Polynomial Time Approximate Sampler for Discretized Dirichlet Distribution

Tomomi Matsui[1], Mitsuo Motoki[2], and Naoyuki Kamatani[3]

[1] 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/

[2] 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

[3] 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 vector 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 bounded by $(1/2)n(n-1)\ln((\Delta-n)\varepsilon^{-1})$ where n is the dimension (the number of parameters), $1/\Delta$ is the grid size for discretization, and ε is the error bound. Thus the obtained bound does not depend on the magnitudes of parameters. We estimate the mixing time by using the path coupling method. When the magnitudes of parameters are large, the log-concavity of the density function implies the rapidity straightforwardly. In the case that parameters are less than 1, the density function is convex and so we need a specified 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 is a distribution over vectors of positive numbers in which the sum total is equal to 1. The distribution often appears as prior and posterior distribution for the multinomial distribution in these methods since the Dirichlet distribution is the conjugate prior of parameters of the multinomial distribution [6].

For example, Niu, Qin, Xu, and Liu proposed a Bayesian haplotype inference method [4], which decides 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 the Dirichlet distribution with some appropriate parameters, parameters of the multinomial distribution is updated from the posterior distribution which is the Dirichlet distribution with updated parameters conditional on the "imputed" events.

Another example is a population structure inferring algorithm by Pritchard, Stephens, and Donnely [5]. Their algorithm is based on MCMC method. 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 parameters. One approach of sampling from the Dirichlet distribution is by rejection (see [3] for example). In this way, the number of required samples from the gamma distribution is equal to the size of the dimension of the 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 paper, we propose a simple Markov chain for sampling a random vector distributed according to a discretized Dirichlet distribution. We show that our Markov chain is rapidly mixing. More precisely, 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 note that this mixing time does not depend on the magnitudes of parameters. We also show experimentally that the required number of steps of our Markov chain is much smaller than our theoretical upper bound of the mixing time.

2 Markov Chain for Approximate Sampler

Dirichlet random vector $P = (P_1, P_2, \ldots, P_n)$ with non-negative 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 $\{(p_1, p_2, \ldots, p_n) \in \mathrm{R}^n \mid p_1 + \cdots + p_n = 1, p_1, p_2, \ldots, p_n > 0\}$ where $\Gamma(u)$ is the gamma function. Throughout this paper, we assume that $n \geq 2$.

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

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

A discretized Dirichlet random vector with non-negative 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 non-negative 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 $t \in \{0, 1, 2, \ldots\}$, transition $X^t \mapsto X^{t+1}$ takes 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 irreducible 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 an upper bound of 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 theorem.

Before showing the above, we discuss the influence of discretization. The stationary distribution of our chain is different from the original Dirichlet distribution because of the discretization. The statistics of the Dirichlet distribution with parameters $(u_1, \ldots, u_n)$ are given as follows. For each random variable P_i, $\mathrm{E}[P_i] = u_i/u_0$ and $\mathrm{Var}[P_i] = \frac{u_i(u_0 - u_i)}{u_0^2(u_0+1)}$ where $u_0 = \sum_i u_i$. For each pair of random variables P_i and P_j with $(i \neq j)$, $\mathrm{Cov}[P_i, P_j] = \frac{-u_i u_j}{u_0^2(u_0+1)}$. For some discretized Dirichlet distributions, we calculated the statistics, $\mathrm{E}_\Delta[P_i]$, $\mathrm{Var}_\Delta[P_i]$, and $\mathrm{Cov}_\Delta[P_i, P_j]$ by a brute force method. Table 1 shows the results.

In the rest of this section, we briefly review the definition of the mixing time and path coupling method. For any probability distribution function π' on Ω,

Table 1. Influence of discretization.

(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
	$\max(\lvert \mathrm{Cov}_\Delta[P_i, P_j] - \mathrm{Cov}[P_i, P_j] \rvert)$	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
	$\max(\lvert \mathrm{Cov}_\Delta[P_i, P_j] - \mathrm{Cov}[P_i, P_j] \rvert)$	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
	$\max(\lvert \mathrm{Cov}_\Delta[P_i, P_j] - \mathrm{Cov}[P_i, P_j] \rvert)$	0.015	0.013	0.013

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

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

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

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

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.

