

Speeding Up Markov Chain Monte Carlo Algorithms

András Faragó

Department of Computer Science
The University of Texas at Dallas
Richardson, Texas
E-mail: farago@utdallas.edu

Abstract

We prove an upper bound on the convergence rate of Markov Chain Monte Carlo (MCMC) algorithms for the important special case when the state space can be aggregated into a smaller space, such that the aggregated chain approximately preserves the Markov property.

Keywords: *Markov Chain Monte Carlo, convergence rate, lumpable and quasi-lumpable Markov chain.*

I Introduction

The Markov Chain Monte Carlo (MCMC) method is often used to solve hard counting, sampling and optimization problems in a number of areas, including physics, combinatorial optimization, computational biology and many others. The success and influence of the method is shown by the fact that one of its variants, the celebrated Metropolis algorithm has been selected as one of the top 10 of all algorithms [1].

In this paper we consider the following frequently occurring application of the MCMC method. Let S be a very large set and A be a subset of it. We would like to estimate the relative size of A , that is, the goal is to obtain

a good estimate of the value

$$p = \frac{|A|}{|S|}.$$

If we can take samples from S uniformly at random, then an obvious estimate with good properties is the relative frequency of the event that the sample falls in A . Unfortunately, in most nontrivial cases of interest, sampling uniformly at random from S is not feasible. The reason is that the large set S is defined *implicitly*. Examples are the set of all matchings in a graph or the set of all feasible solutions to a knapsack problem etc. No efficient method is known to sample *uniformly* at random from such sets. An important practical application is to estimate blocking probabilities in telecommunication networks, see [4, 7].

At this point the MCMC does a very good service. If we define a Markov chain in which the states are the elements of S and the transitions are based on simple local operations, then we can very often obtain a Markov chain with uniform stationary distribution over S . Then, if we run this chain long enough so that it gets close to the stationary distribution, then the state where we stop the chain will be a good approximation of a uniformly distributed random sample over S .

The key difficulty is, however, that we should

run the chain long enough to get sufficiently close to the stationary distribution. This time is often referred to as *mixing time* [8]. If the mixing time grows only polynomially with the size of the problem, e.g. with the size of the graph, then we say that the chain is *rapidly mixing*. Unfortunately, in many cases of interest the mixing time grows exponentially with the problem parameters, so in many important cases the Markov chain is not rapidly mixing.

What we are interested in is whether it is possible to speed up the running time. It is clear that if we want to estimate the size of *any* possible subset, then we really need to get close to the stationary distribution, since only this distribution can guarantee that the probability of the random state falling in the set is really the relative size of the set. On the other hand, if we only want to estimate the relative size of a given subset A , then it is enough for us if we reach a distribution in which the measure of A is close to the stationary measure, but this does not have to hold for every other set. In other words, if π_t denotes the state distribution after t steps and π is the stationary distribution, then we want to choose t such that $|\pi_t(A) - \pi(A)|$ is small, but the same does not have to hold for all other sets. This makes it possible to reduce the required value of t , that is, to speed up the algorithm. In this paper we investigate under what conditions it is possible to obtain such a speed-up. The main result is that if the Markov chain is close to a so called *lumpable* chain, then the speedup is possible.

II Lumpable Markov Chains

We assume the reader is familiar with the fundamental concepts of Markov chains. We use the notation that a Markov chain \mathcal{M} is given by a set S of states and by a transition probability

matrix P , so we use the notation $\mathcal{M} = (S, P)$. We do not include the initial distribution, because it is assumed arbitrary.

An important concept is the *lumpability* of Markov chain. Informally, a chain is lumpable if its states can be aggregated into larger subsets of S , such that the aggregated (lumped) chain remains a Markov chain with respect to the set transition probabilities (i.e, it preserves the property that the future depends on the past only through the present. Let us present now the formal definition.

