STA4502: Topics in Stochastic Processes Winter 2018

Lecture 1: Introduction — March 1, 2018

Lecturer: Jeffrey Rosenthal *Scribe: Mufan (Bill) Li*

1 Introduction to Markov Chains

Before we define a Markov chain, we introduce the following notations and definitions.

Definition 1. We denote $(\mathcal{X}, \mathcal{F})$ our (measurable) **state space**, equipped with appropriate σ-algebra $\mathcal{F}$. Here a state space is called **discrete** if the cardinality of $\mathcal{X}$ is finite or countable, otherwise it's called **continuous**.

Here's a diagram with some examples

$$\text{state space}(\mathcal{X}, \mathcal{F}) \begin{cases} \text{discrete:} \begin{cases} \text{finite:} & \text{e.g. } \mathcal{X} = \{1, 2, 3, \ldots, n\}, \mathcal{F} = \mathcal{P}(\mathcal{X}) \text{ (the power set)} \\ \text{countable:} & \text{e.g. } \mathcal{X} = \mathbb{N} = \{0, 1, 2, \ldots\}, \mathcal{F} = \mathcal{P}(\mathcal{X}) \end{cases} \\ \text{continuous: e.g. } \mathcal{X} = \mathbb{R}^d, \mathcal{F} = \mathcal{B}(\mathbb{R}^d) \text{ (the Borel sets).} \end{cases}$$

Definition 2. $P : \mathcal{X} \times \mathcal{F} \to [0, 1]$ is a **transition probability** if $\forall x \in \mathcal{X}$, $P(x, \cdot) : \mathcal{F} \to [0, 1]$ is a probability measure on $(\mathcal{X}, \mathcal{F})$.

Remark 3. Observe that for a discrete $\mathcal{X}$, $P(x, y)$ is well defined $\forall x, y \in \mathcal{X}$, i.e. it's the probability of going from point x to point y. For a continuous $\mathcal{X}$, we need to use measurable sets $A \in \mathcal{F}$ to talk about transition probability $P(x, A)$ from x to A.

Finally we can define a Markov chain.

Definition 4. We call a sequence of random variables $\{X_k\}_{k=0}^{\infty}$ taking values in $\mathcal{X}$ a **Markov chain** if $\mathbb{P}[X_{k+1} \in A | X_k] = P(X_k, A), \forall k \in \mathbb{N}, A \in \mathcal{F}$.

At the same time, we would like to define the following distributions on $\mathcal{X}$.

Definition 5. We call $\nu = \mathcal{L}(X_0)$ the **initial disitribution**, and denote $\mu_k = \mathcal{L}(X_k)$. A distribution π is called a **stationary distribution** if $\pi(A) = \int_{\mathcal{X}} P(x, A) d\pi(x)$.

In other words, if the Markov chain starts in a stationary distribution ($\nu = \pi$), it will remain in stationarity ($\mu_k = \pi, \forall k \in \mathbb{N}$).

The main goal of this course is study whether or not we have convergence of $\mu_k \to \pi$ in some sense, and if so quantify the "rates" of this convergence. To this end, we will introduce several defintions and basic conditions to guarantee convergence.

2 Convergence Conditions

We start with a couple of definitions.

Definition 6. A dicrete Markov chain is **irreducible** if

$$\forall x, y \in \mathcal{X}, \mathbb{P}[X_k = y \text{ eventually } | X_0 = x] > 0.$$

Equivalently, we can say $\exists m \in \mathbb{N} : P^m(x, y) > 0$, where $P^m(x, y)$ is the transition probability after m steps.

In general, we say Markov chain is ϕ-**irreducible** if $\exists$ non-zero σ-finite measure ϕ on $(\mathcal{X}, \mathcal{F}) : \forall x \in \mathcal{X}, A \in \mathcal{F}$ with $\phi(A) > 0$, we have that $\mathbb{P}[X_k \in A \text{ eventually } | X_0 = x] > 0$.

Remark 7. The general ϕ-irreducibility is not equivalent to the discrete irreducibility. Consider for example ϕ that concentrates only on a single point $x \in \mathcal{X}$, then ϕ-irreducibility only requires $\forall y \in \mathcal{X}, \exists m \in \mathbb{N} : P^m(y, x) > 0$.

Definition 8. For a discrete irreducible Markov chain, a point $x \in \mathcal{X}$ is said to be **aperiodic** if $\gcd(\{n, P^n(x, x) > 0\}) = 1$. The Markov chain is aperiodic if every point is aperiodic.

A general Markov chain with stationary distribution π is **aperiodic** if there does not exist $d \geq 2$ and a partition of size $d + 1$ such that $\mathcal{X} = \left(\bigsqcup_{i=1}^{d} \mathcal{X}_i\right) \bigsqcup N$, where $\bigsqcup$ denotes disjoint union, N is a π-null set, and for π-a.e. $x \in \mathcal{X}_i, P(x, \mathcal{X}_{i+1}) = 1$, except for π-a.e. $x \in \mathcal{X}_d, P(x, \mathcal{X}_1) = 1$.

At this point, we can state our first theorem, the conditions to guarantee convergence.

Theorem 9. If a Markov chain is irreducible (or ϕ-irreducible for the general case), aperiodic, and have a stationary distribution π, then we have for a discrete Markov chain

$$\forall v \text{ initial distribution}, \forall y \in \mathcal{X}, \lim_{k \to \infty} \mu_k(y) = \pi(y),$$

or for a general Markov chain

$$\text{for } \pi - \text{a.e. } x \in \mathcal{X}, \lim_{k \to \infty} \sup_{A \in \mathcal{F}} \left| P^k(x, A) - \pi(A) \right| = 0.$$

Remark 10. For the general chain, this type of convergence is stronger than the typical weak convergence (in distribution), it is known as **convergence in total variation**. The name refers to the total variation distance defined by

$$\mathrm{TV}(\mu_k, \pi) := \sup_{A \in \mathcal{F}} |\mu_k(A) - \pi(A)| .$$

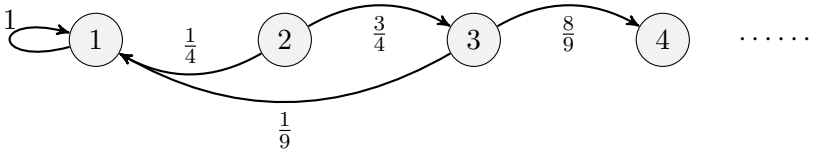
In the discrete case, we also have the following identity

$$\mathrm{TV}(\mu_k, \pi) = \frac{1}{2} \sum_{y \in \mathcal{X}} |\mu_k(y) - \pi(y)| = \sum_{y \in \mathcal{X} : \mu_k(y) > \pi(y)} |\mu_k(y) - \pi(y)|.$$

Therefore we have that for the discrete case, weak convergence implies convergence in total variation.

However, in general this implication is false. Consider the following counter example (by Jeffrey Negrea). Let $\mathcal{X} = [0, 1], \pi = \delta_0$, a point mass at $x = 0$. Define the transition probability as $P(x, x/2) = 1$. Then $\forall x \in \mathcal{X}, \nu = \delta_x$, we have $\mu_k \xrightarrow{d} \delta_0$. However the total variation distance is always 1 as $|\pi(0) - \mu_k(0)| = 1, \forall k$.

Example 11. Let $\mathcal{X} = \{1, 2, \ldots\}, \pi = \delta_1, P(1,1) = 1$. For $n \geq 2$, let $P(n, n+1) = 1 - 1/n^2$, $P(n, 1) = 1/n^2$. See diagram below for a few sample points.



Observe in this case, the chain is not irreducible in the discrete definition, however it is ϕ-irreducible when $\phi = \delta_1$. Similarly, this Markov chain is aperiodic in the general sense, since the only possible node to return to has a period of 1.

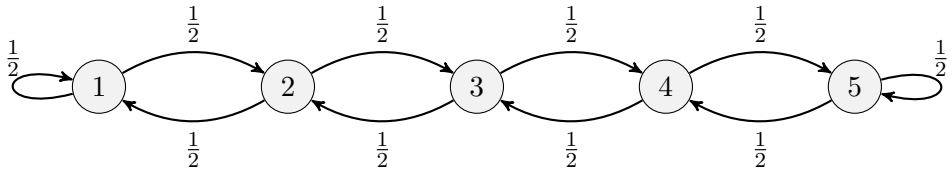
Here we can check for which $x \in \mathcal{X}$ we have convergence to stationary distribution.

$x = 1 \implies \mu_k = \delta_1 \, \forall k$. ✓

$x \geq 2$, in this case we have that $1/n^2$ is summable, which implies $\prod_{n=1}^{\infty}(1 - 1/n^2) > 0$. In other words, there is a positve probability of $X_k \to \infty$. ✗

This implies we have $\mu_k \xrightarrow{\text{TV}} \pi$, but only for π-a.e. $x \in \mathcal{X}$ starting points, which is only $x = 1$.

Example 12. Let $\mathcal{X} = \{1, 2, 3, 4, 5\}$, and define a symmetric random walk, i.e. $P(x, x+1) = P(x, x-1) = 1/2$, except at the ends, we have $P(1,1) = P(5,5) = 1/2$ instead. See diagram below.



This chain is clearly irreducible. It is also aperiodic since for every path $x \to y$, we can stop at 1 or 5 for one additional step, making the gcd 1.

The stationary distribution π is uniform since the chain is reversible, i.e. $P(x, y) = P(y, x)$.

Since all the conditions are satisfied we have

$$\lim_{k \to \infty} \mathbb{P}[X_k = x] = \pi(x) = \frac{1}{5}, \forall x \in \mathcal{X}.$$

Goal 13. Find $k^* \in \mathbb{N}$ such that

$$\sup_{A \in \mathcal{F}} \left| P^{k^*}(x, A) - \pi(A) \right| < 0.01$$

This is called the **quantitative rate of convergence**. Here we remark that finding one particular k^* is already difficult, therefore we are less interested in find the minimal k^*.

To this goal, we will introduce the coupling technique.

