

A non-reversible Markov chain Sampling Method

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 Neal
Statistics and Computer Science
University of Toronto
Canada
radford@cs.toronto.edu

Abstract

We introduce a nonreversible version of the Metropolis algorithm and show how it can improve on the diffusive behavior of the classical algorithm. The running time of some versions is analysed using both probabilistic techniques and an explicit diagonalisation. Numerical examples suggest real improvements over Metropolis in some cases.

Key words : Non-reversible Markov chain, Metropolis algorithm, cutoff phenomena.

Technical report: BU-1385-M

1 Introduction

Given a probability distribution $\pi(x) > 0$ on a finite set \mathcal{X} , the widely used Markov Chain approach [18], [7] draws samples from $\pi(x)$ by running a Markov Chain $K(x, y)$ on \mathcal{X} with stationary distribution $\pi(x)$. For many versions (Metropolis, Hasting, Gibb's sampler), the chain is constructed to be reversible with respect to π :

$$\pi(x)K(x, y) = \pi(y)K(y, x) \quad (1.1)$$

We give some non reversible constructions which can offer considerable speedups. The idea is best understood by a simple example:

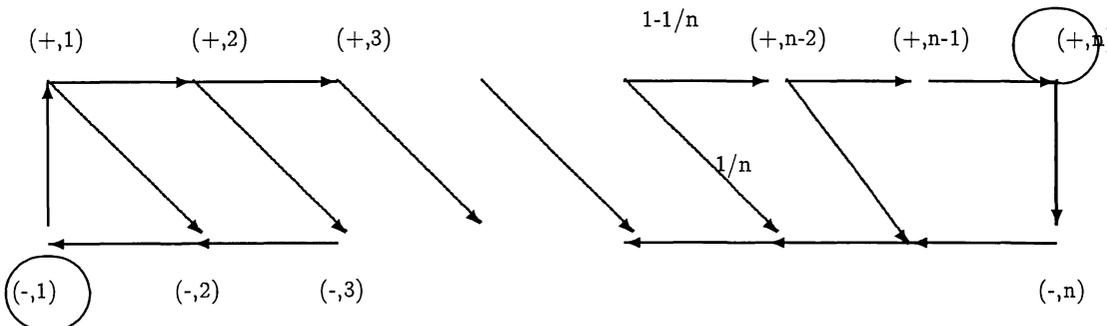
Consider nearest neighbor random walk on an n -point path with holding probabilities $\frac{1}{2}$ at each end.



This walk takes an order of n^2 steps to reach stationarity. This can be seen using the central limit theorem: the walk takes an order of k^2 steps to travel a distance of order k .

We overcome the "diffusive" behavior by introducing two copies of the chain, in one copy the chain will go right $1 - \frac{1}{n}$ of the time. In the second copy it will go left most of the time. The chain switches from copy to copy at rate $\frac{1}{n}$.

Start by labeling upstairs: $(+, 1), (+, 2), \dots, (+, n)$ and downstairs: $(-, 1), (-, 2), \dots, (-, n)$



The transition matrix is thus:

$$\begin{aligned} K((+, i), (+, i + 1)) &= (1 - \frac{1}{n}) & \text{for } 1 \leq i < n & K((-, 1), (-, 1)) &= \frac{1}{n} \\ K((+, i), (-, i + 1)) &= \frac{1}{n} & \text{for } 1 \leq i < n & K((-, 1), (+, 1)) &= 1 - \frac{1}{n} \\ K((-, i), (-, i - 1)) &= (1 - \frac{1}{n}) & \text{for } 1 < i \leq n & K((+, n), (-, n)) &= 1 - \frac{1}{n} \\ K((-, i), (+, i - 1)) &= \frac{1}{n} & \text{for } 1 < i \leq n & K((+, n), (+, n)) &= \frac{1}{n} \end{aligned} \quad (1.2)$$

The new chain has stationary distribution $\frac{1}{2n}$ on the new state space. Indeed the enter and exit weights of all points is 1, the matrix is doubly stochastic and thus it has a uniform stationary distribution. In section 2 we show that this chain reaches stationarity in an order of n steps.

Section 2 also derives the explicit diagonalization which is used to show that for ℓ^2 or χ^2 convergence, $\frac{1}{2}n \log n + cn$ steps are necessary and suffice. We also find the best "flip" rates in this simple example.

In section 3 we show how to generalize this example in two ways:

1. The n point path can be replaced by a connected part of a d dimensional grid.
2. The uniform distribution can be replaced by any distribution $\pi(x) > 0$ on \mathcal{X} .

As examples, we treat a variety of statistical problems (contingency tables and logistic regression) where non-reversible speedups are feasible. We also show how to construct non-reversible versions of the Gibbs sampler or Heat-bath algorithm. Section 4 contains some cautionary notes and a comparison with random choice Monte Carlo.

There have been numerous efforts to get rid of diffusive behavior. The present idea was motivated by hybrid Monte Carlo [16],[19],[14] which introduces auxiliary velocity variables and (in continuous settings) Hamiltonian dynamics to give directions to move in. Our technique is based on an idea of Horowitz[16] developed in this setting. Other techniques for overcoming diffusive behavior are overrelaxation [19], and the multi-grid techniques of Goodman and Sokal [15].

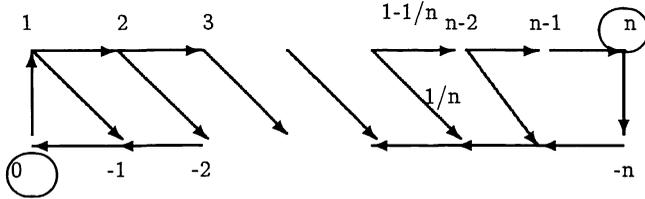
Further, the analysis of section 2 offers one of the few cases known of a natural Markov chain where the total variation and chi-squared relaxation times are different.

Acknowledgments :

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

2 Analysis of the One-dimensional walk

The state space of the walk can be labeled with elements of the circle $\mathbb{Z}_{2n} \pmod{2n}$.



