The Annals of Applied Probability
2000, Vol. 10, No. 3, 726–752

ANALYSIS OF A NONREVERSIBLE MARKOV CHAIN SAMPLER

By Persi Diaconis,[1] Susan Holmes and Radford M. Neal[2]

Stanford University, Stanford University and INRA and University of Toronto

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

1. Introduction. Markov chain sampling methods are commonly used in statistics [33, 32], computer science [31], statistical mechanics [3] and quantum field theory [34, 23]. 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 (e.g., the Metropolis algorithm [25, 13], and the Gibbs sampler [17, 16] with a random scan) the Markov chain constructed is reversible. Some of these methods explore the distribution by means of a diffusive random walk. We use the term "diffusive" for processes like the ordinary random walk on a d-dimensional lattice which require time of order T^2 to travel distance T. Some other common methods, such as the Gibbs sampler with a systematic scan, use a Markov chain that is not reversible, but have diffusive behavior resembling that of a related reversible chain [30].

Some Markov chain methods attempt to avoid the inefficiencies of such diffusive exploration. The Hybrid Monte Carlo method [15] uses an elaborate Metropolis proposal that can make large changes to the state. In a variant of this method due to Horowitz [21], a similar effect is produced using a Markov chain that is carefully designed to be nonreversible. (See [34, 23, 27] for reviews of these methods.) The overrelaxation method [1] also employs a nonreversible Markov chain as a way of suppressing diffusive behavior, as discussed in [29].

In this paper, we analyze a nonreversible Markov chain that does a one-dimensional walk, as an abstraction of these practical sampling methods, particularly that of Horowitz [21]. Gustafson [19] has also recently tried using adaptations of Horowitz's method. We find that the nonreversible walk does indeed converge more rapidly than the usual simple random walk. We ana-

Received June 1997; revised November 1999.

[1] Supported by NSF Grant DMS-95-04379.

[2] Supported in part by a grant from Natural Sciences and Engineering Research Council of Canada.

AMS 1991 *subject classifications*. Primary 60J10; secondary 65U05.

Key words and phrases. Nonreversible Markov chain, Markov chain Monte Carlo, Metropolis algorithm.

lyze 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.

Since this paper was submitted there have been several variations and extensions. Chen, Lovász and Pak [5] show how to use nonreversible walks to achieve speed-ups in a variety of other problems. Hildebrand [20] carries the careful analysis of Section 3 further and analyzes our "V" example of Section 6.1. Mira and Geyer [26] look at other measures of convergence for our basic example of Theorem 1. Finally, Bassiri [2] analyzes a random walk on dihedral groups similar to Theorem 2.

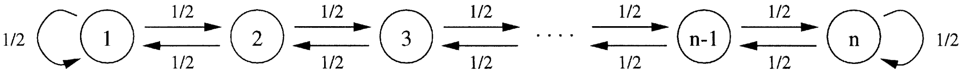
2. Reversible and nonreversible walks in one dimension.

All our examples concern distributions on some finite set, $\mathscr{X}$, with positive probabilities given by $\pi(x)$. We sample from $\pi(x)$ by running an irreducible aperiodic Markov chain on $\mathscr{X}$ 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

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

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 $\mathscr{X} = \{1, 2, 3, \dots, n\}$, and where the desired distribution is uniform: $\pi(x) \equiv 1/n$. 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; i.e., $K(x, y) = 1/2$ for $y = x \pm 1$ and $x, y \in \mathscr{X}$, and $K(1, 1) = K(n, n) = 1/2$ also. The chain can be pictured thus:

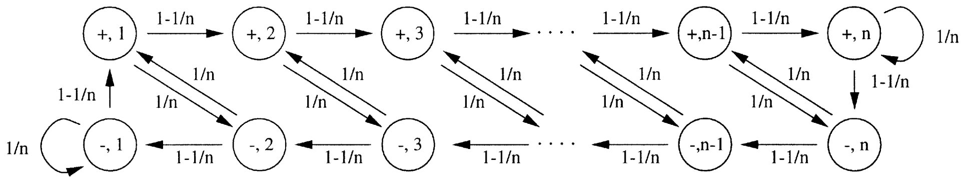


This walk takes on the order of n^2 steps to reach stationarity, since, using the central limit theorem, we see that the walk will take on the order of k^2 steps to travel a distance of order k.