Definition 1 (Lumpable Markov chain)

Let $\mathcal{M} = (S, P)$ be a Markov chain. Let $\mathcal{Q} = \{A_1, \dots, A_m\}$ be a partition of S . The chain \mathcal{M} is called lumpable with respect to \mathcal{Q} if for any initial distribution the relationship

$$\Pr(X_t \in A_j \mid X_{t-1} \in A_{i_1}, \dots, X_{t-k} \in A_{i_k}) = \Pr(X_t \in A_j \mid X_{t-1} \in A_{i_1}) \quad (1)$$

holds for any t, k, j, i_1, \dots, i_k , whenever these conditional probabilities are defined (i.e., the conditions occur with positive probability).

A fundamental result on the lumpability of Markov chains is the following theorem, see [5], Theorem 6.3.2. We use the notation that $p(x, A)$ denotes the probability that the chain moves into a set $A \subseteq S$, given that it is in the state $x \in S$. Note that x itself may or may not be in the set A .

Theorem 1 (Necessary and sufficient condition for lumpability) A Markov chain $\mathcal{M} = (S, P)$ is lumpable with respect to a partition $\mathcal{Q} = \{A_1, \dots, A_m\}$ of S if and only if for any i, j the value of $p(x, A_j)$ is the same for every $x \in A_i$. These common values define the transition probabilities $\hat{p}(A_i, A_j)$ for the lumped

chain, which is a Markov chain with state set \mathcal{Q} and state transition probabilities

$$\hat{p}(A_i, A_j) = p(x, A_j) = \Pr(X_t \in A_j \mid X_{t-1} \in A_i)$$

where x is any state in A_i .

Whenever our Markov chain is lumpable, we can reduce the number of states by the above aggregation, and that is usually advantageous for faster convergence (a precise convergence bound will be proven in Section IV). In the next section we relax the concept of lumpability to broaden the family of considered Markov chains.

III Quasi-Lumpable Markov Chains

Informally, a Markov chain is called quasi-lumpable or ϵ -lumpable, if it may not be perfectly lumpable, but it is “close” to that, as defined formally below.

Definition 2 (ϵ -lumpability) Let $\epsilon \geq 0$. A Markov chain $\mathcal{M} = (S, P)$ is called ϵ -lumpable with respect to a partition $\mathcal{Q} = \{A_1, \dots, A_m\}$ of S if

$$|p(x, A_j) - p(y, A_j)| \leq \epsilon$$

holds for any $x, y \in A_i$ and for any i, j .

Note that if we take $\epsilon = 0$, then we get back the ordinary concept of lumpability. Thus, quasi-lumpability is indeed a relaxation of the original concept. It can also be interpreted in the following way. If $\epsilon > 0$, then the original definition of lumpability may not hold. This means, the aggregated process may not remain Markov. i.e., it does not satisfy (1). On the other hand, if ϵ is small, then the aggregated process will be close to a Markov chain.

What we are interested in is the convergence analysis of quasi lumpable Markov chains. Specifically, we would like to derive a convergence bound for the aggregated process, which may not be Markov, but, in the above sense, it is not far from being Markov. For the analysis this we need to introduce another definition.

Definition 3 (Lower and upper transition matrices) Let $\mathcal{M} = (S, P)$ be a Markov chain which is ϵ -lumpable with respect to a partition $\mathcal{Q} = \{A_1, \dots, A_m\}$. The lower and upper transition matrices $L = [l_{ij}]$ and $U = [u_{ij}]$ are defined as $m \times m$ matrices with entries

$$l_{ij} = \min_{x \in A_i} p(x, A_j) \quad \text{and} \quad u_{ij} = \max_{x \in A_i} p(x, A_j),$$

respectively, for $i, j = 1, \dots, m$.

Note that it always holds (componentwise) that $L \leq P \leq U$. If the chain is lumpable, then these matrices coincide, so then $L = U = P$. If the chain is ϵ -lumpable, then L and U differ at most by ϵ in each entry.

Generally, L and U are not necessarily stochastic matrices¹, as their rows may not sum up to 1.

IV Convergence Analysis

An important concept in Markov chain convergence analysis is the *ergodic coefficient* or *coefficient of ergodicity*, see, e.g., [6].

Definition 4 Let $P = [p_{ij}]$ be an $n \times n$ stochastic matrix. Its ergodic coefficient is defined as

$$\rho(P) = \frac{1}{2} \max_{i,j} \sum_{k=1}^n |p_{ik} - p_{jk}|.$$

¹A vector is called stochastic if each coordinate is nonnegative and their sum is 1. A matrix is called stochastic if each row vector of it is stochastic.

