1 Random walks: an introduction

1.1 Simple random walks on $\mathbb{Z}$

1.1.1 Definitions

Let $(\xi_n, n \geq 1)$ be i.i.d. (independent and identically distributed) random variables such that $\mathbb{P}(\xi_n = +1) = p$ and $\mathbb{P}(\xi_n = -1) = q = 1 - p$, where $0 < p < 1$.

Definition 1.1. The simple random walk on $\mathbb{Z}$ is the random(stochastic) process $(S_n, n \in \mathbb{N})$ defined as $S_0 = 0, S_n = \xi_1 + \cdots + \xi_n, n \geq 1$

Example 1.2. This type of random walk could be associated with the "walk of a drunkard" (Figure 1). The moves are independent, $S_{n+1} = S_n + \xi_{n+1}$ (i.e. at each step, we take an action that has nothing to do with the previous steps). However, the position at step $n+1$ is highly correlated with the position at step n.

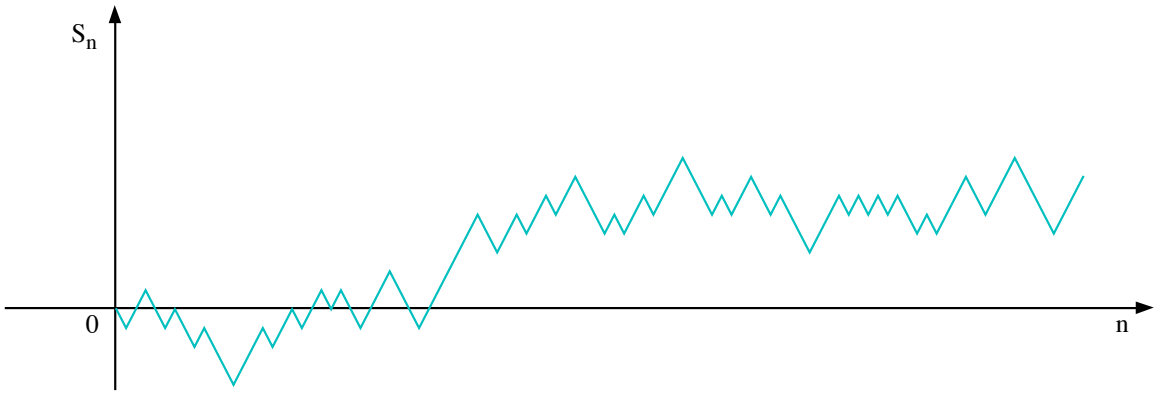


Figure 1: The "walk of a drunkard"

Definition 1.3. If $p = \frac{1}{2}$ the random walk is **symmetric**.

Definition 1.4. If $p \neq \frac{1}{2} (\Rightarrow q \neq \frac{1}{2})$ the random walk is **asymmetric**.

1.1.2 Basic Properties

- **Expectation** (on average, where are we at time n?)

$$\begin{aligned} \mathbb{E}(S_n) &= \mathbb{E}(\xi_1 + \cdots + \xi_n) \\ &= \mathbb{E}(\xi_1) + \cdots + \mathbb{E}(\xi_n) && \text{by linearity of expectation} \\ &= n\mathbb{E}(\xi_1) && \text{identically distributed random variables} \\ \mathbb{E}(\xi_1) &= 1 \cdot p + (-1) \cdot q = p + (p-1) = 2p - 1 \Rightarrow \\ \Rightarrow \mathbb{E}(S_n) &= n(2p-1) \end{aligned}$$

- **Variance**

$$\begin{aligned}
\mathrm{Var}(S_n) &= \mathrm{Var}(\xi_1 + \cdots + \xi_n) \\
&= \mathrm{Var}(\xi_1) + \cdots + \mathrm{Var}(\xi_n) && \text{by independence of the random variables} \\
&= n\mathrm{Var}(\xi_1) && \text{identically distributed random variables} \\
\mathrm{Var}(\xi_1) &= \mathbb{E}({\xi_1}^2) - [\mathbb{E}(\xi_1)]^2 \\
&= p + (1-p) - (2p-1)^2 \\
&= 1 - 4p^2 + 4p - 1 = 4p(1-p) \\
\Rightarrow \mathrm{Var}(S_n) &= 4np(1-p)
\end{aligned}$$

  Note: $\mathrm{Var}(S_n)$ reaches its maximum for $p = \frac{1}{2}$, having $\mathrm{Var}(S_n) = n$

- **Standard Deviation:** $Stdev(S_n) = \sqrt{\mathrm{Var}(S_n)} = \sqrt{4np(1-p)}$

Example 1.5. Examples of symmetric and asymmetric random walks are illustrated below (Figure 2). One can notice that when $p > 1/2$, on average, the trajectory of the walk is "climbing".

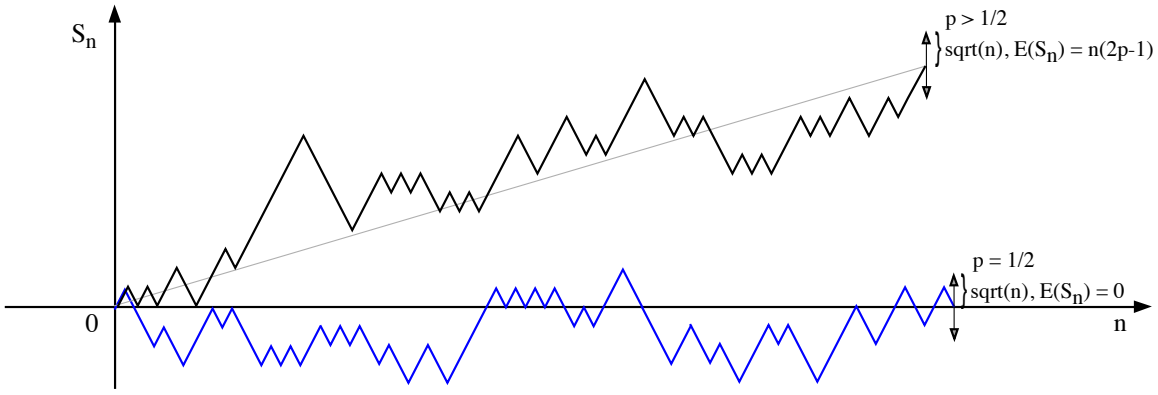


Figure 2: Symmetric vs. Asymmetric Random Walks

Note that it is possible to recenter and renormalize S_n as follows: $S_n = (2p-1)n + \sqrt{4np(1-p)}Z_n$, where $\mathbb{E}(Z_n) = 0$ and $\mathrm{Var}(Z_n) = 1$. However, if we recenter and renormalize, the relation between Z_n and Z_{n+1} will not be as simple as the relation between S_n and S_{n+1}.

1.1.3 Two Important Theorems

Theorem 1.6. The law of large numbers
For a sequence of i.i.d. random variables ξ_i taking values in $\mathbb{R}$,

$$\frac{S_n}{n} \xrightarrow{n\to\infty} \mathbb{E}(\xi_1) = 2p - 1$$

with high probability (this can be taken to mean that the probability of this event goes to 1 as $n \to \infty$), provided that $\mathbb{E}(\xi_1) < \infty$.

Theorem 1.7. The Central Limit Theorem

$$Z_n \xrightarrow{n\to\infty} Z \sim N(0,1)$$

When n goes to infinity, the distribution of Z_n will tend to a Gaussian random variable, with mean 0 and variance 1.

1.1.4 Transition Probabilities of Markov Chains

Let us begin with a slight generalization of random walks, where the random walk does not necessarily start in 0: $S_0 \neq 0$, $S_n = S_0 + \xi_1 \cdots + \xi_n$ and S_0 is independent of ξ_i, $i = 1 \dots n$.

Definition 1.8. One step transition probability
Define the transition probability $p_{i,j}$ as:

$$\begin{aligned} p_{i,j} &= \mathbb{P}(S_{n+1} = j | S_n = i) \\ &= \mathbb{P}(S_n + \xi_{n+1} = j | S_n = i) \\ &= \mathbb{P}(\xi_{n+1} = j - i | S_n = i) \\ &= \mathbb{P}(\xi_{n+1} = j - i) = \begin{cases} p, & \text{if } j - i = +1. \\ q, & \text{if } j - i = -1. \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

From the above mentioned definition, it is possible to see that at each step, there are only two options (either go upwards, or downwards).

Definition 1.9. N-steps transition probabilities
Define the transition probability $p_{i,j}^{(n)}$ as:

$$p_{i,j}^{(n)} = \mathbb{P}(S_n = j | S_0 = i)$$

In order to deduce the value of $p_{i,j}^{(n)}$, let us start by a simplified example.

Example 1.10. Let us compute $p_{0,0}^{(2n)} = \mathbb{P}(S_{2n} = 0 | S_0 = 0)$. Firstly, note that the paths of this type need to contain an even number of steps. In an odd number of steps, it would be impossible to arrive at position 0, if we start the path from 0. Moreover, every path that starts from 0 and ends in 0 in $2n$ steps, needs to go n times up and n times down (the order does not matter). The probability of one such path is $p(\text{path } n \text{ ups}, n \text{ downs}) = p^n q^n$. Moreover, there exist $\binom{2n}{n}$ such paths (in order to count the number of paths, imagine that you have $2n$ boxes and you need to figure out in how many ways you can place the n values in n of the boxes). Therefore, $p_{0,0}^{(2n)} = \binom{2n}{n} p^n q^n$.

Similarly, we can compute $p_{0,2k}^{(2n)} = \mathbb{P}(S_{2n} = 2k | S_0 = 0)$. In this case, we need to go upwards $n + k$ times and downwards $n - k$ times. So, $p_{0,2k}^{(2n)} = \binom{2n}{n+k} p^{n+k} q^{n-k}$.

Note that $p_{0,2k}^{(2n)} = p_{j,2k+j}^{(2n)}$. Translating the initial position does not change anything regarding the total number of paths, or the probability of obtaining such a path. Therefore, $p_{j,2k+j}^{(2n)} = \binom{2n}{n+k} p^{n+k} q^{n-k}$.

1.2 Random walks in $\mathbb{Z}^d$, $(d = 2)$

Definition 1.11.

$$\mathbb{Z}^2 = (i, j) : i, j \in \mathbb{Z}$$

In this case, we will consider only symmetric random walks. Let $(\vec{\xi_n}, n \geq 1)$ be i.i.d. random variables such that:

- $\mathbb{P}(\vec{\xi_n} = (0, 1)) = \mathbb{P}(\vec{\xi_n} = (0, -1)) = \mathbb{P}(\vec{\xi_n} = (1, 0)) = \mathbb{P}(\vec{\xi_n} = (-1, 0)) = \frac{1}{4}$
- $\vec{S_0} = \vec{0}, \vec{S_n} = \vec{\xi_1} + \cdots + \vec{\xi_n}$, where S_n^i is the i^{th} component of $\vec{S_n}$, $i \in \{1, 2\}$

This model could be interpreted as a discrete version of a brownian motion. As we will see in the exercises, the 2-dimensional random walk is far from being the sum of two 1-dimensional random walks.

Random Walks: WEEK 2

1 Markov chains, Recurrence, Transience

1.1 Basic definitions, Chapman-Kolmogorov equation

1.1.1 Basic definitions

Definition 1.1. A Markov chain is a discrete-time stochastic process $(X_n, n \geq 0)$ such that each random variable X_n takes values in a discrete set S (the state space) and such that

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i)$$

for all $n \geq 0$ and all states $i, j, i_0, \ldots, i_{n-1}$. In other words, the Markov property states that for a Markov process, given the present, the future does not depend on the past.

Remarks. - With a discrete countable state space S, we can typically identify S with $\mathbb{N}$ or $\mathbb{Z}$.

- Note that a continuous state space can also be considered.

Definition 1.2. - We say that the chain "is in state i at time n", and makes a "transition from state i to state j" between time n and $n+1$, and we denote the transition probability as

$$\mathbb{P}(X_{n+1} = j \mid X_n = i) \triangleq p_{ij}(n) \triangleq p_{i \to j}(n).$$

- The matrix $(P)_{ij} = p_{ij}$ is called the transition matrix.

Basic properties. Note the following properties on transition probabilities:

(a) $0 \leq p_{ij} \leq 1, \ \forall\, i, j \in S.$

(b) $\sum_{j \in S} p_{ij} = 1.$

Proof. $\sum_{j \in S} p_{ij} = \sum_{j \in S} \mathbb{P}(X_{n+1} = j \mid X_n = i) = \sum_{j \in S} \frac{\mathbb{P}(X_{n+1}=j, X_n=i)}{\mathbb{P}(X_n=i)} = \frac{\mathbb{P}(X_n=i)}{\mathbb{P}(X_n=i)} = 1.$

Definition 1.3. The initial distribution of a Markov chain is given by $\mathbb{P}(X_0 = i) = \pi_i^{(0)} \ \forall\, i \in S.$

Definition 1.4. An homogeneous Markov chain is a Markov chain such that

$$\mathbb{P}(X_{n+1} = j \mid X_n = i) = p_{ij} = \mathbb{P}(X_1 = j \mid X_0 = i)$$

is independent of the time n.

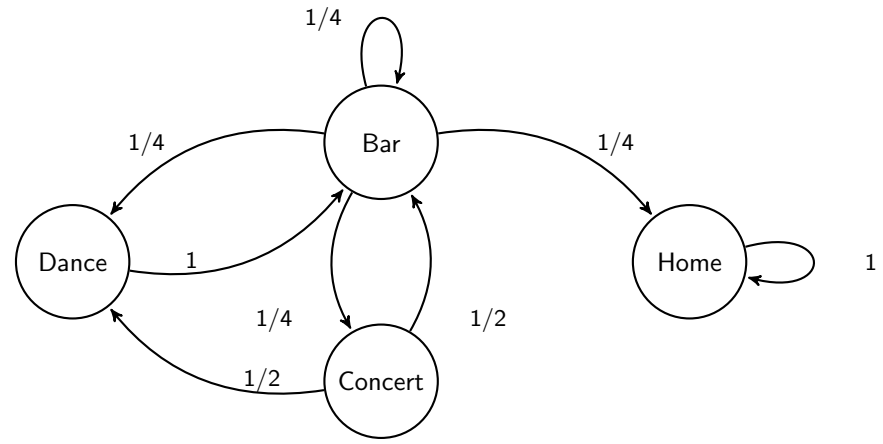
1.1.2 Graphical representation

Example 1.5. (Music festival)
The four states of a student in a music festival are $S = \{$"dancing", "at a concert", "at the bar", "back home"$\}$. Let assume that the student changes state during the festival according to the following transition matrix:

$$P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{matrix} \leftarrow Dance \\ \leftarrow Concert \\ \leftarrow Bar \\ \leftarrow Home \end{matrix}$$

$$\begin{matrix} \nearrow & \nearrow & \nearrow & \nearrow \\ Dance & Concert & Bar & Home \end{matrix}$$

Then this Markov chain can be represented by the following transition graph:



Example 1.6. (Simple symmetric random walk)
Let $(X_n, n \geq 1)$ be i.i.d. random variables such that $\mathbb{P}(X_n = +1) = \mathbb{P}(X_n = -1) = \frac{1}{2}$, and let $(S_n, n \geq 0)$ be defined as $S_0 = 0, S_n = X_1 + \ldots + X_n, \ \forall n \geq 1$. Then $(S_n, n \geq 0)$ is a Markov chain with state space $S = \mathbb{Z}$. Indeed:

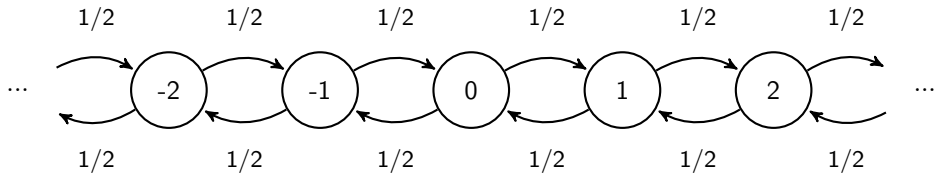
$$\begin{aligned}&\mathbb{P}(S_{n+1} = j | S_n = i, S_{n-1} = i_{n-1}, \ldots, S_0 = i_0) \\ &= \mathbb{P}(X_{n+1} = j - i | S_n = i, S_{n-1} = i_{n-1}, \ldots, S_0 = i_0) = \mathbb{P}(X_{n+1} = j - i)\end{aligned}$$

by the assumption that variables X_n are independent. The chain is moreover time-homogeneous as

$$\mathbb{P}(X_{n+1} = j - i) = \begin{cases} \frac{1}{2} & \text{if } |j - i| = 1 \\ 0 & \text{otherwise} \end{cases}$$

does not depend on n.

Here is the transition graph of the chain:



1.1.3 Time evolution and the Chapman-Kolmogorov equation

We restrict our attention to homogeneous chains and are interested in the study of the evolution of such Markov chains in their long time behavior (i.e. what happens when $n \to +\infty$). The evolution of a Markov chain can be computed since we know the initial distribution $\mathbb{P}(X_0 = i) = \pi_i^{(0)}$, and the transition probabilities $\mathbb{P}(X_1 = j \mid X_0 = i) = p_{ij}$.

We want to compute the distribution of a state at time n:

$$\pi_i^{(n)} = \mathbb{P}(X_n = i),$$

and compute the probability of possible paths (think of the random walk):

$$\mathbb{P}(\text{“path through } i_0, i_1, \ldots, i_n\text{”}) = \mathbb{P}(X_0 = i_0, \ldots, X_n = i_n).$$

To compute this last probability, the Markov property becomes important for simplification:

$$\begin{aligned}
\mathbb{P}(X_n = i_n, \dots, X_0 = i_0) &= \mathbb{P}(X_n = i_n \mid X_{n-1} = i_{n-1}, \dots, X_0 = i_0)\, \mathbb{P}(X_{n-1} = i_{n-1}, ..., X_0 = i_0) \\
&\overset{(a)}{=} \mathbb{P}(X_n = i_n \mid X_{n-1} = i_{n-1})\, \mathbb{P}(X_{n-1} = i_{n-1}, \dots, X_0 = i_0) \\
&\dots \\
&= \mathbb{P}(X_n = i_n \mid X_{n-1} = i_{n-1})\, \mathbb{P}(X_{n-1} = i_{n-1} \mid X_{n-2} = i_{n-2}) \dots \\
&\quad \dots \mathbb{P}(X_1 = i_1 \mid X_0 = i_0) \\
&= \pi^{(0)}_{i_0} p_{i_0 i_1} p_{i_1 i_2} p_{i_2 i_3} \cdots p_{i_{n-1} i_n}
\end{aligned}$$

where (a) comes from the Markov property.

$$\begin{aligned}
\implies \pi^{(n)}_{i_n} = \mathbb{P}(X_n = i_n) &= \sum_{i_0, \dots, i_{n-1} \in S} \mathbb{P}(X_n = i_n, \dots, X_0 = i_0) \\
&= \sum_{i_0, \dots, i_{n-1} \in S} \pi^{(0)}_{i_0} p_{i_0 i_1} p_{i_1 i_2} p_{i_2 i_3} \cdots p_{i_{n-1} i_n} \\
&= \sum_{i_0 \in S} \pi^{(0)}_{i_0} (P^n)_{i_0 i_n} \\
\implies \pi^{(n)}_{i_n} &= (\pi^{(0)} P^n)_{i_n}
\end{aligned}$$

where P^n is the n^{th} power of matrix P, and $\pi^{(0)}$ is the vector of initial probabilities.

To summarize, the study of Markov chains is essentially the study of Equation 1.1.3:

$$\pi^{(n)} = \pi^{(0)} P^n$$

This equation gives the time evolution of the probability distribution in the state. Notice also that $\pi^{(n)} = \pi^{(0)} P^n$ and $\pi^{(n-1)} = \pi^{(0)} P^{n-1}$ so that $\pi^{(n)} = \pi^{(n-1)} P$ or more generally:

$$\pi^{(n+m)} = \pi^{(n)} P^m$$

Theorem 1.7. The Chapman-Kolmogorov equation is simply the equation $P^{m+n} = P^m P^n$ expressed in components (with the transition probabilities):

$$\mathbb{P}(X_{n+m} = j | X_0 = i) = \sum_{k \in S} \mathbb{P}(X_n = k | X_0 = i) \mathbb{P}(X_{n+m} = j | X_n = k)$$

or

$$p^{n+m}_{ij} = \sum_{k \in S} p_{ik}(n) p_{kj}(m)$$

Remarks. The Chapman-Kolmogorov equation expresses a consistency condition that the transition probabilities of the Markov chain has to satisfy. This consistency condition is necessary to have a Markov chain but not sufficient (see e.g. from Grimmett page 219 example 14).

The study of Markov chain is mainly the study of long time behavior. Multiple questions come out:

- When does $\pi^* = \pi^* P$ has a solution ? This is the question of existence of a "stationary distribution" that does not evolve with time (sometime referred to as equilibrium distribution).
- If π^* exists, when is it the case that $\pi^{(n)} \underset{n \to +\infty}{\longrightarrow} \pi^*$? This is called ergodicity.
- And how fast does $\pi^{(n)}$ approach π^* ? This question is important in algorithmic applications. For example, we will see that the spectrum of eigenvalues of P plays an important role.

2 Classification of states of a Markov chain

These questions all concern the long time behavior of the Markov chain. It is useful before turning to these questions to classify the type of states (and chains). We begin with a few easy definitions and terminology and then go over the very important notions of recurrent states and transient states.

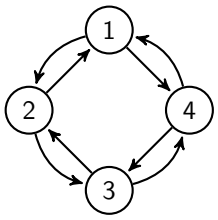
2.1 Quick definitions and terminology

Definition 2.1. A state $j \in S$ is said to be accessible from state i if $p_{ij}(n) > 0$ for some $n \geq 0$.

Definition 2.2. A state $i \in S$ is said to be absorbing if $p_{ii} = 1$. It means that once you reach an absorbing state, you stay there forever (e.g. state "Home" in Example 1.5).

Definition 2.3. A state $i \in S$ is said to be periodic with period d if $d(i) = \text{GCD}(n : p_{ii}(n) > 0)$. If $d = 1$, we say that the state is aperiodic.

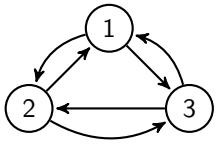
Example 2.4. Let us consider the following Markov chain:



Then $p_{11}(2) > 0$, $p_{11}(4) > 0$, $p_{11}(6) > 0, \ldots$ and $p_{11}(3) = p_{11}(5) = p_{11}(7) = \ldots = 0$. State 1 is thus a periodic state with period $d = 2$. And by symmetry all states have period $d = 2$.

Remarks. When all states are periodic, we say that the "chain is d-periodic".

Example 2.5. Note that for the chain



we have $p_{11}(2) > 0$, $p_{11}(3) > 0$, $p_{11}(4) > 0$, $p_{11}(5) > 0$, ... Here State 1 is aperiodic and we say that $d = 1$.

