Improving the Asymptotic Performance of Markov Chain Monte-Carlo by Inserting Vortices

Yi Sun
IDSIA
Galleria 2, Manno CH-6928, Switzerland
yi@idsia.ch

Faustino Gomez
IDSIA
Galleria 2, Manno CH-6928, Switzerland
tino@idsia.ch

Jürgen Schmidhuber
IDSIA
Galleria 2, Manno CH-6928, Switzerland
juergen@idsia.ch

Abstract

We present a new way of converting a reversible finite Markov chain into a non-reversible one, with a theoretical guarantee that the asymptotic variance of the MCMC estimator based on the non-reversible chain is reduced. The method is applicable to any reversible chain whose states are not connected through a tree, and can be interpreted graphically as inserting vortices into the state transition graph. Our result confirms that non-reversible chains are fundamentally better than reversible ones in terms of asymptotic performance, and suggests interesting directions for further improving MCMC.

1 Introduction

Markov Chain Monte Carlo (MCMC) methods have gained enormous popularity over a wide variety of research fields [6, 8], owing to their ability to compute expectations with respect to complex, high dimensional probability distributions. An MCMC estimator can be based on any ergodic Markov chain with the distribution of interest as its stationary distribution. However, the choice of Markov chain greatly affects the performance of the estimator, in particular the accuracy achieved with a pre-specified number of samples [4].

In general, the efficiency of an MCMC estimator is determined by two factors: i) how fast the chain converges to its stationary distribution, i.e., the mixing rate [9], and ii) once the chain reaches its stationary distribution, how much the estimates fluctuate based on trajectories of finite length, which is characterized by the asymptotic variance. In this paper, we consider the latter criteria. Previous theory concerned with reducing asymptotic variance has followed two main tracks. The first focuses on reversible chains, and is mostly based on the theorems of Peskun [10] and Tierney [11], which state that if a reversible Markov chain is modified so that the probability of staying in the same state is reduced, then the asymptotic variance can be decreased. A number of methods have been proposed, particularly in the context of Metropolis-Hastings method, to encourage the Markov chain to move away from the current state, or its adjacency in the continuous case [12, 13]. The second track, which was explored just recently, studies non-reversible chains. Neal proved in [4] that starting from any finite-state reversible chain, the asymptotic variance of a related non-reversible chain, with reduced probability of back-tracking to the immediately previous state, will not increase, and typically decrease. Several methods have been proposed by Murray based on this idea [5].

Neal's result suggests that non-reversible chains may be fundamentally better than reversible ones in terms of the asymptotic performance. In this paper, we follow up this idea by proposing a new way of converting reversible chains into non-reversible ones which, unlike in Neal's method, are defined on the state space of the reversible chain, with the theoretical guarantee that the asymptotic variance of the associated MCMC estimator is reduced. Our method is applicable to any non-reversible chain whose state transition graph contains loops, including those whose probability of staying in the same state is zero and thus cannot be improved using Peskun's theorem. The method also admits an interesting graphical interpretation which amounts to inserting 'vortices' into the state transition graph of the original chain. Our result suggests a new and interesting direction for improving the asymptotic performance of MCMC.

The rest of the paper is organized as follows: section 2 reviews some background concepts and results; section 3 presents the main theoretical results, together with the graphical interpretation; section 4 provides a simple yet illustrative example and explains the intuition behind the results; section 5 concludes the paper.

2 Preliminaries

Suppose we wish to estimate the expectation of some real valued function f over domain $\mathcal{S}$, with respect to a probability distribution π, whose value may only be known to a multiplicative constant. Let A be a transition operator of an ergodic[1] Markov chain with stationary distribution π, i.e.,

$$\pi(x) A(x \to y) = \pi(y) B(y \to x), \forall x, y \in \mathcal{S}, \tag{1}$$

where B is the reverse operator as defined in [5]. The expectation can then be estimated through the MCMC estimator

$$\mu_T = \frac{1}{T} \sum_{t=1}^{T} f(x_t), \tag{2}$$

where $x_1, \cdots, x_T$ is a trajectory sampled from the Markov chain. The asymptotic variance of μ_T, with respect to transition operator A and function f is defined as

$$\sigma_A^2(f) = \lim_{T \to \infty} T \mathbb{V}[\mu_T], \tag{3}$$

where $\mathbb{V}[\mu_T]$ denotes the variance of μ_T. Since the chain is ergodic, $\sigma_A^2(f)$ is well-defined following the central limit theorem, and does not depend on the distribution of the initial point. Roughly speaking, asymptotic variance has the meaning that the mean square error of the estimates based on T consecutive states of the chain would be approximately $\frac{1}{T}\sigma_A^2(f)$, after a sufficiently long period of "burn in" such that the chain is close enough to its stationary distribution. Asymptotic variance can be used to compare the asymptotic performance of MCMC estimators based on different chains with the same stationary distribution, where smaller asymptotic variance indicates that, asymptotically, the MCMC estimator requires fewer samples to reach a specified accuracy.