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 $(+, 1), (+, 2), \dots, (+, n)$, and the downstairs states $(-, 1), (-, 2), \dots, (-, n)$. To get a uniform stationary distribution, we put holding probabilities of $1/n$ at the two diagonally opposed corners $(+, n)$

and $(-,1)$. The chain can then be pictured thus:



The transition probabilities are as follows:

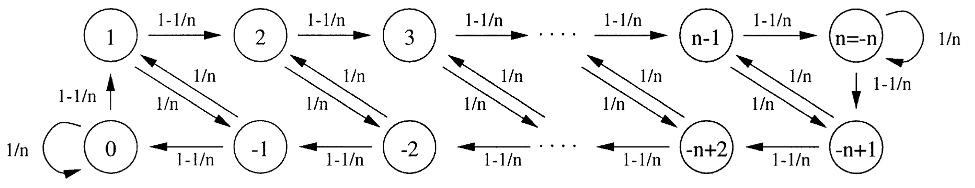
$$
(2.2)\qquad
\begin{aligned}
&K((+,x),(+,x+1)) = 1-\frac{1}{n} \quad \text{for } 1\le x<n, && K((+,n),(-,n)) = 1-\frac{1}{n},\\
&K((+,x),(-,x+1)) = \frac{1}{n} \quad \text{for } 1\le x<n, && K((+,n),(+,n)) = \frac{1}{n},\\
&K((-,x),(-,x-1)) = 1-\frac{1}{n} \quad \text{for } 1< x\le n, && K((-,1),(+,1)) = 1-\frac{1}{n},\\
&K((-,x),(+,x-1)) = \frac{1}{n} \quad \text{for } 1< x\le n, && K((-,1),(-,1)) = \tfrac{1}{n}.
\end{aligned}
$$

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 nonreversible walk can instead be labeled with elements of the circle $\mathbb{Z}_{2n}$ (integers mod $2n$). The walk can then be described equivalently as a Markov chain on $\mathbb{Z}_{2n}$ with transition probabilities

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

Pictorially,



This labeling is more convenient for the proofs.

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

First, however, we analyze 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 nonreversible 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

$$\|K_a^l - \pi\|_{TV} = \max_{\mathscr{A} \subseteq \chi} |K_a^l(\mathscr{A}) - \pi(\mathscr{A})| = \tfrac{1}{2} \sum_{x \in \chi} |K_a^l(x) - \pi(x)|.$$

THEOREM 1. *For any $n \geq 2$, any starting state a, and all $l = 1, 2, \dots,$ the chain* (2.3) *on $\mathbb{Z}_{2n}$ satisfies*

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

for some constant $C > 0$. (*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,*

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

PROOF OF THE CONVERSE. After $l \leq n$ steps, the walk started at state 0 is at l with probability at least $(1 - 1/n)^l \geq (1 - 1/n)^n \geq 1/2n$, and hence, for $n > 2$,

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

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, from any initial state. □

We prove the first part of Theorem 1 in two ways: by a direct probabilistic argument combined with submultiplicativity 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 $m = 4n$,

$$P_a\{X_m = x\} \geq \frac{C}{2n}, \tag{3.1}$$

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

The minorization (3.1) suffices to prove the theorem by an easy argument. Let $K(x, y)$ be a Markov chain on a finite state space $\mathscr{X}$. 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 $\|K_x^l - \pi\|_{\mathrm{TV}} \leq (1-C)^{\lfloor l/m \rfloor}$, 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 [9]; this reference gives results that provide a bound on the total variation. An elementary proof may also be found in [27], Section 3.3. For general m, we apply the above to K^m.

To prove (3.1), let $T_1, T_2, \dots$ be the times that the walks changes sign (i.e., when $x \to -x$ is chosen, including when $x = -x = 0$ or $x = -x = n$). Thus $1 \leq T_1 < T_2 < T_3 < \cdots$. Let A_m be the number of sign change transitions in the sequence up to X_m (i.e., $A_m = i$ when $T_i \leq m < T_{i+1}$). Clearly,

$$P_a\{X_m = x\} \geq P_a\{X_m = x, A_m = 1\} + P_a\{X_m = x, A_m = 2\}.$$

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

$$\text{Given } A_m = 1 : X_m = (m - a + 1) - 2T_1 \ (\mathrm{mod}\ 2n),$$
$$\text{Given } A_m = 2 : X_m = (m + a) + 2(T_1 - T_2) \ (\mathrm{mod}\ 2n).$$

(These equations show the parity problem: after an even number of transitions, the walk will have moved from its start state an odd number of steps if $A_m = 1$ and an even number of steps if $A_m = 2$.)

One can also directly see that

$$P_a\{T_1 = i, A_m = 1\} = \frac{1}{n}\left(1 - \frac{1}{n}\right)^{m-1} \quad \text{for } 1 \leq i \leq m,$$

$$P_a\{T_1 = i, T_2 = j, A_m = 2\} = \frac{1}{n^2}\left(1 - \frac{1}{n}\right)^{m-2} \quad \text{for } 1 \leq i < j \leq m.$$

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

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

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

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

$$P_a\{X_m = x, A_m = 2\} = P_a\{(m+a)+2(T_1-T_2) = x \ (\mathrm{mod}\ 2n), A_m = 2\}$$
$$\geq m\frac{1}{n^2}\left(1-\frac{1}{n}\right)^{m-2} \geq \frac{2^{-7}}{2n}.$$

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

$$\left|\{(i, j)\colon\ j - i = d \ (\mathrm{mod}\ n)\}\right| \geq (m-d) + (m-n-d) \geq m.$$

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 (3.1) and so completes the proof. □

3.2. *A proof using coupling.* Theorem 1 can also be proved by a coupling argument, with $C = 2^{-16}$. We imagine starting chains from all the $2n$ possible initial states. Each of these chains follows the transition probabilities (2.3), but these chains are coupled together by dependencies between their transitions, which encourage the chains to "coalesce," i.e., to all enter the same state and remain in the same state 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 coalesced within l steps [22].