The walk (1.2) is described equivalently by a Markov Chain on \mathbb{Z}_{2n} with kernel:

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

The mass coming into any vertex is the same as the mass coming out. So the chain is doubly stochastic with stationary distribution:

$$\pi(j) = \frac{1}{2n}, \quad 0 \leq j < 2n. \quad (2.4)$$

The chain (2.3) will be analyzed by two methods. First a direct probabilistic argument is combined with sub-multiplicativity to show that order n steps are necessary and sufficient for total variation convergence.

Then section 2.2 gives the eigenvalues and eigenvectors of the chain and uses them to show that $\frac{1}{2}n \log n + cn$ steps are necessary and sufficient for convergence in the χ^2 distance. This is somewhat surprising ; usually the two distances are equivalent in natural problems. In section 2.3 we find the best flip rates (roughly $\sqrt{\log n}/n$)

2.1 A Probability Argument

The first result shows that order n steps are necessary and sufficient for total variation convergence.

Theorem 2. 1 The chain (2.3) on \mathbb{Z}_{2n} satisfies:

$$\|K^\ell - \pi\|_{TV} \leq (1 - C)^{\lfloor \frac{\ell}{4n} \rfloor} \quad C = 8e^{-8} > .002$$

For $n > 2$, and all $\ell = 1, 2, \dots$ and any starting state.

Conversely: The chain started at 0 is not close to random in only n steps:

$$\|K_0^\ell - \pi\|_{TV} > \frac{1}{e^2} \text{ for } \ell \leq n$$

Proof:

After n steps, the walk started at 0 is at n with probability $(1 - \frac{1}{n})^n$. Hence

$$\|K_0^n - \pi\|_{TV} > (1 - \frac{1}{n})^n - \frac{1}{2n} > e^{-(1+\frac{1}{n-1})} > \frac{1}{e^2}$$

For the other direction, let X_m be the position of the walk at time m . We will show that for any starting state α and any state j , when $m = 4n$,

$$P_\alpha\{X_m = j\} \geq \frac{C}{2n} \quad \text{with } C = 8e^{-8} \quad (2.5)$$

The majorization (2.5) suffices to prove the theorem by an easy argument:

Let $\mathbf{K}(x, y)$ be a Markov chain on a finite state space \mathcal{X} . Suppose π is a stationary probability for \mathbf{K} and there are m, C such that $\mathbf{K}^m(x, y) \geq C\pi(y)$, for all x, y . Then $\|\mathbf{K}_x^\ell - \pi\| \leq (1 - C)^{\lfloor \frac{\ell}{m} \rfloor}$, for all ℓ .

To see this, suppose without loss that $m = 1$, then $\mathbf{K}(x, y) = C\pi(y) + \frac{(1-C)\mathbf{K}(x,y) - C\pi(y)}{1-C}$, this presents the chain as a mixture of π and a second chain. If T is the first time that the mixture is chosen, then at time T , the process is stationary and indeed T is a strong stationary time in the sense of [11], and this reference gives results that provide a bound on the total variation. For general m , we apply the above to \mathbf{K}^m .

To prove 2.5, let T_1, T_2, \dots be the times that the walks changes sign, i.e. when $x \rightarrow -x$ is chosen. Thus $1 \leq T_1 < T_2 < T_3 < \dots$. Let A_i be the sequences of length m with exactly i sign changes. Clearly

$$P_\alpha\{X_m = j\} \geq P_\alpha\{\{X_m = j\} \cap A_1(m)\} + P_\alpha\{\{X_m = j\} \cap A_2(m)\}$$

The reason for using both A_1 and A_2 is because of a parity problem. From direct considerations, starting at α , with any m ,

$$\text{on } A_1 : X_m = m - \alpha + 1 - 2T_1 \pmod{2n} \quad (2.6)$$

$$\text{on } A_2 : X_m = 2(T_1 - T_2) + \alpha + m \pmod{2n} \quad (2.7)$$

These equations show the parity problem: on A_2 , the walk is at an even number of steps past its starting position α after an even number of steps.

By sufficiency,

$$P_\alpha\{T_1 = i | A_1(m)\} = \frac{1}{m}, 1 \leq i \leq m$$

and

$$P_\alpha\{T_1 = i, T_2 = j | A_2(m)\} = \frac{1}{\binom{m}{2}}, 1 \leq i \leq j \leq m$$

Further

$$P_\alpha\{A_1(m)\} = m \frac{1}{n} \left(1 - \frac{1}{n}\right)^{m-1}$$

$$P_\alpha\{A_2(m)\} = \binom{m}{2} \frac{1}{n^2} \left(1 - \frac{1}{n}\right)^{m-2}$$

Take $m = 4n$ and consider $j + (m - \alpha + 1)$ even. By directly counting solutions

$$P_\alpha\{X_m = j \cap A_1(m)\} = P_\alpha\{m - \alpha + 1 - 2T_1 = j \pmod{2n} \cap A_1(m)\}$$

$$\geq \frac{1}{m} m \frac{1}{n} \left(1 - \frac{1}{n}\right)^{m-1} \geq \frac{8e^{-8}}{2n}$$

The last inequality used $(1 - t) > e^{-t/(1-t)}$ for $0 < t < 1$ and $n \geq 2$.