Under the ergodic assumption, the asymptotic variance can be written as

$$\sigma_A^2(f) = \mathbb{V}[f] + \sum_{\tau=1}^{\infty} \left(c_{A,f}(\tau) + c_{B,f}(\tau) \right), \tag{4}$$

where

$$c_{A,f}(\tau) = \mathbb{E}_A[f(x_t) f(x_{t+\tau})] - \mathbb{E}_A[f(x_t)] \mathbb{E}[f(x_{t+\tau})]$$

is the covariance of the function value between two states that are τ time steps apart in the trajectory of the Markov chain with transition operator A. Note that $\sigma_A^2(f)$ depends on both A and its reverse operator B, and $\sigma_A^2(f) = \sigma_B^2(f)$ since A is also the reverse operator of B by definition.

In this paper, we consider only the case where $\mathcal{S}$ is finite, i.e., $\mathcal{S} = \{1, \cdots, S\}$, so that the transition operators A and B, the stationary distribution π, and the function f can all be written in matrix form. Let $\pi = [\pi(1), \cdots, \pi(S)]^\top$, $f = [f(1), \cdots, f(S)]^\top$, $A_{i,j} = A(i \to j)$, $B_{i,j} = B(i \to j)$. The asymptotic variance can thus be written as

$$\sigma_A^2(f) = \mathbb{V}[f] + \sum_{\tau=1}^{\infty} f^\top \left(Q A^\tau + Q B^\tau - 2\pi\pi^\top \right) f,$$

[1] Strictly speaking, the ergodic assumption is not necessary for the MCMC estimator to work, see [4]. However, we make the assumption to simplify the analysis.

with $Q = \operatorname{diag}\{\pi\}$. Since B is the reverse operator of A, $QA = B^\top Q$. Also, from the ergodic assumption,

$$\lim_{\tau\to\infty} A^\tau = \lim_{\tau\to\infty} B^\tau = R,$$

where $R = \mathbf{1}\pi^\top$ is a square matrix in which every row is $\pi^\top$. It follows that the asymptotic variance can be represented by Kenney's formula [7] in the non-reversible case:

$$\sigma_A^2(f) = \mathbb{V}[f] + 2(Qf)^\top \left[\Lambda^-\right]_H (Qf) - 2f^\top Qf, \tag{5}$$

where $[\cdot]_H$ denotes the Hermitian (symmetric) part of a matrix, and $\Lambda = Q + \pi\pi^\top - J$, with $J = QA$ being the joint distribution of two consecutive states.

3 Improving the asymptotic variance

It is clear from Eq.5 that the transition operator A affects the asymptotic variance only through term $[\Lambda^-]_H$. If the chain is reversible, then J is symmetric, so that Λ is also symmetric, and therefore comparing the asymptotic variance of two MCMC estimators becomes a matter of comparing their J, namely, if[2] $J \preceq J' = QA'$, then $\sigma_A^2(f) \leq \sigma_{A'}^2(f)$, for any f. This leads to a simple proof of Peskun's theorem in the discrete case [3].

In the case where the Markov chain is non-reversible, i.e., J is asymmetric, the analysis becomes much more complicated. We start by providing a sufficient and necessary condition in section 3.1, which transforms the comparison of asymptotic variance based on arbitrary finite Markov chains into a matrix ordering problem, using a result from matrix analysis. In section 3.2, a special case is identified, in which the asymptotic variance of a reversible chain is compared to that of a non-reversible one whose joint distribution over consecutive states is that of the reversible chain plus a skew-Hermitian matrix. We prove that the resulting non-reversible chain has smaller asymptotic variance, and provide a necessary and sufficient condition for the existence of such non-zero skew-Hermitian matrices. Finally in section 3.3, we provide a graphical interpretation of the result.

3.1 The general case

From Eq.5 we know that comparing the asymptotic variances of two MCMC estimators is equivalent to comparing their $[\Lambda^-]_H$. The following result from [1, 2] allows us to write $[\Lambda^-]_H$ in terms of the symmetric and asymmetric parts of Λ.