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,

$$F_k(x) = \begin{cases} U_{x+k-1}, & \text{if } 1 \leq x \leq n, \\ D_{x+n+k-1}, & \text{if } -n+1 \leq x \leq 0, \end{cases} \tag{3.2}$$

where $U_1, U_2, \ldots$ and $D_1, D_2, \ldots$ 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 coalesce within $4n$ transitions. Iterating, the probability that the chains will not all have coalesced after l transitions is no more than $(1-C)^{\lfloor l/4n \rfloor}$, from which Theorem 1 follows.

We consider the situation where $D_1, D_2, \dots, D_{4n}$ are all zero, and all of $U_1, U_2, \dots, U_{4n}$ 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 coalesce, 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 (3.2) 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 -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 coalesce by time $j \le 4n$ in this situation.

If this situation does not occur, we consider the possibility of the analogous situation involving $D_{4n+1}, D_{4n+2}, \dots, D_{8n}$ and $U_{4n+1}, U_{4n+2}, \dots, U_{8n}$. This leads to the conclusion that the chains will coalesce 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 coalescing by iteration l is no more than $(1-C)^{\lfloor l/4n \rfloor}$, from which Theorem 1 follows. □

4. χ^2 convergence of the nonreversible walk.

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

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

Here $\|f\|_2 = (1/2^n)\sum |f(x)|^2$ and $\|A\|_{2\to 2}$ is the operator norm. For these equivalences, see [12].

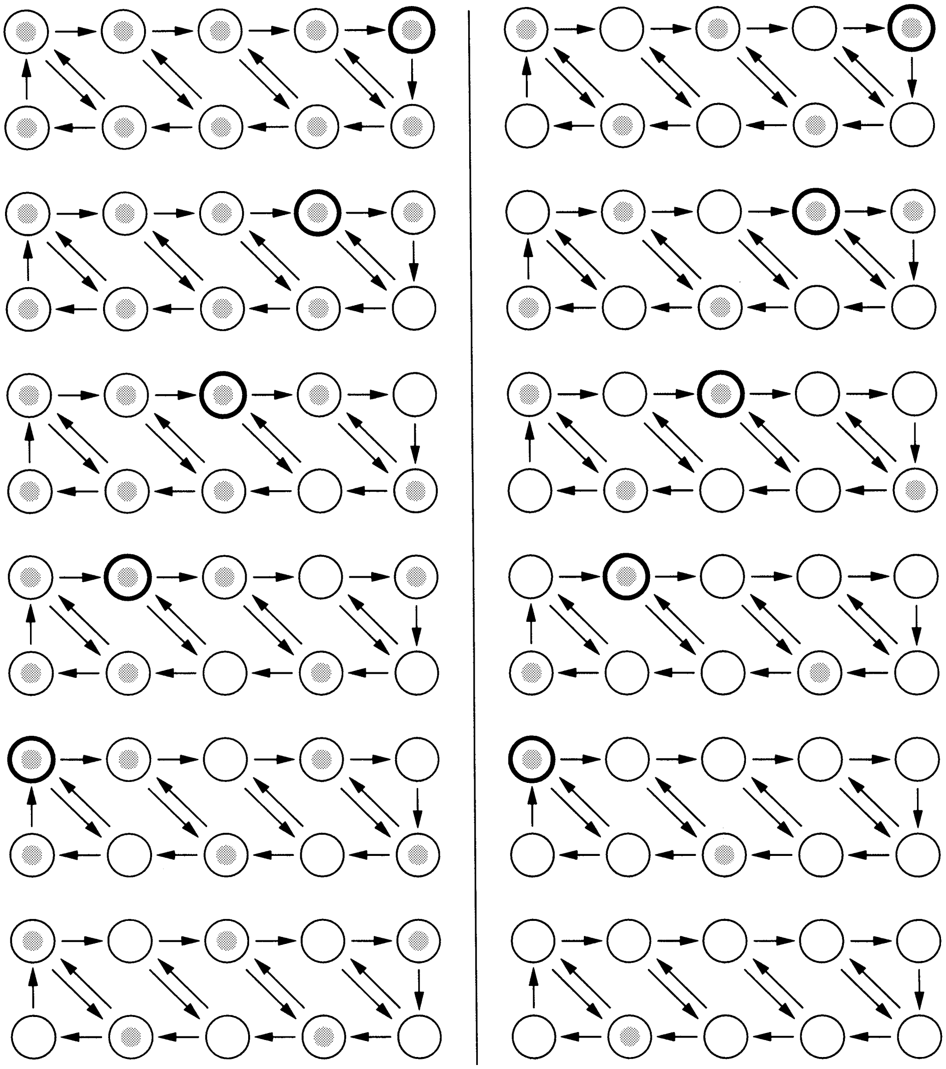


FIG. 1. *Illustration of the coupling proof. The diagrams parallel the one in Section* 1, *for* $n = 5$. *A dot in a state indicates that one or more of the chains from different start states is in that state at the given time*; *to begin, at the top left, no chains have coalesced. A heavy circle indicates 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 coalescence to a single state can also overlap, as long as the second phase starts more than one step after the first.*

