

行政院國家科學委員會專題研究計畫 期中進度報告

馬可夫鏈趨向穩定分佈時間之研究(1/2)

計畫類別：個別型計畫

計畫編號：NSC94-2115-M-009-010-

執行期間：94 年 08 月 01 日至 95 年 07 月 31 日

執行單位：國立交通大學應用數學系(所)

計畫主持人：許元春

報告類型：精簡報告

處理方式：本計畫可公開查詢

中 華 民 國 95 年 5 月 30 日

期中報告-關於 Mixing Time 的一些估計

許元春

在理論及應用上，Mixing Time 是馬可夫鏈收斂到穩定狀態的重要時間。著名的機率學者，例如：Aldous、Diaconis、Saloff-Coste 等，提出許多關於 Mixing Time 的研究與探討。計算科學家 Goldberg 與 Jerrum 在 90 年代，提出 Burnside Process，做為 Pólya 計數理論的一種重要機率數值方式。P. Diaconis 在 2005 的文章裡探討 Bose-Einstein Markov Chain 其 Mixing time 的估計。這是關於 Burnside processes mixing time 初步及重要的結果。我們的工作是探討一般 Burnside processes mixing time 的估計問題。

Mixing Times for Burnside Processes

Wei-Kuo Chen

Department of Applied Mathematics, National Chiao Tung University

Yuan-Chung Sheu

Department of Applied Mathematics, National Chiao Tung University

Abstract

Mixing time is the crucial time for a Markov chain converging to its equilibrium. Several tools have been developed to analyze this important quantity, such as analytic techniques in Hilbert space, coupling methodology. In the last decade, computer theorists Goldberg and Jerrum purposed a special Markov chain, called Burnside process which is an important probability model for Pólya's theory of counting. P. Diaconis 2005 discusses the mixing time of Bose-Einstein Markov chain. However, we still know little about Burnside process. Hence, in this article we want to discuss mixing times for general Burnside processes.

1 Preliminaries

A Markov chain on a finite state space $\mathcal{X}$ can be specified by a kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, which satisfies

$$K(x, y) \geq 0 \quad \text{for all } x, y \in \mathcal{X} \quad \text{and} \quad \sum_{y \in \mathcal{X}} K(x, y) = 1 \quad \text{for all } x \in \mathcal{X}.$$

By the Chapman-Kolmogorov equations, we define the iterated kernel

$$K^n(x, y) = \sum_{z \in \mathcal{X}} K(x, z) K^{n-1}(z, y).$$

A probability measure π define on $\mathcal{X}$ is *stationary* if $\sum_x \pi(x) K(x, y) = \pi(y)$ for all $y \in \mathcal{X}$. Throughout, we assume that the Markov chain is *irreducible*, that is, for every $x, y \in \mathcal{X}$, there exists some $n = n(x, y) > 0$ such that $K^n(x, y) > 0$. Under this assumption K has a unique stationary distribution π with $\min_x \pi(x) > 0$.

The Markov operator associated with K is defined by

$$Kf(x) = \sum_{y \in \mathcal{X}} K(x, y) f(y),$$

and similarly the iterated Markov operator is formulated as

$$K^n f(x) = \sum_{y \in \mathcal{X}} K^n(x, y) f(y).$$

The chain is *reversible* if K satisfies the *detail balance condition*

$$\pi(x) K(x, y) = \pi(y) K(y, x) \quad \text{for every } x, y.$$

A state $x \in \mathcal{X}$ is *aperiodic* if $K^n(x, x) > 0$ for all sufficiently large n , and the chain is aperiodic if every state is aperiodic. For irreducible chains, aperiodicity of a single state implies aperiodicity of every state.

Given a kernel K we can then have a discrete-time Markov chain. With the kernel K , we can also consider the continuous-time Markov chain identifying by the kernel H^t which waits an exponential time before moving. Precisely, let X_n be a given Markov chain. Suppose that N_t is the Poisson distribution with rate 1 and N_t is independent of $\{X_n\}$. Define $Y_t = X_{N_t}$, then

$$H^t(x, y) \equiv E [1_{\{Y_t=y\}} | Y_0 = x] = e^{-t} \sum_{n=0}^{\infty} \frac{t^n}{n!} K^n(x, y)$$

is called the transition kernel of the continuous time Markov chain associated with the kernel K . Consequently, the operator is defined by

$$H^t f(x) = e^{-t} \sum_{n=0}^{\infty} \frac{t^n}{n!} K^n f(x).$$

If we define $u(t, x) = H^t f(x)$, then it solves the heat equation

$$\begin{cases} (\partial_t + (I - K))u(t, x) &= 0 \text{ on } (0, \infty) \times \mathcal{X}, \\ u(0, x) &= f(x). \end{cases}$$

Proposition 1. *Let (K, π) be an irreducible Markov chain on a finite state space $\mathcal{X}$ with stationary distribution π . Then*

$$\lim_{n \rightarrow \infty} H^n(x, y) = \pi(y) \quad \text{for all } x, y.$$

Furthermore, if we assume that K is aperiodic, then

$$\lim_{n \rightarrow \infty} K^n(x, y) = \pi(y) \quad \text{for all } x, y.$$

With the weak conditions on K , we obtain that the Markov chain converges to its stationary distribution. However, this classic result does not tell us any information about the rate of convergence. Hence, the aim of this thesis is to present the general techniques for analyzing the behavior of convergence and several examples will be involved.

In order to quantify the distance between the chain and its stationary distribution, we firstly need to introduce a metric. The most natural and useful one is the total variation distance.

Definition 1. *The total variation distance between two probability distributions μ and ν on $\mathcal{X}$ is defined as*

$$\|\mu - \nu\|_{TV} = \max_{A \subset \mathcal{X}} |\mu(A) - \nu(A)|.$$

Proposition 2. *Let μ and ν be two probability distributions defined on $\mathcal{X}$, then*

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)|.$$

Proof: Let $\mathcal{S} = \{x \in \mathcal{X} : \mu(x) > \nu(x)\}$. For any $A \subset \mathcal{X}$,

$$\begin{aligned} \mu(A) - \nu(A) &= \mu(A \cap \mathcal{S}) - \nu(A \cap \mathcal{S}) - (\nu(A \cap \mathcal{S}^c) - \mu(A \cap \mathcal{S}^c)) \\ &= \sum_{x \in A \cap \mathcal{S}} (\mu(x) - \nu(x)) - \sum_{x \in A \cap \mathcal{S}^c} (\nu(x) - \mu(x)). \end{aligned}$$

Since $\mu - \nu$ is positive on $\mathcal{S}$ and $\nu - \mu$ is nonnegative on $\mathcal{S}^c$, it follows that

$$\mu(A) - \nu(A) \leq \mu(\mathcal{S}) - \nu(\mathcal{S}).$$

Similarly,

$$\nu(A) - \mu(A) \leq \nu(\mathcal{S}^c) - \mu(\mathcal{S}^c).$$

Since μ and ν are probability measures, $\mu(\mathcal{S}) - \nu(\mathcal{S}) = \nu(\mathcal{S}^c) - \mu(\mathcal{S}^c)$. Thus,

$$\begin{aligned} \|\mu - \nu\|_{TV} &= \frac{1}{2} [\mu(\mathcal{S}) - \nu(\mathcal{S}) + \nu(\mathcal{S}^c) - \mu(\mathcal{S}^c)] \\ &= \frac{1}{2} \left[\sum_{x \in \mathcal{S}} (\mu(x) - \nu(x)) + \sum_{x \in \mathcal{S}^c} (\nu(x) - \mu(x)) \right] = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)|. \end{aligned}$$

This completes our proof. □

The techniques in Hilbert space will be very useful in our analysis. Hence, we define the L_p distance.

Definition 2. *For any two measures μ and ν with densities $f(x) = \mu(x)/\pi(x)$ and $g(x) = \nu(x)/\pi(x)$ with respect to the positive measure π , their $L^p(\pi)$ distance is*

$$d_{\pi, p}(\mu, \nu) = \|f - g\|_{L^p(\pi)} \quad \text{for } 1 \leq p \leq \infty.$$

From proposition 2, we could see that $d_{\pi,1}(\mu, \pi) = 2\|\mu - \pi\|_{TV}$. By Jensen's inequality, the function $p \mapsto d_{\pi,p}(\mu, \nu)$ is non-decreasing.

Next, we want to define the *mixing time* which measures the time that the chain needs to go to its equilibrium.

Definition 3. *The total variation mixing time is given by*

$$T(K) = \inf \left\{ n > 0 : \sup_x \|K^n(x, \cdot) - \pi(\cdot)\|_{TV} \leq 1/e \right\}.$$

Definition 4. *The L^p mixing time is given by*

$$T_p(K) = \inf \left\{ n > 0 : \sup_x d_{\pi,p}(K^n(x, \cdot), \pi) \leq 1/e \right\}.$$

Analogous definitions can extend to the continuous time Markov chain H_t and we use the notations T and T_p to denote the mixing times of total variance distance and the L_p distance, respectively. We should note that because of the following proposition, the constant $1/e$ in the definition is chosen for convention.

