Technical Report BU-1385-M, Biometrics Unit, Cornell University

Analysis of a Non-Reversible Markov Chain Sampler

Persi Diaconis
Maths and ORIE
Cornell University
and
Dept of Mathematics
Harvard University

Susan Holmes
Biometrics Unit
Cornell University
and
Unité de Biométrie
INRA-Montpellier
France
sph11@cornell.edu

Radford M. Neal
Dept. of Statistics and Dept. of Computer Science
University of Toronto
Canada
radford@stat.utoronto.ca

Abstract

We analyse the convergence to stationarity of a simple non-reversible Markov chain that serves as a model for several non-reversible Markov chain sampling methods that are used in practice. Our theoretical and numerical results show that non-reversibility can indeed lead to improvements over the diffusive behavior of simple Markov chain sampling schemes. The analysis uses both probabilistic techniques and an explicit diagonalisation.

Keywords: Non-reversible Markov chain, Markov chain Monte Carlo, Metropolis algorithm.

5 June 1997

Acknowledgments

We thank David Aldous, Martin Hildebrand, Brad Mann, and Laurent Saloff-Coste for their help.

1 Introduction

Markov chain sampling methods are commonly used in statistics [30, 29], computer science [28], statistical mechanics [2], and quantum field theory [31, 21]. In all these fields, distributions are encountered that are difficult to sample from directly, but for which a Markov chain that converges to the distribution can easily be constructed. For many such methods (eg, the Metropolis algorithm [23, 11], and the Gibbs sampler [15, 14] with a random scan), the Markov chain constructed is reversible. These methods tend to explore the distribution by means of a diffusive random walk. Some other common methods, such as the Gibbs sampler with a systematic scan, use a Markov chain that is not reversible, but which has diffusive behavior resembling that of a related reversible chain [27].

Some Markov chain methods attempt to avoid the inefficiencies of such diffusive exploration. The Hybrid Monte Carlo method [13] uses an elaborate Metropolis proposal that can make large changes to the state. In a variant of this method due to Horowitz [19], a similar effect is produced using a Markov chain that is carefully designed to be non-reversible. (See [31, 21, 24] for reviews of these methods.) The overrelaxation method [1] also employs a non-reversible Markov chain as a way of suppressing diffusive behavior, as discussed in [25].

In this paper, we analyse a non-reversible Markov chain that does a one-dimensional walk, as an abstraction of these practical sampling methods, particularly that of Horowitz [19]. Gustafson [17] has also recently tried using adaptations of Horowitz's method. We find that the non-reversible walk does indeed converge more rapidly than the corresponding simple random walk. We analyse convergence in total variation distance and in χ^2 distance in some detail, finding that this is one of the few natural instances where total variation and χ^2 relaxation times differ. We then discuss generalizations of the method, and their relationships to other sampling methods, and explore applications to several statistical problems. Finally, we discuss some limitations of these techniques.

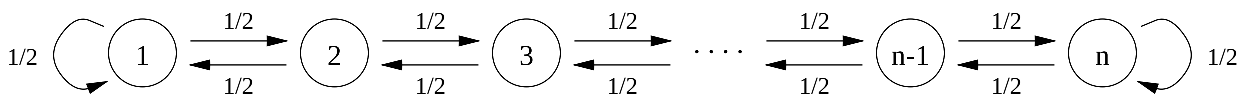
2 Reversible and non-reversible walks in one-dimension

All our examples concern distributions on some finite set, $\mathcal{X}$, with positive probabilities given by $\pi(x)$. We sample from $\pi(x)$ by running an irreducible Markov Chain on $\mathcal{X}$ with transition probabilities $K(x, y)$, constructed so that $\pi(x)$ is the stationary distribution. Such a chain is reversible with respect to π if

$$\pi(x)K(x,y) \;=\; \pi(y)K(y,x), \quad \text{for all } x, y \in \mathcal{X} \tag{2.1}$$

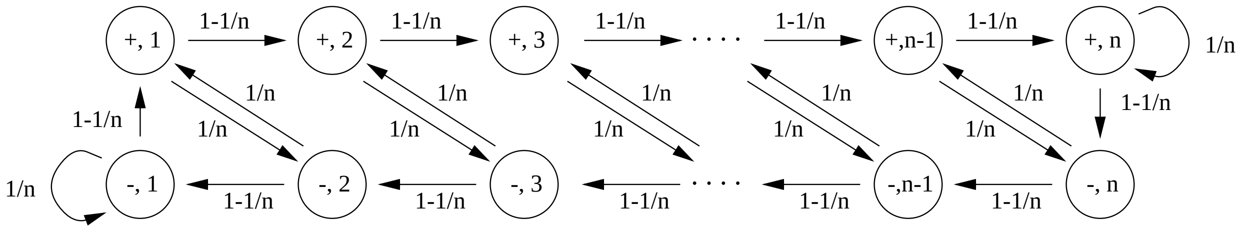
Reversibility is a sufficient, but not necessary, condition for $\pi(x)$ to be a stationary distribution of the chain.

We consider first the simple case where $\mathcal{X} = \{1, 2, 3, \ldots, n\}$, and where the desired distribution is uniform: $\pi(x) = 1/n$. A reversible Markov chain converging to this distribution can be constructed as a nearest neighbor random walk on the n-point path with holding probabilities of 1/2 at each end — ie, $K(x, y) = 1/2$ for $y = x \pm 1$ and $x, y \in \mathcal{X}$, and $K(1, 1) = K(n, n) = 1/2$ also. The chain can be pictured thus:



This walk takes on the order of n^2 steps to reach stationarity, since using the central limit theorem we see that the walk will take on the order of k^2 steps to travel a distance of order k.

We overcome this "diffusive" behavior by introducing two copies of each state, in the "upstairs" copy the chain goes right $1 - 1/n$ of the time. In the "downstairs" copy it goes left $1 - 1/n$ of the time. The chain switches between copies at rate $1/n$. We label the upstairs states $(+, 1)$, $(+, 2)$, ..., $(+, n)$, and the downstairs states $(-, 1)$, $(-, 2)$, ..., $(-, n)$. The chain can then be pictured thus:



The transition probabilities are as follows:

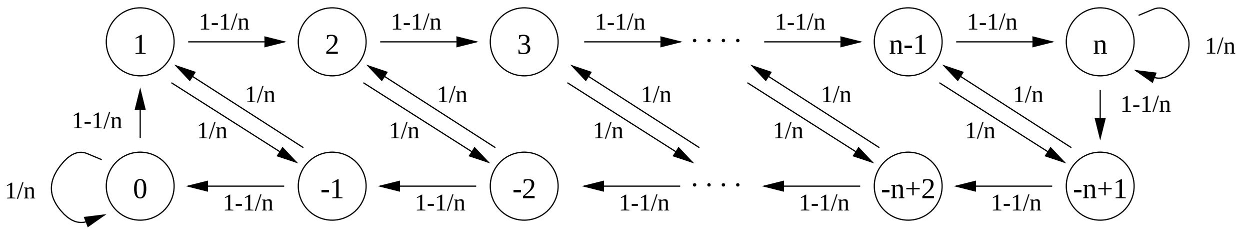
$$
\begin{array}{rclll}
K((+,x),(+,x+1)) & = & 1-\frac{1}{n} & \text{for } 1 \le x < n, & K((+,n),(-,n)) = 1-\frac{1}{n} \\
K((+,x),(-,x+1)) & = & \frac{1}{n} & \text{for } 1 \le x < n, & K((+,n),(+,n)) = \frac{1}{n} \\
K((-,x),(-,x-1)) & = & 1-\frac{1}{n} & \text{for } 1 < x \le n, & K((-,1),(+,1)) = 1-\frac{1}{n} \\
K((-,x),(+,x-1)) & = & \frac{1}{n} & \text{for } 1 < x \le n, & K((-,1),(-,1)) = \frac{1}{n}
\end{array}
\tag{2.2}
$$

The enter and exit weights of all states is 1, the transition matrix is doubly stochastic, and thus the stationary distribution of this chain is uniform on the new state space, with all states having probability $1/2n$. The marginal distribution of just the second component of state (ignoring the $+$ or $-$) is therefore also uniform. This chain is thus an alternative to the simple random walk as a way of sampling from the original state space.

The state space of the non-reversible walk can instead be labeled with elements of the circle $\mathbb{Z}_{2n}$ (mod $2n$). The walk can then be described equivalently as a Markov Chain on $\mathbb{Z}_{2n}$ with transition probabilities:

$$K(x, x+1) = 1 - \frac{1}{n} \qquad K(x, -x) = \frac{1}{n} \tag{2.3}$$

Pictorially:



This labeling is more convenient for the proofs.

In Section 5 we show how to generalize this method to work with a non-uniform distribution (though the efficiency gains may not always carry over to non-uniform distributions). We also discuss generalizations to higher-dimensional grids.

First, however, we analyse the convergence of the chain shown above with respect to total variation distance, in Section 3, and with respect to χ^2 distance, in Section 4. Somewhat surprisingly, these two convergence rates are different.

3 Total variation convergence of the non-reversible walk

Our first result is that order n steps are necessary and sufficient for convergence in total variation distance of the non-reversible walk. Let the distribution after ℓ steps starting from a be K_a^ℓ. The total variation distance is defined as

$$||K_a^\ell - \pi||_{TV} \;=\; \max_{\mathcal{A}\subseteq\mathcal{X}} |K_a^\ell(\mathcal{A}) - \pi(\mathcal{A})| \;=\; \frac{1}{2}\sum_{x\in\mathcal{X}} |K_a^\ell(x) - \pi(x)|$$

Theorem 1 *For any $n \geq 2$, any starting state a, and all $\ell = 1, 2, \ldots$, the chain (2.3) on $\mathbb{Z}_{2n}$ satisfies:*

$$||K_a^\ell - \pi||_{TV} \;\leq\; (1-C)^{\lfloor \ell/4n \rfloor}$$

for some constant C. (The direct proof below shows the theorem for $C = 2^{-7}$, the coupling proof for $C = 2^{-16}$. In both proofs, the constant could easily be improved.)

Conversely, for $n > 2$, the chain started at state 0 is not close to π after only n steps:

$$||K_0^\ell - \pi||_{TV} \;\geq\; \frac{7}{54} \quad \textit{for all } \ell \leq n$$

Proof of the converse: After $\ell \leq n$ steps, the walk started at state 0 is at ℓ with probability at least $(1-\frac{1}{n})^\ell \;\geq\; (1-\frac{1}{n})^n \;\geq\; 1/2n$, and hence, for $n > 2$:

$$||K_0^\ell - \pi||_{TV} \;\geq\; \left(1 - \frac{1}{n}\right)^n - \frac{1}{2n} \;\geq\; \frac{7}{54}$$

using $(1-1/n)^n \geq 8/27$ for $n > 2$ (since $(1-1/n)^n$ increases monotonically with n). The converse is not true for $n = 2$, for which the distribution is exactly uniform after two transitions.

$\diamond$

We prove the first part of Theorem 1 in two ways: by a direct probabilistic argument combined with sub-multiplicativity, and by a coupling argument.

3.1 A direct probabilistic proof

Let X_m be the position of the walk (2.3) at time m. We will show that for any starting state a and any state x, when $m = 4n$,

$$P_a\{X_m = x\} \;\geq\; \frac{C}{2n} \tag{3.4}$$

where here $C = 2^{-7}$.