This χ^2 distance bounds total variation distance through

$$4\|K_x^l - \pi\|_{\mathrm{TV}}^2 \le \chi^2(l).$$

Usually the two distances give essentially the same answers for convergence; Chapter 3 of [7] has many examples. 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

$$K(x,\ x+1) = 1 - \frac{c}{n}, \qquad K(x, -x) = \frac{c}{n}. \tag{4.1}$$

We often regard K as a matrix, with rows and columns corresponding to states $-n$ to $n-1$, in order.

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 $l = (n/2c)(\log n + \theta)$ steps are necessary and sufficient for convergence if c is fixed.

THEOREM 2. *Consider the chain* (4.1) *on $\mathbb{Z}_{2n}$, for fixed $c \in (0, \pi)$. For all sufficiently large n, and all l,*

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

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

In Lemma 2 below we show that for the chain (4.1), $\chi^2(l)$ does not depend on the starting state. If

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

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 preliminary computations 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. The first step is an explicit diagonalization of the underlying transition matrix.

Lemma 1. *For any c, the chain K as defined in* (4.1) *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* $-(1-2c/n)$. *The two-dimensional blocks are*

$$(4.2)\quad P_h = \begin{pmatrix} \left(1-\frac{c}{n}\right)\exp\left(\frac{i\pi h}{n}\right) & \frac{c}{n} \\ \frac{c}{n} & \exp\left(-\frac{i\pi h}{n}\right)\left(1-\frac{c}{n}\right) \end{pmatrix} \quad \textit{for } 1 \le h \le n-1.$$

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

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

The matrix form (4.1) is with respect to the standard basis $\{\delta_h\colon -n \le h < n\}$ of L.

Consider instead the Fourier basis $\{f_h\colon -n < h \le n\}$,

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

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

The subspace L_h spanned by $\{f_h,\, f_{-h}\}$ is invariant under K giving P_h of (4.2) above as the matrix of the restriction of K to L_h. Further, $Kf_0 \equiv (1-c/n) + c/n = 1 \equiv f_0$ and $Kf_n(j) = (-1)^j \times -(1-2(c/n))$, 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* (4.1) *is such that for all x, all x', and all positive l, there is a permutation σ for which*

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

PROOF. Let $\mathscr{C}$ 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. Let $\mathscr{P}$ be the basic Hankel matrix: a $2n \times 2n$ matrix with ones down the antidiagonal. Observe that $K = a\mathscr{C} + b\mathscr{P}\mathscr{C}$ for $a = 1 - c/n,\ b = c/n$.

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

$$K^l = \sum_{i=0}^{2n-1} x_i^l \mathscr{C}^i + \sum_{i=0}^{2n-1} y_i^l \mathscr{P}\mathscr{C}^i$$

with $x_i^l = y_i^l = 0$ if i and l differ mod 2.

This shows that $K^l = \mathscr{C}_1 + \mathscr{P}\mathscr{C}_2$ for circulants $\mathscr{C}_1$ and $\mathscr{C}_2$, and that, further, the nonzero entries in each row of $\mathscr{C}_1$ and of $\mathscr{C}_2$ fall into disjoint subsets. Since each of $\mathscr{C}_1$ and $\mathscr{P}\mathscr{C}_2$ has the same entries in each of its rows, the lemma follows from this.

The claim is proved by induction. It is clearly true when $l = 1$. Furthermore,

$$K^{l+1} = (a\mathscr{C} + b\mathscr{P}\mathscr{C})K^l = (a\mathscr{C} + b\mathscr{P}\mathscr{C}) \sum_{i=0}^{2n-1} (x_i^l \mathscr{C}^i + y_i^l \mathscr{P}\mathscr{C}^i).$$

Using $\mathscr{P}\mathscr{C}\mathscr{P} = \mathscr{C}^{-1},\ \mathscr{C}\mathscr{P}\mathscr{C} = \mathscr{P},\ \mathscr{C}^{2n} = \mathrm{Id}$, and the inductive hypothesis, K^{l+1} can be written in the required form with

$$x_0^{l+1} = ax_{2n-1}^l + by_1^l, \quad x_{2n-1}^{l+1} = ax_{2n-2}^l + by_0^l,\ x_i^{l+1} = ax_{i-1}^l + by_{i+1}^l \quad \text{for } 0 < i < 2n-1,$$

$$y_0^{l+1} = bx_{2n-1}^l + ay_1^l, \quad y_{2n-1}^{l+1} = bx_{2n-2}^l + ay_0^l,\ y_i^{l+1} = bx_{i-1}^l + ay_{i+1}^l \quad \text{for } 0 < i < 2n-1. \quad \square$$

The next lemma gives the basic computational expression needed.