For $j + (m - \alpha + 1)$ odd :

$$P_\alpha\{X_m = j \cap A_2(m)\} = P_\alpha\{m + \alpha + 2(T_1 - T_2) = j \pmod{2n} \cap A_2(m)\}$$

$$\geq \frac{m}{\binom{m}{2}} \binom{m}{2} \frac{1}{n^2} \left(1 - \frac{1}{n}\right)^{m-2} \geq \frac{8e^{-8}}{2n}$$

The first inequality comes from considering the distribution of $2(T_1 - T_2) \pmod{2n}$ given $A_2(m)$. By direct count, for any ℓ ,

$$|\{(i, k) : k - i = \ell \pmod{n}\}| \geq (m - \ell) + (n - \ell)$$

The $m - \ell$ term comes from solutions $(1, \ell + 1), (2, \ell + 2), \dots, (m - \ell, m)$. The $n - \ell$ term comes from solutions $(1, n + \ell + 1), \dots, (n - \ell, m)$. Thus the number of solutions is bounded below by $m = 4n$, uniformly in ℓ . This proves 2.5 and so completes the proof. \diamond

Remark: The bound (2.5) implies

$$\max_{a,b} \left(1 - \frac{\mathbf{K}_a^\ell(b)}{\pi(b)}\right) \leq (1 - C)^{\lfloor \frac{\ell}{4n} \rfloor}$$

In section 2.2 below we show $\max_{a,b} |1 - \frac{\mathbf{K}_a^\ell(b)}{\pi(b)}|$ is only small after $\frac{1}{2}n \log n$ steps.

2.2 ℓ^2 bounds

In this and the following section the ℓ^2 or χ^2 rate of convergence is determined.

Here the ℓ^2 distance can be written:

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

For these equivalences, see [6].

This ℓ^2 distance bounds total variation through

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

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 $\chi^2(\ell)$ convergence.

The walk (2.3) changes direction at rate $\frac{1}{n}$. It is natural to ask how the change rate effects the speed of convergence. For example if the change rate is $\frac{1}{2}$, it is not hard to see that order n^2 steps are necessary and suffice for either total variation or ℓ^2 convergence.

In this section we analyze a one-parameter family of chains on \mathbb{Z}_{2n} :

$$\mathbf{K}(x, y) = \begin{cases} \frac{c}{n} & \text{if } y = -x \\ \frac{1}{n} - \frac{c}{n} & \text{if } y = x + 1 \\ 0 & \text{otherwise} \end{cases} \quad (2.8)$$

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

Theorem 2.2 Consider the chain (2.8) on \mathbb{Z}_{2n} , for fixed $c \in (0, \pi)$, for $\ell = \frac{n}{2c}(\log n + \theta)$ with θ a fixed real, then:

$$\mathbf{A}e^{-\theta} \leq \chi^2(\ell) \leq \mathbf{B}e^{-\theta}$$

with \mathbf{A} and \mathbf{B} are bounded continuous functions of c .

Remark :

In Lemma 2.2 below we show that for the chain (2.8) $\chi^2(\ell)$ does not depend on the starting state. Observe that increasing the change rate $\frac{c}{n}$ decreases the time to stationarity for c in $(0, \pi)$. In section 2.3 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 ℓ^2 convergence.

Theorem 2 will be proved as a sequence of lemmas which are also used in section 2.3, the first step is an explicit diagonalization of the underlying transition matrix:

Lemma 2.1 For any c , the chain K as defined in (2.8) 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:

$$P_h = \begin{pmatrix} (1 - \frac{c}{n})e^{\frac{i\pi h}{n}} & \frac{c}{n} \\ \frac{c}{n} & e^{-\frac{i\pi h}{n}}(1 - \frac{c}{n}) \end{pmatrix}, \quad 1 \leq h \leq n-1 \quad (2.9)$$

Proof.

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

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

The matrix form (2.8) is with respect to the standard basis $\delta_j(k)$ of L .

Consider instead the Fourier basis $f_h(j)$, $0 \leq h < 2n$:

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

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 (2.9) 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 $Kf_n(j) = (-1)^j \times -(1 - \frac{2c}{n})$, proving the lemma. ◇

Lemma 2.1 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 2.2 does not hold for the walk on \mathbb{Z}_{2n+1}

Lemma 2.2 The matrix K of (2.8) satisfies:

$$\{K^\ell(x, y)\}_{0 \leq y < 2n} = \{K^\ell(x', y)\}_{0 \leq y < 2n}$$

This holds for all c, x, x' and all positive ℓ .

Proof:

Let \mathcal{C} be the basic circulant of size $2n$. Thus \mathcal{C} is a $2n \times 2n$ matrix with ones above the diagonal and a one in the lower left corner, zeroes elsewhere. ($\mathcal{C}(i, j) = \delta_{i+1}(j)$). Let \mathcal{P} be the basic Hankel matrix, that is to say, \mathcal{P} has ones down the antidiagonal: (so $\mathcal{P}(i, j) = \delta_i(2n+1-j)$).

Observe that $K = a\mathcal{C} + b\mathcal{P}\mathcal{C}$ for some a, b .

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

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

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

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

This shows that $\mathbf{K}^\ell = \mathcal{C}_1 + \mathcal{P}\mathcal{C}_2$ 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 $\mathcal{P}\mathcal{C}_2$ 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

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

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

The next lemma gives the basic computational expression needed.

Lemma 2.3 For any \mathbf{c} and any starting \mathbf{x} , the chain (2.8) satisfies:

$$\chi^2(\ell) = \text{Trace}(\mathbf{K}^\ell \mathbf{K}^{\ell*}) - 1 = \left(1 - \frac{2\mathbf{c}}{n}\right)^{2\ell} + \sum_{1 \leq \mathbf{h} < n} \mathbf{T}(\mathbf{h}, \ell) \quad (2.10)$$

where $\mathbf{T}(\mathbf{h}, \ell) = \text{Trace}(\mathbf{P}_\mathbf{h}^\ell \mathbf{P}_\mathbf{h}^{\ell*})$ and $\mathbf{P}_\mathbf{h}$ is defined by (2.9).

Proof:

From Lemma 2.2, the entries of any row of \mathbf{K}^ℓ are a permutation of the first row.

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

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

The result now follows from Lemma 2.1. ◇

The following lemma is the heart of the argument, it gives an explicit diagonalization of the 2×2 blocks $\mathbf{P}_\mathbf{h}$. The alternative expressions given for the eigenvalues are needed in section 2.3.

Lemma 2.4 For any \mathbf{c} , $\mathbf{T}(\mathbf{h}, \ell)$ in (2.10) of Lemma 2.3 is given by:

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

with λ_+ and λ_- the eigenvalues of the matrices $\mathbf{P}_\mathbf{h}$, (we have omitted the \mathbf{h} in their symbols to ease the notation).