The importance of the ergodic coefficient lies in its relationship to the convergence rate of the Markov chain. It is well known in Markov chain analysis that the convergence rate is determined by the second largest eigenvalue of the transition matrix (that is, the eigenvalue which has the largest absolute value less than 1), see, e.g., [8]. If this eigenvalue is denoted by λ_1 , then the convergence to stationarity happens at least at a rate of $O(\lambda_1^t)$, where t is the number of steps. It is also known [6] that the ergodic coefficient is always an upper bound on this eigenvalue, it satisfies $\lambda_1 \leq \rho(P) \leq 1$. Therefore, the distance to the stationary distribution is also bounded by $O(\rho(P)^t)$. Thus, the smaller is the ergodic coefficient, the faster is the convergence.

It is also worth mentioning that the ergodic coefficient also has a geometric interpretation. It is the half of the maximum L_1 distance that occurs among the row vectors of the transition matrix. Consequently, if we can reduce the number of states, then we have a better chance for fast convergence, since the maximum distance occurring among fewer vectors of lower dimension is likely to be smaller than the maximum distance among a large number of vectors of higher dimension.

Now we are ready to present the main result, which is a bound on how fast will an ϵ -lumpable Markov chain converge to its stationary distribution on the sets that are in the partition used for defining the ϵ -lumpability of the chain. We are going to discuss the applicability of the result in the next section.

Theorem 2 *Let $\epsilon \geq 0$ and $\mathcal{M} = (S, P)$ be an irreducible, aperiodic Markov chain which is ϵ -lumpable with respect to a partition $\mathcal{Q} = \{A_1, \dots, A_m\}$ of S . Let $\rho < 1$ be a common upper bound on the ergodic coefficients of all stochastic matrices M with $L \leq M \leq U$, where L, U are the lower and upper transition matrices,*

respectively. Let π_0 be any initial probability distribution on S , such that $P(X_t \in A_i) > 0$ for any i and $t = 0, 1, 2, \dots$. Then for every $t \geq 1$ the following estimation holds:

$$\sum_{i=1}^m |\pi_t(A_i) - \pi(A_i)| \leq 2\rho^t + \epsilon m \frac{1 - \rho^t}{1 - \rho}$$

where $\pi = \lim_{t \rightarrow \infty} \pi_t$ is the stationary distribution of \mathcal{M} .

For the proof we need a lemma of D.J. Hartfiel about stochastic vectors and matrices. (Lemma 3.4 on p. 70 in [2], see also [3]):

Lemma 1 (Hartfiel [2, 3]) *Let x, y be n -dimensional stochastic vectors and*

$$B_1, \dots, B_k; C_1, \dots, C_k$$

be $n \times n$ stochastic matrices. If $\rho(B_i) \leq \rho_0$ and $\rho(C_i) \leq \rho_0$ for all i , $1 \leq i \leq k$, then

$$\|xB_1 \dots B_k - yC_1 \dots C_k\| \leq \rho_0^k \|x - y\| + (\rho_0^{k-1} + \dots + 1)\mathcal{E}$$

where $\mathcal{E} = \max_i \|B_i - C_i\|$. The vector norm used is the L_1 -norm $\|x\| = \sum_{i=1}^n |x_i|$ and the matrix norm is

$$\|A\| = \sup_{z \neq 0} \frac{\|zA\|}{\|z\|} = \max_i \sum_{j=1}^n |a_{ij}|$$

for any $n \times n$ real matrix $A = [a_{ij}]$.

Lemma 1 can be proved via induction on k , see [2, 3]. Now, armed with the lemma, we can prove our theorem.

Proof of Theorem 2. Let π_0 be an initial state distribution of the Markov chain \mathcal{M} , let π_t be the corresponding distribution after t steps and $\pi = \lim_{t \rightarrow \infty} \pi_t$ be the stationary distribution of \mathcal{M} . For a set $A \subseteq S$ of states the usual notations $\pi_t(A) = P(X_t \in A)$, $\pi(A) = \lim_{t \rightarrow \infty} \pi_t(A)$ are adopted.