LEMMA 3. *For any c and any starting x, the chain* (4.3) *satisfies*

$$\chi^2(l) = \mathrm{Trace}(K^l K^{l*}) - 1 = \left(1 - \frac{2c}{n}\right)^{2l} + \sum_{1 \le h < n} T(h,\ l), \tag{4.3}$$

where $T(h,\ l) = \mathrm{Trace}(P_h^l P_h^{l^*})$ and P_h is defined by (4.2).

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

Thus $\chi^2(l)$ does not depend on the starting state. We have for any x,

$$\chi^2(l) = 2n\sum_y\left(K^l(x,y) - \frac{1}{2n}\right)^2 = 2n\sum_y (K^l(x,\ y))^2 - 1 = \mathrm{Trace}(K^l K^{l^*}) - 1.$$

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.

LEMMA 4. *For any c, let $\lambda_+(h)$ and $\lambda_-(h)$ be the eigenvalues of the matrices P_h (we often omit the h in their symbols to ease the notation), then*

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

Further for $T(h,\ l)$ defined in (4.3), *if h is such that the eigenvalues have a nonzero imaginary part*,

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

If h is such that the eigenvalues are real,

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

PROOF. This follows from an explicit diagonalization of P_h in (4.2). 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

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

with $\omega = \exp\left(\frac{i\pi h}{n}\right)$, $p = 1 - c/n$, and $q = c/n$. 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

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

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

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

$$T(h,\, l) = \operatorname{Trace}(P_h^l P_h^{l*}) = \sum_i \sum_j r_{ij}\overline{r_{ij}} = \frac{2q^2}{|\lambda_+ - \lambda_-|^2}(|C|^2 + |\lambda_-{}^l - \lambda_+{}^l|^2)$$

$$= \frac{2}{|\beta - \alpha|^2}(|C|^2 + |\lambda_-{}^l - \lambda_+{}^l|^2).$$

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

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

from which

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

If the eigenvalues are real, then

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

and in this case,

$$T(h,\, l) = \frac{2}{|\beta - \alpha|^2}(2|\lambda_-{}^l - \lambda_+{}^l|^2 - (\lambda_+\lambda_-)^l(\alpha\bar{\beta} + \bar{\alpha}\beta - 2))$$

$$= 2(p-q)^l + \frac{4}{|\beta - \alpha|^2}(\lambda_-{}^l - \lambda_+{}^l)^2$$

$$= 2(p-q)^l + \frac{q^2}{q^2 - p^2 \sin^2(\pi h/n)}(\lambda_-{}^l - \lambda_+{}^l)^2.$$

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

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

with $\phi = \operatorname{Arg}(\lambda_-(h))$.

Bounding $2c^2 \sin^2(l\phi)$ 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 expansion is used for $1 < h \leq \varepsilon n$ for suitably small ε. For $\varepsilon n \leq h < n/2$, the denominator is bounded below by $\varepsilon^2 n^2(1+O(1/n))$. Finally, $\sin^2(\pi h/n) = \sin^2(\pi(n-h)/n)$. Combining bounds we have

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

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

and $O(1/n)$ depending on c.

For the lower bound, use the fact that the second term in square brackets is positive for all h so $T(h,l) \geq (1-2c/n)^l$. This completes the proof of Theorem 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 0. The χ^2 distance from stationarity at time l will be at least as big as the single term for the state $x = l \pmod{2n}$. The chance of being in this state will be at least $(1-1/n)^l$ (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 $1/2n$, the contribution to the χ^2 distance from this state will be at least $2n((1-1/n)^l - 1/2n)^2$. When n is large, $(1-1/n)^l \approx e^{-l/n}$. Using this, we can see that after $l = n$ 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 whose transition probabilities, $\widetilde{K}$, are given by

$$\widetilde{K}(x,x) = \tfrac{1}{2} + \tfrac{1}{2}K(x,x), \qquad \widetilde{K}(x,y) = \tfrac{1}{2}K(x,y) \quad \text{for } x \neq y.$$

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 $l/2$ with probability at least $(1-(1/2n))^l$. However, the probability of the chain being in state $l/2$ exactly is smaller than this by a factor of order $\sqrt{l}$. Consequently, the contribution to the χ^2 distance after n steps for 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 have shown in preliminary calculations, 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. Indeed if $c = c(n)$ then for $c(n) \leq \sqrt{\log n}$ order $n \log n/c$ steps appear to be necessary and sufficient to achieve uniformity. For $c(n) \geq \sqrt{\log n}$, order nc steps appear to be necessary and sufficient. Convergence in order n time is not attained because more frequent reversals of direction reintroduce a diffusive aspect into the chain's exploration of the state space.

In Theorem 2 we determined the rate of convergence carefully enough to find the cutoff in the χ^2 distance at $(n/2c)(\log n + \theta)$. Martin Hildebrand [20] 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.