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

Actually, if \mathbf{h} is such that the eigenvalues have an imaginary part:

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

If h is such that the eigenvalues are real:

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

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:

$$\begin{aligned} \alpha &= \frac{\lambda_- - p\omega}{q} = \frac{q}{\lambda_- - p\bar{\omega}} \\ \beta &= \frac{\lambda_+ - p\omega}{q} = \frac{q}{\lambda_+ - p\bar{\omega}} \\ \omega &= e^{\frac{i\pi h}{n}} \quad p = 1 - \frac{c}{n} \quad \text{and} \quad q = \frac{c}{n} \end{aligned}$$

Further we have the identities:

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

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

Calling $C = \beta\lambda_-^\ell - \lambda_+^\ell\alpha$, and $D = -\alpha\lambda_-^\ell + \lambda_+^\ell\beta$, we always have $|C|^2 = |D|^2$. We also always (real and complex cases alike) have $|\alpha\beta|^2 = 1$.

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

$$\begin{aligned} T(\mathbf{h}, \ell) &= \text{Tr} P_h^\ell P_h^{\ell*} = \sum_i \sum_j r_{ij} \bar{r}_{ij} \\ &= \frac{2q^2}{|\lambda_- - \lambda_+|^2} (|C|^2 + |\lambda_- - \lambda_+|^2) \\ &= \frac{2}{|\beta - \alpha|^2} (|C|^2 + |\lambda_- - \lambda_+|^2) \end{aligned}$$

Separating the two cases, using $\lambda_+\lambda_- = (1 - \frac{2c}{n}) = (p - q)$ and denoting by $\rho_h \stackrel{\text{def}}{=} \frac{\pi h}{n}$:

Eigenvalues have an imaginary part:

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

gives:

$$\begin{aligned} T(\mathbf{h}, \ell) &= 2|\lambda_+|^{2\ell} \left(1 + \frac{2q^2 \sin^2 \ell \phi}{|\lambda_+|^{2\ell} \sin^2 \phi}\right) \\ &= 2(p - q)^\ell \left(1 + \frac{2q^2 \sin^2 \ell \phi}{p^2 \sin^2 \rho_h - q^2}\right) \end{aligned}$$

Real Eigenvalues:

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

So that, in this case:

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

This completes the proof of Lemma 2.4. ◇

Proof of Theorem 2:

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

Now, lemma 4 gives

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

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

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

This expansions is used for $1 < \mathbf{h} \leq \epsilon \mathbf{n}$ for suitably small ϵ .

For $\epsilon \mathbf{n} \leq \mathbf{h} < \frac{\mathbf{n}}{2}$, the denominator is bounded below by $\epsilon^2 \mathbf{n}^2 (1 + \mathcal{O}(\frac{1}{\mathbf{n}}))$.

Finally, $\sin^2(\frac{\pi\mathbf{h}}{\mathbf{n}}) = \sin^2(\frac{\pi(\mathbf{n}-\mathbf{h})}{\mathbf{n}})$. Combining bounds we have:

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

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

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

2.3 ℓ^2 bounds with large flip rates

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

Theorem 2.3 For the chain (2.8) with $\mathbf{c} = \mathbf{c}(\mathbf{n})$

a) Suppose $c(n) \leq a\sqrt{\log n}$ for fixed a . Then for any starting state x and for

$$\ell = \frac{An \log n}{c}$$

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

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

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

$$\ell = An\sqrt{\log n}$$

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

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 2.10 we have, for any starting state x , any c and ℓ :

$$\chi^2(\ell) = \text{Trace}(\mathbf{K}^\ell \mathbf{K}^{\ell*}) - 1 = \left(1 - \frac{2c}{n}\right)^{2\ell} + \sum_{1 \leq h < n} \mathsf{T}(h, \ell) \quad (2.11)$$

with for λ_\pm with an imaginary part:

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

$$\mathsf{T}(h, \ell) = 2\left(1 - \frac{2c}{n}\right)^\ell + \left(1 - \left(1 - \frac{n}{c}\right)^2 \sin^2\left(\frac{\pi h}{n}\right)\right)^{-1} (\lambda_+^{2\ell} + \lambda_-^{2\ell} - 5\left(1 - \frac{2c}{n}\right)^\ell) \text{ for } \lambda_\pm \text{ real}$$

and as before

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

For Ω defined as follows. Let $h^* = h^*(c, n)$ be the smallest h so that the eigenvalues are imaginary, we have:

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

Let us first treat the case with

$$c \leq \Delta \log n, \text{ where } \Delta \text{ is a fixed constant} \quad (2.12)$$

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:

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

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 $c \leq \Delta \log n$ provides:

$$\begin{aligned} \lambda_-(h) &= 1 - \frac{\pi^2 h^2}{2nc} + O\left(\frac{h^2}{n^2}\right) \\ \lambda_+(h) &= 1 - \frac{2c}{n} + \frac{\pi^2 h^2}{2nc} + O\left(\frac{h^2}{n^2}\right) \end{aligned}$$

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

$$\sum_{h \leq h^*} T(h, \ell) \leq D \left\{ 2\left(1 - \frac{2c}{n}\right)^\ell + \sum_{h \leq \frac{c}{\pi}} \left(1 - \frac{\pi^2 h^2}{2nc} + O\left(\frac{h^2}{n^2}\right)\right)^{2\ell} + \left(1 - \frac{2c}{n} + \frac{\pi^2 h^2}{2nc} + O\left(\frac{h^2}{n^2}\right)\right)^{2\ell} \right\} \quad (2.13)$$

Essentially the same bounds hold for the large elements in Zone 0:

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

Thus the right hand side of 2.13 (with a different D), bounds the sum over zone 0.

