KERNEL SELECTION FOR CONVERGENCE AND EFFICIENCY IN MARKOV CHAIN MONTE CARLO

CHRISTOPHER C.J. POTTER

24 April 2013

Department of Mathematical Sciences
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Robert H. Swendsen (thesis advisor)
Shlomo Ta'asan (committee chair)
Jack Schaeffer
Michael Widom

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Copyright © 2013 Christopher C. J. Potter

Keywords:

Markov Chain Monte Carlo, detailed balance, acceptance ratio, Metropolis algorithm

Dedicated to people in every nation who die nobly or live humbly for the cause of freedom.

ABSTRACT. Markov Chain Monte Carlo (MCMC) is a technique for sampling from a target probability distribution, and has risen in importance as faster computing hardware has made possible the exploration of hitherto difficult distributions. Unfortunately, this powerful technique is often misapplied by poor selection of transition kernel for the Markov chain that is generated by the simulation.

Some kernels are used without being checked against the convergence requirements for MCMC (total balance and ergodicity), but in this work we prove the existence of a simple proxy for total balance that is not as demanding as detailed balance, the most widely used standard. We show that, for discrete-state MCMC, that if a transition kernel is equivalent when it is "reversed" and applied to data which is also "reversed", then it satisfies total balance. We go on to prove that the sequential single-variable update Metropolis kernel, where variables are simply updated in order, does indeed satisfy total balance for many discrete target distributions, such as the Ising model with uniform exchange constant.

Also, two well-known papers by Gelman, Roberts, and Gilks (GRG)[1, 2] have proposed the application of the results of an interesting mathematical proof to the realistic optimization of Markov Chain Monte Carlo computer simulations. In particular, they advocated tuning the simulation parameters to select an acceptance ratio of 0.234.

In this paper, we point out that although the proof is valid, its result's application to practical computations is not advisable, as the simulation algorithm considered in the proof is so inefficient that it produces very poor results under all circumstances. The algorithm used by Gelman, Roberts, and Gilks is also shown to introduce subtle time-dependent correlations into the simulation of intrinsically independent variables. These correlations are of particular interest since they will be present in all simulations that use multi-dimensional MCMC moves.

CONTENTS

List of Tables i

List of Figures ii

Acknowledgements iii

Preface iv

Part 1. Preliminaries 1

1.1. Measure Theory 1

1.2. Markov Chain Monte Carlo 5

1.3. Metropolis algorithm 14

Part 2. Balance theorems for Sequential and Checkerboard Update 30

2.1. Sequential update for the Ising Model 30

2.2. The checkerboard decomposition 33

2.3. Guaranteeing Detailed Balance for Subset Updates 37

2.4. Guaranteeing Total Balance for Subset Updates 41

Part 3. Effects of Joint Update Schemes 45

3.1. Time-Shifted Cross-Correlations 45

3.2. Efficiency of simulations using joint update scheme 54

3.3. Conclusions 67

Part 4. Conditions for Π-irreducibility and aperiodicity 69

4.1. Fixed cycle subset update 69

4.2. Global update and mixture or mixed cycle subset update 79

References 80

Index 82

List of Tables

2.1 Performance of total balance kernels for Ising model 45

3.1 Time-shifted correlation coefficient for jointly updated simple harmonic oscillators from MCMC and numerical integration 50

3.2 Conditional mean of one oscillator given the other in a high energy state 53

3.3 Optimal acceptance ratio and correlation time for simple harmonic oscillator system 58

3.4 Time to halve mean displacement for symmetric anharmonic oscillator system 59

3.5 Displacement correlation time for asymmetric anharmonic oscillator system 60

LIST OF FIGURES

3.1 Correlation coefficient and conditional expectation tests for explanation of time-shifted cross-correlation 52

3.1 Optimal acceptance ratio and correlation time for simple harmonic oscillator system 57

3.2 Evolution of mean displacement for symmetric anharmonic oscillator system 60

3.3 Optimal acceptance ratio and correlation times for asymmetric anharmonic oscillator system 61

3.4 Effect of proposal distribution on correlation time for simple harmonic oscillator 63

3.5 Gaussian proposed move behavior with increasing dimension 66

4.1 Non-irreducible with connected support 69

4.2 Illustration of proof of irreducibility with connected support 75

4.3 Illustration of proof of irreducibility with disconnected support 77

Acknowledgements

I think the most valuable non-mathematical lesson to be taken from my graduate study is how dependent every serious scholar or researcher must be on the help of others. I consider myself lucky to have come to Carnegie Mellon University with a cadre of great mathematicians only a short walk away. Also a short walk away was my advisor, Robert Swendsen in the Department of Physics. Bob was a very patient but constantly encouraging impetus to my research, and took lots of time to introduce me to the area on which my thesis was based, of which I had little knowledge when I started here at CMU. I also would like to thank Prof. Bill Hrusa, who suggested that I work with Bob as my advisor, and helped me at many points in my time here at CMU both with mathematical issues and administrative ones. It's impossible to overstate the great and taxing work he does for the graduate students in mathematics. The list of faculty across CMU who helped ready me for research work here would be too long to fit here, but special recognition is owed to Giovanni Leoni, Russell Schwartz, Shlomo Ta'asan, Jack Schaeffer, and Noel Walkington.

More than a short walk away is Yuko Okamoto of Nagoya University, whom I thank for suggesting to my advisor (who suggested it to me) taking a look at the 0.234 acceptance ratio result of Gelman, Roberts, and Gilks, which inspired a significant part of the thesis.

There is also a great deal of thanks due to the community of budding mathematicians in the graduate program at Carnegie Mellon. In particular I mention my very patient office sharer, Jing Wang, who endured the steady encroachment of messes of books and papers expanding ever-further from my desk as the years went on, and my habit of hosting gatherings in our office with boisterous discussions of PDE theory or finite element methods or, on particularly unfortunate occasions, religion and politics. Also I would have been much worse off without several productive discussions about my research with Brian Kell, Will Gunther, Jing Liu, Sheila Chandran, Lou Perrotti, and Robert Simione. Dwelling in the midst of so many very smart people is truly humbling.

Finally, I should mention that some of the phrasing in the Dedication is borrowed from a quote by Austrian psychologist Wilhelm Stekel, via J.D. Salinger's famous work *The Catcher In The Rye*.

Preface

Markov Chain Monte Carlo (MCMC) is an indirect method for generating random values from a probability distribution Π which is difficult to sample from directly. This is accomplished by generating a Markov chain, a sequence of values (called terms) with each term depending the value of its predecessor according to some probabilistic rule, called a stochastic kernel. The behavior of a Markov chain is completely determined by the choice of stochastic kernel, so the kernel must be selected so that the terms of the chain have the correct probability distribution Π. This is the concern of guaranteeing convergence.

In practice, we would also like the Markov chain to visit many areas of the sample space in as few terms as possible. Doing so would allow us to estimate statistics for the distribution using a small amount of computer time. It is in this way that we optimize efficiency. As computer simulations using MCMC have found extremely broad uses that go far beyond the original physics applications proposed by Metropolis, et al.[12], the optimization of MCMC has become an important practical issue, as well as an interesting subject of research.

After some preliminary background material in Part 1, we discuss the properties that a kernel must have if it is to produce a Markov chain with the correct distribution. The bare minimum is total balance, a condition which is difficult to verify directly. A stronger condition, called detailed balance, is widely used as a proxy for total balance, but is unnecessarily strict in some cases. The heavily cited work of Manousiouthakis and Deem [17] shows that there is an alternative proxy for total balance; however, this alternative is inapplicable to several stochastic kernels which are commonly used in MCMC, such as the sequential update for the Ising model. In Part 2, we show that under fairly broad conditions, a kernel can satisfy total balance without satisfying detailed balance. We show that this is true for the sequential update and other important kernels. As the sequential update is simpler to program and generally results in faster code than other more complicated update schemes, this result should have wide application in MCMC.

On the topic of MCMC optimization, widely cited papers by Gelman, Roberts, and Gilks[1, 2] have claimed that for MCMC simulations of models with a large number of independent variables, a kernel should be selected which produces an acceptance ratio (an important parameter for an MCMC simulation) with the numerical value of 0.234. The elegant simplicity of the GRG result, as well as the fact that it is based on a mathematical proof, has brought it a great deal of attention. However, in Part 3, we show that the 0.234 rule is of little use for practical applications; the mathematical proof behind it is only valid for an inherently inefficient class of kernels, the global updates, which in practice result in very small changes between successive states in a large number of dimensions, inhibiting the chain's ability to spread quickly through the possible states. Using the 0.234 rule with a global update kernel is valid but inadvisable, while using the rule with other kernels is simply incorrect; there are different "optimal acceptance ratios" for other kernels, as has been shown by previous work [4, 5].

On the other hand, the algorithm considered by Gelman and co-workers[1, 2] does have interesting properties that provide a cautionary tale for commonly-used algorithms. Chief among these properties is the existence of time-shifted cross-correlations. In this phenomenon, variables which are supposed to be independent under the distribution Π are indeed independent from each other within each term of the Markov chain. However, the values of these variables taken from terms separated by a fixed number of time steps may be correlated. We verify the existence of this bizarre and previously unseen phenomenon in Part 3, and offer a likely explanation for its occurrence in an MCMC simulation using the global update kernel.

Finally, in Part 4, we broaden the applicability of a result by Chan and Geyer [10] which addresses the guaranteeing of convergence. Our proof applies to some combinations of distributions Π and choices of stochastic kernels which the previous result did not, chief among these the single-variable update Metropolis kernel using uniform distributions to propose moves.

Part 1. **Preliminaries**

1.1. MEASURE THEORY

Definition 1. A collection $\mathcal{A} \subseteq \mathfrak{P}(\Omega)$ is called a σ-algebra on Ω provided that

(1) $\emptyset \in \mathcal{A}$
(2) For any $A \in \mathcal{A}$, we have $\Omega \backslash A \in \mathcal{A}$
(3) For any sequence $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{A}$, we have $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$

This definition implies also that $\Omega \in \mathcal{A}$, and for any sequence $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{A}$, we have $\bigcap_{n=1}^{\infty} A_n \in \mathcal{A}$. An important example of a σ-algebra is the Borel σ-algebra on Ω, which contains all open and closed sets in Ω, along with the other sets necessary to produce a σ-algebra.

Definition 2. Given a set Ω and σ-algebra $\mathcal{A}$ on Ω, a function $\mu : \mathcal{A} \to [0, \infty]$ is called a measure provided that

(1) For $A, B \in \mathcal{A}$ with $A \subseteq B$, we have $\mu(A) \leq \mu(B)$
(2) For any sequence $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{A}$ with $A_m \cap A_n = \emptyset$ if $m \neq n$, we have

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu\left(A_n\right) \tag{1.1.1}$$

Note that a measure μ may take the value ∞, and we abide by the convention that for any $a \in [0, \infty]$, we have $a \leq \infty$ and $a + \infty = \infty$. A set in the domain $\mathcal{A}$ is said to be μ-measurable. A measure space $(\Omega, \mathcal{A}, \Pi)$ is said to be a probability space, and Π a probability measure (or distribution) provided $\Pi(\Omega) = 1$.

Example 3. An important measure is the counting measure Card (for cardinality). For any Ω, the value $\mathsf{Card}(A)$ is defined for every $A \in \mathfrak{P}(\Omega)$ to be equal to the number of members of A.

Definition 4. Given a measure space $(\Omega, \mathcal{A}, \mu)$ and a set $A \subseteq \Omega$, we say a logical statement $\mathcal{P}(x)$ is true for μ-almost every $x \in A$, provided that the set

$$F := \{x \in A : \mathcal{P}(x) \text{ is false}\} \tag{1.1.2}$$

has $\mu(F) = 0$.

Definition 5. Given a measure space $(\Omega, \mathcal{A}, \mu)$, we define the completion $\left(\Omega, \bar{\mathcal{A}}, \bar{\mu}\right)$ by letting

$$\bar{\mathcal{A}} := \mathcal{A} \cup \{B \in \mathfrak{P}(\Omega) : B \subseteq A \text{ for some } A \in \mathcal{A} \text{ with } \mu(A) = 0\} \tag{1.1.3}$$

and

$$\bar{\mu}(A) := \begin{cases} \mu(A), & A \in \mathcal{A} \\ 0, & A \in \bar{\mathcal{A}} \backslash \mathcal{A} \end{cases} \tag{1.1.4}$$

The measurable space $(\Omega, \mathcal{A}, \mu)$ is said to be complete provided $\bar{\mathcal{A}} = \mathcal{A}$.

Note that this also means that μ is complete if and only if, for every $A, B \in \mathcal{A}$ and $C \in \mathfrak{P}(\Omega)$ with $A \subseteq C \subseteq B$ and $\mu(A) = \mu(B) < \infty$, we have $C \in \mathcal{A}$ and $\mu(C) = \mu(B) = \mu(A)$. That is, a set between two sets of equal finite measure must be measurable and have the same measure as the others.