2.2 Recurrent and transient states

Definition 2.6. A state $i \in S$ is recurrent if $\mathbb{P}(X_n = i \text{ for some } n > 0 \mid X_0 = i) = 1$

Definition 2.7. A state $i \in S$ is transient if $\mathbb{P}(X_n = i \text{ for some } n > 0 \mid X_0 = i) < 1$

Note that in Example 1.5, state "Home" is recurrent (even absorbing), but all other states are transient.

Definition 2.8. Let $f_{ii}(n)$ be the probability that the first return time to i is n when we start at i:

$$\begin{aligned} f_{ii}(n) &= \mathbb{P}(X_1 \neq i, \ldots, X_{n-1} \neq i, X_n = i | X_0 = i) \\ f_{ii}(0) &= 0, \text{ by convention} \end{aligned}$$

Then we define the probability of eventual return as:

$$f_{ii} = \sum_{n=1}^{+\infty} f_{ii}(n)$$

Theorem 2.9. A rule to determine recurrent and transient states is:

(a) state $i \in S$ is recurrent iff $\sum_{n\geq 1} p_{ii}^{(n)} = +\infty$
(b) state $i \in S$ is transient iff $\sum_{n\geq 1} p_{ii}^{(n)} < +\infty$

Proof: We use a generating functions. We define

$$P_{ii}(s) = \sum_{n=0}^{+\infty} s^n p_{ii}(n),\ |s| < 1$$

and

$$F_{ii}(s) = \sum_{n=0}^{+\infty} s^n f_{ii}(n),\ |s| < 1$$

One can show that, for $|s| < 1$:

$$P_{ii}(s) = \frac{1}{1 - F_{ii}(s)}$$

This equation is enough to prove the theorem since, as $s \nearrow 1$ we have

$$P_{ii}(s) \nearrow +\infty \iff f_{ii} = F_{ii}(s = 1) = 1$$

. By Abel's theorem

$$\sum_{n=0}^{+\infty} p_{ii}(n) = \lim_{s\nearrow 1} P_{ii}(s)$$

Thus

$$\sum_{n=0}^{+\infty} p_{ii}(n) \iff f_{ii} = +1$$

Theorem 2.10 (Abel's Theorem)**.** If $a_i \geq 0$ for all i and $G(s) = \sum_{n\geq 0} s^n a_n$ is finite for $|s| < 1$, then

$$\lim_{s\nearrow 1} G(s) = \begin{cases} +\infty & \text{if} \sum_{n\geq 0} a_n = +\infty \\ \sum_{n\geq 0} a_n & \text{if} \sum_{n\geq 0} a_n < +\infty \end{cases}$$

Proof of formula $P_{ii}(s) = \frac{1}{1-F_{ii}(s)}$:
Let the events $A_m = \{X_m = i\}$, $B_r = \{X_1 \neq i, \ldots, X_{r-1} \neq i, X_r = i\}$ $\forall 1 \leq r \leq m$. The B_r are disjoint, thus:

$$\begin{aligned} \mathbb{P}(A_m \mid X_0 = i) &= \sum_{r=1}^{m} \mathbb{P}(A_m \cap B_r \mid X_0 = i) \\ &= \sum_{r=1}^{m} \mathbb{P}(A_m \mid B_r, X_0 = i)\mathbb{P}(B_r \mid X_0 = i) \\ &= \sum_{r=1}^{m} \mathbb{P}(A_m \mid X_r = i)\mathbb{P}(B_r \mid X_0 = i) \\ \implies p_{ii}(m) &= \sum_{r=1}^{m} f_{ii}(r)p_{ii}(m-r) \end{aligned}$$

Multiplying the last equation by s^m with $|s| < 1$ and sum over $m \geq 1$ we get:

$$\begin{aligned}
\implies \sum_{m\geq 1} s^m p_{ii}(m) &= \sum_{m\geq 1} s^m \sum_{r=1}^{m} f_{ii}(r) p_{ii}(m-r) \\
&= \sum_{m\geq 1} \sum_{k+l=m} f_{ii}(k) s^k p_{ii}(l) s^l \\
&= \sum_{k} f_{ii}(k) s^k \sum_{l} p_{ii}(l) s^l \\
\implies P_{ii}(s) - 1 &= P_{ii}(s) F_{ii}(s) \\
\implies P_{ii}(s) &= \frac{1}{1 - F_{ii}(s)}
\end{aligned}$$

Random Walks: WEEK 3

1 Positive/null-recurrence and stationary distribution

1.1 Preliminaries and positive/null-recurrence

Let $T_i = \inf\{n \geq 0 : X_n = i\}$ be the first return time of a random walk starting in $X_0 = i$.

Note that

$$\begin{aligned}\mathbb{P}(T_i < +\infty | X_0 = i) &= \sum_{m\geq 1} \mathbb{P}(T_i = m | X_0 = i) = \sum_{n\geq 1} \mathbb{P}(X_n = i, X_{n-1} \neq i, \ldots, X_0 \neq 1 | X_0 = i) \\ &= \mathbb{P}(X_n = i \text{ for some } n \geq 1 | X_0 = i) = f_{ii}\end{aligned}$$

Recall also the notation (from Lecture 2) for the probability that *the first return time to i is n*

$$f_{ii}(n) = \mathbb{P}(X_n = i, X_{n-1} \neq i, \ldots, X_0 \neq 1 | X_0 = i) = \mathbb{P}(T_i = n | X_0 = i)$$

For a recurrent state, $1 = f_{ii} = \sum_{n\geq 1} f_{ii}(n) = \mathbb{P}(T_i < +\infty | X_0 = i)$ and for a transient state, $f_{ii} < 1$. So for a transient state, the random variable T_i must have some mass at $T_i = +\infty$: $\mathbb{P}(T_i = +\infty | X_0 = i) > 0$. We define the following:

Definition 1.1. The **mean recurrence time** is given by

$$\mu_i \equiv \mathbb{E}(T_i | X_0 = i) = \sum_{n\geq 1} n\, \mathbb{P}(T_i = n | X_0 = i) = \sum_{n\geq 1} n\, f_{ii}(n).$$

Can we compute this quantity? Let us to that end distinguish two cases:

Transient states: We know that for transient states the event $T_i = +\infty$ occurs with some strictly positive probability $P(T_i = +\infty | X_o = i) > 0$. We therefore have that $\mu_i = \mathbb{E}(T_i | X_0 = i) = +\infty$.

Recurrent states: By definition $\sum_{m\geq 1} f_{ii}(m) = 1$ for recurrent states. But $\sum_{n\geq 1} n f_{ii}(n)$ is not necessarily finite. So even though you know for sure the chain will return to state i at some finite time, the average recurrence time may be infinite. Based on this, we say that any recurrent state is either one of the following:

- **Positive-recurrent** if $\mu_i < +\infty$.
- **Null-recurrent** if $\mu_i = +\infty$.

1.2 Irreducible chains

- We say that $i \to j$ if $\exists n \geq 0$ such that $p_{i\to j}(n) > 0$. In words, we say that i *communicates* with j or that j is *accessible* from i.
- We say that $i \longleftrightarrow j$ if $i \to j$ and $j \to i$. In words, i and j *intercommunicate*. Note, however, that this implies that $\exists n \geq 0$ s.t $p_{i\to j}(n) > 0$ and $\exists m \geq 0$ s.t. $p_{j\to i}(m) > 0$ (but it is not necessary that $n = m$ holds).
- If $i \longleftrightarrow j$ then these states are reflexive, symmetric and transitive, meaning:
  - Reflexivity: $i \longleftrightarrow i$, because we always have that at least $p_{ii}(0) = 1$ holds.
  - Symmetry: if $i \longleftrightarrow j$ holds then -of course- also $j \longleftrightarrow i$ holds.

– Transitivity: If $i \longleftrightarrow k$ and $k \longleftrightarrow j$, then $i \longleftrightarrow j$. This statement follows from the Chapman-Kolmogorov equation:

$$p_{i\to j}(n+m) = \sum_{\ell\in\mathcal{S}} p_{i\to\ell}(n)\, p_{\ell\to j}(m) \geq p_{i\to k}(n)\, p_{k\to j}(m).$$

One can always pick an n and m such that the right hand side is strictly positive and hence $\exists n+m$ such that $p_{i\to j}(n+m) > 0$. the same argument shows that there exists n', m' such that $p_{j\to i}(n'+m') > 0$. Thus i and j intercommunicate.

- We can partition the state space $\mathcal{S}$ in *equivalence classes*. For example, the music festival example of Lecture 2 had two such classes: $\{\text{Home}\}$ and $\{\text{Bar}, \text{Dancing}, \text{Concert}\}$. Note that $\{\text{Home}\}$ is here trivially positive-recurrent state (in fact it is absorbing: once you are at home you return to home at every time step with probability one) and that $\{\text{Bar}, \text{Dancing}, \text{Concert}\}$ are transient states (fortunately or unfortunately ?).

Having stated these notions, we are ready to define irreducible chains as follows:

Definition 1.2. A Markov chain is said to be **irreducible** if it has only one equivalence class, i.e., $\forall i, j \in \mathcal{S}\ \exists n, m$ such that $p_{i\to j}(n)\, p_{j\to i}(m) > 0$.

In other words, in an irreducible Markov chain every state is accessible from every state.

Proposition 1.3. Within an equivalence class of a Markov chain or for an irreducible Markov chain it holds that

1. All states i have the same period.
2. All states i are recurrent or all states are transient.
3. If all states i are recurrent, then either they are all null-recurrent or they are all positive-recurrent.

Proof of point 2. Take two states in the same equivalence class, $i \longleftrightarrow j$. Then, from the Chapman-Kolmogorov equation, we deduce the inequality

$$p_{ii}(n+t+m) \geq \underbrace{p_{i\to j}(n)}_{>0}\, p_{j\to j}(t)\, \underbrace{p_{j\to i}(m)}_{>0} \geq \underbrace{\alpha}_{>0} p_{jj}(t).$$

If i is transient, then $\sum_t p_{ii}(t) < +\infty$ (criterion proved in Lecture 2) and thus $\sum_t p_{jj}(n+t+m) < +\infty$ so j is also transient. To complete the proof, we note that the roles of i and j can be interchanged. This way we also get that "if j is transient, then i is transient".

The proof of 1 is similar. The proof of 3 requires more tools that we don't quite have at this point. □

Lemma 1.4. If a Markov chain has a finite state space $\mathcal{S}$ and is irreducible, then all its states are (positive-)recurrent.

Proof. We have the following property:

$$\lim_{n\to\infty} \sum_{j\in\mathcal{S}} p_{i\to j}(n) = 1,$$

but our state space is finite so we can interchange the order:

$$\sum_{j\in\mathcal{S}} \lim_{n\to+\infty} p_{i\to j}(n) = 1. \tag{1}$$

We continue by contradiction. Assume that all $j \in \mathcal{S}$ are transient, then $p_{jj}(n) \to 0$ as $n \to \infty$. Even stronger, we proved in Homework 2 that $p_{ij}(n) \to 0$ as well. This contradicts (1):

$$\sum_{j \in \mathcal{S}} \underbrace{\lim_{n \to +\infty} p_{i \to j}(n)}_{\to 0} \neq 1.$$

So if there is a j that is recurrent and the chain is irreducible, then all $j \in \mathcal{S}$ must be recurrent.

The proof that all states are in fact positive-recurrent requires more tools that we don't yet have. □

1.3 Stationarity

Definition 1.5. A distribution π^* is called **stationary** if it satisfies the equation $\pi^* = \pi^* P$.

It follows immediately that any stationary distribution also satisfies $\pi^* = \pi^* P^k$ for any $k \geq 0$. In particular, if we initialize a chain in the stationary distribution $\pi^{(0)} = \pi^*$ then at any time n, $\pi^{(n)} = \pi^*$ (and this is why π^* is called stationary).

Discussion: For systems with a *finite state space* one can show that the finite stochastic matrix P always has an eigenvalue 1 with left eigenvector with non-negative components $\pi_i \geq 0$[1]. But it may not be unique (as will be clear from the discussion below). Uniqueness requires more conditions on P. For example if P has all its elements strictly positive or if there exists $N \geq 1$ such that P^N has all its elements strictly positive then the standard forms of the Perron-Frobenius theorem imply unicity. However this is not the most general condition. The theory of Markov chains in finite state spaces can be developed through the Perron-Frobenius theorems (at various levels of sophistication) but this is not the route we take in this class because we are also interested in *infinite* (countable) state spaces.

An important theorem is the following.

Theorem 1.6 (existence and uniqueness of a stationary distribution)**.** Consider an irreducible Markov chain. It has a stationary distribution if and only if the chain is positive-recurrent. Moreover, this distribution is unique and takes on the value

$$\pi_i^* = \frac{1}{\mu_i} = \frac{1}{\mathbb{E}(T_i | X_0 = i)}$$

Remark 1.7. Note that μ_i is finite because we assume the chain is positive-recurrent.

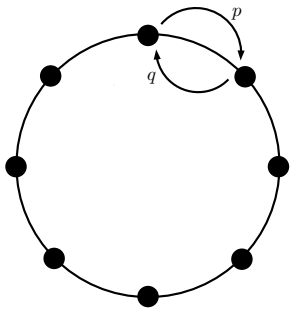
Remark 1.8. Take an irreducible Markov chain with a *finite state space*. Then by Lemma 1.4, we know it must be positive-recurrent. Thus an irreducible Markov chain on a finite state space has a unique stationary distribution $\pi_i^* = 1/\mu_i$.

Remark 1.9. This theorem is very powerful. Indeed suppose you have a chain and you know that it is irreducible. With an infinite state space, it might be difficult to prove directly that it is positive-recurrent, but it might be easier to compute the stationary distribution. Then you immediately can conclude that it is necessarily positive-recurrent.

The proof of the theorem is not easy and we do not go through it here. We rather try to motivate the theorem through the following discussion.

[1] It is clear that the vector with all ones is a right eigenvector with eigenvalue 1. But the existence of a left eigenvector with non-negative components for the same eigenvalue is not obvious.

Example 1.10. Consider the following random walk on a circle, a finite state space:



The stochastic matrix P of the walk is of size $|\mathcal{S}| \times |\mathcal{S}|$ and looks as follows:

$$\begin{pmatrix} 0 & p & 0 & \cdots & \cdots & q \\ q & 0 & p & & & \vdots \\ 0 & q & 0 & & & \vdots \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & 0 & p \\ p & \cdots & & \cdots & q & 0 \end{pmatrix}$$

One can easily verify that $\pi_i^* = \frac{1}{|S|}$ is the stationary distribution. This example suggests that on $\mathbb{Z}^d$ the random walk has no stationary distribution because $|S| \to \infty$ and thus $\pi_i^* = \frac{1}{|S|} \to 0$ would not yield a valid probability distribution. This is true. Indeed the random walk in $\mathbb{Z}^d$ is irreducible and *null recurrent* for $d = 1, 2$ and *transient* for $d \geq 3$, so by the above theorem, it cannot have a stationary distribution.

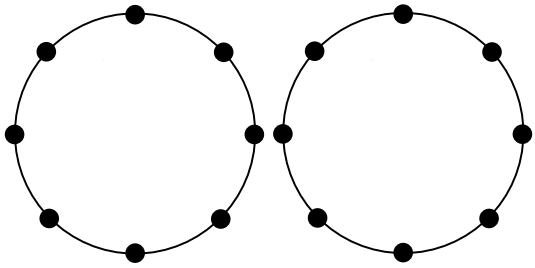
When one verifies that $\pi_i^* = \frac{1}{|S|}$ is the stationary distribution in the example above, one sees that this works because the *columns* sum to one (recall that for a stochastic matrix the *rows* always sum to one). This motivates the following definition.

Definition 1.11. A doubly stochastic matrix is a matrix $P = [p_{ij}]$ with $p_{ij} \geq 0$ of which all the rows and all the columns sum to one.

One can easily see that the mechanism of the example above generalizes to any chain on a finite state space with a doubly stochastic matrix: in this case, $\pi_i^* = \frac{1}{|S|}$ is a stationary distribution because the *columns* sum to one.

Now what about the unicity of the stationary distribution ? The following simple situation suggests that we don't have unicity for reducible chains.

Example 1.12. Consider two separate circles:



The state space is the union of two disconnected finite state spaces, $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$. This Markov chain is *not* irreducible! Its transition matrix shows the following block structure:

$$P = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix}. \tag{2}$$

and consequently $\pi^* = \pi^* P$ has many solutions. An example are all distribution that are computed as follows:

$$\pi^* = \left(\frac{\alpha}{|\mathcal{S}_1|}, \cdots, \frac{\alpha}{|\mathcal{S}_1|}, \frac{\beta}{|\mathcal{S}_2|}, \cdots, \frac{\beta}{|\mathcal{S}_2|} \right), \tag{3}$$

where $\alpha + \beta = 1$ and $\alpha, \beta \geq 0$. The first $|\mathcal{S}_1|$ components correspond to the first circle, the last $|\mathcal{S}_2|$ correspond to the second. The uniform stationary distribution corresponds to $\alpha = \frac{|\mathcal{S}_1|}{|\mathcal{S}_1|+|\mathcal{S}_2|}$ and $\beta = \frac{|\mathcal{S}_2|}{|\mathcal{S}_1|+|\mathcal{S}_2|}$. Also note that extreme cases, such as $\{\alpha = 0, \beta = 1\}$ and $\{\alpha = 1, \beta = 0\}$ are also perfectly valid stationary distributions. The general stationary distributions are convex combinations of the two extremal ones.

Random Walks: WEEK 4

1 Introduction

Let us first recall an important theorem from last time.

Theorem 1.1 (Stationary distribution)**.** Consider an irreducible Markov chain with transition matrix P. It has a stationary distribution, i.e., a state distribution π^* satisfying $\pi^* = \pi^* P$, if and only if the chain is positive-recurrent.

Let $\pi^{(n)}$ denotes the state distribution of the Markov chain at time n. We are interested in the following question: for any given initial distribution $\pi^{(0)}$, does it hold that $\pi^{(n)} \to \pi^*$ when n goes to infinity?

For a Markov chain (X_n), we use the notations $p_{ij} := \mathbb{P}\{X_1 = j | X_0 = i\}$, $p_{ij}(n) := \mathbb{P}\{X_n = j | X_0 = i\}$.

2 The ergodic theorem

Recall that a Markov chain is said to be aperiodic if $gcd\{n : p_{ii}(n) > 0\} = 1$ for any state i.

Definition 2.1 (Ergodicity)**.** A chain is said to be ergodic if it is irreducible, aperiodic and positive recurrent.

We will prove the following theorem.

Theorem 2.2 (Ergodic theorem)**.** An ergodic Markov chain admits a unique stationary distribution π^* by Theorem 1.1. This distribution is also a "limiting distribution" in the sense:

$$\lim_{n\to\infty} \pi_i^{(n)} = \pi_i^*, \forall i \in S$$

Remark 2.3. The state distribution is given by $\pi^{(n)} = \pi^{(0)} P^n$ at any finite time n. The above theorem implies that the limiting distribution does not depend on the initial distribution $\pi^{(0)}$ as $n \to \infty$.

We give an example before starting with the proof.

Example 2.4 (Aperiodicity matters)**.** Consider a Markov chain with two states $\{0, 1\}$. The transition probability is given by

$$p_{00} = 1 - p, \quad p_{01} = p, \quad p_{11} = 1 - q, \quad p_{10} = q$$

It is easy to show that the stationary distribution π^* satisfying

$$\pi^* = \pi^* P = \pi^* \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

has the unique solution $\pi^* = (\frac{q}{q+p}, \frac{p}{q+p})$. As a result of Theorem 2.2, this Markov chain has the limiting distribution π^* for any initial distribution if it is ergodic (i.e. irreducible, aperiodic, postive recurrent). The caveat here is that the assumption on the aperiodicity of the Markov chain is not always satisfied for all p and q.

Suppose $p = q = 1$ and the initial distribution $\pi^{(0)} = (\mathbb{P}\{s = 0\}, \mathbb{P}\{s = 1\}) = (0, 1)$, meaning the chain starts in state 1 with probability one. This Markov chain is not aperiodic. Indeed, we have in this case $\pi^{(1)} = (1, 0)$, $\pi^{(2)} = (0, 1)$, $\pi^{(3)} = (1, 0)$ and so on. We see the state 1 has a positive probability only at even times, i.e., $gcd\{n : p_{11}(n) > 0\} = 2$ thus the chain is periodic with period 2. As a consequence, Theorem 2.2 does not apply to this Markov chain in this case. Indeed when initialized to $\pi^{(0)} = (0, 1)$ the distribution $\pi^{(n)}$ jumps between the two distributions $\pi^{(n)} = (0, 1), \pi^{(n+1)} = (1, 0)$, even for arbitrarily large n, and does not converge to the stationary distribtion π^*.

3 A proof of the ergodic theorem

A proof technique called *coupling* is used for the proof. We first give the main idea of this technique.

3.1 Coupling

Consider an ergodic Markov chain $(X_n, n \geq 0)$. Take a "copy" of the above chain called $(Y_n, n \geq 0)$, i.e., the two chains X and Y are two independent Markov chains having the same state space S and transition matrix P. Then construct a "coupled chain" $(Z_n = (X_n, Y_n), n \geq 0)$.

We have the following properties for the chain Z:

- **The chain Z is a Markov chain** with the transition probability given by

$$\begin{aligned} p_{ij\to k\ell} &:= \mathbb{P}\{Z_{n+1} = (k,\ell) | Z_n = (i,j)\} \\ &= p_{ik} p_{j\ell} \end{aligned}$$

  where $p_{ik} = \mathbb{P}\{X_{n+1} = k | X_n = i\}$. This fact is easy to verify using the Markovity and independence of X and Y.

- **The chain Z is irreducible.**

  Proof. Using the irreducibility and aperiodicity of (X_n), we show in the Appendix that (X_n) satisfies the following property

$$\forall i, j \in S, \exists N(i,j), \text{such that for all } n \geq N(i,j), p_{ij}(n) > 0$$

  Obviously (Y_n) also satisfies this property. Now for any $i, j, k, \ell \in S$, choose $m > \max\{N(i,j), N(k,\ell)\}$, we will have

$$\mathbb{P}\{Z_m = (j,\ell) | Z_0 = (i,k)\} = p_{ij}(m)\, p_{k\ell}(m) > 0$$

  □

- **The chain Z is positive recurrent.**

  Proof. By assumption (X_n) is irreducible and positive-recurrent, hence there exists a stationary distribution π^* for this chain by Theorem 1.1. So does it for the chain (Y_n). Now define a distribution ν^* on $S \times S$ as

$$\nu^*_{i,j} = \pi^*_i \pi^*_j \quad (i,j) \in S \times S$$

  It is easy to check that $\nu^*_{i,j}$ defined above is indeed a stationary distribution for the chain (Z_n), i.e.

$$\nu^*_{i,j} = \sum_{k,\ell} \nu^*_{k,\ell} p_{k\ell \to ij}$$

  using the fact that π^* is a stationary distribution for X, Y. Now use Theorem 1.1 again: since Z is irreducible and has a stationary distribution, it must be positive recurrent. □

Let $Z_0 = (X_0, Y_0) = (i,j)$ and define

$$T_s = \min\{n \geq 1 : Z_n = (s,s)\}$$

for some $s \in S$. In words, this is the first time the trajectories of (X_n) and (Y_n) meet at the state s. The central idea of coupling is that after this meeting time, the distributions of (X_n) and (Y_n) should be the same. Formally we have

Lemma 3.1 (Coupling)**.** Let X, Y, Z and T_s as defined above,. It holds that

$$\mathbb{P}\{X_n = k, T_s < n | Z_0 = (i,j)\} = \mathbb{P}\{Y_n = k, T_s < n | Z_0 = (i,j)\}$$

for any $k \in S$.