Proposition 3. *For $1 \leq p \leq \infty$, then $n \mapsto \sup_{x \in \mathcal{X}} d_{\pi,p}(K^n(x, \cdot), \pi)$ is a nonincreasing sub-additive function. In particular, if*

$$\sup_{x \in \mathcal{X}} d_{\pi,p}(K^m(x, \cdot), \pi) \leq \beta$$

for some fixed integer m and some $\beta \in (0, 1)$ then

$$\sup_{x \in \mathcal{X}} d_{\pi,p}(K^n(x, \cdot), \pi) \leq \beta^{\lfloor n/m \rfloor}.$$

Proof: Use Jensen inequality, we have

$$\begin{aligned} d_{\pi,p}(K^{n+1}(x, \cdot), \pi)^p &= \sum_{y \in \mathcal{X}} \left| \sum_{z \in \mathcal{X}} K(x, z) \left(\frac{K^n(z, y)}{\pi(y)} - 1 \right) \right|^p \pi(y) \\ &\leq \sum_{y \in \mathcal{X}} \sum_{z \in \mathcal{X}} K(x, z) \left| \frac{K^n(z, y)}{\pi(y)} - 1 \right|^p \pi(y) = \sum_{z \in \mathcal{X}} d_{\pi,p}(K^n(z, \cdot), \pi)^p K(x, z) \end{aligned}$$

and

$$\begin{aligned} d_{\pi,p}(K^{n+m}(x, \cdot), \pi)^p &= \sum_{y \in \mathcal{X}} \left| \sum_{z \in \mathcal{X}} \left(\frac{K^n(x, z)}{\pi(z)} - 1 \right) \left(\frac{K^m(z, y)}{\pi(y)} - 1 \right) \pi(z) \right|^p \pi(y) \\ &\leq \sum_{z \in \mathcal{X}} \left| \frac{K^n(x, z)}{\pi(z)} - 1 \right|^p d_{\pi,p}(K^m(z, \cdot), \pi)^p \pi(z). \end{aligned}$$

Now, taking the supremum, we prove that $d_{\pi,p}(K^n(x, \cdot), \pi)^p$ is nonincreasing and subadditive. Next, suppose that n is any nonnegative integer then there exists some k such that $km \leq n < (k+1)m$. So

$$\sup_{x \in \mathcal{X}} d_{\pi,p}(K^n(x, \cdot), \pi) \leq \sup_{x \in \mathcal{X}} d_{\pi,p}(K^{km}(x, \cdot), \pi) \leq \sup_{x \in \mathcal{X}} d_{\pi,p}(K^m(x, \cdot), \pi)^k \leq \beta^k = \beta^{\lfloor n/m \rfloor}.$$

Hence, we prove our assertion. $\square$

Assume that K is an irreducible and aperiodic Markov chain on a finite state space $\mathcal{X}$. We say that K satisfies the so-called *Doebelin* condition if there exists an integer n_0 and a positive constant c such that

$$K^{n_0}(x, y) \geq c\pi(y) \quad \text{for all } x, y.$$

The first result of the rate of convergence involving with Doebelin condition is stated in the following.

Theorem 1. *If (K, π) is an irreducible and aperiodic Markov chain, then*

$$\sup_{x \in \mathcal{X}} \|K^n(x, \cdot) - \pi(\cdot)\|_{TV} \leq (1 - c)^{\lfloor \frac{n}{n_0} \rfloor}$$

where n_0 is a nonnegative integer such that $K^{n_0}(x, y) > 0$ for every x, y and $c = \min_{x, y} \left\{ \frac{K^{n_0}(x, y)}{\pi(y)} \right\}$.

Proof: From the aperiodicity, we have that $K^{n_0}(x, y) > 0$ for every x, y and some n_0 . Define the matrix M with each row equal to the stationary distribution π . Then it is clear that $MK = KM = M$. Take $c = \min_{x, y \in \mathcal{X}} \left\{ \frac{K^{n_0}(x, y)}{M(x, y)} \right\} > 0$ and define

$$N = \begin{cases} \frac{K^{n_0} - cM}{1 - c}, & \text{if } 0 < c < 1, \\ 0, & \text{if } c = 1. \end{cases}$$

If $c = 1$, then $M(x, y) \leq K^{n_0}(x, y)$ for each x, y . If $M(x, y) < K^{n_0}(x, y)$ for some x, y , then $1 = \sum_y M(x, y) < \sum_y K^{n_0}(x, y) = 1$, this is a contradiction. Hence, $M = K^{n_0}$ and we are done.

If $0 < c < 1$, it is easy to check that N is a Markov kernel and $MN = NM = M$ such that

$$\begin{aligned} (N - M)^n &= \sum_{l=0}^n (-1)^l \binom{n}{l} N^{n-l} M^l = \sum_{l=1}^n (-1)^l \binom{n}{l} N^{n-l} M^l + N^n \\ &= \sum_{l=1}^n (-1)^l \binom{n}{l} M + N^n = (1 - 1)^n - M + N^n = (N - M)N^{n-1} \end{aligned}$$

and

$$K^{n_0 n} - M = \sum_{l=1}^n \binom{n}{l} K^{n_0 l} M^{n-n_0 l} = (K^{n_0} - M)^n = (1 - c)^n (N - M)^n.$$

Now, if we set a matrix norm $\|\cdot\|$ by $\|A\| = \max_x \sum_y |A(x, y)|$ for any matrix A and recall that $\|AB\| \leq \|A\| \|B\|$, then

$$\|K^{n_0 n} - M\| \leq (1 - c)^n \|N - M\| \|N\|^{n-1} = (1 - c)^n \|N - M\| \leq 2(1 - c)^n.$$

Finally, for every $m \geq 1$, there exists some $n \geq 0$ such that $n_0 n \leq m < n_0(n + 1)$, and it follows that

$$\|K^m - M\| = \|K^{m-n_0 n} (K^{n_0 n} - M)\| \leq \|K^{n_0 n} - M\| \leq 2(1 - c)^n = 2(1 - c)^{\lfloor \frac{m}{n_0} \rfloor}.$$

So we prove the announced result. □

We can get an upper bound of the total variance distance via the constant c in theorem 1. Usually, it is easy to apply theorem 1. Nonetheless, the estimation is always not sharp enough. From the elementary result in the matrix analysis, we can express our Markov kernel by using its eigenvalues and eigenfunctions. It turns out the bounds for eigenvalues is related with our convergent rate.

Lemma 1. Consider a reversible Markov kernel K with stationary distribution π . Let $\{\lambda_i : i = 0, \dots, |\mathcal{X}| - 1\}$ be the set of all eigenvalues of K and $\{\phi_i : i = 0, \dots, |\mathcal{X}| - 1\}$ be the corresponding orthonormal basis and eigenfunctions in $L^2(\pi)$. Without loss of generality, we assume that $\lambda_0 = 1$ and $\phi_0 = 1$. Then the eigenvalues of K are all real and

$$\frac{K^n(x, y)}{\pi(y)} = \sum_{i=0}^{|\mathcal{S}|-1} \lambda_i^n \psi_i(x) \overline{\psi_i(y)}, \quad (1)$$

$$d_{\pi,2}(K^n(x, \cdot), \pi)^2 = \sum_{y \in \mathcal{S}} \left| \frac{K^n(x, y)}{\pi(y)} - 1 \right|^2 \pi(y) = \sum_{i=1}^{|\mathcal{S}|-1} |\lambda_i|^{2n} |\psi_i(x)|^2, \quad (2)$$

and

$$\frac{H^t(x, y)}{\pi(y)} = \sum_{i=0}^{|\mathcal{S}|-1} e^{-t(1-\lambda_i)} \psi_i(x) \overline{\psi_i(y)}, \quad (3)$$

$$d_{\pi,2}(H^t(x, \cdot), \pi)^2 = \sum_{y \in \mathcal{S}} \left| \frac{H^t(x, y)}{\pi(y)} - 1 \right|^2 \pi(y) = \sum_{i=1}^{|\mathcal{S}|-1} e^{-2t \operatorname{Re}(1-\lambda_i)} |\psi_i(x)|^2. \quad (4)$$

2 Spectral gap

Now, we introduce the L^2 techniques to bound eigenvalues.

Definition 5. For a Markov chain (K, π) on $\mathcal{X}$. The Dirichlet form is given by

$$\mathcal{E}(f, g) = \langle (I - K)f, g \rangle_\pi$$

where $\langle f, g \rangle_\pi = \sum_{x \in \mathcal{X}} f(x) \overline{g(x)} \pi(x)$.

Proposition 4. A simple calculation yields that

(1)

$$\mathcal{E}(f, f) = \left\langle \left(I - \frac{1}{2}(K + K^*) \right) f, f \right\rangle = \frac{1}{2} \sum_{x, y \in \mathcal{X}} |f(x) - f(y)|^2 K(x, y) \pi(x), \quad (5)$$

(2) As a function of t , the derivative of $\|H^t\|_2^2$ is given by

$$\frac{d}{dt} \|H^t f\|_2^2 = -2\mathcal{E}(H^t f, H^t f) \quad (6)$$

where $\|f\|_p = [\sum_{y \in \mathcal{S}} |f(y)|^p \pi(y)]^{\frac{1}{p}}$.

The spectral gap is defined in terms of Dirichlet form.

Definition 6. The spectral gap λ associated with K is given by the formula

$$\lambda = \lambda(K) = \inf \left\{ \frac{\mathcal{E}(f, f)}{\operatorname{Var}(f)} : \operatorname{Var}(f) \neq 0 \right\}.$$

Here $\operatorname{Var}(f) = \sum_{x \in \mathcal{X}} |x - E_\pi[f]|^2 \pi(x)$ and $E_\pi[f] = \sum_{x \in \mathcal{X}} f(x) \pi(x)$.