Using the sets A_1, \dots, A_m of the partition \mathcal{Q} , let us define the stochastic vectors

$$\tilde{\pi}_t = (\pi_t(A_1), \dots, \pi_t(A_m)) \quad (2)$$

for $t = 0, 1, 2, \dots$ and the $m \times m$ stochastic matrices

$$\tilde{P}_t(\pi_0) = [p_t^{(\pi_0)}(i, j)] = [\mathbb{P}(X_{t+1} \in A_j \mid X_t \in A_i)] \quad (3)$$

for $t = 1, 2, \dots$. Let us call them aggregated state distribution vectors and aggregated transition matrices, respectively. Note that although the entries in (3) involve only events of the form $\{X_t \in A_k\}$, they may also depend on the detailed state distribution within these sets, which is in turn determined by the initial distribution π_0 . In other words, if two different initial distributions give rise to the same probabilities for the events $\{X_t \in A_k\}$ for some t , they may still result in different conditional probabilities of the form $\mathbb{P}(X_{t+1} \in A_j \mid X_t \in A_i)$, since the chain is not assumed lumpable in the ordinary sense. This is why the notations $\tilde{P}_t(\pi_0)$, $p_t^{(\pi_0)}(i, j)$ are used. Also note that the conditional probabilities are well defined for any initial distribution allowed by the assumptions of the lemma, since then $\mathbb{P}(X_t \in A_i) > 0$.

For any fixed t the events $\{X_t \in A_i\}$, $i = 1, \dots, m$, are mutually exclusive with total probability 1, therefore, by the law of total probability,

$$\begin{aligned} \mathbb{P}(X_{t+1} \in A_j) &= \\ &= \sum_{i=1}^m \mathbb{P}(X_{t+1} \in A_j \mid X_t \in A_i) \mathbb{P}(X_t \in A_i) \end{aligned}$$

holds for every $j = 1, \dots, m$. This implies $\tilde{\pi}_{t+1} = \tilde{\pi}_t \tilde{P}_t(\pi_0)$, from which

$$\tilde{\pi}_t = \tilde{\pi}_0 \tilde{P}_1(\pi_0) \dots \tilde{P}_t(\pi_0) \quad (4)$$

follows.

We next show that for any $t = 1, 2, \dots$ the matrix $\tilde{P}_t(\pi_0)$ falls between the lower and upper

transition matrices, i.e., $L \leq \tilde{P}_t(\pi_0) \leq M$ holds. Let us use short notations for certain events: for any $i = 1, \dots, m$ and for a fixed $t \geq 1$ set $H_i = \{X_t \in A_i\}$, $H'_i = \{X_{t+1} \in A_i\}$, and for $x \in S$ let $E_x = \{X_t = x\}$. Then $E_x \cap E_y = \emptyset$ holds for any $x \neq y$ and $\sum_{x \in S} E_x = 1$. Applying the definition of conditional probability and the law of total probability, noting that $\mathbb{P}(H_i) > 0$ is provided by the assumptions of the lemma, we get

$$\begin{aligned} p_t^{(\pi_0)}(i, j) &= \mathbb{P}(H'_j \mid H_i) = \\ &= \frac{\mathbb{P}(H'_j \cap H_i)}{\mathbb{P}(H_i)} \\ &= \frac{\sum_{x \in S} \mathbb{P}(H'_j \cap H_i \cap E_x)}{\mathbb{P}(H_i)} \\ &= \frac{\sum_{x \in S} \mathbb{P}(H'_j \mid H_i \cap E_x) \mathbb{P}(H_i \cap E_x)}{\mathbb{P}(H_i)} \\ &= \sum_{x \in S} \mathbb{P}(H'_j \mid H_i \cap E_x) \frac{\mathbb{P}(H_i \cap E_x)}{\mathbb{P}(H_i)} \\ &= \sum_{x \in S} \mathbb{P}(H'_j \mid H_i \cap E_x) \mathbb{P}(E_x \mid H_i). \end{aligned}$$

