アルゴリズム89－6
（2003．3．14）

Dirichlet 分布の rapidly mixing approximate sampler

松井 知己† 元木 光雄‡ 鎌谷 直之§

† 東京大学大学院 情報理工学系研究科 数理情報学専攻
‡ 北陸先端科学技術大学院大学 情報科学研究科 情報処理学専攻
§ 東京女子医科大学 附属膠原病リウマチ痛風センター

概要 Dirichlet 分布は，多項分布の共役事前分布として，ベイズ統計においてよく用いられている．近年，遺伝統計学の発達により，さまざまな形で Dirichlet 分布からのランダム変数の生成が必要になってきている．本稿では，離散化した Dirichlet 分布を生成する Markov chain を取り上げる．この Markov chain の収束速度が $(1/2)n(n-1)\ln((\Delta-n)\varepsilon^{-1})$ であることを，path coupling 法を用いて示した（ただし，n は Dirichlet 分布のパラメータの数，Δ は離散化したときのグリッドサイズ，ε は誤差である）．また，シミュレーションにより，収束速度がさらに速い可能性があることを実験的に示した．

A rapidly mixing approximate sampler of Dirichlet distribution

Tomomi MATSUI† Mitsuo MOTOKI‡ Naoyuki KAMATANI§

† Department of Mathematical Informatics,
Graduate School of Information Science and Technology,
The University of Tokyo, Bunkyo-ku, Tokyo 113-8656, Japan.
http://www.simplex.t.u-tokyo.ac.jp/~tomomi/

‡ Department of Information Processing,
School of Information Science,
Japan Advanced Institute of Science and Technology,
1-1, Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan.
mmotoki@jaist.ac.jp

§ Institute of Rheumatology, Tokyo Women's Medical University,
10-22 Kawada-cho, Shinjuku-ku, Tokyo 162-0054, Japan.

Abstract In this paper, we propose a Markov chain for sampling a random variable distributed according to a discretized Dirichlet distribution. We show that our Markov chain is rapidly mixing, that is, the mixing time of our chain is $(1/2)n(n-1)\ln((\Delta-n)\varepsilon^{-1})$ where n is the dimension (the number of parameters), $1/\Delta$ is the grid size for discretization, and ε is the error bound. We estimate the mixing time by using the path coupling method. When the parameters are large, the log-concavity of the density function implies the rapidity straightforwardly. In the case that parameters are small, the density function is convex and so we need a different approach to use the path coupling method. We also show the rate of convergence of our chain experimentally.

1 Introduction

Statistical methods are widely studied in bioinformatics since they are powerful tools to discover genes causing a (common) disease from a number of observed data. These methods often use EM algorithm, Markov chain Monte Carlo method, Gibbs sampler, and so on. The Dirichlet distribution, a distribution over vectors of positive numbers in which sum is equal to 1, often appears as prior and posterior distribution for the multinomial distribution in these methods since this is the conjugate prior of parameters of the multinomial distribution.

For example, Nil et al. propose a Bayesian haplotype inference method [5], that is, deciding phased (paternal and maternal) individual genotypes probabilistically. This method is based on Gibbs sampler. In their method, the Dirichlet distribution is used to update population haplotype frequencies, i.e., parameters of the multinomial distribution, for each iteration. That is to say, for each iteration starting from a Dirichlet distribution with some appropriate parameters, parameters of the multinomial distribution is updated from the posterior distribution which is a Dirichlet distribution with updated parameters conditional on the "imputed" events, .

Another example is a population structure inferring algorithm by Pritchard et al. [6]. This algorithm is based on MCMC algorithm. For each step of MCMC, the Dirichlet distribution with two distinct sets of parameters are used to sample allele frequencies in each population and admixture proportions for each individual. Similar to the first example, these two sets of parameters are updated at each iteration.

In these examples, the Dirichlet distribution appears with various dimensions and various parameters. Thus we need an efficient algorithm for sampling from the Dirichlet distribution with arbitrary dimensions and arbitrary parameters. One approach of sampling from the Dirichlet distribution is by rejection (see [3] for example). In this way, we have to sample from the gamma distribution as many as the size of dimension of a Dirichlet distribution. Though we can sample from the gamma distribution by using rejection sampling, the ratio of rejection becomes higher as the parameter is smaller. Thus, it does not seems effective way for small parameters.

We employ another approach, the Metropolis algorithm using a Markov chain. In this case, it is important to estimate a mixing rate of Markov chain. Otherwise, the samples may not be distributed according to a desired distribution.

In this paper, we propose a simple Markov chain for sampling a random variable distributed according to a discretized Dirichlet distribution. We show that our Markov chain is rapidly mixing, that is, the mixing rate of our chain is quadratic of the number of the dimension of a Dirichlet distribution and logarithmic of discretizing grid size and the inverse of variation distance. We note that this mixing rate does not depend on the magnitude of parameters. We also show experimentally that the recuired number of steps of our Markov chain is much smaller than our theoretical upper bound of the mixing rate.

2 Markov chain for Approximate Sampler

Dirichlet random vector $P = (P_1, P_2, \ldots, P_n)$ with positive parameters $u_1, \ldots, u_n$ is a vector of random variables that admits the probability density function

$$\frac{\Gamma(\sum_{i=1}^n u_i)}{\prod_{i=1}^n \Gamma(u_i)} \prod_{i=1}^n p_i^{u_i - 1}$$

defined on the set $\{(p_1, p_2, \ldots, p_n) \in \mathrm{R}^n \mid \sum_{i=1}^n p_i = 1,\ p_1, p_2, \ldots, p_n > 0\}$ where $\Gamma(u)$ is the gamma function. Throughout this paper, we assume that $n \geq 2$.

