Notes on Convergence Rate of Markov Chains (Eigen)

Yu Hang Jiang, Tong Liu, Zhiya Lou,
Shanshan Shangguan, Fei Wang, Zixuan Wu
Under the supervision of Jeffrey S. Rosenthal

April 2020

1 Introduction and Motivation

This is a supplementary reading for the paper *Convergence Rate of Markov Chains* by Jeffrey S. Rosenthal. It is meant to help readers have better understanding of the original paper.

1.1 Convergence of Markov Chain

The main motivation of the paper is raised by the following two questions:
Does a Markov chain converge? If it does, what is the rate of convergence?

1.2 Key application: Markov Chain Monte Carlo algorithm

Markov Chain Monte Carlo algorithm (abbreviated as MCMC algorithm) is an approach where a Markov chain is defined in a way that it will converge to a certain probability distribution of interest. That's saying, given a probability distribution, we are able to construct a Markov Chain that converges to this probability distribution.

Examples of such algorithm in applied settings include but not restricted to Gibbs sampler in statistics(a sampling method to get the joint distribution when conditional distributions are known), approximation algorithm in computer science, and stochastic algorithms in physics.

1.3 Other applications

Convergence rates for Markov chains also applied to card shuffling(seven ordinary "riffle" shuffle), and method for generating random matrices to be used for encryption algorithms (in particular random walk on groups).

1.4 Opportunities for further works

Examine some of the applied algorithms which use Markov chain; get bounds on convergence rate.

1.5 Summary of the paper

Section 1-3: preliminary materials
Section 4: basic connection between Markov chains and eigenvalues
Section 5: random walks on groups (not included in this notes)
Section 6: coupling and minorisation conditions

2 Basic definitions

2.1 Definition of a Markov chain

A Markov chain consists of
(i) a measurable state space $\mathcal{X}$
(ii) an initial distribution μ_0 on $\mathcal{X}$
(iii) transition probabilities $P(x, dy)$
(iv) $\int_A(x) = P(x, A)$
is a measurable function of $x \in \mathcal{X}$ for each fixed set $A \subset \mathcal{X}$

2.2 Multi-step transition probability distribution

Definition: μ_k on $\mathcal{X}$ is a set of probabilities of where the Markov chain will be after k steps. And

$$\mu_k(\mathcal{X}) = \int_{\mathcal{X}} P(x, A)\mu_{k-1}(dx)$$

In discrete case, we can write

$$\mu_k(y) = \sum_x P(x, y)\mu_{k-1}(x)$$

If we write μ_k as a row-vector, and P as a matrix with $[P]_{xy} = P(x, y)$, then

$$\mu_k = \mu_{k-1}P = \cdots = \mu_0 P^k$$

2.3 Total variation distance between probability measures

Definition: $||v_1 - v_2|| := \sup_{A\in\mathcal{X}} |v_1(A) - v_2(A)|$[1], where A is a measurable subset of $\mathcal{X}$. In finite case we have

$$||v_1 - v_2|| = \frac{1}{2}\sum_x |v_1(x) - v_2(x)|$$

When $||v_1 - v_2|| = 0$, we can see it as the case v_1 infinitely close to v_2, since the supremum of all the distance among choice of x is 0, all the other elements have smaller distance goes to 0.

Proof. If $\mathcal{X}$ is finite, then

$$|v_1(A) - v_2(A)| = \sum_{x\in A} |v_1(x) - v_2(x)|$$

Let

$$B = \{x \in \mathcal{X} : v_1(x) - v_2(x) \geq 0\}$$

Then clearly the maximum of $|v_1(A) - v_2(A)|$ is achieved either when $A = B$ or $A = B^c$. But

$$\begin{aligned}|v_1(B) - v_2(B)| - |v_1(B^c) - v_2(B^c)| &= v_1(B) - v_2(B) - (-(v_1(B^c) - v_2(B^c)) \\ &= v_1(B) + v_1(B^c) - v_2(B) - v_2(B^c) \\ &= 1 - 1 = 0\end{aligned}$$