Whenever $x \notin A_i$ we have $\mathbb{P}(E_x \mid H_i) = \mathbb{P}(X_t = x \mid X_t \in A_i) = 0$. Therefore, it is enough to take the summation over A_i , instead of the entire S . For $x \in A_i$, however, $H_i \cap E_x = \{X_t \in A_i\} \cap \{X_t = x\} = \{X_t = x\}$ holds, so we obtain

$$\begin{aligned} p_t^{(\pi_0)}(i, j) &= \\ &= \sum_{x \in A_i} \mathbb{P}(X_{t+1} \in A_j \mid X_t = x) \mathbb{P}(X_t = x \mid X_t \in A_i). \end{aligned}$$

Thus, $p_t^{(\pi_0)}(i, j)$ is a weighted average of the $\mathbb{P}(X_{t+1} \in A_j \mid X_t = x)$ probabilities. The weights are $\mathbb{P}(X_t = x \mid X_t \in A_i)$, so they are nonnegative and sum up to 1. Further,

$$l_{ij} \leq \mathbb{P}(X_{t+1} \in A_j \mid X_t = x) \leq u_{ij}$$

must hold, since l_{ij}, u_{ij} are defined as the minimum and maximum values, respectively, of

$$p(x, A_j) = \mathbb{P}(X_{t+1} \in A_j \mid X_t = x)$$

over $x \in A_i$. Since the weighted average must fall between the minimum and the maximum, therefore, we have

$$l_{ij} \leq p_t^{(\pi_0)}(i, j) \leq u_{ij}, \quad (5)$$

that is,

$$L \leq \tilde{P}_t(\pi_0) \leq M \quad (6)$$

for any $t \geq 1$ and for any initial distribution π_0 allowed by the conditions of the Lemma.

Let us now start the chain from an initial distribution π_0 that satisfies the conditions of the Lemma. We are going to compare the arising aggregated state distribution vectors (2) with the ones resulting from starting the chain from the stationary distribution π . Note that, due to the assumed irreducibility of the original chain, $\pi(x) > 0$ for all $x \in S$, so π is also a possible initial distribution that satisfies the conditions $P(X_t \in A_i) > 0$.

When the chain is started from the stationary distribution π , then, according to (4), the aggregated state distribution vector at time t is $\tilde{\pi} \tilde{P}_1(\pi) \dots \tilde{P}_t(\pi)$ where $\tilde{\pi} = (\pi(A_1), \dots, \pi(A_m))$. On the other hand, $P(X_t \in A_i)$ remains the same for all $t \geq 0$ if the chain starts from the stationary distribution. Therefore, we have

$$\tilde{\pi} \tilde{P}_1(\pi) \dots \tilde{P}_t(\pi) = \tilde{\pi} = (\pi(A_1), \dots, \pi(A_m)). \quad (7)$$

When the chain starts from π_0 , then we obtain the aggregated state distribution vector

$$\tilde{\pi}_t = \tilde{\pi}_0 \tilde{P}_1(\pi_0) \dots \tilde{P}_t(\pi_0) \quad (8)$$

after t steps. Now we can apply Lemma 1 for the comparison of (7) and (8). The roles for the quantities in Lemma 1 are assigned as $x = \tilde{\pi}_0$, $y = \tilde{\pi}$, $k = t$, $n = m$, $\rho_0 = \rho$, and, for every $\tau = 1, \dots, k$, $B_\tau = \tilde{P}_\tau(\pi_0)$, $C_\tau = \tilde{P}_\tau(\pi)$. Recall that by (6) we have $L \leq \tilde{P}_\tau(\pi_0) \leq M$ and $L \leq \tilde{P}_\tau(\pi) \leq M$, implying $\rho(\tilde{P}_\tau(\pi_0)) \leq \rho$ and $\rho(\tilde{P}_\tau(\pi)) \leq \rho$, where ρ is the common upper

bound on the ergodic coefficients of all stochastic matrices M with $L \leq M \leq U$, assumed in the conditions of the Lemma. With this role assignment we obtain from Lemma 1

$$\|\tilde{\pi}_0 \tilde{P}_1(\pi_0) \dots \tilde{P}_t(\pi_0) - \tilde{\pi} \tilde{P}_1(\pi) \dots \tilde{P}_t(\pi)\| \leq$$