Notice that given $Z_0 = (i,j)$, we need T_s to be finite in order to make effective use of the above statement. This is justified by the following lemma which, along with the above coupling lemma, will be proved later.

Lemma 3.2. Let Z be an irreducible, recurrent Markov chain and T_s defined as above, then

$$\mathbb{P}\{T_s < \infty | Z_0 = (i,j)\} = 1$$

This means the chains X, Y will eventually become "coupled" with probability one.

3.2 The Proof of the Ergodic Theorem

Equipped with the tool of coupling, we are ready to prove the theorem. Recall $p_{ik}(n) := \mathbb{P}\{X_n = k | X_0 = i\}$, we first show that

$$|p_{ik}(n) - p_{jk}(n)| \to 0 \text{ as } n \to \infty \tag{1}$$

for any $i, j, k \in S$. Given any $s \in S$, we have

$$\begin{aligned}
p_{ik}(n) &= \mathbb{P}\{X_n = k | X_0 = i\} \\
&= \mathbb{P}\{X_n = k | X_0 = i, Y_0 = j\} \quad \text{using independence of } X, Y \\
&= \mathbb{P}\{X_n = k, T_s < n | X_0 = i, Y_0 = j\} + \mathbb{P}\{X_n = k, T_s \geq n | X_0 = i, Y_0 = j\} \\
&= \mathbb{P}\{Y_n = k, T_s < n | X_0 = i, Y_0 = j\} + \mathbb{P}\{X_n = k, T_s \geq n | X_0 = i, Y_0 = j\} \quad \text{coupling, Lemma 3.1} \\
&\leq \mathbb{P}\{Y_n = k | X_0 = i, Y_0 = j\} + \mathbb{P}\{T_s \geq n | X_0 = i, Y_0 = j\} \\
&= \mathbb{P}\{Y_n = k | Y_0 = j\} + \mathbb{P}\{T_s \geq n | X_0 = i, Y_0 = j\} \quad \text{using independence of } X, Y
\end{aligned}$$

Hence we have

$$p_{ik}(n) - p_{jk}(n) \leq \mathbb{P}\{T_s \geq n | X_0 = i, Y_0 = j\}$$

The result in Lemma 3.2 can be rewritten as

$$\sum_{n \geq 1} \mathbb{P}\{T_s = n | Z_0 = (i,j)\} = 1$$

It implies that we must have $\mathbb{P}(T_s \geq n | Z_0(i,j)) \to 0$ for n large enough, otherwise the sum would not converge to 1. This proves we have $p_{ik}(n) - p_{jk}(n) \leq 0_+$ as $n \to \infty$. Finally, notice that the above steps goes through if i, j are switched, i.e. $p_{ik}(n) - p_{jk}(n) \leq 0_+$ as $n \to \infty$. Hence the claim $|p_{ik}(n) - p_{jk}(n)| \to 0$ is proved.

Consider the stationary distribution π^*,

$$\begin{aligned}
\lim_n (\pi_k^* - p_{ik}(n)) &= \lim_n \sum_j \pi_j^* (p_{jk} - p_{ik}(n)) \\
&= \sum_j \pi_j^* \lim_n (p_{jk} - p_{ik}(n)) \qquad (2) \\
&= \sum_j \pi_j^* \cdot 0 \quad \text{using Eq. (1)} \\
&= 0 \qquad (3)
\end{aligned}$$

In the second equality we switched the limit and sum thanks to Lebesgue's dominated convergence theorem: $|\pi_j^*(p_{jk} - p_{ik}(n))| \leq 2\pi_j^*$ which is independent of n and summable.

Finally we have

$$\begin{aligned}
\lim_n \pi_k^{(n)} &= \lim_n \sum_i p_{ik}(n)\pi_i^{(0)} \\
&= \sum_i \lim_n p_{ik}(n)\pi_i^{(0)} \\
&= \sum_i \pi_k^*\pi_i^{(0)} \quad \text{using Eq. (3)} \\
&= \pi_k^*
\end{aligned}$$

Again, in the second equality we switched the limit and sum thanks to Lebesgue's dominated convergence theorem: $p_{ik}(n)\pi_i^{(0)} \leq \pi_i^{(0)}$ which is independent of n and summable. This completes the proof of the ergodic theorem. □

4 Proofs of lemmas

We present the detailed proofs of the previously used lemmas in this section.

4.1 Proof of Lemma 3.1

We start with the L.H.S of the expression

$$\begin{aligned}
\mathbb{P}\{X_n = k, T_s < n | Z_0 = (i,j)\} &= \sum_\ell \mathbb{P}\{X_n = k, Y_n = \ell, T_s < n | Z_0 = (i,j)\} \\
&= \sum_\ell \sum_{m=1}^{n-1} \mathbb{P}\{X_n = k, Y_n = \ell | T_s = m, Z_0 = (i,j)\}\, \mathbb{P}\{T_s = m | Z_0 = (i,j)\} \\
&= \sum_\ell \sum_{m=1}^{n-1} \mathbb{P}\{X_n = k, Y_n = \ell | T_s = m, Z_0 = (i,j)\}\, \mathbb{P}\{T_s = m | Z_0 = (i,j)\}
\end{aligned}$$

Notice we have

$$\begin{aligned}
&\mathbb{P}\{X_n = k, Y_n = \ell | T_s = m, Z_0 = (i,j)\} \\
&= \mathbb{P}\{X_n = k, Y_n = \ell | X_m = s, Y_m = s, Z_0 = (i,j), (X_i, Y_i) \neq (s,s) \text{ for any } i < m\} \\
&= \mathbb{P}\{X_n = k, Y_n = \ell | X_m = s, Y_m = s\} \quad \text{Markovity of } Z = (X,Y)
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{P}\{X_n = k, T_s < n | Z_0 = (i,j)\} &\leq \sum_\ell \sum_{m=1}^{n-1} \mathbb{P}\{X_n = k, Y_n = \ell | X_m = s, Y_m = s\}\, \mathbb{P}\{T_s = m | Z_0 = (i,j)\} \\
&= \sum_{m=1}^{n-1} \sum_\ell p_{sk}(n-m) p_{s\ell}(n-m)\, \mathbb{P}\{T_s = m | Z_0 = (i,j)\} \\
&= \sum_{m=1}^{n-1} p_{sk}(n-m)\, \mathbb{P}\{T_s = m | Z_0 = (i,j)\}
\end{aligned}$$

Notice we could have the same derivation to obtain

$$\mathbb{P}\{Y_n = k, T_s < n | Z_0 = (i,j)\} = \sum_{m=1}^{n-1} p_{sk}(n-m)\, \mathbb{P}\{T_s = m | Z_0 = (i,j)\}$$

simply replacing X by Y. This completes the proof. □

4.2 Proof of Lemma 3.2

Let m be the smallest time such that $p_{ss\to ij}(m) > 0$. By irreducibilty we know that a finite such time exists. This means that the event that the chain goes from (s,s) to (i,j) directly has non zero probability. Now, we claim that the following inequality holds:

$$\underbrace{p_{ss\to ij}(n)}_{\text{Z leaves } (s,s) \text{ to } (i,j)} \cdot \underbrace{(1-\mathbb{P}(T_s<\infty)|Z_0=(i,j))}_{Z \text{ never goes back to } (s,s)} \leq \underbrace{1-\mathbb{P}(T_s<\infty|Z_n=(s,s))}_{Z \text{ levaves } (s,s) \text{ and never comes back}}$$

The RHS is the probability of the event that Z leaves (s,s) and nevers come back; LHS is the probability of the event Z goes from (s,s) to (i,j) directly, then leaves (i,j) but never comes back to (s,s). Obviously the event of the LHS is included in the event of the RHS (the event of the left hand side implies the event of the right hand side). Thus the probability of the LHS is smaller than the probability of the RHS.

The RHS equals zero because Z is recurrent and the first term on the LHS is non-zero because Z is irreducible. This means we must have

$$1-\mathbb{P}(T_s<\infty)|Z_0=(i,j))=0$$

which proves the lemma. □

5 Appendix

In the proof of irreducibility of Z we made use of the following technical statement:

Lemma 5.1. Let X be irreducible and aperiodic. Then

$$\forall i,j\in S, \exists N(i,j), \text{such that for all } n\geq N(i,j), p_{ij}(n)>0$$

Notice that this is stronger than pure irreducibility because we want $p_{ij}(n) > 0$ for all large enough n (given i,j). This is why aperiodicity is needed. The proof is slightly technical (and has not much to do with probability); but we present it here for completeness.

Proof. For an irreducible aperiodic chain we have for all states $1 = \gcd\{n : p_{jj}(n) > 0\}$. Thus we can find a set of integers $r_1, \cdots, r_k$ such that $p_{jj}(r_k) > 0$ and $1 = \gcd\{r_1, \ldots, r_k\}$.

Claim: for any $r > M$ with M large enough (depending possibly on $r_1, \ldots, r_k$) we can find integers $a_1, \ldots, a_k \in \mathbb{N}$ that are solution of

$$r = a_1 r_1 + \cdots + a_k r_k$$

This claim will be justified at the end of the proof not to disrupt the flow of the main idea.

Since the chain is irreducible, for all i,j we can find some time m such that $p_{ij}(m) > 0$. By the Chapman-Kolmogorov equation we have

$$\begin{aligned} p_{ij}(r+m) &= \sum_{k\in\mathcal{S}} p_{ik}(m)p_{kj}(r) \\ &\geq p_{ij}(m)p_{jj}(r) \end{aligned}$$

Using Chapman-Kolmogorov again and again,

$$\begin{aligned} p_{jj}(r) &= p_{jj}(a_1r_1+\cdots+a_kr_k) \\ &= \sum_{\ell_1,\cdots,\ell_k\in\mathcal{S}} p_{j\ell_1}(a_1r_1)p_{\ell_1\ell_2}(a_2r_2\cdots p_{\ell_k j}(a_kr_k) \\ &\geq p_{jj}(a_1r_1)p_{jj}(a_2r_2)\cdots p_{jj}(a_kr_k) \\ &\geq (p_{jj}(r_1))^{a_1}(p_{jj}(r_2))^{r_2}\cdots(p_{jj}(r_k))^{a_k} \end{aligned}$$

We conclude that

$$p_{ij}(r+m) \geq p_{ij}(m)(p_{jj}(r_1))^{a_1}(p_{jj}(r_2))^{r_2}\cdots(p_{jj}(r_k))^{a_k} > 0$$

We have obtained that for all $r > M$, M large enough we have $p_{ij}(r+m) > 0$. Thus we have that $p_{ij}(n) > 0$ *for all* $n > N(i,j) \equiv M + m$. Note that in the above construction M depends on j and m depends on i, j.

It remains to justify the *claim*. For simplicity we do this for $k = 2$. Let $\gcd(a,b) = 1$. We show that for $c > ab$ the equation $c = ax_0 + by_0$ has non-negative solutions (x_0, y_0). If we were allowing negative integers this claim would follow from Bézout theorem. But here we want non-negative solutions (and maybe that we dont remenber Bézout theorem anyway?) so we give an explicit argument.

Take $c = ax + by \bmod a$. Then $c \equiv by \bmod a$. Since a and b are coprime the inverse b^{-1} exists mod a, so $y \equiv b^{-1}c \bmod a$. Take the smallest integer $y_0 = b^{-1}c \bmod a$ and try now to solve $c = ax + by_0$ for x. Note that $c - by_0 > 0$ because $c > ba$. Note also that $c - by_0$ is divisible by a (since $y_0 - b^{-1}c \equiv 0 \bmod a$). Therefore doing the Euclidean division of $c - by_0$ by a we find x_0 non negative and satisfing $c - by_0 = ax_0$. We have thus found our solution (x_0, y_0).

□

Random Walks: WEEK 5

1 Hitting Times

Let $(X_n, n \geq 0)$ be a time-homogeneous Markov Chain with state space S. Let $i \in S, A \subset S$ (note that A is *any* subset of S) : $H_A = \inf\{n \geq 0 : X_n \in A\}$.

Example 1.1.

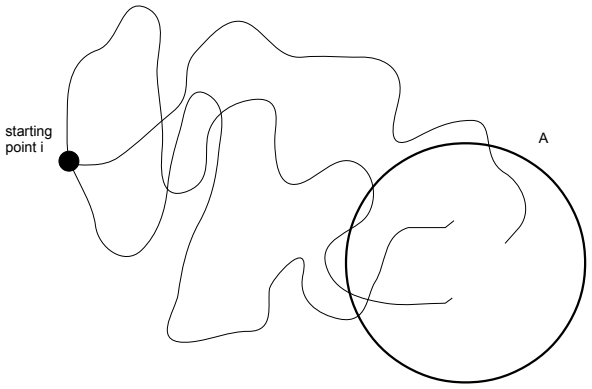


Figure 1: Three realisations of a random walk starting at point i arriving in subset A

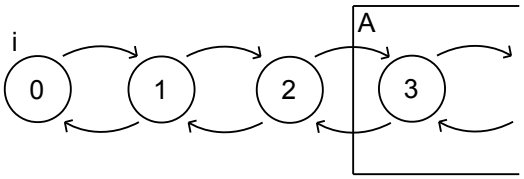


Figure 2: Markov chain for which it's easier to compute μ_{iA}

Definition 1.2 (mean Hitting Time)**.** $\mu_{iA} = \mathbb{E}(H_A | X_0 = i)$.

Remark 1.3. If $\mathbb{P}(H_A < +\infty | X_0 = i) < 1$, then $\mu_{iA} = +\infty$

Example 1.4. $A = \{0\}$ and $\mathbb{P}(H_A < +\infty | X_0 = 1) < 1$

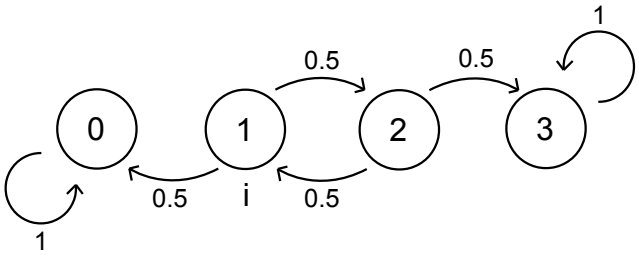


Figure 3: Markov chain with two absorbing states

Theorem 1.5. Assume that $\mathbb{P}(H_A < +\infty | X_0 = i) = 1 \; \forall i \in S$. Then the row vector $(\mu_{iA}, i \in S)$ is the *minimal*$^{(*)}$ *non-negative* solution of:

$$\begin{cases} \mu_{iA} & = 0 \quad \text{if } i \in A \\ \mu_{iA} & = 1 + \sum_{j \in A^c} p_{ij} \mu_{jA} \quad \text{if } i \in A^c \end{cases} \tag{1}$$

$(*)$ i.e. if $(\nu_{iA}, i \in S)$ is the solution of the same equation, then $\nu_{iA} \geq \mu_{iA}, \forall i \in S$.

Proof. (without the minimality condition)

- if $i \in A$, $\mu_{iA} = 0$: trivial, by definition of H_A
- if $i \in A^c$, we have to develop the expression of μ_{iA}:

$$\begin{aligned}
\mu_{iA} &= \mathbb{E}(H_A|X_0 = i) \\
&= \sum_{n\geq 0} n\,\mathbb{P}(H_A = n|X_0 = i) \\
&= \underbrace{0 + 1\cdot\mathbb{P}(X_1 \in A|X_0 = i)}_{\text{first two terms}} + \sum_{n\geq 2} n\,\mathbb{P}(H_A = n|X_0 = i) \\
&= \mathbb{P}(X_1 \in A|X_0 = i) + \sum_{n\geq 2} n \sum_{j\in A^c} \mathbb{P}(H_A = n, X_1 = j|X_0 = i) \\
&= \mathbb{P}(X_1 \in A|X_0 = i) + \sum_{m\geq 1} (m+1) \sum_{j\in A^c} \underbrace{\mathbb{P}(H_A = m+1|X_1 = j, X_0 = i)}_{(*)} \underbrace{\mathbb{P}(X_1 = j|X_0 = i)}_{p_{ij}} \\
&= \mathbb{P}(X_1 \in A|X_0 = i) + \sum_{m\geq 1} (m+1) \sum_{j\in A^c} \mathbb{P}(H_A = m|X_0 = j)p_{ij} \\
&= \mathbb{P}(X_1 \in A|X_0 = i) + \sum_{j\in A^c} p_{ij} \left[\underbrace{\sum_{m\geq 1} \mathbb{P}(H_A = m|X_0 = j)}_{=\mathbb{P}(H_A<\infty|X_0=j)=1} + \underbrace{\sum_{m\geq 1} m\,\mathbb{P}(H_A = m|X_0 = j)}_{=\mathbb{E}(H_A|X_0=j)=\mu_{jA}} \right] \\
&= \underbrace{\mathbb{P}(X_1 \in A|X_0 = i) + \underbrace{\sum_{j\in A^c} p_{ij}\cdot 1}_{=\mathbb{P}(X_1\in A^c|X_0=i)}}_{=1} + \sum_{j\in A^c} p_{ij}\mu_{jA} \\
&= 1 + \sum_{j\in A^c} p_{ij}\mu_{jA}
\end{aligned}$$

$(*) \overset{\text{Markov}}{=} \mathbb{P}(H_A = m+1|X_1 = j) \overset{\text{time hom.}}{=} \mathbb{P}(H_A = m|X_0 = j)$ □

Example 1.6 (Symmetric random walk on the circle)**.** Recall that it has the following stransition matrix

$$P = \begin{bmatrix} 0 & 1/2 & 0 & \dots & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 & \dots & 0 \\ 0 & 1/2 & \ddots & \ddots & & \vdots \\ \vdots & 0 & \ddots & & & 0 \\ 0 & \dots & & & 0 & 1/2 \\ 1/2 & 0 & \dots & 0 & 1/2 & 0 \end{bmatrix}$$

(and that $\pi_i^* = 1/N$)

Let $T_0 = \inf\{n \geq 1 : X_1 = 0\}$ be the first return time to zero. What about the value of $\mathbb{E}(T_0|X_0 = 0)$? We will use the result of Theorem 1.5 in order to avoid tedious computations.

$$\mathbb{E}(T_0|X_0 = 0) = 1 + \mathbb{E}(H_0|X_0 = 1) = 1 + \mu_{10}$$

Let $i \in S$ and $A = \{0\}$. Is it true that $\mathbb{P}(H_A < +\infty|X_0 = i) = 1$? Yes, because S is finite and the chain is irreducible, so it is positive-recurrent.

We will apply the previous theorem and this implies solving a set of linear equations:

$$\left[\begin{array}{ll} \mu_{00} & = 0 \\ \mu_{10} & = 1 + \sum_{j\neq 0} p_{ij}\mu_{j0} = 1 + \frac{1}{2}\mu_{00} + \frac{1}{2}\mu_{20} = 1 + \frac{1}{2}\mu_{20} \\ \mu_{20} & = 1 + \frac{1}{2}\mu_{j-1,0} + \frac{1}{2}\mu_{j+1,0} \\ & \vdots \\ \mu_{j0} & = 1 + \frac{1}{2}\mu_{10} + \frac{1}{2}\mu_{30} \\ & \vdots \\ \mu_{N-1,0} & = 1 + \frac{1}{2}\mu_{N-2,0} \end{array}\right.$$

which can be solved like this (for example):

$$\left[\begin{array}{ll} \mu_{20} & = 2\mu_{10} - 2 \\ \mu_{30} & = 2\mu_{20} - 2 - \mu_{10} = 4\mu_{10} - 4 - 2 - \mu_{10} = 3\mu_{10} - 6 \\ \mu_{40} & = 2\mu_{30} - 2 - \mu_{20} = 4\mu_{10} - 12 \\ & \vdots \end{array}\right.$$

By induction, one can prove that $\mu_{i0} = i\mu_{10} - i(i-1)$ for $1 \leq i \leq N-1$. Using the last line we find $(N-1)\mu_{10} - (N-1)(N-2) = 1 + \frac{1}{2}((N-2)\mu_{10} - (N-2)(N-3)) \Rightarrow \mu_{10} = N-1$ and thus we get

$$\mathbb{E}(T_0|X_0 = 0) = 1 + \mathbb{E}(H_0|X_0 = 1) = 1 + \mu_{10} = 1 + N - 1 = N$$

Example 1.7 (Symmetric random walk on $\mathbb{Z}$)**.**

$$T_0 = \inf\{n \geq 1 : X_n = 0\} = 1 + \mathbb{E}(H_0|X_0 = 1) = 1 + \mu_{10}$$

As before, we get by induction $\mu_{i0} = i\mu_{10} - i(i-1)$ for $i \geq 1$. But what is the value of μ_{10}?

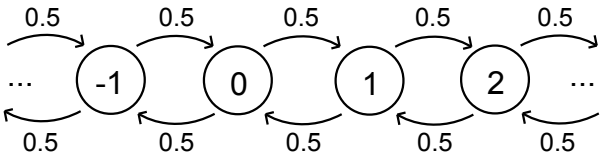


Figure 4: Symmetric random walk on $\mathbb{Z}$

Recall that $(\mu_{iA}, i \in S)$ a *non-negative* solution of (1). If we choose μ_{10} finite, this implies there will be an i sufficiently large for which $\mu_{i0} < 0$: contradiction. So the only remaining option is $\mu_{10} = +\infty$, which implies that $\mu_{i0} = +\infty$, $\forall i \geq 1$ and that the chain is *null-recurrent* (but notice that we are not yet able to compute $\mathbb{P}(T_0 = n|X_0 = 0)$).

Another approach

Reminder: Let $(X_n, n \geq 0)$ be an irreducible Markov chain. Then $(X_n, n \geq 0)$ admits a stationary distribution if and only if it is positive-recurrent. In this case, $\pi_i^* = \frac{1}{\mathbb{E}(T_i|X_0=i)} \forall i \in S$ (where $T_i = \inf\{n \geq 1 : X_n = i\}$), so π^* is also *unique*.

We now set out to prove

$$\pi_i^* = \frac{1}{\mathbb{E}(T_i|X_0 = i)} \tag{2}$$

Lemma 1.8. Let X be a random variable with values in $\mathbb{N}^* = \{1, 2, 3, \ldots\}$, then $\mathbb{E}(X) = \sum_{n\geq 1} \mathbb{P}(X \geq n)$

Proof.