3 The Coupling Inequality

If X, Y are jointly defined random variables, then we can bound the total variation distance by the following steps

$$\begin{aligned}
\|\mathcal{L}(X) - \mathcal{L}(Y)\|_{\text{TV}} &= \sup_{A \in \mathcal{F}} |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)| \\
\text{partition each event } \ldots &= \sup_{A \in \mathcal{F}} |\mathbb{P}(X \in A, X = Y) + \mathbb{P}(X \in A, X \neq Y) \\
&\quad - \mathbb{P}(Y \in A, X = Y) - \mathbb{P}(Y \in A, X \neq Y)| \\
\{X \in A, X = Y\} = \{Y \in A, X = Y\} \implies &= \sup_{A \in \mathcal{F}} |\mathbb{P}(X \in A, X \neq Y) - \mathbb{P}(Y \in A, X \neq Y)| \\
&\leq \mathbb{P}(X \neq Y),
\end{aligned}$$

where a factor of 2 is not required in the last step since both probabilities are non-negative.

Example 14. (Apply to Markov Chains) Here we start with a Markov chain $\{X_k\}_{k=0}^{\infty}$, and we make a copy of it denoted $\{Y_k\}_{k=0}^{\infty}$, with the joint distribution specified later.

Usually (although not exclusively), we will let $Y_0 \sim \pi$, i.e. start in the stationary distribution, therefore $Y_k \sim \pi$, i.e. remains in stationarity. This implies

$$\|\mu_k - \pi\|_{\text{TV}} = \|\mathcal{L}(X_k) - \mathcal{L}(Y_k)\|_{\text{TV}} \leq \mathcal{P}(X_k \neq Y_k).$$

Proof intuition: here we let X_k, Y_k move together, i.e. if $X_{k+1} = X_k \pm 1$, then we also have $Y_{k+1} = Y_k \pm 1$; while at the end points, one of the chains must remain in the same node, hence reducing the "distance" by 1. Eventually, the two chains will "converge" to the same value.

Challenge 15. Use the coupling technique above to find k^* such that

$$\sup_{A \in \mathcal{F}} \left| P^{k^*}(x, A) - \pi(A) \right| < 0.01$$

STA4502: Topics in Stochastic Processes Winter 2018

Lecture 2: Minorization Condition — March 7, 2018

Lecturer: Jeffrey Rosenthal *Scribe: Louis Bélisle*

4 Recap of previous lecture

A Markov Chain is a sequence $\{X_k\}$ in a space $\mathcal{X}$, transition probability P, initial distribution $\nu = \mu_0$, where the k-th step is distributed following $\mu_k = \mathcal{L}(X_k)$. It may have a stationary distribution π such that $\pi P = \pi$.

Theorem 16. *If the chain is irreducible and aperiodic for π-a.e. $x = X_0$, then $\|\mu_k - \pi\|_{TV} \to 0$*

Remark 17. It is possible to show that the Total Variation function is non-increasing. Start by noticing that P is a weak contraction operator. In "hand-wavy" form,

$$|P| < 1 \Rightarrow \|\mu_{k+1} - \pi\| = \|(\mu_k - \pi)P\| \leq \|\mu_k - \pi\| \cdot \|P\|$$

Proposition 18 (Roberts and Rosenthal, 2004)**.** 1. $\|\nu_1(.) - \nu_2(.)\| = \sup_{f:\mathcal{X}\to[0,1]} |\int f \mathrm{d}\nu_1 - \int f \mathrm{d}\nu_2|$

2. $\|\nu_1(.) - \nu_2(.)\| = \frac{1}{b-a} \sup_{f:\mathcal{X}\to[a,b]} |\int f \mathrm{d}\nu_1 - \int f \mathrm{d}\nu_2|$ for any $a < b$ and in particular $\|\nu_1(.) - \nu_2(.)\| = \frac{1}{2} \sup_{f:\mathcal{X}\to[-1,1]} |\int f \mathrm{d}\nu_1 - \int f \mathrm{d}\nu_2|$

3. If π is stationary for a Markov chain kernel P, then $\|P^n(x,.) - \pi(.)\|$ is non-increasing in n, i.e., $\|P^n(x,.) - \pi(.)\| \leq \|P^{n-1}(x,.) - \pi(.)\|$ for $n \in \mathbb{N}$

4. More generally, letteing $(\nu_i P)(A) = \int \nu_i(\mathrm{d}x) P(x, A)$, we always have $\|(\nu_1 P)(.) - (\nu_2 P)(.)\| \leq \|\nu_1(.) - \nu_2(.)\|$.

5. Let $t(n) = 2 \sup_{x\in\mathcal{X}} \|P^n(x,.) - \pi(.)\|$, where $\pi(.)$ is stationary. the t is submultiplicative, i.e., $t(m+n) \leq t(m)t(n)$ for $n, m \in \mathbb{N}$.

6. if $\mu(.)$ and $\nu(.)$ have densities g and h, respectively, with respect to some σ-finite measure $\rho(.)$ and $M = \max(g, h)$ and $m = \min(g, h)$, then

$$\|\mu(.) - \nu(.)\| = \frac{1}{2} \int_{\mathcal{X}} (M - m) \mathrm{d}\rho = 1 - \int_{\mathcal{X}} m \mathrm{d}\rho$$

7. Given probability measures $\mu(.)$ and $\nu(.)$, there are jointly defined random variables X and Y such that $X \sim \mu(.)$ and $Y \sim \nu(.)$ and $P[X = Y] = 1 - \|\mu(.) - \nu(.)\|$.

Proof. Ref: Roberts and Rosenthal, 2004. General State Space Markov Chains and MCMC Algorithms. □

Then we saw the coupling inequality and introduced the purpose of this course: studying the speed of convergence of a Markov Chain. This means:

$$\text{For any } \epsilon > 0, \text{ say } \epsilon = 0.01, \text{ find } k^* \text{ such that } \|\mu_k - \pi\|_{TV} \leq \epsilon.$$

4.1 Challenge Solution

Let $\mathcal{X} = \{1, 2, 3, 4, 5\}$ and $P(x, \cdot)$ follow a single-step random walk with holding, referring back to challenge 15 which stems from example 12. We know it has a stationary distribution $\pi = \text{Unif}(\mathcal{X})$. Using the coupling inequality,

$$\begin{aligned} \|\mu_k - \pi\| &\leq P(X_k \neq Y_k) \\ &\leq \left(\frac{7}{8}\right)^{\lfloor k/4 \rfloor} \\ &< 0.01 \text{ if } k \geq 140 \end{aligned}$$

This value of k gives a number of steps in the chain that will guaranty that the result is within a "reasonable" distance of its stationary distribution. We can find tighter bounds for k^*, the tightest exposed in class having been found by numerical exponentiation of P to yield a $k^* = 39$. Next, we will present different ways to get bounds on k^*.

5 Minorization Condition

Goal 19. *The goal is to find more efficient ways of finding the speed of convergence of a Markov chain, other than trial and error. Using the Minorization Condition is similar in a way as thinking about coupling.*

Condition 20 (Rosenthal,1995)**.** A Markov chain with transition kernel $P(x, \mathrm{d}y)$ on a state space $\mathcal{X}$ is said to satisfy a *minorization condition* if there is a probability measure $\rho(\cdot)$ on $\mathcal{X}$, a positive integer k_0, and $\epsilon > 0$, such that

$$P^{k_0}(x, A) \geq \epsilon \rho(A), \ \forall x \in \mathcal{X},$$

for all measurable subsets $A \subseteq \mathcal{X}$.

The condition requires every state in the state space to be within reach of any other state. We can then minorize the transition probability with a density $\rho(\cdot)$ scaled by a parameter ϵ. This is equivalent to finding a sliver of a probability distribution where all the transition probabilities "overlap" with each other (see Figure 1 for illustration). This can fail because we may not have an overlap in common for all possible values of $x \in \mathcal{X}$ (see Observation 24).

Remark 21. Why is this similar to coupling? Because coupling is trying to make two Markov chains become equal, while the minorization condition is showing us how this can be done.

Remark 22. The overlap suggests how to create the joint distribution. We know that the marginals need to satisfy the Markov Chain conditions, but the joint distribution can be specified to fit our needs.

Proposition 23 (Coupling under Minorization Condition)**.** Given $X_{n-1} = x$ and $Y_{n-1} = y$,

$$\text{if } x \neq y, \begin{cases} \text{With probability} = \epsilon, \text{ choose } z \sim \rho(\cdot), \text{ and set } X_n = Y_n = z \\ \text{With probability} = (1-\epsilon), \text{ choose } \begin{cases} X_n \sim \frac{1}{1-\epsilon}(P(x,\cdot) - \epsilon\rho(\cdot)) \\ Y_n \sim \frac{1}{1-\epsilon}(P(y,\cdot) - \epsilon\rho(\cdot)) \end{cases} \end{cases}$$

otherwise, if $x = y$, leave them together and choose $X_n = Y_n \sim P(x,\cdot)$

For a matter of convenience, in the case of $x \neq y$ where we choose X_n and Y_n separately (i.e. not setting them equal to z) we often take the two distributions of X_n and Y_n to be conditionally independent from each other. This completely defines the joint distribution of the two Markov processes.

Therefore, the distribution of X_n becomes $\epsilon\rho(\cdot) + \frac{1}{1-\epsilon}(P(x,\cdot) - \epsilon\rho(\cdot))$. Similarly for Y_n which implies

$$Pr(Y = X) \geq \epsilon$$

For this coupling, $P(\text{“becoming equal at step } n\text{”}) \geq \epsilon$, i.e., the probability of becoming equal at step n is larger or equal to ϵ, therefore,

$$\|\mu_k - \pi\| \leq P(X_k \neq Y_k) \leq (1-\epsilon)^k$$

If the minorization condition is satisfied, then the above inequality would allow us to find a k^* that is indicative of the speed of convergence.

Observation 24. It is possible to have a Markov chain where not all states are reachable within one step of any other state (think of our example 12). However, with a Markov chain that we know converges to a stationary distribution, it is possible to create an analogous chain that consists of a small power of the transition kernel P that makes all states reachable within one “step” of this power.

