a randomised 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 9.1. *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 $Z_n = \sqrt{\pi(S_n^\sharp)}/\pi(S_n)$. 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 (41), we get:

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

$$\leq \hat{\mathbb{E}}_x\left(\frac{\sqrt{\pi(S_n^\sharp)}}{\pi(S_n)}\right) = \hat{\mathbb{E}}_x(Z_n).$$

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 9.5. *Let $f : [0, \infty) \to [0, 1]$ 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 $n \geq \displaystyle\int_\delta^{L_0} \frac{2dz}{zf(z/2)}$ 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

$$L_{n+1} \leq L_n(1 - g(L_n)) \tag{42}$$

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 9.1. Let us compute $\hat{\mathbb{E}}_x(Z_{n+1}/Z_n|S_n)$.

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

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 9.5 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 $\hat{\mathbb{E}}(Z_n) = \chi(\mu_n, \pi)$, 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 9.1.

□

9.3 Application: the isoperimetric profile

Theorem 9.1 can be used to prove a bound on mixing times which slightly more intuitive than the above, and which is given in terms of the *isoperimetric* or *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

$$t_{\text{mix}}(\varepsilon) \leq t_{\text{rel}} \log((\pi_* \varepsilon)^{-1}) \leq 8I^{-2}(\log \frac{1}{\pi^*} + \log \frac{1}{\varepsilon}). \tag{43}$$

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 9.2. *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 (43).

Proof. The proof is essentially a consequence of Theorem 9.1, and of the following lemma which relates the conductance Φ_S to the boundary gauge $\Psi(S)$ used in the previous theorem:

Lemma 9.6. *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 9.2. 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

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

and therefore:

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

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

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

then

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

and therefore,

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

This finishes the proof of Theorem 9.2. □

10 Coupling from the past: a method for 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 the 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.

10.1 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 10.1. *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 $C/(1+\exp(\beta\Delta H))$:

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

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.

10.2 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 $U_{-1}, U_{-2}, \ldots$ 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 $U_{-T}, \ldots, U_0$ 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 $U_{-2T}, \ldots, U_{-T}, \ldots, U_{-1}$, 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 10.2. *This algorithm terminates almost surely in finite time. Moreover, $X \overset{d}{=} \pi$.*

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

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

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

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 $T^{\star}$ 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^0_{-T} 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 10.3.

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

In particular, this means that coupling from the past is very efficient, since of course T_* 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^* 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 10.1. *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 $X^k_- = F^k_0(\widehat{-1})$ is different from $X^k_+ = F^k_0(\widehat{1})$, it must be the case that

$$h(X^k_-) \leq h(X^k_+) - 1$$

since we know that $X^k_- \preceq X^k_+$. Therefore,

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

from which the inequality follows. □

Lemma 10.2. *The quantity $\mathbb{P}(T^\star > k)$ is submultiplicative: for $k_1, k_2 \geq 0$:*

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

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. □

Lemma 10.3. *For all $k \geq 1$,*

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

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

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

and thus

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

This proves the lemma. □

Proof. (of Theorem 10.3). 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 10.1,

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

i.e., $\mathbb{P}(T^\star \leq k) \geq 1 - 1/e$. Thus by Lemma 10.3

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

as claimed. □

11 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 [2] 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 11.1. (Aldous 1983 [2]) *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 11.2. (Bayer-Diaconis 1992 [5]) *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 11.3. (Bayer-Diaconis 1992 [5]) *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:

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

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

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

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 [5] is wholeheartedly recommended !

We will present the key ideas that lead to the proof of Aldous' results (Theorem 11.1.) 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 11.1. *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:

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

then there are two rising sequences:

$$\begin{array}{ccccccccccccc} 1 & & 2 & & & 3 & & 4 & 5 & & 6 & & \\ & 7 & & 8 & 9 & & 10 & & & 11 & & 12 & 13 \end{array}$$

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

$$X'_m =_d X_m^{-1}$$

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

11.1 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 11.1. *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