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{m\geq 1} m\,\mathbb{P}(X = m) = \sum_{m\geq 1} m\,\mathbb{E}(1_{\{x=m\}}) \overset{(*)}{=} \mathbb{E}\left(\sum_{m\geq 1} m\,1_{\{X=m\}}\right) = \mathbb{E}\left(\sum_{m\geq 1}\left(\sum_{n=1}^{m} 1\right) 1_{\{X=m\}}\right) \\
&\overset{(*)}{=} \mathbb{E}\left(\sum_{n\geq 1}\sum_{m\geq n} 1_{\{X=m\}}\right) \overset{(*)}{=} \sum_{n\geq 1} \mathbb{E}(1_{\{X=n \text{ or } X=n+1 \text{ or } X=n+2\ldots\}}) = \sum_{n\geq 1} \mathbb{E}(1_{\{X\geq n\}}) = \sum_{n\geq 1} \mathbb{P}(X \geq n)
\end{aligned}$$

$(*)$ using Fubini's theorem to swap the order of expectation and sums □

Proof of (2).
We want to compute

$$\mathbb{E}(T_i|X_0 = i) = \sum_{n\geq 1} \mathbb{P}(T_i \geq n|X_0 = i)$$

as shwon by the above lemma. Since we know the result of the theorem we want to show, let us multiply both sides by π_i^*.

$$\pi_i^*\,\mathbb{E}(T_i|X_0 = i) = \pi_i^* \sum_{n\geq 1} \mathbb{P}(T_i \geq n|X_0 = i)$$

In order to prove (2), we have to show that the above quantity is equal to 1.

Notice that we are looking for the conditional expectation of T_i, so we can choose any distribution for X_0. In particular, we will choose the stationary distribution for the purpose of the proof. Mathematically, we express it as $\mathbb{P}(X_0 = i) = \pi_i^*$, which implies that $\mathbb{P}(X_n = i) = \pi_i^*$, $\forall n \geq 0$. So

$$\begin{aligned}
\pi_i^*\,\mathbb{E}(T_i|X_0 = i) &= \sum_{n\geq 1} \mathbb{P}(T_i \geq n|X_0 = i)\,\mathbb{P}(X_0 = i) = \sum_{n\geq 1} \mathbb{P}(T_i \geq n, X_0 = i) \\
&= \mathbb{P}(T_i \geq 1, X_0 = i) + \sum_{n\geq 2} \mathbb{P}(T_i \geq n, X_0 = i) = \mathbb{P}(X_0 = i) + \sum_{n\geq 2} \mathbb{P}(T_i \geq n, X_0 = i)
\end{aligned}$$

(noticing that $\{T_i \geq n\} = \{X_1 \neq i, X_2 \neq i, \ldots, X_{n-1} \neq i\}$). We continue developing our equality:

$$\begin{aligned}
\pi_i^*\mathbb{E}(T_i|X_0 = i) &= \mathbb{P}(X_0 = i) + \sum_{n\geq 2} \mathbb{P}(\underbrace{X_0 = i, X_1 \neq i, X_2 \neq i, \ldots, X_{n-1} \neq i)}_{=\{X_1\neq i,\ldots,X_{n-1}\neq i\}\setminus\{X_0\neq i,\ldots,X_{n-1}\neq i\}} \\
&= \mathbb{P}(X_0 = i) + \sum_{n\geq 2} \left(\mathbb{P}(X_1 \neq i, \ldots, X_{n-1} \neq i) - \mathbb{P}(X_0 \neq i, \ldots, X_{n-1} \neq i)\right) \\
&\overset{(*)}{=} \mathbb{P}(X_0 = i) + \sum_{n\geq 2} \left(\underbrace{\mathbb{P}(X_0 \neq i, \ldots, X_{n-2} \neq i)}_{=a_{n-2}} - \underbrace{\mathbb{P}(X_0 \neq i, \ldots, X_{n-1} \neq i)}_{=a_{n-1}}\right) \\
&= \mathbb{P}(X_0 = i) + \sum_{n\geq 2} a_{n-2} - a_{n-1} \\
&= \mathbb{P}(X_0 = i) + a_0 - a_1 + a_1 - a_2 + a_2 + \ldots - \lim_{n\to\infty} a_n \\
&= \mathbb{P}(X_0 = i) + a_0 - \lim_{n\to\infty} a_n
\end{aligned}$$

Notice that $(*)$ holds, as we assumed that the chain is in stationary distribution $\forall n$. Also, $\lim_{n\to\infty} a_n = \mathbb{P}(\underbrace{X_0 \neq i, \ldots, X_n \neq i, \ldots}_{\text{the chain never comes back to } i}) = 0$, because the chain is recurrent.

Finally, we find

$$\pi_i^* \, \mathbb{E}(T_i | X_0 = i) = \mathbb{P}(X_0 = i) + \underbrace{a_0}_{=\mathbb{P}(X_0 \neq i)} - 0 = 1 \quad \text{so} \quad \pi_i^* = \frac{1}{\mathbb{E}(T_i | X_0 = i)}$$

□

Example 1.9 (Asymmetric random walk on the circle with $p + q = 1$)**.** The transition matrix is given by

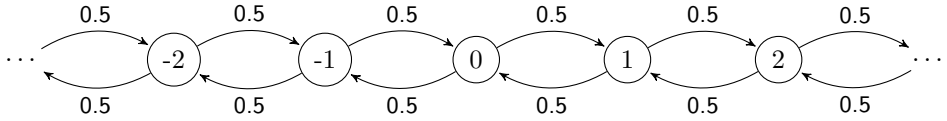
$$P = \begin{bmatrix} 0 & p & 0 & \ldots & 0 & q \\ q & 0 & p & 0 & \ldots & 0 \\ 0 & q & \ddots & \ddots & & \vdots \\ \vdots & 0 & \ddots & & & 0 \\ 0 & \ldots & & & 0 & p \\ p & 0 & \ldots & 0 & q & 0 \end{bmatrix}$$

In this case, we have seen that $\pi_i^* = 1/N$, irrespective of the value of p. We know that this chain is irreducible and positive-recurrent. Thus $\mathbb{E}(T_0 | X_0 = 0) = \frac{1}{\pi_0^*} = N$. The expected return time to 0 does therefore not depend on the value of p, which was a priori not so clear!

Random Walks: WEEK 6

1 Reflection principle

First, let us recall a few things about the symmetric random walk on $\mathbb{Z}$. We denote by $(S_n, n \geq 0)$ the simple symmetric random walk on $\mathbb{Z}$.



We know that this chain is irreducible and recurrent, that is:

$$f_{00} = \mathbb{P}(T_0 < +\infty | S_0 = 0) = 1$$

More than that, we know that it is null-recurrent, meaning:

$$\mathbb{E}(T_0 | S_0 = 0) = +\infty$$

How could we compute $f_{00}(2n) = \mathbb{P}(T_0 = 2n | S_0 = 0)$?

Due to the fact that we can translate the random walk along $\mathbb{Z}$ without actually changing anything, we will use the following notation:

$$p_{j-i}(n) = p_{ij}(n) = \mathbb{P}(S_n = j | S_0 = i)$$

The reflection principle is the proof technique that we are going to use to prove the following statement.

Theorem 1.1. Let $T_0 = \inf\{n \geq 1 : S_n = 0\}$. Then $\mathbb{P}(T_0 > 2n | S_0 = 0) = p_0(2n)$.

The left-hand side of the equality is the probability of never coming back to 0 before $2n$ steps. The right-hand side is the probability of being at 0 at time $2n$. From this theorem, we can then compute $f_{00}(2n)$ (left as an exercise).

Proof. First of all, let us shift the starting index by using the symmetry of the random walk.

$$\begin{aligned}
\mathbb{P}(T_0 > 2n | S_0 = 0) &= \mathbb{P}(T_0 > 2n, S_1 = +1 | S_0 = 0) + \mathbb{P}(T_0 > 2n, S_1 = -1 | S_0 = 0) \\
&= \mathbb{P}(T_0 > 2n | S_1 = +1, S_0 = 0)\, \mathbb{P}(S_1 = +1 | S_0 = 0) + \\
&\quad \mathbb{P}(T_0 > 2n | S_1 = -1, S_0 = 0)\, \mathbb{P}(S_1 = -1 | S_0 = 0) \\
&= \frac{1}{2}\, \mathbb{P}(T_0 > 2n | S_1 = +1) + \frac{1}{2}\, \mathbb{P}(T_0 > 2n | S_1 = -1)
\end{aligned}$$

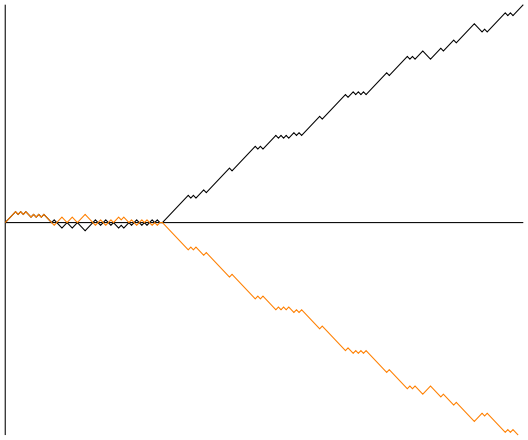
By symmetry, $\mathbb{P}(T_0 > 2n | S_1 = +1) = \mathbb{P}(T_0 > 2n | S_1 = -1)$ and we get:

$$\mathbb{P}(T_0 > 2n | S_0 = 0) = \mathbb{P}(T_0 > 2n | S_1 = 1)$$

Let us now distinguish paths depending on the point they are at time $2n$. Note that it has to be positive and even.

$$\begin{aligned}
\mathbb{P}(T_0 > 2n | S_1 = 1) &= \mathbb{P}(S_2 \neq 0, \ldots, S_{2n} \neq 0 | S_1 = 1) \\
&= \sum_{k \geq 1} \Big\{ \mathbb{P}(S_2 \neq 0, \ldots, S_{2n-1} \neq 0, S_{2n} = 2k | S_1 = 1) \Big\} \\
&= \sum_{k \geq 1} \Big\{ \mathbb{P}(S_{2n} = 2k | S_1 = 1) - \mathbb{P}(S_{2n} = 2k, \exists 2 \leq m \leq 2n-1 : S_m = 0 | S_1 = 1) \Big\}
\end{aligned}$$

The first term in the sum is simply the probability to be at $2k$ after $2n-1$ steps starting from 1. The second term is the probability to be at $2k$ after $2n-1$ steps starting from 1, but after hitting the 0-axis at some point. For any such path, we can draw a "mirror" path, as shown in the following graph, that will end up to be at $-2k$ after $2n-1$ steps.



The mirror path coincides with the original path until it hits zero. Then, it is a mirrored version of the original path.

Since any path that starts from 1 and ends up in $-2k$ has to cross the 0-axis, this shows that the number of paths described in the second term of the sum is exactly the number of paths that start from 1 and end in $-2k$ after $2n-1$ steps.

From this, we can further simplify the sum:

$$\begin{aligned}\mathbb{P}(T_0 > 2n | S_1 = 1) &= \sum_{k\geq 1} \Big\{ \mathbb{P}(S_{2n} = 2k | S_1 = 1) - \mathbb{P}(S_{2n} = -2k | S_1 = 1) \Big\} \\ &= \sum_{k\geq 1} \Big\{ p_{2k-1}(2n-1) - p_{2k+1}(2n-1) \Big\}\end{aligned}$$

This is a telescopic sum whose terms are null after some index, because $p_{2k+1}(2n-1) = 0$ for $k \geq n$. Therefore, the only remaining term is $p_1(2n-1)$ and we get:

$$\mathbb{P}(T_0 > 2n | S_1 = 1) = p_1(2n-1)$$

Finally, by the same argument of symmetry as in the beginning, we can conclude:

$$\mathbb{P}(T_0 > 2n | S_0 = 0) = p_0(2n)$$

□

2 Consequences

2.1 Null-recurrence of the symmetric random walk on $\mathbb{Z}$

We now have a third proof of the fact that the simple symmetric random walk on $\mathbb{Z}$ is null-recurrent, i.e.

$$\mathbb{E}(T_0 | S_0 = 0) = +\infty$$

Proof. By using the lemma from last time:

$$\mathbb{E}(T_0|S_0 = 0) = \sum_{n\geq 1} \mathbb{P}(T_0 \geq n|S_0 = 0) \geq \sum_{n\geq 1} \mathbb{P}(T_0 > 2n|S_0 = 0)$$

Now we just apply the theorem:

$$\mathbb{E}(T_0|S_0 = 0) \geq \sum_{n\geq 1} p_0(2n)$$

Recall that $p_0(2n) \sim \frac{1}{\sqrt{\pi n}}$. Since $\sum_{n\geq 1} \frac{1}{\sqrt{\pi n}}$ diverges to $+\infty$, then $\sum_{n\geq 1} p_0(2n)$ also diverges to $+\infty$ and $\mathbb{E}(T_0|S_0 = 0) = +\infty$. □

2.2 The arcsine law

When we average everything, a symmetric random walk on $\mathbb{Z}$ will spend half its time above the 0-axis and half its time below. But what will actually typically happen is that the random walk will either spend most of its time above the 0-axis or most of its time below. We express this with the following theorem.

Definition 2.1. We define $L_{2n} = \sup\{0 \leq m \leq 2n : S_m = 0\}$ the time of last visit to 0 before $2n$.

Theorem 2.2. For n and k large (typically, $k = xn, 0 < x < 1$):

$$\mathbb{P}(L_{2n} = 2k|S_0 = 0) \sim \frac{1}{\pi\sqrt{k(n-k)}}$$

It means that a typical trajectory will cross 0 either at beginning or at the end.

Proof.

$$\begin{aligned}
\mathbb{P}(L_{2n} = 2k|S_0 = 0) &= \mathbb{P}(S_m \neq 0 \text{ for } m \in \{2k+1, \ldots, 2n\}, S_{2k} = 0|S_0 = 0) \\
&= \mathbb{P}(S_m \neq 0 \text{ for } m \in \{2k+1, \ldots, 2n\}|S_{2k} = 0)\, \mathbb{P}(S_{2k} = 0|S_0 = 0) \\
&= \mathbb{P}(S_m \neq 0 \text{ for } m \in 1, \ldots, 2(n-k)|S_0 = 0)\, p_0(2k) \\
&= p_0(2(n-k))\, p_0(2k) \text{ (by the theorem)} \\
&\sim \frac{1}{\sqrt{\pi(n-k)}} \frac{1}{\sqrt{\pi k}} \\
&\sim \frac{1}{\pi\sqrt{k(n-k)}}
\end{aligned}$$

□

2.3 Law of the iterated logarithm

We will not prove it, but we can also get the following result:

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = +1\right) = \mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = -1\right) = 1$$

This provides an envelope that the random walk will almost surely keep hitting.

3 Reversible chains

The ergodic theorem provides us with a nice convergence result, that is $\lim_{n\to\infty} p_{ij}(n) = \pi_j^*$ for any $i, j \in S$. But for the purpose of any practical application, we would like to know more about the rate at which this convergence occurs. We will start by talking about reversible chains and detailed balanced.

Definition 3.1. An ergodic Markov chain $(X_n, n \geq 0)$ is said to be reversible if its stationary distribution π^* satisfies the following detailed balance equation:

$$\pi_i^* p_{ij} = \pi_j^* p_{ji} \quad \forall i, j \in S$$

Remarks.

- We can still talk about reversibility if the chain is only irreducible and positive-recurrent.
- If one assumes a that the chain is in stationary distribution from the start then the backwards chain $X_n, X_{n-1}, \ldots$ has the same transition probabilities as the original chain, hence the name "reversible".
- If π^* satisfies the detailed balance equation, then $\pi^* = \pi^* P$.
- The reciprocal statement is wrong, as we will see in some counter-examples.
- We do not have general conditions that ensure that the detailed balance equation is satisfied.

Example 3.2 (Ehrenfest urns process)**.** Consider 2 urns with N numbered balls. At each step, we pick uniformly at random a number between 1 and N, take the ball with this number and put it in the other urn. The state is the number of balls in the right urn. The transition probabilities are the following:

$$\begin{aligned} p_{i,i+1} &= \frac{N-i}{N} \\ p_{i,i-1} &= \frac{i}{N} \end{aligned}$$

If we try to solve the detailed balance equation we get:

$$\begin{aligned} & \pi_{i+1}^* = \frac{p_{i,i+1}}{p_{i+1,i}} \pi_i^* = \frac{N-i}{i+1} \pi_i^* \\ \Rightarrow \quad & \pi_{i+1}^* = \frac{(N-i)(N-i-1)\ldots N}{(i+1)i(i-1)\ldots 2} \pi_0^* = \frac{N!}{(N-i-1)!(i+1)!} \pi_0^* \\ \Rightarrow \quad & \pi_{i+1}* = \binom{N}{i+1} \pi_0^* \end{aligned}$$

which leads to the conclusion that $\pi_0^* = \frac{1}{2^N}$ (see Homework 4),This process is therefore reversible.

Example 3.3. All irreducible birth-death processes (as also studied in Homework 4) satisfy the detailed balance equation.

Example 3.4. If for any $i, j \in S$, we have $p_{ij} > 0$ and $p_{ji} = 0$, then the chain is not reversible.

Example 3.5 (Random walk on the circle)**.** We know that the stationary distribution for the cyclic random walk on a circle with transition probabilities p and q ($p+q=1$) is simply the uniform distribution, $\pi_i^* = \frac{1}{N}$. To be verified, the detailed balance equation requires $\pi_i^* p = \pi_{i+1}^* q \Leftrightarrow p = q = \frac{1}{2}$, which is not the case in general.

Random Walks: WEEK 7

1 Rate of convergence, spectral gap and mixing times

1.1 Setup, motivation and assumptions

Let $(X_n, n \geq 0)$ be a time-homogeneous, ergodic Markov chain on a state space $\mathcal{S}$ and let $\mathbf{P}$ be its transition matrix. Therefore, there exists a limiting and stationary distribution which we call π. In other words, $p_{ij}(n) \xrightarrow[n\to\infty]{} \pi_j, \forall i, j \in \mathcal{S}$ (limiting distribution) and $\pi = \pi\mathbf{P}$ (stationary distribution).

The question we ask is: *For what values of n is $p_{ij}(n)$ "really close" to π_j? In other words, "how fast" does $p_{i,\cdot}(n)$ converge to its limiting distribution π?* The answer to this question is useful for practical applications (see for example Section 6.14 of Grimmett & Stirzaker) where it is not enough to only know what happens when $n \to \infty$; in some cases, we also need to have a notion of how soon the behavior of a Markov chain becomes similar to the behavior at infinity.

We make the following simplifying assumptions:

1. $\mathcal{S}$ is finite ($|\mathcal{S}| = N$).
2. The detailed balance equation is satisfied:

$$\pi_i p_{ij} = \pi_j p_{ji} \quad \forall i, j \in \mathcal{S}. \tag{1}$$

1.2 Total variation norm and convergence of distribution

Here, we consider *convergence of distribution*. We want to see when $P_{ij}(n) = \mathbb{P}(X_n = j | X_0 = i)$ and $\pi_j = \mathbb{P}(X_* = j)$ "get close to each other". To clarify what this means, we introduce a new object, the *total variation norm*.

Definition 1.1. Total variation norm. Let $\mu = (\mu_i, i \in \mathcal{S})$, $\nu = (\nu_i, i \in \mathcal{S})$ such that $\mu_i \geq 0$, $\nu_i \geq 0$, $\sum_{i\in\mathcal{S}} \mu_i = 1$, $\sum_{i\in\mathcal{S}} \nu_i = 1$. We define the total variation norm of μ and ν as $\frac{1}{2}\sum_{i\in\mathcal{S}} |\mu_i - \nu_i|$ and we denote it by $\|\mu - \nu\|_{TV}$.

Note: It is easy to check that $\|\mu - \nu\|_{TV} \in [0, 1]$.

In what follows, we will find an upper-bound on $\|\mathbf{P}_i^n - \pi\|_{TV}$. By studying how fast this upper-bound goes to 0, we will find out how fast $\|\mathbf{P}_i^n - \pi\|_{TV} = \frac{1}{2}\sum_{i\in\mathcal{S}} |p_{ij}(n) - \pi_j|$ goes to 0.

1.3 Eigenvalues and eigenvectors of P

Define a new matrix $\mathbf{Q}$ as follows:

$$q_{ij} = \sqrt{\pi_i}\, p_{ij}\, \frac{1}{\sqrt{\pi_j}}, \qquad \forall i, j \in \mathcal{S}$$

Two observations: 1. $q_{ii} = p_{ii},, \forall i \in \mathcal{S}$, 2. $q_{ij} \geq 0$, but $\sum_{j\in\mathcal{S}} q_{ij} \neq 1$ in general.

Proposition 1.2. $\mathbf{Q}$ is symmetric.

Proof.

$$q_{ji} = \sqrt{\pi_j} p_{ji} \frac{1}{\sqrt{\pi_i}} = \frac{1}{\sqrt{\pi_i \pi_j}} \pi_j p_{ji} \overset{(*)}{=} \frac{1}{\sqrt{\pi_i \pi_j}} \pi_i p_{ij} = q_{ij}.$$

where $(*)$ follows from the detailed balance equation. □

Since $\mathbf{Q}$ is symmetric, we can use the spectral theorem to conclude the following:

Proposition 1.3. There exist real numbers $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{N-1}$ (the eigenvalues of $\mathbf{Q}$) and vectors $\mathbf{u}^{(0)}, \ldots, \mathbf{u}^{(N-1)} \in \mathbb{R}^N$ (the eigenvectors of $\mathbf{Q}$) such that $\mathbf{Q}\mathbf{u}^{(k)} = \lambda_k \mathbf{u}^{(k)}, \forall k = 0, ..., N-1$. Moreover, $\mathbf{u}^{(0)}, \ldots, \mathbf{u}^{(N-1)}$ forms an orthonormal basis of $\mathbb{R}^n$.

Proposition 1.4. Define the vector $\phi^{(k)} = \left(\frac{\mathbf{u}_j^{(k)}}{\sqrt{\pi_j}}, j \in \mathcal{S} \right)$. Then,

$$\mathbf{P}\phi^{(k)} = \lambda_k \phi^{(k)}.$$

Proof. By Proposition 1.3, for every $k = 0, ..., N-1$

$$\begin{aligned}
&\mathbf{Q}\mathbf{u}^{(k)} = \lambda_k \mathbf{u}^{(k)} && \Leftrightarrow \sum_{j \in \mathcal{S}} q_{ij} \mathbf{u}_j^{(k)} = \lambda_k \mathbf{u}_i^{(k)} \\
\Leftrightarrow &\sum_{j \in \mathcal{S}} \sqrt{\pi_i} p_{ij} \frac{1}{\sqrt{\pi_j}} \mathbf{u}_j^{(k)} = \lambda_k \mathbf{u}_i^{(k)} && \Leftrightarrow \sum_{j \in \mathcal{S}} p_{ij} \underbrace{\left(\frac{\mathbf{u}_j^{(k)}}{\sqrt{\pi_j}} \right)}_{=\phi_j^{(k)}} = \lambda_k \underbrace{\left(\frac{\mathbf{u}_i^{(k)}}{\sqrt{\pi_i}} \right)}_{=\phi_i^{(k)}} && \Leftrightarrow \mathbf{P}\phi^{(k)} = \lambda_k \phi^{(k)}
\end{aligned}$$

□

Proposition 1.4 says that the eigenvalues of $\mathbf{P}$ are $\lambda_0, \lambda_1, ..., \lambda_{N-1}$ (the same as those of $\mathbf{Q}$) and the eigenvectors of $\mathbf{P}$ are $\phi^{(0)}, ..., \phi^{(N-1)}$. Note that $\phi^{(0)}, ..., \phi^{(N-1)}$ is not in general an orthonormal basis of $\mathbb{R}^n$.