The majorization (3.4) suffices to prove the theorem by an easy argument. Let $K(x,y)$ be a Markov chain on a finite state space $\mathcal{X}$. Suppose π is a stationary distribution for K and there are m, C such that $K^m(x,y) \geq C\pi(y)$, for all x, y. Then $||K_x^\ell - \pi||_{TV} \leq (1-C)^{\lfloor \ell/m \rfloor}$, for all ℓ.

To see this, suppose without loss that $m = 1$, then write

$$K(x,y) \;=\; C\pi(y) \;+\; (1-C)\left[\frac{K(x,y) - C\pi(y)}{1-C}\right]$$

This presents the transition probabilities as a mixture with π as one component. If T is the first time that a transition chooses π from this mixture, then at time T, the process is stationary. Indeed, T is a strong stationary time in the sense of [7]; this reference gives results that provide a bound on the total variation. An elementary proof may also be found in [24, Section 3.3]. For general m, we apply the above to K^m.

To prove 3.4, let $T_1, T_2, \ldots$ be the times that the walks changes sign (ie, when $x \to -x$ is chosen, including when $x = -x = 0$). Thus $1 \le T_1 < T_2 < T_3 < \cdots$. Let A_m be the number of sign change transitions in the sequence up to X_m (ie, $A_m = i$ when $T_i \le m < T_{i+1}$). Clearly,

$$P_a\{X_m = x\} \;\ge\; P_a\{X_m = x,\, A_m = 1\} \;+\; P_a\{X_m = x,\, A_m = 2\}$$

We must look both when $A_m = 1$ and when $A_m = 2$ because of a parity problem. From direct considerations, starting at a, with any m,

$$\begin{aligned} \text{Given } A_m = 1: \quad X_m &= (m - a + 1) - 2T_1 \pmod{2n} \\ \text{Given } A_m = 2: \quad X_m &= (m + a) + 2(T_1 - T_2) \pmod{2n} \end{aligned}$$

These equations show the parity problem: After an even number of transitions, the walk will be an even number of steps past its starting state whenever $A_m = 2$.

One can also directly see that

$$\begin{aligned} P_a\{T_1 = i,\, A_m = 1\} &= \frac{1}{n}\left(1 - \frac{1}{n}\right)^{m-1} \quad \text{for } 1 \le i \le m \\ P_a\{T_1 = i,\, T_2 = j,\, A_m = 2\} &= \frac{1}{n^2}\left(1 - \frac{1}{n}\right)^{m-2} \quad \text{for } 1 \le i < j \le m \end{aligned}$$

Now take $m = 4n$. If $(m - a + 1) - x$ is even,

$$\begin{aligned} P_a\{X_m = x,\, A_m = 1\} &= P_a\{(m - a + 1) - 2T_1 = x \pmod{2n},\, A_m = 1\} \\ &\ge \frac{1}{n}\left(1 - \frac{1}{n}\right)^{m-1} \ge \frac{2^{-7}}{2n} \end{aligned}$$

The first inequality follows from the existence of at least one value of T_1 in the range 1 to m for which $(m - a + 1) - 2T_1 = x \pmod{2n}$. The last inequality uses $(1 - 1/n)^n \ge 1/4$ for $n \ge 2$.

If $(m - a + 1) - x$ is odd, then $(m + a) - x$ is even, and

$$\begin{aligned} P_a\{X_m = x,\, A_m = 2\} &= P_a\{(m + a) + 2(T_1 - T_2) = x \pmod{2n},\, A_m = 2\} \\ &\ge m\,\frac{1}{n^2}\left(1 - \frac{1}{n}\right)^{m-2} \ge \frac{2^{-7}}{2n} \end{aligned}$$

Here, the first inequality comes from counting the number of values for T_1 and T_2 that make $(m + a) + 2(T_1 - T_2) = x \pmod{2n}$, given that $A_m = 2$. We can find this from the following count, for any d with $0 \le d < n$:

$$|\{(i, j) : j - i = d \pmod{n}\}| \;\ge\; (m - d) + (m - n - d) \;\ge\; m$$

The $m - d$ term comes from solutions $(1, d + 1)$, $(2, d + 2)$, $\ldots$, $(m - d, m)$. The $m - n - d$ term comes from solutions $(1, n + d + 1)$, $\ldots$, $(m - n - d, m)$. Thus the number of solutions is bounded below by $m = 4n$, uniformly in d. This proves 3.4 and so completes the proof.

$\diamond$

3.2 A proof using coupling

Theorem 1 may also be proved by a coupling argument. We imagine starting chains from all the $2n$ possible initial states. Each of these chains follows the transition probabilities (2.3), but dependencies are introduced between the chains to encourage them to "couple" — to all enter the same state, and remain in identical states thereafter. The total variation distance between the distribution after ℓ steps, from any starting state, and the stationary distribution, π, is bounded by the probability that not all the chains will have coupled within ℓ steps [20].

Let $X_{a,k}$ be the position of the chain started at state a after k transitions. We define the transitions (on $\mathbb{Z}_{2n}$) as follows:

$$X_{a,k} = \begin{cases} X_{a,k-1} + 1 & \text{if } F_k(X_{a,k-1}) = 0 \\ -X_{a,k-1} & \text{if } F_k(X_{a,k-1}) = 1 \end{cases}$$

Here $F_k(x)$ controls whether or not a sign change transition occurs at step k for any chain that is in state x. We define $F_k(x)$ in terms of a stream of indicators that move from right to left through the "upstairs" states, along with a corresponding stream moving from left to right through the "downstairs" states:

$$F_k(x) = \begin{cases} U_{x+k-1} & \text{if } 1 \le x \le n \\ D_{x+n+k-1} & \text{if } -n+1 \le x \le 0 \end{cases} \tag{3.5}$$

where U_1, U_2, ... and D_1, D_2, ... are independent Bernoulli random variables taking the value 1 with probability $1/n$.

Clearly, the definition of $F_k(x)$ results in the probability of a sign change transition being $1/n$. Furthermore, since the sign change indicators move in the opposite direction to the states in the chain, the decisions whether to change sign within any single chain are independent from one time to another. Transitions within any single chain are therefore as in chain (2.3).

We now show that with probability at least $C = 2^{-16}$, the chains started from all $2n$ possible initial states will couple within $4n$ transitions. Iterating, the probability that the chains will not all be coupled after ℓ transitions is no more than $(1-C)^{\lfloor \ell/4n \rfloor}$, from which Theorem 1 follows.

We consider the situation where D_1, D_2, ..., D_{4n} are all zero, and all of U_1, U_2, ..., U_{4n} are also zero, except that $U_i = 1$ and $U_j = 1$ for some i and j such that $n \le i < j \le 3n$ and $j - i$ is odd and greater than one. There are $n^2 - n$ such i, j pairs. The probability of such a situation arising is therefore

$$(n^2 - n)\,\frac{1}{n^2}\left(1 - \frac{1}{n}\right)^{8n-2} > 2^{-16}$$

using $(1 - 1/n)^n \ge 1/4$ for $n \ge 2$.

When this situation does occur, the chains from all starting states will couple, as illustrated in Figure 1. Suppose that i is even, and hence j is odd (the argument proceeds analogously in the reverse situation). A chain started in a state a for which a is odd will then not be affected by the indicator $U_i = 1$, since (3.5) implies that U_i can affect this chain only if at some time, k, we have $i = (a + k - 1) + k - 1$, which is not possible if i is even and a is odd. Such a chain will be affected by the indicator $U_j = 1$, however. Indeed, as a result of indicator $U_j = 1$, all such chains will be in state -1 after transition j, as illustrated on the right of Figure 1.

On the other hand, chains started in a state a for which a is even will be affected by the indicator $U_i = 1$, and subsequently also by the indicator $U_j = 1$. In detail, all these chains will be in state -1 after transition i, as illustrated on the left of Figure 1. The effect of $U_j = 1$ does not interfere with this, as long as $j - i > 1$. The chain started in state $-1 - i$ (which is odd) will also be in state

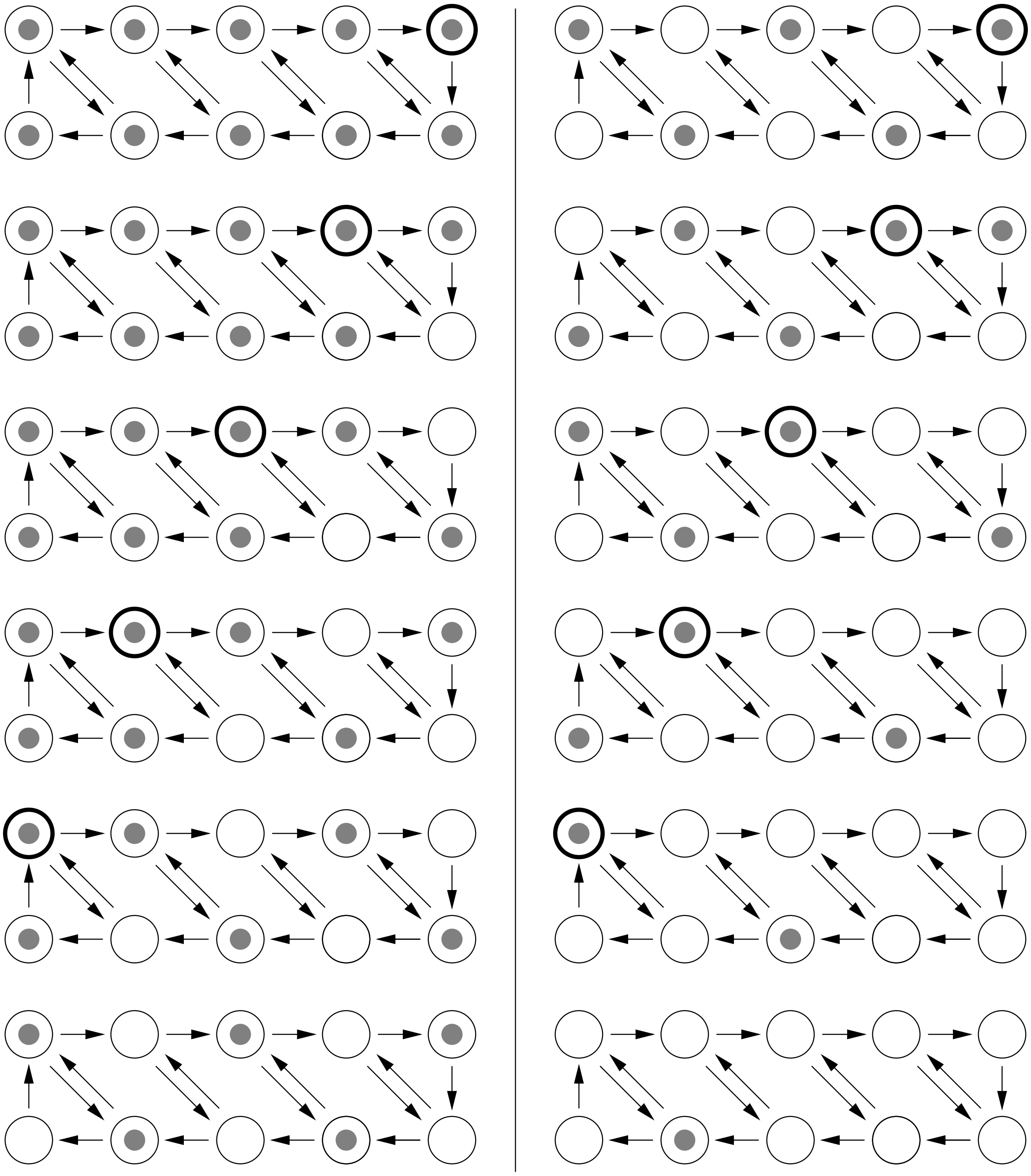


Figure 1: Illustration of the coupling proof. The diagrams parallel the one in Section 1, for $n = 5$. A dot in a state indicates that one or more of the chains from different start states is in that state at the given time; to begin, at the top left, no chains have coupled. A heavy circle indicates that the next transition for chains at the indicated state will be a sign change. The diagrams on the left show such a sign change indicator propagating to the left, and in the process moving all chains to a subset of states. The diagrams on the right show a second such propagation, occurring some time later, which has the effect of moving all chains to a single state. These two phases leading to coupling can also overlap, as long as the second phase starts more than one step after the first.