An important complete measure is the Lebesgue measure on $\mathbb{R}$, denoted by $\mathcal{L}^1$, defined for the collection of Lebesgue-measurable sets $\mathcal{M}_1$, which includes all the Borel sets along with some (but not all) other subsets of $\mathbb{R}$. For "visualizable" sets, the Lebesgue measure corresponds to the length of a set on the real line. A more rigorous definition for $A \in \mathcal{M}_1$ is

$$\mathcal{L}^1(A) := \inf \left\{ \sum_{n=1}^{\infty} s_n : \{x_n\}_{n=1}^{\infty} \in \mathbb{R}^{\mathbb{N}}, \{s_n\}_{n=1}^{\infty} \in (0, \infty)^{\mathbb{N}}, A \subseteq \bigcup_{n=1}^{\infty} (x_n, x_n + s_n) \right\} \tag{1.1.5}$$

that is, the largest number which is always smaller than the sum of the lengths of open intervals which cover the set A.

Definition 6. Given measure spaces $(\Omega, \mathcal{A}, \mu)$ and $(\Psi, \mathcal{B}, \nu)$, we can define the product space $(\Omega \times \Psi, \mathcal{A} \times \mathcal{B}, \mu \times \nu)$ by

$$(\mu \times \nu)(A \times B) := \mu(A)\nu(B) \tag{1.1.6}$$

for $A \in \mathcal{A}$ and $B \in \mathcal{B}$.

Another important measurable space is that of the N-dimensional Lebesgue measure $\left(\mathbb{R}^N, \mathcal{M}_N, \mathcal{L}^N\right)$. This is, roughly speaking, the area of a set in the plane for $N = 2$, volume for $N = 3$, and so on. It is rigorously defined recursively to be the completion of the product measurable space $\left(\mathbb{R}^N, \mathcal{M}_{N-1} \times \mathcal{M}_1, \mathcal{L}^{N-1} \times \mathcal{L}^1\right)$ for $N > 1$.

Definition 7. Given a measure space $(\Omega, \mathcal{A}, \mu)$ and a topology τ for Ω, we define the support of the measure μ by

$$\mathsf{supp}_\tau(\mu) := \{x \in \Omega : \mu(A) > 0 \text{ for every set } A \text{ open in } \tau \text{ with } x \in A\} \tag{1.1.7}$$

(Usually we will omit the subscript when the topology is obvious.) Note that the support is always closed. For a separable metric space (such as $\mathbb{R}^N$), every open cover has a countable subcover, so

$$\mu\left(\Omega \backslash \mathsf{supp}(\mu)\right) = 0 \tag{1.1.8}$$

since the complement of the support is the union of countably many open sets with measure zero. Thus $\mu(\Omega) = \mu\left(\mathsf{supp}(\mu)\right)$.

Definition 8. Given a measure space $(\Omega, \mathcal{A}, \mu)$, we say the measure μ is σ-finite provided that there is some sequence $\{A_n\}_{n=1}^{\infty}$ of μ-measurable sets with $\mu(A_n) < \infty$ for every $n \in \mathbb{N}$, such that

$$\Omega = \bigcup_{n=1}^{\infty} A_n \tag{1.1.9}$$

Example 9. If $\Omega = \mathbb{R}$, the counting measure Card is not σ-finite, since the union of countably many finite sets is always countable, and $\mathbb{R}$ is uncountable. However, the

Lebesgue measure $\mathcal{L}^1$ is σ-finite since

$$\mathbb{R} = \bigcup_{n=1}^{\infty} ([-n, -n+1] \cup [n-1, n]) \tag{1.1.10}$$

and each of the unioned sets has Lebesgue measure 2.

Definition 10. Given a measure space $(\Omega, \mathcal{A}, \mu)$ and a measure ν on $\mathcal{A}$, we say that μ is absolutely continuous with respect to ν, and write $\mu \ll \nu$, provided that for any $A \in \mathcal{A}$ with $\nu(A) = 0$, we also have $\mu(A) = 0$.

Example 11. Given a measure space $(\Omega, \mathcal{A}, \mu)$ and a set $A \in \mathcal{A}$ with $\mu(A) > 0$, the *uniform distribution* on A is a probability measure defined for $B \in \mathcal{A}$ by

$$\mathsf{Unif}_A(B) := \frac{\mu(A \cap B)}{\mu(A)} \tag{1.1.11}$$

Clearly we always have $\mathsf{Unif}_A \ll \mu$.

Theorem 12. (Radon-Nikodym) *Let a measure space $(\Omega, \mathcal{A}, \mu)$ be given, along with another measure ν on $\mathcal{A}$. If μ and ν are σ-finite and $\mu \ll \nu$, then there is some function $m : \Omega \to \mathbb{R}$, called the density function of μ with respect to the basis measure ν, such that*

$$\mu(A) = \int_A m(x)\nu(dx) \tag{1.1.12}$$

Note that since, for any $B \subseteq A$ with $\nu(B) = 0$ we have

$$\mu(A) = \int_{A \setminus B} m(x)\nu(dx) + \int_B m(x)\nu(dx) = \int_{A \setminus B} m(x)\nu(dx) \tag{1.1.13}$$

regardless of the values of $m(x)$ for $x \in B$, the density function is well-defined only for sets with strictly positive ν-measure. If the basis measure is $\mathcal{L}^N$, for example, it is impossible to specify the value of the density function at a single point in isolation.

Definition 13. Given a measure space $(\Omega, \mathcal{A}, \mu)$ and a measure ν on $\mathcal{A}$, we say that μ is singular with respect to ν, and write $\mu \perp \nu$, provided that there is some $A \subseteq \Omega$ such that $\mu(B) = 0$ for every $B \in \mathcal{A} \cap \mathfrak{P}(A)$, and $\nu(C) = 0$ for every $C \in \mathcal{A} \cap \mathfrak{P}(\Omega \backslash A)$.

Theorem 14. (Lebesgue decomposition) *Let a measure space $(\Omega, \mathcal{A}, \mu)$ be given, along with another measure ν on $\mathcal{A}$. If μ and ν are σ-finite, then there are unique signed measures μ_{AC} and μ_{sing} on $\mathcal{A}$ such that*

$$\mu = \mu_{\mathsf{AC}} + \mu_{\mathsf{sing}} \tag{1.1.14}$$

and $\mu_{\mathsf{AC}} \ll \nu$ and $\mu_{\mathsf{sing}} \perp \nu$.

1.2. Markov Chain Monte Carlo

In most practical situations, MCMC is used to produce a sequence of samples from a state space Ω distributed according to a distribution Π from which it is difficult to generate samples directly. This is accomplished by producing a Markov process with Π as its equilibrium distribution. For a probability space with $\mathcal{A}$ the collection of measurable sets, a discrete-time Markov process consists of a sequence of states $X(t) \in \Omega$ for time steps $t \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, satisfying

$$\mathsf{Pr}\left[X(t+1) \in A | X(t)\right] = \mathsf{Pr}[X(t+1) \in A | X(t), X(t-1), \ldots, X(0)] \tag{1.2.1}$$

for every set $A \in \mathcal{A}$. That is, the distribution of $X(t+1)$ depends solely on the value of $X(t)$, and information about the values of earlier states in the chain does not alter the distribution of $X(t+1)$. This is termed the Markov property.

1.2.1. **Notation.** The dependence of $X(t+1)$ on $X(t)$ can be characterized by a stochastic kernel, which represents the probability of moving from a state x to some state in a measurable set A in one time step. In general, a mapping $K : \Omega \times \mathcal{A} \to [0,1]$ is a *stochastic kernel* provided that

- $K(x, \cdot)$ is a measure on the collection $\mathcal{A}$ for all $x \in \Omega$,
- $K(\cdot, A)$ is a measurable function for all $A \in \mathcal{A}$,

- $K(x, \Omega) = 1$ for all $x \in \Omega$.

The product of kernels K_1 and K_2 represents the probability of moving from state x to a state in the set A by applying the kernel K_1 and then K_2. This is denoted by K_1K_2 and defined formally as

$$(K_1K_2)(x, A) := \int_{y \in \Omega} K_1(x, dy) K_2(y, A) \tag{1.2.2}$$

for continuous state spaces Ω, and

$$(K_1K_2)(x, A) := \sum_{y \in \Omega} K_1(x, y) K_2(y, A) \tag{1.2.3}$$

for discrete spaces Ω. (For the remainder of this section, we will assume continuous state spaces; the definitions for discrete state spaces are analogous, replacing integrals with sums.)

We define for $n \in \mathbb{N}_0$ the nth iterate of the kernel K, denoted by $K^n(x, A)$, which represents the probability of moving from x to some state in A by applying the kernel K exactly n times. For $n \in \mathbb{N} \cap [2, \infty)$, we define

$$K^n(x, A) \;:=\; \int_{y_1 \in \Omega} K(x, dy_1)\, K(y_{n-1}, A) \prod_{i=2}^{n-1} K(y_{i-1}, dy_i) \tag{1.2.4}$$

and by convention set $K^1(x, A) := K(x, A)$ and $K^0(x, A) := \mathbf{1}_A(x)$, the characteristic function of A.

Supposing H is another probability measure on $\mathcal{A}$, we also define a notation for the probability of a random variable distributed according to H taking a value in the set A after the application of the kernel K. This is defined formally as

$$(HK)(A) \;:=\; \int_{\Omega} K(x, A)\, H(dx) \tag{1.2.5}$$

For a Markov chain X, we can formulate the dependence of $X(t+1)$ on its immediate predecessor $X(t)$ using a stochastic kernel P_t, called the transition kernel at time step t.

For all $t \in \mathbb{N}_0$, $x \in \Omega$, and $A \in \mathcal{A}$, we have

$$\mathsf{Pr}[X(t+1) \in A | X(t) = x] = P_t(x, A) \tag{1.2.6}$$

If $P_t = P_0$ for all $t \in \mathbb{N}_0$, so that the transition kernel is independent of time step, we say the process X is time-homogeneous (or stationary) and denote the transition kernel as P. We can then write

$$\mathsf{Pr}[X(t+n) \in A | X(t) = x] = P^n(x, A) \tag{1.2.7}$$

for any $n \in \mathbb{N}_0$. Finally, we denote the distribution of a Markov process at time $t \in \mathbb{N}_0$ by H_t, so that

$$H_t(A) := \mathsf{Pr}[X(t) \in A] \tag{1.2.8}$$

We define the total variation of the probability measures G and H defined on the collection $\mathcal{A}$ by

$$\delta(G, H) := \frac{1}{2} \sup_{A \in \mathcal{A}} |G(A) - H(A)| \tag{1.2.9}$$

1.2.2. **Discrete-state process.** If $\mathsf{supp}(\Pi)$ is countable, then the target distribution Π is considered discrete, regardless of the cardinality of Ω. For example, if $\Omega = \mathbb{R}$ and $\mathcal{A} = \mathfrak{P}(\mathbb{R})$, a valid probability measure on $\mathcal{A}$ is defined by

$$\Pi(A) := \begin{cases} \sum_{n \in A \cap \mathbb{N}} 2^{-n}, & A \cap \mathbb{N} \neq \emptyset \\ 0, & A \cap \mathbb{N} = \emptyset \end{cases} \tag{1.2.10}$$

for all $A \in \mathcal{A}$, and we would consider this a discrete distribution even though Ω is uncountable. Of course, it would be more natural to define this probability measure for $\Omega = \mathbb{N}$ and $\mathcal{A} = \mathfrak{P}(\mathbb{N})$ by

$$\Pi(A) := \sum_{n \in A} 2^{-n} \tag{1.2.11}$$

for all $A \in \mathcal{A}$, so we will assume for discrete distributions that our sample space is $\Omega_+ := \mathsf{supp}(\Pi)$. Note that this means $\Pi(\{x\}) > 0$ for all $x \in \Omega_+$.

1.2.2.1. *Definition.* If the distribution Π is discrete, then for each $t \in \mathbb{N}_0$ and $A \in \mathcal{A}$, we have

$$P_t(x, A) = \sum_{y \in A} P_t\left(x, \{y\}\right) \tag{1.2.12}$$

so the transition kernel is completely determined by the values of $P_t\left(x, \{y\}\right)$ for $x, y \in \Omega$. It seems sensible in the discrete-state case to define the mass kernel and mass function

$$p_t(x, y) := P_t\left(x, \{y\}\right) \tag{1.2.13}$$

$$h_t(x) := H_t\left(\{x\}\right) \tag{1.2.14}$$

We then have

$$\begin{aligned}
\mathsf{Pr}[X(t+1) = x] &= \left(H_t P_t\right)\left(\{x\}\right) \\
&= \sum_{y \in \Omega} h_t(y) p_t(x, y) \\
&= h_t(x) \left(1 - \sum_{y \in \Omega \setminus \{x\}} p_t(x, y)\right) + \sum_{y \in \Omega \setminus \{x\}} h_t(y) p_t(x, y) \\
&= H_t\left(\{x\}\right) + \sum_{y \in \Omega \setminus \{x\}} \left[h_t(y) p_t(y, x) - h_t(x) p_t(x, y)\right]
\end{aligned} \tag{1.2.15}$$

This leads to the discrete-state master equation

$$\mathsf{Pr}[X(t+1) = x] - \mathsf{Pr}[X(t) = x] = \sum_{y \in \Omega \setminus \{x\}} \left[h_t(y) p_t(y, x) - h_t(x) p_t(x, y)\right] \tag{1.2.16}$$

1.2.2.2. *Conditions for convergence to equilibrium distribution* Π. Let Π be the discrete target distribution which we wish to sample from. Suppose that X is a time-homogeneous, that is $P_t = P_0$ for all $t \in \mathbb{N}_0$. Then define $p := p_0$ and $\pi(x) := \Pi(\{x\})$. In this discrete case,

it is sensible to work with the mass functions h_t and π rather than the distributions H_t and Π.