Path coupling lemma [Bubley and Dyer [1]]

Let G be a directed graph with vertex set Ω and arc set $A \subseteq \Omega \times \Omega$. Let $\ell : A \to \mathrm{Z}_{++}$ be a positive integer 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 a 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

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

Then the mixing time $\tau(\varepsilon)$ of the original Markov chain $\mathcal{M}$ satisfies $\tau(\varepsilon) \leq (1-\beta)^{-1} \ln(D/\varepsilon)$ 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 graph $G = (\Omega, A)$ whose vertex set is equivalent to the state space Ω. There exists a directed arc from a 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 indices $\{i, j\} \in \{1, 2, \ldots, n\}$ 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 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 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 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

$$\Omega' \stackrel{\text{def.}}{=} \left\{ \begin{array}{ll} ((1, b-1), (1, b)), & ((2, b-2), (2, b-1)), \cdots, ((b-1, 1), (b-1, 2)), \\ ((1, b-1), (2, b-1)), & ((2, b-2), (3, b-2)), \cdots, ((b-1, 1), (b, 1)) \end{array} \right\}.$$

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

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

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 1 *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 corresponding functions have 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 an 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 are faithful copy of the original chain $\mathcal{M}$. Marginal distributions of X', Y' satisfy that

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

$$
\begin{aligned}
&+C_{b+1}\sum_{l=1}^{k} l^{u_i-1}(b-l+1)^{u_j-1} - C_b \sum_{l=1}^{k-1} l^{u_i-1}(b-l)^{u_j-1} \\
&= C_{b+1}\sum_{l=1}^{k} l^{u_i-1}(b-l+1)^{u_j-1} - C_{b+1} \sum_{l=1}^{k-1} l^{u_i-1}(b-l+1)^{u_j-1} \\
&= C_{b+1} k^{u_i-1}(b-k+1)^{u_j-1}.
\end{aligned}
$$

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

(Case 4) The 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 $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega^2$ adjacent on the graph G define above, i.e., $(\boldsymbol{x}, \boldsymbol{y}) \in A$, we put the length of the edge $(\boldsymbol{x}, \boldsymbol{y})$ 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 a shortest path on G from $\boldsymbol{x}'$ to $\boldsymbol{y}'$ where the length of a 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 Cases 1, 3 and 4, the distance from X' to Y' is equal to 1. When Case 2 occurred, 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 settings of our simulations are as follows. Through all simulations, we use Mersenne Twister[7] 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. For each Markov chain process, we chose a random seed deterministically and transitions are executed 50 steps. The initial state is an integer vector in Ω obtained by rounding $(\Delta/n, \cdots, \Delta/n)$. The running time of 10^9 processes, i.e., 5×10^{10} steps, is between 10$\sim$30 hours.

First, we show results on the relation between parameters and mixing time. We fixed the dimension n to 4 and the discretizing grid size $1/\Delta$ to 1/100. 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 case $(1,1,1,1)$ corresponds to the uniform distribution over Ω. In Fig. 1, along the vertical axis we give the total variation distance ε, and the horizontal axis means the number of transitions of chains from the initial state. As Fig. 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 the number of samples (10^9) from Markov chains, that is, the total variation distance has a positive lower bound for each vector of parameters. Fig. 1 shows that 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 time is less than the case that all values of a parameter are less than 1.

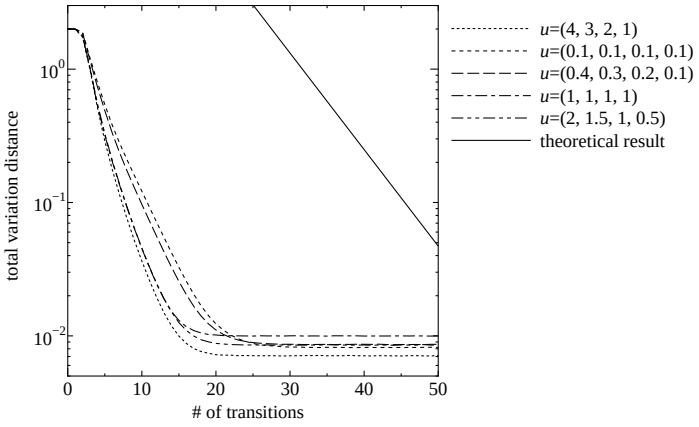


Fig. 1. Relation between mixing time and the magnitude of parameters.

Next, we confirm how the discretizing value Δ contribute to the mixing time. We fixed the dimension n to 4 again and the parameter to $(1,1,1,1)$. We chose Δ from 10, 20, 50, 100, and 200. In Fig. 2, we plotted the total variation distance ε for each Δ. This figure shows that Δ will have little contribution to the mixing time. More specifically, until the decrease of ε is saturated, the ratios of decreasing have little difference for each Δ. In the proof of Theorem 1, the term $(\Delta - n)$ is artificially introduced as the diameter of the graph $G = (\Omega, A)$. These experimental results, however, suggest that the mixing time does not depend on

Δ. This property is substantiated by the fact that the diameter of our chain is bounded by n and independent of Δ.