1.4 Main results

Facts about the λ's and ϕ's (without proof):

1. $\lambda_0 = 1, \quad \phi^{(0)} = [1, ..., 1]^T$.
2. $|\lambda_k| \leq 1, \quad \forall k = 1, ..., N-1$.
3. $\lambda_1 < 1$ and $\lambda_{N-1} > -1$.

Definition 1.5. $\lambda_* = \max\limits_{1 \leq k \leq N-1} |\lambda_k| \, (< 1)$

Theorem 1.6. Rate of convergence. Under all the assumptions made, it holds that:

$$\|\mathbf{P}_i^n - \pi\|_{TV} \leq \frac{\lambda_*^n}{2\sqrt{\pi_i}} \quad \forall i \in \mathcal{S}, \forall n \geq 1 \tag{2}$$

Theorem 1.6 says that $\|\mathbf{P}_i^n - \pi\|_{TV} = \frac{1}{2} \sum_{j \in \mathcal{S}} |p_{ij}(n) - \pi_j|$ is decaying exponentially fast to 0 as $n \to \infty$.

Definition 1.7. Spectral gap. $\gamma = 1 - \lambda_*$.

Note: $\lambda_*^n = (1-\gamma)^n \leq e^{-\gamma n}$ (since $1 - x \leq e^{-x}, \quad \forall x \geq 0$). This shows that if the spectral gap is large, convergence is fast; if it is small, convergence is slow.

Definition 1.8. Mixing time. For a given $\epsilon > 0$, define $T_\epsilon = \inf\{n \geq 1 : \max\limits_{i \in \mathcal{S}} \|\mathbf{P}_i^n - \pi\|_{TV} \leq \epsilon\}$.

Remark. One can show that the sequence $\max_{i \in \mathcal{S}} \|\mathbf{P}_i^n - \pi\|_{TV}$ is decreasing in n, so the above definition makes sense.

1.5 Examples

Example 1.9. Walk on a circle.

Remember from previous week that the stationary distribution is $\pi_j = \frac{1}{N},\ \forall j \in \mathcal{S}$, and that $p = q = \frac{1}{2}$ implies detailed balance.

$$\mathbf{P} = \begin{bmatrix} 0 & 1/2 & 0 & \cdots & 0 & 1/2 \\ 1/2 & 0 & 1/2 & \cdots & 0 & 0 \\ 0 & 1/2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1/2 \\ 1/2 & 0 & 0 & \cdots & 1/2 & 0 \end{bmatrix}$$

The eigenvalues of $\mathbf{P}$ are $\lambda_k = \cos\left(\frac{2k\pi}{N}\right),\ \forall k = 0, ..., N-1$ (note that the eigenvalues are not ordered here).

If N is even, the chain is periodic of period 2, therefore not ergodic, therefore a limiting distribution does not exist (and the spectral gap is equal to zero)..

If N is odd,

$$\lambda_* = \left|\cos\left(\frac{2\pi(N-1)/2}{N}\right)\right| = \left|\cos\left(\pi\left(1 - \frac{1}{N}\right)\right)\right| = \cos\left(\frac{\pi}{N}\right)$$

The spectral gap is:

$$\gamma = 1 - \cos\frac{\pi}{N} \simeq 1 - \left(1 - \frac{\pi^2}{2N^2}\right) = \frac{\pi^2}{2N^2}$$

because $\cos(x) \simeq 1 - x^2/a$ close to $x = 0$. The spectral gap is therefore $\mathcal{O}\left(\frac{1}{N^2}\right)$. To compute the mixing time T_ϵ, we use Theorem 1.6:

$$\max_{i\in\mathcal{S}} \|\mathbf{P}_i^n - \pi\|_{TV} \leq \frac{\lambda_*^n}{2\sqrt{\pi_i}} = \frac{e^{-\gamma n}}{2\sqrt{1/N}} \simeq \frac{\sqrt{N}}{2}\exp\left(-\frac{\pi^2 n}{2N^2}\right)$$

This goes fast to 0 if $n \gg N^2$, for example if $n = cN^2 \log n$, with $c > 0$ a sufficiently large constant. This confirms our intuition that the larger the circle is, the longer we have to wait for the chain to reach equilibrium.

Example 1.10. Complete graph of N vertices.

$$p_{ij} = \begin{cases} 0 & \text{if } i = j \\ \frac{1}{N-1} & \text{otherwise} \end{cases}$$

$$\mathbf{P} = \frac{1}{N-1}\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{bmatrix}$$

The stationary distribution is uniform:

$$\pi_i = \frac{1}{N},\ \forall i \in \mathcal{S}$$

The eigenvalues of $\mathbf{P}$ are $\lambda_0 = 1,\ \lambda_k = -\frac{1}{N-1},\ \forall 1 \leq k \leq N-1 \Rightarrow \lambda_* = \frac{1}{N-1}$. The spectral gap is therefore $\gamma = 1 - \frac{1}{N-1} \sim \mathcal{O}(1)$.

To compute the mixing time T_ϵ, we use Theorem 1.6:

$$\|\mathbf{P}_i^n - \pi\|_{TV} \leq \frac{\lambda_*^n}{2\sqrt{1/N}} = \frac{\sqrt{N}}{2}\left(\frac{1}{N-1}\right)^n = \frac{\sqrt{N}}{2}\exp\left(n\log\frac{1}{N-1}\right) = \frac{\sqrt{N}}{2}\exp(-n\log(N-1))$$

This is of order ϵ for n finite.

Random Walks: WEEK 8

1 Rate of convergence: proofs

1.1 Reminder

Let $(X_n,\, n \geq 0)$ be a Markov Chain with state space S and transition matrix P, and consider the following assumptions:

- X is ergodic (irreducible, aperiodic and positive-recurrent). so there exists a stationary distribution π and it is a limiting distribution as well.
- The state space S is finite, $|S| = N$.
- Detailed balance holds ($\pi_i p_{ij} = \pi_j p_{ji} \quad \forall i, j \in S$).

Statement 1.1. Under these assumptions, we have seen that there exist numbers $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{N-1}$ and vectors $\phi^{(0)}, \phi^{(1)}, \ldots, \phi^{(N-1)} \in \mathbb{R}^N$ such that

$$P\phi^{(k)} = \lambda_k \phi^{(k)}, \quad k = 0, \ldots, N-1$$

and $\phi_j^{(k)} = \frac{u_j^{(k)}}{\sqrt{\pi_j}}$, where $u^{(0)}, \cdots, u^{(N-1)}$ is an orthonormal basis of $\mathbb{R}^N$ ($u^{(k)}$ are the eigenvectors of the symmetric matrix Q where $q_{ij} = \sqrt{\pi_i}\, p_{ij}\, \frac{1}{\sqrt{\pi_j}}$). Note that the $\phi^{(k)}$ do not usually form an orthnormal basis of $\mathbb{R}^N$.

Facts

1. $\phi_j^{(0)} = 1 \quad \forall j \in S, \quad \lambda_0 = 1 \quad$ and $\quad |\lambda_k| \leq 1 \quad \forall k \in \{0, \cdots, N-1\}$
2. $\lambda_1 < +1$ and $\lambda_{N-1} > -1$

Definition 1.2. Let us define $\lambda_* = \max\limits_{k \in \{1, \cdots, N-1\}} |\lambda_k| = \max(|\lambda_1|, |\lambda_{N-1}|)$. The spectral gap is defined as $\gamma = 1 - \lambda_*$.

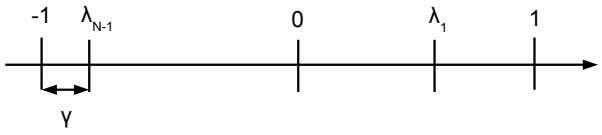


Figure 1: Spectral gap

Theorem 1.3. Under all the assumptions made above, we have

$$\|P_i^n - \pi\|_{TV} = \frac{1}{2} \sum_{j \in S} |p_{ij}(n) - \pi_j| \leq \frac{1}{2\sqrt{\pi_i}}\, \lambda_*^n \leq \frac{1}{2\sqrt{\pi_i}}\, e^{-\gamma n}, \quad \forall j \in S,\, n \geq 1$$

1.2 Proof of Fact 1

Let us first prove that $\phi_j^{(0)} = 1 \quad \forall j \in S$ and $\lambda_0 = 1$.

Consider $\phi_j^{(0)} = 1 \quad \forall j \in S$; we will prove that $(P\phi^{(0)})_i = \phi_i^{(0)}$ (so $\lambda_0 = 1$):

$$(P\phi^{(0)})_i = \sum_{j \in S} p_{ij} \underbrace{\phi_j^{(0)}}_{=1} = \sum_{j \in S} p_{ij} = 1 = \phi_i^{(0)}$$

Also, we know that $\phi_i^{(0)} = \frac{u_i^{(0)}}{\sqrt{\pi_i}}$, so $u_i^{(0)} = \sqrt{\pi_i}\,\phi_i^{(0)} = \sqrt{\pi_i}$. The norm of $u^{(0)}$ is therefore equal to 1:

$$\|u^{(0)}\|^2 = \sum_{i \in S} (u_i^{(0)})^2 = \sum_{i \in S} \pi_i = 1$$

Let us then prove that $|\lambda_k| \leq 1 \quad \forall k \in \{0, \cdots, N-1\}$.

Let $\phi^{(k)}$ be the eigenvector corresponding to λ_k. We define i to be such that $|\phi_i^{(k)}| \geq |\phi_j^{(k)}| \quad \forall j \in S$ ($|\phi_i^{(0)}| > 0$ because an eigenvector cannot be all-zero). We will use $P\phi^{(k)} = \lambda_k \phi^{(k)}$ in the following:

$$|\lambda_k \phi_i^{(k)}| = \left|(P\phi^{(k)})_i\right| = \left|\sum_{j \in S} p_{ij} \phi_j^{(k)}\right| \leq \sum_{j \in S} p_{ij} \underbrace{|\phi_j^{(k)}|}_{\leq |\phi_i^{(k)}|, \forall j \in S} \leq |\phi_i^{(k)}| \underbrace{\sum_{j \in S} p_{ij}}_{=1}$$

So we have $|\lambda_k||\phi_i^{(k)}| \leq |\phi_i^{(k)}|$, which implies that $|\lambda_k| \leq 1$, as $|\phi_i^{(k)}| > 0$. □

1.3 Proof of Fact 2

We want to prove that $\lambda_1 < +1$ and $\lambda_{N-1} > -1$, which together imply that $\lambda_* < 1$.
By the assumptions made, we know that the chain is irreducible, aperiodic and finite, so $\exists n_0 > 1$ such that $p_{ij}(n) > 0$, $\forall i, j \in S, \forall n \geq n_0$.

$\underline{\lambda_1 < +1}$.
Assume ϕ is such that $P\phi = \phi$: we will prove that ϕ can only be a mutliple of $\phi^{(0)}$, which implies that the eigenvalue $\lambda = 1$ has a unique eigenvector associated to it, so $\lambda_1 < 1$. Take i such that $|\phi_i| \geq |\phi_j|$, $\forall j \in S$, and let $n \geq n_0$.

$$\phi_i = (P\phi)_i = (P^n\phi)_i \overset{(*)}{=} \sum_{j \in S} p_{ij}(n)\,\phi_j$$

so

$$|\phi_i| = \left|\sum_{j \in S} p_{ij}(n)\,\phi_j\right| \leq \sum_{j \in S} p_{ij}(n) \underbrace{|\phi_j|}_{\leq |\phi_i|} \leq |\phi_i| \underbrace{\sum_{j \in S} p_{ij}(n)}_{1} = |\phi_i|$$

So we have $|\phi_i| \leq \sum_{j \in S} p_{ij}(n)|\phi_j| \leq |\phi_i|$. To have equality, we clearly need to have $|\phi_i| = |\phi_j|$, $\forall j \in S$ (because $p_{ij}(n) > 0$ for all i, j and $\sum_{j \in S} p_{ij} = 1$ for all $i \in S$). Because $(*)$ is satisfied, we also have $\phi_i = \sum_{j \in S} p_{ij}(n)\phi_j$, which in turn implies that $\phi_j = \phi_i$ for all $j \in S$. So the vector ϕ is constant. □

$\underline{\lambda_{\mathbf{N-1}} > -\mathbf{1}}$.
Assume there exists $\phi \neq 0$ such that $P\phi = -\phi$: we will prove that this is impossible, showing therefore that no eigenvalue can take the value -1. Take i such that $|\phi_i| \geq |\phi_j|$, $\forall j \in S$ and let n odd be such that $n \geq n_0$.

Now, as $P^n\phi = P\phi = -\phi$, we have $-\phi_i \overset{(*)}{=} \sum_{j\in S} p_{ij}(n)\,\phi_j$ and $|\phi_i| \leq \sum_{j\in S} p_{ij}(n)\,|\phi_j| \leq |\phi_i|$. So, as above, we need to have $|\phi_j| = |\phi_i|$, for all $j \in S$ and then, thanks to (*), $\phi_j = -\phi_i$, for all $j \in S$. This implies that $\phi_i = -\phi_i = 0$, and leads to $\phi_j = 0$ for all $j \in S$, which is impossible. □

1.4 Proof of the theorem

We will first use the Cauchy-Schwarz inequality which states that

$$\left|\sum_{j\in S} a_j b_j\right| \leq \left(\sum_{j\in S} a_j^2\right)^{1/2} \left(\sum_{j\in S} b_j^2\right)^{1/2}$$

so as to obtain

$$\begin{aligned}
\|P_i^n - \pi\|_{TV} &= \frac{1}{2}\sum_{j\in S} \underbrace{\left(\frac{p_{ij}(n) - \pi_j}{\sqrt{\pi_j}}\right)}_{a_j} \underbrace{\sqrt{\pi_j}}_{b_j} \leq \frac{1}{2}\left(\sum_{j\in S}\left(\frac{p_{ij}(n)}{\sqrt{\pi_j}} - \sqrt{\pi_j}\right)^2\right)^{1/2} \underbrace{\left(\sum_{j\in S}\pi_j\right)^{1/2}}_{1} \\
&= \frac{1}{2}\left(\sum_{j\in S}\left(\frac{p_{ij}(n)}{\sqrt{\pi_j}} - \sqrt{\pi_j}\right)^2\right)^{1/2}
\end{aligned}$$

Lemma 1.4.

$$\frac{p_{ij}(n)}{\sqrt{pi_j}} - \sqrt{\pi_j} = \sqrt{\pi_j}\sum_{k=1}^{N-1} \lambda_k^n \phi_i^{(k)} \phi_j^{(k)}$$

Proof. Remember that $u^{(0)}, \cdots, u^{(N-1)}$ is an orthonormal basis of $\mathbb{R}^N$, so we can write for any $v \in \mathbb{R}^N$ $v = \sum_{k=0}^{N-1}(v^T u^{(k)})u^{(k)}$ i.e. $v_j = \sum_{k=0}^{N-1}(v^T u^{(k)})u_j^{(k)}$. For a fix i, take $v_j = \frac{p_{ij}(n)}{\sqrt{\pi_j}}$. We obtain

$$(v^T u^{(k)}) = \sum_{j\in S} \frac{p_{ij}(n)}{\sqrt{\pi_j}} u_j^{(k)} = \sum_{j\in S} p_{ij}(n)\phi_j^{(k)} = (P^n\phi^{(k)})_i = \lambda_k^n \phi_i^{(k)}$$

which in turn implies

$$v_j = \frac{p_{ij}(n)}{\sqrt{\pi_j}} = \sum_{k=0}^{N-1} \lambda_k^n \phi_i^{(k)} u_j^{(k)} = \sum_{k=0}^{N-1} \lambda_k^n \phi_i^{(k)} \phi_j^{(k)} \sqrt{\pi_j} = \underbrace{\lambda_0^n \phi_i^{(0)} \phi_j^{(0)}}_{1} \sqrt{\pi_j} + \sqrt{\pi_j} \sum_{k=1}^{N-1} \lambda_k^n \phi_i^{(k)} \phi_j^{(k)}$$

□

Let us continue with the proof of the theorem using this lemma.

$$\begin{aligned}
\|P_i^n - \pi\|_{TV} &\leq \frac{1}{2}\left(\sum_{j\in S}\left(\frac{p_{ij}(n)}{\sqrt{\pi_j}} - \sqrt{\pi_j}\right)^2\right)^{1/2} = \frac{1}{2}\left(\sum_{j\in S}\left(\sqrt{\pi_j}\sum_{k=1}^{N-1}\lambda_k^n \phi_i^{(k)}\phi_j^{(k)}\right)^2\right)^{1/2} \\
&= \frac{1}{2}\left(\sum_{j\in S}\pi_j \sum_{k,l=1}^{N-1}\lambda_k^n\phi_i^{(k)}\phi_j^{(k)}\lambda_l^n\phi_i^{(l)}\phi_j^{(l)}\right)^{1/2} = \frac{1}{2}\left(\sum_{k,l=1}^{N-1}\lambda_k^n\phi_i^{(k)}\lambda_l^n\phi_i^{(l)}\sum_{j\in S}\pi_j\phi_j^{(k)}\phi_j^{(l)}\right)^{1/2} \\
&= \frac{1}{2}\left(\sum_{k=1}^{N-1}\lambda_k^{2n}(\phi_i^{(k)})^2\right)^{1/2}
\end{aligned}$$

where we have used the fact that $\sum_{j\in S}\pi_j\phi_j^{(k)}\phi_j(l) = \sum_{j\in S}u_j^{(k)}u_j^{(l)} = (u^{(k)})^T u^{(l)} = \delta_{kl}$. Remembering now that $|\lambda_k| \leq \lambda_*$, we obtain

$$\|P_i^n - \pi\|_{TV} \leq \frac{1}{2}\,\lambda_*^n\left(\sum_{k=1}^{N-1}(\phi_i^{(k)})^2\right)^{1/2}$$

In order to compute the term in parentheses, remember again that $v_j = \sum_{k=0}^{N-1}(v^T u^{(k)})\,u_j^{(k)}$ for every $v \in \mathbb{R}^N$, so by chosing $v = e_i$, i.e., $v_j = \delta_{ij} = \begin{cases}1 & \text{if } i = j\\ 0 & \text{otherwise}\end{cases}$, we obtain:

$$v^T u^{(k)} = u_i^{(k)} \quad \text{and} \quad \delta_{ij} = \sum_{k=0}^{N-1} u_i^{(k)}u_j^{(k)}$$

For $i = j$, we get $\delta_{ii} = 1 = \sum_{k=0}^{N-1}(u_i^{(k)})^2 = \sum_{k=0}^{N-1}\pi_i\,(\phi_i^{(k)})^2$, so

$$\sum_{k=1}^{N-1}(\phi_i^{(k)})^2 = \sum_{k=0}^{N-1}(\phi_i^{(k)})^2 - \underbrace{(\phi_i^{(0)})^2}_{1} = \frac{1}{\pi_i} - 1 \leq \frac{1}{\pi_i}$$

which leads to the inequality

$$\|P_i^n - \pi\|_{TV} \leq \frac{\lambda_*^n}{2\sqrt{\pi_i}}$$

and therefore completes the proof. □

1.5 Lazy Random Walks

Adding self-loops to a Markov chain makes it a priori "lazy". Surprisingly perhaps, this might in some cases speed up the convergence to equilibrium!

Adding self-loops of weight $\alpha \in (0,1)$ to every state has the following impact on the transition matrix: assuming P is the transition matrix of the initial Markov chain, the new transition matrix $\widetilde{P}$ becomes

$$\widetilde{P} = \alpha\, I + (1-\alpha)\, P$$

As a consequence:

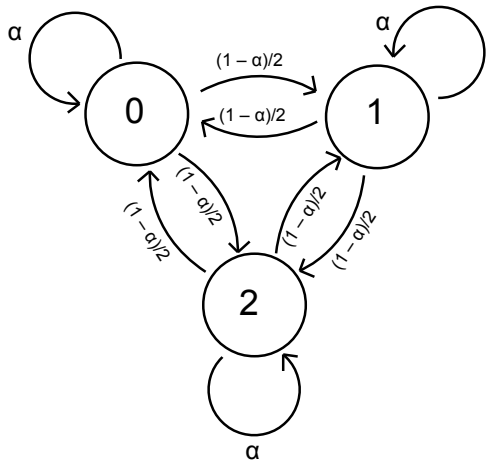
- The eigenvalues also change from λ_k to $\widetilde{\lambda}_k = \alpha + (1-\alpha)\lambda_k$, which in some case reduces the value of $\lambda_* = \max_{1\leq k\leq N-1}|\lambda_k|$. The spectral gap being equal to $\gamma = 1 - \lambda_*$, we obtain that by reducing λ_*, we might increase the spectral gap as well as the convergence rate to equilibirum.

- Note that λ_0 stays the same: $\widetilde{\lambda}_0 = \alpha\,\lambda_0 + (1-\alpha) = 1$, as well as the stationary distribution π:

$$\pi\widetilde{P} = \pi\left(\alpha I + (1-\alpha)P\right) = \alpha\pi + (1-\alpha)\underbrace{\pi P}_{=\pi} = \pi$$

Example 1.5. Random walk on the circle with $N = 3$:

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix} \xrightarrow{\text{add } \alpha} \widetilde{P} = \begin{pmatrix} \alpha & \frac{1-\alpha}{2} & \frac{1-\alpha}{2} \\ \frac{1-\alpha}{2} & \alpha & \frac{1-\alpha}{2} \\ \frac{1-\alpha}{2} & \frac{1-\alpha}{2} & \alpha \end{pmatrix}$$



Example 1.6 (PageRank)**.** The principle of the search enginge of Google works as follows. One can represent the internet as a graph with the hyperlinks being the edges and the webpages being the vertices. We define the transition probabilties of a random walk on this graph as

$$p_{ij} = \frac{1}{d_i} \quad \forall j \text{ connected to } i$$

where d_i is the degree of page i.

The principle is that the most popular pages are the webpages visited the most often. If π is the stationary distribution of the above random walk, then π_i is a good indicator of the popularity of page i. We therefore need to solve $\pi = \pi P$. In practice however, due to the size of the state space, solving this linear system takes too long in real time. What PageRank does is to compute instead $\pi_0 P^n$ for some initial distribution π_0 and a small value of n, which is meant to give a good approximation of the stationary distribution π. The quality of the approximation is of course directly linked to the rate of convergence to equilibrium. Adding self-loops of weight α to the graph seems to help in this case: the practical value chosen for α is around 15%.

Random Walks: WEEK 9

1 The cut-off phenomenon

1.1 Summary of the two previous lectures

Recall that we are considering a Markov chain $(X_n,\ n \geq 0)$ with transition matrix P and a *finite* state space S, with $|S| = N$. We assume that the chain is *ergodic* (irreducible, aperiodic and positive-recurrent), thus there is a unique stationary and limiting distribution π, with $\pi = \pi P$, and $p_{ij}(n) \xrightarrow[n\to\infty]{} \pi_j,\ \forall i, j \in S$. Finally, we assume that the detailed balance equation is satisfied: $\pi_i\, p_{ij} = \pi_j\, p_{ji},\ \forall i, j \in S$.