We seek to guarantee the convergence of h_t to π in the sense of total variation, that is, $\delta(H_t, \Pi) \to 0$ as $t \to \infty$. It is sufficient to show that p is ergodic and has π as an equilibrium mass function. By Theorem 2 in Tierney [9], since Ω is countable, we have p ergodic if it is π-irreducible and has an aperiodic state with positive π-probability. For p to be π-irreducible requires that, for any $x, y \in \Omega$ with $\pi(y) > 0$, we must have some $n \in \mathbb{N}_0$ such that $P^n(x, \{y\}) > 0$. This means that for any pair of states it is possible to travel from one to the other in some finite time.

If a state $x \in \Omega$ can only occur during time steps exactly $d \in \mathbb{N}$ apart, we say that x is periodic with period d. Formally, this requires that for all $c \in \mathbb{N} \cap [1, d)$,

$$\mathsf{Pr}[X(t+d) = x | X(t) = x] > 0 \tag{1.2.17}$$

$$\mathsf{Pr}[X(t+c) = x | X(t) = x] = 0 \tag{1.2.18}$$

A state is said to be aperiodic if it is not periodic with respect to any period d.

For p to have π as an equilibrium distribution, we need to ensure that if $h_{t_0} = \pi$ for some $t_0 \in \mathbb{N}_0$, then $h_t = \pi$ for all $t \geq t_0$. That is, the distribution of $X(t)$ does not change once it matches the target distribution π. From the master equation (1.2.16) this will be satisfied if and only if

$$0 = \sum_{y \in \Omega \setminus \{x\}} [\pi(y)p(y,x) - \pi(x)p(x,y)] \tag{1.2.19}$$

for each $x \in \Omega$. This condition is total balance, and may be cumbersome to verify. A stronger condition, which is often easier to verify, is detailed balance, which requires that each term of the sum vanish. That is,

$$\pi(y)p(y,x) = \pi(x)p(x,y) \tag{1.2.20}$$

for all $x, y \in \Omega$.

1.2.3. **Continuous-state process.** In contrast to the discrete distributions, we have the continuous (or diffuse) distributions where $\Pi(\{x\}) = 0$ for all $x \in \Omega$. If Π is to be a probability measure, this obviously requires that the sample space Ω and $\mathsf{supp}(\Pi)$ be uncountable. The most common continuous distributions in applications have $\Omega \subseteq \mathbb{R}^N$ and $\Pi \ll \mathcal{L}^N$, but other situations are easily constructed, for example the Cantor measure on $\mathbb{R}$ which is not only not absolutely continuous with respect to $\mathcal{L}^N$ but singular. The following applies to the more general situations as well.

1.2.3.1. *Definition.* Suppose the sample space $\Omega \subseteq \mathbb{R}^N$ and we have a vector-valued Markov process $\mathbf{X}$. Then for any $A \in \mathcal{A}$ we have

$$\begin{aligned} \mathsf{Pr}[\mathbf{X}(t+1) \in A] &= \int_\Omega P_t(\mathbf{x}, A) H_t(d\mathbf{x}) \\ &= \int_A \left[1 - P_t(\mathbf{x}, \Omega \backslash A)\right] H_t(d\mathbf{x}) + \int_{\Omega \backslash A} P_t(\mathbf{y}, A) H_t(d\mathbf{y}) \\ &= H_t(A) + \int_{\Omega \backslash A} P_t(\mathbf{y}, A) H_t(d\mathbf{y}) - \int_A P_t(\mathbf{x}, \Omega \backslash A) H_t(d\mathbf{x}) \end{aligned} \tag{1.2.21}$$

where H_t again represents the instantaneous distribution of $\mathbf{X}(t)$. This gives rise to the continuous-state master equation:

$$\mathsf{Pr}[\mathbf{X}(t+1) \in A] - \mathsf{Pr}[\mathbf{X}(t) \in A] = \int_{\Omega \backslash A} P_t(\mathbf{y}, A) H_t(d\mathbf{y}) - \int_A P_t(\mathbf{x}, \Omega \backslash A) H_t(d\mathbf{x}) \tag{1.2.22}$$

The master equation can be thought of as representing the net "probability flow" into the set $A \in \mathcal{A}$.

1.2.3.2. *Convergence to equilibrium distribution* Π. Suppose $\mathbf{X}$ is time-homogeneous. The process $\mathbf{X}$ converges to an equilibrium distribution Π as $t \to \infty$, in the sense of total variation on $(\Omega, \mathcal{A})$, that is if P is ergodic and Π is an invariant measure for P. From Theorem 1 in Tierney [9], if the latter condition holds, we can guarantee ergodicity as well if P is Π-irreducible and aperiodic.

Irreducibility. The kernel P is Π-irreducible if, for any $\mathbf{x} \in \Omega$, and $A \in \mathcal{A}$ with $\Pi(A) > 0$, there is some $n = n(\mathbf{x}, A) \in \mathbb{N}$ for which $P^n(\mathbf{x}, A) > 0$. Essentially, Π-irreducibility means that it is possible to enter any positive-probability set at some fixed time later in the process, regardless of the current state.

Aperiodicity. We say P is periodic if, there is some $d \in \mathbb{N}_0$ and disjoint sets $A_1, A_2, \ldots, A_d \in \mathcal{A}$ for which $P\left(\mathbf{x}, \Omega \backslash A_{j+1}\right) = 0$ whenever $\mathbf{x} \in A_j$. That is, the process may return to each set A_j once in every d time steps, but not at other time steps. If P is not periodic we say it is aperiodic. A sufficient condition for aperiodicity is that, for every $A \in \mathcal{A}$ and $\mathbf{x} \in A$ there exists $t_0 = t_0(\mathbf{x}, A) \in \mathbb{N}_0$ for which $P^t(\mathbf{x}, A) > 0$ for all $t \geq t_0$.

Invariant measure. For the distribution Π to be an invariant measure for P, we must ensure that, if there is some $t_0 \in \mathbb{N}_0$ for which $\mathbf{X}(t_0)$ has distribution Π, then $\mathbf{X}(t)$ must have distribution Π for all $t \geq t_0$. That is, $\Pi = \Pi P$. From the master equation (1.2.22) we see that this will be the case if and only if

$$\int_{\Omega \backslash A} P(\mathbf{y}, A)\, \Pi(d\mathbf{y}) = \int_A P(\mathbf{x}, \Omega \backslash A)\, \Pi(d\mathbf{x}) \tag{1.2.23}$$

for all $A \in \mathcal{A}$. This condition is total balance, but can be cumbersome to ensure in practice.

Detailed balance. As in the discrete case, it may be useful to find a criterion sufficient to guarantee total balance as in (1.2.23), but which only depends on the localized behavior of P and Π. Suppose $\Pi \ll \mu$ for some σ-finite measure μ on $\mathcal{A}$. Then by the Radon-Nikodym Theorem, we can find a density function π such that

$$\Pi(A) = \int_A \pi(\mathbf{x})\, \mu(d\mathbf{x}) \tag{1.2.24}$$

for all $A \in \mathcal{A}$.[1] We would like to proceed analogously to the discrete-state case above to obtain an equation similar to (1.2.20). Unfortunately, if $P\left(\mathbf{x}, \{\mathbf{x}\}\right) > 0$ for some $\mathbf{x} \in \Omega$, as will be the case for the Metropolis kernel discussed later, we know that P is not absolutely

[1] In the common continuous case where $\Pi \ll \mathcal{L}^N$, we can replace $\mu(d\mathbf{x})$ with $d\mathbf{x}$ and this is just ordinary Lebesgue integration.

continuous with respect to μ and thus does not have a density function with respect to μ. However, there is a relatively painless workaround.

Suppose that the process is never trapped in its current state, that is, $P(\mathbf{x}, \{\mathbf{x}\}) < 1$ for all $\mathbf{x} \in \Omega$. This is certainly the case for a continuous distribution Π, if P is Π-irreducible.[2] Then define the off-diagonal transition kernel as

$$P^\circ(\mathbf{x}, A) \;:=\; P(\mathbf{x}, A \backslash \{\mathbf{x}\}) \tag{1.2.25}$$

for $A \in \mathcal{A}$. This is technically a substochastic kernel, not stochastic, since we may have $P^\circ(\mathbf{x}, \Omega) < 1$ for some $\mathbf{x} \in \Omega$. The difference is not important for our applications, however.

This allows us to write

$$P(\mathbf{x}, A) = P^\circ(\mathbf{x}, A) + \mathbf{1}_A(\mathbf{x}) P(\mathbf{x}, \{\mathbf{x}\}) \tag{1.2.26}$$

where $\mathbf{1}_A$ is the characteristic function of A. If $P^\circ(\mathbf{x}, \cdot) \ll \mu$ for every $\mathbf{x} \in \Omega$, then we can find a density function $p(\mathbf{x}, \cdot)$ such that

$$P^\circ(\mathbf{x}, A) \;=\; \int_A p(\mathbf{x}, \mathbf{y})\, \mu(d\mathbf{y}) \tag{1.2.27}$$

for all $A \in \mathcal{A}$. Note that while π is a probability density function, $p(\mathbf{x}, \cdot)$ is not, unless $P(\mathbf{x}, \{\mathbf{x}\}) = 0$.

Since the second term in (1.2.26) will vanish for $\mathbf{x} \notin A$, this allows us to write the total balance relation (1.2.23) as

$$\int_{\Omega \backslash A} \int_A p(\mathbf{y}, \mathbf{x}) \pi(\mathbf{y})\, \mu(d\mathbf{x})\, \mu(d\mathbf{y}) \;=\; \int_A \int_{\Omega \backslash A} p(\mathbf{x}, \mathbf{y}) \pi(\mathbf{x})\, \mu(d\mathbf{y})\, \mu(d\mathbf{x}) \tag{1.2.28}$$

$$= \int_{\Omega \backslash A} \int_A p(\mathbf{x}, \mathbf{y}) \pi(\mathbf{x})\, \mu(d\mathbf{x})\, \mu(d\mathbf{y}) \tag{1.2.29}$$

[2] If $\Pi(\Omega \backslash \{\mathbf{x}\}) = 0$ for some $\mathbf{x} \in \Omega$, then Π is discrete.

where we have used Fubini's theorem in the second line. A sufficient condition for (1.2.29) to hold is that the two integrands be equal, that is,

$$\pi(\mathbf{y})p(\mathbf{y},\mathbf{x}) = \pi(\mathbf{x})p(\mathbf{x},\mathbf{y}) \tag{1.2.30}$$

for μ almost every $\mathbf{x},\mathbf{y} \in \Omega$. This condition is detailed balance, which implies total balance. It is often used to show that Π is an invariant measure for P since it is usually easier to verify in practice.

Thus, irreducibility, aperiodicity, and detailed balance are sufficient to guarantee convergence of the process $\mathbf{X}$ to a unique equilibrium distribution Π as $t \to \infty$, for Π almost every $\mathbf{X}(0) \in \Omega$.

1.3. METROPOLIS ALGORITHM

A common class of algorithms for producing a Markov process with equilibrium distribution Π are the Metropolis algorithms, the first variant of which was described in [12]. For the continuous state case, the algorithm assumes that we have a measurable space $(\Omega, \mathcal{A}, \mu)$ with the measure μ being σ-finite, and $\Pi \ll \mu$. In this case we have a density function as defined in (1.2.24).

1.3.1. **Algorithm definition.** Let us assume that a measurable space $(\Omega, \mathcal{A}, \mu)$ has already been given with μ being a σ-finite measure, and our target distribution Π satisfies $\Pi \ll \mu$.

In the case of a discrete distribution Π, we can use $\mu = \mathsf{Card}$ (the counting measure), and the density π will just be the probability mass function

$$\pi(\mathbf{x}) := \Pi(\{\mathbf{x}\}) \tag{1.3.1}$$

for $\mathbf{x} \in \Omega$. We also have, for any function $f : \Omega \to \mathbb{R}$ and $A \subseteq \Omega$,

$$\int_A f(\mathbf{x}) \mathsf{Card}(d\mathbf{x}) = \sum_{\mathbf{x} \in A} f(\mathbf{x}) \tag{1.3.2}$$

so the reader is encouraged to mentally substitute masses and sums in the following when Π is a discrete distribution.