$$\rho^t \|\tilde{\pi}_0 - \tilde{\pi}\| + \mathcal{E} \sum_{k=0}^{t-1} \rho^k$$

where $\mathcal{E} = \max_\tau \|P_\tau(\pi_0) - P_\tau(\pi)\|$ and the norms are as in Lemma 1. Taking (7) and (8) into account yields

$$\|\tilde{\pi}_t - \tilde{\pi}\| = \quad (9)$$

$$\sum_{i=1}^m |\pi_t(A_i) - \pi(A_i)| \leq \rho^t \|\tilde{\pi}_0 - \tilde{\pi}\| + \mathcal{E} \sum_{k=0}^{t-1} \rho^k.$$

Thus, it only remains to estimate $\|\tilde{\pi}_0 - \tilde{\pi}\|$ and \mathcal{E} . Given that $\tilde{\pi}_0, \tilde{\pi}$ are both stochastic vectors, we have $\|\tilde{\pi}_0 - \tilde{\pi}\| \leq \|\tilde{\pi}_0\| + \|\tilde{\pi}\| \leq 2$. Further,

$$\mathcal{E} = \max_\tau \|P_\tau(\pi_0) - P_\tau(\pi)\| =$$

$$\max_\tau \max_i \sum_{j=1}^m |p_\tau^{(\pi_0)}(i, j) - p_\tau^{(\pi)}(i, j)| \leq \epsilon m,$$

since (5) holds for any considered π_0 (including π), and, by the definition of ϵ -lumpability, $u_{ij} - l_{ij} \leq \epsilon$. Substituting the estimations into (9), we obtain

$$\sum_{i=1}^m |\pi_t(A_i) - \pi(A_i)| \leq$$

$$2\rho^t + \epsilon m \sum_{k=0}^{t-1} \rho^k = 2\rho^t + \epsilon m \frac{1 - \rho^t}{1 - \rho}$$

proving the Theorem.

V Discussion

The main application opportunity of our result is the following. Assume we want to estimate

the measure of a set $A \subseteq S$ by the MCMC method, so we define a Markov chain \mathcal{M} on S , as usual, such that the chain has uniform stationary distribution. But this chain typically has a huge state space and may not be rapidly mixing, so it does not provide us with a fast algorithm. Assume now that we can also define a partition \mathcal{Q} of S , such that $A \in \mathcal{Q}$. If \mathcal{M} is ϵ -lumpable with respect to \mathcal{Q} for some $\epsilon \geq 0$, then we can apply the bound of Theorem 2 to estimate the time for achieving a good approximation for the measure of A (assuming that ϵ is small). In this way we can possibly lower the time needed to run the MCMC algorithm. Of course, it also assumes that we have a good estimation for the ergodic coefficient ρ , but that is only needed for the *aggregated setting*, not for the original (huge) state space! This may prove to be a major advantage in certain situations.

References

- [1] I. Beichl and F. Sullivan, “The Metropolis Algorithm”, *Computing in Science and Engineering*, 2(2000), pp. 65-69.
- [2] D.J. Hartfiel, *Markov Set-Chains*, Lecture Notes in Mathematics 1695, Springer-Verlag, 1998.
- [3] D.J. Hartfiel, “Results on Limiting Sets of Markov Set-Chains”, *Linear Algebra and its Applications*, 195(1993), pp. 155-163.
- [4] F.P. Kelly, “Loss Networks”, *Annals of Applied Probability*, Vol. 1, No. 3, 1991, pp. 319-378.
- [5] J.G. Kemeny and J.L. Snell, *Finite Markov Chains*, Van Nostrand Reinhold, New York, 1960. (Later editions: Springer, 1976, 1983)
- [6] M. Kijima, *Markov Processes for Stochastic Modeling*, Chapman & Hall, 1997.
- [7] G. Louth, M. Mitzenmacher and F.P. Kelly, “Computational Complexity of Loss Networks”, *Theoretical Computer Science*, 125(1994), pp. 45-59.
- [8] A. Sinclair, *Algorithms for Random Generation and Counting*, Birkhäuser, Boston 1993.