Under these assumptions, the transition matrix P has N eigenvectors $\phi^{(0)}, \cdots, \phi^{(N-1)} \in \mathbb{R}^N$, and N corresponding eigenvalues $1 = \lambda_0 > \lambda_1 \geq \cdots \geq \lambda_{N-1} > -1$ such that $P\phi^{(k)} = \lambda_k \phi^{(k)}\ \forall 0 \leq k \leq N-1$, We define also $\lambda_* := \max_{1\leq k\leq N-1} |\lambda_k| = \max\{|\lambda_1|, |\lambda_{N-1}|\}$.

In the previous lecture, we studied the rate of convergence of the distribution of this Markov chain towards π. If the initial state is $X_0 = i$, at time-step n the probability distribution is given by P_i^n. We measure the distance between this distribution and π in terms of the total variation (TV) distance: $\|P_i^n - \pi\|_{TV} := \frac{1}{2}\sum_{j\in S} |p_{ij}(n) - \pi_j|$. We proved the following bound:

Theorem 1.1 (Rate of Convergence)**.** Under the above assumptions,

$$\|P_i^n - \pi\|_{TV} \leq \frac{\lambda_*^n}{2\sqrt{\pi_i}},\ \forall i \in S, n \geq 1$$

Notice that this upper bound decays exponentially in terms of n. Today, we study a *reciprocal statement*, which gives a corresponding lower bound (under additional assumptions), that also decays exponentially in terms of n. We study examples where these two bounds are tight or loose. We also analyze an example of the *cut-off phenomenon*, where the actual TV distance does not have a smooth exponential decay in terms of n, but rather a sudden drop from 1 to 0 in a short interval.

1.2 Reciprocal statement

Before we present the statement, we need to better understand the concept of total variation distance.

Proposition 1.2. Let μ, ν be two probability distributions over the state space S. The following are equivalent definitions for the total variation distance $\|\mu - \nu\|_{TV}$:

$$\begin{aligned}
\|\mu - \nu\|_{TV} &= \frac{1}{2}\sum_{j\in S} |\mu_j - \nu_j| && (1)\\
&= \max_{A\subset S} |\mu(A) - \nu(A)| \qquad \text{where } \mu(A) = \sum_{j\in A} \mu_j && (2)\\
&= \frac{1}{2}\max_{\phi: S\to[-1,1]} |\mu(\phi) - \nu(\phi)| \qquad \text{where } \mu(\phi) = \sum_{j\in S} \mu_j \phi_j && (3)
\end{aligned}$$

Proving these equivalences is left as an exercise (suggestion: show that (1) $\leq$ (2) $\leq$ (3) $\leq$ (1)).

In the theorem below, we will use these facts about the eigenvectors of the transition matrix P, which have already been seen in class:

1. $\lambda_0 = 1$ and $\phi^{(0)} = (1, \cdots, 1)^T$;
2. $\phi_j^{(k)} = \dfrac{u_j^{(k)}}{\sqrt{\pi_j}}$ where $(u^{(k)})^T u^{(l)} = \delta_{kl}$.

Theorem 1.3 (Reciprocal Statement)**.** Under the above assumptions, and with the additional condition $|\phi_0^{(k)}| = 1,\ |\phi_j^{(k)}| \le 1,\ \forall k \in \{0, \cdots, N-1\}, j \in S$, we have

$$\|P_0^n - \pi\|_{TV} \ge \frac{\lambda_*^n}{2},\ \forall n \ge 1$$

Proof. Because of the additional assumption that all eigenvectors satisfy $|\phi_j^{(k)}| \le 1,\ \forall j, k$, we can use the third definition of the TV distance:

$$\|P_0^n - \pi\|_{TV} = \frac{1}{2} \max_{\phi: S \to [-1,1]} |P_0^n(\phi) - \pi(\phi)| \ge \frac{1}{2} \max_{1 \le k \le N-1} |P_0^n(\phi^{(k)}) - \pi(\phi^{(k)})|$$

We do not include $k = 0$ in the above formula, because for $k = 0$, $|P_0^n(\phi^{(0)}) - \pi(\phi^{(0)})| = 1 - 1 = 0$. We compute the two terms on the right-hand side separately:

$$P_0^n(\phi^{(k)}) = \sum_{j \in S} p_{0j}(n) \phi_j^{(k)} = (P^n \phi^{(k)})_0 = (\lambda_k^n \phi^{(k)})_0 = \lambda_k^n \phi_0^{(k)}$$

and

$$\pi(\phi^{(k)}) = \sum_{j \in S} \pi_j \phi_j^{(k)} = \sum_{j \in S} \pi_j \phi_j^{(k)} \underbrace{(\phi_j^{(0)})}_{=1} = \sum_{j \in S} u_j^{(k)} u_j^{(0)} = (u^{(k)})^T u^{(0)} = 0 \text{ (for } k \ne 0)$$

Therefore,

$$\|P_0^n - \pi\|_{TV} \ge \frac{1}{2} \max_{1 \le k \le N-1} |P_0^n(\phi^{(k)}) - \pi(\phi^{(k)})| = \frac{1}{2} \max_{1 \le k \le N-1} |\lambda_k^n| \underbrace{|\phi_0^{(k)}|}_{=1} = \frac{1}{2} \max_{1 \le k \le N-1} |\lambda_k^n| = \frac{1}{2} \lambda_*^n$$

□

1.3 Examples

The upper and lower bounds for $\|P_0^n - \pi\|_{TV}$ seem to be very close, as they only differ by a multiplicative factor of $1/\sqrt{\pi_0}$. But if this factor is very large with respect to the spectral gap $\gamma = 1 - \lambda_*$, the bounds are no longer tight. In the next two examples, it is easy to check that all the assumptions are satisfied, including the condition in the reciprocal statement.

Example 1.4 (Random walk over odd cycle)**.** Consider the random walk over an odd cycle, where $S = \{0, \cdots, N-1\}$, N is odd, and $p_{ij} = \frac{1}{2}$ if $i - j \equiv \pm 1 \mod N$, 0 otherwise. As we proved in an exercise, in this case $\lambda_* = \cos(\pi/N) \approx 1 - \frac{\pi^2}{2N^2}$ (when N is large), so $\lambda_*^n \approx \exp(-\frac{\pi^2 n}{2N^2})$ (when n is large); and on the other hand $\pi_j = \frac{1}{N}\ \forall j \in S$. Therefore:

$$\frac{1}{2} \exp\left(-\frac{\pi^2 n}{2N^2}\right) \le \|P_0^n - \pi\|_{TV} \le \frac{\sqrt{N}}{2} \exp\left(-\frac{\pi^2 n}{2N^2}\right)$$

The upper and lower bounds become arbitrarily close to zero for $n = \Omega(N^2)$ and $n = O(N^2 \log N)$, respectively. Thus the bounds are tight, and describe well the true behaviour of the TV distance.

Example 1.5 (Lazy walk over the binary cube)**.** Fix a number d and consider the following variant of the Ehrenfest urns example: there are d balls numbered from 1 to d, partitioned over two urns which are labeled '0' and '1'. At each step, we pick a number t between 0 and d uniformly at random: if we pick $t = 0$, we do nothing; else we move ball t to the opposite urn. Each possible configuration of the balls is described by a vector $x \in \{0,1\}^d$, where x_t indicates which urn contains ball t. Hence the state space is $S = \{0,1\}^d$, with $N = |S| = 2^d$, and the transition probabilities are $p_{xy} = \frac{1}{d+1}$ if $x = y$ or x and y differ in exactly one coordinate, 0 otherwise.

Another way to look at this problem is as a lazy random walk over the d-dimensional binary cube. We make the walk lazy (i.e. with self-loops) because otherwise it would be periodic. We denote by 0 the all-zero vector in $\{0,1\}^d$, and index the 2^d eigenvalues and eigenvectors with elements $z \in \{0,1\}^d$.

Lemma 1.6. The eigenvalues and eigenvectors of the transistion probability matrix are

$$\lambda_z = 1 - \frac{2|z|}{d+1}, \text{ where } |z| = \text{number of non-zero components in } z$$

and

$$\phi_x^{(z)} = (-1)^{z \cdot x} \ \forall x \in \{0,1\}^d, \text{ where } z \cdot x = \textstyle\sum_{1 \le t \le d} z_t x_t$$

Proof.

$$(P\phi^{(z)})_x = \sum_{y \in S} p_{xy} \phi_y^{(z)} = \frac{1}{d+1} \phi_x^{(z)} + \frac{1}{d+1} \sum_{1 \le t \le d} \phi_{x+e_t}^{(z)}$$

where

$$\phi_{x+e_t}^{(z)} = (-1)^{z \cdot (x+e_t)} = (-1)^{z \cdot x} (-1)^{z \cdot e_t} = \phi_x^{(z)} (-1)^{z_t}$$

Thus

$$(P\phi^{(z)})_x = \frac{1}{d+1} \phi_x^{(z)} \left(1 + \sum_{1 \le t \le d} (-1)^{z_t} \right) = \frac{1}{d+1} \phi_x^{(z)} (1 + d - 2|z|) = \left(1 - \frac{2|z|}{d+1} \right) \phi_x^{(z)}$$

which proves that

$$P\phi^{(z)} = \left(1 + \frac{2|z|}{d+1} \right) \phi^{(z)}$$

□

Notice in particular that $\lambda_0 = 1$, $\phi^{(0)} = (1, \cdots, 1)^T$, and $|\phi_x^{(z)}| = 1, \ \forall x, z \in \{0,1\}^d$. The eigenvalues have high multiplicities: for $1 \le t \le d$, the eigenvalue $\lambda = 1 - \frac{2t}{d+1}$ corresponds to $\binom{d}{t}$ eigenvectors, namely all those $\phi^{(z)}$ with $|z| = t$. The eigenvectors are also orthogonal:

$$(\phi^{(z)})^T \phi^{(w)} = \sum_{x \in S} \phi_x^{(z)} \phi_x^{(w)} = \sum_{x \in S} (-1)^{z \cdot x} (-1)^{w \cdot x} = \sum_{x \in S} (-1)^{x \cdot (z+w)} = \begin{cases} 2^N & \text{if } z = w \\ 0 & \text{otherwise} \end{cases}$$

Finally, in this case, we obtain $\lambda_* = 1 - \frac{2}{d+1}$, so $\lambda_*^n \approx \exp(-\frac{2n}{d+1})$ for large d and n. On the other hand, the limiting distribution is uniform, so $\pi_x = 2^{-d} \ \forall x \in S$. Therefore,

$$\frac{1}{2} \exp\left(-\frac{2n}{d+1} \right) \le \|P_0^n - \pi\|_{TV} \le 2^{\frac{d}{2}-1} \exp\left(-\frac{2n}{d+1} \right)$$

The upper and lower bounds become arbitrarily close to zero for $n = \Omega(d)$ and $n = O(d^2)$, respectively. Thus, the bounds are loose, and do not really capture the true behaviour of the TV distance. It turns out that in this example, the TV distance behaves in an unexpected way.

1.4 Cut-off phenomenon

When the cut-off phenomenon occurs, the value of the TV distance does not decay smoothly and exponentially, but rather it rapidly drops in a short interval. More concretely, there is a mixing time τ such that the distance remains close to 1 when $n < \tau$, and rapidly converges to 0 when $n > \tau$. We will prove that the last example observes the cut-off phenomenon, with mixing time $\tau = \frac{d+1}{4} \log d$.

Proposition 1.7. Let c be a large positive constant. In the last example:

$$\text{If } n = \frac{d+1}{4}(\log d + c), \text{ then } \|P_0^n - \pi\|_{TV} \to 0 \text{ as } c \text{ increases.}$$
$$\text{If } n = \frac{d+1}{4}(\log d - c), \text{ then } \|P_0^n - \pi\|_{TV} \to 1 \text{ as } c \text{ increases.}$$

Proof of the first statement. Assume that $n = \frac{d+1}{4}(\log d + c)$. In what follows, the first inequality was proved in lecture notes 8, within the proof of the main theorem:

$$\begin{aligned}
\|P_0^n - \pi\|_{TV} &\le \frac{1}{2}\left(\sum_{z \in S\setminus\{0\}} \lambda_z^{2n} \underbrace{\left(\phi_0^{(z)}\right)^2}_{=1}\right)^{1/2} = \frac{1}{2}\left(\sum_{t=1}^{d} \binom{d}{t}\left(1 - \frac{2t}{d+1}\right)^{2n}\right)^{1/2} \\
&\le \frac{1}{2}\left(2\sum_{t=1}^{\lceil d/2\rceil} \binom{d}{t}\left(1 - \frac{2t}{d+1}\right)^{2n}\right)^{1/2} \le \frac{1}{\sqrt{2}}\left(\sum_{t=1}^{\lceil d/2\rceil} \frac{d^t}{t!}\exp\left(-\frac{4tn}{d+1}\right)\right)^{1/2} \\
&\le \frac{1}{\sqrt{2}}\left(\sum_{t=1}^{\infty} \frac{1}{t!}\exp\left(t\log d - \frac{4tn}{d+1}\right)\right)^{1/2} = \frac{1}{\sqrt{2}}\left(\sum_{t=1}^{\infty} \frac{1}{t!}e^{-tc}\right)^{1/2} \\
&= \frac{1}{\sqrt{2}}\left(\exp(e^{-c}) - 1\right)^{1/2} \approx \frac{1}{\sqrt{2}}\left(e^{-c}\right)^{1/2} = \frac{1}{\sqrt{2}}e^{-c/2}
\end{aligned}$$

Notice that the final expression approaches 0 exponentially as c increases, and does not depend on d. □

Proof of the second statement. Assume that $n = \frac{d+1}{4}(\log d - c)$. We use the second equivalent definition of the TV distance:

$$\|P_0^n - \pi\|_{TV} = \max_{A \subset S} |P_0^n(A) - \pi(A)| \ge |P_0^n(A) - \pi(A)|, \quad \forall A \in S$$

To obtain a good lower bound, we want to pick an appropriate set $A \subset S$ such that $P_0^n(A) \approx 0$ (for the chosen value of n) and $\pi(A) \approx 1$. As we will see, such a set is in the "center" of S: observe indeed that if the random variable X_∞ is distributed according to π, then $\mathbb{E}(|X_\infty|) = \frac{d}{2}$. The idea is therefore to include in A all states x with $|x| \approx \frac{d}{2}$. More concretely, we define $f : S \to \mathbb{Z}$ as $f(x) = d - 2|x| = \sum_{t=1}^{d}(-1)^{x_t}$, and $A = \{x \in S : |f(x)| \le \beta\sqrt{d}\} = \{x \in S : \left||x| - \frac{d}{2}\right| \le \frac{\beta}{2}\sqrt{d}\}$, where β is a parameter to be defined later.

Claim 1. $\pi(A) \ge 1 - \beta^{-2}$.

Proof.

$$\pi(A) = \mathbb{P}(X_\infty \in A) = \mathbb{P}\left(|f(X_\infty)| \le \beta\sqrt{d}\right) = 1 - \mathbb{P}\left(|f(X_\infty)| > \beta\sqrt{d}\right) \ge 1 - \frac{\mathbb{E}\left(f^2(X_\infty)\right)}{\beta^2 d}$$

by Chebychev's inequality. The expected value of $f^2(X_\infty)$ is given by

$$\mathbb{E}(f^2(X_\infty)) = \sum_{x\in S} f^2(x)\,\pi_x = 2^{-d}\sum_{x\in S}\left(\sum_{t=1}^{d}(-1)^{x_t}\right)^2 = 2^{-d}\sum_{s,t=1}^{d}\sum_{x\in S}(-1)^{x_s}(-1)^{x_t}$$
$$= 2^{-d}\sum_{s,t=1}^{d}(\phi^{(e_s)})^T\phi^{(e_t)} = 2^{-d}\sum_{s,t=1}^{d}2^d\delta_{s,t} = 2^{-d}\,d\,2^d = d$$

So finally, we obtain $\pi(A) \geq 1 - \frac{d}{\beta^2 d} = 1 - \beta^{-2}$. □

Claim 2. $P_0^n \leq (e^{c/2} - \beta)^{-2}$.

Proof. In order to show that $P_0^n(A)$ is small for the chosen value of n (i.e. in order to show that in n steps, the chain does not have the time, starting from position 0, to reach the "center" of the state space S), we need to analyse the distribution of the random variable X_n conditioned on the starting point $X_0 = 0$. For this, it is convenient to use $\mathbb{P}_0(\cdot)$, $\mathbb{E}_0(\cdot)$ and $\mathrm{Var}_0(\cdot)$ as shorthand notations for $\mathbb{P}(\cdot|X_0 = 0)$, $\mathbb{E}(\cdot|X_0 = 0)$ and $\mathrm{Var}(\cdot|X_0 = 0)$, respectively. We then obtain

$$P_0^n(A) = \mathbb{P}_0(X_n \in A) = \mathbb{P}_0\left(|f(X_n)| \leq \beta\sqrt{d}\right) = \mathbb{P}_0\left(|f(X_n) - \mathbb{E}_0\left(f(X_n)\right) + \mathbb{E}_0\left(f(X_n)\right)| \leq \beta\sqrt{d}\right)$$
$$\leq \mathbb{P}_0\left(|f(X_n) - \mathbb{E}_0\left(f(X_n)\right)| \leq \beta\sqrt{d} - \mathbb{E}_0\left(f(X_n)\right)\right) \leq \frac{\mathrm{Var}_0\left(f(X_n)\right)}{\left(\beta\sqrt{d} - \mathbb{E}_0\left(f(X_n)\right)\right)^2}$$

where we have again used Chebychev's inequality. The expectation can be computed as follows:

$$\mathbb{E}_0\left(f(X_n)\right) = \sum_{x\in S} p_{0x}(n)f(x) = \sum_{x\in S} p_{0x}(n)\sum_{t=1}^{d}(-1)^{x_t} = \sum_{t=1}^{d}\sum_{x\in S} p_{0x}(n)\phi_x^{(e_t)} = \sum_{t=1}^{d}\left(P^n\phi^{(e_t)}\right)_0$$
$$= \sum_{t=1}^{d}\lambda_{e_t}^n\phi_0^{(e_t)} = \sum_{t=1}^{d}\left(1 - \frac{2}{d+1}\right)^n = d\left(1 - \frac{2}{d+1}\right)^n \approx d\,\exp\left(-\frac{2n}{d+1}\right)$$
$$= d\,\exp\left(\frac{c - \log d}{2}\right) = \sqrt{d}\,e^{c/2}$$

In a similar manner, the variance is given by

$$\mathrm{Var}_0(f(X_n)) = \mathbb{E}_0(f(X_n)^2) - \mathbb{E}(f(X_n))^2 \approx \sum_{x\in S} p_{0x}(n)f(x)^2 - d\,e^c = \sum_{x\in S} p_{0x}(n)\sum_{s,t=1}^{d}(-1)^{x_s+x_t} - d\,e^c$$
$$= \sum_{s,t=1}^{d}\sum_{x\in S} p_{0x}(n)\phi_x^{(e_s+e_t)} - d\,e^c = \sum_{s,t=1}^{d}\left(P^n\phi^{(e_s+e_t)}\right)_0 - d\,e^c$$
$$= \sum_{t=1}^{d}\lambda_0^n\phi_0^{(0)} + \sum_{s\neq t}\lambda_{e_s+e_t}^n\phi_0^{(e_s+e_t)} - d\,e^c = d + d(d-1)\left(1 - \frac{4}{d+1}\right)^n - d\,e^c$$
$$\approx d + d(d-1)\exp(c - \log d) - d\,e^c \approx d$$

for large d. Gathering these last three computations together, we obtain

$$P_0^n(A) \leq \frac{d}{\left(\beta\sqrt{d} - \sqrt{d}\,e^{c/2}\right)^2} = (e^{c/2} - \beta)^{-2}$$

Joining the two claims together finally leads to the inequality

$$\|P_0^n - \pi\|_{TV} \geq |P_0^n(A) - \pi(A)| = \pi(A) - P_0^n(A) \geq 1 - \beta^{-2} - (e^{c/2} - \beta)^{-2}$$

which can be made arbitrarily close to 1 by first chosing β large and then c such that $e^{c/2} \gg \beta$. □

Random Walks: WEEK 10

1 Card Shuffling

1.1 Motivation

- Comparaison between computer and man dealt cards rejected uniformity for the man dealt cards as for the distribution of the 4 suits. This has an impact in competitive play.
- How to insure our deck is well shuffled? Sharp transition phenomenon. Dovetail shuffle requires 7 shuffles, whereas the overhand shuffle requires 2500 for a regular deck

We will analyze two different ways to shuffle cards and discuss the rate of convergence of $\|P^n - \pi\|_{TV}$ to 0, as well as associated cut off phenomenon.

We will always work on a finite state space.

1.2 Preliminaries

Definition 1.1. Let S be a finite set and μ and ν be probability distributions on S with $X \sim \mu$ and $Y \sim \nu$. A *coupling* is a probability ξ on $S \times S$ with marginals μ and ν, i.e. $\sum_y \xi(x,y) = \mu(x)$ and $\sum_x \xi(x,y) = \nu(y)$.

Remark 1.2. Recall that $\|\mu - \nu\|_{TV} = \sup_A |\mu(A) - \nu(A)|$.

Lemma 1.3 (Coupling Lemma)**.** *For all couplings as above, we have*

$$\|\mu - \nu\|_{TV} \leq \mathbb{P}(X \neq Y)$$

Moreover, there always is a coupling that achieves equality. (Exercise)

Proof. Let $\Delta := \{(s,s) : s \in S\}$ be the diagonal and ξ be a coupling of μ and ν.

$$\begin{aligned}
|\mu(A) - \nu(A)| &= |(\xi(A \times S) - \xi(S \times A)| \\
&= |\xi(A \times S \bigcap \Delta) + \xi(A \times S \bigcap \Delta^c) - \xi(S \times A \bigcap \Delta) - \xi(S \times A \bigcap \Delta^c)| \\
&= |\xi(A \times S \bigcap \Delta^c) - \xi(S \times A \bigcap \Delta^c)| \\
&\leq \max(\xi(A \times S \bigcap \Delta^c), \xi(S \times A \bigcap \Delta^c)) \leq \xi(\Delta^c) = \mathbb{P}(X \neq Y)
\end{aligned}$$

Both terms on the third line are less in absolute value than $\xi(\Delta^c)$, hence the result.

□

Definition 1.4. We define the distance function as $d(n) = \max_{i \in S} \|P_i^n - \pi\|$ where π is the stationary distribution.

2 Random to Top Shuffling

The current state of a deck of cards can be represented as an element of S_{52}, the group of permutation on 52 elements (the symetric group). By convention, i is the number of the card (say ace of spaces), and $\sigma(i)$ is it's position in the deck, 1 being the top.

At each step, pick a card uniformly at random (say the ace of spades), remove it and place it on top of the deck.

Now consider two decks $X_0 \sim Id$ and $Y_0 \sim Unif$. We let both decks evolve according to the RtT transition probabilities. In particular, Y_n is uniform for all times n.