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

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 $m = 2\log_2 n$. □

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

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

and $P(T > m) \to 0$ 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 $1 \le j \le n-1$) if $\sigma(j) > \sigma(j+1)$. Let

$$\text{Des}(\sigma) = \#\text{descents of } \sigma = \sum_j a_j \tag{44}$$

where a_j is the indicator of the event that σ has a decent at j. It is trivial to observe that

$$R(\sigma) = \mathrm{Des}(\sigma^{-1}) - 1$$

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 11.2. *Let $\sigma =_d U$. Then*

$$E(Des(\sigma)) = (n-1)/2 \text{ and } \operatorname{var} Des(\sigma) \sim n/12 \tag{45}$$

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, $\mathrm{Des}(\sigma)$ 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

$$\mathrm{Des}(X'_m) \leq 2^m - 1$$

If $m = (1-\varepsilon)\log_2 n$ then $\mathrm{Des}(X'_m) \leq n^{1-\varepsilon}$ and thus this is incompatible with (45). 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

$$E(\mathrm{Des}(X'_m)) = \frac{n-1}{2} - n^\beta \tag{46}$$

with $\beta > 1/2$, while the variance of $\mathrm{Des} X'_m$ stays $O(n)$. This will imply again that the total variation distance is approximately 1 in this regime. Indeed, (45) implies that for a uniform permutation, the number of descents is $n/2 \pm \sqrt{n}$. Here, (46) 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 modifitcation of the Birthday problem.

Lemma 11.3. *Throw n balls in M urns, and suppose $M \sim n^\alpha$. Let*

$$U_n = \#\{j \leq n : \textit{ball } j \textit{ and ball } i \textit{ fall in same urn, for some } i < j\}$$

Then

$$E(U_n) \sim \frac{1}{2} n^{2-\alpha} \text{ and } \operatorname{var}(U_n) \sim \frac{1}{2} n^{2-\alpha} \tag{47}$$

There surely is a central limit theorem, too.

To prove (46), 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 $(a_j, j$ 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:

$$E(\mathrm{Des}(X'_m)|J_m) = \frac{n-1}{2} - \#J_m$$

(each integer gives us probability 1/2 of being a descent, except those who are in J_m). Also,

$$\mathrm{var}\,\mathrm{Des}(X'_m) = O(n)$$

Now, to conclude, remark that $\#J_m =_d U_n$ in equation (47) and thus

$$E(\#J_m) \sim \frac{1}{2} n^{2-\alpha}.$$

Since $\beta = 2 - \alpha > 1/2$, the lower-bound is proved.

11.2 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 [2] 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, $\mathrm{Des}X'_m$ 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 11.4. *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 11.4 and the above comment thus imply that the total variation distance between the law of $\mathrm{Des}X'_m$ and $\mathrm{Des}\sigma$ (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, $\mathrm{Des}(X'_m)$ is a sufficient statistics for X'_m, and it is obviously so for a uniform permutation as well. On the other hand,

Lemma 11.5. *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.

11.3 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 riffling 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

$$d(m) = \sum_{\pi \in \mathcal{S}_n} \left(P^m(\pi) - \frac{1}{n!} \right)^+ = \sum_{\pi \in \mathcal{S}_n} \frac{1}{n!} (n! P^m(\pi) - 1)^+ \tag{48}$$

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

$$n! P^m(\pi) = f_n(h) := \exp\left(\frac{-h}{c\sqrt{n}} - \frac{1}{24c^2} - \frac{1}{2} \left(\frac{h}{cn} \right)^2 + O(1/n) + O(h/n) \right) \tag{49}$$

Let h^* be defined by

$$h \le h^* \iff P^m(\pi) \ge 1/n!$$

This h^* tells us what are the nonzero terms in (48). Now, by setting the exponent equal to 0 in (49), we obtain

$$h^* = -\frac{\sqrt{n}}{24c} + \frac{1}{12c^3} + B + O(1/\sqrt{n}) \tag{50}$$

It follows that

$$d(m) = \sum_{-n/2 \le h \le h^*} \frac{R_{nh}}{n!}(f_n(h) - 1)$$

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 [26]. Tanny and Stanley show the remarkable formula that if $X_1, \dots, X_n$ are i.i.d. uniform on $(0,1)$

$$\frac{a_{nj}}{n!} = P(j \le X_1 + \dots X_n \le j + 1) \tag{51}$$

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 (45).