-1 at time i. As seen above, this chain, and hence also all the chains for which a is even, will be in state -1 at time j. We therefore see that all chains couple at time $j \leq 4n$ in this situation.

If this situation does not occur, we consider the possibility of the analogous situation involving $D_{4n+1}, D_{4n+2}, \ldots, D_{8n}$ and $U_{4n+1}, U_{4n+2}, \ldots, U_{8n}$. This leads to the conclusion that the chains will couple at some time from $4n+1$ to $8n$ with probability at least $C = 2^{-16}$. Iterating this argument, we see that the probability of the chains not coupling by iteration ℓ is no more than $(1-C)^{\lfloor \ell/4n \rfloor}$, from which Theorem 1 follows. $\diamond$

4 χ^2 convergence of the non-reversible walk

In this section, we determine the χ^2 rate of convergence of the non-reversible walk. The χ^2 (or ℓ^2) distance can be written as

$$\chi^2(\ell) = \max_x \sum_y \frac{(K^\ell(x,y) - \pi(y))^2}{\pi(y)} = \max_x \left\| \frac{K_x^\ell}{\pi} - 1 \right\|_2^2 = \|K^\ell - \pi\|_{2\to 2}^2$$

For these equivalences, see [10].

This χ^2 distance bounds total variation distance through

$$4\|K_x^\ell - \pi\|_{TV}^2 \leq \chi^2(\ell)$$

Usually the two distances give essentially the same answers for convergence. The present example is one of the few where they differ: As shown above, order n steps are necessary and suffice for total variation convergence; as shown below, order $n \log n$ steps are necessary and suffice for χ^2 convergence. We explain why this should be in Section 4.2.

The walk (2.3) changes direction at rate $1/n$. It is natural to ask how the change rate effects the speed of convergence. For example if the change rate is $1/2$, it is not hard to see that order n^2 steps are necessary and suffice for either total variation or χ^2 convergence. We will therefore analyze a one-parameter family of chains on $\mathbb{Z}_{2n}$ that generalize (2.3), with transition probabilities:

$$K(x, x+1) = 1 - \frac{c}{n} \qquad K(x, -x) = \frac{c}{n} \tag{4.6}$$

For any c in $(0, n)$ these chains have uniform stationary distribution, $\pi(x) = 1/2n$.

4.1 Bounds on the χ^2 distance

The main theorem of this section determines fairly sharp bounds on the χ^2 distance after ℓ steps. As explained after the statement, it shows that $\ell = \frac{n}{2c}(\log n + \theta)$ steps are necessary and sufficient for convergence.

Theorem 2 *Consider the chain (4.6) on $\mathbb{Z}_{2n}$, for fixed $c \in (0, \pi)$. For all sufficiently large n, and all ℓ:*

$$2n\left(1 - \frac{2c}{n}\right)^\ell \leq \chi^2(\ell) \leq \left(1 - \frac{2c}{n}\right) + 2n\left(1 - \frac{2c}{n}\right)^\ell \left\{1 + A(c) + O\left(\frac{1}{n}\right)\right\}$$

$$\textit{with } A(c) = \sum_{h=1}^{\infty} \frac{4c^2}{\pi^2 h^2 - c^2}$$

In Lemma 2 below we show that for the chain (4.6) $\chi^2(\ell)$ does not depend on the starting state. If

$$\ell = \frac{n}{2c}(\log n + \theta)$$

the lead term is asymptotic to $2e^{-\theta}$. So if θ is large (e.g. $\theta = 10$) the distance is small while if θ is small (e.g. $\theta = -10$), the distance is large. For ℓ in the crucial range, the time to stationarity is *decreasing* with increasing c. In Section 4.3 we determine the best value of c in $(0, n)$. Roughly this is $c = \sqrt{\log n}$. Then order $n\sqrt{\log n}$ steps are necessary and suffice for χ^2 convergence.

Theorem 2 will be proved as a sequence of lemmas which are also used in Section 4.3. The first step is an explicit diagonalization of the underlying transition matrix:

Lemma 1 *For any c, the chain K as defined in (4.6) is unitarily similar to a block diagonal matrix with two one-dimensional blocks at each extreme and $(n-1)$ two dimensional blocks. The one dimensional blocks have entries 1 and $-(1-\frac{2c}{n})$. The two dimensional blocks are:*

$$P_h = \begin{pmatrix} \left(1-\dfrac{c}{n}\right)e^{\frac{i\pi h}{n}} & \dfrac{c}{n} \\ \dfrac{c}{n} & e^{-\frac{i\pi h}{n}}\left(1-\dfrac{c}{n}\right) \end{pmatrix}, \qquad \textit{for } 1 \le h \le n-1 \tag{4.7}$$

Proof: The matrix K may be thought of as an operator on L, the $2n$-dimensional vector space of functions $f : \mathbb{Z}_{2n} \to \mathbb{C}$, via

$$Kf(j) = \sum_k K(j,k)f(k)$$

The matrix form (4.6) is with respect to the standard basis $\delta_j(k)$ of L.

Consider instead the Fourier basis $f_h(j)$, $-n < h \le n$:

$$\begin{aligned} f_0(j) &= 1 \\ f_h(j) &= e^{\frac{2\pi i h j}{2n}}, \quad 1 \le h < n \\ f_{-h}(j) &= e^{-\frac{2\pi i h j}{2n}}, \quad 1 \le h < n, \\ f_n(j) &= (-1)^j \end{aligned}$$

This basis, multiplied by $\frac{1}{\sqrt{2n}}$ is a unitary change preserving ℓ^2 norms.

The subspace L_h spanned by $[f_h, f_{-h}]$ is invariant under K giving P_h of (4.7) above as the matrix of the restriction of K to L_h. Further, $Kf_0 = (1-\frac{c}{n}) + \frac{c}{n} = 1 = f_0$ and $Kf_n(j) = (-1)^j \times -(1-2\frac{c}{n})$, proving the lemma.

◇

Lemma 1 reduces the computations to two by two matrices. It is of course equivalent to a treatment via representations of the dihedral group.

The next lemma shows that the initial starting state does not matter. Indeed, all rows of any power of the matrix K have the same entries (in permuted order). We find this surprising since the walk is not symmetric enough for us to see the result from invariance considerations. Indeed Lemma 2 does not hold for the walk on $\mathbb{Z}_{2n+1}$

Lemma 2 *For any c, the matrix K of (4.6) is such that for all x, all x', and all positive ℓ, there is a permutation σ for which*

$$K^\ell(x,y) = K^\ell(x', \sigma(y))$$

Proof: Let $\mathcal{C}$ be the basic circulant of size $2n$: a $2n \times 2n$ matrix with ones above the diagonal, a one in the lower left corner, and zeroes elsewhere (ie, $\mathcal{C}(i,j) = \delta_j(i+1 \,(\mathrm{mod}\, 2n))$. Let $\mathcal{P}$ be the basic Hankel matrix: a $2n \times 2n$ matrix with ones down the anti-diagonal (ie, $\mathcal{P}(i,j) = \delta_j(2n+1-i)$). Observe that $K = a\mathcal{C} + b\mathcal{PC}$ for some a, b.

Note that we have $\mathcal{PCPC} = Id$ and $\mathcal{P}^2 = Id$.

We claim that there are scalars x_i^ℓ and y_i^ℓ such that:

$$K^\ell = \sum_{i=0}^{2n-1} x_i^\ell \mathcal{C}^i + \sum_{i=0}^{2n-1} y_i^\ell \mathcal{PC}^i$$

with $x_i^\ell = y_i^\ell = 0$ if i and ℓ differ (mod 2).

This shows that $K^\ell = \mathcal{C}_1 + \mathcal{PC}_2$ for circulants $\mathcal{C}_1$ and $\mathcal{C}_2$, and that further, the non-zero entries in each row fall into disjoint subsets. Since $\mathcal{C}_1$ and $\mathcal{PC}_2$ have the same entries in each row, this proves the statement.

The claim is proved by induction, being clearly true when $\ell = 1$, and generally

$$\begin{aligned} (a\mathcal{C} + b\mathcal{PC})K^\ell &= (a\mathcal{C} + b\mathcal{PC}) \sum_{i=0}^{2n-1} (x_i^\ell \mathcal{C}^i + y_i^\ell \mathcal{PC}^i) \\ x_0^{\ell+1} = a x_{2n-1}^\ell + b y_1^\ell \qquad & x_{2n-1}^{\ell+1} = a x_{2n-2}^\ell + b y_0^\ell \\ x_{i+1}^{\ell+1} &= a x_i^\ell + b y_{i+2}^\ell \\ y_i^{\ell+1} &= b x_{i-1}^\ell + a y_{i+1}^\ell \end{aligned}$$

Using the inductive hypothesis, the claim and so, the lemma is proved.

◇

The next lemma gives the basic computational expression needed.

Lemma 3 *For any c and any starting x, the chain (4.6) satisfies:*

$$\chi^2(\ell) \;=\; \mathrm{Trace}(K^\ell K^{\ell *}) - 1 \;=\; \left(1 - \frac{2c}{n}\right)^{2\ell} + \sum_{1 \le h < n} T(h,\ell) \tag{4.8}$$

where $T(h,\ell) = \mathrm{Trace}(P_h^\ell P_h^{\ell *})$ and P_h is defined by (4.7).

Proof: From Lemma 2, the entries of any row of K^ℓ are a permutation of the first row.

Thus $\chi^2(\ell)$ does not depend on the starting state. We have

$$\chi^2(\ell) = 2n \sum_y (K^\ell(x,y))^2 - 1 = \mathrm{Trace}(K^\ell K^{\ell *}) - 1$$

The result now follows from Lemma 1.

◇

The following lemma is the heart of the argument, it gives an explicit diagonalization of the 2×2 blocks P_h. The alternative expressions given for the eigenvalues are needed in Section 4.3.