For Zone 1 we use the techniques of section 2.2 to get the bound

$$\sum_{\text{Zone 1}} T(h, \ell) \leq 2n\left(1 - \frac{2c}{n}\right)^\ell \left(1 + \sum_{h > h^*} \frac{4c^2}{\pi^2 h^2 - c^2}\right) \quad (2.14)$$

All claims follow from the claims (2.13) and (2.14);

Consider first $c \leq A'\sqrt{\log n}$ and take $\ell = A\frac{n \log n}{c}$. This choice makes the bound 2.14 small for A large; certainly the first and last terms in 2.13 are small also:

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

These bounds give part (a).

For part (b), the sum (2.14) is bounded above by a constant times $n e^{-\frac{2c\ell}{n}}$. Here, $\ell = A n \sqrt{\log n}$ and $c \geq a \sqrt{\log n}$, so this term is small for A large. For the sum in (2.13), 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. \diamond

Remarks:

In theorem 2 we determined the rate of convergence carefully enough to find the cutoff in the ℓ^2 distance at $\frac{n}{2c}(\log n + \theta)$. We do not know whether similar cutoffs hold in total variation distance. Martin Hildebrand (personal communication) 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 2.1.

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

3 Extensions and Examples

This section gives non reversible versions of the Metropolis algorithm for general stationary distributions and general state spaces. We first treat a general stationary distribution on a one-dimensional path and then show how to use this for a more general state space which can be decomposed into "lines" of an appropriate sort.

3.1 Stationary Distribution on a path

Let $\pi(x)$ be given on $\mathcal{X} = \{1, 2, \dots, n\}$. As in section 1 extend the state space to

$$\tilde{\mathcal{X}} = \{(\epsilon, x), \epsilon = \pm 1, x \in \mathcal{X}\}$$

Let $\tilde{\pi} = \frac{\pi(x)}{2}$, construct \tilde{M} on $\tilde{\mathcal{X}}$ in two stages. The second stage depends on a parameter θ which can be any fixed value in $(0,1)$.

Algorithm:

1. From (ϵ, x) , try to move to $(-\epsilon, x + \epsilon)$ via a standard Metropolis step. Thus

$$\left\{ \begin{array}{l} \text{if } \pi(x + \epsilon) \geq \pi(x) \text{ go there,} \\ \text{otherwise flip a } \frac{\pi(i+\epsilon)}{\pi(i)} \text{ coin,} \end{array} \right\} \left\{ \begin{array}{l} \text{if it comes up heads- go there} \\ \text{otherwise stay put} \end{array} \right.$$

If $x + \epsilon$ is outside \mathcal{X} the chain stays at (ϵ, x) .

2. Following the first step the chain is in (ϵ', x') . With probability $1 - \theta$ the chain moves to $(-\epsilon', x')$. With probability θ the chain stays at (ϵ', x') .

Proposition 3.1 The chain \tilde{M} described above is an irreducible aperiodic chain on $\tilde{\mathcal{X}}$ with stationary distribution $\tilde{\pi}(\epsilon, x) = \frac{\pi(x)}{2}$

Proof:

Both stages above are Markov chains with $\tilde{\pi}$ as their stationary distribution. The first stage as the usual construction of the reversible Metropolis algorithm. The second stage by invariance: $\tilde{\pi}(\epsilon, x) = \tilde{\pi}(-\epsilon, x)$, Since $0 < \pi(x) < 1$ (provided $n \geq 2$) the chain \tilde{M} is connected. Since $\tilde{M}((+1, n), (+1, n)) = \theta$, the chain is aperiodic. This completes the proof. ◇

Remarks:

1. The idea behind this example was abstracted from Horowitz [16]. It can be applied to general state spaces. For example on \mathbb{R} , to sample from $\pi(\mathbf{dx})$ introduce two copies of \mathbb{R} . Run the Metropolis algorithm based on 2 base chains; one with a drift to the right, one with a drift to the left.
2. The same idea can be used to make directed versions of other reversible chains. For example, suppose that each $i \in \{1, 2, \dots, n\}$ is associated with a neighborhood $\mathbf{N}(i)$. The usual heat bath (Gibb's sampler) method samples from π restricted to $\mathbf{N}(i)$. Instead of using symmetric neighborhoods eg $\mathbf{N}(i) = \{i - 1, i + 1\}$, one could use an asymmetric neighborhood eg $\mathbf{N}(i) = \{i + 1, i + 2\}$ for the one chain and $\mathbf{N}(i) = \{i - 1, i - 2\}$ for the other chain constructed as above.
3. There is nothing special about working with two copies of \mathcal{X} , the fiber algorithm of section 3.2 can be seen as working with 2^d copies of a base space.

3.2 General Finite State Spaces: the fiber algorithm

Let π be a positive probability measure on the finite state space \mathcal{X} . In what follows, we will assume that \mathcal{X} is partitioned into ordered "lines" in various directions. Our walk will proceed from x by choosing a direction and then a step along the line through x . The reader may find it helpful to consider an $m \times n$ grid with horizontal lines of size n and vertical lines of size m . As in our first example, we introduce two copies of each line and run a non reversible Markov chain.

A class of examples where this structure arises naturally is described in section 3.3.

Suppose \mathcal{X} is given with a collection of partitions P_1, P_2, \dots, P_I . Here, for each i , there is a partition

$$P_i = \{P_{ij}\}_{j=1, \dots, J_i} \text{ having } \cup_j P_{ij} = \mathcal{X}, P_{ij} \cap P_{ij'} = \emptyset, j \neq j'$$

The index i delineates a direction. The parts P_{ij} are called lines in direction i . We suppose that each line P_{ij} is linearly ordered.

Further, we suppose that \mathcal{X} is connected in the sense that for each x, y in \mathcal{X} there is $x_0 = x, x_1, \dots, x_\ell = y$ such that x_i, x_{i+1} are in a common line.

To complete the description, let $\{\omega_i\}_{i=1}^I$ be a positive probability distribution on $\{1, 2, \dots, I\}$ and $\{\theta_i\}_{i=1}^I$ satisfy $0 < \theta_i < 1$. The $\{\omega_i\}$ are used to choose directions, the $\{\theta_i\}$ are flip rates in direction i .