Hence $|v_1(B) - v_2(B)| = |v_1(B^c) - v_2(B^c)|$. Therefore

$$\begin{aligned}\sup_{A\in\mathcal{X}} |v_1(A). - v_2(A)|| &= |v_1(B) - v_2(B)| = |v_1(B^c) - v_2(B^c)| \\ &= \frac{1}{2}(|v_1(B) - v_2(B)| + |v_1(B^c) - v_2(B^c)|) \\ &= \frac{1}{2}(\sum_{x\in B} |v_1(x) - v_2(x)| + \sum_{x\in B^c} |v_1(x) - v_2(x)|) \\ &= \frac{1}{2}\sum_{x\in\mathcal{X}} |v_1(x) - v_2(x)|.\end{aligned}$$

□

Moreover, for any $\mathcal{X}$,

$$||v_1 - v_2|| = \frac{1}{2}\sup_{f:\mathcal{X}\to C, |f|\leq 1} |E_{v_1}(f) - E_{v_2}(f)| = \sup_{f:X\to R,\ 0\leq f\leq 1} |E_{v_1}(f) - E_{v_2}(f)|$$

(Here, we take test functions instead of test sets)

[1] Intuitively it measures the maximum difference between the probabilities assigned to a single event by the two distributions. Maybe it is $1/2\int_{\mathcal{X}} |v_1(x) - v_2(x)|dx$ just like in finite case?

3 The simplest non-trivial example

3.1 Settings

$$\mathcal{X} = \{0, 1\},\ \mu_0 = (1, 0)$$
$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

Then

$$\mu_k(0) = \frac{q}{p+q} + (1 - \frac{q}{p+q})(1-p-q)^k$$
$$\mu_k(1) = \frac{p}{p+q} - (1 - \frac{q}{p+q})(1-p-q)^k$$

3.2 Observations

(1) Let $\pi = (\frac{q}{p+q}, \frac{p}{p+q})$. Assume $|1-p-q| < 1$, then

$$||\mu_k - \pi|| = |(\frac{q}{p+q})(1-p-q)^k| \to 0$$

(decrease exponentially quickly to 0, with rate governed by $(1-p-q)$.)

(2) The limiting distribution π is a stationary distribution: $\pi P = P$, and thus corresponds to a left-eigenvector of the matrix P with eigenvalue 1. It is easily seen that any limiting distribution π for any Markov chain must be a stationary distribution(since $\mu_k = \mu_{k-1}P$)

(3) The convergence fails when $p = q \in \{0, 1\}$. If $p = q = 0$ the Markov chain is decomposable, meaning that the space $\mathcal{X}$ contains two-empty disjoint closed subsets. If $p = q = 1$ then this Markov chain is periodic. (different from class, we say it is periodic if the space contains disjoint subsets $\mathcal{X}_1, \ldots, \mathcal{X}_d$ such that for any $x \in \mathcal{X}_j$, $P(x, \mathcal{X}_{j+1} = 1)$). If our Markov chain is indecomposible and aperiodic, then it converges exponentially quickly.

(4) The eigenvalues of the matrix P are 1 and $1-p-q$. We have a connection between trivial eigenvalues and non-trivial eigenvalues

(5) Define

$$\beta = \sum_y \min_x P(x, y)$$

Then $\beta = \min\{p+q, 2-p-q\}$. Then $1-\beta = |1-p-q|$ is the absolute value of the non-trivial eigenvalue as above. The relationship will be explored in Section 5 via the method of "coupling"

(6) This Markov chain is reversible. It guarantee all eigenvalues will be real (so diagonalizable). But not all Markov chain has such property. It will be discussed in Section 6

(7) When $p = q$, this corresponds to a simple random walk on group $Z/2Z$ with step distribution $Q(1) = p$ and $Q(0) = 1 - p$. Then $E_Q((-1)^x) = -p + (1 - p) = 1 - p - q$. So for simple random walk on groups, the eigenvalues can be computed by taking expected values with respect to Q. This is discussed in Section 4.

4 The eigenvalue connection

Now Consider finite space $\mathcal{X}$. Since $\mu_k = \mu_0 P^k.$, we do not want it to blow up. Naturally we should consider eigenvalue. Since left eigenvalues and right eigenvalues are the same(because eigenvalues of its transpose are the same)

Fact 1 Any stochastic matrix P has an eigenvalue 1.