We emphasize some fact on the spectral gaps. Recall that the adjoint operator K^* of K defined by $K^*(x, y) = \frac{\pi(y)K(y, x)}{\pi(x)}$ has the same stationary distribution as K . Hence, from the definition of spectral gap, it follows that $\lambda(K) = \lambda(K^*)$. More important,

Proposition 5. *If K is reversible, then λ is the smallest non-zero eigenvalue of $I - K$. In general, λ is the smallest non-zero eigenvalue of $I - \frac{1}{2}(K + K^*)$.*

Proof: We only prove for K being reversible. Assume that K is reversible and let λ_i be the eigenvalues of K with

$$-1 \leq \lambda_{|\mathcal{X}|-1} \leq \dots \leq \lambda_1 < \lambda_0 = 1,$$

and ϕ_i be the corresponding orthonormal basis and eigenfunctions of λ_i in $L^2(\pi)$. For any $f \in L^2(\pi)$, $\text{Var}_\pi(f) \neq 0$, f can be expressed as $f(x) = \sum_{i=0}^{|\mathcal{S}|-1} c_i \psi_i(x)$ where $c_0, \dots, c_{|\mathcal{S}|-1} \in \mathbb{C}$. So

$$\frac{\mathcal{E}(f, f)}{\text{Var}_\pi(f)} = \frac{\mathcal{E}(f - c_0, f - c_0)}{\text{Var}_\pi(f - c_0)} = \frac{\sum_{i=1}^{|\mathcal{X}|-1} |c_i|^2 (1 - \lambda_i)}{\sum_{i=1}^{|\mathcal{X}|-1} |c_i|^2 - \|f - c_0\|_1^2} \geq \frac{\sum_{i=1}^{|\mathcal{X}|-1} |c_i|^2 (1 - \lambda_1)}{\sum_{i=1}^{|\mathcal{X}|-1} |c_i|^2} = 1 - \lambda_1.$$

Hence, $\lambda \geq 1 - \lambda_1$. Now, let W be the subspace generated by $\{\phi_0, \phi_1\}$ and for any element $f \in W$, $\|f\|_2 = 1$ and $\pi(f) = 0$. Again, any $f \in W$ can be expressed as $f(x) = c_0 \psi_0(x) + c_1 \psi_1(x)$. Therefore,

$$\lambda \leq \inf \{\mathcal{E}(\phi, \phi) : \phi \in W\} \leq \mathcal{E}(f, f) = \sum_{i=0}^1 |c_i|^2 (1 - \lambda_i) = |c_1|^2 (1 - \lambda_1) \leq 1 - \lambda_1.$$

Thus, $\lambda = 1 - \lambda_1$ and this completes our proof. $\square$

The following theorem establishes the relationship between the rate of convergence and the spectral gap.

Theorem 2. *For every vector f in $\mathbb{C}^{|\mathcal{S}|}$, we have*

$$\|H^t f - \pi(f)\|_2^2 \leq e^{-2\lambda t} \text{Var}(f).$$

Proof: Let $u(t) = \|H^t f - \pi(f)\|_2^2 = \|H^t(f - \pi(f))\|_2^2$. Hence we have, by (6),

$$\begin{aligned} u'(t) &= -2\mathcal{E}(H^t(f - \pi(f)), H^t(f - \pi(f))) \\ &\leq -2\lambda \text{Var}(H^t(f - \pi(f))) = -2\lambda u(t). \end{aligned}$$

This implies that $u(t) \leq e^{-2\lambda t} u(0)$. Hence, we are done. $\square$

Proposition 6. *Let K be a Markov kernel with the spectral gap λ . Then we have*

$$d_{\pi,2}(H^t(x, \cdot), \pi) = \|h^t(x, \cdot) - 1\|_2 \leq \frac{e^{-\lambda t}}{\sqrt{\pi(x)}} \quad (7)$$

and

$$|H^t(x, y) - \pi(y)| \leq \sqrt{\frac{\pi(y)}{\pi(x)}} e^{-\lambda t}. \quad (8)$$

Proof: Due to the fact $(K^*)^n(x, y) = \frac{\pi(y)K^n(y, x)}{\pi(x)}$, the associated semigroup $\tilde{H}^t = e^{-t(I-K^*)}$ has kernel

$$\tilde{H}^t(x, y) = \frac{\pi(y)H^t(y, x)}{\pi(x)} = (H^t)^*(x, y).$$

Consider the densities of $H^t(x, y)$ and $\tilde{H}^t(x, y)$ with respect to π ,

$$h^t(x, y) = \frac{H^t(x, y)}{\pi(y)} \quad \text{and} \quad \tilde{h}^t(x, y) = \frac{\tilde{H}^t(x, y)}{\pi(y)} = h^t(y, x).$$

Note that $h^t(x, y) = \frac{\tilde{H}^t(y, x)}{\pi(x)} = \tilde{H}^t f(y)$, where $f(z) = \frac{1}{\pi(x)}\delta_x(z)$. Hence, use the fact $\lambda(K) = \lambda(K^*)$,

$$\|h^t(x, \cdot) - 1\|_2 = \|\tilde{H}^t f - \pi(f)\|_2 \leq e^{-\lambda t} \sqrt{\text{Var}(f)} = e^{-\lambda t} \sqrt{\frac{1 - \pi(x)}{\pi(x)}} \leq \frac{e^{-\lambda t}}{\sqrt{\pi(x)}}.$$

Moreover, we have

$$\begin{aligned} |h^t(x, y) - 1| &= \left| \sum_z \left(h^{t/2}(x, z) - 1 \right) \left(h^{t/2}(z, y) - 1 \right) \pi(z) \right| \\ &\leq \|h^{t/2}(x, \cdot) - 1\|_2 \|h^{t/2}(y, \cdot) - 1\|_2 \leq \frac{e^{-\lambda t}}{\sqrt{\pi(x)\pi(y)}}. \end{aligned}$$

Hence, we finish our proof. $\square$

Proposition 6 is one of the fundamental results of finite Markov chain theory. In particular, it tells us

Theorem 3. *Let K be an irreducible Markov chain and*

$$\omega = \min \{ \text{Re}(\xi) : \xi \neq 0 \text{ is an eigenvalue of } I - K \}.$$

Then

$$\forall 1 \leq p \leq \infty, \lim_{t \rightarrow \infty} \frac{-1}{t} \log \left(\max_{x \in \mathcal{S}} d_{\pi, p}(H^t(x, \cdot), \pi) \right) = \omega. \quad (9)$$

For $1 \leq p \leq 2$, we have

$$\frac{1}{\omega} \leq T_p \leq \frac{1}{2\lambda} \left(2 + \log \frac{1}{\pi_*} \right), \quad (10)$$

whereas, for $2 < p \leq \infty$,

$$\frac{1}{\omega} \leq T_p \leq \frac{1}{\lambda} \left(1 + \log \frac{1}{\pi_*} \right), \quad (11)$$

where $\pi_* = \min_{x \in \mathcal{X}} \pi(x)$.

Proof: We may prove (10) and (11) easily by using inequalities (7) and (8) and Jensen inequality. To prove (9), we need a fact in real analysis. Suppose that $1 \leq p, q \leq \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$, then

$$\|H^t - \Pi\|_{q \rightarrow \infty} = \max_{x \in \mathcal{X}} d_{\pi, p}(H^t(x, \cdot), \pi),$$

where the matrix $H^t - \Pi$ with $\Pi(x, y) = \pi(y)$ is regarded as a linear operator from $L^q(\pi)$ to $L^\infty(\pi)$ and $\|H^t - \Pi\|_{q \rightarrow \infty} \equiv \sup \{ \|(H^t - \Pi)f\|_\infty : \|f\|_q \leq 1 \}$ is the corresponding operator norm. Hence the left hand side of (10) and (11) are obtained by noting that the spectral radius of operator $H^t - \Pi$ is $e^{-\omega t}$ and the fact

$$\|H^t - \Pi\|_{q \rightarrow \infty} \geq \|H^t - \Pi\|_{q \rightarrow q} \geq \frac{\|(H^t - \Pi)\nu\|_{L^q(\pi)}}{\|\nu\|_{L^q(\pi)}} = e^{-\omega t} \frac{\|\nu\|_{L^q(\pi)}}{\|\nu\|_{L^q(\pi)}} = e^{-\omega t},$$

where ν is the eigenvector associated to $e^{-t\omega}$.