With these ingredients specified, a chain \tilde{M} can be defined on $\tilde{\mathcal{X}} = \mathbb{Z}_2^I \times \mathcal{X}$. Suppose the chain is currently at (z, x) . The construction proceeds in 3 stages:

- Choose $i \in \{1, \dots, I\}$ with probability ω_i .
- Given i , suppose x in P_{ij} . If $z_i = \epsilon$ try to move to $-z_i$ and the successor x^ϵ of x in P_{ij} .
If $\pi(x^\epsilon) \geq \pi(x)$ accept the move.
If $\pi(x^\epsilon) < \pi(x)$ flip a coin with success rate $\frac{\pi(x^\epsilon)}{\pi(x)}$, if success, accept the move, if not the chain stays at (z, x) . At this stage the chain is at (z', x') .
- Change the i th coordinate z' of z back to $-z_i'$ with probability $1 - \theta_i$ and keep it unchanged with probability θ_i .

Proposition 3.2 For a connected set of partitions into linearly ordered lines the chain \tilde{M} is aperiodic, connected with stationary distribution $\tilde{\pi} = \frac{\pi(x)}{2^I}$ on $\tilde{\mathcal{X}}$.

Proof:

The chain is a mixture of I chains, each of which will be shown to have the claimed stationary distribution. Suppose $\{S_j\}_{j=1}^J$ is a partition of \mathcal{X} . The last two steps above define a chain on $\mathbb{Z}_2 \times \mathcal{X}$ driven by

$\{S_j\}_{j=1}^J$. This chain is not connected (if $J > 1$). But proposition 1 above shows that on each component S_j the chain has stationary distribution for any partition and flip rate θ .

The general stationarity result 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. ◇

Remarks:

1. Again, it is easy to generalize the construction to Euclidean and more general spaces. Consider a probability density $f(x)$ on \mathbb{R}^d take P_i the partition of \mathbb{R}^d into lines parallel to the i th coordinate axis. For each i consider two random walks with opposite drifts as base chains for the Metropolis chain in this coordinate.
2. The construction above points to a potential drawback: we must find directions and ordered lines. This is easy to do for naturally given grids. Less obvious examples are given in section 3.4. The following example shows some of the problems and possibilities.

3.3 Three Examples

This section shows how the ideas above specialize in three examples: a non uniform distribution on a path, versions of Fishers exact test for contingency tables and ranked data.

3.3.1 A non-uniform distribution on a path

Consider sampling from a distribution π on $\{1, 2, 3 \dots, n\}$. We take π to be V-shaped as in Figure 1.

$$\pi(j) = z(2|j - \frac{n}{2}| + 1), 1 \leq j \leq n, z \text{ a normalizing constant}$$

This π has 2 “peaks” and one might expect the usual Metropolis algorithm with base chain nearest neighbor random walk to get stuck and be unable to cross from peak to peak. While this is true for exponential peaks, things are better for polynomial peaks. For the linear peaks in Figure 1, available theory [7] shows that $n^2 \log n$ steps are necessary and sufficient for the usual Metropolis chain to reach stationarity.

Figure 2 shows how the worst case total variation distance decreases when $n = 100$ as a function of ℓ , the number of steps taken. It shows both the ordinary Metropolis and our directed version. Evidently the directed version converges faster.

Indeed by 5000 steps the ordinary Metropolis chain does not have a good chance of crossing over the valley, ($n^2 \log n \doteq 46,000$ for $n = 100$). We believe the nonreversible version of the Metropolis algorithm of section 3.1 reaches stationarity in order n^2 steps. The computations shown below are **not** based on Monte Carlo runs but rather on exact treatment achieved by raising the transition matrix to successive powers and then calculating the total variation distance to stationarity for each starting state.

3.3.2 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 [10] who give statistical motivation. Diaconis and Gangolli [12] give a host of other applications. Even for I, J small, the size of the