For any integer $\Delta \geq n$, we discretize Ω with grid size $1/\Delta$ and obtain a discrete set of integer vectors Ω defined by

$$\Omega \stackrel{\text{def.}}{=} \left\{ (p_1, p_2, \ldots, p_n) \in \mathrm{Z}^n \mid p_i > 0\ (\forall i),\ \sum_{i=1}^n p_i = \Delta \right\}.$$

The discretized Dirichlet random vector with positive parameters $u_1, \ldots, u_n$ is a random vector $X = (X_1, \ldots, X_n) \in \Omega$ with the distribution

$$\Pr[X = (x_1, \ldots, x_n)] = g(\boldsymbol{x}) \stackrel{\text{def.}}{=} C_\Delta \prod_{i=1}^n (x_i/\Delta)^{u_i - 1}$$

where C_Δ is the partition function (normarizing constant) defined by $(C_\Delta)^{-1} \stackrel{\text{def.}}{=} \sum_{\boldsymbol{x} \in \Omega} \prod_{i=1}^n (x_i/\Delta)^{u_i - 1}$.

For any integer $b \geq 2$, we introduce a set of 2-dimensional integer vectors $\Omega(b) \stackrel{\text{def.}}{=} \{(Y_1, Y_2) \in \mathrm{Z}^2 \mid Y_1, Y_2 > 0,\ Y_1 + Y_2 = b\}$ and a distribution function $f_b(Y_1, Y_2 \mid u_i, u_j) : \Omega(b) \to [0, 1]$ with positive parameters u_i, u_j defined by

$$f_b(Y_1, Y_2 \mid u_i, u_j) \stackrel{\text{def.}}{=} C(u_i, u_j, b) Y_1^{u_i - 1} Y_2^{u_j - 1}$$

where $(C(u_i, u_j, b))^{-1} \stackrel{\text{def.}}{=} \sum_{(Y_1, Y_2) \in \Omega(b)} Y_1^{u_i - 1} Y_2^{u_j - 1}$ is the partition function.

We describe our Markov chain $\mathcal{M}$ with state space Ω. At each time $t \in \{0, 1, 2, \ldots\}$, transitions take place as follows.

Step 1: Pick a mutually distinct pair of indices $\{i, j\} \subseteq \{1, 2, \ldots, n\}$ uniformly at random.

Step 2: Put $b = X_i^t + X_j^t$. Pick $(Y_1, Y_2) \in \Omega(b)$ according to the distribution function $f_b(Y_1, Y_2 \mid u_i, u_j)$.

Step 3: Put $X_k^{t+1} = \begin{cases} Y_1 & (k = i), \\ Y_2 & (k = j), \\ X_k^t & (\text{otherwise}). \end{cases}$

Clearly, this chain is irereducible and aperiodic. Since the detailed balance equations hold, the stationary distribution of the above Markov chain $\mathcal{M}$ is $g(\boldsymbol{x})$.

The following theorem is a main result of this paper, which shows the mixing time of our chain.

Theorem 1. *The mixing time $\tau(\varepsilon)$ of Markov chain $\mathcal{M}$ satisfies*

$$\tau(\varepsilon) \leq (1/2) n(n-1) \ln((\Delta - n)\varepsilon^{-1}).$$

In the rest of this paper, we prove the above by using the path coupling method.

Before showing the above lemma, we briefly review the definition of the mixing time and path coupling method. For any probability distribution function π' on

Ω, define the *total variation distance* between the stationary distribution function g of $\mathcal{M}$ and π' to be

$$\mathrm{D_{TV}}(g,\pi') \stackrel{\text{def.}}{=} \max_{\Omega'\subseteq\Omega}\left|\sum_{\boldsymbol{x}\in\Omega'} g(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\Omega'}\pi'(\boldsymbol{x})\right| = \frac{1}{2}\sum_{\boldsymbol{x}\in\Omega}|g(\boldsymbol{x})-\pi'(\boldsymbol{x})|.$$

If the initial state of the chain $\mathcal{M}$ is $\boldsymbol{x}\in\Omega$, we denote the distribution of the chain at time t by $X^t_{\boldsymbol{x}} : \Omega \to [0,1]$, i.e.,

$$P^t_{\boldsymbol{x}}(\boldsymbol{y}) \stackrel{\text{def.}}{=} \Pr[X^t = \boldsymbol{y} \mid X^0 = \boldsymbol{x}] \quad (\forall \boldsymbol{y}\in\Omega).$$

The rate of convergence to stationary from the initial state $\boldsymbol{x}$ may be measured by

$$\tau_{\boldsymbol{x}}(\varepsilon) \stackrel{\text{def.}}{=} \min\{t \mid \mathrm{D_{TV}}(g, P^{t'}_{\boldsymbol{x}}) \le \varepsilon \quad \text{for all } t' \ge t\}$$

where the error bound ε is a given positive constant. The *mixing time* $\tau(\varepsilon)$ of $\mathcal{M}$ is defined by

$$\tau(\varepsilon) \stackrel{\text{def.}}{=} \max_{\boldsymbol{x}\in\Omega}\tau_{\boldsymbol{x}}(\varepsilon),$$

which is independent of the initial state.