Proof. the vector u with $u_1 = \cdots = u_n = 1$ is a right-eigenvector corresponding to eigenvalue 1 of P. □

Fact 2 Suppose we have eigenvalues $\lambda_0, \ldots, \lambda_{n-1}$ such that $\lambda_0 = 1$. Consider $\lambda_* := \max\{\lambda_1, \ldots, \lambda_{n-1}\}$. Then $\lambda_* \leq 1$. Furthermore, if $P(x, y) > 0$ for all x, y, we have $\lambda_* < 1$

Proof. Let $v(x)$ be the biggest entry. Then we have

$$|\lambda v(x)| = (Pv)_x = |\sum_{y=1}^{n} P(x, y)v(y)| \leq |v(x)|$$

So $\lambda \leq 1$.
Suppose $P(x, y) > 0$, the equality holds only if $v(x)$ is a constant. But in this case it is v_0. Therefore $\lambda_* < 1$ □

Fact 3 Suppose P satisfies $\lambda_* < 1$. Then there is a unique stationary distribution π on $\mathcal{X}$ and, given an initial distribution μ_0 and any point $x \in \mathcal{X}$, there is a constant $C_x > 0$ such that

$$|\mu_k(x) - \pi(x)| \leq C_x k^{J-1} (\lambda_*)^{k-J+1}$$

If P is diagonalizable, we have

$$|\mu_k(x) - \pi(x)| \leq \sum_{m=1}^{n-1} |a_m v_m(x)||\lambda_m|^k \leq (\sum_{m=1}^{n-1} |a_m v_m(x)|)(\lambda_*)^k$$

where initial distribution can be expressed as linear combination of corresponding eigenvectors

$$\mu_0 = a_0 v_0 + \cdots + a_{n-1} v_{n-1}$$

If the eigenvectors v_i are orthogonal in $L^2(\pi)$, i.e. if $\sum_x v_i(x)\overline{v_j(x)}\pi(x) = \delta_{ij}$, we get the further bound,

$$\sum_x |\mu_k(x) - \pi(x)|^2 \pi(x) = \sum_{m=1}^{n-1} |a_m|^2 |\lambda_m|^{2k} \leq (\sum_{m=1}^{n-1} |a_m|^2)(\lambda_*)^k$$

Interpretation: Now we easily get a new bound using eigenvalues. When $\lambda_* < 1$, $|\mu_k(x) - \pi(x)| \to 0$ as $k \to \infty$, the Markov Chain converges exponentially quickly. $\mu_k(x) \to a_0 v_0 = \pi(x)$. Since $\sum_x \pi(x) = 1$, $a_0 = (\sum_y v_0(y))^{-1}$. Indeed, the stationary distribution does not depend on the initial distribution μ_0

Meanwhile, $\lambda_* < 1$ means the eigenvalue 1 has only one multiplicity, so the Markov chain has a unique stationary distribution π as the corresponding eigenvector. Conversely, if $\lambda_* = 1$, the eigenvalue 1 has at least 2 multiplicity, and also assume P is diagonalizable, then each multiplicity has a corresponding eigenvector, contradicts the property of unique stationary distribution.

Fact 4 A finite Markov chain satisfies $\lambda_* < 1$ if and only if it is both indecomposable and aperiodic

Proof. Necessity: assume it is decomposable, with disjoint subspaces $\mathcal{X}_1$ and $\mathcal{X}_2$. Then P is like

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Then there are two vectors with eigenvalue 1

Assume it is periodic. Then the matrix is like

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

Let

$$v = (e^{2\pi i/3}, e^{4\pi i/3}, 1)$$

Then $e^{2\pi i/d}$ is an eigenvalue.(Intuitively, the matrix is a permutation, which moves x_i to x_{j+1}, so we can just make $\frac{v(x_{j+1})}{v(x_j)}$ to be a fixed number)(can not pick a random ratio because we want $v(x_1)/v(x_n) = r$) In this case $\lambda_* = 1$

Sufficiency: 2.4.4. If a chain is indecomposible and aperiodic, then for any state $i, j \in S$, there is a $n_0(i,j)$ such that $p_{ij}^n > 0$ for any $n > n_0$. Find the biggest n_0. Then P^n has all positive entries. □

Note: Irreducible is a stronger property than indecomposable.