From then on, it is essentially a game of algebraic manipulations to obtain Theorem 11.3. We refer the interested reader to p. 308 of [5] for details.

References

[1] Aldous, D.J., 1982. Some inequalities for reversible Markov chains. *Journal of the London Mathematical Society*, 2(3), pp.564-576.

[2] D. Aldous (1983). Random walks on finite groups and rapidly mixing Markov chains. *Sminaire de Probabilits XVII. Lecture Notes in Math.* 986, 243–297. Springer, New-York.

[3] Aldous, D. and Diaconis, P., 1986. Shuffling cards and stopping times. *American Mathematical Monthly,* pp.333–348.

[4] 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`.

[5] D. Bayer and P. Diaconis (1992). Trailing the dovetail shuffle to its lair. *Ann. Probab.*, 2, 294-313.

[6] Berestycki, N. and Durrett, R., 2008. Limiting behavior for the distance of a random walk. *Electron. J. Probab*, 13, pp.374–395.

[7] Berestycki, N., Lubetzky, E., Peres, Y., and Sly, A. (2015). Random walks on the random graph. arXiv preprint arXiv:1504.01999.

[8] Bollobás, B. and de la Vega, F., 1982. The diameter of random regular graphs. *Combinatorica*, 2(2), pp.125–134.

[9] Diaconis, P. 1988. *Group Representations in Probability and Statistics*, Lecture Notes - Monograph Series, vol. 11, Inst. Math. Stat., Hayward, CA.

[10] P. Diaconis and M. Shahshahani, Generating a random permutation with random transpositions. *Z. Wahrsch. Verw. Gebiete* **57:2** (1981), 159–179.

[11] P. Diaconis and L. Saloff-Coste. Comparison techniques for random walks on finite groups. *Ann. Probab.,* 21, 2131–2156 (1993).

[12] Ding, J., Kim, J.H., Lubetzky, E. and Peres, Y., 2011. Anatomy of a young giant component in the random graph. *Rand. Struct. Algor.*, 39(2), pp.139–178.

[13] Ding, J., Lee, J.R. and Peres, Y., 2012. Cover times, blanket times, and majorizing measures. *Ann. Math.*, 175(3), pp.1409–1471.

[14] R. Durrett. *Random graph dynamics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2010.

[15] M. Jerrum and A. Sinclair. Approximating the permanent. *SIAM Journal on Computing*, 18, 1149–1178 (1989).

[16] Lacoin, H., 2013. Mixing time and cutoff for the adjacent transposition shuffle and the simple exclusion. arXiv preprint arXiv:1309.3873.

[17] D. Levin, Y. Peres and E. Wilmer. 2009. *Markov chains and mixing times.* Americian Mathematical Society.

[18] Lubetzky, E. and Peres, Y., 2015. Cutoff on all Ramanujan graphs. arXiv preprint arXiv:1507.04725.

[19] E. Lubetzky and A. Sly. Cutoff phenomena for random walks on random regular graphs. *Duke Math. J.*, 153(3):475–510, 2010.

[20] B. Morris and Y. Peres. Evolving sets, mixing and heat kernel bounds. *Probab. Theor. Rel. Fields*, 133: 245–266 (2005).

[21] R. I. Oliveira. Mixing and hitting times for finite Markov chains. *Electron. J. Probab.*, 17(70), 12, 2012.

[22] Peres, Y. and Sousi, P., 2015. Mixing times are hitting times of large sets. *J. Theor. Probab.*, 28(2), pp.488–519.

[23] J. Propp and D. Wilson (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Rand. Struct. Algor.*, 9, pp. 223–252

[24] Saloff-Coste, L. 1997. *Lectures on finite Markov chains*, Lectures on Probability Theory and Statistics, Ecole dEté de Probabilités de Saint-Flour XXVI - 1996, pp. 301-413

[25] L. Saloff-Coste, 2003. *Random Walks on Finite Groups.* In: H. Kesten, ed. *Probability on Discrete Structures*, Encyclopaedia of Mathematical Sciences (110), Springer.

[26] R. Stanley (1977). Eulerian partitions of a unit hypercube. In: *Higher Combinatorics* (M. Aigner, ed.) Reidel, Dordecht.