Next, we define a special Markov process with respect to $\mathcal{M}$ called joint process. A *joint process* of $\mathcal{M}$ is a Markov chain (X^t, Y^t) defined on $\Omega\times\Omega$ satisfying that each of $(X^t), (Y^t)$, considered marginally, is a faithful copy of the original Markov chain $\mathcal{M}$. More precisely, we require that

$$\Pr[X^{t+1} = \boldsymbol{x}' | (X^t, Y^t) = (\boldsymbol{x},\boldsymbol{y})] = P_{\mathcal{M}}(\boldsymbol{x},\boldsymbol{x}'),$$
$$\Pr[Y^{t+1} = \boldsymbol{y}' | (X^t, Y^t) = (\boldsymbol{x},\boldsymbol{y})] = P_{\mathcal{M}}(\boldsymbol{y},\boldsymbol{y}'),$$

for all $\boldsymbol{x},\boldsymbol{y},\boldsymbol{x}',\boldsymbol{y}' \in \Omega$ where $P_{\mathcal{M}}(\boldsymbol{x},\boldsymbol{x}')$ and $P_{\mathcal{M}}(\boldsymbol{y},\boldsymbol{y}')$ denotes the transition probability from $\boldsymbol{x}$ to $\boldsymbol{x}'$ and from $\boldsymbol{y}$ to $\boldsymbol{y}'$ of the original Markov chain $\mathcal{M}$, respectively.

Lemma 1 (Path coupling lemma[1]). *Let G be a directed graph with vertex set Ω and arc set $A\subseteq\Omega\times\Omega$. Let $\ell : A \to \mathbf{Z}_{++}$ be a positive length function defined on the arc set. We assume that G is strongly connected. For any ordered pair of vertices $(\boldsymbol{x},\boldsymbol{x}')$ of G, the distance from $\boldsymbol{x}$ to $\boldsymbol{x}'$, denoted by $\mathrm{d}(\boldsymbol{x},\boldsymbol{x}')$, is the length of the shortest path from $\boldsymbol{x}$ to $\boldsymbol{x}'$, where the length of a path is the sum of the lengths of arcs in the path. Suppose that there exists a joint process $(X,Y)\mapsto(X',Y')$ with respect to $\mathcal{M}$ satisfying that*

$$1 > \exists\beta > 0,\ \forall (X,Y)\in A,\ E[\mathrm{d}(X',Y')] \le \beta \mathrm{d}(X,Y).$$

Then the mixing time $\tau(\varepsilon)$ of the original Markov chain $\mathcal{M}$ satisfies $\tau(\varepsilon)\le (1-\beta)^{-1}\ln(D/\varepsilon)$ where D denotes the diameter of G, i.e., the distance of a farthest (ordered) pair of vertices.

3 Analysis of Mixing Time

In this section, we define the joint process and analyze the mixing time by using path coupling method. First, we introduce a directed garph $G=(\Omega,A)$ whose vertex set is equivalent to the state space Ω. There exists a directed arc from state (vertex) $\boldsymbol{x}$ to $\boldsymbol{y}$ if and only if $\|\boldsymbol{x}-\boldsymbol{y}\|_1 \stackrel{\text{def.}}{=} (|x_1-y_1|+\cdots+|x_n-y_n|) = 2$. Thus the set A of arcs of G is defined by

$$A \stackrel{\text{def.}}{=} \{(\boldsymbol{x},\boldsymbol{y}) \mid \boldsymbol{x},\boldsymbol{y}\in\Omega, \|\boldsymbol{x}-\boldsymbol{y}\|_1 = 2\}.$$

Clearly, G is strongly connected

Now we define the joint process with state space $\Omega\times\Omega$. For any adjacent pair of states $(\boldsymbol{x},\boldsymbol{y})\in A$, the joint process does the following. Without loss of generality, we can assume that $x_1 = y_1+1, x_2 = y_2-1, x_3 = y_3,\ldots,x_n = y_n$. The transition of the joint process $(\boldsymbol{x},\boldsymbol{y})\mapsto(X',Y')$ is defined as follows.

Step 1: Pick a pair of mutually distinct indicies $\{i,j\}\in\{1,2,\ldots,n\}$ uniformly at random.

Step 2: For any index $i'\in\{1,2,\ldots,n\}\setminus\{i,j\}$, set $X'_{i'} = x_{i'}$, $Y'_{i'} = y_{i'}$. Pick $((X'_i,X'_j),(Y'_i,Y'_j))$ from the set $\Omega(x_i+x_j)\times\Omega(y_i+y_j)$ according to the following transition rule.

(Case 1) The case that the pair of indices $\{i,j\}$ picked at Step 1 satisfies $\{1,2\}\cap\{i,j\} = \emptyset$.

It is easy to see that the equality $x_i+x_j = y_i+y_j$ holds. At Step 2, we pick (X'_i,X'_j) according to the distribution function $f_{(x_i+x_j)}(X'_i,X'_j \mid u_i,u_j)$ and put $(Y'_i,Y'_j) = (X'_i,X'_j)$. Here we note that the pair of states satisfies $(X',Y')\in A$.

(Case 2) The case that the pair of indices $\{i,j\}$ picked at Step 1 satisfies $\{1,2\} = \{i,j\}$.

At Step 2, we pick (X',Y') in the same way with Case 1. In this case, the pair of states satisfies $X' = Y'$.

(Case 3) The case that the pair of indices $\{i,j\}$ picked at Step 1 satisfies $\{1,2\}\cap\{i,j\} = \{2\}$.