This means, we can find a k_0 such that, if

$$P^{k_0}(x,\cdot) \geq \epsilon\rho(x,\cdot), \ \forall x \in \mathcal{X},$$

then

$$\|P^{k_0}(x,\cdot) - \pi\| \leq \|(P^{k_0})^{\lfloor k/k_0 \rfloor}(x,\cdot) - \pi\| \leq (1-\epsilon)^{\lfloor k/k_0 \rfloor}$$

Example 25. For our example 12 from Lecture 1, we do not immediately satisfy the minorization condition because not all states are reachable from a particular starting point. However, within 4 steps, we have a positive probability to reach any point for every starting state. So we can use $P^4(x,\cdot)$ as our “chain” that satisfies the minorization condition. Within 4 steps, we have at least a probability $1/4^4 = 1/16$ of reaching any other state. We can thus choose $\epsilon = 1/16$. Then to choose a distribution $\rho(\cdot)$, we have many options:

1. If we decide to take $\rho(\cdot) = \delta_3(\cdot)$, i.e., a point mass at state 3, then

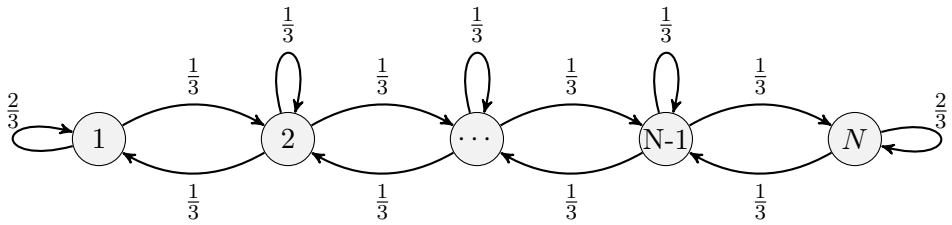
$$P^4(x,3) \geq \frac{1}{16}\delta_3(\cdot), \ \forall x \Rightarrow \|P^{k_0}(x,\cdot) - \pi\| \leq \left(\frac{15}{16}\right)^{\lfloor k/4 \rfloor} \leq 0.01 \Rightarrow k^* = 288$$

2. If we decide to take $\rho(\cdot) = \text{Unif}(\mathcal{X})$, i.e., the discrete uniform distribution over $\mathcal{X}$, then

$$P^4(x,\cdot) \geq \frac{5}{16}\text{Unif}(\mathcal{X}), \ \forall x \Rightarrow \|P^{k_0}(x,\cdot) - \pi\| \leq \left(\frac{11}{16}\right)^{\lfloor k/4 \rfloor} \leq 0.01 \Rightarrow k^* = 52$$

Challenge 26. Take a new MC similar to example 12, i.e., single-step random walk over $\mathcal{X} = \{1, 2, \ldots, N\}$, for $N \in \mathbb{N}$ but where the transition probabilities are

$$\begin{aligned} \Pr(\text{Go Left}) &= 1/3 \\ \Pr(\text{Stay Put}) &= 1/3 \\ \Pr(\text{Go Right}) &= 1/3 \end{aligned}$$



Then

1. Find k^* with $N = 5$
2. What is k^* with $N \to \infty$ (gets arbitrarily large)

5.1 Method to find minorization components

Optimally, we would take

$$\epsilon\rho(y) = \min_{x \in \mathcal{X}} P(x, y), \ \forall y \in \mathcal{X},$$

which leads us to choose a particular ϵ and create the $\rho(\cdot)$ such that it is a probability distribution that fits the criteria for the minorization condition. One way to build such elements is the following:

$$\text{Discrete: } \begin{cases} \epsilon = \sum_y \min_x P(x, y) \\ \rho(y) = \frac{\min_x P(x,y)}{\sum_y \min_x P(x,y)} \end{cases}$$

$$\text{Continuous: } \begin{cases} \epsilon = \int_y \inf_x P(x, \mathrm{d}y) \\ \rho(y) = \frac{\inf_x P(x,\mathrm{d}y)}{\int_y \inf_x P(x,\mathrm{d}y)} \end{cases}$$

5.2 Continuous state space: an application of the minorization condition

Example 27. Let $\mathcal{X} = [0, 2]$. Let the transition probability from state $x \in \mathcal{X}$ to a subset $A \subseteq \mathcal{X}$ be

$$P(x, A) = N(x, 1; A) + r(x)\delta_x(A)$$

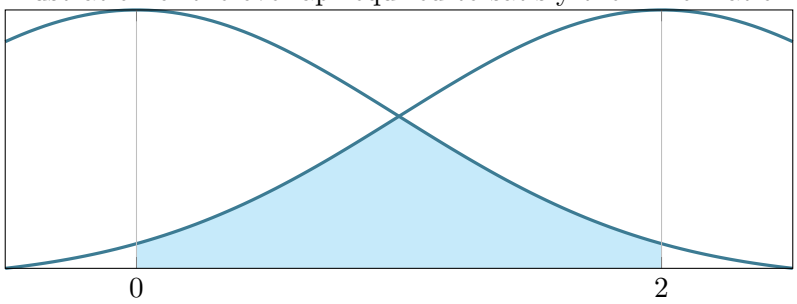
where $N(x, 1; A) = \Pr(z \in A)$ with $z \sim N(x, 1)$, and where $r(x) = 1 - N(x, 1; \mathcal{X})$, the probability that a draw from $N(x, 1)$ falls outside $\mathcal{X}$. (This corresponds to the Metropolis-Hastings algorithm with $\pi = \text{Unif}[0, 2]$.)

Remark 28. *This transition probability is reversible with respect to* $\pi = \text{Unif}[0,2]$*, i.e., if we start in a neighbourhood of* x*, the probability of jumping in a neighbourhood of* y *is* the same as *if we had started in neighbourhood of* y *and measured the probability of jumping in a neighbourhood of* x*.* $\forall\, x, y \in \mathcal{X}$,

$$\begin{aligned}\pi(x)P(x,y) =& \pi(y)P(y,x),\ \textit{(Discrete)}\\ \pi(dx)P(x,dy) =& \pi(dy)P(y,dx),\ \textit{(Continuous)}\end{aligned}$$

In this situation, we have special case where the Uniform distribution guarantees $\pi(\mathrm{d}x) = \pi(\mathrm{d}y)$ and the symmetry of the Normal distribution guarantees $P(x, \mathrm{d}y) = P(y, \mathrm{d}x)$.

Figure 1: Illustration of the overlap required to satisfy the minorization condition



To be able to use a minorization argument, we must verify 2 things:

1. The Markov chain converges
   (a) This chain is ϕ-irreducible under $\phi = \text{Lebesgue}|_{[0,2]}$
   (b) It is aperiodic since $N(\cdot)$ covers all the domain $[0, 2]$.
2. The minorization condition is satisfied
   (a) we can find $\epsilon = \int_y g(y)\mathrm{d}y$ where $g(y) \leq f(x, y)\, \forall x, y$.

Then, we will be able to find a value k^* such that, $\forall\, k \geq k^*$, $\|\mu_k - \pi\|_{\text{TV}} < 0.01$. To construct ϵ, it helps to think of the "worst case" scenario for the location of x and Y. In this case, take $X = 0$ and $Y = 2$ (as represented in Figure 1). The shaded area represents $\epsilon\rho(\cdot)$. Then,

$$\begin{aligned}\forall x, y,\ P(x,\mathrm{d}y) \geq& \min[P(0,\mathrm{d}y), P(2,\mathrm{d}y)]\\ \Rightarrow \epsilon =& \int_y \min[P(0,\mathrm{d}y), P(2,\mathrm{d}y)]\\ =& (\Phi(2) - \Phi(1)) + (\Phi(-1) - \Phi(-2))\\ =& 2\,(\Phi(2) - \Phi(1))\\ \geq& 0.27\\ \therefore \|\mu_k - \pi\|_{\text{TV}} \leq& (1-\epsilon)^k = (0.73)^k\\ <& 0.01 \ \text{ if } k \geq 15\end{aligned}$$

So take $\epsilon = 0.23$ and $k^* = 15$. In this case, we do not need to know the exact form of $\rho(\cdot)$, but by construction we know $\rho(\cdot)$ has density

$$f(y) = \frac{\min[N(0,1;y), N(2,1;y)]}{2\left(\Phi(2) - \Phi(1)\right)} \mathbb{I}_{\{y \in \mathcal{X}\}}.$$

6 Eigenvectors and eigenvalues: first concept

We know our distribution at step k is $\mu_k = \mu_0 P^k$ with $|\mathcal{X}| = d$. Suppose we coudl find λ_i, v_i such that $v_i P = \lambda_i v_i$ for $i = 0, 1, \ldots, d-1$. If we represent μ_0 as

$$\mu_0 = a_0 v_0 + a_1 v_1 + \ldots + a_{d-1} v_{d-1},$$

then we could find values for λ_i's such that

$$\mu_k = \mu_0 P^k = a_0 (\lambda_0)^k v_0 + a_1 (\lambda_1)^k v_1 + \ldots + a_{d-1} (\lambda_{d-1})^k v_{d-1}.$$

where we would usually take $\lambda_0 = 1$, $v_0 = \pi$, $a_0 = 1$ (by relabeling, since we know $\pi P = \pi$) and we will have $|\lambda_m| < 1$ for $m > 0$, which will give us bounds on convergence.

STA4502: Topics in Stochastic Processes **Winter 2018**

Lecture 3: Eigenvalue Connection — March 14, 2018

Lecturer: Jeffrey Rosenthal *Scriber: Yuenan Joseph Cai*

Challenge 26 Solution

Starting from any state, there is at least $\frac{1}{3^{N-1}}$ probability to get to any other state in $N-1$ steps. This suggests using Minorization technique with $\epsilon\rho(\cdot) = \frac{1}{3^{N-1}}$ where $\rho(\cdot)$ is a uniform distribution on $\mathcal{X}$. Therefore, $\epsilon = \frac{N}{3^{N-1}}$. For $\delta > 0$, find a bound k^* s.t. $||\mu_k - \pi||_{TV} \leq (1-\epsilon)^{\lfloor k/(N-1)\rfloor} \leq \delta$.

$$\Rightarrow k^* \geq (N-1)(\frac{\ln(\delta)}{\ln(1-\frac{N}{3^{N-1}})} + 1).$$

When $N = 5$ and $\delta = 0.01$, $k^* = 294$. Can this be improved?