Lemma 4 *For any c, $T(h,\ell)$ in (4.8) of Lemma 3 is given by:*

$$T(h,\ell) \;=\; \frac{2}{|\lambda_- - \lambda_+|^2} \left\{ (1 - \tfrac{2c}{n} + \tfrac{2c^2}{n^2}) \left|(\lambda_-{}^\ell - \lambda_+{}^\ell)\right|^2 + \left|\lambda_+{}^{\ell+1} - \lambda_-{}^{\ell+1}\right| - \tfrac{2c}{n}(\lambda_-{}^\ell - \lambda_+{}^\ell)\overline{(\lambda_+{}^{\ell+1} - \lambda_-{}^{\ell+1})} \right\}$$

with λ_+ and λ_- the eigenvalues of the matrices P_h, (we have omitted the h in their symbols to ease the notation).

$$\lambda_\pm \;=\; \left(1-\frac{c}{n}\right)\left(\cos\frac{\pi h}{n} \pm \sqrt{\frac{c^2}{n^2(1-\frac{c}{n})^2} - \sin^2(\frac{\pi h}{n})}\right)$$

More specifically, if h is such that the eigenvalues have an imaginary part:

$$T(h,\ell) \;=\; 2\left(1-\frac{2c}{n}\right)^\ell \left[1+\frac{2c^2\sin^2(\ell\phi)}{n^2((1-\frac{c}{n})^2\sin^2(\frac{\pi h}{n})-\frac{c^2}{n^2})}\right], \quad \textit{with } \phi = \mathrm{Arg}(\lambda_-(h))$$

If h is such that the eigenvalues are real:

$$T(h,\ell) \;=\; 2\left(1-\frac{2c}{n}\right)^\ell + \left[1-\left(1-\frac{n}{c}\right)^2\sin^2\left(\frac{\pi h}{n}\right)\right]^{-1}\left(\lambda_+{}^{2\ell}+\lambda_-{}^{2\ell}-2\left(1-\frac{2c}{n}\right)^\ell\right)$$

Proof : This follows from an explicit diagonalization of P_h in (2.9). We give some details; throughout we write B for the matrix whose columns are the eigenvectors of P_h associated to λ_- and λ_+.

$$B = \begin{pmatrix} 1 & 1 \\ \alpha & \beta \end{pmatrix}$$

where α and β satisfy:

$$\alpha \;=\; \frac{\lambda_- - p\omega}{q} \;=\; \frac{q}{\lambda_- - p\bar{\omega}}, \qquad \beta \;=\; \frac{\lambda_+ - p\omega}{q} \;=\; \frac{q}{\lambda_+ - p\bar{\omega}}$$

where $\omega = e^{\frac{i\pi h}{n}}$, $p = 1-\dfrac{c}{n}$, and $q = \dfrac{c}{n}$. Further we have the identities:

$$B^{-1} \;=\; \frac{1}{\beta-\alpha}\begin{pmatrix} \beta & -1 \\ -\alpha & 1 \end{pmatrix}, \qquad \frac{1}{\beta-\alpha} \;=\; \frac{q}{\lambda_+ - \lambda_-}$$

Define

$$\Gamma^\ell = \begin{bmatrix} \lambda_-{}^\ell & 0 \\ 0 & \lambda_+{}^\ell \end{bmatrix} \quad \text{and} \quad R = P_h^\ell = B\Gamma^\ell B^{-1} = \begin{pmatrix} \beta\,\lambda_-{}^\ell - \lambda_+{}^\ell\alpha & \lambda_-{}^\ell - \lambda_+{}^\ell \\ (\alpha\beta)\,(\lambda_-{}^\ell - \lambda_+{}^\ell) & -\alpha\,\lambda_-{}^\ell + \lambda_+{}^\ell\beta \end{pmatrix}$$

Letting $C = \beta\,\lambda_-{}^\ell - \lambda_+{}^\ell\alpha$ and $D = -\alpha\,\lambda_-{}^\ell + \lambda_+{}^\ell\beta$, we always have $|C|^2 = |D|^2$. For real and complex cases alike, we also have $|\alpha\beta|^2 = 1$.

So that in the general case, whether real or complex, the following formula is valid:

$$\begin{aligned} T(h,\ell) \;&=\; \text{Trace}\,(P_h^\ell P_h^{\ell*}) \;=\; \sum_i\sum_j r_{ij}\overline{r_{ij}} \;=\; \frac{2q^2}{|\lambda_- - \lambda_+|^2}\left(|C|^2 + |\lambda_- - \lambda_+|^2\right) \\ &=\; \frac{2}{|\beta-\alpha|^2}\left(|C|^2 + |\lambda_- - \lambda_+|^2\right) \end{aligned}$$

We now separate the two cases, and use $\lambda_+\lambda_- = 1 - \frac{2c}{n} = p - q$. If the eigenvalues have an imaginary part, then

$$|C|^2 \;=\; |\lambda_+|^{2\ell}(2+|\beta-\alpha|^2) \;-\; (\lambda_+{}^{2\ell}+\lambda_-{}^{2\ell})$$

from which

$$T(h,\ell) \;=\; 2|\lambda_+|^{2\ell}\left(1+\frac{2q^2\sin^2\ell\phi}{|\lambda_+|^2\sin^2\phi}\right) \;=\; 2(p-q)^\ell\left(1+\frac{2q^2\sin^2\ell\phi}{p^2\sin^2(\frac{\pi h}{n})-q^2}\right)$$

If the eigenvalues are real, then

$$|C|^2 \;=\; ({\lambda_+}^{2\ell}+{\lambda_-}^{2\ell}) \;-\; (\lambda_+\lambda_-)^\ell(\alpha\bar\beta+\bar\alpha\beta)$$

and in this case:

$$\begin{aligned} T(h,\ell) &= \frac{2}{|\beta-\alpha|^2}(2|\lambda_- - \lambda_+|^2 - (\lambda_+\lambda_-)^\ell(\alpha\bar\beta+\bar\alpha\beta-2)) \\ &= 2(p-q)^\ell + \frac{4}{|\beta-\alpha|^2}(\lambda_- - \lambda_+)^2 \\ &= 2(p-q)^\ell + \frac{q^2}{q^2-p^2\sin^2(\frac{\pi h}{n})}(\lambda_- - \lambda_+)^2 \end{aligned}$$

After slight rearrangement, these give the formulas in Lemma 4.

◇

Proof of Theorem 2:

From lemma 4 we see that for $c \in (0,\pi)$ fixed and n sufficiently large, all the eigenvalues $\lambda_\pm(h)$ are complex, for $1 \le h \le n-1$.

Now, lemma 4 gives

$$T(h,\ell) \;=\; 2\left(1-\frac{2c}{n}\right)^\ell\left[1+\frac{2c^2\sin^2(\ell\phi)}{n^2((1-\frac{c}{n})^2\sin^2(\frac{\pi h}{n})-\frac{c^2}{n^2})}\right], \quad \text{with} \;\; \phi = \mathrm{Arg}(\lambda_-(h))$$

Bounding $2c^2\sin^2(\ell\phi)$ by $2c^2$ and using Taylor expansions for the denominator:

$$n^2\left[\left(1-\frac{c}{n}\right)^2\sin^2\left(\frac{\pi h}{n}\right)-\frac{c^2}{n^2}\right] \;=\; \left(1-\frac{c}{n}\right)^2 h^2\left[\pi^2+O\left(\left(\frac{h}{n}\right)^2\right)\right]-c^2 \;=\; h^2\pi^2-c^2+O\left(\left(\frac{h}{n}\right)^2\right)$$

This expansions is used for $1 < h \le \epsilon n$ for suitably small ϵ. For $\epsilon n \le h < \frac{n}{2}$, the denominator is bounded below by $\epsilon^2 n^2(1+O(\frac{1}{n}))$. Finally, $\sin^2(\frac{\pi h}{n}) = \sin^2(\frac{\pi(n-h)}{n})$. Combining bounds we have:

$$\chi^2(\ell) \;\le\; \left(1-\frac{2c}{n}\right)^{2\ell} \;+\; 2n\left(1-\frac{2c}{n}\right)^\ell\left\{1+A(c)+O\left(\frac{1}{n}\right)\right\}, \quad \text{with} \;\; A(c) = \sum_{h=1}^\infty \frac{4c^2}{\pi^2h^2-c^2}$$

and $O(\frac{1}{n})$ depending on c.

For the lower bound, use the fact that the second term in square brackets is positive for all h so $T(h,\ell) \ge (1-\frac{2c}{n})^\ell$. This completes the proof of theorem 2.2.

◇

4.2 Why χ^2 convergence takes order $n\log n$ steps

It is a bit surprising that the χ^2 convergence rate of the walk (2.3) is slower than its total variation convergence rate. This phenomenon can be traced to the deterministic behavior of the chain in the absence of sign change transitions.

For simplicity, take $c = 1$, and suppose that the chain starts in state at 0. The χ^2 distance from stationarity at time ℓ will be at least as big as the single term for the state $x = \ell \pmod{2n}$. The chance of being in this state will be at least $(1 - \frac{1}{n})^\ell$ (this is the chance of not having done any sign change transitions up to time ℓ). If this is greater than the stationary probability of $\frac{1}{2n}$, the contribution to the χ^2 distance from this state will be at least $2n((1 - \frac{1}{n})^\ell - \frac{1}{2n})^2$. When n is large, $(1 - \frac{1}{n})^\ell \approx e^{-\ell/n}$. Using this, we can see that after $\ell = n$ transitions, the χ^2 distance from stationarity is of order n. Only for ℓ of order $n \log n$ does the distance become small.

Preliminary computations indicate that χ^2 the convergence time can be reduced to order n by introducing a holding probability of $1/2$ in each state. That is, we use a new chain with transition probabilities

$$K(x, x) = \frac{1}{2} \qquad K(x, x+1) = \frac{1}{2} - \frac{1}{2n} \qquad K(x, -x) = \frac{1}{2n} \tag{4.9}$$

The holding probability of $1/2$ fuzzes out the behavior of the chain in the absence of a sign change transition. After ℓ transitions, this chain when started in state 0 will be in the vicinity of state $\ell/2$ with probability at least $(1 - \frac{1}{2n})^\ell$. However, the probability of the chain being in state $\ell/2$ exactly is smaller than this by a factor of order $\sqrt{\ell}$. Consequently, the contribution to the χ^2 distance after n steps of state $n/2$ is of order 1, not of order n, and the behavior of the original chain explained above is avoided. Thus, in terms of χ^2 distance, the holding probability of $1/2$ actually "speeds up" the chain, though convergence in terms of total variation distance is slowed down by a factor of two.

One might instead attempt to improve the χ^2 convergence rate by increasing the probability of a sign change transition. As we show in the next section, using a higher flip rate can indeed improve the χ^2 convergence time, but only to order $n\sqrt{\log n}$, not to order n. This is because the more frequent reversals of direction re-introduce a diffusive aspect into the chain's exploration of the state space.

4.3 χ^2 bounds with large flip rates

In this section we bound the rate of convergence in χ^2 distance when c is allowed to grow with n. The main results show that increasing the flip rates speeds up the chain for $c = c(n)$ up to order $\sqrt{\log n}$. Taking larger c then slows things down.

Theorem 3 *For the chain (4.6) with $c = c(n)$, and n sufficiently large,*

a) Suppose $c(n) \le a\sqrt{\log n}$ for fixed a. Then for any starting state x and for $\ell = \dfrac{An \log n}{c}$,

$$B_1 e^{-b_1 A} \le \chi^2(\ell) \le B e^{-bA}$$

with B_1, b_1, B, b positive continuous functions of a alone.

b) For $a\sqrt{\log n} < c(n) < a'\sqrt{\log n}$, any starting state x and $\ell = An\sqrt{\log n}$,

$$B_1 e^{-b_1 A} \le \chi^2(\ell) \le B e^{-bA}$$

with B_1, b_1, B, b positive continuous functions of a and a' alone.

c) For $c \geq a'\sqrt{\log n}$, any starting state x and $\ell = Anc$,

$$B_1 e^{-b_1 A} \leq \chi^2(\ell) \leq B e^{-bA}$$

with B_1, b_1, B, b positive continuous functions of a' alone.