Without loss of generality, we can assume that $i=2$. Set $b = x_i+x_j$. Clearly, the equality $y_i+y_j = b+1$ holds. We introduce the distribution function defined on the set $\Omega(b)\times\Omega(b+1)$ which is used at Step 2 in this case. We define the set Ω' of states which may have positive probability by

$$\Omega' \stackrel{\text{def.}}{=} \{((1,b-1),(1,b)),\ldots,((b-1,1),(b-1,2))\} \cup \{((1,b-1),(2,b-1)),\ldots,((b-1,1),(b,1))\}.$$

We set $\Pr[((X'_i,X'_j),(Y'_i,Y'_j)) = ((x'_i,x'_j),(y'_i,h'_j))] = 0$, $\forall((x'_i,x'_j),(y'_i,y'_j))\in\Omega(b)\times\Omega(b+1)\setminus\Omega'$. For each element in Ω', the corresponding probability is defined

by

$$\Pr[((X_i', X_j'), (Y_i', Y_j')) = ((k, b-k), (k+1, b-k))]$$
$$= C_b \sum_{l=1}^{k} l^{u_i-1}(b-l)^{u_j-1} - C_{b+1} \sum_{l=1}^{k} l^{u_i-1}(b-l+1)^{u_j-1}$$

$$\Pr[((X_i', X_j'), (Y_i', Y_j')) = ((k, b-k), (k, b-k+1))]$$
$$= C_{b+1} \sum_{l=1}^{k} l^{u_i-1}(b-l+1)^{u_j-1} - C_b \sum_{l=1}^{k-1} l^{u_i-1}(b-l)^{u_j-1}$$

where $k \in \{1, 2, \ldots, b-1\}$ and $C_b = C(u_i, u_j, b)$, $C_{b+1} = C(u_i, u_j, b+1)$. (Here we note that for any sequence of real numbers $\{\kappa_l\}$, we define $\sum_{l=L}^{U} \kappa_l = 0$, if $L > U$.) Each pair of states $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega'$ satisfies that $(\boldsymbol{x}', \boldsymbol{y}') \in A$.

To complete the description of Case 3, we need to show that the above probability is non-negative and the sum total is equal to 1. It is easy to see that the sum total is equal to 1. The following lemma shows the non-negativity.

Lemma 2. *If the parameters u_i and u_j are non-negative, the inequalties*

$$\Pr[((X_i', X_j'), (Y_i', Y_j')) = ((k, b-k), (k+1, b-k))] \geq 0, \tag{1}$$

$$\Pr[((X_i', X_j'), (Y_i', Y_j')) = ((k, b-k), (k, b-k+1))] \geq 0, \tag{2}$$

hold for each $k \in \{1, 2, \ldots, b-1\}$.

The proof of the above lemma is complicated and described in Appendix. Here we note that when $u_i, u_j \geq 1$, the correspoding functions has log-concavity, and so we can show the non-negativity in an ordinary way. However, at least one of parameters is less than 1, the function is neither log-concave nor concave. If both parameters are less that 1, the corresponding function is convex and so we cannot apply the ordinary method to show the non-negativity of the transition probability of joint process. See Appendix for detail.

Next, we show that marginal distributions of the joint process is a faithful copy of the original Markov chain $\mathcal{M}$. Marginal distributions of X, Y satisfy that

$$\begin{aligned}
&\Pr\left[(X_i', X_j') = (k, b-k) \text{ and } (Y_i', Y_j') \in \Omega(b+1)\right] \\
&\quad = \Pr\Big[((X_i', X_j'), (Y_i', Y_j')) \\
&\qquad = ((k, b-k), (k+1, b-k))\Big] \\
&\quad + \Pr\Big[((X_i', X_j'), (Y_i', Y_j')) \\
&\qquad = ((k, b-k), (k, b-k+1))\Big] \\
&\quad = C_b k^{u_i-1}(b-k)^{u_j-1},
\end{aligned}$$

$$\begin{aligned}
&\Pr\left[(X_i', X_j') \in \Omega(b) \text{ and } (Y_i', Y_j') = (k, b-k+1)\right] \\
&\quad = \Pr\Big[((X_i', X_j'), (Y_i', Y_j')) \\
&\qquad = ((k-1, b-k+1), (k, b-k+1))\Big] \\
&\quad + \Pr\Big[((X_i', X_j'), (Y_i', Y_j')) \\
&\qquad = ((k, b-k), (k, b-k+1))\Big] \\
&\quad = C_{b+1} k^{u_i-1}(b-k+1)^{u_j-1}.
\end{aligned}$$

Lastly, we note that the pair of picked states satisfies that $(X', Y') \in A$.

(Case 4) The case that the pair of indices $\{i, j\}$ picked at Step 1 satisfies $\{1, 2\} \cap \{i, j\} = \{1\}$.

We choose $(X', Y') \in \Omega(b+1) \times \Omega(b)$ where $b = y_i + y_j$ in a similar way as Case 3. The procedure is obtained by substituting the indices 1 and 2, and states $\boldsymbol{x}$ and $\boldsymbol{y}$ simultaneously in Case 3. In this case, the picked pair of states also satisfies that $(X', Y') \in A$.

Now we completed the description of the transition procedure of joint process. In the rest of this section, we show a proof of the theorem.

Proof of Theorem 1. For any pair of states $(\boldsymbol{x}, \boldsymbol{y}) \in A$ adjacent on the graph G define above, we put the length of the edge is equal to 1. Then the distance from a state $\boldsymbol{x}' \in \Omega$ to $\boldsymbol{y}' \in \Omega$, denoted by $\mathrm{d}(\boldsymbol{x}', \boldsymbol{y}')$, is equal to the length of the shortest path on G from $\boldsymbol{x}'$ to $\boldsymbol{y}'$ where the length of the path is equal to the number of edges contained in the path. For any state $\boldsymbol{x} \in \Omega$, we define $\mathrm{d}(\boldsymbol{x}, \boldsymbol{x}) = 0$. It is clear that the diameter of the graph G, the distance between a farthest pair of vertices, is equal to $\Delta - n$.

