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

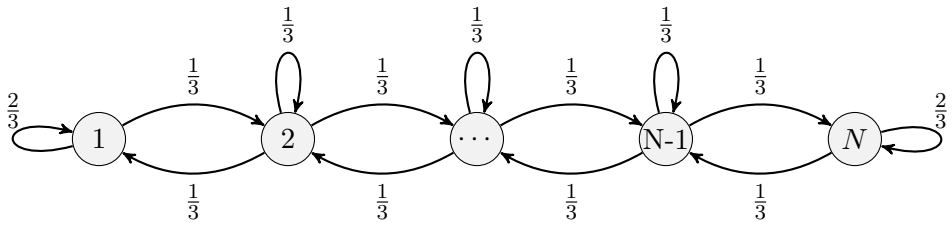
$$P^4(x,3) \geq \frac{1}{16}\delta_3(\cdot), \ \forall x \Rightarrow \|P^{k_0}(x,\cdot) - \pi\| \leq \left(\frac{15}{16}\right)^{\lfloor k/4 \rfloor} \leq 0.01 \Rightarrow k^* = 288$$

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

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

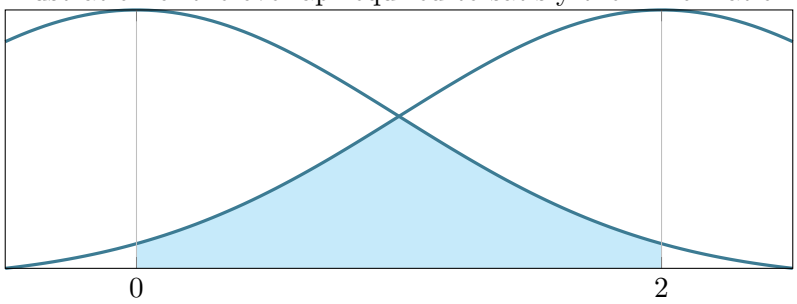
where $N(x, 1; A) = \Pr(z \in A)$ with $z \sim N(x, 1)$, and where $r(x) = 1 - N(x, 1; \mathcal{X})$, the probability that a draw from $N(x, 1)$ falls outside $\mathcal{X}$. (This corresponds to the Metropolis-Hastings algorithm with $\pi = \text{Unif}[0, 2]$.)

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.