5 Coupling and minorisation conditions

5.1 Coupling construction

Let X, Y be two random variables on some space $\mathcal{X}$. If we write $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ for their probability distributions, then

$$||\mathcal{L}(X) - \mathcal{L}(Y)|| \leq \sup_A |P(X \in A, X \neq Y) - P(Y \in A, X \neq Y)| \leq P(X \neq Y)$$

Consider a Markov chain P on $\mathcal{X}$ with (i) $X_0 \sim \mu_0$
(ii) $Y_0 \sim \pi$
(iii)$P(X_{k+1} \in A|X_k) = P(X_k, A)$
(iv)$P(Y_{k+1} \in A|Y_k) = P(Y_k, A)$
(v) There is a random time T (so called coupling time) such that $X_k = Y_k$ for all $k \geq T$ (i.e. X and Y are the same after time T)

We do not know the joint law except after some time T(the coupling time). We know

$$||\mu_k - \pi|| = ||\mathcal{L}(X_k) - \mathcal{L}(Y_k)|| \leq P(X_k \neq Y_k) \leq P(T > k)$$

So if we can find a coupling as above, we get an immediate bound on $||\mu_k - \pi||$ in terms of the tail probabilities of the coupling time T
For the last inequality, $X_k \neq Y_k \Rightarrow k \leq T$ and the converse is not true. This is because $X_k \neq Y_k$ is saying X and Y are not the same at time k, and therefore we haven't reached time T yet (by definition of T). In other words, the set $T \geq k$ is larger than the set $X_k \neq Y_k$
Remark: for an arbitrary chain, we define a coupling with initial state (μ_0, x_0) such that X_k and Y_k are marginally updated by P. Then it suffices to check when X_k converges from Y_k(The benefit is we can just consider the transition probability we constructed)

5.2 Uniform minorisation conditions

Suppose a Markov chain satisfies an inequality of the form

$$P^{k_0}(x, A) \geq \beta\xi(A)$$

where $x \in R$ is a subset, $\beta > 0$ and ξ is a probability distribution.
The inequality above is called a minorisation condition for a Markov chain. It says that the transition probabilities from a set R all have common overlap of at least size β. (Is this sentence saying that all of the transition probabilities from R should be greater than or equal to β?)

For later sections, we will only consider the uniform case where $R = \mathcal{X}$ (uniform because holds for all $x \in \mathcal{X}$) and we set $k_0 = 0$ for simplicity.

Now we'll use uniform minorisation condition to a valid coupling (X_k, Y_k) as follows: First, define $X_0 \sim \mu_0, Z_0 \sim \pi$ independently. Given X_k, Z_k, choose X_{k+1}, Z_{k+1} by flipping an independent coin that has probability β coming up heads:
a) If the coin is heads, we'll force $X_{k+1}, Z_{k+1} = z$, where $z \in \mathcal{X}$ distributed independently according to $\xi(\cdot)$
b) If the coin is tails, then we choose X_{k+1}, Z_{k+1} independently with

$$P(X_{k+1} \in A) = \frac{P(X_k, A) - \beta\xi(A)}{1 - \beta}$$

$$P(Z_{k+1} \in A) = \frac{P(Z_k, A) - \beta\xi(A)}{1 - \beta}$$

We call this residuals (for leftover probabilities). How come?

Recall the condition iii) and iv) above. We have to choose probability that precisely so that $P(X_{k+1} \in A|X_k) = P(X_k, A)$ (same for z)

$$\begin{aligned} P(X_{k+1} \in A|X_k) &= P(X_k, \text{head})P(X_{k+1} \in A|X_k, \text{head}) + P(X_k, \text{tail})P(X_{k+1} \in A|X_k, \text{tail}) \\ &= \beta\xi(A) + (1 - \beta)P(X_{k+1} \in A|X_k, \text{tail}) = P(X_k, A) \end{aligned}$$

To satisfy the equation above, we will get exactly the residual probability in b) for tail situation. Meanwhile, minorisation condition is important here, it ensures the valid probability distribution, $P(X_k, A) - \beta\xi(A) > 0$.