A Note on Coupling under Minorization Condition

The coupling method does not modify the marginal transition probabilities of $\{X_n\}$ and $\{Y_n\}$. Recall (X_n, Y_n) is jointly updated in the following manner:

- If $X_{n-1} = x \neq y = Y_{n-1}$, $\begin{cases} \text{w.p. } \epsilon\text{, choose } X_n = Y_n \sim \rho(\cdot) \\ \text{w.p. } 1-\epsilon\text{, choose } X_n \sim \frac{1}{1-\epsilon}(P(x,\cdot) - \epsilon\rho(.)) \text{ and } Y_n \sim \frac{1}{1-\epsilon}(P(y,\cdot) - \epsilon\rho(\cdot)) \end{cases}$;

- Otherwise, let $X_n = Y_n \sim P(x,\cdot)$.

In the nontrivial case $X_{n-1} = x \neq y = Y_{n-1}$,

$$P(X_n \in \cdot | X_{n-1} = x) = \epsilon\rho(\cdot) + (1-\epsilon)\Big(\frac{1}{1-\epsilon}(P(x,\cdot) - \epsilon\rho(\cdot))\Big) = P(x,\cdot).$$

Similarly, $P(Y_n \in \cdot | Y_{n-1} = y) = P(y,\cdot)$.

Example 29. (Coupling - Card Shuffling) Suppose a deck of cards is to be shuffled by taking the top card and place it randomly back into the deck. How long will it take to well scrambled the deck (i.e. "almost" equally likely to obtain any card arrangement)?

To make our analysis easier, consider an alternative shuffling method that takes a card at random from the deck and place it on top. It can be shown that the "random-to-top" shuffling method is equivalent to "top-to-random" method [1].

How to apply coupling? Consider using two desks of cards, a well mixed deck on the left and the one to be shuffled on the right. We can choose a card randomly from a deck, find the same card in the other one, and place both cards on top of the respective deck. Despite having the same card drawn from both decks, it is still a random draw from each in terms of marginals. Once all the cards have been selected once, both decks must be in the same order. Now the question can be viewed as a coupon collector's problem in finding the probability that more than k shuffles are needed to touch all n cards. Formally, the left deck starts in

[1] They are random walks on a group. One way to relate the two methods is that they are "time reversal" version of each other ($\widetilde{p}(x,y) = p(y,x)\frac{\pi(y)}{\pi(x)}$, where $\widetilde{p}$ is the new description under "random-to-top" method). In terms of group operations, each "top-to-random" draw t can be matched by a "random-to-top" draw and $\widetilde{p}(x,xt) = p(x,xt^{-1})$. We have $\widetilde{\mu}_k(x) = \mu_k(x^{-1})$ where x and x^{-1} are a permutation operation and its inverse.

stationary distribution, and after k steps,

$$\begin{aligned}
&||\mu_k - \pi||_{TV} \\
\leq &P(T > k, \text{ where } T \text{ is the time taken to select all cards at least once}) \\
= &P(\text{Have not touch all cards by time } k) \\
= &P(\cup_{i=1}^{n} \text{Have not touch card } i \text{ by time } k) \\
\leq &\sum_{i=1}^{n} P(\text{Have not touch card } i \text{ by time } k) \\
= &n(1 - \frac{1}{n})^k \\
\leq &ne^{-\frac{k}{n}} \\
= &e^{-(\frac{k}{n} - \ln(n))}
\end{aligned}$$

If $k = cn\ln(n)$, then $||\mu_k - \pi||_{TV} \leq n^{1-c}$. The bound is small when $c > 1$ and n is large. For a deck with 52 suit cards, $||\mu_k - \pi||_{TV} \leq 0.01$ when $c \approx 2.2$, or $k_* \approx 452$.

6 Eigenvectors and Eigenvalues

In this section we will study the connections between Markov chains and eigenvalues. Assume a Markov Chain on finite state space $\mathcal{X}$ of size d, the transition probability P is diagonalizable [2] with elements in $\mathbb{C}$. Recall that the eigenpairs (λ_m, v_m) of P satisfies $v_m P = \lambda_m v_m$ for $m = 0, 1, ..., d-1$. Since $\pi P = \pi$ if π is a stationary distribution, we can always assign $(1, \pi)$ to (λ_0, v_0).

For an initial distribution written in terms basis of (left) eigenvectors $\mu_0 = a_0 v_0 + ... + a_{d-1} v_{d-1}$, the k-step distribution is $\mu_k = \mu_0 P^k = a_0 \lambda_0^k v_0 + ... + a_{d-1}\lambda_{d-1}^k v_{d-1}$. Let $\lambda_* = \max_{i \geq 1} |\lambda_i|$, then

$$\begin{aligned}
&|\mu_k(x) - \pi(x)| \\
= &|a_1 \lambda_1^k v_1(x) + ...a_{d-1}\lambda_{d-1}^k v_{d-1}(x)| \\
\leq &|\lambda_1|^k |a_1 v_1(x)| + ... + |\lambda_{d-1}|^k |a_{d-1} v_{d-1}(x)| \\
\leq &|\lambda_*|^k \big(|a_1 v_1(x)| + ... + |a_{d-1} v_{d-1}(x)|\big) \\
= &C_{\mu_0, x} |\lambda_*|^k
\end{aligned}$$

Remark 30. If $|\lambda_1|, ..., |\lambda_{d-1}| < 1$, then as $k \to \infty$, $\mu_k \to a_0 \pi$ where $a_0 = 1$. For the other direction, if the Markov Chain with stationary distribution π is irreducible and aperiodic, then $|\lambda_i| < 1 \ \forall i \geq 1$. A few comments:

1. We may have other $|\lambda_i| = 1$ for some $i \in \{1, ..., d-1\}$ when the chain is periodic. Consider an example on $\{1, 2\}$ of period 2:
$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$
The eigenpairs are $\{\big(1, (1, 1)\big), \big(-1, (-1, 1)\big)\}$. In this case the k-step distribution does not converge to the stationary distribution $(\frac{1}{2}, \frac{1}{2})$ for general μ_0. Instead we require periodic version of the Markov Chain Convergence Theorem.

[2] If P is not diagonalizable (when some eigenvalues have multiplicity ≥ 2), then we obtain the Jordan canonical form.

2. If the chain is not irreducible, we may still be able to find irreducible sub-chains where the statement still holds.

Next we will review a few results from Linear Algebra

Definition 31. ($L^2(\pi)$ norm) The $L^2(\pi)$ norm between v and w is $< v, w >_{L^2(\pi)} := \sum_{x\in\mathcal{X}} v(x)w(x)\pi(x)$.

If the eigenvectors $\{v_i\}$ are orthonormal in $L^2(\pi)$, then $< v_i, v_j >_{L^2(\pi)} = \delta_{ij}$. We seek for conditions in which P is diagonalizable w.r.t. some orthonormal vectors.

Definition 32.
1. The adjoint operator P^* of P satisfies $< v, wP > = < vP^*, w >$.
2. P is normal if $PP^* = P^*P$.
3. P is self-adjoint if $P = P^*$.

Fact 33.
1. If P is self-adjoint, then it is also normal, with real eigenvalues.
2. If P is either normal or self-adjoint, then there exists an orthonormal basis $\{v_i\}$.
3. If π is a uniform distribution, then $P^* = P^\dagger$ where $P^\dagger$ is the conjugate transpose of P. Furthermore P is self-adjoint iff P is symmetric.

Fact 34. P is self-adjoint iff the chain is reversible w.r.t P.

Fact 35. Using Cauchy–Schwarz inequality, one can show that

$$
\begin{aligned}
&||\mu_k - \pi||_{TV} \\
=&\frac{1}{2}\sum_x |\mu(x) - \pi(x)| \\
=&\frac{1}{2}\sum_x |\mu(x) - \pi(x)|\sqrt{\pi(x)})\frac{1}{\sqrt{\pi(x)}} \\
\leq&\frac{1}{2}\sqrt{\sum_x (\mu_k(x) - \pi(x))^2\pi(x)\sum_x \frac{1}{\pi(x)}} \\
\leq&\frac{1}{2}\sqrt{\sum_x (\mu_k(x) - \pi(x))^2\pi(x)\frac{n}{\min_x \pi(x)}} \\
=&\frac{1}{2}\sqrt{\frac{n}{\min_x \pi(x)}}||\mu_k - \pi||_{L^2(\pi)}
\end{aligned}
$$

Fact 36. The result below will allow us to quantify the convergence rate in terms of eigenvalues.

$$\begin{aligned}
&||\mu_k - \pi||^2_{L^2(\pi)} \\
=& < \mu_k - \pi, \mu_k - \pi >_{L^2(\pi)} \\
=& < \sum_{i=1}^{n-1} a_i \lambda_i^k v_i, \sum_{i=1}^{n-1} a_i \lambda_i^k v_i >_{L^2(\pi)} \\
=& \sum_{i,j=1}^{n-1} a_i a_j \lambda_i^k \lambda_j^k < v_i, v_j >_{L^2(\pi)} \\
=& \sum_{i=1}^{n-1} a_i^2 \lambda_i^{2k} \\
\leq& \lambda_*^{2k} \sum_{i=1}^{n-1} a_i^2 \\
\leq& \lambda_*^{k} \sum_{i=1}^{n-1} a_i^2
\end{aligned}$$

Remark 37. In continuous case, we are interested in the operator norm $||P_0||_{L^2(\pi)\to L^2(\pi)}$ [3]. The notation $P_0 := P|_{\pi^\perp}$ stands for P restricted to signed measure of total mass 0 (analog of $\{v_1, ..., v_{d-1}\}$, which are orthogonal elements of $v_0 = \pi$). In general, $||P||_{L^2(\pi)\to L^2(\pi)} = 1$. If $||P_0|| < 1$, the chain is geometric ergodic.