Lemma 1 *If a matrix X is invertible, then $[X^-]_H^- = [X]_H + [X]_S^\top [X]_H^- [X]_S$, where $[X]_S$ is the skew Hermitian part of X.*

From Lemma 1, it follows immediately that in the discrete case, the comparison of MCMC estimators based on two Markov chains with the same stationary distribution can be cast as a different problem of matrix comparison, as stated in the following proposition.

Proposition 1 *Let A, A' be two transition operators of ergodic Markov chains with stationary distribution π. Let $J = QA$, $J' = QA'$, $\Lambda = Q + \pi\pi^\top - J$, $\Lambda' = Q + \pi\pi^\top - J'$. Then the following three conditions are equivalent:*

1) $\sigma_A^2(f) \leq \sigma_{A'}^2(f)$ *for any f*

2) $[\Lambda^-]_H \preceq \left[(\Lambda')^-\right]_H$

3) $[J]_H - [J]_S^\top [\Lambda]_H^- [J]_S \preceq [J']_H - [J']_S^\top [\Lambda']_H^- [J']_S$

Proof. First we show that Λ is invertible. Following the steps in [3], for any $f \neq 0$,

$$\begin{aligned} f^\top \Lambda f &= f^\top [\Lambda]_H f = f^\top \left(Q + \pi\pi^\top - J\right) f \\ &= \frac{1}{2}\mathbb{E}\left[\left(f(x_t) - f(x_{t+1})\right)^2\right] + \mathbb{E}[f(x_t)]^2 > 0, \end{aligned}$$

[2] For symmetric matrices X and Y, we write $X \preceq Y$ if $Y - X$ is positive semi-definite, and $X \prec Y$ if $Y - X$ is positive definite.

thus $[\Lambda]_H \succ 0$, and Λ is invertible since $\Lambda f \neq 0$ for any $f \neq 0$.

Condition 1) and 2) are equivalent by definition. We now prove 2) is equivalent to 3). By Lemma 1,

$$\left[\Lambda^-\right]_H \preceq \left[\left(\Lambda'\right)^-\right]_H \Longleftrightarrow [\Lambda]_H + [\Lambda]_S^\top [\Lambda]_H^- [\Lambda]_S \succeq [\Lambda']_H + [\Lambda']_S^\top [\Lambda']_H^- [\Lambda']_S ,$$

the result follows by noticing that $[\Lambda]_H = Q + \pi\pi^\top - [J]_H$ and $[\Lambda]_S = -[J]_S$. ■

3.2 A special case

Generally speaking, the conditions in Proposition 1 are very hard to verify, particularly because of the term $[J]_S^\top [\Lambda]_H^- [J]_S$. Here we focus on a special case where $[J']_S = 0$, and $[J']_H = J' = [J]_H$. This amounts to the case where the second chain is reversible, and its transition operator is the average of the transition operator of the first chain and the associated reverse operator. The result is formalized in the following corollary.

Corollary 1 *Let T be a reversible transition operator of a Markov chain with stationary distribution π. Assume there is some H that satisfies*

Condition I. $\mathbf{1}^\top H = 0$, $H\mathbf{1} = 0$, $H = -H^\top$, *and*[3]

Condition II. $T \pm Q^- H$ *are valid transition matrices.*

Denote $A = T + Q^- H$, $B = T - Q^- H$, then

1) *A preserves π, and B is the reverse operator of A.*

2) $\sigma_A^2(f) = \sigma_B^2(f) \leq \sigma_T^2(f)$ *for any f.*

3) *If $H \neq 0$, then there is some f, such that $\sigma_A^2(f) < \sigma_T^2(f)$.*

4) *If $A_\varepsilon = T + (1+\varepsilon) Q^- H$ is valid transition matrix, $\varepsilon > 0$, then $\sigma_{A_\varepsilon}^2(f) \leq \sigma_A^2(f)$.*

Proof. For 1), notice that $\pi^\top T = \pi^\top$, so

$$\pi^\top A = \pi^\top T + \pi^\top Q^- H = \pi^\top + \mathbf{1}^\top H = \pi^\top,$$

and similarly for B. Moreover

$$QA = QT + H = (QT - H)^\top = \left(Q\left(T - Q^- H\right)\right)^\top = (QB)^\top,$$

thus B is the reverse operator of A.