Now, we turn to prove (9). Note first that the limit can be rewritten as

$$\lim_{t \rightarrow \infty} \frac{-1}{t} \log (\|H^t - \Pi\|_{q \rightarrow \infty}) = \omega,$$

where $p^{-1} + q^{-1} = 1$. Again the previous fact, $\|H^t - \Pi\|_{q \rightarrow \infty} \geq e^{-\omega t}$, shows that

$$\limsup_{t \rightarrow \infty} \frac{-1}{t} \log (\|H^t - \Pi\|_{q \rightarrow \infty}) \leq \omega.$$

It remains to show that

$$\liminf_{t \rightarrow \infty} \frac{-1}{t} \log (\|H^t - \Pi\|_{q \rightarrow \infty}) \geq \omega.$$

We recall the following fact in matrix analysis: for any $\epsilon > 0$ and any matrix A , there exists a submultiplicative matrix norm $\|\cdot\|$ such that $\|A\| \leq \rho(A) + \epsilon$, where $\rho(A) = \max\{|\xi| : \xi \text{ is an eigenvalue of } A\}$. Let $A = H^1 - \Pi$, then we have

$$\|H^1 - \Pi\| \leq \rho(H^1 - \Pi) + \epsilon.$$

Then

$$\|H^t - \Pi\| \leq \|H^{\lfloor t \rfloor} - \Pi\| \|H^{t - \lfloor t \rfloor}\| \leq C \|(H - \Pi)^{\lfloor t \rfloor}\| \leq C (\rho(H - \Pi) + \epsilon)^{\lfloor t \rfloor},$$

where $C = \max_{s \in [0, 1]} \|H^s\|$. Note C indeed exists because $\|H^t\|$ is a continuous function, this can be proven by using the equivalence of matrix norm.

Therefore,

$$\liminf_{t \rightarrow \infty} \frac{-1}{t} \log \|H^t - \Pi\| \geq \liminf_{t \rightarrow \infty} \frac{-\log C}{t} + \frac{-\lfloor t \rfloor}{t} \log (e^{-\omega} + \epsilon) \geq -\log(e^{-\omega} + \epsilon).$$

Since any two matrix norm are equivalent, the above limit also holds for the operator norm $\|\cdot\|_{q \rightarrow \infty}$. At last let $\epsilon \rightarrow 0$, we get (9) and we finish our proof. $\square$

Remark 1. In general, $\lambda \leq \omega$. Furthermore, if K is reversible, then the equality holds.

3 Coupling

In this section, we want to develop the elementary idea and result of coupling method to bound our total variance distance. Suppose that $(X_n)_{n \geq 0}$ be a discrete time Markov chain defined on a finite state space $\mathcal{X}$. Let K and π be its kernel and stationary distribution, respectively.

Proposition 7. $\|K^n(x, \cdot) - \pi\|_{TV} \leq \max_{y \in \mathcal{X}} \|K^n(x, \cdot) - K^n(y, \cdot)\|_{TV}$

Proof: Since π is stationary, $\pi(x) = \sum_{y \in \mathcal{X}} \pi(y) K^n(y, x)$ for every $x \in \mathcal{X}$ and $n \in \mathbb{N}$,

$$\begin{aligned} \|K^n(x, \cdot) - \pi\|_{TV} &= \max_A \left| \sum_z \pi(z) (K^n(x, A) - K^n(z, A)) \right| \\ &\leq \sum_{z \in \mathcal{X}} \pi(z) \max_{y \in \mathcal{X}} \|K^n(x, \cdot) - K^n(y, \cdot)\|_{TV} = \max_{y \in \mathcal{X}} \|K^n(x, \cdot) - K^n(y, \cdot)\|_{TV}. \end{aligned}$$

□

For any $x, y \in \mathcal{X}$, a *coupling* is a joint process $(X_{1,n}^x, X_{2,n}^y)_{n \geq 0}$ such that

$$P(X_n \in \cdot | X_0 = x) = \mathbf{P}(X_{1,n}^x \in \cdot) \quad \text{and} \quad P(X_n \in \cdot | X_0 = y) = \mathbf{P}(X_{2,n}^y \in \cdot)$$

for every $n \geq 0$. If there exists a stopping time $T_{x,y} \leq \infty$, such that

$$X_{1,n}^x = X_{2,n}^y, \quad \text{for } T_{x,y} \leq n < \infty,$$

and we call that $T_{x,y}$ is a coupling time.

Proposition 8.

$$\|P(X_n \in \cdot | X_0 = x) - P(X_n \in \cdot | X_0 = y)\|_{TV} \leq \mathbf{P}(T_{x,y} \geq n + 1)$$

Proof: For every A ,

$$\begin{aligned} &P(X_n \in A | X_0 = x) - P(X_n \in A | X_0 = y) \\ &= \mathbf{P}(X_{1,n}^x \in A) - \mathbf{P}(X_{2,n}^y \in A) \\ &= \mathbf{P}(X_{1,n}^x \in A, X_{1,n}^x \neq X_{2,n}^y) - \mathbf{P}(X_{2,n}^y \in A, X_{1,n}^x \neq X_{2,n}^y) \\ &\leq \mathbf{P}(X_{1,n}^x \in A, X_{1,n}^x \neq X_{2,n}^y) \leq \mathbf{P}(T_{x,y} \geq n + 1). \end{aligned}$$

Now, similarly,

$$P(X_n \in A | X_0 = y) - P(X_n \in A | X_0 = x) \leq \mathbf{P}(T_{x,y} \geq n + 1).$$

Thus, by the definition of total variance distance, we are done.

□

In applying coupling methodology there are two issues.

- (i) First we need to specify the coupling.
- (ii) Second we need to analyze the coupling time.

The most common strategy for constructing coupling is via Markov couplings. We describe this procedure as follows. Suppose that the underlying chain has state space $\mathcal{X}$, and kernel K . Let $\tilde{K}$ be a new Markov chain on $\mathcal{X} \times \mathcal{X}$ satisfying the following conditions. For each pair (x, y) with $x \neq y$,

$$\tilde{K}((x, y), (x', y')) \quad \text{has marginals } K(x, \cdot) \text{ and } K(y, \cdot),$$

in other words $\sum_{y'} \tilde{K}((x, y), (x', y')) = K(x, x')$ and $\sum_{x'} \tilde{K}((x, y), (x', y')) = K(y, y')$. And suppose that

$$\tilde{K}((x, x), (x', y')) = \begin{cases} K(x, x'), & \text{if } x' = y', \\ 0, & \text{if } x' \neq y'. \end{cases}$$

Let $(X_{1,n}^x, X_{2,n}^y)$ be the Markov chain on $\mathcal{X} \times \mathcal{X}$ associated with Markov kernel $\tilde{K}$ and initial position (x, y) . Then this defines a coupling on K . We can see that if $X_{1,n}^x = X_{2,n}^y = z$, then $X_{1,m}^x = X_{2,m}^y$ for every $m \geq n+1$. Indeed, for every $m \geq n+1$,

$$\mathbf{P} \left((X_{1,m}^x, X_{2,m}^y) = (z', z') | (X_{1,n}^x, X_{2,n}^y) = (z, z) \right) = \tilde{K}^{m-n}((z, z), (z', z')) = K^{m-n}(z, z').$$

Therefore, $T_{x,y} \equiv \min \left\{ n \geq 0 : X_{1,n}^x = X_{2,n}^y \right\}$ is a coupling time. This construction gives a *natural Markov coupling*.

4 Mixing Times for Burnside Processes

In this section, we want to introduce a special Markov chain, called *Burnsider process*, purposed by computer theorists. Some convergent behaviors will be given by using the techniques we established above. Firstly, we give a background of this chain.

Assume that we have a cube and want to paint its faces with two colors red and blue. A natural question to ask is how many distinguishable configurations are? In this case, we can easily check out the answer. However, how about a n regular polyhedron painted with k different colors for arbitrary n and k ? It seems very hard to figure out, even for $k = 2$ and any n . Let G be a permutation group, the *cycle index polynomial* is defined by

$$P_G(z_1, \dots, z_{|G|}) = \frac{1}{|G|} \sum_{g \in G} z_1^{c_1(g)} \dots z_{|G|}^{c_{|G|}(g)},$$

where c_i means the number of i -cycles. Pólya theory of enumeration says that the number of distinguishable configurations is equal to

$$P_G(k, \dots, k),$$

where G is the permutation group induced by the rigid motion of the given n regular polyhedron. Nevertheless, Goldberg [4], a computer theorist, proved that to calculate the cycle index polynomial is intractable even when G is an abelian 2-group.

Instead Goldberg and Jerrum, [5] purposed a Markov chain, called *Burnside process*, to sample an orbit uniform at random. Let $\mathcal{X}$ be a finite state space and G be a finite permutation group. Assume that G is a group action on $\mathcal{X}$, that is, there exists a map $(\cdot, \cdot) : \mathcal{X} \times G \rightarrow \mathcal{X}$ by

- (1) $(x, e) = x$ for every $x \in \mathcal{X}$ and e the identity of G ;
- (2) $((x, g), h) = (x, gh)$ for every $g, h \in G$.