1.3.1.1. *Metropolis kernel.* In the basic kernel of the Metropolis algorithm, we select for each $\mathbf{x} \in \Omega$ a proposal distribution $Q(\mathbf{x}, \cdot)$ on $\mathcal{A}$ and define an acceptance probability function for $\mathbf{x}, \mathbf{y} \in \Omega$

$$a(\mathbf{x},\mathbf{y}) := \begin{cases} 1, & \pi(\mathbf{x}) \le \pi(\mathbf{y}) \\ \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}, & \pi(\mathbf{x}) > \pi(\mathbf{y}) \end{cases} \tag{1.3.3}$$

$$= \begin{cases} \min\left\{1, \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\right\}, & \pi(\mathbf{x}) > 0 \\ 1, & \pi(\mathbf{x}) = 0 \end{cases} \tag{1.3.4}$$

where π is the density function of Π with respect to basis μ. We then define the Metropolis transition kernel

$$\begin{aligned} P(\mathbf{x},A) &:= \int_A a(\mathbf{x},\mathbf{y})\, Q(\mathbf{x},d\mathbf{y}) + \mathbf{1}_A(\mathbf{x}) \int_\Omega [1 - a(\mathbf{x},\mathbf{w})] Q(\mathbf{x},d\mathbf{w}) \\ &= \begin{cases} \int_A a(\mathbf{x},\mathbf{y})\, Q(\mathbf{x},d\mathbf{y}), & \mathbf{x} \notin A \\ 1 - \int_{\Omega\setminus A} a(\mathbf{x},\mathbf{w})\, Q(\mathbf{x},d\mathbf{w}), & \mathbf{x} \in A \end{cases} \end{aligned} \tag{1.3.5}$$

Note that π is only well-defined for sets of positive μ-measure. In isolation, $\pi(\mathbf{x})$ and $\pi(\mathbf{y})$ can take any real value, and thus so can $a(\mathbf{x},\mathbf{y})$. If $Q(\mathbf{x},\cdot) \not\ll \mu$, this fact can create ambiguity about the meanings of the integrals in (1.3.5). If $\mu(A) = 0$ and $Q(\mathbf{x},A) > 0$, then the values of $a(\mathbf{x},\mathbf{y})$ are totally arbitrary for all $\mathbf{y} \in A$ and the value of the integral is not well-defined. Thus we will insist that $Q(\mathbf{x},\cdot) \ll \mu$.

This also implies that the off-diagonal kernel $P^\circ(\mathbf{x},\cdot)$ as defined in (1.2.25) will be absolutely continuous with respect to μ. We denote by $p(\mathbf{x},\cdot)$ and $q(\mathbf{x},\cdot)$ the density functions of $P^\circ(\mathbf{x},\cdot)$ and $Q(\mathbf{x},\cdot)$ with respect to basis μ. Under these restrictions, we will have

$$p(\mathbf{x},\mathbf{y}) = q(\mathbf{x},\mathbf{y}) a(\mathbf{x},\mathbf{y}) \tag{1.3.6}$$

In many cases, we choose a probability measure Q^* on $\mathcal{A}$ and define the proposal distribution Q by

$$Q(\mathbf{x}, A) := Q^*(A - \mathbf{x}) \tag{1.3.7}$$

for all $\mathbf{x} \in \Omega$ and $A \in \mathcal{A}$. The resulting kernel is called a random-walk Metropolis kernel, and is by far the most common type of proposal distribution in practice, with Q^* usually being either a Gaussian distribution, or a uniform distribution on some useful kind of set in the space (for example, an interval in $\mathbb{R}$ or a hypersphere or rectangle in $\mathbb{R}^N$).

1.3.1.2. *Subset updates.*

Single-variable mixture kernel. If we are to achieve Π-irreducibility for a kernel P, that kernel must be able to propose changes to every variable in the system. A common, simple approach is to combine several "subkernels", each of which proposes a change to one variable.

Define for $j \in \mathbb{N} \cap [1, N]$ the subkernel P_j, which can change only the variable x_j, Denoting by p_j the density (or mass, in the discrete case) function for P_j, we can write this condition as

$$p_j(\mathbf{x}, \mathbf{y}) > 0 \implies x_k = y_k \text{ whenever } k \neq j \tag{1.3.8}$$

During each time step, we apply a randomly-chosen subkernel, with the probability of applying the kernel P_j during any time step being equal to some fixed $\zeta_j \in [0, 1]$. This produces the single-variable mixture kernel

$$P = \sum_{j=1}^{N} \zeta_j P_j \tag{1.3.9}$$

If P_j is a Metropolis kernel, P satisfies detailed balance as in (1.2.30). If, for every $j \in \mathbb{Z} \cap [1, N]$, there is some $r_j > 0$ for which $Q_j\left(\mathbf{x}, \mathbf{x} + \delta \mathbf{e}^{(j)}\right) > 0$ whenever $|\delta| < r_j$ for all $\mathbf{x} \in \Omega$, then this kernel is Π-irreducible as well. [9] The very weak conditions necessary to guarantee convergence of the time-varying distribution H_t to the target distribution Π make this a popular and simple approach to Markov Chain Monte Carlo. However, it

also has drawbacks. First, it requires the generation of an extra random number on each step. Second, the existence of long durations in the process $\mathbf{X}$ where a particular variable is not updated at all may cause the efficiency to suffer. Third, under some distributions Π, there are groups of variables highly correlated with one another, and it would be more natural to update these variables together, rather than separately. Other kinds of updates have been developed to address these drawbacks.

General Metropolis subkernels. A general way of producing combinations of kernels is to define several proposal distributions, each of which only affect a subset of the system variables, an *update set*. In the single-variable update described above, the update sets are singletons. In general, we have ℓ update sets $U_i \subseteq \mathbb{N} \cap [1, N]$, with $i \in \mathbb{N} \cap [1, \ell]$. We define for each i the subspace

$$V_i = \left\{ \mathbf{z} \in \Omega | z_j = 0 \text{ for all } j \notin U_i \right\} \tag{1.3.10}$$

and select a proposal distribution Q_i such that

$$Q_i(\mathbf{x}, A) = Q_i\left(\mathbf{x}, A \cap (V_i + \mathbf{x})\right) \tag{1.3.11}$$

for each $\mathbf{x} \in \Omega$ and $A \in \mathcal{A}$. We also define the measures ν_i on $\mathcal{A}$, with $\nu_i(A)$ equal to the $|U_i|$-dimensional Lebesgue measure of $A \cap V_i$.

For each i, we can define a Metropolis subkernel $P_i(\mathbf{x}, \cdot)$ by

$$\begin{aligned} P_i(\mathbf{x}, A) &= \int_{(V_i+\mathbf{x})\cap A} a(\mathbf{x}, \mathbf{y})\, Q_i(\mathbf{x}, d\mathbf{y}) + \mathbf{1}_A(\mathbf{x}) \int_{(V_i+\mathbf{x})\setminus A} [1 - a(\mathbf{x}, \mathbf{w})] Q_i(\mathbf{x}, d\mathbf{w}) \\ &= \begin{cases} \int_{(V_i+\mathbf{x})\cap A} a(\mathbf{x}, \mathbf{y})\, Q_i(\mathbf{x}, d\mathbf{y}), & \mathbf{x} \notin A \\ 1 - \int_{(V_i+\mathbf{x})\setminus A} a(\mathbf{x}, \mathbf{w})\, Q_i(\mathbf{x}, d\mathbf{w}), & \mathbf{x} \in A \end{cases} \end{aligned} \tag{1.3.12}$$

If $Q_i(\mathbf{x}, \cdot) \ll \nu_i(\mathbf{x}, \cdot)$, then we can find a density function for $Q_i(\mathbf{x}, \cdot)$ with respect to $\nu_i(\mathbf{x}, \cdot)$. We will denote it by $q_i(\mathbf{x}, \cdot)$. In this case, we can satisfy (1.3.11) by requiring that $q_i(\mathbf{x}, \mathbf{y}) = 0$ for $\mathbf{y} \notin V_i(\mathbf{x})$. Then

$$P_i^\circ(\mathbf{x}, A) = P_i(\mathbf{x}, A \backslash \{\mathbf{x}\}) = \int_{V_i(\mathbf{x}) \cap A \backslash \{\mathbf{x}\}} a(\mathbf{x}, \mathbf{y}) q_i(\mathbf{x}, \mathbf{y}) \tag{1.3.13}$$

so the off-diagonal density is

$$p_i(\mathbf{x}, \mathbf{y}) = a(\mathbf{x}, \mathbf{y}) q_i(\mathbf{x}, \mathbf{y}) \tag{1.3.14}$$

Independence sampler kernels. It is also possible to combine Metropolis transition kernels with independence samplers (also known as heat bath or Gibbs samplers). The independence sampler is so named because the change to the variables in its update set does not depend on the current values of those variables. Formally, if P_i is an independence sampler kernel, then for any $\mathbf{x}, \mathbf{y} \in \Omega$ we have

$$\mathbf{x} - \mathbf{y} \in V_i \Longrightarrow P_i(\mathbf{x}, \cdot) = P_i(\mathbf{y}, \cdot) \tag{1.3.15}$$

This property makes the independence sampler useful for escaping from regions of concentrated probability, where a random walk kernel may spend an inordinate number of time steps.

Suppose we define the subset kernel P_i to have density function equal to the conditional density for Π, given the values of all variables outside the ith update set. That is

$$p_i(\mathbf{x}, \mathbf{y}) := \begin{cases} \pi(\mathbf{y}) / Z_i(\mathbf{x}), & \mathbf{y} - \mathbf{x} \in V_i \\ 0, & \text{otherwise} \end{cases} \tag{1.3.16}$$

with the normalizing "constant"

$$Z_i(\mathbf{x}) := \int_{V_i} \pi(\mathbf{x}+\mathbf{z})\nu_i(d\mathbf{z}) \tag{1.3.17}$$

Then P_i is an independence sampler kernel, and for all $\mathbf{x}, \mathbf{y} \in \Omega$ we have

$$\pi(\mathbf{x})p_i(\mathbf{x},\mathbf{y}) = \begin{cases} \pi(\mathbf{x})\pi(\mathbf{y})/Z_i(\mathbf{x}), & \mathbf{y}-\mathbf{x} \in V_i \\ 0, & \text{otherwise} \end{cases} \tag{1.3.18}$$

$$= \pi(\mathbf{y})p_i(\mathbf{y},\mathbf{x}) \tag{1.3.19}$$

so P_i satisfies detailed balance with respect to Π.

Combining the kernels. Suppose that the U_i cover the set $\mathbb{N} \cap [1,\ell]$, so that each variable x_j is updated by at least one of the subkernels. A *subset update* is defined by a set of k finite sequences of members of $\mathbb{N} \cap [1,\ell]$, with the sequences denoted by $s^1, s^2, \ldots, s^k$, and the length of the sequence s^j denoted by r_j, and a corresponding sequence of probabilities $\zeta_1, \zeta_2, \ldots, \zeta_k \in [0,1]$ adding to 1. During each time step, we select one sequence s^j with probability ζ_j, and apply the subkernels in the order specified by the sequence s^j. Thus the full transition kernel becomes

$$P(\mathbf{x}, A) := \sum_{j=1}^{k} \zeta_j \left(P_{s_1^j} P_{s_2^j} \cdots P_{s_{r_j}^j} \right)(\mathbf{x}, A) \tag{1.3.20}$$

using the kernel product defined in (1.2.2). In the continuous-state case, the off-diagonal density becomes

$$p(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{k} \zeta_j \int_\Omega \cdots \int_\Omega p_{s_1^j}\left(\mathbf{x}, \mathbf{z}^{(1)}\right) p_{s_{r_j-1}^j}\left(\mathbf{z}^{(r_j-2)}, \mathbf{y}\right) d\mathbf{z}^{(1)} \prod_{i=2}^{r_j-2} p_{s_i^j}\left(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}\right) d\mathbf{z}^{(i)} \tag{1.3.21}$$

while in the discrete-state case we have

$$p(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{k} \zeta_j \sum_{\mathbf{z}^{(r_j-2)}\in\Omega} \cdots \sum_{\mathbf{z}^{(1)}\in\Omega} p_{s_1^j}\left(\mathbf{x},\mathbf{z}^{(1)}\right) p_{s_{r_j-1}^j}\left(\mathbf{z}^{(r_j-2)},\mathbf{y}\right) \prod_{i=2}^{r_j-2} p_{s_i^j}\left(\mathbf{z}^{(i-1)},\mathbf{z}^{(i)}\right) \tag{1.3.22}$$

There are some important classes of subset updates that Tierney analyzes in [9]. If we have $k = 1$ then the subset update is a *fixed cycle*, which applies each subkernel in a fixed order, given by s^1, during each time step. If $r_j = 1$ for all j, then this is a *mixture*, where for each time step, one subkernel is selected at random and applied. The kernel in (1.3.9) is a mixture. If the sequences s^j include every possible permutation of $\mathbb{N} \cap [1, \ell]$, then we have a *mixed cycle*.

If the U_i are disjoint, that is, if each variable x_j is affected by exactly one of the P_i transition kernels, then we say the transition kernel P is a *partition update*. A partition update with $\ell = N$ must have a singleton for each update set; in this case we have a single-variable update (sometimes called variable-at-a-time Metropolis). We call a partition update that is not a single-variable update a *joint update*.

One special case is where $k = 1$ and $r_0 = 1$, that is, there is only one subkernel, and it is applied during each time step. This is a mixture, a mixed cycle, and a fixed cycle all at once. This approach is called the *global update*, as the lone subkernel is able to update every variable in the system. It has fairly nice properties on paper, but its efficiency in practice is extremely poor, and as we shall see in Part (3.2), it quickly becomes poorer as the system size increases. The global update is assumed for GRG's proof of 0.234 as the optimal acceptance ratio as $N \to \infty$, giving this result limited relevance to realistic Monte Carlo simulations.