5. Generalizations and relationships to other methods.

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

5.1. *Nonuniform distributions in one dimension.*

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

$$\widetilde{\mathscr{X}} = \{(z, x)\colon z \in \{-1, +1\}, x \in \mathscr{X}\}.$$

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 $\widetilde{\mathscr{X}}$. 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 $\widetilde{\mathscr{X}}$ 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. □

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 nonreversible walk of Section 2, which was analyzed in Section 3 and 4. The more general chain described here was abstracted from Horowitz [21], 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 [19], 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.

5.2. *General finite state spaces: the fiber algorithm.* Suppose that our state space, $\mathscr{X}$, 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 nonreversible 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 $\mathscr{X}$ we are given a collection of partitions $P_1, P_2, \dots, P_d$. That is, for each $i = 1, \dots, d$, there is a partition $P_i = \{P_{ij}\}_{j=1}^{J_i}$ for which $\bigcup_j P_{ij} = \mathscr{X}$ and $P_{ij} \cap P_{ij'} = \varnothing$ 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 $\mathscr{X}$ is connected in the sense that for each x, y in $\mathscr{X}$ there is a path $x_0 = x, x_1, \dots, x_l = y$ 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 $\mathscr{X}$, from which we wish to sample.

We now define a Markov chain $\widetilde{M}_d$ on an extended state space, $\widetilde{\mathscr{X}} = \{-1, +1\}^d \times \mathscr{X}$. This chain is parameterized by 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 $\{1, \dots, d\}$ 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 $x^* = x +_i z$, where $x +_i z$ is the successor of x in P_{ij} if $z_i = +1$, or the predecessor of x in P_{ij} if $z_i = -1$. If this successor or predecessor does not exist, reject the move. Otherwise, accept the move to $x^* = x +_i z$ 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 rejected, 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* $\tilde{\mathscr{X}}$.

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 $\mathscr{X}$. The last two steps above define a chain on $\{-1, +1\} \times \mathscr{X}$ 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. □

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 nonreversible 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, $\mathscr{X}$, has N points, and let $\pi^* = \max_x \pi(x)$. Liu [24] shows that

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

We consider two examples for which $\mathscr{X} = \{1, \dots, n\}^d$, 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 $\pi(x) = z\exp(-(x_1 + x_2 + \cdots + x_d))$. The normalizing constant, z, is bounded uniformly in n for fixed d and Liu's [24] 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 [13] 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 nonnegative 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 $\pi(x) = zp(x)$. For large $n, z \sim a_\alpha^* n^{|\alpha^*|+d}$. Thus $\pi^* \sim c/n^d$, for c bounded. Now, Liu's result shows that the chain M_u reaches stationarity in a bounded number of steps. The analysis in [13] 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 nonreversible methods.* The generalizations above extend the simple nonreversible 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 nonreversible 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 nonreversible methods that are presently used in quantum field theory [34, 23] and in some statistical applications [28, 18].

The one-dimensional walk of Section 5.1 is closely related to the "guided Monte Carlo" method of Horowitz [21]. The context is rather different, however. Horowitz's method applies to continuous state spaces (e.g., $\mathbb{R}^d$) 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 nonreversible, 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 nonreversible chain that reverses direction only when a rejection occurs. (In Horowitz's method, θ is fixed at zero; an effect similar to a nonzero 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 [15], 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, 29] is another way of constructing a nonreversible Markov chain, which can avoid diffusive behavior in many situations. As with the nonreversible walks discussed in this paper, the overrelaxation method uses transitions composed of steps that are individually reversible, but which produce a nonreversible chain when applied in sequence.

6. Examples of sampling using nonreversible chains.

This section shows how the methods of Sections 5.1 and 5.2 can be applied in three examples: a nonuniform 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 $\{1, 2, 3, \dots, n\}$, 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 (i.e., 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 [13] shows that order $n^2 \log n$ steps are necessary and sufficient for the usual Metropolis chain to reach stationarity. Preliminary work of Hildebrand [20] 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 (i.e., 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 (possibly 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.