Proof: From Lemma 3 we have, for any starting state x, any c and ℓ:

$$\chi^2(\ell) = \text{Trace}(K^\ell K^{\ell *}) - 1 = \left(1 - \frac{2c}{n}\right)^{2\ell} + \sum_{1 \leq h < n} T(h, \ell)$$

From Lemma 4, when $\lambda_\pm$ has an imaginary part:

$$T(h, \ell) = 2\left(1 - \frac{2c}{n}\right)^\ell \left[1 + \frac{2c^2 \sin^2(\ell\phi)}{n^2((1-\frac{c}{n})^2 \sin^2(\frac{\pi h}{n}) - \frac{c^2}{n^2})}\right], \quad \text{with } \phi = \text{Arg}(\lambda_-(h))$$

and when $\lambda_\pm$ is real:

$$T(h, \ell) = 2\left(1 - \frac{2c}{n}\right)^\ell + \left[1 - \left(1 - \frac{n}{c}\right)^2 \sin^2\left(\frac{\pi h}{n}\right)\right]^{-1} \left(\lambda_+{}^{2\ell} + \lambda_-{}^{2\ell} - 2\left(1 - \frac{2c}{n}\right)^\ell\right)$$

Here, we will write $\lambda_\pm$ from Lemma 4 as

$$\lambda_\pm = \left(1 - \frac{c}{n}\right)\left(\cos\frac{\pi h}{n} \pm \sqrt{\frac{c^2}{n^2(1-\frac{c}{n})^2} - \sin^2(\frac{\pi h}{n})}\right) = \left(1 - \frac{c}{n}\right)\left(\cos\frac{\pi h}{n} \pm \sqrt{\Omega}\right)$$

where Ω defined as follows:

$$\Omega = \frac{c^2}{n^2(1-\frac{c}{n})^2} - \sin^2\left(\frac{\pi h}{n}\right) = \frac{c^2 - \pi^2 h^2}{n^2} + 2\frac{c^3}{n} + O\left(\frac{h^4}{n}\right).$$

Let $h^* = h^*(c, n)$ be the smallest h so that the eigenvalues are imaginary. We will first treat the case where $c \leq \Delta \log n$, where Δ is a fixed constant. Then $h^* = \frac{c}{\pi} + O(1)$, and we partition the sum composing $\chi^2(\ell)$ into two zones:

Zone 0: $h < h^*, h > n - h^*$, here the eigenvalues of P_h are real.

Zone 1: $h^* \leq h \leq n - h^*$, here the eigenvalues of P_h have imaginary parts.

In Zone 0 we have to approximate the various terms that appear in $T(h, \ell)$. Using Taylor expansions we have:

$$\left(1 - \frac{\pi^2 h^2}{c^2}\right)^{-1} - 2\, c\pi^2 h^2 \left(c^2 - \pi^2 h^2\right)^{-1} \left(1 - \frac{\pi^2 h^2}{c^2}\right)^{-1} n^{-1} + O(n^{-2})$$

$$\left(1 - \left(1 - \frac{n}{c}\right)^2 \left(\sin(\frac{\pi\, h}{n})\right)^2\right)^{-1} = 1 + \frac{\pi^2 h^2}{c^2} + O\left(\frac{h^4}{c^2 n^2}\right), \quad 1 \leq h \leq h^*$$

This is thus bounded by a constant (possibly depending on Δ) for all c and for $h \leq h^*$.

Further, Taylor expansions for the eigenvalues in Zone 0 with $c \leq \Delta \log n$ provide:

$$\lambda_-(h) \;=\; 1 - \frac{\pi^2 h^2}{2nc} + O\left(\frac{h^2}{n^2}\right), \qquad \lambda_+(h) \;=\; 1 - \frac{2c}{n} + \frac{\pi^2 h^2}{2nc} + O\left(\frac{h^2}{n^2}\right)$$

From these bounds we see that for some $D = D(\Delta)$,

$$\sum_{h \leq h^*} T(h,\ell) \;\leq$$

$$D\left\{2\left(1-\frac{2c}{n}\right)^{\ell} + \sum_{h \leq \frac{c}{\pi}} \left[\left(1 - \frac{\pi^2 h^2}{2nc} + O\left(\frac{h^2}{n^2}\right)\right)^{2\ell} + \left(1 - \frac{2c}{n} + \frac{\pi^2 h^2}{2nc} + O\left(\frac{h^2}{n^2}\right)\right)^{2\ell}\right]\right\} \quad (4.10)$$

Essentially the same bounds hold for the large elements in Zone 0: if $h' = n - h$, then $\lambda_+(h') = -1 + \frac{\pi^2 h^2}{2nc} + O(\frac{h^2}{n^2})$, and $\lambda_-(h') = -1 + \frac{2c}{n} - \frac{\pi^2 h^2}{2nc} + O(\frac{h^2}{n^2})$. Thus the right hand side of 4.10 (with a different D), bounds the sum over Zone 0.

For Zone 1 we use the techniques of section 4.1 to get the bound

$$\sum_{\text{Zone } 1} T(h,\ell) \;\leq\; 2n\left(1-\frac{2c}{n}\right)^{\ell}\left(1 + \sum_{h > h^*} \frac{4c^2}{\pi^2 h^2 - c^2}\right) \quad (4.11)$$

All claims follow from (4.10) and (4.11). Consider first $c \leq A'\sqrt{\log n}$ and take $\ell = A\frac{n \log n}{c}$ This choice makes the bound 4.11 small for A large; certainly the first and last terms in 4.10 are small also:

$$\sum_{h? \leq c} \left(1 - \frac{\pi h^2}{2nc}\right)^{2\ell} \;\leq\; ce^{-\pi A h^2 \log n \,/\, c^2} \;\leq\; ce^{-\pi A h^2 \,/\, A'^2}$$

These bounds give part (a).

For part (b), the sum (4.11) is bounded above by a constant times $ne^{-\frac{2c\ell}{n}}$. Here, $\ell = An\sqrt{\log n}$ and $c \geq a\sqrt{\log n}$, so this term is small for A large. For the sum in (4.10), again the first and the last terms are handled by the argument above, for the middle term:

$$\sum_{h \leq \frac{c}{\pi}} \left(1 - \frac{\pi h^2}{2nc}\right)^{2\ell} \;\leq\; \sum_{h=1}^{\infty} e^{-\pi^2 h^2 A \sqrt{\log n} \,/\, c} \;\leq\; \sum_{h=1}^{\infty} e^{-\pi^2 h^2 A \,/\, a}$$

This being small for A large.

The argument for part (c) is similar, now the terms in zone 0 dominate and ℓ of order nc suffices to make all parts small.

This completes the proof if $c \leq \Delta \log n$. A similar, slightly easier argument suffices for larger c, we omit further details.

◇

In Theorem 2 we determined the rate of convergence carefully enough to find the cutoff in the χ^2 distance at $\frac{n}{2c}(\log n + \theta)$. Martin Hildebrand [18] has shown us preliminary results which imply that with flip rates c/n, and $c = c(n)$ tending to infinity, order cn steps are necessary and suffice for convergence in total variation distance. His argument uses the probabilistic tools as in Section 3.1 and shows that there is no cutoff phenomenon in total variation.

In Theorem 3 we have been content to determine rougher bounds. Preliminary computations show a cutoff of a more complicated type.

5 Generalizations and relationships to other methods

In this section, we show some ways in which the non-reversible walk of Section 2 can be generalized, and discuss relationships to previous sampling methods that exploit non-reversibility.

5.1 Non-uniform distributions in one dimension

We first show how to generalize the non-reversible one-dimensions walk to sample from a non-uniform distribution. Let $\pi(x)$ be a strictly positive distribution on $\mathcal{X} = \{1, 2, \ldots, n\}$. As in Section 2, we extend the state space to

$$\widetilde{\mathcal{X}} \;=\; \{(z,\, x) \;:\; z \in \{-1, +1\},\; x \in \mathcal{X}\}$$

The probabilities on the extended state space are given by $\widetilde{\pi}(z,\, x) = \;\pi(x)/2$.

We now construct a chain $\widetilde{M}$ that will sample from $\widetilde{\pi}$ on $\widetilde{\mathcal{X}}$. Each transition of $\widetilde{M}$ involves two steps. The second step depends on a parameter θ, which can be any fixed value in (0,1).

Transitions for chain $\widetilde{M}$:

1. From $(z,\, x)$, try to move to $(-z,\, x+z)$ via a standard Metropolis step. This proposal is symmetric, and so should be accepted with probability

   $$a((z,\, x)) \;=\; \min\left[1,\; \frac{\widetilde{\pi}(-z,\, x+z)}{\widetilde{\pi}(z,\, x)}\right] \;=\; \min\left[1,\; \frac{\pi(x+z)}{\pi(x)}\right]$$

   If $x+z$ is outside the range 1 to n, we set $a((z,\, x)) = 0$. We randomly accept the proposal with probability $a((z,\, x))$, and set the state after Step 1 to $(z',\, x') \;=\; (-z, x+z)$ if the proposal is accepted, or to $(z',\, x') \;=\; (z,\, x)$ if the proposal is rejected.

2. With probability $1-\theta$, the chain moves to $(-z',\, x')$; otherwise (with probability θ), the chain stays at $(z',\, x')$.

Proposition 1 *The chain $\widetilde{M}$ described above is an irreducible aperiodic chain on $\widetilde{\mathcal{X}}$ with stationary distribution $\widetilde{\pi}(z,\, x) \;=\; \pi(x)/2$.*

Proof: Both steps in the transitions for $\widetilde{M}$ leave the distribution $\widetilde{\pi}$ invariant: the first step because it follows the usual construction of the Metropolis algorithm, the second because $\widetilde{\pi}(z, x) = \widetilde{\pi}(-z, x)$. Since $0 < \pi(x) < 1$ (provided $n \geq 2$) the chain $\widetilde{M}$ is connected. Indeed there is positive probability of going from one state to another after $n+1$ steps. Since the probability of $\widetilde{M}$ remaining at state $(+1, n)$ is $\theta > 0$, the chain is aperiodic. This completes the proof.

$\diamond$

Note that the combined effect of the two steps making up a transition of $\widetilde{M}$ is such that with probability $1-\theta$, the chain will move either to state $(z,\, x+z)$, if the proposal in Step 1 is accepted, or to state $(-z,\, x)$, if this proposal is rejected. If we choose a small value for θ, the chain will therefore tend to continue moving in one direction until such time as a rejection occurs.

If $\pi(x)$ is uniform, one can easily see that chain $\widetilde{M}$ with $\theta = 1/n$ reduces to the non-reversible walk of Section 2, which was analysed in Sections 3 and 4. The more general chain described here was abstracted from Horowitz [19], as discussed further in Section 5.4.

The same idea can be applied to general state spaces. For example, to sample from $\pi(dx)$, on $\mathbb{R}$, an extended state space consisting of two copies of $\mathbb{R}$ could be used. One could then define two Metropolis base chains, one with a drift to the right, one with a drift to the left. This has been tried by Gustafson [17], who found that it produces moderate improvements over random walk Metropolis when used in a component-by-component updating scheme for sampling from a multivariate distribution.