Throughout, for simplicity we denote $(x, g) = x^g$. This action splits $\mathcal{X}$ into disjoint orbits,

$$\mathcal{X} = \mathcal{O}_{x_1} \cup \mathcal{O}_{x_2} \cup \dots \cup \mathcal{O}_{x_z},$$

where $\mathcal{O}_x \equiv \{x^g : g \in G\}$ and $\mathcal{O}_{x_i} \cap \mathcal{O}_{x_j} = \emptyset$ for every $i \neq j$. Here, the number is given by *Burnside lemma*,

$$z = \frac{1}{|G|} \sum_{g \in G} |\mathcal{X}_g|,$$

where $\mathcal{X}_g \equiv \{z \in \mathcal{X} : z^g = z\}$ the fix-point set of g . Now, Burnside process is described by the following. Let x be the underlying state, we choose $g \in G$ with

$x^g = x$ uniformly at random and then for this g we choose y with $y^g = y$ uniformly at random. Thus, the chain moves from x to y . Therefore, the transition matrix and its stationary distribution is formulated by

$$K(x, y) = \frac{1}{|G_x|} \sum_{g \in G_x \cap G_y} \frac{1}{|\mathcal{X}_g|} \quad \text{and} \quad \pi(x) = \frac{1}{z|\mathcal{O}_x|}, \quad (12)$$

where $G_x \equiv \{g \in G : x^g = x\}$ is called the stabilizer of x . This was the genesis of Burnside process. Throughout, we assume further

$$K(x, y) = K(x^g, y^g) \quad \text{for each } x, y \in \mathcal{X} \text{ and } g \in G.$$

Iteratively, it implies that

$$K^n(x, y) = K^n(x^g, y^g) \quad \text{for each } x, y \in \mathcal{X} \text{ and } g \in G.$$

Let $\overline{\mathcal{X}}$ be the collection of G -orbits of $\mathcal{X}$. Then we have a corresponding lumped Markov chain $\overline{K}$ defined on $\overline{\mathcal{X}} \times \overline{\mathcal{X}}$ by setting

$$\overline{K}(\mathcal{O}_x, \mathcal{O}_y) = \sum_{z \in \mathcal{O}_y} K(x, z)$$

where $\mathcal{O}_x$ stands for the the orbits containing the state x . This definition is unambiguous, since for every $x_1, x_2 \in \mathcal{O}_x$, $K(x_1, y) = K(x_1^h, y^h) = K(x_2, y^h)$ for some h , and $\mathcal{O}_y = \mathcal{O}_{y^h}$. It is clear that $\overline{K}$ is a reversible Markov chain with uniform stationary distribution $\overline{\pi} = 1/z$. For simplicity, write $K_x^n(y) = K^n(x, y)$ and $\overline{K}_{\mathcal{O}_x}^n(\mathcal{O}) = K^n(\mathcal{O}_x, \mathcal{O})$ and we arrange the distinct orbits in order $\overline{\mathcal{X}} = \{\mathcal{O}_i\}_{i=1}^z$.

Lemma 2. *In general, $\overline{K}_{\mathcal{O}_x}^n(\mathcal{O}) = \sum_{z \in \mathcal{O}} K_x^n(z)$. Furthermore, if $x \in \mathcal{X}$ such that $K(x, z) = K(x, z^g)$ for every $g \in G$ and $z \in \mathcal{X}$, then*

$$\overline{K}_{\mathcal{O}_x}^n(\mathcal{O}) = |\mathcal{O}|K_x^n(y)$$

some $y \in \mathcal{O}$ and $n > 0$.

Proof: Firstly, we prove the first equation. Clearly if $n = 1$, it is just the definition of $\overline{K}_{\mathcal{O}_x}$. Assume that $\overline{K}_{\mathcal{O}_x}^n(\mathcal{O}) = \sum_{y \in \mathcal{O}} K_x^n(y)$ some $n \geq 1$, then

$$\begin{aligned} \overline{K}_{\mathcal{O}_x}^{n+1}(\mathcal{O}) &= \sum_{i=1}^z \overline{K}_{\mathcal{O}_x}^n(\mathcal{O}_i) \overline{K}(\mathcal{O}_i, \mathcal{O}) = \sum_{i=1}^z \left[\sum_{y \in \mathcal{O}_i} K_x^n(y) \right] \left[\sum_{z \in \mathcal{O}} K(y, z) \right] \\ &= \sum_{z \in \mathcal{O}} \left[\sum_{i=1}^z \sum_{y \in \mathcal{O}_i} K_x^n(y) K(y, z) \right] = \sum_{z \in \mathcal{O}} \sum_{y \in \mathcal{X}} K_x^n(y) K(y, z) = \sum_{z \in \mathcal{O}} K_x^{n+1}(z). \end{aligned}$$

By induction we prove the first equation. Now, assume that $\mathcal{O} = \{y^g : g \in G\}$. Since $K^n(x, y) = K^n(x, y^g)$ for each g , it follows that

$$K_{\mathcal{O}_x}^n(\mathcal{O}) = \sum_{z \in \mathcal{O}} K^n(x, z) = |\mathcal{O}|K^n(x, y).$$

So we are done. $\square$

Proposition 9. *Suppose that $x \in \mathcal{X}$ such that $K(x, z) = K(x, z^g)$ for every $g \in G$ and $z \in \mathcal{X}$. For any $n \in \mathbb{N}$,*

$$\|K_x^n - \pi\|_{TV} = \|\overline{K}_{\mathcal{O}_x}^n - \overline{\pi}\|_{TV}. \quad (13)$$

Proof. From lemma 2. and the fact that $\pi(\cdot)$ is constant on every orbit, this proposition is easy to check. Indeed,

$$\begin{aligned}\|\bar{K}_{\mathcal{O}_x}^n - \bar{\pi}\|_{TV} &= \frac{1}{2} \sum_{j=1}^z |\bar{K}_{\mathcal{O}_x}^n(\mathcal{O}_j) - \bar{\pi}(\mathcal{O}_j)| = \frac{1}{2} \sum_{j=1}^z \left| |\mathcal{O}_j| K^n(x, y_j) - |\mathcal{O}_j| \pi(y_j) \right| \\ &= \frac{1}{2} \sum_{j=1}^z \sum_{y \in \mathcal{O}_j} |K^n(x, y) - \pi(y)| = \|K_x^n - \pi\|_{TV}.\end{aligned}$$

where $y_j \in \mathcal{O}_j$. This completes our proof. $\square$

4.1 General Burnside Process

Suppose that $\mathcal{X}$ is a finite state space and G is a finite group acting on $\mathcal{X}$. Let (K, π) be the corresponding Burnside process and $(\bar{K}, \bar{\pi})$ be the lumped chain defined as before.

Lemma 3 (exercise 9, section 1.8, [6]). *For any two subgroups G_1, G_2 of a finite group G ,*

$$|G_1 G_2| = \frac{|G_1| |G_2|}{|G_1 \cap G_2|}$$

where $G_1 G_2 = \{gh : g \in G_1, h \in G_2\}$.

Proposition 10. *For every $n \geq 1$, $\max_x \|K_x^n - \pi\|_{TV} \leq \left(1 - \frac{1}{|G|}\right)^n$.*

Proof: Consider the Doeblin bound, by using Cauchy-Schwarz inequality and lemma 3.,

$$\begin{aligned}\frac{K(x, y)}{\pi(y)} &= \frac{1}{|G_x|} \left(\sum_{g \in G_x \cap G_y} \frac{1}{|\mathcal{X}_g|} \right) \left(\frac{1}{|G|} \sum_g |\mathcal{X}_g| \right) \left(\frac{|G|}{|G_y|} \right) \\ &\geq \frac{1}{|G_x| |G_y|} \left(\sum_{i=1}^{|G_x \cap G_y|} |\mathcal{X}_{g_i}| \right) \left(\sum_{i=1}^{|G_x \cap G_y|} \frac{1}{|\mathcal{X}_{g_i}|} \right) \\ &\geq \frac{|G_x \cap G_y|^2}{|G_x| |G_y|} = \frac{|G_x \cap G_y|}{|G_x G_y|} \geq \frac{1}{|G|},\end{aligned}$$

where $G_x \cap G_y = \{g_i\}_{i=1}^{|G_x \cap G_y|}$. Hence, from theorem 1, we have for every $n \geq 1$,

$$\max_x \|K_x^n - \pi\|_{TV} \leq \left(1 - \frac{1}{|G|}\right)^n.$$

$\square$

Proposition 11.

$$\max_{\mathcal{O} \in \bar{\mathcal{X}}} \|\bar{K}_{\mathcal{O}}^n - \bar{\pi}\|_{TV} \leq \left(1 - \frac{1}{|\mathcal{X}|}\right)^n.$$

Proof: We use the methodology of coupling to prove this. Define

$$\tilde{K}((\mathcal{O}_i, \mathcal{O}_j), (\mathcal{O}_{i'}, \mathcal{O}_{j'})) = \begin{cases} \bar{K}(\mathcal{O}_i, \mathcal{O}_{i'}) \bar{K}(\mathcal{O}_j, \mathcal{O}_{j'}), & \text{if } \mathcal{O}_i \neq \mathcal{O}_j, \\ \bar{K}(\mathcal{O}_i, \mathcal{O}_{i'}), & \text{if } \mathcal{O}_i = \mathcal{O}_j \text{ and } \mathcal{O}_{i'} = \mathcal{O}_{j'}, \\ 0, & \text{otherwise.} \end{cases}$$

In other words, $\mathcal{O}_i$ and $\mathcal{O}_j$ move independently until they hit and then move together. It is easy to check that $\tilde{K}$ is a Markovian coupling. Therefore, suppose that $(X_n, Y_n) = (\mathcal{O}_i, \mathcal{O}_j)$, we may observe that

- (i) if $\mathcal{O}_i = \mathcal{O}_j$, $X_{n+1} = Y_{n+1}$.
- (ii) if $\mathcal{O}_i \neq \mathcal{O}_j$,

$$\begin{aligned} P(X_{n+1} = Y_{n+1} | (X_n, Y_n) = (\mathcal{O}_i, \mathcal{O}_j)) &= \sum_{\mathcal{O}} \tilde{K}((\mathcal{O}_i, \mathcal{O}_j), (\mathcal{O}, \mathcal{O})) \\ &= \sum_{\mathcal{O}} \bar{K}(\mathcal{O}_i, \mathcal{O}) \bar{K}(\mathcal{O}_j, \mathcal{O}). \end{aligned}$$