For 2), $\sigma_A^2(f) = \sigma_B^2(f)$ follows from Eq.5. Let $J' = QT$, $J = QA$. Note that $[J]_S = H$,

$$J' = QT = \frac{1}{2}(QA + QB) = [QA]_H = [J]_H ,$$

and $[\Lambda]_H \succ 0$ thus $H^\top [\Lambda]_H^- H \succeq 0$ from Proposition 1. It follows that $\sigma_A^2(f) \leq \sigma_T^2(f)$ for any f.

For 3), write $X = [\Lambda]_H$,

$$\left[\Lambda^-\right]_H = \left(X + H^\top X^- H\right)^- = X^- - X^- H^\top \left(X + HX^- H^\top\right)^- HX^- .$$

Since $X \succ 0$, $HX^- H^\top \succeq 0$, one can write $\left(X + HX^- H^\top\right)^- = \sum_{s=1}^S \lambda_s e_s e_s^\top$, with $\lambda_s > 0, \forall s$. Thus

$$H^\top \left(X + HX^- H^\top\right)^- H = \sum\nolimits_{s=1}^S \lambda_s He_s \left(He_s\right)^\top .$$

Since $H \neq 0$, there is at least one s^*, such that $He_{s^*} \neq 0$. Let $f = Q^- X He_{s^*}$, then

$$\begin{aligned}
\frac{1}{2}\left[\sigma_T^2(f) - \sigma_A^2(f)\right] &= (Qf)^\top \left[X^- - \left(X + H^\top X^- H\right)^-\right](Qf) \\
&= (Qf)^\top X^- H^\top \left(X + HX^- H^\top\right)^- HX^- (Qf) \\
&= \left(He_{s^*}\right)^\top \sum\nolimits_{s=1}^S \lambda_s He_s \left(He_s\right)^\top \left(He_{s^*}\right) \\
&= \lambda_s \left\|He_{s^*}\right\|^4 + \sum\nolimits_{s \neq s^*} \lambda_s \left(e_{s^*}^\top H^\top He_s\right)^2 > 0.
\end{aligned}$$

[3] We write $\mathbf{1}$ for the S-dimensional column vector of 1's.

For 4), let $\Lambda_\varepsilon = Q + \pi\pi^\top - QA_\varepsilon$, then for $\varepsilon > 0$,

$$\left[\Lambda_\varepsilon^-\right]_H = \left(X + (1+\varepsilon)^2 H^\top X^- H\right)^- \preceq \left(X + H^\top X^- H\right)^- = \left[\Lambda^-\right]_H,$$

by Eq.5, we have $\sigma^2_{A_\varepsilon}(f) \leq \sigma^2_A(f)$ for any f. ■

Corollary 1 shows that starting from a reversible Markov chain, as long as one can find a non-zero H satisfying Conditions I and II, then the asymptotic performance of the MCMC estimator is guaranteed to improve. The next question to ask is whether such an H exists, and, if so, how to find one. We answer this question by first looking at Condition I. The following proposition shows that any H satisfying this condition can be constructed systematically.

Proposition 2 *Let H be an S-by-S matrix. H satisfies Condition I if and only if H can be written as the linear combination of $\frac{1}{2}(S-1)(S-2)$ matrices, with each matrix of the form*

$$U_{i,j} = u_i u_j^\top - u_j u_i^\top,\ 1 \leq i < j \leq S-1.$$

Here $u_1, \cdots, u_{S-1}$ are $S-1$ non-zero linearly independent vectors satisfying $u_s^\top \mathbf{1} = 0$.

Proof. Sufficiency. It is straightforward to verify that each $U_{i,j}$ is skew-Hermitian and satisfies $U_{i,j}\mathbf{1} = 0$. Such properties are inherited by any linear combination of $U_{i,j}$.

Necessity. We show that there are at most $\frac{1}{2}(S-1)(S-2)$ linearly independent bases for all H such that $H = -H^\top$ and $H\mathbf{1} = 0$. On one hand, any S-by-S skew-Hermitian matrix can be written as the linear combination of $\frac{1}{2}S(S-1)$ matrices of the form

$$V_{i,j} : \{V_{i,j}\}_{m,n} = \delta(m,i)\delta(n,j) - \delta(n,i)\delta(m,j),$$

where δ is the standard delta function such that $\delta(i,j) = 1$ if $i = j$ and 0 otherwise. However, the constraint $H\mathbf{1} = 0$ imposes $S-1$ linearly independent constraints, which means that out of $\frac{1}{2}S(S-1)$ parameters, only

$$\frac{1}{2}S(S-1) - (S-1) = \frac{1}{2}(S-1)(S-2)$$

are independent.