One can also use this idea to make directed versions of other reversible chains. For example, suppose that each $x \in \{1, 2, , \ldots, n\}$ is associated with a neighborhood $N(x)$. The usual Gibbs sampler (heatbath) method samples from $\pi(x)$ restricted to $N(x)$. Instead of using symmetric neighborhoods, such as $N(x) = \{x-1,\, x+1\})$, one could instead use two asymmetric neighborhoods, such as $N_+(x) = \{x+1,\, x+2\}$ and $N_-(x) = \{x-1,\, x-2\}$, which are applied to two copies of the state space.

There is nothing special about working with two copies of $\mathcal{X}$, however. The fiber algorithm we present next can be seen as working with 2^d copies of a base space.

5.2 General finite state spaces: The fiber algorithm

Suppose that our state space, $\mathcal{X}$, can be partitioned in various ways into ordered "lines", with each partition corresponding to a "direction". We can then define a walk that proceeds from state x by choosing one of these directions and then making a step along the corresponding line that passes through x. As before, we will make these steps in a non-reversible manner. As a simple example, consider an $m \times n$ grid with horizontal lines of size n and vertical lines of size m. Other examples where this structure arises naturally are described in Sections 6.2 and 6.3.

In detail, suppose that along with $\mathcal{X}$ we are given a collection of partitions $P_1, P_2, \ldots, P_d$. That is, for each $i = 1, \ldots, d$, there is a partition $P_i = \{P_{ij}\}_{j=1}^{J_i}$ for which $\cup_j P_{ij} = \mathcal{X}$ and $P_{ij} \cap P_{ij'} = \emptyset$ for $j \neq j'$. Each index i corresponds to a direction. The parts P_{ij} are called the lines in direction i. We suppose that each line P_{ij} is linearly ordered. Further, suppose that $\mathcal{X}$ is connected in the sense that for each x, y in $\mathcal{X}$ there is a path $x_0 = x, x_1, \ldots, x_\ell = y$ such that each pair x_i, x_{i+1} are in a common line.

Finally, let π be a positive probability measure on the finite state space $\mathcal{X}$, from which we wish to sample.

We now define a Markov chain $\widetilde{M}_d$ on an extended state space, $\widetilde{\mathcal{X}} = \{-1, +1\}^d \times \mathcal{X}$. This chain is parameterized by a a set of positive probabilities, $\{w_i\}_{i=1}^d$, for choosing each of the d directions, and by a set of flip rates in the various directions, $\{\theta_i\}_{i=1}^d$, satisfying $0 < \theta_i < 1$. Each transition of the chain proceeds in three steps, as follows, supposing the chain is currently at (z, x):

Transitions for chain $\widetilde{M}_d$:

1. Randomly choose i from $\{1, \ldots, d\}$ according to the probabilities w_i.

2. Given this i, find the j for which x is in P_{ij}. Then try to move to $x^* = x + z$, where $x + z$ is the successor of x in P_{ij} if $z = +1$, or the predecessor of x in P_{ij} if $z = -1$.

   Accept the move to $x^* = x + z$ with probability $\min[1, \pi(x^*)/\pi(x)]$. If this move is accepted, the new state becomes (z^*, x^*), where z^* is the same as z except that $z_i^* = -z_i$. If the move is not accepted, the state is unchanged. Either way, call the state at this point (z', x').

3. With probability $1 - \theta_i$, negate the ith coordinate of z'; otherwise (with probability θ_i) keep all of z' unchanged. Keep all of x' unchanged regardless.

Proposition 2 *For a connected set of partitions into linearly ordered lines, the chain $\widetilde{M}_d$ above is aperiodic and irreducible, with stationary distribution $\tilde{\pi}(z, x) = \pi(x)\, 2^{-d}$ on $\widetilde{\mathcal{X}}$.*

Proof: The chain is a mixture of d chains, each of which will be shown to have the claimed stationary distribution. Suppose $\{P_{ij}\}_{j=1}^{J_i}$ is one of the partitions of $\mathcal{X}$. The last two steps above define a chain on $\{-1,+1\}\times\mathcal{X}$ driven by this ith partition. This chain is not connected (if $J_i > 1$). But proposition 1 above applied to each component, P_{ij}, shows that $\tilde{\pi}$ is a stationary distribution, for any flip rate θ_i.

Stationarity of $\tilde{\pi}$ with respect to the overall chain follows, since a convex combination of chains with a common stationary distribution has again this same stationary distribution.

The combinatorial connectedness condition translates into irreducibility of the chain. Finally each line in the chain offers holding probabilities at both ends so the chain is aperiodic. This completes the proof.

$\diamond$

Again, it is easy to generalize this construction to Euclidean and more general spaces. For example, to sample from a probability density $f(x)$ on $\mathbb{R}^d$, take P_i to be the partition of $\mathbb{R}^d$ into lines parallel to the ith coordinate axis, and for each i, consider two random walks with opposite drifts as proposals for Metropolis updates in this coordinate.

The potential difficulty with the fiber algorithm is that appropriate sets of lines must be found, preferably ones which will be effective in eliminating diffusive behavior. For a naturally-given grid, it is easy to define lines, but if the distribution is supported only on a connected subset of the grid, these lines might not be effective in eliminating diffusive behavior. Lines can also be defined in less obvious ways, as in the examples of Sections 6.2 and 6.3. Note that simulation of the chain above does not require that the lines be constructed explicitly, only that it be possible to move from the current point on a line to its successor or predecessor.

5.3 Comparison with iid Metropolis methods

It is instructive to compare the non-reversable algorithms described above with the Metropolis algorithm based on a uniform proposal distribution, independent of the current state, with the usual acceptance criterion being used to produce the desired stationary distribution, $\pi(x)$. Call this iid Metropolis chain M_u.

Suppose that the state space, $\mathcal{X}$, has N points, and let $\pi^* = \max_x \pi(x)$. Liu [22] shows that

$$\|M_u^\ell - \pi\|_{TV} \le \left(1 - \frac{1}{N\pi^*}\right)^\ell$$

We consider two examples for which $\mathcal{X} = \{1, \ldots, n\}^d$, for some n and d, and hence $N = n^d$. The fiber method of Section 5.2 could be applied to these examples in an obvious way, using "lines" along which just one of the d coordinates varies. Choosing θ_i of order $1/n$ would seem appropriate.

Example 1: Let $\pi(x) = z\, e^{-(x_1+x_2+\cdots+x_d)}$. The normalizing constant, z, is bounded uniformly in n for fixed d and the bound shows that order $n^d e^{-d}$ transitions are sufficient for stationarity. It is not hard to prove a lower bound showing that they are necessary as well. Thus here the iid Metropolis is slow. The analysis in [11] shows that the classical Metropolis algorithm, (and presumably the fiber algorithm as well) reaches stationarity in order nd steps for this example.

Example 2: Let $p(x)$ be a polynomial with non negative coefficients and maximum degree $|\alpha^*| = \alpha_1^* + \alpha_2^* + \cdots + \alpha_d^*$, for example $p(x) = x_1 + x_2 + \cdots + x_d$ or $p(x) = x_1 x_2 \cdots x_d$. Let $\pi(x) = z\, p(x)$. For large n, $z \sim a_\alpha^* n^{|\alpha^*|+d}$. Thus $\pi^* \sim \frac{c}{n^d}$, for c bounded. Now, Liu's result shows that the chain M_u reaches stationarity in a bounded number of steps. The analysis in [11] shows that the classical Metropolis algorithm requires order n^2 steps to reach stationarity. In line with the results of Section 3 we conjecture that order n steps are necessary and suffice for the directed walk.

5.4 Relationships to other non-reversible methods

The generalizations above extend the simple non-reversible walk of Section 2 to problems that may be of practical interest. Still, in several respects, these methods are not as general or as sophisticated as the practical non-reversible methods that inspired this investigation. The advantage of looking at simpler methods is of course the possibility of more detailed analysis. We briefly discuss here some relationships between the methods of this paper and non-reversible methods that are presently used in quantum field theory [31, 21] and in some statistical applications [26, 16].

The one-dimensional walk of Section 5.1 is closely related to the "guided Monte Carlo" method of Horowitz [19]. The context is rather different, however. Horowitz's method applies to continuous state spaces (eg, $\mathbb{R}^d$), and assumes that the partial derivatives of the density function with respect to the coordinates can be computed. As in the methods of this paper, this state space is extended, by the inclusion of "momentum" variables, equal in number to the original "position" variables, with independent Gaussian distributions. A Hamiltonian dynamical system is defined, which when simulated moves the state along a contour of the probability density in the extended state space. The volume-preserving property of Hamiltonian dynamics ensures that this motion leaves the desired distribution invariant. When combined with other suitable updates to the momentum variables, this can lead to an ergodic Markov chain that samples from the desired distribution. The chain is non-reversible, with the momentum acting to keep the chain moving in one direction for a substantial period of time.

The relationship to the walks on discrete spaces discussed in this paper comes about from the necessity of simulating the Hamiltonian dynamics using some discretization of time into steps. When using such a discretization, the probability density will no longer be exactly constant along the path. This error is corrected using a Metropolis step, as in Step 1 of the transitions in Section 5.1. As in Step 2 there, the trick of negating the direction after the Metropolis step (which itself proposes a negation) produces a non-reversible chain that reverses direction only when a rejection occurs. (In Horowitz's method, θ is fixed at zero; an effect similar to a non-zero θ is produced by other means.)

The result is similar to the fiber algorithm of Section 5.2, with sets of "lines" that are trajectories of the discretized dynamics. This elaborate construction has two advantages over simpler schemes. First, the trajectories will in many cases follow the high-probability regions of the state space, even when these regions are not aligned with the coordinate axes, and may indeed be curved. In contrast, a simple scheme based on coordinate lines will tend to behave diffusively when there are strong dependencies that prevent large movements in any one direction. Second, the rejection rate can be controlled by adjusting the size of the time step used in simulating the dynamics. A high rejection rate that would lead to frequent reversals of direction can thereby be avoided.

Horowitz's method was derived from the "Hybrid Monte Carlo" method [13], in which the dynamics is simulated for many time steps, with a Metropolis acceptance criterion being applied to the final state. The Markov chain for this method is reversible, but diffusive behavior is nevertheless avoided, if the simulated trajectories are long enough to move to distant parts of the distribution. The method of Section 5.1 could also be modified so that several steps were done before applying the Metropolis criterion (though literally stepping in this fashion makes sense only if states can be visited only by stepping through them in sequence). This approach is potentially advantageous when state probabilities vary substantially over short distances, but these variations tend to cancel over longer distances, as is typically the case for the discretization error in a simulation of Hamiltonian dynamics.

Overrelaxation [1, 25] is another way of constructing a non-reversible Markov chain, which can avoid diffusive behavior in many situations. Like the non-reversible walks discussed in this paper, the overrelaxation method uses transitions composed of steps that are individually reversible, but which produce a non-reversible chain when applied in sequence.

6 Examples of sampling using non-reversible chains

This section shows how the methods of Sections 5.1 and 5.2 can be applied in three examples: a non-uniform distribution in one dimension, contingency tables with specified marginal distributions, and distributions of permutations.

6.1 A V-shaped distribution in one dimension

We have tried applying the algorithm of Section 5.1 to several V-shaped distributions on the state space $\{1, 2, 3, \ldots, n\}$, with probabilities of the form

$$\pi(x) \;=\; \frac{1}{Z}\left(2\left|x-\frac{n}{2}\right|+C\right)$$

where Z is the appropriate normalizing constant. The value of the constant C determines how small the probability is at the bottom of the V is (ie, at state $n/2$).