Consequently, we have

$$\begin{aligned} P_{i,j}(X_2 \neq Y_2) &= \sum_{i',j'} P(X_2 \neq Y_2 | (X_1, Y_1) = (\mathcal{O}_{i'}, \mathcal{O}_{j'})) P_{i,j}((X_1, Y_1) = (\mathcal{O}_{i'}, \mathcal{O}_{j'})) \\ &= \sum_{i' \neq j'} P(X_2 \neq Y_2 | (X_1, Y_1) = (\mathcal{O}_{i'}, \mathcal{O}_{j'})) P_{i,j}((X_1, Y_1) = (\mathcal{O}_{i'}, \mathcal{O}_{j'})) \\ &= \sum_{i' \neq j'} [1 - P(X_2 = Y_2 | (X_1, Y_1) = (\mathcal{O}_{i'}, \mathcal{O}_{j'}))] P_{i,j}((X_1, Y_1) = (\mathcal{O}_{i'}, \mathcal{O}_{j'})) \\ &\leq \max_{i' \neq j'} \left(1 - \sum_{\mathcal{O}} \bar{K}(\mathcal{O}_{i'}, \mathcal{O}) \bar{K}(\mathcal{O}_{j'}, \mathcal{O}) \right). \end{aligned}$$

Inductively, it implies that

$$P_{i,j}(X_n \neq Y_n) \leq \left(\max_{i' \neq j'} \left(1 - \sum_{\mathcal{O}} \bar{K}(\mathcal{O}_{i'}, \mathcal{O}) \bar{K}(\mathcal{O}_{j'}, \mathcal{O}) \right) \right)^n.$$

On the other hand, for $x \in \mathcal{O}_{i'}$ and $y \in \mathcal{O}_{j'}$,

$$\bar{K}(\mathcal{O}_{i'}, \mathcal{O}_{j'}) = \frac{1}{|G_x|} \sum_{g \in G_x} \frac{|\mathcal{O}_{j'} \cap \mathcal{X}_g|}{|\mathcal{X}_g|} \geq \frac{1}{|\mathcal{X}|},$$

and

$$1 - \sum_{\mathcal{O}} \bar{K}(\mathcal{O}_{i'}, \mathcal{O}) \bar{K}(\mathcal{O}_{j'}, \mathcal{O}) = \sum_{\mathcal{O}} \bar{K}(\mathcal{O}_{i'}, \mathcal{O}) (1 - \bar{K}(\mathcal{O}_{j'}, \mathcal{O})) \leq \max_{\mathcal{O}_{j'}, \mathcal{O}} (1 - \bar{K}(\mathcal{O}_{j'}, \mathcal{O}))$$

then by the proposition 8, it implies

$$\max_{\mathcal{O} \in \mathcal{X}} \|K_{\mathcal{O}}^n - \pi\|_{TV} \leq \left(1 - \frac{1}{|\mathcal{X}|} \right)^n.$$

Assume that G is an abelian group. Since $|G_x| = \frac{|G|}{|\mathcal{O}_x|} = \frac{|G|}{|\mathcal{O}_{xg}|} = G_{xg}$ and $G_x \subseteq G_{xg}$, it follows that $G_x = G_{xg}$. Hence, □

$$K(x, y) = \frac{1}{|G_x|} \sum_{g \in G_x \cap G_y} \frac{1}{|\mathcal{X}_g|} = \frac{1}{|G_{xg}|} \sum_{g \in G_{xg} \cap G_{yh}} \frac{1}{|\mathcal{X}_g|} = K(x^g, y^h),$$

and for every $n \geq 1$, and $x \in \mathcal{X}$,

$$\|K_x^n - \pi\|_{TV} = \|\bar{K}_{\mathcal{O}_x} - \bar{\pi}\|_{TV}.$$

Furthermore, from lemma 2. $\bar{K}$ is symmetric.

Corollary 1. *If G is an abelian group, then*

$$\max_{x \in \mathcal{X}} \|K_x^n - \pi\|_{TV} \leq \left(1 - \frac{1}{|\mathcal{X}|} \right)^n.$$

4.2 Bose-Einstein Markov chain

Suppose that n balls put into k boxes. Put $\mathcal{X} = \{1, \dots, k\}^n$ the state space. In other words, if $x = (x_1, \dots, x_n) \in \mathcal{X}$, it means that the i -th ball is put in the x_i -th box. Assume that $G = S_n$ is the symmetric group acting on $\mathcal{X}$ by $y = x^g$ where $y_j = x_i$ whenever $j = g(i)$. Hence, we have a Burnside process defined on $\mathcal{X}$ by following eq. (12). In particular, the stationary distribution of the lumped chain is given by the well-known *Bose-Einstein* distribution,

$$\pi = \frac{1}{\binom{n+k-1}{k-1}},$$

which is used to describe the distribution of the energy of state in statistical physics. As we can see that the formula of transition kernel K of Burnside process is very complicated. It seems hard to calculate every single term of K exactly. However, it turns out in Diaconis [3] that Pólya theory of counting prides a crucial bridge for doing that. Based on his idea, the following presents a formula for the kernel of the lumped chain.

Definition 7. A partition λ of a nonnegative integer n is a sequence $(\lambda_1, \lambda_2, \dots, \lambda_k) \in \mathbb{N}^k$ so that $\sum_{i=1}^k \lambda_i = n$ and $\lambda_i \geq \lambda_{i+1}$. Here, we shall write $\lambda \vdash n$ to denote λ is a partition of n . We use the notation $b_i(\lambda)$ to denote the total number of i appearing in λ . Precisely, $1b_1(\lambda) + 2b_2(\lambda) + \dots + nb_n(\lambda) = n$.

Lemma 4. Suppose that P_{S_n} is the cycle index polynomial of the symmetry group S_n , then

i)

$$P_{S_n}(x_1, x_2, \dots, x_n) = \sum_{\lambda \vdash n} \frac{x_1^{b_1} x_2^{b_2} \dots x_n^{b_n}}{\prod_{i=1}^n b_i! i^{b_i}}. \quad (14)$$

ii)

$$\sum_{n=0}^{\infty} P_{S_n}(x_1, x_2, \dots, x_n) z^n = \prod_{n=1}^{\infty} \exp(z^n x_n / n). \quad (P_{S_0} \equiv 1) \quad (15)$$

Proof: These two results are classical and can be found in many textbooks of combinatorial mathematics, for example, chapter E. "Pólya theory of counting" in [2]. $\square$

Suppose that $f : \mathbb{R}^k \rightarrow \mathbb{R}$ can be expressed as a Taylor series. We define the notation

$$(t_1^{x_1} t_2^{x_2} \dots t_k^{x_k}) [f(t_1, t_2, \dots, t_k)]$$

to be the coefficient of $t_1^{x_1} t_2^{x_2} \dots t_k^{x_k}$ of $f(t_1, t_2, \dots, t_k)$ where $x_1, x_2, \dots, x_k \in \mathbb{N} \cup \{0\}$.

Theorem 4. Consider the Burnside random walk with S_n acting on $\mathcal{X} = [k]^n$. Let x and y be two states in $\mathcal{X}$. Suppose that x and y have x_i and y_i balls in box i , respectively. More precisely, $x_1 + x_2 + \dots + x_k = y_1 + y_2 + \dots + y_k = n$. Then the transition matrix of its lumped chain can be formulated as the following

$$\bar{K}(\mathcal{O}_x, \mathcal{O}_y) = (t_1^{x_1} t_2^{x_2} \dots t_k^{x_k}) \left[P_{A_y} \left(\frac{\sum_{j=1}^k t_j^1}{k}, \frac{\sum_{j=1}^k t_j^2}{k}, \dots, \frac{\sum_{j=1}^k t_j^n}{k} \right) \right] \quad (16)$$

where P_{A_y} is the cycle index polynomial of $S_{y_1} \times S_{y_2} \times \dots \times S_{y_k}$.