1.3.2. **Use with systems of independent degrees of freedom.**

1.3.2.1. *Mutually independent variables.* For certain combinations of Π and the q_i proposal densities, the cycle type of subset update may produce a Markov chain $\mathbf{X}$ which does not satisfy detailed balance. Since the target distribution Π is usually dictated by the

problem under consideration, care must be taken in choosing the proposal distributions q_i, and detailed balance must be examined for each case separately. However, there is one simple case where detailed balance of the Metropolis algorithm is guaranteed: a system of independent variables with a partition update kernel, that is, with disjoint update sets.

Theorem 15. *Suppose the target density π is the joint density of mutually independent variables, that is,*

$$\pi(\mathbf{x}) = \prod_{j=1}^{N} \pi_j\left(x_j\right) \tag{1.3.23}$$

and P is a partition update kernel with $k = 1$ (that is, a fixed cycle). Suppose the proposal densities have the form

$$q_i(\mathbf{z}, \mathbf{y}) = \begin{cases} 0, & z_k \neq y_k \text{ for some } k \notin U_i \\ \prod_{j \in U_i} q_i^{(j)}\left(z_j, y_j\right), & \text{otherwise} \end{cases} \tag{1.3.24}$$

where each $q_i^{(j)}$ is symmetric. Then detailed balance as in (1.2.30) holds.

Proof. Since the update sets are disjoint, there is exactly one sequence of intermediate states $\mathbf{z}^{(i)}$ which leads from $\mathbf{x}$ to $\mathbf{y}$.

$$z_j^{(i)} = \begin{cases} x_j, & j \in U_r \text{ and } r \geq i \\ y_j, & j \in U_r \text{ and } r < i \end{cases} \tag{1.3.25}$$

for each $1 \leq j \leq N$. So we have

$$p(\mathbf{x}, \mathbf{y}) = \prod_{i=0}^{\ell-1} p_i\left(\mathbf{z}^{(i)}, \mathbf{z}^{(i+1)}\right) \tag{1.3.26}$$

By the symmetry of the q_i, we also have

$$\begin{aligned}
\pi(\mathbf{x})p(\mathbf{x},\mathbf{y}) &= \pi\left(\mathbf{z}^{(0)}\right) p_1\left(\mathbf{z}^{(0)},\mathbf{z}^{(1)}\right) \prod_{i=1}^{\ell-1} p_i\left(\mathbf{z}^{(i)},\mathbf{z}^{(i+1)}\right) \\
&= \pi\left(\mathbf{z}^{(1)}\right) p_1\left(\mathbf{z}^{(1)},\mathbf{z}^{(0)}\right) \prod_{i=1}^{\ell-1} p_i\left(\mathbf{z}^{(i)},\mathbf{z}^{(i+1)}\right) \\
&= p_1\left(\mathbf{z}^{(1)},\mathbf{z}^{(0)}\right) \pi\left(\mathbf{z}^{(2)}\right) p_2\left(\mathbf{z}^{(2)},\mathbf{z}^{(1)}\right) \prod_{i=2}^{\ell-1} p_i\left(\mathbf{z}^{(i)},\mathbf{z}^{(i+1)}\right) \\
&\vdots \\
&= \left[\prod_{i=0}^{\ell-1} p_i\left(\mathbf{z}^{(i+1)},\mathbf{z}^{(i)}\right)\right] \pi\left(\mathbf{z}^{(\ell)}\right) \\
&= \pi(\mathbf{y})p(\mathbf{y},\mathbf{x}) \qquad (1.3.27)
\end{aligned}$$

□

1.3.2.2. *Alternatives to Monte Carlo.* The proof of GRG is valid only for systems of independent degrees of freedom[3], whose target distribution is of the form

$$\pi(\mathbf{x}) = \prod_{j=1}^{N} \pi_j\left(x_j\right) \qquad (1.3.28)$$

Thus for any random variable

$$X(\mathbf{x}) = \sum_{i=1}^{N} X_i\left(x_i\right) \qquad (1.3.29)$$

we have

$$\mathbb{E}[X] = \sum_{i=1}^{N} \mathbb{E}\left[X_i\right] \qquad (1.3.30)$$

so that the expectation for sum random variables can be computed by evaluating an expectation over each degree of freedom separately. This consists merely of evaluating N

[3]Indeed, the hypotheses of GRG's proof are even more restrictive, requiring not merely independent but also identically distributed variables.

single-variable integrals. Thus an MCMC simulation would produce much poorer accuracy than standard numerical integration algorithms using the same amount of computer time. However, the existence of an independent means of computing expectations makes these systems useful for testing new MCMC algorithms and an important teaching tool. In fact, in this paper we will use just such a technique in Section 3 to verify the time-shifted cross-correlation development in joint updating simulations.

1.3.2.3. *Behavior of updating scheme.* In the special case of a system of independent degrees of freedom as in (1.3.28), the probability of the jth degree of freedom being changed during a joint updating sweep is

$$P_j = \min\left\{1, \prod_{k\in U_i} \frac{\pi_k\left(x_k + y_k^{(i)}\right)}{\pi_k\left(x_k\right)}\right\} \tag{1.3.31}$$

where U_i is the updating set containing j. In this case, the presence of the product over k means we do not have formal independence of update probability of one degree of freedom from the other $m-1$ degrees of freedom in its updating set. The acceptance of each degree of freedom's proposed move depends on the values of, and proposed moves for, every degree of freedom in the updating set. However, this does not affect the detailed balance or ergodicity of the system, so it remains a reversible Markov process, and thus the sampled values of x_k and x_j (for $k \neq j$) at the same time are statistically independent, as the process at equilibrium faithfully replicates the target distribution of independent degrees of freedom at each time step.

In the case of the system of independent degrees of freedom, the single updating scheme causes (1.3.31) to reduce to

$$P_j = \min\left\{1, \frac{\pi_j\left(x_j + y_j^{(j)}\right)}{\pi_j\left(x_j\right)}\right\} \tag{1.3.32}$$

so the single updating of each degree of freedom is formally independent from the current values of and proposed moves for the other degrees of freedom.

1.3.2.4. *Updating scheme used by GRG.* The paper of GRG assumes a global update kernel. In their two papers in 1996 and 1997, Gelman et al. considered the optimization of an MCMC algorithm with this type of transition kernel as the total number of degrees of freedom $N \to \infty$.[1, 2] While in certain types of simulations, there may be significant advantages to grouping certain variables into update sets, joint update kernels with update sets larger than 4 variables are rarely used.[5] The most common reason one would prefer a joint update for a particular system is the existence of significant correlations among certain groups of variables under the target distribution Π, so the GRG hypothesis of independence is far from being satisfied for these distributions. In addition, we will see in Section 3.2.2.1 (and can be seen in GRG's data tables despite their lack of comment) that for large m, a joint update kernel (of which global update is a special case with $m = N$) requires more sweeps by a factor of m to produce the same level of accuracy as a single-variable update kernel on the same system. A single-variable update sweep consumes at most twice the computer time of a joint updating sweep.[4]Thus the use of a global update kernel as $N \to \infty$, even with the optimal acceptance ratio favored by GRG, is inferior in terms of computer time when compared to single-variable update kernels.

1.3.3. **Applicability to mixed distributions.** In the foregoing, we discussed approaches to guaranteeing convergence for discrete distributions and for continuous distributions; we did not discuss the situation where Π is a mixed distribution, that is, where $\Pi(\{x\}) = 0$ for some, but not all, $x \in \mathsf{supp}(\Pi)$. The Metropolis algorithm cannot directly deal with such cases since it requires a consistent density function π with respect to a single basis measure. However, in applications of MCMC we are usually interested only in computing statistics for the distribution Π, so all we really need to do is find a way to compute expectation values.

[4]In practice, random numbers for deciding acceptance are drawn from a uniform distribution, which is much more efficient to sample from than the Gaussian distribution. We must sample the proposed moves from the Gaussian distribution if GRG's proof is to be valid for our simulation. So the time penalty factor for a single updating sweep is actually closer to $1 + \gamma$, where γ is the ratio of time for uniform sampling to time for Gaussian sampling. This is usually far less than the theoretical maximum factor of 2 in a GRG-friendly simulation.

If we can write Π as

$$\Pi = \sum_{i=1}^{n} \Pi_i \tag{1.3.33}$$

with none of the Π_i being the zero measure, then we can normalize these measures to produce n probability measures

$$\Pi_i^*(A) := \frac{\Pi_i(A)}{\Pi_i(\Omega)} \tag{1.3.34}$$

for all $A \in \mathcal{A}$. If each distribution Π_i^* is separately treatable with the Metropolis algorithm, then we have

$$\mathbb{E}_\Pi[X] = \int_\Omega X(\omega)\Pi(d\omega) \tag{1.3.35}$$

$$= \sum_{i=1}^{n} \Pi_i(\Omega) \int_\Omega X(\omega)\Pi_i^*(d\omega) \tag{1.3.36}$$

$$= \sum_{i=1}^{n} \Pi_i(\Omega)\mathbb{E}_{\Pi_i^*}[X] \tag{1.3.37}$$

For a mixed distribution Π, we can define the set of points with positive probability as

$$\Omega_+ := \{x \in \Omega : \Pi(\{x\}) > 0\} \tag{1.3.38}$$

and

$$\Pi_{\mathsf{disc}}(A) := \begin{cases} \sum_{x \in A \cap \Omega_+} \Pi(\{x\}), & A \cap \Omega_+ \neq \emptyset \\ 0, & A \cap \Omega_+ = \emptyset \end{cases} \tag{1.3.39}$$

which is a discrete distribution forming the probability space $(\Omega_+, \mathfrak{P}(\Omega_+), \Pi_{\mathsf{disc}})$. Then

$$\Pi_{\mathsf{cont}} := \Pi - \Pi_{\mathsf{disc}} \tag{1.3.40}$$

is necessarily a continuous distribution.

In most applications where mixed distributions occur, we have $\Pi_{\mathsf{cont}} \ll \mathcal{L}^N$. If this is not the case, another problem can arise at this point: Π_{cont} may not be absolutely continuous with respect to a σ-finite measure on $\mathcal{A}$, and thus will not be treatable using Metropolis directly. By the Lebesgue Decomposition Theorem, we have

$$\Pi_{\mathsf{cont}} = \Pi_{\mathsf{AC}} + \Pi_{\mathsf{SC}} \tag{1.3.41}$$

with $\Pi_{\mathsf{AC}} \ll \mathcal{L}^N$ and $\Pi_{\mathsf{SC}} \perp \mathcal{L}^N$. It may also be possible to further decompose the "singular continuous" part of the measure as

$$\Pi_{\mathsf{SC}} = \Pi_{\mathsf{FD}} + \Pi_{\mathsf{LD}} \tag{1.3.42}$$

where Π_{LD} (lower dimension) is a measure whose support has dimension $N-1$ or lower. Such measures would be singular with respect to $\mathcal{L}^N$ but may be the push-forward of a measure which is absolutely continuous with respect to $\mathcal{L}^n$ for some $n \in \mathbb{N} \cap [1, N)$; this part of the probability is treatable using Metropolis and Gibbs sampling on the lower-dimensional space. If the remaining part of the measure, Π_{FD} (full dimension), is nonzero, using the Metropolis algorithm would require finding a more obscure sequence of σ-finite measures singular with respect to $\mathcal{L}^N$ along with methods for sampling from these distributions. The Cantor measure discussed in [14] would be one example of a σ-finite measure on $\mathbb{R}$ which is singular with respect to $\mathcal{L}^1$, and it may be sampled from using the transformation method [15], as its cumulative distribution function is fairly simple.[5] This is not typical for singular-continuous distributions, however.

1.3.4. **Criterion for minimizing error.** The most important criterion for efficient MCMC simulations was derived by Müller-Krubmhaar and Binder[3] in 1973. They proved that if the purpose of the simulation was to estimate a variable X over T discrete MCMC steps,

[5]There would be a potential for bias if using the transformation method, since values of the Cantor function in $[0,1]$ with terminating binary representations (which includes all IEEE-standard representations) always correspond to members of the Cantor set that are at the endpoint of one of the excluded intervals. Such numbers form a countable subset of the uncountable Cantor set, and are thus atypical.

the simulation should be designed to minimize the correlation time τ_Y, which is defined through the autocorrelation function

$$\phi_Y(s) := \frac{\mathsf{Cov}_t[Y(X(t)), Y(X(t+s))]}{\mathsf{Var}_t[Y(X(t))]} \tag{1.3.43}$$

The correlation time is then defined[6] as

$$\tau_Y := \frac{1}{2} + \sum_{s=1}^{\infty} \phi_Y(s) \tag{1.3.44}$$

Müller-Krubmhaar and Binder[3] showed that the variance of $\bar{Y}$, the estimate of Y from the MCMC simulation of T time steps, is related linearly to the ratio τ_Y/T for large T. We present a proof here as MKB's proof is geared toward a different purpose.