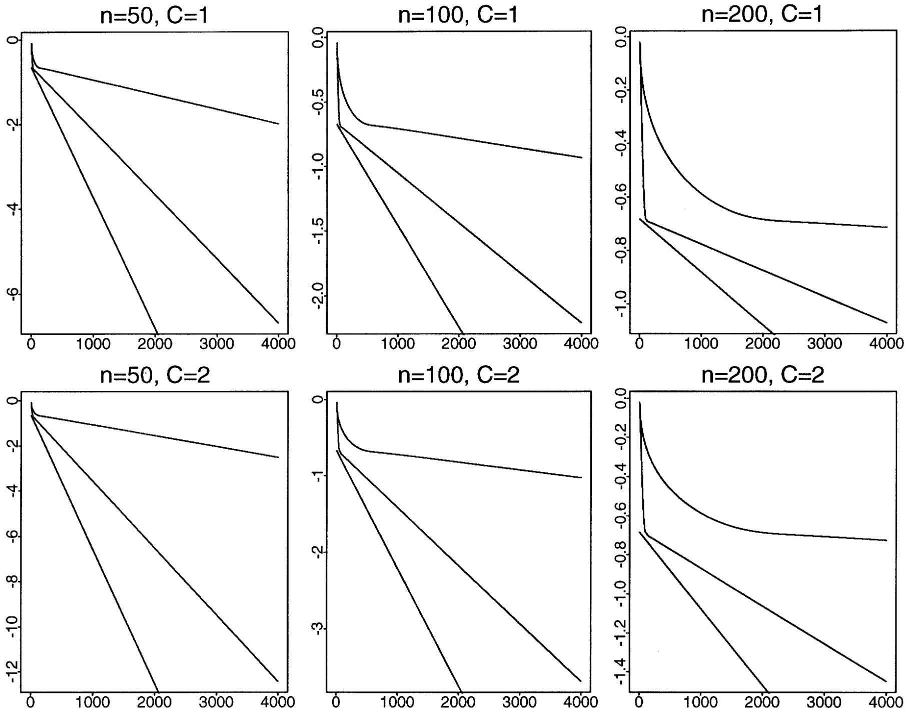


FIG. 2. *Convergence of Metropolis* (*top*), *directed* (*middle*), *and ideal* (*bottom*) *sampling methods on various V-shaped distributions* (*specified by n and C*). *The horizontal axis gives the number of transitions of a chain started in state* 1 [*for the directed method, state* $(+, 1)$]. *The 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.*)

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 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

	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

FIG. 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).*

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 nonnegative integer entries. This problem was posed by Diaconis and Efron [8] who give statistical motivation. Diaconis and Gangolli [10] 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 [14] 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 [6] for tables with large row and column sums and by Diaconis and Saloff-Coste [12] 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 [14] for statistical motivation.

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

The problem is to sample from the uniform distribution on $\mathscr{X}$. The random walk approach of [14] 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 $\mathscr{X}$, there is a positive integer, l, indices $i_1, i_2, \ldots, i_l$, and signs $z_1, z_2, \ldots, z_l$ in $\{\pm 1\}$ such that

$$x' = x + \sum_{j=1}^{l} z_j v_{i_j} \quad \text{and} \quad x + \sum_{j=1}^{a} z_j v_{i_j} \geq 0 \quad \text{for } 1 \leq a \leq l.$$

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

The Markov chain for sampling from $\mathscr{X}$ operates as follows: when in state x, choose one of the v_i at random, and choose z uniformly from $\{\pm 1\}$, then move to $x + zv_i$ provided this is in $\mathscr{X}$, and 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 $\{x + jv_i\}_{j \in \mathbb{Z}} \cap \mathscr{X}$ partition $\mathscr{X}$ as x varies. Varying i gives a collection of "directed" partitions, $P_1, P_2, \dots, P_k$, which satisfy the conditions of Proposition 2.

6.3. *Permutations.* Let $\mathscr{X} = S_n$ be the set of permutations on n letters, and let $d(\sigma, \eta)$ be a metric on S_n. To fix ideas, consider

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

A nonuniform probability distribution on $\mathscr{S}_n$ (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 $\sigma_0 = \text{id}$, so that the distribution $\pi(\sigma)$ is largest at $\sigma = \text{id}$ 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 [7].

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 S_n. Let H be a subgroup of S_n and P_H the partition of 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. $H = S_n$. 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 [4, 11]. 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. $H = \{\text{id}, (1, 2)\}$. 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 $\{\text{id}, (x, y)\}$. We see that with these choices the directed method 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 $(\sigma, \eta\sigma, \eta^2\sigma, \dots, n^{k-1}\sigma)$ where k is the order of η. For a practical version of the algorithm, choose a small collection of

permutations $\eta_1, \eta_2, \dots, \eta_K$ that generate S_n and use these to generate partitions $P_1, P_2, \dots, P_k$. 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 [7], nor when such random transpositions are used as a Metropolis proposal for sampling from a distribution over permutations of exponential form [13].

7. Scope and limitations of nonreversible sampling. We have shown in this paper that nonreversibility can be a desirable property of Markov chain sampling method. This conclusion accords with observations of the behavior of some practical nonreversible sampling methods [28, 19] and some previous theory (e.g., [23]).

The methods we discuss have some limitations, however. As illustrated in Section 6.1, any local algorithm, including the nonreversible 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 nonreversible 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 nonreversible 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 [14], is to consider the four cells in these row and columns as a 2×2 table and choose 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 [14].

Another limitation is that the $(\mathit{diameter})^2$ convergence time associated with reversible random walks applies to uniform or relatively flat stationary distributions. When the distribution is highly nonuniform, a nonreversible walk might have little or no advantage. For example, available theory [13] 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 behavior 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 nondiffusive methods such as Hybrid Monte Carlo [15], Horowitz's method [21] and overrelaxation [1] can sample much more efficiently than Gibbs sampling and simple Metropolis methods [29]. A simple nonreversible 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 [21]. The challenge is to find other such methods, especially for discrete state spaces where dynamical methods cannot be applied.

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

REFERENCES

[1] ADLER, S. L. (1981). Over-relaxation method for the Monte Carlo evaluation of the partition function for multiquadratic actions. *Phys. Rev. D* **23** 2901–2904.