6.1 A Few Motivating Examples

Example 38. (Frog Walk) Suppose there are n lily pads arranged in a circle. A frog starts at pad 0 and at each step, with equal probability, either moves clockwise, counter-clockwise or remains at where it was. The transition matrix associated with the Markov Chain on $\mathcal{X} = \{0, 1, ..., n-1\} = \mathbb{Z}/(n)$ is:

$$P = \begin{bmatrix}
1/3 & 1/3 & 0 & ... & & 0 & 1/3 \\
1/3 & 1/3 & 1/3 & 0 & ... & & 0 \\
0 & 1/3 & 1/3 & 1/3 & 0 & ... & 0 \\
... & & & & & & ... \\
0 & ... & 0 & 1/3 & 1/3 & 1/3 & 0 \\
0 & ... & & 0 & 1/3 & 1/3 & 1/3 \\
1/3 & 0 & ... & & 0 & 1/3 & 1/3
\end{bmatrix}$$

This chain is reversible w.r.t the uniform distribution on $\mathcal{X}$ and therefore, $\pi(x) = \frac{1}{N}$ is a stationary distribution. For fixed ϵ, find k^* such that $\forall k \geq k^*$, $||\mu_{k^*} - \pi||_{TV} \leq \epsilon$; in particular, solve for $n = 1000$ and $\epsilon = 0.01$.

This is a random walk on an abelian group which will be covered in the following lecture. One may simplify the notations in terms of step distributions for random walk on abelian groups.

$$Q(-1) = Q(0) = Q(1) = \frac{1}{3}.$$

[3] Norm of operator A is defined as $||A|| = \sup_{u\neq 0} \frac{||Au||}{||u||}$

Example 39. (Bit Flipping) Suppose the state space $\mathcal{X} = (\mathbb{Z}/(2))^d$ represents the set of bits of length d (for example $d = 8$, $x = (0, 0, 1, 1, 1, 0, 1, 0) \in \mathcal{X}$). Define a Markov Chain with the following transition probabilities:

$$\begin{cases} \text{w.p. } \frac{1}{d+1}, \text{ do nothing to the list} \\ \text{w.p. } \frac{d}{d+1}, \text{ change a random bit} \end{cases}.$$

Equivalently,

$$\begin{cases} \text{w.p. } \frac{1}{d+1}, \text{ set } X_n = X_{n-1} \\ \text{w.p. } \frac{d}{d+1}, \text{ set } \begin{cases} X_{n,i} = X_{n-1,i}, \ i \neq j \\ X_{n,i} = 1 - X_{n-1,i}, \ i = j \end{cases} \end{cases} \quad \text{where } j \text{ is chosen uniformly from } \{1, ..., d\}.$$

This chain is also reversible w.r.t the uniform distribution on $\mathcal{X}$ and therefore, $\pi(x) = \frac{1}{2^d}$ is a stationary distribution. How fast does the k-th step distribution converge to π?

STA4502: Topics in Stochastic Processes Winter 2018

Lecture 4: Random Walks on Groups — March 21, 2018

Lecturer: Jeffrey Rosenthal *Scribe: Tiantian Zheng*

An aside on the choice of norm

In our last lecture we obtained bounds on the convergence rate in terms of eigenvalues. The norm we used for this exercise was the $L^2(\pi)$ norm, defined between vectors v and w as:

Definition 31. $\langle v, w\rangle_{L^2(\pi)} := \sum_{x\in\mathcal{X}} v(x)\overline{w(x)}\pi(x)$

Under this norm, P is reversible iff $\langle v, Pw\rangle = \langle Pv, w\rangle$, $\forall v, w$.

Other norms might also be used, such as:

Definition 40. $\langle v, w\rangle^* := \sum_{x\in\mathcal{X}} \frac{v(x)\overline{w(x)}}{\pi(x)}$

This norm can be thought of as acting on densities instead of vectors w.r.t π as

$$\langle v, w\rangle^* = \sum_{x\in\mathcal{X}} \frac{v(x)}{\pi(x)}\frac{\overline{w(x)}}{\pi(x)}\pi(x)$$

In this case, P is reversible iff $\langle v, wP\rangle^* = \langle vP, w\rangle^*$, $\forall v, w$.

Challenge 41. Check requirements for reversibility under the two norm definitions.

In both cases, P being self adjoint implies that there exists an orthonormal basis $\{v_i\}$ which can be used to improve bounds on the convergence rate. Furthermore, the norm according to Def. 40 has some nice properties when we consider how it bounds the total variation distance:

$$\begin{aligned} 2\|\mu_k - \pi\|_{TV} &= \sum_{x\in\mathcal{X}} |\mu_k(x) - \pi(x)| \\ &= \sum_{x\in\mathcal{X}} |\frac{\mu_k(x)}{\pi(x)} - 1|\pi(x) \\ &= \|\mu_k - \pi\|_{L^1(*)} \\ &\leq \|\mu_k - \pi\|_{L^2(*)} \end{aligned}$$

where the inequality comes from the Cauchy-Schwarz inequality.

This bound does not depend on π, whereas using the $L^2(\pi)$ norm, the coefficient for the bound depends on π:

$$2\|\mu_k - \pi\|_{TV} \leq \sqrt{\frac{n}{\min_x \pi(x)}}\|\mu_k - \pi\|_{L^2(\pi)}$$

However, as we will see, in the case of random walks on groups, π is uniform on $\mathcal{X}$, and both definitions of the norm work, and we will keep using the $L^2(\pi)$ norm.

7 Random Walks on Groups

In the last section we saw that we could get various bounds on convergence in terms of eigenvalues and eigenvectors of the transition matrix P. However, in general, it is usually hard to find these if $\mathcal{X}$ is large.

In the case of random walks on groups, however, it is always possible to obtain explicit forms for the eigenvalues and eigenvectors.

Definition 42. A random walk on a group is a Markov chain on a state space of some general discrete group $\mathcal{X}$. Transition probabilities are written as $P(x, y) = Q(x^{-1}y)$ where $Q(\cdot)$ is some fixed step distribution on $\mathcal{X}$. The increment distributions defined by Q are i.i.d.

Example 43. Our previous example of card shuffling is a random walk on the group S_n, the symmetric group of permutations.

Example 44. Random walk on $\mathbb{Z}$:

- For the random walk with $\frac{1}{2} - \frac{1}{2}$ probabilities of moving by +1 and -1:

$$Q(+1) = Q(-1) = \frac{1}{2}$$

- For the random walk with $\frac{1}{3} - \frac{1}{3} - \frac{1}{3}$ probabilities of moving by +1, 0 and -1:

$$Q(+1) = Q(0) = Q(-1) = \frac{1}{3}$$

It is possible to work with continuous groups, e.g. $O(n)$, the orthogonal group, containing the set of all $n \times n$ orthogonal matrices. For now, we restrict our discussion to finite, abelian groups. As the law of composition is commutative on these groups, we use addition notation, i.e. replace $Q(x^{-1}y)$ with $Q(y - x)$ to represent $P(x, y)$.

Fact 45. A random walk P on a finite group always has $\pi = \text{Unif}(\mathcal{X})$, i.e. $\pi(x) = \frac{1}{n}, \forall x \in \mathcal{X}$, since P is doubly stochastic (i.e. $\sum_x P(x, y) = 1, \forall y$).

Proof.

$$\sum_x P(x, y) = \sum_x Q(y - x) = \sum_z Q(z) = 1$$

□

Fact 46. Finite abelian groups are always of the form $\mathcal{X} = \mathbb{Z}/(n_1) \times \mathbb{Z}/(n_2) \times ... \times \mathbb{Z}/(n_r)$

Example 47. Frog walk: A frog jumps on a circular arrangement of 20 lilypads, each time moving clockwise or counterclockwise by one lilypad. The state space is given by $\mathcal{X} = \mathbb{Z}/(20)$.

Example 48. Bit flipping: A set of bits of length d can have each bit, or no bit flipped at each timestep with probability $\frac{1}{d+1}$.

- The state space is given by $\mathcal{X} = \big(\mathbb{Z}/(2)\big)^d$.
- The step distribution is given by $Q(0) = Q(\mathbf{e}_1) = ... = Q(\mathbf{e}_d)$, where $\mathbf{e}_i = (0, 0, 0, ..., 0, 1, 0, ...)$, i.e. all entries are 0, except for the ith entry, which is replaced by 1.

For the above examples, it is not immediately obvious what the eigenvalues and eigenvectors are. To derive these, we introduce **characters**, which are the beginnings of representation theory of groups.

7.1 Characters

Definition 49. $\chi_m : \mathcal{X} \to \mathbb{C}$ is a character defined for $m = (m_1, m_2, ..., m_r) \in \mathcal{X}$,

$$\chi_m(x) = e^{2\pi i(\frac{m_1 x_1}{n_1} + \frac{m_2 x_2}{n_2} + ... \frac{m_r x_r}{n_r})}$$

Note 50. (some identities)

1. $\chi_m(x + y) = \chi_m(x)\chi_m(y)$
2. $\chi_m(0) = 1$
3. $|\chi_m(x)| = 1$
4. $\chi_m(-x) = \overline{\chi_m(x)}$
5. $\sum_{m\in\mathcal{X}} \chi_m(x) = \begin{cases} n, x = 0 \\ 0, x \neq 0 \end{cases} = n\delta_{x0}$, where $n = n_1 n_2 ... n_r = |\mathcal{X}|$
6. $\langle \chi_m, \chi_j \rangle_{L^2(\pi)} = \sum_{x\in\mathcal{X}} \chi_m(x)\overline{\chi_j(x)}\pi(x) = \sum_{x\in\mathcal{X}} \chi_m(x)\chi_j(x)\frac{1}{n} = \begin{cases} 1, m = j \\ 0, m \neq j \end{cases} = \delta_{mj}$

Identity 5 for $x \neq 0$ follows from the fact that $\chi_m(x)$ are equally distributed on the unit circle in the complex plane, or alternatively, noting that this is a product of geometric sums that evaluate to 0

$$\sum_{m\in\mathcal{X}} \chi_m(x) = \sum_{m\in\mathcal{X}} \Big(\prod_{j=1}^{r} e^{2\pi i \frac{m_j x_j}{n_j}} \Big) = \prod_{j=1}^{r} \Big(\sum_{m_j=0}^{n_j-1} e^{\frac{2\pi i m_j x_j}{n_j}} \Big) = \prod_{j=1}^{r} \frac{1 - e^{2\pi i x_j}}{1 - e^{\frac{2\pi i x_j}{n_j}}} = 0$$

In Identity 6, we have made use of the fact that $\pi = \text{Unif}(\mathcal{X})$, and when $m \neq j$, the sum reduces to $\sum_x \chi_{m-j}(x)$, which is zero for the same reason as in Identity 5.