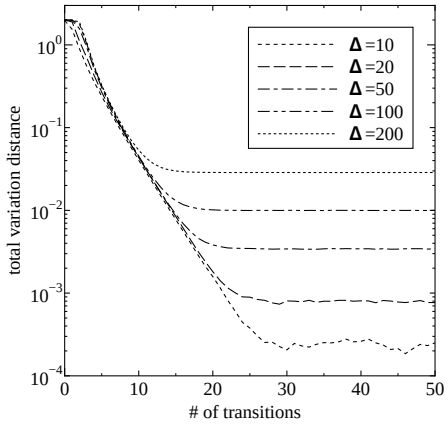
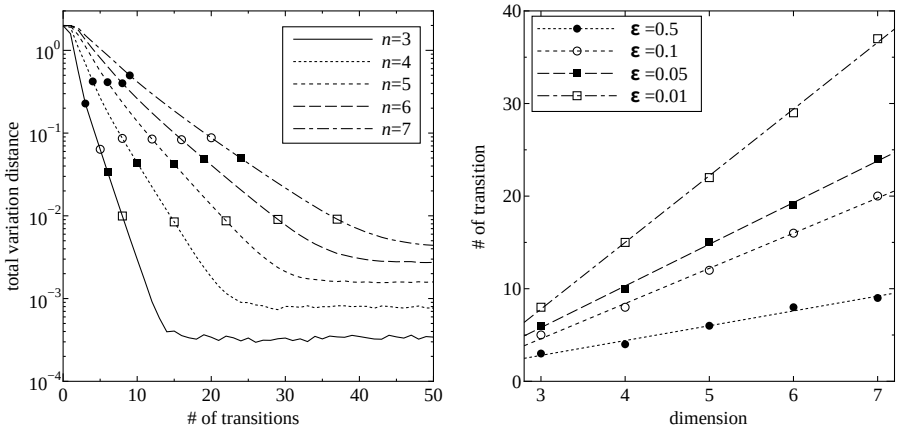


Fig. 2. Relation between Δ and the mixing time.



(a): Dimension and total variation distance. (b): Dimension and mixing time.

Fig. 3. Relation between number of transitions and total variation distance.

Finally, we checked the relation between the dimension n and the mixing time. Because of restriction of the memory, we fixed the discretizing grid size $1/\Delta$ to $1/20$ and chose the dimension n between 3 and 7. We also fixed each parameter to 1. We show all results in Fig. 3. Since our purpose is to compare the mixing time and dimension, we picked up the first time instance that the total variation distance ε exceeds 0.1, 0.5, 0.05, and/or 0.01. These picked time instances are marked in Fig. 3. In Fig. 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 time 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. We showed that our Markov chain is rapidly mixing by using coupling method. Our upper bound of the mixing time does not depend on the magnitudes of parameters. When parameters are less than 1, the corresponding density function is convex and so, ordinary technique related to log-concavity is not applicable. We have shown the required property in Appendix. Our computational experiences indicates that the mixing time of the chain is much smaller than our theoretical upper bound.

Appendix

Proof of Lemma 1
Since 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 inequality

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

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

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

Similarly, we can also show that

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

By introducing the function $h : \{2, 3, \ldots, b\} \to \mathrm{R}$ defined by $h(l) \stackrel{\mathrm{def.}}{=} \left(1 - \frac{1}{l}\right)^{u_i - 1} - \frac{C_{b+1}}{C_b}$, we have the following equality and inequality

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

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

Since $u_i - 1 \geq 0$, the function $h(l)$ 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

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

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

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

(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

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

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;

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

Then it is enough to show that the function

$$H_1(b,\alpha_i,\alpha_j,k) \stackrel{\text{def.}}{=} \left(1+\tfrac{1}{b-k}\right)^{\alpha_j}(b-k) + \left(1+\tfrac{1}{k}\right)^{\alpha_i} k - \left(\tfrac{b}{b-1}\right)^{\alpha_i}(b-1)$$

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

$$H_1(b,\alpha_i,\alpha_j,k) \geq H_1(b,\alpha_i,-1,k) = \tfrac{(b-k)^2}{b-k+1} + \left(1+\tfrac{1}{k}\right)^{\alpha_i} k - \left(\tfrac{b}{b-1}\right)^{\alpha_i}(b-1).$$

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

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

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 respect to $\alpha_i \leq 0$. Thus we have

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

□

References

1. Bubley, R., M. Dyer, M.: 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. Bubley, R.: Randomized Algorithms : Approximation, Generation, and Counting, Springer-Verlag, New York, 2001.
3. Durbin, R., Eddy, R., Krogh, A., Mitchison, G.: Biological sequence analysis: probabilistic models of proteins and nucleic acids, Cambridge Univ. Press, 1998.
4. Niu, T., Qin, Z. S., Xu, X., Liu, J. S.: Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms, Am. J. Hum. Genet., **70** (2002) 157–169.
5. Pritchard, J. K., Stephens, M., Donnely, P.: Inference of population structure using multilocus genotype data, Genetics, **155** (2000) 945–959.
6. Robert, C. P.: The Bayesian Choice, Springer-Verlag, New York, 2001.
7. Mersenne Twister Home Page, http://www.math.keio.ac.jp/~matumoto/mt.html