[2] BASSIRI, F. (1997). Random walks on finite groups of multiplicity two. Ph.D. dissertation, Harvard Univ.

[3] BINDER, K. (1979). *Monte Carlo Methods in Statistical Physics.* Springer, Berlin.

[4] CONWAY, J., SLOANE, N. and WILKS, A. (1989). Gray codes for reflection groups. *Graphs Combin.* **5** 315–325.

[5] CHEN, F., LOVÁSZ, L. and PAK, I. (1999). Lifting Markov chains to speed up mixing. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing* 275–281. ACM Press, New York.

[6] CHUNG, F., GRAHAM, R. and YAU, S. T. (1996). On sampling with Markov chains. *Random Structures Algorithms* **9** 55–77.

[7] DIACONIS, P. (1988). *Group Representations in Probability and Statistics.* IMS, Hayward, CA.

[8] DIACONIS, P. and EFRON, B. (1987). Probabilistic-geometric theorems arising from the analysis of contingency tables. In *Contributions to the Theory and Application of Statistics: A Volume in Honor of Herbert Solomon* (A. E. Gelfand ed.) 103–125. Academic Press, Boston.

[9] DIACONIS, P. and FILL, J. (1990). Strong stationary times via a new form of duality. *Ann. Probab.* **18** 1483–1522.

[10] DIACONIS, P. and GANGOLLI, A. (1996). Rectangular arrays with fixed margins. In *Finite Markov Chain Renaissance* 15–42. Springer, New York.

[11] DIACONIS, P. and HOLMES, S. (1994). Gray codes for randomization procedures. *Statist. Comput.* **4** 207–302.

[12] DIACONIS, P. and SALOFF-COSTE, L. (1994). Moderate growth and random walk on finite groups. *Geom. Funct. Anal.* **4** 1–36.

[13] DIACONIS, P. and SALOFF-COSTE, L. (1998). What do we know about the Metropolis algorithm? *J. Comput. System Sci.* **57** 20–36.

[14] DIACONIS, P. and STURMFELS, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26** 363–397.

[15] DUANE, S., KENNEDY, A. D., PENDLETON, B. J. and ROWETH, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195** 216–222.

[16] GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

[17] GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intelligence* **6** 721–741.

[18] GUSTAFSON, P. (1997). Large hierarchical Bayesian analysis of multivariate survival data. *Biometrics* **53** 230–242.

[19] GUSTAFSON, P. (1998). A guided walk Metropolis algorithm. *Statist. Comput.* **8** 357–364.

[20] HILDEBRAND, M. (1997). Rates of convergence for a non-reversible Markov chain sampler. Preprint. Available from http://math.albany.edu:8000/~martinhi/.

[21] HOROWITZ, A. M. (1991). A generalized guided Monte Carlo algorithm. *Phys. Lett. B* **268** 247–252.

[22] LINDVALL, T. (1992). *Lectures on the Coupling Method.* Wiley, New York.

[23] KENNEDY, A. D. (1990). The theory of hybrid stochastic algorithms. In *Probabilistic Methods in Quantum Field Theory and Quantum Gravity* (P. H. Damgaard, H. Hüffel and A. Rosenblum, eds.). Plenum, New York.

[24] LIU, J. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statist. Comput.* **6** 113–119.

[25] METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.

[26] MIRA, A. and GEYER, C. J. (1999). Ordering Monte Carlo Markov chains. Technical Report 632, School of Statistics, Univ. Minnesota.

[27] NEAL, R. M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Dept. Computer Science, Univ. Toronto. Available from http://www.cs.utoronto.ca/~radford/.

[28] NEAL, R. M. (1996). *Bayesian Learning for Neural Networks. Lecture Notes in Statist.* **118** Springer, New York.

[29] NEAL, R. M. (1998). Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation. In *Learning in Graphical Models* (M. I. Jordan, ed.) 205–225. Kluwer, Dordrecht.

[30] ROBERTS, G. O. and SAHU, S. K. (1997). Updating schemes, correlation structure, blocking and parameterisation for the Gibbs sampler. *J. Roy. Statist. Soc. B* **59** 291–317.

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

[32] SMITH, A. F. M. and ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc.* **55** 3–23.

[33] TIERNEY, L. (1994). Markov Chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762.

[34] TOUSSAINT, D. (1989). Introduction to algorithms for Monte Carlo simulations and their application to QCD. *Comput. Phys. Comm.* **56** 69–92.

P. DIACONIS
DEPARTMENT OF STATISTICS
AND DEPARTMENT OF MATHEMATICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

S. HOLMES
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
AND
UNITÉ DE BIOMÉTRIE
INRA-MONTPELLIER
FRANCE
E-MAIL: susan@stat.stanford.edu

R. M. NEAL
DEPARTMENT OF STATISTICS
AND DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF TORONTO
10 KING'S COLLEGE ROAD
CANADA M5S 3G4
E-MAIL: radford@stat.utoronto.ca