Theorem 16. (Binder, Muller-Krumbhaar) *Suppose $X(t)$ is a Markov chain with the distribution of $X(t)$ constant with respect to t. Let Y be a random variable on Ω whose mean is estimated by*

$$\bar{Y} := \frac{1}{T} \sum_{t=1}^{T} Y(X(t)) \tag{1.3.45}$$

for each particular simulation of $X(t)$ for T time steps. Then

$$\mathsf{Var}\left[\bar{Y}\right] = 2\left(\frac{\tau_Y}{T}\right) \mathsf{Var}_\Pi[Y] + O\left[T^{-2}\right] \tag{1.3.46}$$

where the variance of $\bar{Y}$ is taken over possible simulations.

[6] Muller-Krumbhaar and Binder do not add $1/2$ to the sum; we do so here to simplify the variance relation (1.3.52).

Proof. We have

$$\begin{aligned}
\mathsf{Var}\left[\bar{Y}\right] &= \frac{1}{T^2}\mathsf{Var}\left[\sum_{t=1}^{T} Y(X(t))\right] && (1.3.47)\\
&= \frac{1}{T^2}\left[\sum_{t=1}^{T}\left(\mathsf{Var}[Y(X(t))] + 2\sum_{s=1}^{T-t}\mathsf{Cov}[Y(X(t)), Y(X(t+s))]\right)\right] && (1.3.48)\\
&= \frac{\mathsf{Var}[Y]}{T}\left(\frac{1}{T}\sum_{t=1}^{T}\left(1 + 2\sum_{s=1}^{T-t}\frac{\mathsf{Cov}[Y(X(0)), Y(X(s))]}{\mathsf{Var}[Y]}\right)\right) && (1.3.49)\\
&= \frac{\mathsf{Var}[Y]}{T}\left(1 + \frac{2}{T}\sum_{s=1}^{T}\sum_{t=1}^{T-s}\phi_Y(s)\right) && (1.3.50)\\
&= \frac{\mathsf{Var}[Y]}{T}\left(1 + 2\sum_{s=1}^{T}\left(1 - \frac{s}{T}\right)\phi_Y(s)\right) && (1.3.51)
\end{aligned}$$

since there are T terms in the sum over t in (1.3.49) and $T - s$ terms in the sum over t in (1.3.50). Since

$$2\tau_Y = 1 + 2\sum_{s=1}^{\infty}\phi_Y(s) \tag{1.3.52}$$

we have for $T > 1$

$$\begin{aligned}
\left|\mathsf{Var}\left[\bar{Y}\right] - 2\left(\frac{\tau_Y}{T}\right)\mathsf{Var}[Y]\right| &= \frac{2\mathsf{Var}[Y]}{T^2}\left[\sum_{s=1}^{T} s\phi_Y(s) + \sum_{s=T+1}^{\infty}\phi_Y(s)\right] && (1.3.53)\\
&= \frac{2\mathsf{Var}[Y]}{T^2}\left[\sum_{s=1}^{\infty} s\phi_Y(s) + \sum_{s=T+1}^{\infty}(1-s)\phi_Y(s)\right] && (1.3.54)\\
&\leq \frac{2\mathsf{Var}[Y]}{T^2}\sum_{s=1}^{\infty} s\phi_Y(s) && (1.3.55)
\end{aligned}$$

□

Thus the correlation time is linearly related to the number of time steps necessary for the simulation estimate's variance to equal the variance of the random variable itself. Since the correlation time depends on the individual behavior of the autocorrelation function for

the variable Y, a single simulation is likely to have different correlation times for different measured variables.

While they do not cite the work of Muller-Krubmhaar and Binder, GRG do define the reciprocal of this correlation time as the "efficiency" of a simulation in their numerical experiments.

Part 2. **Balance theorems for Sequential and Checkerboard Update**

2.1. Sequential update for the Ising Model

Working with the discrete case only, that is

$$\Omega = \bigcup_{n=1}^{\infty} \{x_i\} \tag{2.1.1}$$

Manousthakis and Deem [17] prove that there is some invariant measure Π for a transition kernel P if we assume a few properties for the transition matrix T defined by

$$T_{ij} := p\left(x_i, x_j\right) \tag{2.1.2}$$

In order to have a valid chain, we must have the transition probabilities from a given state to all the other states add up to 1. Thus we require that the transition matrix T be stochastic, that is,

$$\sum_j T_{ij} = 1 \tag{2.1.3}$$

for all states $i = 1, \ldots, |\Omega|$. In [17] they additionally require that $T^\top$ be stochastic, that is,

$$\sum_i T_{ij} = 1 \tag{2.1.4}$$

for all states $j = 1, \ldots, |\Omega|$. The latter condition demands that the transition probabilities into a given state from all the other states add up to 1. This is hardly an "obvious feature" of an update scheme, and indeed is not satisfied for some simple and common schemes, including the sequential update which the authors of [17] seek to rehabilitate.

2.1.1. **Sequential Metropolis.** Consider the one-dimensional Ising model with $N \geq 3$ spins and periodic boundary conditions, simulated with a sequential single-variable update Metropolis algorithm. Here we have $\Omega = \{-1, 1\}^N$, for 2^N possible states, and

$$\pi(\mathbf{x}) := Z^{-1} e^{-\beta E(\mathbf{x})} \tag{2.1.5}$$

where Z is a normalizing constant, β is a parameter related to temperature, and the energy is defined as

$$E(\mathbf{x}) := -J\left(x_N x_1 + \sum_{i=2}^{N} x_{i-1}x_i\right) \tag{2.1.6}$$

for some constant $J \in \mathbb{R}$. The transition kernel density for the update of the kth spin will be

$$p_k(\mathbf{x},\mathbf{y}) = \begin{cases} 0, & x_\ell \neq y_\ell \text{ for some } \ell \neq k \\ 1, & E(\mathbf{y}) \leq E(\mathbf{x}) \\ \exp\left[\beta(E(\mathbf{x}) - E(\mathbf{y}))\right], & E(\mathbf{y}) > E(\mathbf{x}) \end{cases} \tag{2.1.7}$$

2.1.1.1. *Failure of detailed balance.* Let $\beta J = 1/4$. Consider the state $\mathbf{x}$ where $x_j = 1$ for all $1 \leq j \leq N$ and the state $\mathbf{y}$ where $y_1 = 1$ and $y_j = -1$ for $2 \leq j \leq N$. Note that $\pi(\mathbf{x}) = Z^{-1}e$ and $\pi(\mathbf{y}) = Z^{-1}$.

The transition from $\mathbf{x}$ to $\mathbf{y}$ requires a hold in the first position, a conditional flip in the second, and automatic flips in all the rest. Thus $p(\mathbf{x},\mathbf{y}) = e^{-1}\left(1 - e^{-1}\right)$. However, the transition from $\mathbf{y}$ to $\mathbf{x}$ is impossible, since the first position in $\mathbf{y}$ will always be flipped. So $p(\mathbf{y},\mathbf{x}) = 0$ and thus detailed balance as in (1.2.20) cannot be satisfied.

2.1.1.2. *Failure of stochastic transpose.* Let $k = 2, ..., N-1$ be given. As in M-D we will denote the transition matrix for updating the kth position by $B^{(k)}$. Suppose a state with up spins in the $k-1, k, k+1$ positions is labeled j. The kth kernel will automatically move to state j from any state that has a down spin in the kth position and either (a) two up spins or (b) an up spin and a down spin as its two neighbors. There are 2^{N-3} states satisfying (a), and 2^{N-2} states satisfying (b), so the transition matrix $B^{(k)}$ must have

$$\sum_{i=1}^{2^N} B_{ij}^{(k)} \geq 2^{N-3} + 2^{N-2} > 1 \tag{2.1.8}$$

which means $\left(B^{(k)}\right)^\top$ is not stochastic.

Nor does the full transition matrix

$$T = \prod_{k=1}^{N} B^{(k)} \tag{2.1.9}$$

have a stochastic transpose. For example, let $\beta J = 1/4$. For each integer $0 \leq \ell \leq N$, let j_ℓ be the label of the state with the first $N - \ell$ spins down and the remaining ℓ spins up. We consider the transition probabilities from the various j_ℓ to j_N which has all spins up. We obtain $T_{j_0 j_N} = e^{-1}$ and $T_{j_\ell j_N} = \left(1 - e^{-1}\right)^\ell$, so that

$$\begin{aligned} \sum_{i=1}^{2^N} T_{i j_N} &\geq e^{-1} + \sum_{\ell=1}^{N} \left(1 - e^{-1}\right)^\ell && (2.1.10) \\ &= e^{-1} + (e-1)\left[1 - \left(1 - e^{-1}\right)^N\right] && (2.1.11) \end{aligned}$$

where we have used the geometric sum formula. This expression is strictly increasing with N, and has a value of 1 when $N = 1$, so the full kernel's transition matrix does not have a stochastic transpose for $N > 1$.

2.1.2. **Sequential Gibbs sampler.** For a one-dimensional Ising model with $N \geq 3$ spins, the single-variable Gibbs sampler has transition kernel

$$p_k(\mathbf{x}, \mathbf{y}) = \begin{cases} 0, & x_\ell \neq y_\ell \text{ for some } \ell \neq k \\ \frac{\pi(\mathbf{y})}{\pi(\mathbf{x}) + \pi(\mathbf{x}')}, & \text{otherwise} \end{cases} \tag{2.1.12}$$

where $\mathbf{x}' := \mathbf{x} - 2x_k \mathbf{e}^{(k)}$ is the result of flipping the kth spin in $\mathbf{x}$. Let $1 < k < N$ be given. Suppose j is the label for the state $\mathbf{x}'$. Then

$$\sum_i B_{ij}^{(k)} = p_k(\mathbf{x}, \mathbf{x}') + p_k(\mathbf{x}', \mathbf{x}') = \frac{2\pi(\mathbf{x}')}{\pi(\mathbf{x}) + \pi(\mathbf{x}')} \tag{2.1.13}$$

and this will only be equal to 1 if $\pi(\mathbf{x}) = \pi(\mathbf{x}')$. This does indeed occur if the spins in positions $k-1$ and $k+1$ are opposite. However, if $(x_{k-1}, x_k, x_{k+1}) = (1, 1, 1)$, then we

have $\left(x'_{k-1}, x'_k, x'_{k+1}\right) = (1, -1, 1)$ and thus

$$\frac{\pi(\mathbf{x}')}{\pi(\mathbf{x})} = e^{-2\beta J} \tag{2.1.14}$$

and thus $B^{(k)}$ can only have a stochastic transpose if $\beta J = 0$, which only happens if temperature is infinite or there is no interaction between spins. Thus, the result of M-D is irrelevant when dealing with this common probability distribution and the sequential application of the two most common samplers for it.

2.2. THE CHECKERBOARD DECOMPOSITION

A common substitute for the sequential update Metropolis algorithm is the checkerboard decomposition update. In general, this decomposition produces m disjoint sets $C_j \subseteq \mathbb{Z} \cap [0, N)$, such that for any $j \in \mathbb{Z} \cap [0, m)$ and $a, b \in C_j$, we have x_a and x_b conditionally independent with respect to Π given the values of all variables x_c with $c \notin C_j$. A useful theorem for checkerboard updating is the following

Theorem 17. *Let $(P_i)_{i=0}^{k-1}$ be a sequence of subset kernels of the form (1.3.20), with each satisfying the detailed balance condition (2.3.1). Suppose that $C \subseteq \mathbb{Z} \cap [0, k)$ has the properties that the update sets U_{i_1} and U_{i_2} are disjoint for any distinct $i_1, i_2 \in C$, and variables x_j with $j \in \bigcup_{i \in C} U_i$ are conditionally independent given the values of all variables x_ℓ with $\ell \notin \bigcup_{i \in C} U_i$. Then any composition of the subset kernels P_i for $i \in C$, in which each subset kernel is applied once, satisfies detailed balance, regardless of the order in which the subset kernels are applied.*

Proof. Let $(i_m)_{m=0}^{|C|-1}$ be a permutation of C. Define a composition kernel P to be the kernel resulting from applying the subset updates P_i in the order specified by $(i_m)_{m=0}^{|C|-1}$. Given $\mathbf{x}, \mathbf{y} \in \Omega$, there is only one sequence of intermediate states $\left(\mathbf{z}^{(m)}\right)_{m=1}^{|C|-1}$ which leads from $\mathbf{x}$ to $\mathbf{y}$, and this is given by

$$\mathbf{z}_j^{(m)} = \begin{cases} x_j, & j \in U_{i_n} \text{ with } n > m \\ y_j, & \text{otherwise} \end{cases} \tag{2.2.1}$$

that is, $\mathbf{z}^{(m)}$ matches $\mathbf{x}$ in positions updated by a kernel after P_{i_m}, matches $\mathbf{y}$ in positions updated by P_{i_m} or a previous kernel, and matches both $\mathbf{x}$ and $\mathbf{y}$ in positions not updated by any of the kernels. Since the update sets are conditionally independent, the off-diagonal density (or mass if Π is discrete) is given by

$$p(\mathbf{x},\mathbf{y}) = p_{i_0}\left(\mathbf{x},\mathbf{z}^{(1)}\right) p_{i_{|C|-1}}\left(\mathbf{z}^{(|C|-1},\mathbf{y}\right) \prod_{m=1}^{|C|-1} p_{i_m}\left(\mathbf{z}^{(m)},\mathbf{z}^{(m+1)}\right) \tag{2.2.2}$$