It follows therefore from Identity 6 that $\{\chi_m\}$ are orthonormal. It remains to be shown that they are eigenvectors.

$$\begin{aligned}\big(\overline{\chi_m}P\big)(y) &= \sum_{x\in\mathcal{X}} \overline{\chi_m(x)}P(x,y)\\ &= \sum_{x\in\mathcal{X}} \chi_m(-x)P(x,y)\\ &= \sum_{x\in\mathcal{X}} \chi_m(-x)Q(y-x)\end{aligned}$$

Making the change of variable $z = y - x, -x = z - y$:

$$\begin{aligned}\big(\overline{\chi_m}P\big)(y) &= \sum_{z\in\mathcal{X}} \chi_m(z-y)Q(z)\\ &= \sum_{z\in\mathcal{X}} \chi_m(z)\chi_m(-y)Q(z)\\ &= \overline{\chi_m(y)}\sum_{z\in\mathcal{X}} \chi_m(z)Q(z)\\ &= E_Q(\chi_m)\overline{\chi_m(y)}\end{aligned}$$

Therefore, $\{\chi_m\}$ is the set of eigenvectors, with corresponding eigenvectors, $\{\lambda_m\}$ being the expectation of the characters under Q. As usual, we set $\lambda_0 = 1$ and define $\lambda_* = \max_{m\neq 0}|\lambda_m|$.

Finally we want to be convinced that in the case when the random walker starts in a designated position, i.e. $\mu_0 = \delta_0(\cdot)$ is a point mass, we can still write μ_0 as a linear combination of the eigenvectors of P. I.e. we want to show that $\mu_0 = \sum_m a_m v_m$ for some set of complex coefficients $\{a_m\}$.

This can be done by simply observing that $a_m = \frac{1}{n}$ since $\sum_m \overline{\chi_m(x)} = n\delta_{x0} = \sum_m v_m$. This leads us to conclude that $\mu_0 - \pi = \frac{1}{n}(\sum_m v_m - \mathbf{1}) = \frac{1}{n}(\sum_{m\neq 0} v_m)$.

Therefore $\mu_k = \frac{1}{n}\sum_m(\lambda_m)^k v_m$, and $\mu_k - \pi$. From this we obtain

$$\sum_{x\in\mathcal{X}}|\mu_k(x) - \pi(x)|^2\pi(x) = \sum_{m\neq 0}|a_m|^2|\lambda_m|^{2k}$$

as $\{v_m\}$ are orthonormal.

And as $\pi(x) = \frac{1}{n} = a_m$

$$\sum_{x\in\mathcal{X}}|\mu_k(x) - \pi(x)|^2 = \frac{1}{n}\sum_{m\neq 0}|\lambda_m|^{2k}$$

We therefore obtain a bound on the total variation distance

$$\begin{aligned}\Big(2\|\mu_k-\pi\|_{TV}\Big)^2 &= \Big(\sum_{x\in\mathcal{X}}|\mu_k(x)-\pi(x)|\Big)^2\\ &= \Big(n\sum_{x\in\mathcal{X}}|\mu_k(x)-\pi(x)|\pi(x)\Big)^2\\ &= n^2\Big(\langle\mu_k-\pi,\mathbf{1}\rangle\Big)^2\\ &\le n^2\|\mu_k-\pi\|^2_{L^2(\pi)}\|\mathbf{1}\|^2_{L^2(\pi)}\\ &= n^2\|\mu_k-\pi\|_{L^2(\pi)}\\ &= \sum_{m\neq 0}|\lambda_m|^{2k}\end{aligned}$$

where again the inequality comes from the Cauchy-Schwarz inequality.

Conclusion 51. $\|\mu_k-\pi\|_{TV}\le\frac{1}{2}\sqrt{\sum_{m\neq0}|\lambda_m|^{2k}}\le\frac{\sqrt{n-1}}{2}(\lambda_*)^k$

7.2 Application to examples

7.2.1 Frog walk

$\mathcal{X}=\mathbb{Z}/(n),\qquad Q(0)=Q(1)=Q(-1)=\frac{1}{3}$

$$\begin{aligned}\chi_m(x) &= e^{2\pi i(\frac{mx}{n})}\\ \lambda_m &= E_Q\big(\overline{\chi_m}\big)\\ &= \frac{1}{3}\overline{\chi_m}(0)+\frac{1}{3}\overline{\chi_m}(1)+\frac{1}{3}\overline{\chi_m}(-1)\\ &= \frac{1}{3}(1)+\frac{1}{3}e^{\frac{-2\pi im}{n}}+\frac{1}{3}e^{\frac{2\pi im}{n}}\\ &= \frac{1}{3}+\frac{2}{3}\cos(\frac{2\pi m}{n})\end{aligned}$$

$m=0$ corresponds to $\lambda_m=1$. It can be seen that as m increases, $\cos(\frac{2\pi m}{n})$ decreases, then increases back towards 1, but cannot exceed $\cos(\frac{2\pi}{n})$. The value for λ_* is therefore $\frac{1}{3}+\frac{2}{3}\cos\big(\frac{2\pi}{n}\big)$.

$$\begin{aligned}\|\mu_k-\pi\| &\le \frac{\sqrt{n}}{2}\big(\frac{1}{3}+\frac{2}{3}\cos(\frac{2\pi}{n})\big)^k\\ &= \frac{\sqrt{n}}{2}\big(1-\frac{2}{3}(1-\cos(\frac{2\pi}{n}))\big)^k\end{aligned}$$

Assuming $n\ge3$, we have that for $0\le x\le\sqrt{6}$, $\cos(x)\le1-\frac{x^2}{4}$. Further, $1-x\le e^{-x}$, therefore,

$$\|\mu_k - \pi\| \leq \frac{\sqrt{n}}{2} e^{-\frac{2\pi^2}{3n^2}k}$$

For $n = 1000$, this gives $k_* = 1120000$. This bound requires k to be on the order of $n^2\log(n)$. Since we know all the eigenvalues, we can obtain a tighter bound

$$\begin{aligned}
\|\mu_k - \pi\|^2 &\leq \frac{1}{4}\sum_{m=1}^{n-1} |\lambda_m|^{2k} \\
&\leq \sum_{m=1}^{\lceil\frac{n-1}{4}\rceil} e^{-\frac{4\pi^2 m^2}{3n^2}k} \\
&\leq \sum_{m=1}^{\infty} e^{-\frac{4\pi^2 m}{3n^2}k} \\
&= \frac{e^{-\frac{4\pi^2}{3n^2}k}}{1 - e^{-\frac{4\pi^2}{3n^2}k}}
\end{aligned}$$

Which for $n = 1000$, gives $k_* = 351000$. This bound now scales with n^2. How much tighter still can we make this bound? To answer this question, we look at the lower bound for convergence. First note that as $E_{\mu_k}(\chi_m)$ is the eigenvalue of P^k corresponding to the eigenvector $\overline{\chi_m}$, it is equal to the kth power of the corresponding eigenvalue of P:

$$E_{\mu_k}(\chi_m) = (E_Q(\chi_m))^k$$

We therefore have

$$\begin{aligned}
\|\mu_k - \pi\| &= \frac{1}{2}\sup_{|f|\leq 1} |E_{\mu_k}(f) - E_\pi(f)| \\
&\geq \frac{1}{2}|E_{\mu_k}(\chi_1) - 0| \\
&= \frac{1}{2}|E_Q(\chi_1)|^k \\
&= \frac{1}{2}\left(\frac{1}{3} + \frac{2}{3}\cos\left(\frac{2\pi}{n}\right)\right)^k
\end{aligned}$$

where the inequality comes from the fact that the supremum of a set must be greater than or equal to any member of that set.

so for $n = 1000$, $k_* \geq 290000$, therefore our previous bound cannot be improved by much more.

STA4502: Topics in Stochastic Processes Winter 2018

Lecture 5: Introduction — March 28, 2018

Lecturer: Jeffrey Rosenthal *Scribe: Jeffrey Negrea*

1 Last Example of Random Walks on Groups

1.1 Bit-Flipping

Recall the bit-flipping example:

$$\begin{aligned}\mathcal{X} &= (\mathbb{Z}/2)^d \\ Q &= \text{Unif}(\{\text{id}\} \cup \{e_j : j \in [d]\})\end{aligned}$$

where e_j frobnicates the jth bit, leaving the other $j-1$ bits unchanged, and $\text{id} = 0$ is the identity element.

We already derived the characters for this group:

$$\chi_m(x) = \exp\left(2\pi i \sum_{i=d}^{n} \frac{m_i x_i}{2}\right) = (-1)^{\langle m,\ x\rangle}$$

The eigenvalues of P may then be computed

$$\begin{aligned}\lambda_m &= \mathop{\mathbb{E}}_{X\sim Q}[\chi_m(X)] \\ &= \sum_{x\in\mathcal{X}} Q(x)\chi_m(x) \\ &= Q(0)\chi_m(0) + \sum_{i=1}^{d} Q(e_i)\chi_m(e_i) \\ &= \frac{1}{d+1}\left(1 + \sum_{i=1}^{d}(-1)^{\langle m,\ e_i\rangle}\right) \\ &= \frac{1}{d+1}\left(1 - N(m) + (d - N(m))\right) \\ &= 1 - \frac{2N(m)}{d+1}\end{aligned}$$

where $N(m) = \langle m,\ \mathbf{1}\rangle$ is the number of 1s in m, since $\langle m,\ e_i\rangle$ is 1 if $m_i = 0$ and is -1 otherwise.