On the other hand, selecting two non-identical vectors from $u_1, \cdots, u_{S-1}$ results in $\binom{S-1}{2} = \frac{1}{2}(S-1)(S-2)$ different $U_{i,j}$. It has still to be shown that these $U_{i,j}$ are linearly independent.

Assume

$$0 = \sum_{1\leq i<j\leq S-1} \kappa_{i,j} U_{i,j} = \sum_{1\leq i<j\leq S-1} \kappa_{i,j}\left(u_i u_j^\top - u_j u_i^\top\right), \forall \kappa_{i,j} \in \mathbb{R}.$$

Consider two cases: Firstly, assume $u_1, \cdots, u_{S-1}$ are orthogonal, i.e., $u_i^\top u_j = 0$ for $i \neq j$. For a particular u_s,

$$\begin{aligned} 0 &= \sum_{1\leq i<j\leq S-1} \kappa_{i,j} U_{i,j} u_s = \sum_{1\leq i<j\leq S-1} \kappa_{i,j}\left(u_i u_j^\top - u_j u_i^\top\right) u_s \\ &= \sum_{1\leq i<s} \kappa_{i,s} u_i \left\|u_s^\top u_s\right\| + \sum_{s<j\leq S-1} \kappa_{s,j} u_j \left\|u_s^\top u_s\right\|. \end{aligned}$$

Since $\left\|u_s^\top u_s\right\| \neq 0$, it follows that $\kappa_{i,s} = \kappa_{s,j} = 0$, for all $1 \leq i < s < j \leq S-1$. This holds for any u_s, so all $\kappa_{i,j}$ must be 0, and therefore $U_{i,j}$ are linearly independent by definition. Secondly, if $u_1, \cdots, u_{S-1}$ are not orthogonal, one can construct a new set of orthogonal vectors $\tilde{u}_1, \cdots, \tilde{u}_{S-1}$ from $u_1, \cdots, u_{S-1}$ through Gram–Schmidt orthogonalization, and create a different set of bases $\tilde{U}_{i,j}$. It is easy to verify that each $\tilde{U}_{i,j}$ is a linear combination of $U_{i,j}$. Since all $\tilde{U}_{i,j}$ are linearly independent, it follows that $U_{i,j}$ must also be linearly independent. ■

Proposition 2 confirms the existence of non-zero H satisfying Condition I. We now move to Condition II, which requires that both $QT + H$ and $QT - H$ remain valid joint distribution matrices, i.e.

all entries must be non-negative and sum up to 1. Since $\mathbf{1}^\top (QT + H)\,\mathbf{1} = 1$ by Condition I, only the non-negative constraint needs to be considered.

It turns out that not all reversible Markov chains admit a non-zero H satisfying both Condition I and II. For example, consider a Markov chain with only two states. It is impossible to find a non-zero skew-Hermitian H such that $H\mathbf{1} = 0$, because all 2-by-2 skew-Hermitian matrices are proportional to $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$.

The next proposition gives the sufficient and necessary condition for the existence of a non-zero H satisfying both I and II. In particular, it shows an interesting link between the existence of such H and the connectivity of the states in the reversible chain.

Proposition 3 *Assume a reversible ergodic Markov chain with transition matrix T and let $J = QT$. The* state transition graph *$\mathcal{G}_T$ is defined as the undirected graph with node set $\mathcal{S} = \{1, \cdots, S\}$ and edge set $\{(i,j) : J_{i,j} > 0, 1 \le i < j \le S\}$. Then there exists some non-zero H satisfying Condition I and II, if and only if there is a loop in $\mathcal{G}_T$.*

Proof. Sufficiency: Without loss of generality, assume the loop is made of states $1, 2, \cdots, N$ and edges $(1,2), \cdots, (N-1, N), (N, 1)$, with $N \ge 3$. By definition, $J_{1,N} > 0$, and $J_{n,n+1} > 0$ for all $1 \le n \le N-1$. A non-zero H can then be constructed as

$$H_{i,j} = \begin{cases} \varepsilon, & \text{if } 1 \le i \le N-1 \text{ and } j = i+1, \\ -\varepsilon, & \text{if } 2 \le i \le N \text{ and } j = i-1, \\ \varepsilon, & \text{if } i = N \text{ and } j = 1, \\ -\varepsilon, & \text{if } i = 1 \text{ and } j = N, \\ 0, & \text{otherwise.} \end{cases}$$

Here

$$\varepsilon = \min_{1 \le n \le N-1} \left\{ J_{n,n+1},\, 1 - J_{n,n+1},\, J_{1,N},\, 1 - J_{1,N} \right\}.$$

Clearly, $\varepsilon > 0$, since all the items in the minimum are above 0. It is trivial to verify that $H = -H^\top$ and $H\mathbf{1} = 0$.