Lemma 2.1. *Observe that once a card has been selected, it's rank in the deck is the same for both X and Y.*

At step n, if ace of space is selected, $X_n(Ace) = Y_n(Ace) = 1$. If it has been picked before, all cards above it are the same in both decks.

- A card above the Ace is picked. Since the cards above the Ace are the same in both decks, $X_n(Ace) = X_{n-1}(Ace) = Y_{n-1}(Ace) = Y_n(Ace)$.
- A card below the Ace is picked (it is below in both decks) and so $X_n(Ace) = X_{n-1}(Ace) + 1 = Y_{n-1}(Ace) + 1 = Y_n(Ace)$.

In particular, letting $\tau = \inf\{n \geq 0 : \text{all cards have been selected at least once}\}$, we have $X_\tau = Y_\tau$.

$$d(n) \leq \mathbb{P}(X_n \neq Y_n) \leq \mathbb{P}(\tau > n)$$

This is the classical coupon collector problem: after having touched j cards, $0 \leq j \leq N-1$, the time until a new card is picked is G_j, a geometric random variable with parameter $1 - \frac{j}{N}$. (it will take time k with probability $\frac{j}{N}^{k-1}(1 - \frac{j}{N})$).

Using the linearity of expectation, we get

$$\mathbb{E}(\tau) = \sum \mathbb{E}(G_j) = \frac{N}{N} + \frac{N}{N-1} + \ldots + \frac{N}{1} \approx N \log N$$

Moreover, these G_j are independent and so we can sum up the variances:

$$Var(\tau) = \sum_j Var(G_j) = \sum_j \frac{j}{N}/(1 - \frac{j}{N})^2 = \sum_j \frac{Nj}{(N-j)^2}$$

Changing the index of summation to $i = N - j$ and summing of i, we get $N^2 \sum \frac{1}{i^2} - N \sum \frac{1}{i} \leq KN^2$.

With $n = (1 + \varepsilon)\mathbb{E}\tau$ and Chebyshev's inequality, we get

$$\mathbb{P}(\tau > (1+\varepsilon)N \log N) \simeq \mathbb{P}(\tau - \mathbb{E}(\tau) > \varepsilon N \log N) \leq \frac{Var(\tau)}{(N \log N)^2} = \frac{K}{\varepsilon^2 (\log N)^2}$$

which tends to zero for large N, so

$$d(n) \leq \mathbb{P}(\tau > n) \to 0$$

□

2.1 Lower bound

Let $i_1, i_2, \dots i_j$ be the j bottom cards, ordered from the top. We will now look at the event $A_j := \{i_1 > i_2 > \dots > i_j\}$.

Since the stationary distribution is uniform, we have $\pi(A_j) = \frac{1}{j!}$.

However, if many cards have not been touched yet, this event has a high probability for the deck X. The untouched cards are all at the bottom and conserve their initial ordering.

Observe that $\mathbb{P}(A_j) \geq \mathbb{P}(j$ bottom cards have not been touched$\}$.

Using the same kind of reasoning as for the upper bound, we get the following lemma.

Lemma 2.2. *Let $\varepsilon < 1$. For any $\varepsilon < \delta < 1$, if $n = \varepsilon N \log N$, at least $N^{1-\delta}$ cards have not been touched by time n with probability going to 1 as $N \to \infty$.*

Indeed, $\mathbb{P}$(bottom j cards have not been touched) $= 1 - \mathbb{P}(N - j$ cards have been touched). As before, this last event is a sum of geometric random variables. The hitting time τ' for $N - j$ cards to have been touched has expected value $\mathbb{E}(\tau') = N(\log N - \log j)$ and $Var(\tau') = KN^2$. So

$$\mathbb{P}(\tau' - N \log N \geq \varepsilon N \log N) \leq \frac{KN^2}{(\varepsilon N \log N)^2} = \frac{K}{\varepsilon^2 (\log N)^2}$$

□

Let now $n = (1 - \varepsilon) N \log N$ and A be the event that N^ε of the bottom cards are ordered. By the above lemma, $\mathbb{P}(X_n \in A) = 1 - o(1)$, while for a uniformly distributed permutation, $\pi(A) = 1/(N^\varepsilon!)$. We conclude

$$d(t) \geq |\mathbb{P}(X_n \in A) - \mathbb{P}(Y_n \in A)| = |1 - o(1) - \frac{1}{N^\varepsilon!}| \to 1$$

This proves the lower bound. □

3 Dovetail Shuffles

We will describe the Gilbert-Shannon-Reeds probabilistic model of card shuffling, which empirically matches well with the way people do a dovetail shuffle. (we could also discuss a-shuffles with only minor changes (GSR are 2-shuffles).

1. GSR: A deck of N cards is split in 2 decks, the size of the left deck being a binomial with parameters $\frac{1}{2}$ and N ($N = 52$). The two packets are then mixed together, the probability of a card coming from the given deck being proportional to the size of the deck: i.e. if the packet sizes are A and B, then the next card will come from the left hand packet with probability $\frac{A}{A+B}$.
2. Uniform: All possible ways of cutting a deck into 2 packets and then interleaving the packets are equally likely.
3. Geometric: Place N points on the interval $[0, 1]$, independently and uniformly. The map $x \to 2x (mod\ 1)$ maps $[0, 1]$ to itself and rearranges the points. This induces a permutation and induces a probability distribution on S_N.
4. Inverse: Go through the deck, assigning values of 0 or 1 uniformly and independently to each cards. Assemble the cards according to their value.

GSR and Inverse models of card shuffling are equivalent:

Claim 1: The size of the deck is always binomial
-GSR: It is clear that the deck size is always binomial.
-Inverse: The deck size is binomial, because sum of iid. Bern(1/2) is binomial.

Claim 2: All possible interleavings are equally likely given the packet sizes
-Inverse: This claim is clear in inverse shuffling.
-GSR: The chance of any specific left right drop is $\frac{A(A-1)...1B(B-1)...1}{A+B(A+B-1)...1} = \binom{A+B}{A}^{-1}$.

3.1 Cutoff time

We will show that the cutoff time τ is between $\log N$ and $2\log N$ ($\log N \leq \tau \leq 2\log N$). More sophisticated arguments would allow us to identify the constant as 3/2.

It turns out that the inverse shuffle is much easier to analyse. Let X'_n and X_n be the Markov chains associated with the inverse and the forward shuffles, respectively. We use the group structure to observe that

$$X'_n = g'_1 g'_2 \dots g'_n = g_1^{-1} \dots g_n^{-1} = (g_n \dots g_1)^{-1}$$

We see that $X_n^{-1} =_d X'_n$. We deduce that the mixing time of the forward and backward shuffles are the same, and even $d(n) = d'(n)$

At each step (shuffle), we are assigning iid 0-1 random variables to each card. Hence each card c is assigned a binary expansion $D(c)$ and the cards are ordered according to this binary expansion (two cards retain the original ordering if they have the same binary expansion). Since these binary expansions (of length the number of shuffles n) are iid, if they all are different, the deck is uniformly distributed. This is a reformulation of the birthday problem, and leads to the conclusion that if $n >> 2\log N$ (i.e. $n = (2+\varepsilon)\log N$), then all labels are distinct.

Now recall $d(n) < \mathbb{P}(\tau' > n)$ with $T\tau'$ the first time at which all labels $D_n(c)$ are distinct. Moreover, by the above reasoning, $\mathbb{P}(T > n) \to 0$ if $n = (2+\varepsilon)\log N$.

3.2 Lower bound (heuristic)

We will look at the number of descents. We say that σ has a descent at j if $\sigma(j) > \sigma(j+1)$. It turns out that the probability of a given permutation under the inverse shuffle depends uniquely on the number of descents. We will expand on this point later.

We will argue that if $n = (1-\varepsilon)\log N$, then the expected number of descents for X'_n and U are different, and so they cannot be close in total variation (the number of descents, as well as the number of rising sequences, is a sufficient statistic).

Lemma 3.1. *For σ uniformly distributed, $\mathbb{E}(Des(\sigma)) = (N-1)/2$ and $Var(Des(\sigma)) \sim N/12$.*

Consider the deck X'_n. Each distinct binary expansion can only create one descent (cards with the same binary expansion retain their original ordering, so no descent). Therefore $Des(X'_n) \leq 2^n$. We conclude that if the number of shuffles $n = (1-\varepsilon)\log N$, then $Des(X'_n) \leq N^{1-\varepsilon}$, and this is incompatible with the fact that for a uniform distribution of the deck, we have lemma 3.1. The two distributions concentrate on permutations with very different number of descents, hence the total variation distance must be large.

We have a very accurate description of the probability of a given permutation.

Definition 3.2. The *rising sequences* of a permutation are the maximal subsets of successive card labels, in increasing order from left to right.

Theorem 3.3. *For a deck initially ordered, we have after n shuffles,*

$$\mathbb{P}(X_n = \sigma) = \frac{1}{2^{nN}} \binom{2^n + N - R(\sigma)}{N}$$

where $R(\sigma)$ is the number of rising sequences/

Using this, Bayer and Diaconis (1992) were able to analyze in great detail the behaviour around the cutoff time of $3/2 \log N$.

Theorem 3.4. *Let $n = \log(N^{3/2}) + c.$. Then*

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

where Φ is the cumulative distribution function of a standard Gaussian random variable.

Random Walks: WEEK 11

1 Markov chain Monte Carlo algorithms

MCMC are a class of algorithms that allow to sample from distributions that typically have large state spaces or are analytically intractable. This is a very lively subject of research still nowadays. The most well known and first MCMC algorithm goes back to work in the context of physics N. Metropolis, A.W. Rosenbluth, M.N. Teller and E. Teller in 1953.[1] The original method was then further developed by W. K. Hastings in 1970.[2] Here we will focus on the Metropolis chain. Another popular and very important MCMC algorithm based on the "Glauber dynamics" devised by R. Glauber in 1963 in the context of Ising spin models[3]. Glauber dynamics will form the subject of the exercises. There exist various other MCMC algorithms that we will not discuss at all.

Before introducing the Metropolis chain we first wish to recall the traditional sampling method and the basic application to Monte Carlo integration.

1.1 Traditional sampling method

Let X be a random variable following a Bernoulli distribution of parameter p. We have $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$. We assume that we know how to generate a random variable $\mathcal{U}$ uniformly distributed in $[0; 1]$. One way to generate a value for X is to generate $\mathcal{U}$ and set $X = 0$ if $\mathcal{U} \leq p$ and $X = 1$ if $\mathcal{U} > p$.

More generally, if we want to sample from the distribution of a variable X such that $\forall i \in \{0; 1; 2; 3\}, \mathbb{P}_X(X = i) = p_i$, we can set:

$$X = \begin{cases} 0, & \text{if } \mathcal{U} \leq p_0. \\ 1, & \text{if } p_0 < \mathcal{U} \leq p_0 + p_1. \\ 2, & \text{if } p_0 + p_1 < \mathcal{U} \leq p_0 + p_1 + p_2. \\ 3, & \text{if } p_0 + p_1 + p_2 < \mathcal{U}. \end{cases} \tag{1}$$

Definition 1.1. The cumulative distribution of X is $F_X(x) = \mathbb{P}_X(X \leq x)$.

We observe that what we have written more or less amounts to set $X = F_X^{-1}(\mathcal{U})$.

This is the good idea to generalize to the case were we want to sample from a continuous distribution (that is $\forall x \in [a; b], \mathbb{P}_X(X = x) = p(x)$). Indeed $\mathbb{P}(F_X^{-1}(\mathcal{U}) \leq x) = \mathbb{P}(\mathcal{U} \leq F_X(x)) = F_X(x)$ (because $\mathcal{U}$ is uniform). So X and $F_X^{-1}(\mathcal{U})$ have the same cumulative distribution, so they also have the same distribution, and we can sample from X using $F_X^{-1}(\mathcal{U})$.

1.1.1 Application: Monte Carlo integration

Given a function f and a probability distribution p on $[a, b]$, we want to compute $I = \int_a^b f(x)p(x)dx$. For this, we sample $x_1, ..., x_n$ from the probability distribution p, and we compute $\hat{I} = \frac{1}{n}\sum_{j=1}^n f(x_j)$.

We have $\mathbb{E}_{x_1,...,x_n \sim p}(\hat{I}) = \mathbb{E}_{x \sim p}(f(x))$, and we can prove that $\mathrm{Var}(\hat{I}) = \frac{1}{n}\mathrm{Var}(f(x))$.

It is possible, to generalize to functions of two or more variables. However, computing the cumulative distribution can quickly become impractical, as we are going to see in the examples below and the traditional sampling methods are not good enough.

[1] "Equations of State Calculations by Fast Computing Machines". Journal of Chemical Physics 21 (6): 1087 - 1092.
[2] "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". Biometrika 57 (1): 97 - 109.
[3] "Time dependent statistics of the Ising model", J. Math. Phys. 4 (1963) 294

1.2 Typical models and distributions with large state spaces

Example 1.2. The first example is related to the coloring problem. For this problem, we consider a graph $G = (V, E)$ with $V = \{1, 2, ..., N\}$ and an alphabet of color $\{1, ..., q\}$. We want to color the vertices so that no two adjacent vertices share the same color. Let $x_v \in \{1, ..., q\}$ be the color of vertex v. What we want is $\forall (v, w) \in E, x_v \neq x_w$. A coloring that respect this rule is called a *proper* coloring or an *allowed* coloring.

It is possible to show that if the maximum degree Δ of a vertice is such that $\Delta + 1 \leq q$, there exist at least one proper coloring (later we will consider q is sufficiently greater than Δ so that the existence of proper colorings is not an issue).

Now, the probability distribution from which we want to sample is the uniform probability distribution on the set Ω of proper colorings:

$$\mathbb{P}_X(X = (x_v)_{v\in V}) = \frac{\mathbf{1}((x_v)_{v\in V} \text{is a proper coloring})}{|\Omega|}. \tag{2}$$

The problem of identifying the set of all proper colorings is hard. For example we do not even know how to determine the cardinality of this set $|\Omega|$. So the distribution, although uniform on Ω, is falsely simple, and sampling is not obvious.

Example 1.3. For the second example, we will consider the Ising model. We have a graph $G = (V, E)$ with $V = \{1, 2, ..., N\}$, and an alphabet $\sigma = \{1, -1\}$. Variables σ_v are "attached" to the vertices $v \in V$. These variables are called spins. We consider spin assignments $\underline{\sigma} = (\sigma_1, \ldots, \sigma_N) \in \{1, -1\}^N$. If you wish you can think of those as functions from V to the state space $S = \{1, -1\}^N$. The distribution we are interested in is:

$$\mu(\underline{\sigma}) = \frac{1}{Z} \exp(\beta(\sum_{(v,w)\in E} J_{vw}\sigma_v\sigma_w + \sum_{v\in V} h_v\sigma_v)), \tag{3}$$

where $\beta > 0$, $J_{vw}, h_v \in \mathbb{R}$ and Z is a normalization constant (called partition function), that is:

$$Z = \sum_{\underline{\sigma}\in S} \exp(\beta(\sum_{(v,w)\in E} J_{vw}\sigma_v\sigma_w + \sum_{v\in V} h_v\sigma_v)) \tag{4}$$

Nobody knows how to compute Z for large N because it is given by a sum over 2^N states (except for special graphs like a one-dimensional line or a tree). Here again, it is not obvious to sample from this distribution.

Origin of the Ising model: The Ising model (1920) was originally invented to model magnetic materials and such models are of utmost importance in physics and statistical mechanics. The vertices of the graph are the atoms of a crystal[4], the spins are the magnetic moments of the atoms. Each magnetic moment behaves like a little magnet oriented south-north = +1 or north-south= -1. When the material is at a temperature β^{-1} the magnetic moments behave like random variables distributed according to the Gibbs distribution $\mu(\underline{\sigma})$. The real numbers J_{vw} are related to the mutual interaction between neighboring magnetic moments, and h_v is a bias related to their interaction with magnetic fields. But these models have also found interpretations and applications independent from physics, e.g. in image processing, social networks, voter models etc. They are also much studied by mathematicians, specially in probability theory.

Voter model interpretation: Let us briefly give a simplistic voter model interpretation. Each vertex v is a person that votes $\sigma_v = \pm 1$ (think of your favorite yes/no societal issue). Persons v and w related by an edge are friends ($J_{vw} > 0$) or enemies ($J_{vw} < 0$). Persons not related by an edge don't know each other. In our model the society is a bit simple minded: friends with $J_{vw} > 0$ tend to vote similarly, while enemies with $J_{vw} < 0$ tend to vote in opposite ways. The biases h_v (positive or negative) may model

[4]For the simplest crystalline arrangement of atoms the graph would be a cubic grid of size $L \times L \times L = N$ where L is the linear dimension of the sample.

a prior opinion that each person has a priori and influences his decision. Sampling from $\mu(\underline{\sigma})$ would correspond to evaluate the voting pattern of the population.

1.3 General philosophy of MCMC (Markov Chain Monte Carlo) algorithms

We have a distribution $\mathbb{P}(X = i)$ for $i \in S$ in some huge state space. We think of $\mathbb{P}$ as the stationary distribution of a Markov chain and call it π_i.

Now, we are confronted with an inverse problem: given π, we want to construct an ergodic Markov chain P such that π is a limit stationary distribution. Our sampling method will be to start at an initial condition i_0 and let the chain run for n time steps. This gives a sample i_n (for each run). Since $(P^n)_{ij} \to \pi_j$, as $n \to +\infty$, the samples i_n are approximately distributed according to π.

1.4 Metropolis chain (1953)

We consider a state space S and a *base Markov chain* with transition probabilities ψ_{ij}. We start at some state i, and generate a next state j with distribution ψ_{ij}. Then, we accept the move $i \to j$ with a probability a_{ij} that we will have to define (this means in particular that we stay in state i with probability $1 - a_{ij}$). Therefore,

$$p_{ij} = \begin{cases} \psi_{ij} a_{ij}, & \text{if} i \neq j. \\ \psi ii + \sum_{j \neq i} \psi_{ij}(1 - a_{ij}) = 1 - \sum_{k \in S \setminus i} \psi_{ik} a_{ik}, & \text{if } i = j. \end{cases} \tag{5}$$

We want $P_{ij}^{(n)} \to \pi_j$.

In particular, we want a stationary distribution to exist. The *Metropolis rule* is a choice of a_{ij} that ensures the existence of the stationary distribution by satisfying a stronger condition, namely the detailed balance equations: $\pi_i p_{ij} = \pi_j p_{ji}$.

In the case of a symmetric base chain $\psi_{ij} = \psi_{ji}$ the a_{ij} are defined by the following equation:

$$a_{ij} = \min(1, \frac{\pi_j}{\pi_i}) \tag{6}$$

For non-symmetric base chain the generalization (due to Hastings) is

$$a_{ij} = \min(1, \frac{\pi_j \psi_{ji}}{\pi_i \psi_{ij}}) \tag{7}$$

The intuition behind this is that a transition $i \to j$ should be accepted if we want j to be more likely than i. Otherwise, we allow the transition with some probability, to avoid getting stuck in a local extremum of π.

Verification of detailed balance condition: It is easy to verify that the detailed balance equations are verified (we leave this check to the reader; see also exercises).

Convergence of Metropolis chain: If the resulting Metropolis chain is irreducible then, since we have ensured existence of a stationary distribution, we are sure that it is positive recurrent (see first fundamental theorem lecture 4). Moreover if the chain is also aperiodic we can apply the ergodic theorem: $(P^n)_{ij} \to \pi_i$ as $n \to +\infty$ (see ergodic theorem lecture 4).

Thus when the Metropolis chain is irreducible and aperiodic it is ergodic. This allows to produce samples distributed according to π by letting the chain run long enough (each run gives one sample). This is basic idea of MCMC. However this does not tell us how long we should let the chain run in order to produce non-biased samples. We will adress this problem in the next lecture.

Remarks: We do not give here general conditions that ensure irreducibility and aperiodicity of the Metropolis chain. In pratice this is best checked on a case by case basis. We point out that irreducibility is not always ensured. In fact for many theoretical computer science problems it is not always possible to satisfy this condition and MCMC may fail. For example in coloring this may happen if the state space of proper colorings is not "connected". Aperiodicity is a less serious problem because the rules often allow for self-loops. Otherwise one can modify them by introducing ad-hoc self-loops with small probabilities.

Example 1.4. General ideas on optimization This toy example serves to illustrate some general and important ideas that find their origin in the Metropolis rule.

We consider some function $f(i)$ for $i \in \mathbb{Z}$ that we want to minimize (in general the function can live on a complicated state space of course). Here we have in mind a function that is bounded below, goes to infinity for $i \to \pm\infty$ and has a huge number of local minima. The global minimum might be attained at one or many points also.

We want to search the global minima but this is not easy because many algorithms get stuck in local minima. For example a local greedy search (that goes in a direction if the function if the value of the function is smaller in that direction), will only find a local minimum (except if by chance you started close enough to a global minimum). Indeed once you reach a local minimum with a purely local greedy search you cannot go lower.

What we want to do is sample from:

$$\pi_\infty(i) = \frac{\mathbf{1}(i \in \text{set of global minima})}{\#\text{global minima}}$$

Instead we will sample from

$$\pi_\beta(i) = \frac{e^{-\beta f(i)}}{Z}$$

where Z is a normalization constant

$$Z = \sum_i e^{-\beta f(i)}$$

If β is large enough number, π_β is close to π_∞ and sampling from π_β gives us an idea of the global minima of $f(.)$.

The choice of the base chain ψ_{ij} does not really matter as long as the resulting (Metropolis) chain is irreducible and aperiodic. For instance, we can take the chain for which at each state i, there is a $1/2$ probability to go at $i+1$ and a $1/2$ probability to go at $i-1$.

Now, according to the Metropolis rule, if we are about to go from i to $j \in \{i+1, i-1\}$:

- If $f(i) > f(j)$, we accept the transition.
- otherwise, we accept the transition only with probability $e^{-\beta(f(j)-f(i))}$

It is not difficult to convince one-self that the resulting chain is irreducible. There is also a probability that we do not accept a move, thus there are self-loops which ensure aperiodicity.

How to choose β? On one hand we want β large to approximate the original problem which is to sample π_∞. On the other hand with β "too large" we will "loose irreducibility" (strictly speaking at $\beta = +\infty$ you don't have irreducibility) and the mixing time will typically become so large that the chain becomes useless. Tuning β to the right value thus has to take into account these two conflicting goals and is an art.[5]

[5]Simulated annealing is an elaboration on these ideas. One follows a special schedule where one takes a sequence of "inverse temperatures" $\beta_1 < \beta_2 < \dots \beta_R$ and times durations $\tau_1, \tau_2, \dots, \tau_R$, and one runs the chain at inverse temperature β_1 for a duration τ_1, then at inverse temperature β_2 for a duration τ_2 etc. The name "annealing" comes from the process of fabrication of some materials by slowly cooling the material so that its atoms do not become stuck in undesirable states which could introduce defects. Simulated annealing is an important optimization procedure that was inspired from material science.