Thus $\lambda_\star = 1 - \frac{2}{d+1}$, which is realised when $N(m) = 1$, for example when $m = e_1$.

Using the crude $\lambda - \star$-based method, we have the following bound for the total variation distance of the marginal distribution of the chain to stationarity:

$$\|\mu_k - \pi\|_{\text{TV}} \leq \frac{\sqrt{|\mathcal{X}|}}{2}\lambda_\star^k = \frac{\sqrt{|\mathcal{X}|}}{2}\left(1 - \frac{2}{d+1}\right)^k$$

For $d = 1000$ we get that $k_\star = 175243$ is sufficient for $\|\mu_{k_\star} - \pi\|_{\mathrm{TV}} \leq 0.01$

Using the more refined summation-based method, we have the following bound for the total variation distance of the marginal distribution of the chain to stationarity:

$$\begin{aligned}
\|\mu_k - \pi\|_{\mathrm{TV}} &\leq \frac{1}{2}\sqrt{\sum_{m \in \mathcal{X}\setminus\{0\}} |\lambda_m|^{2k}} \\
&\leq \frac{1}{2}\sqrt{\sum_{m \in \mathcal{X}\setminus\{0\}} \left|1 - \frac{2N(m)}{d+1}\right|^{2k}} \\
&\leq \frac{1}{2}\sqrt{\sum_{n=1}^{d} \binom{d}{n} \left|1 - \frac{2n}{d+1}\right|^{2k}}
\end{aligned}$$

For $d = 1000$ we get that $k_\star = 3684$ is sufficient for $\|\mu_{k_\star} - \pi\|_{\mathrm{TV}} \leq 0.01$. This was calculated with the following `R` script:

```
bins = choose(1000,1:1000)
pows = abs(1-2*(1:1000)/1001)
tv.bound = function(k){1/2 * sqrt(sum(bins * pows ^ (2 * k))) -0.01}
k.star = ceiling(uniroot(tv.bound,c(0,176000))$root)
```

We can also get a lower bound for the total variation distance from stationarity, which gives a necessary number of steps through the chain:

$$\begin{aligned}
\|\mu_k - \pi\|_{\mathrm{TV}} &\geq \frac{1}{2}\left|\mathop{\mathbb{E}}_{X\sim Q}[\chi_{e_1}]\right|^k \\
&= \frac{1}{2}\left(1 - \frac{2}{d+1}\right)^k
\end{aligned}$$

For $d = 1000$ we get that $k_\star \geq 1957$ is necessary for $\|\mu_{k_\star} - \pi\|_{\mathrm{TV}} \leq 0.01$.

Putting these together we get that, for $d = 1000$, the true $k_\star$ is between 1957 and 3684.

2 Drift and Minorisation Conditions

Recall the uniform minorisation condition:

If $P(x, \cdot) \geq \epsilon\rho(\cdot)$ for all $x \in \mathcal{X}$ for some $\epsilon > 0$ and some probability measure ρ on $\mathcal{X}$, then the markov chain is *uniformly geometrically ergodic*. That is to say, for any initial probability measure μ_0 and for any $k \in \mathbb{N}$:

$$\|\mu_k - \pi\|_{\mathrm{TV}} \leq (1 - \epsilon)^k \ .$$

The universal quantification over all initial measures seems to be nice mathematically, but restricts our analysis to Markov chains which converge to stationarity uniformly. In order to be able to analyse markov chains without uniform convergence properties we need to develop new tools. Te following example will be used to illustrate non-uniform ergodicity in this section:

Example 1 (Canonical non-uniformly ergodic example: AR(1)-process). . A particular gaussian autoregresive process of order 1 is given by:

Let $\mathcal{X} = \mathbb{R}$ and let $P(x, \cdot) \equiv \mathcal{N}(\cdot\,; \frac{x}{2}, \frac{3}{4})$.

This kernel "pulls" the chain back to 0 on each step. The farther the chain is from 0 the longer it will take to return to a neighbourhood of 0.

Does this process have a stationary distribution? Yes! The stationary distribution is $\mathcal{N}(0, 1)$. We verify this below. Suppose that $X_n \sim \mathcal{N}(0, 1)$, then;

$$\begin{aligned} X_n \perp\!\!\!\perp Z = X_{n+1} - \frac{X_n}{2} &\sim \mathcal{N}(0, \frac{3}{4}) \\ \implies X_{n+1} &\sim \mathcal{N}(0, 1)\ . \end{aligned}$$

Since this example is so simple, we could directly bound its total variation distance from stationarity. Since the methods used wouldn't generalise, this would not be instructive. In this section we will develop generally applicable techniques, and then apply them to this example.

In this example, we cannot get uniform minorisation for all $x \in \mathcal{X}$ since, taking any two x sufficiently far apart w,e could show that no uniform minorising probability measure exists for any $\epsilon > 0$.

2.1 Drift and minorisation derivation

Instead of minorising uniformly over the whole state space, we may instead attempt to minorise only uniformly over some subset of the state space. More precisely we will attempt to find $C \subset \mathcal{X}$ such that $P(x, \cdot) \geq \epsilon\rho(\cdot)$ for all $x \in C$. We will call such a C a "small set".

We cannot use the same construction as in the uniform case — we need to first allow both copies of the chain to reach the small set, then hope that the chains couple.

2.1.1 Coupling Construction

The coupling is constructed as follows. Let $X_0 \sim \mu_0$ and let $Y_0 \sim \pi$. At stage n, given the coupled X_n and Y_n and $Z_n \sim \text{Bernoulli}(\epsilon)$ we determine the coupled X_{n+1} and Y_{n+1} by:

if	$X_n = Y_n$	then	$X_{n+1} = Y_{n+1} \sim P(X_n, \cdot)$	
else if	$(X_n, Y_n) \in C^2 \ \wedge \ Z_n = 1$	then	$X_{n+1} = Y_{n+1} \sim \rho(\cdot)$	
else if	$(X_n, Y_n) \in C^2 \ \wedge \ Z_n = 0$	then	$X_{n+1} \sim \frac{P(X_n, \cdot) - \epsilon\rho(\cdot)}{1-\epsilon} = R(X_n, \cdot)$	
			$\perp\!\!\!\perp\ Y_{n+1} \sim \frac{P(Y_n, \cdot) - \epsilon\rho(\cdot)}{1-\epsilon} = R(Y_n, \cdot)$	
else	$(X_n, Y_n) \notin C^2$	then	$X_{n+1} \sim P(X_n, \cdot)$	
			$\perp\!\!\!\perp\ Y_{n+1} \sim P(Y_n, \cdot)$	

2.1.2 Coupling Inequality

To bound the distance from stationarity we use the coupling inequality:

$$\|\mu_k - \pi\|_{\mathrm{TV}} \leq \mathbb{P}(X_k \neq Y_k) \ .$$

Choose $j \in [k]$. Let $N_k = |\left\{m \in [k] : (X_m, Y_m) \in C^2\right\}|$. We can then decompose the RHS above as:

$$\begin{aligned}\mathbb{P}(X_k \neq Y_k) &= \mathbb{P}(X_k \neq Y_k, N_{k-1} \geq j) + \mathbb{P}(X_k \neq Y_k, N_{k-1} < j) \\ &\leq (1-\epsilon)^j + \mathbb{P}(X_k \neq Y_k, N_{k-1} \leq j-1)\end{aligned}$$

We need some new techniques to bound $\mathbb{P}(X_k \neq Y_k, N_{k-1} \leq j-1)$, the probability that we have an insufficient number of chances to couple and do not couple.

2.1.3 Drift conditions

The new trick will be to introduce a "drift condition". There are univariate and bivariate versions of drift conditions. In this course we will only examine bivariate versions.

We introduce the forward expectation operator $\overline{P}$ defined by

$$\overline{P}h(x,y) = \mathbb{E}[h(X_1, Y_1)\big|(X_0, Y_0)] = (x,y)$$

Suppose that we have a function $h : \mathcal{X}^2 \to [1, \infty)$ and $\alpha > 1$ such that

$$\overline{P}h(x,y) \leq \frac{h(x,y)}{\alpha} \qquad \forall (x,y) \notin C^2 \tag{1}$$

Then h is called a drift function and (**??**) is called a drift . The key idea is that, on average, h gets smaller per step of the Markov chain. Often times we will use an additive, symmetric drift function of the form $h(x,y) = 1 + V(x) + V(y)$ for $V : \mathcal{X} \to [0, \infty)$. For the AR(1) example, we will use $V(x) = x^2$ and so $h(x,y) = 1 + x^2 + y^2$.

2.1.4 Final coupling bound

Suppose we have a drift condition as well as a local minorisation condition. Then, for $B \geq 1$,

$$\begin{aligned}&\mathbb{P}(X_k \neq Y_k, N_{k-1} \leq j-1) \\ \text{equality if } B \neq 1 \qquad &\leq \mathbb{P}(X_k \neq Y_k, B^{-N_{k-1}} \geq B^{-(j-1)}) \\ &= \mathbb{P}\left((\mathbf{1}_{X_k \neq Y_k} B^{-N_{k-1}}) \geq B^{-(j-1)}\right) \\ \text{Markov's Ineq.} \qquad &\leq B^{j-1}\mathbb{E}\left[\mathbf{1}_{X_k \neq Y_k} B^{-k-1}\right] \\ &\leq B^{j-1}\alpha^{-k}\mathbb{E}\left[\mathbf{1}_{X_k \neq Y_k}\alpha^k B^{-N_{k-1}} h(X_k, Y_k)\right]\end{aligned}$$

Let $M_k = \mathbf{1}_{X_k \neq Y_k}\alpha^k B^{-N_{k-1}} h(X_k, Y_k)$and let $B = 1 \vee \left[\alpha(1-\epsilon)\sup_{(x,y)\in C^2} \mathbb{E}_{\substack{X_1 \sim R(x,\cdot) \\ Y_1 \sim R(y,\cdot)}}[h(X_1, Y_1)]\right]$.