Necessity: Assume there are no loops in $\mathcal{G}_T$, then all states in the chain must be organized in a tree, following the ergodic assumption. In other word, there are exactly $2\,(S-1)$ non-zero off-diagonal elements in J. Plus, these $2\,(S-1)$ elements are arranged symmetrically along the diagonal and spanning every column and row of J.

Because the states are organized in a tree, there is at least one leaf node s in $\mathcal{G}_T$, with a single neighbor s'. Row s and column s in J thus looks like $r_s = [\cdots, p_{s,s}, \cdots, p_{s,s'}, \cdots]$ and its transpose, respectively, with $p_{s,s} \ge 0$ and $p_{s,s'} > 0$, and all other entries being 0.

Assume that one wants to construct a some H, such that $J \pm H \ge 0$. Let h_s be the s-th row of H. Since $r_s \pm h_s \ge 0$, all except the s'-th elements in h_s must be 0. But since $h_s\mathbf{1} = 0$, the whole s-th row, thus the s-th column of H must be 0.

Having set the s-th column and row of H to 0, one can consider the reduced Markov chain with one state less, and repeat with another leaf node. Working progressively along the tree, it follows that all rows and columns in H must be 0. ■

The indication of Proposition 3 together with 2 is that all reversible chains can be improved in terms of asymptotic variance using Corollary 1, except those whose transition graphs are trees. In practice, the non-tree constraint is not a problem because almost all current methods of constructing reversible chains generate chains with loops.

3.3 Graphical interpretation

In this subsection we provide a graphical interpretation of the results in the previous sections. Starting from a simple case, consider a reversible Markov chain with three states forming a loop. Let $u_1 = [1, 0, -1]^\top$ and $u_2 = [0, 1, -1]^\top$. Clearly, u_1 and u_2 are linearly independent and $u_1^\top \mathbf{1} = u_2^\top \mathbf{1} = 0$. By Proposition 2 and 3, there exists some $\varepsilon > 0$, such that $H = \varepsilon U_{12}$ satis-

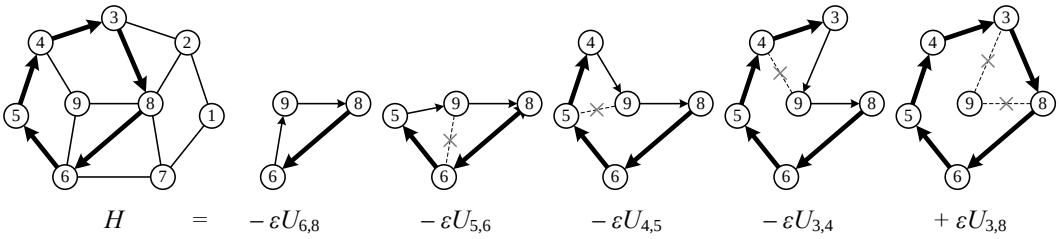

Figure 1: Illustration of the construction of larger vortices. The left hand side is a state transition graph of a reversible Markov chain with $S = 9$ states, with a vortex $3 \to 8 \to 6 \to 5 \to 4$ of strength ε inserted. The corresponding H can be expressed as the linear combination of $U_{i,j}$, as shown on the right hand side of the graph. We start from the vortex $8 \to 6 \to 9 \to 8$, and add one vortex a time. The dotted lines correspond to edges on which the flows cancel out when a new vortex is added. For example, when vortex $6 \to 5 \to 9 \to 6$ is added, edge $9 \to 6$ cancels edge $6 \to 9$ in the previous vortex, resulting in a larger vortex with four states. Note that in this way one can construct vortices which do not include state 9, although each $U_{i,j}$ is a vortex involving 9.