Next, we estimate the expectation of the distance from X' to Y' obtained by applying the transition procedure of the joint process to an adjacent pair of states $(\boldsymbol{x}, \boldsymbol{y}) \in A$. Without loss of generality, we can assume that the pair $(\boldsymbol{x}, \boldsymbol{y})$ satisfies that $x_1 = y_1 + 1, x_2 = y_2 - 1, x_3 = y_3, \ldots, x_n = y_n$.

In (Case 1),(Case 3) and (Case 4), the distance from X' to Y' is equal to 1. When (Case 2) occured, the distance from X' to Y' decreases to 0. Since the probability of the event that (Case 2) is selected is equal to $2/(n(n-1))$, the expectation of the distance $\mathrm{E}[\mathrm{d}(X', Y')]$ becomes to $1 - 2/(n(n-1))$.

Path coupling theorem [1, 2] shows that the mixing time $\tau(\varepsilon)$ satisfies $\tau(\varepsilon) \leq (1/2)n(n-1)\ln((\Delta - n)\varepsilon^{-1})$. □

4 Experimental study

In this section, we show some simulation results. The overview of the simulations is that we run the Markov chain a number of times and compare the proportion of occurrence with the stational distribution for each vector at each transition step of Markov chain.

(u_1, u_2, u_3, u_4)	maximum difference of statistic	Δ		
		10	50	100
$(1, 1, 1, 1)$	$\lvert \mathrm{E}_\Delta[P_i] - \mathrm{E}[P_i] \rvert$	0	0	0
	$\lvert \mathrm{Var}_\Delta[P_i] - \mathrm{Var}[P_i] \rvert$	0.015	0.003	0.0015
	$\lvert \mathrm{Cov}_\Delta[P_i, P_j] - \mathrm{Cov}[P_i, P_j] \rvert$	0.005	0.001	0.0005
$(4, 3, 2, 1)$	$\max(\lvert \mathrm{E}_\Delta[P_i] - \mathrm{E}[P_i] \rvert)$	0.051	0.0092	0.0046
	$\max(\lvert \mathrm{Var}_\Delta[P_i] - \mathrm{Var}[P_i] \rvert)$	0.0036	0.00049	0.00023
	$\max\left(\lvert \mathrm{Cov}_\Delta[P_i, P_j] - \mathrm{Cov}[P_i, P_j] \rvert\right)$	0.0080	0.0074	0.0073
$(0.1, 0.1, 0.1, 0.1)$	$\lvert \mathrm{E}_\Delta[P_i] - \mathrm{E}[P_i] \rvert$	0	0	0
	$\lvert \mathrm{Var}_\Delta[P_i] - \mathrm{Var}[P_i] \rvert$	0.11	0.071	0.061
	$\lvert \mathrm{Cov}_\Delta[P_i, P_j] - \mathrm{Cov}[P_i, P_j] \rvert$	0.035	0.024	0.020
$(0.4, 0.3, 0.2, 0.1)$	$\max(\lvert \mathrm{E}_\Delta[P_i] - \mathrm{E}[P_i] \rvert)$	0.13	0.10	0.092
	$\max(\lvert \mathrm{Var}_\Delta[P_i] - \mathrm{Var}[P_i] \rvert)$	0.090	0.055	0.045
	$\max\left(\lvert \mathrm{Cov}_\Delta[P_i, P_j] - \mathrm{Cov}[P_i, P_j] \rvert\right)$	0.051	0.042	0.040
$(2, 1.5, 1, 0.5)$	$\max(\lvert \mathrm{E}_\Delta[P_i] - \mathrm{E}[P_i] \rvert)$	0.079	0.029	0.019
	$\max(\lvert \mathrm{Var}_\Delta[P_i] - \mathrm{Var}[P_i] \rvert)$	0.014	0.0032	0.0019
	$\max\left(\lvert \mathrm{Cov}_\Delta[P_i, P_j] - \mathrm{Cov}[P_i, P_j] \rvert\right)$	0.015	0.013	0.013

Table 1: the difference of statistics

We note that the stational distribution of Markov chain is quite different from the original Dirichlet distribution because of discretizing. For example, we show some differences of statistics in Table 1. The statistics of Dirichlet distribution with parameters $(u_1, \ldots, u_n)$ is given as follows. Let $u_0 = \sum_i u_i$. For each i, the expectation of p_i, $\mathrm{E}[p_i]$, is u_i/u_0 and the variance $\mathrm{Var}[p_i]$ is $\frac{u_i(u_0-u_i)}{u_0^2(u_0+1)}$. For each i, j, the covariance between p_i and p_j, $\mathrm{Cov}[p_i, p_j]$, is given by $\frac{-u_i u_j}{u_0^2(u_0+1)}$. On the other hands, we can calculate the statistics, $\mathrm{E}_\Delta[p_i]$, $\mathrm{Var}_\Delta[p_i]$, and $\mathrm{Cov}_\Delta[p_i, p_j]$, of the discretized Dirichlet distribution by a brute force.

Here, we describe the settings of our simulations. Since the behavior of Markov chain depends on random numbers, it is important to choose a good pseudo-random generator. Through all simulations, we use Mersenne Twister[4] as a pseudo-random generator. We run these simulations on the PC Linux machine with following specifications.