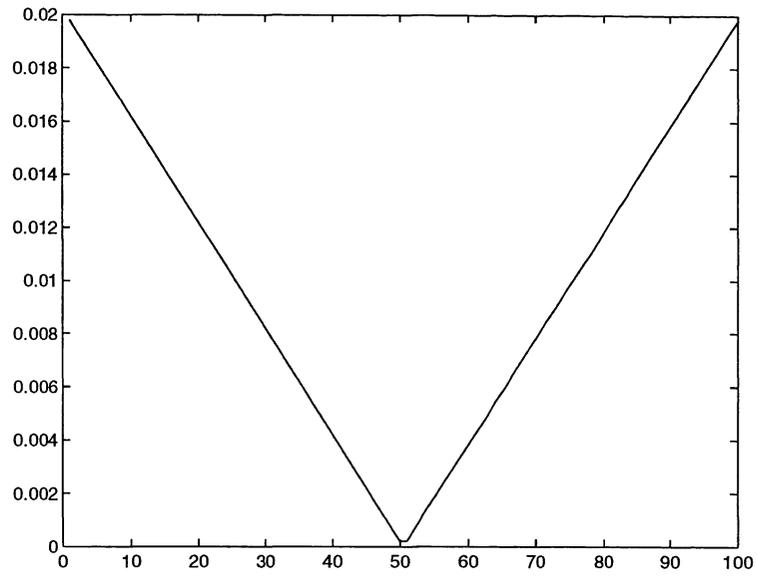


Figure 1: Linear Valley Stationary Distribution on $n=100$ points

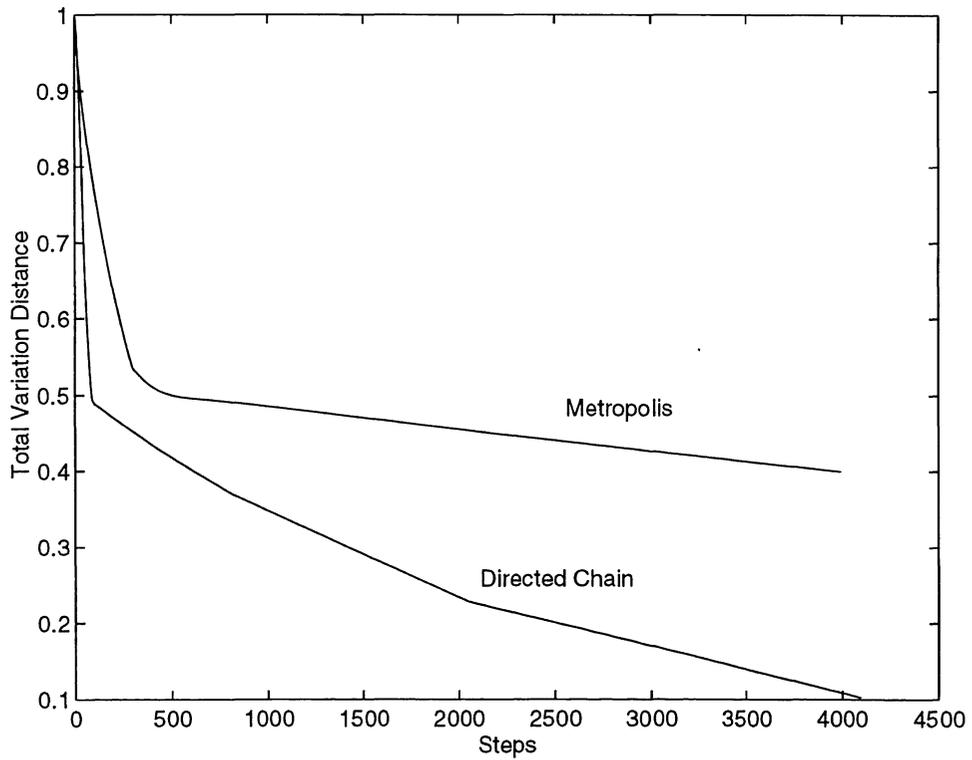


Figure 2: Comparisons of the two algorithms

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 [8] suggested the following algorithm for generating random tables:

- Choose a pair of different rows at random
- Choose a pair of different columns at random
- As long as it doesn't make any table value negative make the following change to the 2 by 2 square thus defined : $\begin{pmatrix} + & - \\ - & + \end{pmatrix}$ or $\begin{pmatrix} - & + \\ + & - \end{pmatrix}$ choosing one of the above with probability $\frac{1}{2}$

This is a symmetric, connected aperiodic Markov chain 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-Yau [3] for tables with large row and column sums and by Diaconis and Saloff-Coste [6] for small values of I and J.

We can apply the ideas of section 3.2 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.

Diaconis and Sturmfels have extended the reversible walk described above for tables to a host of other statistical problems. We give a general description referring to [8] for statistical motivation.

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

The problem is to choose from the uniform distribution on \mathcal{X} . In [8] a random walk approach is suggested.

A Markov Basis is a set of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k \in \mathbb{Z}^n$, which satisfy:

1. $\mathbf{A}\mathbf{v}_i = 0$
2. For $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ there are a positive integer ℓ and indices i_1, i_2, \dots, i_ℓ and signs $\epsilon_1, \epsilon_2, \dots, \epsilon_\ell$ in $\{\pm 1\}$ such that:

$$\mathbf{x}' = \mathbf{x} + \sum_{j=1}^{\ell} \epsilon_j \mathbf{v}_{i_j} \text{ and } \mathbf{x} + \sum_{j=1}^{\ell} \epsilon_j \mathbf{v}_{i_j} \geq 0 \text{ for } 1 \leq a \leq \ell$$

Condition (1) says that $\mathbf{A}(\mathbf{x} + \mathbf{v}_i) = \mathbf{y}$, condition (2) says that there is a path between each \mathbf{x} and \mathbf{x}' in \mathcal{X} , adding or subtracting \mathbf{v}_i , staying in \mathcal{X} .

This allows a Markov Chain approach to sampling from \mathcal{X} . From \mathbf{x} , choose \mathbf{v}_i at random and move to $\mathbf{x} + \mathbf{v}_i$, if this is in \mathcal{X} , otherwise stay at \mathbf{x} . This chain reduces to the chain described above for tables with an appropriate choice of \mathbf{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 3.2. Observe that the lines $\{\mathbf{x} + j\mathbf{v}_i\}_{j \in \mathbb{Z}} \cap \mathcal{X}$ partition \mathcal{X} as \mathbf{x} varies. Varying i gives a collection of partitions \mathcal{P}_i . Then $\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_k$ are 'directed' partitions satisfying the conditions of Proposition 1.

We have run these speed-ups for the table above and have found them to work much faster than the original ± 1 walk.

3.3.3 Ranked Data

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

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

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

$$\pi(\sigma) = z\theta^{d(\sigma, \sigma_0)}, \text{ } z \text{ a normalizing constant}$$

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 when for instance $n = 52$.

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 [4].

To pursue the present approach we must find a collection of ordered partitions. One natural construction uses the group structure of \mathcal{S}_n . Let H be a subgroup and P_H the partition of \mathcal{S}_n into cosets of H . Taking all conjugates, $H^\pi = \pi^{-1}H\pi$ gives a neat family of partitions, we consider 3 special cases.

Case 1 Take $H = \mathcal{S}_n$, 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 (see Sloane[2] or [13]). This linearizes the problem so that the method of section 3.1 can be used. This is not a foolish approach: if the walk is started off at the identity it should be reasonably efficient.

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

Case 3 Take H as the cyclic group generated by a single permutation η .

Now the block of the partition containing σ is $(\sigma, \eta\sigma, \eta^2\sigma, \dots, \eta^{k-1}\sigma)$ where k is the order of η .

For a practical version of the algorithm choose a small collection of permutations $\eta_1, \eta_2, \dots, \eta_k$ that generate \mathcal{S}_n and use these to generate partitions P_1, P_2, \dots, P_k .

The walk is connected.

We remark in closing that the usual random walks on the symmetric group (generation of uniformly distributed random permutations) does NOT exhibit diffusive behavior (see [4]) for a review.