fies Condition I and II, with $U_{1,2} = u_1 u_2^\top - u_2 u_1^\top$. Write $U_{1,2}$ and $J + H$ in explicit form,

$$U_{1,2} = \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}, \quad J + H = \begin{bmatrix} p_{1,1} & p_{1,2} + \varepsilon & p_{1,3} - \varepsilon \\ p_{2,1} - \varepsilon & p_{2,2} & p_{2,3} + \varepsilon \\ p_{3,1} + \varepsilon & p_{3,2} - \varepsilon & p_{3,3} \end{bmatrix},$$

with $p_{i,j}$ being the probability of the consecutive states being i, j. It is clear that in $J + H$, the probability of jumps $1 \to 2$, $2 \to 3$, and $3 \to 1$ is increased, and the probability of jumps in the opposite direction is decreased. Intuitively, this amounts to adding a 'vortex' of direction $1 \to 2 \to 3 \to 1$ in the state transition. Similarly, the joint probability matrix for the reverse operator is $J - H$, which adds a vortex in the opposite direction. This simple case also gives an explanation of why adding or subtracting non-zero H can only be done where a loop already exists, since the operation requires subtracting ε from all entries in J corresponding to edges in the loop.

In the general case, define $S - 1$ vectors $u_1, \cdots, u_{S-1}$ as

$$u_s = [0, \cdots, 0, \underset{s\text{-th element}}{1}, 0, \cdots, 0, -1]^\top.$$

It is straightforward to see that $u_1, \cdots, u_{S-1}$ are linearly independent and $u_s^\top \mathbf{1} = 0$ for all s, thus any H satisfying Condition I can be represented as the linear combination of $U_{i,j} = u_i u_j^\top - u_j u_i^\top$, with each $U_{i,j}$ containing 1's at positions (i, j), (j, S), (S, i), and -1's at positions (i, S), (S, j), (j, i). It is easy to verify that adding $\varepsilon U_{i,j}$ to J amounts to introducing a vortex of direction $i \to j \to S \to i$, and any vortex of N states ($N \geq 3$) $s_1 \to s_2 \to \cdots \to s_N \to s_1$ can be represented by the linear combination $\sum_{n=1}^{N-1} U_{s_n, s_{n+1}}$ in the case of state S being in the vortex and assuming $s_N = S$ without loss of generality, or $U_{s_N, s_1} + \sum_{n=1}^{N-1} U_{s_n, s_{n+1}}$ if S is not in the vortex, as demonstrated in Figure 1. Therefore, adding or subtracting an H to J is equivalent to inserting a number of vortices into the state transition map.

4 An example

Adding vortices to the state transition graph forces the Markov chain to move in loops following pre-specified directions. The benefit of this can be illustrated in the following example. Consider a reversible Markov chain with S states forming a ring, namely from state s one can only jump to $s \oplus 1$ or $s \ominus 1$, with $\oplus$ and $\ominus$ being the mod-S summation and subtraction. The only possible non-zero H in this example is of form $\varepsilon \sum_{s=1}^{S-1} U_{s,s+1}$, corresponding to vortices on the large ring.

We assume uniform stationary distribution $\pi(s) = \frac{1}{S}$. In this case, any reversible chain behaves like a random walk. The chain which achieves minimal asymptotic variance is the one with the probability of both jumping forward and backward being $\frac{1}{2}$. The expected number of steps for this chain to reach the state $\frac{S}{2}$ edges away is $\frac{S^2}{4}$. However, adding the vortex reduces this number to