Machine: Dell Precision 450

CPU: Intel Xeon 2.8GHz (FSB 533MHz) × 2

OS: RedHat Linux 8.0 (Kernel 2.4.18-14smp)

Memory: Dual channel PC2100 DDR SDRAM 2GByte

Compiler: Intel C++ Compiler 7.0

For each simulation, we ran 10^9 processes of our Markov chain with deterministically chosen random seed from the unique vector where each element is approximately $1/n$. For each Markov chain process, we executed 50 steps. The running time of 10^9 processes, i.e., 5×10^{10} steps, is between 10 hours and 30 hours.

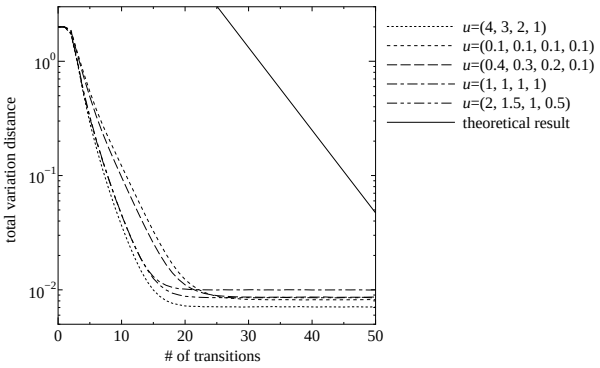


Figure 1: changing parameters of the Dirichlet

First, we show results on the relation between parameters and mixing rate. We fixed the dimension n to 4 and the discretizing grid size Δ to 100. We selected parameters from $(1, 1, 1, 1)$, $(4, 3, 2, 1)$, $(2, 1.5, 1, 0.5)$, $(0.1, 0.1, 0.1, 0.1)$, and $(0.4, 0.3, 0.2, 0.1)$. We note that the first one means the uniform distribution over Ω. In Figure 1, along the vertical axis we give the total variation distance ε, and the horizontal axis means the number of transition of chains from the unique initial state. As Figure 1 shows, the decrease of total variation distance are saturated at about 10^{-2}, though it must descend constantly. This is caused by the limitation of number of processes of Markov chains, that is, the difference of probability has lower bound for each vector. Thus, the larger number of executions we run, the smaller the difference will be. Aside from this saturation, we can see that if the value of a parameter is greater than or equal to 1, the mixing rate is less than the case that all values of a parameter are less than 1.

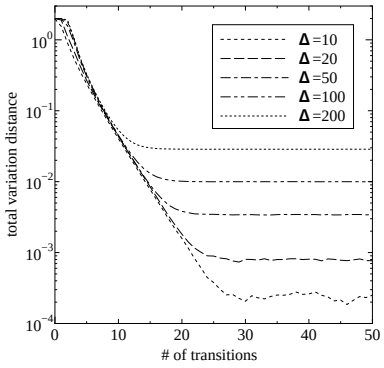


Figure 2: relation between Δ and the mixing rate

Next, we confirm how the discretizing value Δ contribute to the mixing rate. We fixed the dimension n to 4 again, the parameter to $(1, 1, 1, 1)$. We chose Δ from 10, 20, 50, 100, and 200. In Figure 2, we plotted the total variation distance ε for each discretizing grid size Δ. This figure shows that Δ will have little contribution to the mixing rate. More specifically, until the decrease of ε is saturated, the ratios of decreasing have little difference for each Δ. In theoretical result, $\log(\Delta - n)$ is caused by the diameter of the $G = (\Omega, A)$, which is artificially introduced to estimate the upper bound of the mixing rate. These experimental results, however, suggest us that the diameter does not depend on Δ. This suggestion is substantiated by the fact that the diameter of our chain is bounded by n.

Finally, we checked the relation between the dimension and the mixing rate. Because of restriction of memory, we fixed the discretizing grid size Δ to 20 and chose the dimension n between 3 and 7. We also fixed each parameter to 1. We show all results in Figure 3(a). Since our purpose is compare the mixing rate and dimension, we picked up the first step of transition that the total variation distance ε exceed 0.1, 0.5, 0.05, and 0.01. These picked points are marked in Figure 3(a). In Figure 3(b), we show the results for each ε. Though accurate consideration cannot be made because of the insufficient range of dimension, our results indicate that the mixing rate is $\Theta(n)$ rather than $\Theta(n^2)$.

5 Conclusion

In this paper, we proposed a Markov chain whose stationary distribution is a discretized Dirichlet distribution function. We showed that our Markov chain is rapidly mixing by using path coupling method. Our simulations indicates that the mixing time of the chain is much smaller than our theoretical upper bound.

References

[1] R. Bubley and M. Dyer, Path coupling: A technique for proving rapid mixing in Markov chains, 38th Annual Symposium on Foundations of Computer Science IEEE, San Alimitos, (1997) 223–231.

[2] R. Bubley Randomized Algorithms : Approximation, Generation, and Counting, Springer-Verlag, New York, 2001.

[3] R. Durbin, R. Eddy, A. Krogh, and G. Mitchison. Biological sequence analysis: probabilistic models of proteins and nucleic acids, Cambridge Univ. Press, 1998.

[4] Mersenne Twister Home Page, http://www.math.keio.ac.jp/~matumoto/mt.html

[5] T. Niu, Z. S. Qin, X. Xu, and J. S. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms, Am. J. Hum. Genet. 70:157–169, 2002.