Proof of Theorem: Recall that the lumped chain is originally defined by

$$\overline{K}(\mathcal{O}_x, \mathcal{O}_y) = \frac{1}{|G_x|} \left(\sum_{z \in \mathcal{O}_y} \sum_{g \in G_x \cap G_z} \frac{1}{|\mathcal{X}_g|} \right) = \frac{1}{|G_x|} \sum_{g \in G_x} \left(\sum_{z \in \mathcal{O}_y} 1_{\{z^g=z\}} \right) \frac{1}{|\mathcal{X}_g|}.$$

If $z^g = z$ then z is constant on every cycle of g . On the other hand, since $z \in \mathcal{O}_y$,

$$\sum_{z \in \mathcal{O}_y} 1_{\{z^g=z\}} = \sum_{\lambda_i \vdash y_i, \forall i=1,2,\dots,k} \prod_{i=1}^n \binom{c_i(g)}{b_i(\lambda_1) b_i(\lambda_2), \dots, b_i(\lambda_k)}.$$

Therefore, the transition matrix can be expressed as

$$\sum_{\lambda_i \vdash y_i, \forall i=1,2,\dots,k} \left[\frac{1}{x_1! x_2! \dots x_k!} \sum_{g \in G_x} \prod_{i=1}^n \binom{c_i(g)}{b_i(\lambda_1) b_i(\lambda_2), \dots, b_i(\lambda_k)} \frac{1}{k^{c(g)}} \right]. \quad (17)$$

To attain our purpose, from ii) in lemma 4, we use the identity,

$$\prod_{j=1}^k \left[\sum_{i=0}^{\infty} t_j^i P_{S_i}(u_1, u_2, \dots, u_i) \right] = \prod_{i=1}^{\infty} \exp \left(\frac{u_i}{i} \sum_{j=1}^k t_j^i \right). \quad (18)$$

Now, for each $\lambda_i \vdash y_i$, let $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_k)$, and define the differential operator

$$\left(\frac{\partial}{\partial u} \right)^\lambda = \frac{\partial^{\sum_{i=1}^n \sum_{j=1}^k b_i(\lambda_j)}}{\partial u_1^{\sum_{j=1}^k b_1(\lambda_j)} \partial u_2^{\sum_{j=1}^k b_2(\lambda_j)} \dots \partial u_n^{\sum_{j=1}^k b_n(\lambda_j)}}.$$

Thus, on the one hand,

$$\left(\frac{\partial}{\partial u} \right)^\lambda P_{G_x}(u_1, u_2, \dots, u_n) = \frac{c_\lambda(u)}{x_1! x_2! \dots x_k!} \sum_{g \in G_x} \prod_{i=1}^n \binom{c_i(g)}{b_i(\lambda_1) b_i(\lambda_2) \dots b_i(\lambda_k)} u_i^{c_i(g)} \quad (19)$$

where $c_\lambda(u) = \prod_{i=1}^n \prod_{j=1}^k u_i^{-b_i(\lambda_j)} b_i(\lambda_j)!$. And on the other hand,

$$\left(\frac{\partial}{\partial u} \right)^\lambda \prod_{i=1}^{\infty} \exp \left(\frac{u_i}{i} \sum_{j=1}^k t_j^i \right) = \left[\prod_{i=1}^{\infty} \exp \left(\frac{u_i}{i} \sum_{j=1}^k t_j^i \right) \right] \left[\prod_{i=1}^n \left(\frac{1}{i} \sum_{j=1}^k t_j^i \right)^{\sum_{l=1}^k b_i(\lambda_l)} \right]. \quad (20)$$

Therefore, combine eqs.(19), (20) together, it follows

$$\begin{aligned} & \frac{1}{x_1! x_2! \dots x_k!} \sum_{g \in G_x} \prod_{i=1}^n \binom{c_i(g)}{b_i(\lambda_1) b_i(\lambda_2) \dots b_i(\lambda_k)} \frac{1}{k^{c(g)}} \\ &= c_\lambda(1/k, 1/k, \dots, 1/k)^{-1} (t_1^{x_1} t_2^{x_2} \dots t_k^{x_k}) \left[\left(\frac{\partial}{\partial u} \right)^\lambda \prod_{i=1}^n \sum_{j=0}^{\infty} t_j^i P_{S_j}(1/k, 1/k, \dots, 1/k) \right] \\ &= c_\lambda(1/k, 1/k, \dots, 1/k)^{-1} (t_1^{x_1} t_2^{x_2} \dots t_k^{x_k}) \left[\left(\frac{\partial}{\partial u} \right)^\lambda \prod_{i=1}^{\infty} \exp \left(\frac{\sum_{j=1}^k t_j^i}{ik} \right) \right] \\ &= c_\lambda(1/k, 1/k, \dots, 1/k)^{-1} (t_1^{x_1} t_2^{x_2} \dots t_k^{x_k}) \left[\prod_{i=1}^n \left(\frac{1}{i} \sum_{j=1}^k t_j^i \right)^{\sum_{l=1}^k b_i(\lambda_l)} \right] \\ &= (t_1^{x_1} t_2^{x_2} \dots t_k^{x_k}) \left\{ \prod_{i=1}^k \prod_{l=1}^n \left[\left(\frac{\sum_{j=1}^k t_j^l}{k} \right)^{b_l(\lambda_i)} \frac{1}{b_l(\lambda_i)! l^{b_l(\lambda_i)}} \right] \right\}. \end{aligned}$$

Finally, from i) in lemma 4, we are done if we rearrange eq. (16) suitably. $\square$

Remark 2. Particularly, if $k = n$ and $x_i = 1$ for each i , then

$$\overline{K}(\mathcal{O}_x, \mathcal{O}_y) = \binom{n}{y_1 y_2 \cdots y_n} \frac{1}{n^n}.$$

For general n and k , suppose x is constant then

$$\overline{K}(\mathcal{O}_x, \mathcal{O}_y) = \binom{n}{y_1 y_2 \cdots y_n} \frac{1}{n!} \prod_{i=1}^k \frac{\Gamma(y_i + \frac{1}{k})}{\Gamma(\frac{1}{k})}.$$

Proposition 12. *The transition matrix $\overline{K}$ is symmetric.*

Proof: From the representation of $\overline{K}$ in theorem 4 and eq. (4), this property holds clearly, indeed,

$$\begin{aligned} \overline{K}(\mathcal{O}_x, \mathcal{O}_y) &= \sum_{\lambda_i \vdash x_i, \forall i=1, 2, \dots, k} \left[\frac{1}{y_1! y_2! \cdots y_k!} \sum_{g \in G_y} \prod_{i=1}^n \binom{c_i(g)}{b_i(\lambda_1) b_i(\lambda_2), \dots, b_i(\lambda_k)} \frac{1}{k^{c(g)}} \right] \\ &= \overline{K}(\mathcal{O}_y, \mathcal{O}_x). \end{aligned}$$

$\square$

We can easily check that if x is constant, then for every $y \in \mathcal{X}$ and $g \in S_n$,

$$K(x, y) = K(x, y^g).$$

By proposition 9, 10 and 11, it follows that

$$\|K_x^l - \pi\|_{TV} \leq \left(1 - \frac{1}{\min\{n!, k^n\}}\right)^l. \quad (21)$$

Diaconis [3] proved that

$$\|K_x^l - \pi\|_{TV} \leq \left(1 - \frac{1}{k!}\right)^l. \quad (22)$$

If k is large, then (21) is better than (22). Aldous [1] gives a remarkable upper bound of this chain by using an inspired coupling method. Precisely, he proved that

Theorem 5 (Aldous). *For every state $x \in [k]^n$,*

$$\|K^l(x, \cdot) - \pi(\cdot)\|_{TV} \leq n \left(1 - \frac{1}{k}\right)^l.$$

In other words, this means that the spectral gap λ of Burnside process is bounded below by $\frac{1}{k}$. Recall that equation (9) says

$$\lambda = \lim_{t \rightarrow \infty} \frac{-1}{t} \log \left(2 \sup_{x \in \mathcal{X}} \|H^t(x, \cdot) - \pi(\cdot)\|_{TV} \right).$$

Therefore, for each x ,

$$\begin{aligned} 2\|H^t(x, \cdot) - \pi(\cdot)\|_{TV} &= \sum_y |H^t(x, y) - \pi(y)| \\ &= \sum_y \left| e^{-t} \sum_{l=0}^{\infty} \frac{t^l}{l!} [K^l(x, y) - \pi(y)] \right| \\ &\leq e^{-t} \sum_{l=0}^{\infty} \frac{t^l}{l!} \sum_y |K^l(x, y) - \pi(y)| \\ &\leq 2ne^{-t} \sum_{l=0}^{\infty} \frac{t^l}{l!} \left(1 - \frac{1}{k}\right)^l = 2ne^{-t} e^{t(1-1/k)} = 2ne^{-t/k} \end{aligned}$$

Plug this result into eq. (9), we have

$$-\frac{1}{t} \log \left(2 \sup_x \|H^t(x, \cdot) - \pi(\cdot)\|_{TV} \right) \geq \frac{-\log(2ne^{-t/k})}{t} = \frac{-\log 2n + t/k}{t} = \frac{-\log 2n}{t} + \frac{1}{k},$$

and now, passing to the limit, it follows $\lambda \geq \frac{1}{k}$.

Hence, with the upper bounded of the spectral gap, we can immediately get the mixing time for L^p distance from theorem 3.

Corollary 2. *For this Burnside process K , the mixing time has an upper bound,*

$$T_p \leq \begin{cases} \frac{k}{2} (2 + n \log k), & \text{if } 1 \leq p \leq 2 \\ k (1 + n \log k), & \text{if } 2 < p \leq \infty, \end{cases}$$

where

$$T_p \equiv \inf \left\{ t > 0 : \max_{x \in \mathcal{X}} d_{\pi,p}(H^t(x, \cdot), \pi) \leq 1/e \right\}.$$

4.3 Other Burnside Processes

Suppose that S_n be the symmetry group and $G = \langle h \rangle$ where $h = (1, 2, \dots, m)$ and $m < n$ is a prime. Consider the Burnside process with the group G acting on the state space $\mathcal{X} = [k]^n$. Observe that $x^g = x$ for some g in G if and only if x is constant on h , i.e. $x_1 = x_2 = \dots = x_m$. This observation is easy to show. Indeed, since $|g| = m$ and

$$x_1 = x_{g(1)} = x_{g^2(1)} = \dots = x_{g^{i-1}(1)} = x_{g^i(1)} = \dots = x_{g^{m-1}(1)} = x_1,$$

it follows that $x_1 = x_2 = \dots = x_m$.