Since distributions of this form have two "peaks", separated by low-probability states, one might expect the usual Metropolis algorithm with nearest neighbor proposals to have difficulty crossing from peak to peak. This is certainly true for exponential peaks, but things are somewhat better for polynomial peaks. For the linear peaks, as in the distribution above, available theory [11] shows that order $n^2 \log n$ steps are necessary and sufficient for the usual Metropolis chain to reach stationarity. Preliminary work of Hildebrand [18] suggests that order n^2 are necessary and suffice for convergence of the directed algorithm. Here, we show some numerical results that are consistent with such asymptotic behavior.

We tried using the following three methods to sample from V-shaped distributions:

1. The random walk Metropolis method, with nearest-neighbor proposal distribution (ie, from state x, we propose either $x-1$ or $x+1$, each with probability $1/2$).

2. The directed sampling method of Section 5.1, with switching probability of $\theta = 1/n$.

3. An "ideal" sampling method, for which the bottleneck at the bottom of the V is the only impediment to sampling. Each transition for this method consists of two steps. The first step applies only if the state is in the range 1 to $n/2$ (inclusive); it changes the state to one chosen from the stationary distribution conditional on the state being in this range. The second step is then applied if the (possible changed) state is in the range $n/2$ to n (inclusive); it too changes the state to one chosen from the stationary distribution conditional on the state being in this range.

All three methods were started from state 1 (for the directed method, the extended state $(+, 1)$).

The convergence in total variation over 4000 transitions for each of these methods is shown in Figure 2, for V-shaped distributions with various values of n and C. These plots were produced by

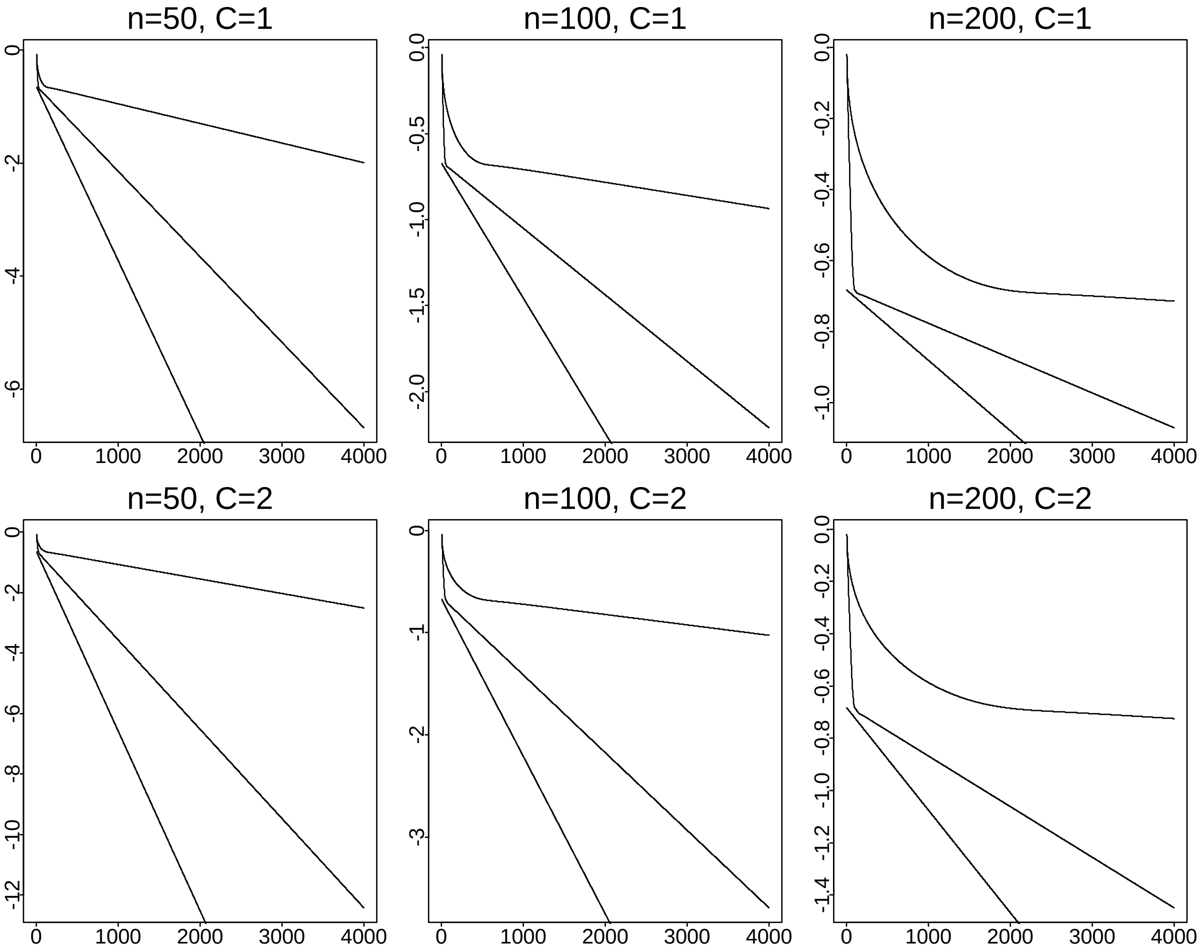


Figure 2: Convergence of Metropolis (top), directed (middle), and ideal (bottom) sampling methods on various V-shaped distributions (specified by n and C). The horizontal axis gives the number of transitions of a chain started in state 1 (for the directed method, state $(+, 1)$). The vertical axis gives the log of the total variation distance from the stationary distribution. (For the directed method, this is for the marginal distribution on the orginal state space; the total variation distance for the extended state space is very nearly the same.)

	Ideal	*Directed*	*Metropolis*	*Min. Prob.*
$C = 1,\ n = 50:$	0.00308	0.00151	0.000347	0.000769
$C = 1,\ n = 100:$	0.000785	0.000386	0.0000763	0.000196
$C = 1,\ n = 200:$	0.000198	0.0000979	0.0000170	0.0000495
$C = 2,\ n = 50:$	0.00593	0.00295	0.000479	0.00148
$C = 2,\ n = 100:$	0.00154	0.000758	0.000102	0.000385
$C = 2,\ n = 200:$	0.000392	0.000193	0.0000220	0.0000980

Figure 3: Convergence rates of the three methods, for various V-shaped distributions. The rate is the value of r for which total variation distance goes down with t in proportion to e^{-rt}, asymptotically. The last column is the minimum probability in the distribution (at the bottom of the V).

successive multiplication of a vector of probabilities by the transition matrix for the method, not by simulation.

For all distributions, the ideal method was best, followed by the directed method, with the Metropolis method being worst. Figure 3 gives numerical convergence rates for each method and distribution. These were measured from the slope of the lines in Figure 2 at iteration 4000, except for the Metropolis method with $n = 200$, for which the chain was continued up to iteration 10000 in order to obtain an accurate answer. The figure also gives the minimum probability for each distribution ($\pi(n/2)$, the probability at the bottom of the V).

The convergence rate for the ideal method is always four times the probability of the state at the bottom of the V. The rate for the directed method is always slightly less than twice the minimum probability (and hence slightly less than half that of the ideal method). The Metropolis method is always slower than the directed method, by factors ranging from 4.35 to 8.77 for the runs shown in the figures. The difference is greater for larger values of n and of C. For $C = 0.1$ and $n = 100$, we found that the directed method was faster than Metropolis by a factor of only 2.34, and for $C = 0.01$ and $n = 100$, it was faster by a factor of only 2.02.

The results in the limit as $C \to 0$ (with n fixed) can be explained by assuming that all the methods will in this case reach stationarity within a peak in much less time than is typically needed to move from one peak to the other (passing through the lowest-probability state). In this situation, what matters is the probability of moving between peaks; the convergence rate will just be twice this probability. The ideal method will move between peaks whenever it is in state $n/2$ after either the first or second step of its transition. The probability of such a move is therefore $2\pi(n/2)$. The directed method makes such a move whenever it is in state $(+, n/2)$ or $(-, n/2)$, which occurs with probability $\pi(n/2)$. The Metropolis method makes a move between peaks only half of the time when it is in state $n/2$, since it may jump back the way it came; its probability of moving between peaks is thus $\pi(n/2)/2$.

We therefore see that when there are extreme barriers to movement between peaks ($C \to 0$), the directed method has only a factor of two advantage over the random walk Metropolis method. However, when the barriers are more moderate (larger values of C), the advantage of the directed method over Metropolis is larger, and grows with n. The data shown in Figure 3, along with additional data for $n = 26$, are consistent with an order n^2 convergence rate for the directed method, and with the expected $n^2 \log n$ convergence rate for the Metropolis method.

6.2 Contingency tables

Consider the problem of generating a random $I \times J$ table with fixed row and column sums and non-negative integer entries. This problem was posed by Diaconis and Efron [6] who give statistical motivation. Diaconis and Gangolli [8] give a host of other applications. Even for small I and J, the size of the state space can be huge. Consider the 4×4 table below:

	Black	Brunette	Red	Blonde
Brown	68	20	15	5
Blue	119	84	54	29
Hazel	26	17	14	14
Green	7	94	10	16

There are approximately 10^{15} tables with these same margins.

Diaconis and Sturmfels [12] suggested the following algorithm for generating random tables:

1. Randomly choose a pair of different rows and a pair of different columns.

2. Choose one of the following two changes to the 2 by 2 square thus defined, with equal probabilities:

$$\begin{pmatrix} + & - \\ - & + \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} - & + \\ + & - \end{pmatrix}$$

3. Make the chosen change, unless it would result in a table value becoming negative.

This defines a Markov chain that is a symmetric, connected, and aperiodic, with uniform stationary distribution on the set of all tables with the given row and column sums.

The walk described above has a diffusive behavior taking an order $(Diameter)^2$ steps to reach stationarity. This is proved by Chung, Graham, and Yau [4] for tables with large row and column sums and by Diaconis and Saloff-Coste [10] for small values of I and J.

One can try to avoid this diffusive behavior by applying the method of Section 5.2 in an obvious way, taking the lines to be determined by a pair of rows and columns and moving along these lines in a directed fashion. We have done this, and found that the directed method does indeed work much faster than the reversible random walk.

A host of other statistical problems can also be solved by an extension of the random walk algorithm given above. We give a general description here; see [12] for statistical motivation.

Let $\mathcal{X} = \{x \in \mathbb{N}^n : Ax = y\}$, where A is a specified $m \times n$ matrix with non-negative entries, and y is an m-vector with non-negative entries. In applications, $\mathcal{X}$ will be finite and non-empty.

The problem is to sample from the uniform distribution on $\mathcal{X}$. The random walk approach of [12] is defined in terms of a set of Markov Basis vectors, $v_1, v_2, \ldots v_k \in \mathbb{Z}^n$, which satisfy:

(1) $Av_i = 0$.

(2) For any x and x' in $\mathcal{X}$, there is a positive integer, ℓ, indices $i_1, i_2, \ldots i_\ell$, and signs $z_1, z_2, \ldots z_\ell$ in $\{\pm 1\}$ such that:

$$x' \;=\; x + \sum_{j=1}^{\ell} z_j v_{ij} \quad \text{and} \quad x + \sum_{j=1}^{a} z_j v_{ij} \;\geq\; 0 \;\; \text{for } 1 \leq a \leq \ell$$

Condition (1) ensures that $A(x + v_i) = y$ when $x \in \mathcal{X}$. Condition (2) says there is a path between each x and x' in $\mathcal{X}$, found by adding or subtracting v_i while staying in $\mathcal{X}$.