By repeatedly applying (2.3.1) for each subset kernel P_{i_m}, we obtain

$$\begin{aligned}
\pi(\mathbf{x})p(\mathbf{x},\mathbf{y}) &= \pi(\mathbf{x})p_{i_0}\left(\mathbf{x},\mathbf{z}^{(1)}\right) p_{i_{|C|-1}}\left(\mathbf{z}^{(|C|-1)},\mathbf{y}\right) \prod_{m=1}^{|C|-2} p_{i_m}\left(\mathbf{z}^{(m)},\mathbf{z}^{(m+1)}\right) \\
&= p_{i_0}\left(\mathbf{z}^{(1)},\mathbf{x}\right) p_{i_{|C|-1}}\left(\mathbf{z}^{(|C|-1)},\mathbf{y}\right) \pi\left(\mathbf{z}^{(1)}\right) \prod_{m=1}^{|C|-2} p_{i_m}\left(\mathbf{z}^{(m)},\mathbf{z}^{(m+1)}\right) \\
&= p_{i_0}\left(\mathbf{z}^{(1)},\mathbf{x}\right) p_{i_{|C|-1}}\left(\mathbf{z}^{(|C|-1)},\mathbf{y}\right) p_{i_1}\left(\mathbf{z}^{(2)},\mathbf{z}^{(1)}\right) \pi\left(\mathbf{z}^{(2)}\right) \prod_{m=2}^{|C|-2} p_{i_m}\left(\mathbf{z}^{(m)},\mathbf{z}^{(m+1)}\right) \\
&\vdots \\
&= p_{i_0}\left(\mathbf{z}^{(1)},\mathbf{x}\right) \pi\left(\mathbf{z}^{(|C|-1)}\right) p_{i_{|C|-1}}\left(\mathbf{z}^{(|C|-1)},\mathbf{y}\right) \prod_{m=1}^{|C|-2} p_{i_m}\left(\mathbf{z}^{(m+1)},\mathbf{z}^{(m)}\right) \\
&= \pi(\mathbf{y})p_{i_0}\left(\mathbf{z}^{(1)},\mathbf{x}\right) p_{i_{|C|-1}}\left(\mathbf{y},\mathbf{z}^{(|C|-1)}\right) \pi\left(\mathbf{z}^{(1)}\right) \prod_{m=1}^{|C|-2} p_{i_m}\left(\mathbf{z}^{(m)},\mathbf{z}^{(m+1)}\right) \\
&= \pi(\mathbf{y})p(\mathbf{y},\mathbf{x})
\end{aligned} \tag{2.2.3}$$

□

Thus the updating of all variables in C_j, while holding the rest of the variables constant, satisfies detailed balance. However, we will see that a sequence of such updates of $C_0, C_1, \ldots, C_m$ need not satisfy detailed balance.

2.2.1. **Application to Ising model.** The "checkerboard" name arises from its application to the two-dimensional Ising model, where we have a state space of matrices

$$\Omega := \{-1, 1\}^{L \times L} \tag{2.2.4}$$

and Π defined by

$$\Pi(\mathbf{X}) = Z^{-1} \exp\left(-\beta E(\mathbf{X})\right) \tag{2.2.5}$$

with β and Z defined as in (2.1.5). The notation for the energy function is much neater if we define $[[n]]$ to be the positive remainder of n divided by L, that is, it satisfies $0 \leq [[n]] < L$ and $n = cL + [[n]]$ for some $c \in \mathbb{Z}$. Then the energy function is

$$\begin{aligned} E(\mathbf{X}) &:= -J \sum_{\langle (a,b),(c,d) \rangle} X_{ab} X_{cd} && (2.2.6) \\ &= -J \sum_{i,j} X_{ij} \left[X_{[[i+1]],j} + X_{i,[[j+1]]} \right] && (2.2.7) \end{aligned}$$

where $\langle (a,b),(c,d) \rangle$ holds if the positions (a,b) and (c,d) are "nearest neighbors" in the lattice, with periodic boundary conditions taken into account. That is, the 0th row and the $(L-1)$th row are considered adjacent, and likewise with the columns. If L is even, the checkerboard decomposition is accomplished by defining

$$\begin{aligned} C_0 &:= \{(i,j) | i + j \text{ is even}\} && (2.2.8) \\ C_1 &:= \{(i,j) | i + j \text{ is odd}\} && (2.2.9) \end{aligned}$$

The nearest neighbors of any position in C_0 are all in C_1, and vice versa. We think of C_0 as the red squares of the checkerboard and C_1 as the black squares (or vice versa). Thus

$$E(\mathbf{X}) = -J \sum_{(i,j)\in C_0} \left[X_{ij} \sum_{\langle (a,b),(i,j)\rangle} X_{ab} \right] \tag{2.2.10}$$

$$=: -J \sum_{(i,j)\in C_0} E_{(i,j)} \left(X_{ij}, X_{[[i+1]],j}, X_{[[i-1]],j}, X_{i,[[j+1]]}, X_{i,[[j-1]]} \right) \tag{2.2.11}$$

and

$$\Pi(\mathbf{X}) = Z^{-1} \prod_{(i,j)\in C_0} \exp\left[-\beta E_{(i,j)} \left(X_{ij}, X_{[[i+1]],j}, X_{[[i-1]],j}, X_{i,[[j+1]]}, X_{i,[[j-1]]} \right) \right] \tag{2.2.12}$$

so the X_{ij} with $(i,j) \in C_0$ are conditionally independent given all X_{ab} with $(a,b) \in C_1$, since the only non-given variable in each factor in the product is X_{ij}.

The checkerboard decomposition update is usually implemented as a fixed cycle, where we update all the positions in C_0 and then update all the positions in C_1. This gives the sole update sequence

$$s^0 := \{(0,0),(0,2),\ldots,(0,L-2),(1,1),(1,3),\ldots,(L-1,L-1),(0,1),(0,3),\ldots,(L-1,L-2)\} \tag{2.2.13}$$

This update scheme is often claimed to satisfy detailed balance, but this is not so. Let $L \times L$ matrices $\mathbf{X}$ and $\mathbf{Y}$ be defined by $X_{ij} = 1$ and $Y_{ij} = (-1)^{i+j}$ for all $i,j \in \mathbb{Z} \cap [0,L)$. Under the checkerboard update we have $P(\mathbf{X},\mathbf{Y}) = \left(1-e^{-8J}\right)^{L^2/2} e^{-4JL^2}$, which is produced by first leaving unchanged all the $L^2/2$ spins in C_0, and then flipping each of the $L^2/2$ spins in C_1. However, $P(\mathbf{Y},\mathbf{X}) = 0$, as the first spin $Y_{0,0}$ must flip from +1 to -1 since this leads to a more probable state, making it impossible to reach $\mathbf{X}$. Thus we cannot have detailed balance.

2.3. GUARANTEEING DETAILED BALANCE FOR SUBSET UPDATES

We now introduce a few methods by which an update scheme satisfying detailed balance can be constructed from subset kernels. Let the notation $j \rightleftarrows u$ indicate that s^j and s^u are reverse sequences; that is, $r_j = r_u$ and $s^j_i = s^u_{r_u - i}$ for $i \in \mathbb{Z} \cap [0, r_u - 1]$. Then we have the following.

Theorem 18. *Suppose P is a subset update kernel of the form (1.3.20) and each subset kernel P_i satisfies*

$$\pi(\mathbf{x}) p_i(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y}) p_i(\mathbf{y}, \mathbf{x}) \tag{2.3.1}$$

for all $\mathbf{x}, \mathbf{y} \in \Omega$. Suppose that for every sequence $s^j \in S$, there is $s^u \in S$ such that $j \rightleftarrows u$ and $\zeta_u = \zeta_j$. Then P satisfies detailed balance.

Proof. Suppose Π is a continuous distribution with $\Omega \subseteq \mathbb{R}^N$. Let $j \in \mathbb{Z} \cap [0, k)$ be given, and find $u \in \mathbb{Z} \cap [0, k)$ such that $j \rightleftarrows u$. Then

$$\pi(\mathbf{x})p_{s^j}(\mathbf{x},\mathbf{y}) = \pi(\mathbf{x})\int_\Omega \cdots \int_\Omega p_{s_1^j}\left(\mathbf{x},\mathbf{z}^{(1)}\right) p_{s_{r_j-1}^j}\left(\mathbf{z}^{(r_j-2)},\mathbf{y}\right) d\mathbf{z}^{(1)} \prod_{i=2}^{r_j-2} p_{s_i^j}\left(\mathbf{z}^{(i-1)},\mathbf{z}^{(i)}\right) d\mathbf{z}^{(i)} \tag{2.3.2}$$

$$= \int_\Omega \cdots \int_\Omega \pi\left(\mathbf{z}^{(1)}\right) p_{s_1^j}\left(\mathbf{z}^{(1)},\mathbf{x}\right) p_{s_{r_j-1}^j}\left(\mathbf{z}^{(r_j-2)},\mathbf{y}\right) d\mathbf{z}^{(1)} \prod_{i=2}^{r_j-2} p_{s_i^j}\left(\mathbf{z}^{(i-1)},\mathbf{z}^{(i)}\right) d\mathbf{z}^{(i)} \tag{2.3.3}$$

$$\vdots \tag{2.3.4}$$

$$= \int_\Omega \cdots \int_\Omega p_{s_1^j}\left(\mathbf{z}^{(1)},\mathbf{x}\right) \pi(\mathbf{y}) p_{s_{r_j-1}^j}\left(\mathbf{y},\mathbf{z}^{(r_j-2)}\right) d\mathbf{z}^{(1)} \prod_{i=2}^{r_j-2} p_{s_i^j}\left(\mathbf{z}^{(i)},\mathbf{z}^{(i-1)}\right) d\mathbf{z}^{(i)} \tag{2.3.5}$$

$$= \pi(\mathbf{y})\int_\Omega \cdots \int_\Omega p_{s_{r_u-1}^u}\left(\mathbf{z}^{(1)},\mathbf{x}\right) p_{s_1^u}\left(\mathbf{y},\mathbf{z}^{(r_u-2)}\right) d\mathbf{z}^{(1)} \prod_{i=2}^{r_u-2} p_{s_{r_u-i}^u}\left(\mathbf{z}^{(i-1)},\mathbf{z}^{(i)}\right) d\mathbf{z}^{(i)} \tag{2.3.6}$$

$$= \pi(\mathbf{y})p_{s^u}(\mathbf{y},\mathbf{x}) \tag{2.3.7}$$

Similarly we have

$$\pi(\mathbf{x})p_{s^u}(\mathbf{x},\mathbf{y}) = \pi(\mathbf{y})p_{s^j}(\mathbf{y},\mathbf{x}) \tag{2.3.8}$$

If $j \rightleftarrows j$ for all j, this means

$$\begin{aligned} \pi(\mathbf{x})p(\mathbf{x},\mathbf{y}) &= \pi(\mathbf{x})\sum_{j=0}^{k-1}\zeta_j p_{s^j}(\mathbf{x},\mathbf{y}) \\ &= \pi(\mathbf{y})\sum_{j=0}^{k-1}\zeta_j p_{s^j}(\mathbf{y},\mathbf{x}) \\ &= \pi(\mathbf{y})p(\mathbf{y},\mathbf{x}) \end{aligned} \tag{2.3.9}$$

which gives detailed balance. Otherwise, we have

$$\begin{aligned}
\pi(\mathbf{x})p(\mathbf{x},\mathbf{y}) &= \sum_{j=0}^{k-1} \zeta_j \pi(\mathbf{x}) p_{s^j}(\mathbf{x},\mathbf{y}) \\
&= \sum_{\substack{j<u \\ j \leftrightarrows u}} \zeta_j \pi(\mathbf{x}) \left[p_{s^j}(\mathbf{x},\mathbf{y}) + p_{s^u}(\mathbf{x},\mathbf{y})\right] + \sum_{j \rightleftarrows j} \zeta_j \pi(\mathbf{x}) p_{s^j}(\mathbf{x},\mathbf{y}) \\
&= \sum_{\substack{j<u \\ j \rightleftarrows u}} \zeta_j \pi(\mathbf{y}) \left[p_{s^u}(\mathbf{y},\mathbf{x}) + p_{s^j}(\mathbf{y},\mathbf{x})\right] + \sum_{j \rightleftarrows j} \zeta_j \pi(\mathbf{y}) p_{s^j}(\mathbf{y},\mathbf{x}) \\
&= \pi(\mathbf{y})p(\mathbf{y},\mathbf{x}) \qquad (2.3.10)
\end{aligned}$$

If Π is a discrete distribution, replace the integrals in (2.3.2) with sums over Ω and the rest of the proof follows. □

Corollary 19. *Let P be a subset update kernel of the form (1.3.20) with each subset kernel P_i satisfying (2.3.1). Then P satisfies detailed balance provided that at least one of the following conditions hold:*

(1) (mixture) $r_j = 1$ for each $j \in \mathbb{Z} \cap [0, \ell)$
(2) (mixed cycle) S is the set of all permutations of $\mathbb{Z} \cap [0, \ell)$ and $\zeta_j = 1/\ell!$ for all j.
(3) (palindrome cycle) $k = 1$ and $0 \rightleftarrows 0$

Proof. If condition 1 holds, each sequence s^j has length 1 and thus $j \rightleftarrows j$ for all j. If condition 3 holds, there is only one sequence s^j and $j \rightleftarrows j$ for that sequence. If condition 2 holds, note that the reverse sequence of a permutation of $\mathbb{Z} \cap [0, \ell)$ is just another permutation, so reverse sequences will both be in S with equal selection probability $1/\ell!$. Thus, in each case we can apply Proposition 5. □

Mixtures of subset kernels satisfying detailed balance have long been known to satisfy detailed balance, which is why mixtures are commonly used in practical MCMC simulations. The other two conditions have not been linked to balance until now. Mixed cycles are less common because they require more random numbers to be generated per time step.