[6] J. K. Pritchard, M. Stephens, and P. Donnely. Inference of population structure using multilocus genotype data, Genetics, 155:945–959, 2000.

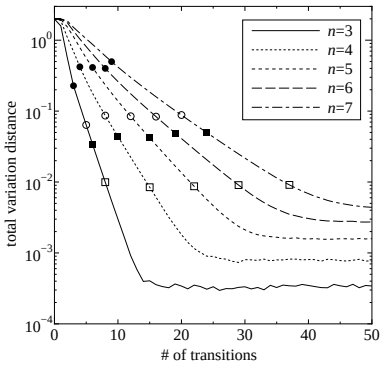
Appendix

Proof of Lemma 2. The inequalities (1) and (2) are symmetric in terms of u_i and u_j, we only need to show one of the inequalities. In the following, we discuss the ineaulity

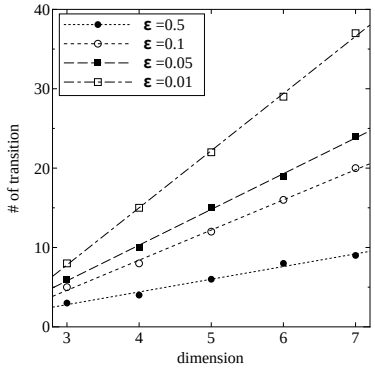
$$\Pr\left[((X'_i, X'_j), (Y'_i, Y'_j)) = ((k, b-k), (k, b-k+1))\right] \geq 0.$$

From the definition of the transition probability of the joint process, we have

$$\begin{aligned}
&\Pr\left[((X'_i, X'_j), (Y'_i, Y'_j)) = ((k, b-k), (k, b-k+1))\right] \\
&= C_{b+1} \sum_{l=1}^{k} l^{u_i-1}(b-l+1)^{u_j-1} \\
&\quad - C_b \sum_{l=1}^{k-1} l^{u_i-1}(b-l)^{u_j-1} \\
&= \left(1 - C_{b+1} \sum_{l=k+1}^{b} l^{u_i-1}(b-l+1)^{u_j-1}\right) \\
&\quad - \left(1 - C_b \sum_{l=k}^{b-1} l^{u_i-1}(b-l)^{u_j-1}\right) \\
&= \sum_{l=k+1}^{b} C_b l^{u_i-1}(b-l+1)^{u_j-1}\left(\left(1-\frac{1}{l}\right)^{u_i-1} - \frac{C_{b+1}}{C_b}\right).
\end{aligned}$$



(a) # of transitions v.s. total variation distance



(b) dimension v.s. mixing rate

Figure 3: relation between dimension and mixing rate

Thus we can show that

$$\begin{aligned}
&\Pr\left[((X_i',X_j'),(Y_i',Y_j')) = ((k,b-k),(k,b-k+1))\right] \\
=&C_{b+1}\sum_{l=1}^{k} l^{u_i-1}(b-l+1)^{u_j-1} - C_b\sum_{l=1}^{k-1} l^{u_i-1}(b-l)^{u_j-1} \\
\geq&C_{b+1}\sum_{l=2}^{k} l^{u_i-1}(b-l+1)^{u_j-1} \\
&- C_b\sum_{l=2}^{k}(l-1)^{u_i-1}(b-l+1)^{u_j-1} \\
=&\sum_{l=2}^{k} C_b l^{u_i-1}(b-l+1)^{u_j-1}\left(\frac{C_{b+1}}{C_b} - \left(1-\frac{1}{l}\right)^{u_i-1}\right).
\end{aligned}$$

By introducing the function $h : \{2,3,\ldots,b\} \to \mathrm{R}$ defined by $h(l) = \left(1-\frac{1}{l}\right)^{u_i-1} - \frac{C_{b+1}}{C_b}$, we can show that

$$\begin{aligned}
\Pr\left[((X_i',X_j'),(Y_i',Y_j')) = ((k,b-k),(k,b-k+1))\right] \\
= \sum_{l=k+1}^{b} C_b l^{u_i-1}(b-l+1)^{u_j-1}h(l) \quad (3) \\
\geq -\sum_{l=2}^{k} C_b l^{u_i-1}(b-l+1)^{u_j-1}h(l). \quad (4)
\end{aligned}$$

(a) The case that $u_i \geq 1$.

Since $u_i - 1 \geq 0$, the function $h_1(l)$ is monotone non-decreasing. When $h(k) \geq 0$ holds, we have $0 \leq h(k) \leq h(k+1) \leq \cdots \leq h(b)$, and so (3) implies the non-negativity

$$\begin{aligned}
\Pr\left[((X_i',X_j'),(Y_i',Y_j')) = ((k,b-k),(k,b-k+1))\right] \\
= \sum_{l=k+1}^{b} C_b l^{u_i-1}(b-l+1)^{u_j-1}h(l) \geq 0.
\end{aligned}$$

If $h(k) < 0$, then inequalities $h(2) \leq h(3) \leq \cdots \leq h(k) < 0$ hold, and so (4) implies that

$$\begin{aligned}
\Pr\left[((X_i',X_j'),(Y_i',Y_j')) = ((k,b-k),(k,b-k+1))\right] \\
\geq -\sum_{l=2}^{k} C_b l^{u_i-1}(b-l+1)^{u_j-1}h(l) \geq 0.
\end{aligned}$$

(b) The case that $0 \leq u_i \leq 1$.