Lemma 5. *Let $x = (x_1, x_2, \dots, x_n) \in \mathcal{X}$ and $g \in G$, then*

$$G_x = \begin{cases} G, & \text{if } x \text{ is constant on } h, \\ \{e\}, & \text{otherwise,} \end{cases}$$

$$|\mathcal{O}_x| = \begin{cases} 1, & \text{if } x \text{ is constant on } h, \\ m, & \text{otherwise,} \end{cases}$$

and

$$|\mathcal{X}_g| = \begin{cases} k^{n-m+1}, & \text{if } g \neq e, \\ k^n, & \text{if } g = e. \end{cases}$$

Lemma 6. *The transition matrices of Burnside process and its lumped chain are given respectively by*

$$K(x, y) = \frac{1}{mk^n} \begin{cases} (m-1)k^{m-1} + 1, & \text{if both } x \text{ and } y \text{ are constant on } h, \\ 1, & \text{if } x \text{ is constant on } h \text{ and } y \text{ is not constant on } h, \\ m, & \text{if } x \text{ is not constant on } h. \end{cases}$$

and

$$\bar{K}(O_x, O_y) = \frac{1}{mk^n} \begin{cases} (m-1)k^{m-1} + 1, & \text{if both } x \text{ and } y \text{ are constant on } h, \\ m, & \text{if there is exactly one of } x \text{ and } y \text{ be constant on } h, \\ m^2, & \text{if neither } x \text{ nor } y \text{ are constant on } h. \end{cases}$$

Proof: Recall that the transition matrix of the Burnside process is formulated by

$$K(x, y) = \frac{1}{|G_x|} \sum_{g \in G_x \cap G_y} \frac{1}{|\mathcal{X}_g|}.$$

And the lumped chain is $\bar{K}(\mathcal{O}_x, \mathcal{O}_y) = |\mathcal{O}_y|K(x, y)$. From lemma 5, we can easily get what we want. $\square$

With this transition matrix $\bar{K}$, we can calculate several important quantities of this chain. Let λ be the spectral gap of this Burnside process.

Proposition 13.

$$\frac{1}{m} \leq \lambda \leq \frac{1}{m} + \frac{1}{k^{m-1}} - \frac{1}{mk^{m-1}}.$$

Proof: Firstly, from proposition 4. we see that

$$\max_{x \in \mathcal{X}} \|K_x^l - \pi\|_{TV} \leq \left(1 - \frac{1}{m}\right)^l.$$

Hence, by using 3 and following the same argument as above we could see that the spectral gap λ of K is bounded below by $1/m$. On the other hand, set a test function $f : \mathcal{X} \rightarrow \mathbb{R}$ by

$$f(x) = \begin{cases} 1, & \text{if } x \text{ is constant on } h, \\ 0, & \text{otherwise.} \end{cases}$$

and the number of disjoint orbits is

$$z = \frac{1}{|G|} \sum_{g \in G} |\mathcal{X}_g| = \frac{1}{m} [k^n + (m-1)k^{n-m+1}] = \frac{k^n}{m} \left[1 + \frac{m-1}{k^{m-1}}\right]$$

then it follows that

$$\frac{\mathcal{E}(f, f)}{\text{Var}(f)} = \frac{\frac{1}{k^n}}{\frac{k^n}{m} \left[1 + \frac{m-1}{k^{m-1}}\right]} = \frac{1}{m} + \frac{1}{k^{m-1}} - \frac{1}{mk^{m-1}}.$$

From the definition of spectral gap, we get the assertion. $\square$

Next, we turn to calculate the lower bound for the total variation distance.

Proposition 14. *If x is constant on h and $k \geq 8$, then for each $1 \leq l \leq m$,*

$$\left(1 - \frac{8}{k}\right) \left(1 - \frac{1}{m}\right)^l \leq \|\bar{K}_x^l - \pi\|_{TV}.$$

Proof: Define a square matrix A of dimension $k^n \times k^n$ by

$$A_{xy} = \begin{cases} \frac{m-1}{mk^{n-m+1}} + \frac{1}{mk^n}, & \text{if both } x \text{ and } y \text{ are constant on } h, \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to see that $K^l \geq A^l$ for every $l \geq 1$, and a simple calculation yields

$$A_{xy}^l = k^{(n-m+1)(l-1)} \left[\frac{m-1}{mk^{n-m+1}} + \frac{1}{mk^n} \right]^l = \frac{1}{k^{n-m+1}} \left[1 - \frac{1}{m} \left(1 - \frac{1}{k^{m-1}} \right) \right]^l,$$

whenever both x and y are constant on h . Set $\mathcal{A} = \{x : x \text{ is constant on } h\}$. Then

$$\|K_x^l - \pi\|_{TV} \geq \sum_{y \in \mathcal{A}} \left[K^l(x, y) - \frac{1}{z} \right] \geq \sum_{y \in \mathcal{A}} \left[A_{xy}^l - \frac{1}{z} \right].$$

Since

$$\begin{aligned}
\sum_{y \in \mathcal{A}} \left[A_{xy}^l - \frac{1}{z} \right] &= k^{n-m+1} \left[\frac{1}{k^{n-m+1}} \left[1 - \frac{1}{m} \left(1 - \frac{1}{k^{m-1}} \right) \right]^l - \frac{m}{k^n} \left[1 + \frac{m-1}{k^{m-1}} \right]^{-1} \right] \\
&= \left[\left(1 - \frac{1}{m} \right) + \frac{1}{mk^{m-1}} \right]^l - \frac{1}{\left[\left(1 - \frac{1}{m} \right) + \frac{k^{m-1}}{m} \right]} \\
&\quad (\text{use binomial formula}) \\
&\geq \left(1 - \frac{1}{m} \right)^l + \frac{l}{mk^{m-1}} \left(\frac{m}{m-1} \right) \left(1 - \frac{1}{m} \right)^l - \frac{m}{k^{m-1}} \\
&\geq \left(1 - \frac{1}{m} \right)^l + \frac{l}{mk^{m-1}} \left(1 - \frac{1}{m} \right)^l - \frac{m}{k^{m-1}}.
\end{aligned}$$

Assume that $l \leq m$. The last inequality becomes

$$\begin{aligned}
&\left(\frac{4m^2 - l}{mk^{m-1}} + \frac{mk^{m-1} + l - 4m^2}{mk^{m-1}} \right) \left(1 - \frac{1}{m} \right)^l + \frac{l}{mk^{m-1}} \left(1 - \frac{1}{m} \right)^l - \frac{m}{k^{m-1}} \\
&= \left(\frac{mk^{m-1} + l - 4m^2}{mk^{m-1}} \right) \left(1 - \frac{1}{m} \right)^l + \frac{4m}{k^{m-1}} \left(1 - \frac{1}{m} \right)^l - \frac{m}{k^{m-1}} \\
&\geq \left(\frac{mk^{m-1} + l - 4m^2}{mk^{m-1}} \right) \left(1 - \frac{1}{m} \right)^l + \frac{4m}{k^{m-1}} \left(1 - \frac{1}{m} \right)^m - \frac{m}{k^{m-1}} \\
&\geq \left(\frac{mk^{m-1} + l - 4m^2}{mk^{m-1}} \right) \left(1 - \frac{1}{m} \right)^l + \frac{4m}{k^{m-1}} \left(1 - \frac{1}{2} \right)^2 - \frac{m}{k^{m-1}} \\
&= \left(\frac{mk^{m-1} + l - 4m^2}{mk^{m-1}} \right) \left(1 - \frac{1}{m} \right)^l \\
&\geq \left(1 - \frac{4m}{k^{m-1}} \right) \left(1 - \frac{1}{m} \right)^l \\
&\geq \left(1 - \frac{8}{k} \right) \left(1 - \frac{1}{m} \right)^l.
\end{aligned}$$

We note that $\left(1 - \frac{1}{m} \right)^m$ and $-\frac{4m}{k^{m-1}}$ are increasing in m . Hence, we complete our proof. $\square$

As a conclusion, if k is large, then from proposition 13, the spectral gap of K is almost equal to $\frac{1}{m}$. Therefore, theorem 3 tells us that the mixing time of the total variation distance of the continuous time Markov chain is bounded below by m . While in proposition 14, the discrete time Markov chain also have the same lower bound m for the total variation distance.

References

- [1] Aldous, D. and Fill, J.A. (2001) *Reversible Markov chains and random walks on graphs*, Book in preparation.
- [2] Beckenbach, E. (1968) *Applied Combinatorial Mathematics*. Wiley, New York.
- [3] Diaconis, P. (2005) *Analysis of a Bose-Einstein Markov Chain*. Ann. I. H. Poincaré 41, 409-418.
- [4] Goldberg, L. (1993) *Automating Pólya theory: the computational complexity of the cycle index polynomial*. Info and Computation 105, 268-288.

- [5] Goldberg, L. A. and Jerrum M. (2002) *The "Burnside Process" converges slowly*, Combinatorics, Probability, and Computing 11, 21-34.
- [6] Jacobson, N. (1973) *Basic Algebra I*.
- [7] Jerrum, M. (1993) *Uniform sampling modulo a group a symmetries using Markov chain simulation* DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 10, 37-47.
- [8] Saloff-Coste, L. (1997) *Lectures on finite Markov chains*, Ecole d'Eté de Probabilités de Saint Four XXVI-1996, Lect. Notes in Math. 1665, Springer, Berlin.