Let us give a ballpark estimate to choose β correctly. Note that this is just a qualitative idea which can only serve as a first guide when these ideas are applied to specific problems. To choose β, we decide that we want to spend a $1-\epsilon$ fraction of time in global minima. Recall (see e.g homework 4) that π_i is the average fraction of time that the chain spends in state i when it has reached the stationary distribution. Thus we set

$$1-\epsilon \approx \sum_{i \text{ global minimum}} \pi_\beta(i)$$

Let $f_0 = \min_{i\in\mathbb{Z}} f(i)$ be the global minimum and $f_1 = \min_{i\in\mathbb{Z}, f(i)\neq f_0} f(i)$, $f_2 = \min_{i\in\mathbb{Z}, f(i)\neq f_0, f_1} f(i)$, ... be the local minima. Let N_0, N_1, N_2, ... the number of points were the minima f_0, f_1, f_2, ... are reached. We have

$$\begin{aligned}
\sum_{i \text{ global minimum}} \pi_\beta(i) &= \frac{1}{Z} N_0 e^{-\beta f_0} \\
&\approx \frac{1}{1 + \frac{N_1}{N_0} e^{-\beta(f_1-f_0)} + \frac{N_2}{N_0} e^{-\beta(f_2-f_0)} + \dots} \text{ (we neglect the terms for which } f(i) > f_1) \\
&\approx 1 - \frac{N_1}{N_0} e^{-\beta(f_1-f_0)} \text{ (because } \beta \text{ is large and } f_1 - f_0 > 0)
\end{aligned}$$

This gives us the following constraint on β:

$$\beta \approx \frac{1}{f_1 - f_0} \log\left(\frac{N_0}{\epsilon N_1}\right) \tag{8}$$

2 Metropolis Algorithm for The Coloring Problem

Now we shall study the Metropolis algorithm for solving the problem we introduced in Example 1.2. Recall that as long as $q \geq \Delta + 1$ (q is the number of available colors and Δ is the maximum degree of the graph) there exist at least one proper coloring of the graph $G = (V, E)$ (in other words, $\Omega \neq \emptyset$). We assume $q \geq 3\Delta$ and derive upper-bounds on the mixing time of the Metropolis chain.

The Metropolis algorithm is as follows:

1. Start with a proper coloring $\underline{x} = (x_v, v \in V) \in \Omega$.
2. Select a vertex $v \in V$ uniformly at random (with probability $\frac{1}{N}$).
3. Select a color $c \in \{1, 2, \dots, q\}$ uniformly at random (with probability $\frac{1}{q}$).
4. If c is allowed at vertex v, set $x_v = c$, otherwise do nothing!

First of all observe that the above algorithm is a Metropolis algorithm. The base chain is the one that moves from a state $\underline{x}$ to $\underline{y}$ if they differ at most at one color with positive probability (this is a symmetric chain). The acceptance probability as we derived in (6) is:

$$a_{\underline{x}\to\underline{y}} = \min\left(1, \frac{\pi(\underline{y})}{\pi(\underline{x})}\right) = \mathbf{1}(\underline{y} \text{ is proper}),$$

(since $\pi(\underline{x}) = \frac{\mathbf{1}(\underline{x} \text{ is proper})}{|\Omega|}$) which is exactly what is done at step 4.

Another important question is what if we don't know a proper coloring to start with at step 1? In fact, it turns out that we could start with *any* coloring and the algorithm will still work.

The important challenge here is that the chain produced by the above algorithm is not guaranteed to be irreducible, hence may not be ergodic. We will prove bounds on the mixing time (for $q \geq 3\Delta$) of the chain without using the ergodic theorem.

Remark: In fact when one is able to prove constructive good bounds on the mixing time it is often the case that we directly obtain information that is much stronger than the ergodic theorem. In this respect it is good to keep in mind that the ergodic theorem is a "structure theorem" that does not by itself to prove constructive results on a chain.

Let

$$d(n) \triangleq \max_{\underline{x}\in\Omega} \|P^n(\underline{x}) - \pi\|_{\mathrm{TV}}$$

(where $P^n(\underline{x})$ is the row of the n step transition matrix indexed by $\underline{x}$ – i.e. the distribution of the state at time n starting at $\underline{x}$) and

$$T_\epsilon \triangleq \inf\{n \geq 0 : d(n) \leq \epsilon\}$$

be the mixing time of the chain.

Theorem 2.1. The mixing time of the Metropolis chain for the coloring is upper-bounded as

$$T_\epsilon \leq \frac{1}{1-3\Delta/q} N \left(\log(N) + |\log(\epsilon)|\right)$$

Proof. The proof is based on the idea of grand coupling. We start two chains at two different initial states $\underline{x}$ and $\underline{y}$ but we choose the same vertices v and colors c (at steps 2 and 3 of the algorithm) for both of them to move to the next states. We hence have two Markov chains $\underline{X}(n)$ and $\underline{Y}(n)$. We will show

$$\mathbb{P}\left(\underline{X}(n) \neq \underline{Y}(n)\right) \leq N e^{-\frac{n}{N}(1-3\Delta/q)} \tag{9}$$

Having shown (9), defining $T_c \triangleq \min\{m \geq 0 : \underline{X}(m) = \underline{Y}(m)\}$, we will have

$$\mathbb{P}(T_c > n | \underline{X}(0) = \underline{x}, \underline{Y}(0) = \underline{y}) \leq \mathbb{P}\left(\underline{X}(n) \neq \underline{Y}(n)\right) \leq N e^{-\frac{n}{N}(1-3\Delta/q)}.$$

Since

$$d(n) \leq \max_{\underline{x},\underline{y}} \mathbb{P}(T_c > n | \underline{X}(0) = \underline{x}, \underline{Y}(0) = \underline{y})$$

the proof will be complete.[6]

Let $\rho(\underline{x}, \underline{y}) \triangleq \sum_{v\in V} \mathbf{1}(x_v \neq y_v)$ denote the Hamming distance between $\underline{x}$ and $\underline{y}$.

Preliminary Attempt: Suppose $\underline{x}$ and $\underline{y}$ differ only at one vertex v_0 (hence $\rho(\underline{x}, \underline{y}) = 1$). Let's compute $\mathbb{E}[\rho(\underline{X}(1), \underline{Y}(1))]$. Since $\rho(\underline{X}(1), \underline{Y}(1)) \in \{0, 1, 2\}$,

$$\mathbb{E}[\rho(\underline{X}(1), \underline{Y}(1)) - 1] = (-1) \times \mathbb{P}(\rho(\underline{X}(1), \underline{Y}(1)) = 0) + 1 \times \mathbb{P}(\rho(\underline{X}(1), \underline{Y}(1)) = 2)$$

Equivalently,

$$\mathbb{E}[\rho(\underline{X}(1), \underline{Y}(1))] = 1 + (-1) \times \mathbb{P}(\rho(\underline{X}(1), \underline{Y}(1)) = 0) + 1 \times \mathbb{P}(\rho(\underline{X}(1), \underline{Y}(1)) = 2)$$

Now:

- $\rho(\underline{X}(1), \underline{Y}(1)) = 0$ if and only if the color of v_0 is changed (otherwise, whatever happens $\underline{X}(1)$ and $\underline{Y}(1)$ will still differ at v_0). Therefore,

$$\mathbb{P}(\rho(\underline{X}(1), \underline{Y}(1)) = 0) = \underbrace{\frac{1}{N}}_{\text{selecting } v_0} \underbrace{\frac{\#\text{ of allowed colors at } v_0}{q}}_{\text{choosing an allowed color}} \geq \frac{1}{N}\frac{q-\Delta}{q}$$

  Consequently,

$$\mathbb{E}[\rho(\underline{X}(1), \underline{Y}(1))] \leq 1 - \frac{1}{N}\frac{q-\Delta}{q} + 1 \times \mathbb{P}(\rho(\underline{X}(1), \underline{Y}(1)) = 2)$$

[6]The explicit proof of this very last inequality is ommitted here. A general proof follows by an application of the tools developped in previous lectures, and is not related to the coloring problem per se.

- $\rho(\underline{X}(1), \underline{Y}(1)) = 2$ if and only if at step 2 of the algorithm a vertex $w_0 \in \partial v_0$ (a neighbor of v_0) is chosen (otherwise both chains will either be recolored or remain unchanged – hence they will still differ at only one vertex v_0) and recolor $\underline{X}$ at w_0 and *don't* recolor $\underline{Y}$ at w_0 (or vice versa).

  Let ν be the set of colors at $\partial w_0 \setminus \{v_0\}$ (the neighborhood of w_0 excluding v_0). In order not to recolor $\underline{Y}$, we should pick $c \in \nu \cup \{y_{v_0}\}$. On the other side, to recolor $\underline{X}$ we should have $c \notin \nu \cup \{x_{v_0}\}$. This is included in the event that $c = y_{v_0}$.[7] Hence,

$$\mathbb{P}(\rho(\underline{X}(1), \underline{Y}(1)) = 2) \leq 2\frac{\Delta}{N}\frac{1}{q}$$

  (the factor of 2 comes from the other possibility of recoloring $\underline{Y}$ and keeping $\underline{X}$ unchanged). Hence:

$$\mathbb{E}[\rho(\underline{X}(1), \underline{Y}(1))] \leq 1 - \frac{1}{N}\frac{q-\Delta}{q} + 2\frac{\Delta}{N}\frac{1}{q} = 1 - \frac{1}{N}(1 - 3\Delta/q)$$

General Setting: Now suppose $\rho(\underline{x}, \underline{y}) = r$. We know that in the extended set of states, we have a sequence of neighbor states

$$\underline{x}_0 = \underline{x}, \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_{r-1}, \underline{x}_r = \underline{y}$$

such that $\rho(\underline{x}_{k-1}, \underline{x}_k) = 1$, for $\forall k = 1, 2, \ldots, r$. Suppose we run $r+1$ Metropolis grand coupled chains starting at all those $r+1$ states. Since

$$\rho(\underline{X}(1), \underline{Y}(1)) \leq \sum_{k=1}^{r} \rho(\underline{X}_{k-1}(1), \underline{X}_k(1))$$

(the triangle inequality) we have,

$$\mathbb{E}[\rho(\underline{X}(1), \underline{Y}(1))] \leq \sum_{k=1}^{r} \mathbb{E}[\rho(\underline{X}_{k-1}(1), \underline{X}_k(1))] \leq r\left[1 - \frac{1}{N}(1 - 3\Delta/q)\right]$$

using our previous results. Repeating the above, we can conclude that

$$\mathbb{E}[\rho(\underline{X}(n), \underline{Y}(n))] \leq r\left[1 - \frac{1}{N}(1 - 3\Delta/q)\right]^n \leq re^{-\frac{n}{N}(1-3\Delta/q)} \leq Ne^{-\frac{n}{N}(1-3\Delta/q)}$$

Finally, equation 9 follows by Markov inequality (and the fact that $\{\underline{X}(n) \neq \underline{Y}(n)\} = \{\rho(\underline{X}(n), \underline{Y}(n)) \geq 1\}$). □

[7] In other words we have shown that (the event) $\rho(\underline{X}(1), \underline{Y}(1)) = 2$ implies $c = y_{v_0}$.

Random Walks: WEEK 12

1 Coupling From The Past

Coupling From The Past is a method introduced by James Propp and David Wilson in 1996 which allows us to assert if a Markov chain has reached the stationary distribution or not.

In order to study this method, we need a tool called *random mapping representation of a Markov chain*. So far we used to define a (time-homogeneous) Markov chain by a matrix of transition probabilities $P = [p_{i\to j}]$ where $p_{i\to j} = \mathbb{P}(X_{n+1} = j | X_n = i)$. Alternatively one can represent a Markov chain as

$$X_{n+1} = \Phi(X_n, U_n)$$

where $\Phi(\cdot,\cdot)$ is a cleverly chosen function and $(U_n, n = 1, 2, \dots)$ is a sequence of intelligently chosen i.i.d random variables which is furthermore independent of $(X_n, n = 1, 2, \dots)$. It is clear that

$$p_{i\to j} = \mathbb{P}(\Phi(i, U_n) = j)$$

Proposition 1.1. Every Markov chain has a random mapping representation.

Proof. We assume U_ns are uniform random variables in $[0, 1]$ (we denote this as $U_n \sim \mathcal{U}[0, 1]$) and construct $\Phi(\cdot,\cdot)$ such that $p_{i\to j} = \mathbb{P}(\Phi(i, U_n) = j)$ for any arbitrary set of transition probabilities $p_{i\to j}$.

Define

$$F_{i\to k} \triangleq \sum_{j=1}^{k} p_{i\to j}, \qquad \forall i, k \in S$$

(where S is the state space) and set

$$\Phi(i, u) \triangleq j \cdot \mathbf{1}\left(F_{i\to j-1} < u \le F_{i\to j}\right).$$

We hence have:

$$\mathbb{P}(\Phi(i, U_n) = j) = \mathbb{P}(F_{i\to j-1} < U_n \le F_{i\to j}) = F_{i\to j} - F_{i\to j-1} = p_{i\to j}. \qquad \square$$

Remark: In general there may exist many different random mapping representations for a particular chain. In the above proof we just constructed *one of* these representations.

1.1 Forward Coupling

Suppose we take two copies of a Markov chain X_n and Y_n. Their random mapping representations are:

$$X_{n+1} = \Phi(X_n, U_n)$$
$$Y_{n+1} = \Phi(Y_n, U_n).$$

In general, the U_n's used in the chains X_n and Y_n are two independent samples. However, *if we use the same samples of U_n for updating $X_n \to X_{n+1}$ and $Y_n \to Y_{n+1}$,* we will impose *grand coupling* between those chains. We used grand coupling to prove upper-bounds on the mixing time of the Metropolis chain for graph coloring.

Now suppose we start $|S|$ copies of the chain $X_n(1), X_n(2), \dots$ and update them using the same samples of U_n (i.e. we establish pairwise grand coupling). This situation is called *forward coupling*.

One may propose that in order to verify whether the chain is in stationary distribution or not, we shall wait until all the chains coalesce and after coalescence the chain is in the stationary distribution. Unfortunately this is not the case as we will see in the following examples:

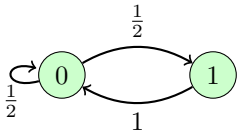


Figure 1: Markov chain of Example 1.2

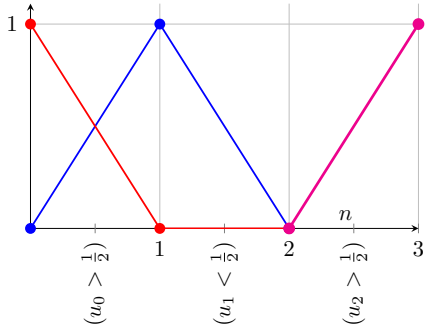


Figure 2: Two copies of the chain considered in Example 1.2.

Example 1.2. Consider the Markov chain of Figure 1. A random mapping representation of this chain (using $U_n \sim \mathcal{U}[0,1]$) is

$$\Phi(0,u) = \begin{cases} 0 & \text{if } 0 \le u \le \frac{1}{2}, \\ 1 & \text{if } \frac{1}{2} < u \le 1, \end{cases}$$
$$\Phi(1,u) = 0.$$

It is easy to check that coalescence always happens at state 0. Indeed the only way to get $\Phi(0,u) = \Phi(1,u)$ is to have $0 \le u \le 1$ implying $\Phi(0,u) = \Phi(1,u) = 0$. For example, consider the situation depicted in Figure 2. That is to say, $(\pi_0(T_c) = 1, \pi_1(T_c) = 0)$ where T_c is the coalescence time. However, the stationary distribution is $(\pi_0^* = \frac{2}{3}, \pi_1^* = \frac{1}{3})$. Therefore the chains are not in the stationary distribution when they coalesce. They are also not in the stationary distribution after the coalescence time.

The choice of random mapping representation can even lead to the situations where we do not have coalescence at all.

Example 1.3. Consider the Markov chain of Figure 3. One possible candidate for its random mapping representation (still assuming $U_n \sim \mathcal{U}[0,1]$) is

$$\Phi(i,u) = \begin{cases} i & \text{if } u \le \frac{1}{2}, \\ 1-i & \text{if } u > \frac{1}{2}. \end{cases}$$

Using this mapping, the two chains will never coalesce (see Figure 4a for example).

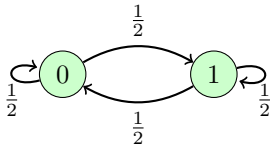
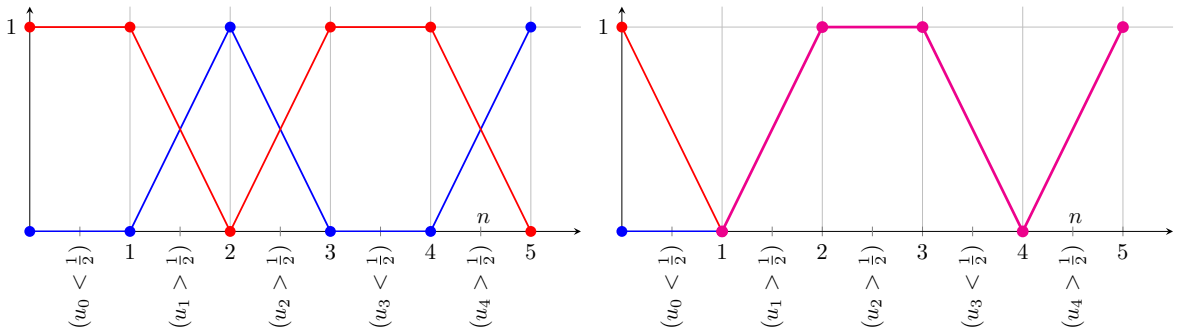


Figure 3: Markov chain of Example 1.3



(a) Coalescence does not happen.

(b) Coalescence happens immediately.

Figure 4: The choice of random mapping representation can change the coalescence time.

However, if we pick another random mapping representation as,

$$\Phi(i,u) = \begin{cases} 0 & \text{if } u \leq \frac{1}{2}, \\ 1 & \text{if } u > \frac{1}{2}, \end{cases}$$

with the same realization of U_ns the coalescence will happen in one step (see Figure 4b).

1.2 Coupling From The Past

Surprisingly, modifying the idea of forward coupling will lead to a criterion to check whether the chain is in the stationary distribution. The idea is called *coupling form the past* and is formalized as follows (the algorithm is known as Propp-Wilson algorithm):

1. Set $T_0 = -1$.
2. Generate u_{T_0}.
3. Start the experiment at all states $i \in S$ at T_0 and update $X_{n+1} = \Phi\left(X_n(T_0, i), u_n\right)$ for $n = T_0, T_0+1, \ldots, -1$ ($X_n(T_0, i)$ denotes the state of the chain initiated at state i at T_0 at time n).
4. Check coalescence at time $n = 0$: If $X_0(T_0, i)$ is independent of i (in other words, $X_0(T_0, i) = X_0(T_0, j)$ for $\forall i, j \in S$), $X_0(T_0, i)$ is the output and the algorithm terminates. If not, set $T_0 \leftarrow T_0 - 1$ and return to step 2.

We will see that the distribution of $X_0(T_0, i)$ is *exactly* the stationary distribution of the Markov chain.

In order to prove the above mentioned claim let's first setup some notation. Given the state of the chain at time n_1 and the sequence $u_{n_1}, u_{n_1+1}, \ldots, u_{n_2-1}$ we can determine the state at time n_2. In other words, the state at time n_2 is a deterministic function of the state at time n_1 and the sequence $u_{n_1}, u_{n_1+1}, \ldots, u_{n_2-1}$. We denote this relationship as

$$x_{n_2} \triangleq F_{n_1}^{n_2}(x_{n_1}; u_{n_1}, u_{n_1+1}, \ldots, u_{n_2-1}).$$

We define the event of coalescence at time L as

$$A_0^L \triangleq \left\{F_0^L(i; U_0, U_1, \ldots, U_{L-1}) \text{ is independent of } i \text{ for } \forall i \in S\right\}.$$

Theorem 1.4. Suppose X_n is an ergodic Markov chain and fix a random mapping representation of this chain. If $\exists L < \infty$ such that $\mathbb{P}(A_0^L) > 0$ (i.e. coalescence happens at finite time) then the Propp-Wilson algorithm returns $X_0(T_0, i)$ that is independent of i and $X_0(T_0, i) \sim \pi^*$.

Proof. We first need to check that the algorithm terminates in finite time with probability 1. To this end, we will show:

$$\mathbb{P}\left(\bigcup_{k\geq 1} A_{-kL}^{-(k-1)L}\right) = 1.$$

Because of Markovity, the events $A_{-kL}^{-(k-1)L}$ are i.i.d. Hence,

$$\begin{aligned}\mathbb{P}\left(\bigcup_{k\geq 1} A_{-kL}^{-(k-1)L}\right) &= 1 - \mathbb{P}\left(\bigcap_{k\geq 1} \overline{A_{-kL}^{-(k-1)L}}\right) \\ &= 1 - \prod_{k\geq 0} \mathbb{P}\left(\overline{A_{-kL}^{-(k-1)L}}\right) \\ &= 1 - \lim_{n\to\infty} \left(1 - \mathbb{P}(A_0^L)\right)^n = 1,\end{aligned}$$

since $\mathbb{P}(A_0^L) > 0$ by assumption.

It remains to show that $X_0(T_0, i) \sim \pi^*$ (when the algorithm stops). To this end, let π denote the distribution of $X_0(T_0, i)$. We will show that $\pi P = \pi$.

If we continued the algorithm for one more round, that is to say, we generated u_{T_0-1} and computed $X_0(T_0 - 1, i)$ for $\forall i \in S$, because of coalescence we would have $X_0(T_0 - 1, i) = X_0(T_0, i)$ (simply because regardless of u_{T_0-1}, $X_{T_0}(T_0 - 1, i)$ corresponds to the starting point of one of the coupled chains that are coalesced at time 0 by assumption). On the other side, time-homogeneity implies $X_0(T_0 - 1, i) \sim \pi P$. Therefore $\pi P = \pi$ and the proof is complete. □

1.2.1 Propp-Wilson Algorithm in Practice: Monotone Coupling From the Past

At first glance, Propp-Wilson algorithm seems to be useless when the state space is huge; we need to keep track of $|S|$ copies of the chain in order to generate one sample distributed according to π^*. For example in case of Ising model we need to run 2^N chains simultaneously which is not practically feasible.

However, if we have a partial ordering in the state space and we have a random mapping representation that preserves this ordering, that is to say, if $x \preceq y$, then $\Phi(x, u) \preceq \Phi(y, u)$ for $\forall x, y \in S$, we only need to keep track of *two* copies of the chain; the ones starting from two extremal states. This is called *monotone coupling from the past.*

Example 1.5. For the Ising model a natural partial order is $\underline{\sigma} \preceq \underline{\tau}$ iff $\sigma_v \leq \tau_v, i = v, \cdots, N$. One can show for the ferromagnetic case (defined by $J_{vw} \geq 0$ that the Glauber updates are monotone for this partial order. If you try to check this you will see that this (exercise!) is just because the tanh is a monotone increasing function. In this situation it is enough to check the coalescence for the minimal and maximal states $\underline{\sigma}_{min} = (\sigma_v = -1, v = 1, \cdots, N)$ and $\underline{\sigma}_{max} = (\sigma_v = +1, v = 1, \cdots, N)$.