Finally, let T be the first time we toss up head, that is the coupling time that X and Z will be at same step. Define $Y_k = Z_k$ $(k \leq T)$, but after coupling time, $Y_k = X_k$ for $k > T$. Therefore, we shall consider (X_k, Y_k) as coupling with coupling time T. It is again important to think about the minorisation condition based on our assumptions, because of the inequality, we can successfully construct such coupling, and then get a neat result, that we call the Markov Chain is geometric ergodic.

Fact 5 Suppose a Markov chain satisfies $P(x, A) \geq \beta\xi(A)$ for all $x \in \mathcal{X}$ and for all measurable subsets $A \subseteq \mathcal{X}$. Then given any initial distribution μ_0 and stationary distribution π, we have

$$||\mu_k - \pi|| \leq (1 - \beta)^k$$

Note that as long as a Markov chain satisfies the above condition (uniform minorisation conditions), we can define the combined chain (X_k, Y_k) with coupling time T, then we see that $P(T > k) = (1 - \beta)^k$

We have $\beta\xi(x) = \beta(x_1, x_2, \ldots, x_n) \leq \{\min(P(x,1)), \ldots, \min P(x,n)\}$ by the minorisation condition. Sum the entries together, we get $\beta * 1 \leq \sum_{y\in\mathcal{X}} \min_{x\in\mathcal{X}} P(x,y)$ In the infinite case, we instead take integral. Then the largest value of beta will be $\sum_{y\in\mathcal{X}} \min_{x\in\mathcal{X}} P(x,y)$.

Example. We let

$$P(x, dy) = \frac{1-x-y}{\frac{3}{2}+x}dy = 1 + \frac{-\frac{1}{2}+y}{\frac{3}{2}+x}.$$

If $y < \frac{1}{2}$ then we have $\inf(P)$ when x is the smallest; if $y > \frac{1}{2}$ then we have $\inf(P)$ when x is the largest. Therefore we may take $\beta = \frac{29}{30}$, and by Fact 5, we conclude that $||\mu_k - \pi|| \leq (\frac{1}{30})^k$.

5.3 Compare to Markov Forgetting Lemma

Here is an interesting connection to Markov Forgetting Lemma(in discrete case):

If a Markov Chain is irreducible and aperiodic, and has stationary distribution π_i. then for all $i, j, k \in S$, $|P^n(i,k) - P^n(j,k)| \to 0$ as $n \to \infty$
So that after a long time n, it doesn't matter where the chain started from, they will eventually come at the same step, that is, converge to stationary distribution.

In the proof of Markov Forgetting lemma, we require coupling for two independent random variables X_k, Y_k, and their joint Markov chain (X_k, Y_k) where their joint transition probability is $\bar{P}_{(ik)(jl)} = P_{(ik)}P_{(jl)}$, joint stationary distribution $\bar{\pi}_{(ij)} = \pi_i\pi_j$, since they are independent. However, in our construction of coupling above, X_k and Y_k is not independent. They depends on the coin's situation.

5.4 Comparison between two methods

In Coupling, we apply this approach to Markov Chain that has stationary distribution π, the bounding condition is easier to compute. In section 4, eigenvalue approach can be applied to any Markov Chain, and after we found out the eigenvalues, we can verify if unique stationary distribution exists. But in both cases, even we don't know the exact stationary distribution, given ϵ, we can get a large enough k for the bound, compute μ_k to approximate π since they are "ϵ close" now.

6 Generalization of Minorisation to Subset of State Space

In Section 5, when we defined the minorisation condition as:

$$P^{k_0}(x, A) \geq \beta\xi(A) \quad x \in R,\ A \subseteq \mathcal{X}$$

for some $R \subseteq \mathcal{X}$, we decide to set R to be $\mathcal{X}$. Now, we are going to explore the general case when R can be any subset of $\mathcal{X}$. The bound we used before is not valid here anymore, and there are several approaches to solve this problem. We are able to bound the probability of escaping from R if R can be very large, or we can use "drift conditions". Also, we can update regeneration times from X_k and try to bound the "times since the last generation".

7 Appendix

7.1 Integral with respect to a probability measure

We first define such integral for simple functions, then construct a sequence of simple functions which converges to the original function, and take integrals of then plus use dominated convergence theorem
e.g.

$$\int f(x)d\mu(x) = \sum f_i \mu(x_i, x_{i+1})$$

For probability μ, this means $P(x \in (x_i, x_{i+1}))$