We claim that M_k is a supermartingale with respect to the filtration generated by $\{(X_k, Y_k) : k \in \mathbb{N}\}$. In light of this claim, we have:

$$\begin{aligned}\mathbb{P}(X_k \neq Y_k, N_{k-1} \leq j-1) &\leq B^{j-1}\alpha^{-k}\mathbb{E}[M_0] \\ &= B^{j-1}\alpha^{-k}\mathbb{E}[h(X_0, Y_0)]\end{aligned}$$

We verify the submartingale claim below. The proof relies on the fact the the N_{k-1} term in the definition of M_k is predictable.

Case 1: The chain coupled by time $k+1$, so that $X_{k+1} = Y_{k+1}$. Then $M_{k+1} = 0 \leq M_k$.

Case 2: The chains were not coupled and did not have a chance to couple at time $k+1$, i.e.: $(X_k, Y_k) \notin C^2$ and $X_{k+1} \neq Y_{k+1}$. Then $N_{k-1} = N_k$ (no new chance to couple) so that

$$\begin{aligned}&\mathbb{E}[M_{k+1}\big|(X_k, Y_k)] \\ &\quad= \mathbb{E}[\mathbf{1}_{X_k = Y_k} M_{k+1}\big|(X_k, Y_k)] \\ &\qquad+ \mathbb{E}[\mathbf{1}_{(X_k,Y_k)\notin C^2}\mathbf{1}_{X_k \neq Y_k}\ M_{k+1}\big|(X_k, Y_k)] \\ &\qquad+ \mathbb{E}[\mathbf{1}_{(X_k,Y_k)\in C^2}\mathbf{1}_{X_k \neq Y_k}\ M_{k+1}\big|(X_k, Y_k)]\end{aligned}$$

The first term is 0 since if the chain is coupled at time k then it is coupled at time $k+1$ so $M_{k+1} = 0$.

The second term can be bounded by:

$$\begin{aligned}\mathbb{E}[\mathbf{1}_{(X_k,Y_k)\notin C^2}\mathbf{1}_{X_k\neq Y_k} M_{k+1}\big|(X_k, Y_k)] &\leq \mathbb{E}[\mathbf{1}_{(X_k,Y_k)\notin C^2 \wedge X_k \neq Y_k}\ \alpha^{k+1}B^{-N_k}h(X_{k+1}, Y_{k+1})\big|(X_k, Y_k)] \\ &= \alpha\mathbb{E}[\mathbf{1}_{(X_k,Y_k)\notin C^2 \wedge X_k \neq Y_k}\ \alpha^{k}B^{-N_{k-1}}h(X_{k+1}, Y_{k+1})\big|(X_k, Y_k)] \\ &= \mathbf{1}_{(X_k,Y_k)\notin C^2} M_k\ \frac{\mathbb{E}[h(X_{k+1}, Y_{k+1})\big|(X_k, Y_k)]}{h(X_k, Y_k)/\alpha} \\ &\leq \mathbf{1}_{(X_k,Y_k)\notin C^2} M_k\end{aligned}$$

The third term can be bounded by:

$$\begin{aligned}\mathbb{E}[\mathbf{1}_{(X_k,Y_k)\in C^2}\mathbf{1}_{X_k\neq Y_k} M_{k+1}\big|(X_k, Y_k)] &\leq \mathbb{E}[\mathbf{1}_{(X_k,Y_k)\in C^2 \wedge X_k \neq Y_k \wedge Z_k = 0}\ \alpha^{k+1}B^{-N_k}h(X_{k+1}, Y_{k+1})\big|(X_k, Y_k)] \\ &\leq \frac{\alpha}{B}\mathbb{E}[\mathbf{1}_{(X_k,Y_k)\in C^2 \wedge X_k \neq Y_k \wedge Z_k = 0}\ \alpha^{k}B^{-N_{k-1}}h(X_{k+1}, Y_{k+1})\big|(X_k, Y_k)] \\ &= \mathbf{1}_{(X_k,Y_k)\in C^2} M_k \frac{\mathbb{E}[\mathbf{1}_{Z_k=0}h(X_{k+1}, Y_{k+1})\big|(X_k, Y_k)]}{Bh(X_k, Y_k)/\alpha} \\ &= \mathbf{1}_{(X_k,Y_k)\in C^2} M_k \frac{(1-\epsilon)\mathbb{E}[h(X_{k+1}, Y_{k+1})\big|(X_k, Y_k, Z_k = 0)]}{Bh(X_k, Y_k)/\alpha} \\ &\leq \mathbf{1}_{(X_k,Y_k)\in C^2} M_k\end{aligned}$$

The last step follows from the choice of B. Combining these, we get the supermartingale property for M_k.

2.1.5 Drift and minorisation theorem

Theorem 2. *If a Markov chain has a stationnary distribution, π, and a local minorisation condition of the form:*

$$\exists(C \in \Sigma_{\mathcal{X}},\ \epsilon > 0,\ \rho \in \mathcal{M}(\Sigma_{\mathcal{X}}))\ :\ (\pi(C) > 0 \text{ and } (x \in C \implies P(x,\cdot) \geq \epsilon\rho(\cdot)))\ ,$$

and a drift condition of the form:

$$(\exists h : \mathcal{X}^2 \to [1,\infty),\ \alpha > 0)\ :\ ((x,y) \notin C \times C \implies \overline{P}h(x,y) \leq \alpha^{-1}h(x,y))$$

then for any $j \in [k]$ we have

$$\|\mu_k - \pi\|_{TV} \leq (1-\epsilon)^j + \alpha^{-k}B^{j-1}\mathbb{E}h(x_0,y_0)\ ,$$

where $B = 1 \vee \left[\alpha(1-\epsilon)\sup_{(x,y)\in C^2} \mathop{\mathbb{E}}_{\substack{X_1\sim R(x,\cdot)\\ Y_1\sim R(y,\cdot)}}[h(X_1,Y_1)]\right]$.

Example 3 (Canonical non-uniformly ergodic example: AR(1)-process — continued). We will choose $C = [-\sqrt{3},\sqrt{3}]$ and $h(x,y) = 1 + x^2 + y^2$. Then we get:

$$\begin{aligned}
\epsilon &= \int_{\mathbb{R}} \inf_{x\in C} \mathcal{N}(\mathrm{d}y\ ;\ \frac{x}{2}, \frac{3}{4}) \\
&= \mathbb{P}(|\mathcal{N}(0,1)| \geq 1) \\
&= 0.3173105\ ,
\end{aligned}$$

and, for $(x,y) \notin C^2$ we have $h(x,y) \geq 4$, so then:

$$\begin{aligned}
\overline{P}h(x,y) &= 1 + \left(\frac{x^2}{4} + \frac{3}{4}\right) + \left(\frac{y^2}{4} + \frac{3}{4}\right) \\
&= \frac{9 + h(x,y)}{4} = \frac{\frac{9}{4}4 + h(x,y)}{4} \\
&\leq \frac{h(x,y)}{4}\left(1 + \frac{9}{4}\right) \\
&= \frac{13}{16}h(x,y)
\end{aligned}$$

which means we can take

$$
\begin{aligned}
\alpha &= \frac{16}{13} \\
B &= \frac{16}{13}(1-0.3173105) \sup_{(x,y)\in C^2} \mathop{\mathbb{E}}_{\substack{X_1\sim R(x,\cdot)\\ Y_1\sim R(y,\cdot)}} [h(X_1,Y_1)] \\
&\le \frac{16}{13}(1-0.3173105) \sup_{(x,y)\in C^2} \left(1+\overline{R}[x^2]+\overline{R}[y^2]\right) \\
&= \frac{16}{13}(1-0.3173105) \sup_{(x,y)\in C^2} \left(1+\frac{\overline{P}[x^2]-\epsilon \mathop{\mathbb{E}}_{W\sim\rho}[W]}{1-\epsilon}+\frac{\overline{P}[y^2]-\epsilon \mathop{\mathbb{E}}_{W\sim\rho}[W]}{1-\epsilon}\right) \\
&= \frac{16}{13}(1-0.3173105) \sup_{(x,y)\in C^2} \left(1+\frac{\overline{P}[x^2]}{1-\epsilon}+\frac{\overline{P}[y^2]}{1-\epsilon}\right) \\
&= \frac{16}{13}(1-0.3173105) \left(1+\frac{3/4+3/4}{1-\epsilon}+\frac{3/4+3/4}{1-\epsilon}\right) \\
&= \frac{16}{13}(1-0.3173105) \left(1+\frac{3}{1-\epsilon}\right) \\
&= \frac{16}{13}(4-0.3173105) \\
&= 4.532541 \le 4.6
\end{aligned}
$$

Take $j = \lfloor k/10 \rfloor$

Then we get the following bound for $\mu_0 = \delta_0$:

$$
\|\mu_k - \pi\|_{\mathrm{TV}} \le (0.683)^{\lfloor k/10 \rfloor} + \left(\frac{13}{16}\right)^k 4.6^{\lfloor k/10 \rfloor - 1}\ 2 \tag{2}
$$

Since $\mathbb{E}h(X_0, y_0) = 1 + 0 + \mathbb{E}Y_0^2 = 2$

Finally, for $k_\star = 130$ we have $\|\mu_{k_\star} - \pi\|_{\mathrm{TV}} \le 0.01$, which was computed with the following `R` script:

```
tv.bound = function(k){(1-2*pnorm(-1))^floor(k/10) +
                       (13/16)^k * ((16/13) * (4- 0.3173105))^(floor(k/10)-1)*2 -0.01 }
k.star = ceiling(uniroot(tv.bound,c(0,1000))$root)
```