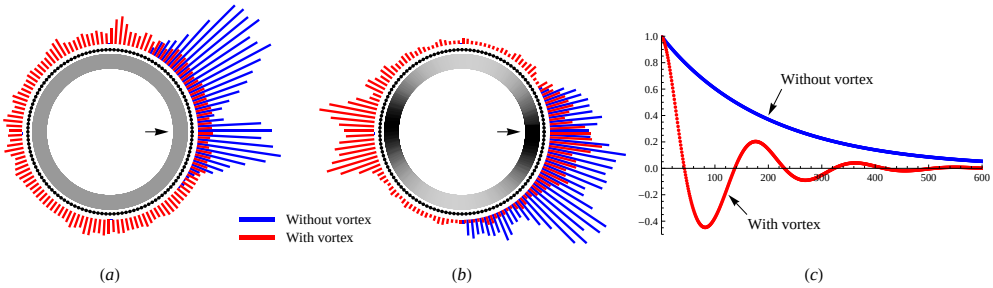

Figure 2: Demonstration of the vortex effect: (a) and (b) show two different, reversible Markov chains, each containing 128 states connected in a ring. The equilibrium distribution of the chains is depicted by the gray inner circles; darker shades correspond to higher probability. The equilibrium distribution of chain (a) is uniform, while that of (b) contains two peaks half a ring apart. In addition, the chains are constructed such that the probability of staying in the same state is zero. In each case, two trajectories, of length 1000, are generated from the chain with and without the vortex, starting from the state pointed to by the arrow. The length of the bar radiating out from a given state represents the relative frequency of visits to that state, with red and blue bars corresponding to chains with and without vortex, respectively. It is clear from the graph that trajectories sampled from reversible chains spread much slower, with only $1/5$ of the states reached in (a) and $1/3$ in (b), and the trajectory in (b) does not escape from the current peak. On the other hand, with vortices added, trajectories of the same length spread over all the states, and effectively explore both peaks of the stationary distribution in (b). The plot (c) show the correlation of function values (normalized by variance) between two states τ time steps apart, with τ ranging from 1 to 600. Here we take the Markov chains from (b) and use function $f(s) = \cos\left(4\pi \cdot \frac{s}{128}\right)$. When vortices are added, not only do the absolute values of the correlations go down significantly, but also their signs alternate, indicating that these correlations tend to cancel out in the sum of Eq.5.

roughly $\frac{S}{2\varepsilon}$ for large S, suggesting that it is much easier for the non-reversible chain to reach faraway states, especially for large S. In the extreme case, when $\varepsilon = \frac{1}{2}$, the chain cycles deterministically, reducing asymptotic variance to zero. Also note that the reversible chain here has zero probability of staying in the current state, thus cannot be further improved using Peskun's theorem.

Our intuition about why adding vortices helps is that chains with vortices move faster than the reversible ones, making the function values of the trajectories less correlated. This effect is demonstrated in Figure 2.

5 Conclusion

In this paper, we have presented a new way of converting a reversible finite Markov chain into a non-reversible one, with the theoretical guarantee that the asymptotic variance of the MCMC estimator based on the non-reversible chain is reduced. The method is applicable to any reversible chain whose states are not connected through a tree, and can be interpreted graphically as inserting vortices into the state transition graph.

The results confirm that non-reversible chains are fundamentally better than reversible ones. The general framework of Proposition 1 suggests further improvements of MCMC's asymptotic performance, by applying other results from matrix analysis to asymptotic variance reduction. The combined results of Corollary 1, and Propositions 2 and 3, provide a specific way of doing so, and pose interesting research questions. Which combinations of vortices yield optimal improvements for a given chain? Finding one of them is a combinatorial optimization problem. How can a good combination be constructed in practice, using limited history and computational resources?

References

[1] R.P. Wen, "Properties of the Matrix Inequality", Journal of Taiyuan Teachers College, 2005.

[2] R. Mathias, "Matrices With Positive Definite Hermitian Part: Inequalities And Linear Systems", http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.1768, 1992.

[3] L.H. Li, "A New Proof of Peskun's and Tierney's Theorems using Matrix Method", Joint Graduate Students Seminar of Department of Statistics and Department of Biostatistics, Univ. of Toronto, 2005.

[4] R.M. Neal, "Improving asymptotic variance of MCMC estimators: Non-reversible chains are better", Technical Report No. 0406, Department of Statistics, Univ. of Toronto, 2004.

[5] I. Murray, "Advances in Markov chain Monte Carlo methods", M. Sci. thesis, University College London, 2007.

[6] R.M. Neal, "Bayesian Learning for Neural Networks", Springer, 1996.

[7] J. Kenney and E.S. Keeping, "Mathematics of Statistics", van Nostrand, 1963.

[8] C. Andrieu, N. de Freitas, A. Doucet, and M.I. Jordan, "An Introduction to MCMC for Machine Learning", Machine Learning, 50, 5-43, 2003.

[9] Szakdolgozat, "The Mixing Rate of Markov Chain Monte Carlo Methods and some Applications of MCMC Simulation in Bioinformatics", M.Sci. thesis, Eotvos Lorand University, 2006.

[10] P.H. Peskun, "Optimum Monte-Carlo sampling using Markov chains", Biometrika, vol. 60, pp. 607-612, 1973.

[11] L. Tierney, "A note on Metropolis Hastings kernels for general state spaces", Ann. Appl. Probab. 8, 1-9, 1998.

[12] S. Duane, A.D. Kennedy, B.J. Pendleton and D. Roweth, "Hybrid Monte Carlo", Physics Letters B, vol.195-2, 1987.

[13] J.S. Liu, "Peskun's theorem and a modified discrete-state Gibbs sampler", Biometria, vol.83, pp.681-682, 1996.