The Markov chain for sampling from $\mathcal{X}$ operates as follows: When in state x, choose one of the v_i at random, and move to $x + v_i$ provided this is in $\mathcal{X}$, otherwise stay at x. This chain reduces to the chain described above for tables with an appropriate choice of A. It appears to have diffusive behavior in general.

The above set of problems can be solved more rapidly using the fiber algorithm of Section 5.2. Observe that the lines $\{x + jv_i\}_{j \in \mathbb{Z}} \cap \mathcal{X}$ partition $\mathcal{X}$ as x varies. Varying i gives a collection of "directed" partitions, $P_1, P_2, \ldots, P_k$, which satisfy the conditions of Proposition 2.

6.3 Permutations

Let $\mathcal{X} = \mathcal{S}_n$ be the set of permutations on n letters, and let $d(\sigma, \eta)$ be a metric on $\mathcal{S}_n$. To fix ideas, consider

$$d(\sigma, \eta) \;=\; \sum |\sigma(i) - \eta(i)| \qquad \text{(Spearman's footrule)}$$

A non uniform probability distribution on $\mathcal{S}_n$ (Mallow's model), can be constructed as follows:

$$\pi(\sigma) \;=\; \theta^{d(\sigma,\sigma_0)} \,/\, Z$$

where Z is the appropriate normalizing constant. In the model above, $0 < \theta \leq 1$ is fixed, as is the location parameter σ_0. Again, just to fix ideas, consider $\sigma_0 = \mathrm{id}$, so that the distribution $\pi(\sigma)$ is largest at $\sigma = \mathrm{id}$ and falls off exponentially.

The problem is to draw samples from π, for instance when $n = 52$.

One approach is to use the Metropolis algorithm with base chain random transpositions. This seems to work well even in the uniform case ($\theta = 1$). Some analyses and references to background literature appear in [5].

To apply the directed method of Section 5.2 we must find a collection of ordered partitions. One natural construction uses the group structure of $\mathcal{S}_n$. Let H be a subgroup of $\mathcal{S}_n$ and P_H the partition of $\mathcal{S}_n$ into cosets of H. Taking all conjugates, $H^\sigma = \sigma^{-1} H \sigma$ gives a neat family of partitions. We consider three special cases:

1. $H = \mathcal{S}_n$. There is only one block in the partition. This must be ordered. One method is to use lexicographical order. A second method uses a Gray code based on transpositions [3, 9]. This linearizes the problem so that the method of Section 5.1 can be used. This is not a foolish approach; if the walk is started off at the identity it should be reasonably efficient.

2. $H = \{\mathrm{id},\, (1,2)\}$. Now the block of P_H containing the permutation σ consists of $\{\sigma,\, (1,2)\sigma\}$. Running over all the conjugates gives blocks of the form $\{\mathrm{id}, (x,y)\}$. We see that with these choices our generalized Metropolis algorithm reduces to the random transpositions algorithm described previously.

3. H is the cyclic group generated by a single permutation η. Now the block of the partition containing σ is $(\sigma,\, \eta\sigma,\, \eta^2\sigma,\, \ldots,\, \eta^{k-1}\sigma)$ where k is the order of η. For a practical version of the algorithm choose a small collection of permutations $\eta_1,\, \eta_2,\, \ldots,\, \eta_K$ that generate $\mathcal{S}_n$ and use these to generate partitions $P_1, P_2, \ldots, P_k$. This walk is connected.

We remark in closing that diffusive behavior does *not* occur when generating uniformly distributed random permutations by successive transpositions of randomly chosen pairs [5], nor when such random transpositions are used as a Metropolis proposal for sampling from a distribution over permutations of exponential form [11].

7 Scope and limitations of non-reversible sampling

We have shown in this paper that non-reversibility can be a desirable property of a Markov chain sampling method. This conclusion accords with observations of the behaviour of some practical non-reversible sampling methods [26, 17] and some previous theory (eg, [21]).

The methods we discuss have some limitations, however. As illustrated in Section 6.1, any local algorithm, including the non-reversible walk, can effectively get stuck when sampling from a multimodal distribution with extreme barriers to movement between peaks. Even with less extreme barriers, we saw that the non-reversible walk provided only a modest ($\log n$) improvement over a reversible walk for the V-shaped distribution. This is expected — no algorithm can overcome multimodality without some input of information that would allow the peaks to be located.

A more serious limitation is that the most general algorithm, of Section 5.2, must use suitable "lines" that proceed in various "directions" in the underlying state space. These may be difficult to find. If such directions are found, it may also be possible to use them to construct other algorithms that are even better than the non-reversible walk. One possibility is an iid Metropolis algorithm, as discussed in Section 5.3. For the contingency table example of Section 6.2, where a direction was specified by a pair of rows and a pair of columns, an alternative, implemented in [12], is to consider the four cells in these row and columns as a 2×2 table and chooses uniformly among all the 2×2 tables with the same margins. This is easy to do, since such a 2×2 table is specified by one entry, which varies between easily computed bounds. A similar comment holds for the more general problems described in [12].

Another limitation is that the $(diameter)^2$ convergence time associated with reversible random walks applies to uniform or relatively flat stationary distributions. When the distribution is highly non-uniform, a non-reversible walk might have little or no advantage. For example, available theory [11] shows that when a random-walk Metropolis algorithm is used to sample from a distribution on a low dimensional grid having exponential peaks, the walk basically heads directly for the nearest peak. Thus if the stationary distribution is unimodal order *diameter* steps suffice for stationarity.

This is not necessarily the whole story, however. Even if a random-walk Metropolis method heads toward the mode when started from a state far out in the tails of the distribution, it may nevertheless suffer from diffusive behaviour when exploring the high-probability portion of the state space. This can be seen in the simple case of a multivariate Gaussian distribution with high positive correlations, where non-reversible methods such as Hybrid Monte Carlo [13], Horowitz's method [19], and overrelaxation [1] can sample much more efficiently than Gibbs sampling and simple Metropolis methods [25]. A simple non-reversible walk using "lines" in the coordinate directions will not necessarily be adequate for such a situation, however.

Because of these limitations, directed walks may be most useful when the states making up a line have approximately equal probabilities, and when it is not easy to directly sample from a line, perhaps because the states within the line can be located only in a sequential fashion. This is essentially the situation with Horowitz's dynamical method [19]. The challenge is to find other such methods, especially for discrete state spaces where dynamical methods cannot be applied.

References

[1] Adler, S. L. (1981) "Over-relaxation method for the Monte Carlo evaluation of the partition function for multiquadratic actions", *Physical Review D*, vol. 23, pp. 2901-2904.

[2] Binder, K. (1979) *Monte Carlo Methods in Statistical Physics*, Berlin: Springer-Verlag.

[3] Conway, J., Sloane, N., Wilks, A. (1989) "Gray codes for reflection groups", *Graphs and Combinatorics*, vol. 5, pp. 315-325.

[4] Chung, F., Graham, R., Yau, S. T. (1996) "On sampling with Markov chains", *Random Structures and Algorithms*, vol. 9, pp. 55-77.

[5] Diaconis P. (1988) *Group Representations in Probability and Statistics*, Hayward, California: Institute of Mathematical Statistics.

[6] Diaconis P. and Efron, B. (1987) "Probabilistic-geometric theorems arising from the analysis of contingency tables", in A. E. Gelfand (editor), *Contributions to the Theory and Application of Statistics: A Volume in Honor of Herbert Solomon*, Boston: Academic Press.

[7] Diaconis P. and Fill J. (1990) "Strong stationary times via a new form of duality", *Annals of Probability*, vol. 18, pp. 1483-1522.

[8] Diaconis P. and Gangolli, A. (1996) "Rectangular arrays with fixed margins", in *Finite Markov Chain Renaissance*, Springer-Verlag IMA series, pp. 15-42.

[9] Diaconis, P. and Holmes, S. (1994) "Gray codes for randomization procedures" *Statistics and Computing*, vol. 4, pp. 207-302.

[10] Diaconis P. and Saloff-Coste L. (1992) "Moderate growth and random walk on finite groups", *Geometry and Functional Analysis*, vol. 4, pp. 1-36.

[11] Diaconis P. and Saloff-Coste L. (1995) "What do we know about the Metropolis algorithm?", *Procedings of the 27th Symposium on Theory of Computing*, pp. 112-129.

[12] Diaconis P. and Sturmfels B. (1997) "Algebraic algorithms for sampling from conditional distributions", to appear in *Annals of Statistics.*

[13] Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987) "Hybrid Monte Carlo", *Physics Letters B*, vol. 195, pp. 216-222.

[14] Gelfand, A. E. and Smith, A. F. M. (1990) "Sampling-based approaches to calculating marginal densities", *Journal of the American Statistical Association*, vol. 85, pp. 398-409.

[15] Geman, S. and Geman, D. (1984) "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721-741.

[16] Gustafson, P. (1997a) "Large hierarchical Bayesian analysis of multivariate survival data", *Biometrics*, vol. 53, pp. 230-242.

[17] Gustafson, P. (1997b) "A guided walk Metropolis algorithm", preprint.

[18] Hildebrand, M. (1997) "Rates of Convergence for a Directed Version of the Metropolis Algorithm", preprint.

[19] Horowitz, A. M. (1991) "A generalized guided Monte Carlo algorithm", *Physics Letters B*, vol. 268, pp. 247-252.

[20] Lindvall, T. (1992) *Lectures on the Coupling Method*, New York: Wiley.

[21] Kennedy, A. D. (1990) "The theory of hybrid stochastic algorithms", in P. H. Damgaard, *et al.* (editors) *Probabilistic Methods in Quantum Field Theory and Quantum Gravity*, New York: Plenum Press.

[22] Liu, J. (1996) "Metropolized independent sampling with comparisons to rejection sampling and importance sampling", *Statistics and Computing*, vol. 6, pp. 113-119.

[23] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953) "Equation of state calculations by fast computing machines", *Journal of Chemical Physics*, vol. 21, pp. 1087-1092.

[24] Neal, R. M. (1993) *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 144 pages. Available from the author's home page at `http://www.cs.utoronto.ca/~radford/`.

[25] Neal, R. M. (1995) "Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation", Technical Report No. 9508, Dept. of Statistics, University of Toronto, 22 pages. Available from the author's home page at `http://www.cs.utoronto.ca/~radford/`.

[26] Neal, R. M. (1996) *Bayesian Learning for Neural Networks* (Lecture Notes in Statistics No. 118), New York: Springer-Verlag.

[27] Roberts, G. O. and Sahu, S. K. (1997) "Updating schemes, correlation structure, blocking and parameterisation for the Gibbs sampler", to appear in the *Journal of the Royal Statistical Society B*.

[28] Sinclair, A. (1993) *Algorithms for Random Generation and Counting: A Markov Chain Approach*, Boston: Birkhäuser.

[29] Smith, A. F. M. and Roberts, G. O. (1993) "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods", *Journal of the Royal Statistical Society*, vol. 55, pp. 3-23.

[30] Tierney, L. (1994) "Markov Chains for exploring Posterior Distributions", *Annals of Statistics* vol. 22, pp. 1701-1762

[31] Toussaint, D. (1989) "Introduction to algorithms for Monte Carlo simulations and their application to QCD", *Computer Physics Communications*, vol. 56, pp. 69-92.