4 Cautionary Remarks

This section collects together caveats and pointers to competitive algorithms :

1. Our most general algorithm depends on having "directions" in the underlying state space. These may be difficult to find and, if available, they may be used for yet better algorithms. For example, consider the contingency table example of section 3.3. There, a direction was specified by a choice of a pair of rows and a pair of columns. In our implementation, a directed walk was taken in this direction. An alternative (implemented in [8]) considers the 4 fixed cells as a 2×2 table and chooses uniformly among all the 2×2 tables with the same margins.

This is easy to do, a 2×2 table being specified by one entry which varies between easily computed bounds. A similar comment holds for the more general problems described in [8].

2. The $(\text{diameter})^2$ behavior is associated with random walk and so to uniform or relatively flat stationary distributions.

Available theory [7] shows that when the Metropolis algorithm is used following a random walk on a low dimensional grid to generate from a stationary distribution with exponential peaks the walk basically heads directly for the nearest peak. Thus if the stationary distribution is unimodal order diameter steps suffice for stationarity. For multimodal distributions, any local algorithm, including ours, can effectively get stuck.

3. It is instructive to compare the present algorithm with an iid sampling Metropolis algorithm.

This is based on iid uniform choices from the state space followed by the usual Metropolis step to give stationary distribution $\pi(x)$. Call this chain $M_u(x, y)$. Suppose the state space has N points and let $\pi^* = \max_x \pi(x)$. Then Liu [17] shows

$$\|M_u - \pi\| \leq \left(1 - \frac{1}{N\pi^*}\right)^k$$

Let us use Liu's result on

$$\mathcal{X}(n, d) = \{(i_1, \dots, i_d), 1 \leq i_j \leq n, j = 1, \dots, d\}, |\mathcal{X}| = n^d$$

Example 1

Take $\pi(i) = ze^{-(i_1+i_2+\dots+i_d)}$. Then the normalizing constant is bounded uniformly in n for fixed d and the bound shows that order $n^d e^{-d}$ are sufficient for stationarity. It is not hard to prove a lower bound showing they are necessary as well. Thus here the iid Metropolis is slow. The analysis in [7] shows that the classical Metropolis algorithm, (and presumably the present algorithm) reaches stationarity in order nd steps for this example.

Example 2

Let $p(x) = \sum_{\alpha} a_{\alpha} x^{\alpha}$ be a polynomial with non negative coefficients and maximum degree $|\alpha^*| = \alpha_1^* + \alpha_2^* + \dots + \alpha_d^*$, for example $p(x) = x_1 + x_2 + \dots + x_d$ or $p(x) = x_1 x_2 \dots x_d$. Let $\pi(i) = zp(i)$ on $\mathcal{X}(n, d)$. Here 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 [7] shows that the classical Metropolis algorithm requires order n^2 steps to reach stationarity. In line with the results of section 2 we conjecture that order n steps are necessary and suffice for the directed algorithms.

References

- [1] Bélisle, C.J., Romeijn, H. E., Smith, R.L., (1993)
Hit-and-run algorithms for generating multivariate distributions,
Math. Oper. Research, 18, 255-266.
- [2] Conway J., Sloane N., Wilks A. (1989)
Gray codes for reflection groups.
Graphs. Combin. ,5 315-325.
- [3] Chung,F. Graham R., Yau (1997)
Some tables paper to appear.
- [4] Diaconis P. (1986)
Group representations in probability and statistics.
IMS, Hayward.

- [5] Diaconis P. and Saloff-Coste L. (1993)
Comparison theorems for reversible Markov chains.
Ann. Appl. Prob. 3, 696-730.
- [6] Diaconis P. and Saloff-Coste L. (1992)
Moderate growth and random walk on finite groups.
G.A.F.A., 4, 1-36.
- [7] Diaconis P. and Saloff-Coste L. (1997)
What do we know about the Metropolis algorithm?
to appear in J.C.S.S.
- [8] Diaconis P. and Sturmfels B. (1997)
Algebraic algorithms for sampling from conditional distributions,
to appear in Ann. Stat. .
- [9] Diaconis P., Eisenbud D. and Sturmfels B. (1996)
Lattice Walks and Primary Decomposition, to appear in Rotafest.
- [10] Diaconis P., Efron, B. (1989)
Probabilistic-geometric theorems arising from the analysis of contingency tables.
Contributions to the theory and application of statistics
Volume in honour of Herb Solomon.
- [11] Diaconis P., Fill J. (1990)
Strong Stationary Times via a new form of Duality
Ann. Prob., 18, 1483-1522.
- [12] Diaconis P., Gangolli (1996)
Rectangular arrays with fixed margins.
in Finite Markov Chain Renaissance, Springer Verlag -IMA series pp15-42
- [13] Diaconis P. and Holmes S., (1994)
Gray Codes for Randomization Procedures.
Statistics and Computing vol. 4 no 4., 207-302
- [14] Duane S., Kennedy I., Pendleton B., Roweth D. (1987)
Hybrid Monte Carlo,
Phys. Letters B, 195, 216-222.
- [15] Goodman J. and Sokal, A. (1989)
Multigrid Monte Carlo methods,
Phys. Rev. D 40, 2035
- [16] Horowitz A. (1991)
A generalized guided Monte Carlo algorithm,
Phys. Letters B, 268, 247-252.
- [17] Liu, J. (1996)
Metropolized independent sampling with comparisons to rejection sampling and importance sampling.
Statistics and Computing, 6, 113-119.
- [18] Metropolis N., Rosenbluth A., Rosenbluth M., Teller A., Teller E. (1953)
Equation of State calculations by fast computing machines,
J. Chem. Phys., 21, 1087-

- [19] Neal R. (1996)
Bayesian Learning for Neural Networks.
Lecture Notes in Statistics, vol.118, Springer Verlag.
- [20] Smith R.L. (1984)
Efficient Monte Carlo procedures for generating points distributed uniformly over bounded regions.
Oper. Research, 32, 1296-1308.
- [21] StatXact Software (1992),
commercial software package, by Cytel Company.