Since $u_i - 1 \leq 0$, the function $h(l)$ is monotone non-increasing. If the inequality $h(b) \geq 0$ hold, we have $h(2) \geq h(3) \geq \cdots \geq h(b) \geq 0$ and inequality (3) implies the non-negativity

$$\begin{aligned}
\Pr\left[((X_i',X_j'),(Y_i',Y_j')) = ((k,b-k),(k,b-k+1))\right] \\
= \sum_{l=k+1}^{b} C_b l^{u_i-1}(b-l+1)^{u_j-1}h(l) \geq 0.
\end{aligned}$$

In the rest of this section, we show that $h(b) = (\frac{b-1}{b})^{u_i-1} - \frac{C_{b+1}}{C_b} \geq 0$.

We define a function $H_0(b,\alpha_i,\alpha_j)$ by

$$H_0(b,\alpha_i,\alpha_j) = (b-1)^{\alpha_i}C_{b+1}^{-1} - b^{\alpha_i}C_b^{-1}.$$

It is clear that if the condition $[-1 \leq \forall\alpha_i \leq 0,\ -1 \leq \forall\alpha_j,\ \forall b \in \{2,3,4,\ldots\},\ H_0(b,\alpha_i,\alpha_j) \geq 0]$ holds, we obatin the required result that $h(b) \geq 0$ for each $b \in \{2,3,4,\ldots\}$. Now we transform the function

$H_0(b, \alpha_i, \alpha_j)$ and obtain another expression as follows;

$$
\begin{aligned}
&H_0(b, \alpha_i, \alpha_j) \\
&= (b-1)^{\alpha_i} \sum_{k=1}^{b} k^{\alpha_i} (b-k+1)^{\alpha_j} - b^{\alpha_i} \sum_{k=1}^{b-1} k^{\alpha_i} (b-k)^{\alpha_j} \\
&= \sum_{k=1}^{b-1} \Bigg[(b-1)^{\alpha_i} k^{\alpha_i} (b-k+1)^{\alpha_j} \left(\frac{b-k}{b-1} \right) \\
&\qquad + (b-1)^{\alpha_i} (k+1)^{\alpha_i} (b-k)^{\alpha_j} \left(\frac{k}{b-1} \right) \\
&\qquad - b^{\alpha_i} k^{\alpha_i} (b-k)^{\alpha_j} \Bigg] \\
&= \sum_{k=1}^{b-1} \frac{(b-1)^{\alpha_i} k^{\alpha_i} (b-k)^{\alpha_j}}{b-1} \Bigg[\left(1 + \frac{1}{b-k} \right)^{\alpha_j} (b-k) \\
&\qquad\qquad + \left(1 + \frac{1}{k} \right)^{\alpha_i} k \\
&\qquad\qquad - \left(\frac{b}{b-1} \right)^{\alpha_i} (b-1) \Bigg].
\end{aligned}
$$

Then it is enough to show that the function

$$
\begin{aligned}
H_1(b, \alpha_i, \alpha_j, k) &\stackrel{\text{def.}}{=} \left(1 + \frac{1}{b-k} \right)^{\alpha_j} (b-k) \\
&\quad + \left(1 + \frac{1}{k} \right)^{\alpha_i} k - \left(\frac{b}{b-1} \right)^{\alpha_i} (b-1)
\end{aligned}
$$

is nonnegative for any $k \in \{1, 2, \ldots, k-1\}$. Since $1 + 1/(b-k) > 1$ and $\alpha_j \geq -1$, we have

$$
\begin{aligned}
&H_1(b, \alpha_i, \alpha_j, k) \geq H_1(b, \alpha_i, -1, k) \\
&\quad = \frac{(b-k)^2}{b-k+1} + \left(1 + \frac{1}{k} \right)^{\alpha_i} k - \left(\frac{b}{b-1} \right)^{\alpha_i} (b-1).
\end{aligned}
$$

We differentiate the function H_1 by α_i, we obtain the following

$$
\begin{aligned}
&\frac{\partial}{\partial \alpha_i} H_1(b, \alpha_i, -1, k) \\
&\qquad = \left(1 + \frac{1}{k} \right)^{\alpha_i} k \log \left(1 + \frac{1}{k} \right) \\
&\qquad\quad - \left(\frac{b}{b-1} \right)^{\alpha_i} (b-1) \log \left(\frac{b}{b-1} \right) \\
&\qquad = \left(1 + \frac{1}{k} \right)^{\alpha_i} \log \left(1 + \frac{1}{k} \right)^{k} \\
&\qquad\quad - \left(1 + \frac{1}{b-1} \right)^{\alpha_i} \log \left(1 + \frac{1}{b-1} \right)^{(b-1)}.
\end{aligned}
$$

Since k, b is a pair of positive integers satisfying $1 \leq k \leq b-1$, the non-positivity of α_i implies $0 \leq (1 + 1/k)^{\alpha_i} \leq (1 + 1/(b-1))^{\alpha_i}$ and $0 \leq \log(1 + 1/k)^k \leq \log(1 + 1/(b-1))^{b-1}$. Thus the function $H_1(b, \alpha_i, -1, k)$ is monotone non-decreasing with respest to $\alpha_i \leq 0$. Thus we have

$$
\begin{aligned}
&H_1(b, \alpha_i, -1, k) \\
&\qquad \geq H_1(b, 0, -1, k) \\
&\qquad = \frac{(b-k)^2}{b-k+1} + \left(1 + \frac{1}{k} \right)^0 k - \left(\frac{b}{b-1} \right)^0 (b-1) \\
&\qquad = \frac{1}{b-k+1} \geq 0.
\end{aligned}
$$

□