Fixed cycles in general do not satisfy detailed balance. However, condition 3 shows that this can be remedied by appending the cycle steps in reverse order to produce a palindrome cycle. This workaround would appear to come at the cost of doubling the amount of variable-updating per sample, as the cycle now has length $2N$, but this is actually not so. Let a collection of subset update kernels $\{P_i\}_{i=0}^{\ell-1}$ be given. Suppose that we produce one fixed palindrome cycle with the sole updating sequence of length $2N$:

$$s^0 = \{0, 1, \ldots, N-2, N-1, N-1, N-2, \ldots, 1, 0\} \tag{2.3.11}$$

and another with the sole updating sequence of length $2N$:

$$\hat{s}^0 = \{N-1, N-2, \ldots, 1, 0, 0, 1, \ldots, N-2, N-1\} \tag{2.3.12}$$

Note that these two fixed cycle updates prescribe exactly the same sequence of variable updates, merely shifted by N time steps. The only difference is that in the first sequence we sample between updates of x_0, while in the second we sample between the updates of x_{N-1}. Since sampling from either sequence (after equilibration) will produce the same averages for all random variables, we can actually sample both between the x_0 updates and between the x_{N-1} updates to obtain more precise averages.

Note that this double-sampling trick will not work if we produce a fixed palindrome cycle with the following updating sequence of length $2N-1$:

$$s^0 = \{0, 1, \ldots, N-2, N-1, N-2, \ldots, 1, 0\} \tag{2.3.13}$$

though the resulting chain will still satisfy detailed balance as it is a palindrome cycle.

2.3.1. **Recovering detailed balance.** Using Corrolary 19 and Theorem 18, we can recover detailed balance for the checkerboard decomposition. This can be accomplished in several ways.

- Mixture: randomly select whether to update spins in C_0 or to update spins in C_1 and then sample.
- Mixed cycle: update both C_0 and C_1 during each time step and then sample, but randomly select (with equal probability) which update is done first.
- Palindrome cycle, length 3: update all spins in C_0, then all spins in C_1, then all spins in C_0, then sample.
- Palindrome cycle, length 4: update all spins in C_0, then all spins in C_1, then all spins in C_1 again and finally all spins in C_0 again. In this case we can use the double-sampling trick to sample after the first C_1 update as well as after the second C_0 update.

Note that interchanging C_0 and C_1 in any of the above methods will still preserve detailed balance.

2.4. Guaranteeing Total Balance for Subset Updates

If we only seek a fixed cycle update kernel P which satisfies total balance, the restrictions can be loosened somewhat. To simplify notation, we will define the "reverse" of a state $\mathbf{z} \in \Omega$. Let $f : \mathbb{N} \cap [1, N] \to \mathbb{N} \cap [1, N]$ be a bijection such that $f(j) \in U_{\ell - i}$ if $j \in U_i$. Then the reverse of $\mathbf{z}$ with respect to f and P is denoted $\bar{\mathbf{z}}$ and defined so that $\bar{z}_j := z_{f(j)}$.

Lemma 20. *Suppose $P = \prod_{i=1}^{\ell} P_i$ and each P_i satisfies detailed balance with respect to Π as in (2.3.1), the update sets U_i are of equal size, and*

$$p_i(\mathbf{x}, \mathbf{y}) = p_{\ell - i}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \tag{2.4.1}$$

for all $i \in \mathbb{N} \cap [1, \ell]$ and $\mathbf{x}, \mathbf{y} \in \Omega$. Then

$$\pi(\mathbf{x}) p(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y}) p(\bar{\mathbf{y}}, \bar{\mathbf{x}}) \tag{2.4.2}$$

Proof. From (2.3.6) in Theorem 18 we have that

$$\begin{aligned}
\pi(\mathbf{x})p(\mathbf{x},\mathbf{y}) &= \pi(\mathbf{y})\int_\Omega \cdots \int_\Omega p_1\left(\mathbf{z}^{(1)},\mathbf{x}\right)\left[\prod_{i=2}^{\ell-1} p_i\left(\mathbf{z}^{(i)},\mathbf{z}^{(i-1)}\right)d\mathbf{z}^{(i)}\right] p_{\ell-1}\left(\mathbf{y},\mathbf{z}^{(\ell-1)}\right)d\mathbf{z}^{(1)} \\
&= \pi(\mathbf{y})\int_\Omega \cdots \int_\Omega p_{\ell-1}\left(\overline{\mathbf{z}^{(1)}},\bar{\mathbf{x}}\right)\left[\prod_{i=2}^{\ell-1} p_{\ell-i}\left(\overline{\mathbf{z}^{(i)}},\overline{\mathbf{z}^{(i-1)}}\right)d\mathbf{z}^{(i)}\right] p_1\left(\bar{\mathbf{y}},\overline{\mathbf{z}^{(\ell-1)}}\right)d\mathbf{z}^{(1)} \\
&= \pi(\mathbf{y})p(\bar{\mathbf{y}},\bar{\mathbf{x}}) \qquad (2.4.3)
\end{aligned}$$

□

This leads to the following theorem.

Theorem 21. *Suppose* Π *is a discrete probability measure. Suppose* $P = \prod_{i=1}^{\ell} P_i$ *and each* P_i *satisfies detailed balance with respect to* Π *as in (2.3.1), the update sets* U_i *are of equal size, and*

$$p_i(\mathbf{x},\mathbf{y}) = p_{\ell-i}(\bar{\mathbf{x}},\bar{\mathbf{y}}) \qquad (2.4.4)$$

for all $i \in \mathbb{N} \cap [1,\ell]$ *and* $\mathbf{x},\mathbf{y} \in \Omega$*. Then* P *satisfies total balance with respect to* Π*.*

Proof. Suppose the target distribution is continuous. Let $\mathbf{x} \in \Omega$ be given. Aiming to satisfy the expression (1.2.23) of total balance, in view of Lemma 20 we have

$$\begin{aligned}
\sum_{\mathbf{y}\in\Omega}\left[\pi(\mathbf{y})p(\mathbf{y},\mathbf{x}) - \pi(\mathbf{x})p(\mathbf{x},\mathbf{y})\right] &= \sum_{y\in\Omega}\pi(\mathbf{x})\left[p(\bar{\mathbf{x}},\bar{\mathbf{y}}) - p(\mathbf{x},\mathbf{y})\right] && (2.4.5)\\
&= \pi(\mathbf{x})\left[\sum_{\mathbf{y}\in\Omega} p(\bar{\mathbf{x}},\bar{\mathbf{y}}) - \sum_{\mathbf{y}\in\Omega} p(\mathbf{x},\mathbf{y})\right] && (2.4.6)\\
&= \pi(\mathbf{x})\left[\sum_{\mathbf{y}\in\Omega} p(\bar{\mathbf{x}},\mathbf{y}) - \sum_{\mathbf{y}\in\Omega} p(\mathbf{x},\mathbf{y})\right] && (2.4.7)\\
&= 0 && (2.4.8)
\end{aligned}$$

since the "reverse" map $\mathbf{z} \mapsto \bar{\mathbf{z}}$ is a bijection, and thus both sums in the third line are equal to 1. □

The hypotheses of Theorem 21 are not difficult to satisfy for many target distributions Π. For example, suppose the target distribution is of the form

$$\pi(\mathbf{x}, \mathbf{y}) = \prod_{j=1}^{N} \pi_j(\mathbf{x}, \mathbf{y}) \tag{2.4.9}$$

where we define (with $[[n]]$ again representing the positive remainder of n when divided by the system size N)

$$\pi_j(\mathbf{x}) := \pi^* \left(x_j, x_{[[j+d_1]]}, x_{[[j-d_1]]}, \ldots, x_{[[j+d_s]]}, x_{[[j-d_s]]} \right) \tag{2.4.10}$$

for some fixed $d_1, \ldots, d_s$. Suppose further that

$$\pi^* \left(x_j, x_{[[j-d_1]]}, x_{[[j+d_1]]}, \ldots, x_{[[j-d_s]]}, x_{[[j+d_s]]} \right) = \pi^* \left(x_j, x_{[[j+d_1]]}, x_{[[j-d_1]]}, \ldots, x_{[[j+d_s]]}, x_{[[j-d_s]]} \right) \tag{2.4.11}$$

for all $\mathbf{x} \in \Omega$. Each of these statements is true for the Ising model in any number of dimensions, so long as we index the spin variables in a natural way, ie in spatial order for the one dimensional model, by rows then columns (or columns then rows) for the two dimensional model, etc. For the d-dimensional model, with L sites in each row, column, etc, we would then have $s = d$ and $d_m = L^{m-1}$ and

$$\pi_j(\mathbf{x}) = Z^{-1/N} \exp \left[-\frac{1}{2} J \beta x_j \sum_{m=0}^{d-1} \left(x_{[[j-L^m]]} + x_{[[j+L^m]]} \right) \right] \tag{2.4.12}$$

for all i. Then we can satisfy total balance with the sequential update consisting of single-variable Metropolis subkernels, so long as we have symmetric proposal densities $q_i(\mathbf{x}, \cdot)$ such that $q_{N-i}(\mathbf{x}, \mathbf{y}) = q_i(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ for all i. To see this, define the reverse of $\mathbf{x}$ in the natural way, that is $\bar{x}_j := x_{N-j}$. Then we have

$$\pi_{N-i}(\mathbf{x}) = \pi_i(\bar{\mathbf{x}}) \tag{2.4.13}$$

Thus we also have

$$\begin{aligned} p_{N-i}(\mathbf{x},\mathbf{y}) &= q_{N-i}(\mathbf{x},\mathbf{y}) \min\left\{1, \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\right\} \\ &= q_{N-i}(\mathbf{x},\mathbf{y}) \min\left\{1, \frac{\pi_{N-i}(\mathbf{y})}{\pi_{N-i}(\mathbf{x})} \prod_{m=1}^{s} \frac{\pi_{N-i+d_m}(\mathbf{y})\pi_{N-i-d_m}(\mathbf{y})}{\pi_{N-i+d_m}(\mathbf{x})\pi_{N-i-d_m}(\mathbf{x})}\right\} \\ &= q_i(\bar{\mathbf{x}},\bar{\mathbf{y}}) \min\left\{1, \frac{\pi_i(\bar{\mathbf{y}})}{\pi_i(\bar{\mathbf{x}})} \prod_{m=1}^{s} \frac{\pi_{i-d_m}(\bar{\mathbf{y}})\pi_{i+d_m}(\bar{\mathbf{y}})}{\pi_{i-d_m}(\bar{\mathbf{x}})\pi_{i+d_m}(\bar{\mathbf{x}})}\right\} \\ &= p_i(\bar{\mathbf{x}},\bar{\mathbf{y}}) \end{aligned} \tag{2.4.14}$$

so we can apply Theorem 21 in this case. The checkerboard decomposition for the Ising model will also have total balance. This is shown by defining the reverse of $\mathbf{x}$ to satisfy $\bar{x}_j = x_{j-1}$ for all j, that is, the "circular left shift".

2.4.1. **Efficiency of kernels having total balance.** We tested the efficiency of various techniques satisfying total balance on a two-dimensional Ising model with $L = 32$ rows and columns. From the results in Table 2.1 for estimating the expectation value of magnetization

$$M := \frac{1}{L^2} \sum_{i,j} \sigma_{ij} \tag{2.4.15}$$

we see that the best correlation times are produced by the mixed cycle of single-variable updates, where we update each site in a randomly-selected order during each time step. When computer time on this machine is taken into account, however, the shortest fixed-cycle updates (sequential and checkerboard), where the sites are simply updated in the same order during every time step, are more efficient. This is due to the need to generate $L^2 - 1$ extra random numbers to determine the order for a mixed cycle during each step.

kernel	Var $[\bar{M}]$	τ_M	τ_M * comp time (ns)
sitewise mixture	5.98e-4	49.29	1613
sitewise mixed cycle	4.69e-4	27.48	1755
sequential forward	5.82e-4	38.62	1261
sequential forward-backward	4.90e-4	31.67	2025
checkerboard mixture	3.84e-3	328.52	5561
checkerboard mixed cycle	1.61e-3	105.53	3438
checkerboard 0-1	9.20e-4	39.74	1289
checkerboard 0-1-0	1.75e-3	187.23	9016
checkerboard 0-1-1-0	2.80e-3	397.54	25347

TABLE 2.1. Performance of various kernels for estimating $\mathbb{E}[M]$ on two-dimensional Ising model with $L = 32$ and $T = 640000$ sampled time steps, with $\beta^{-1} = 2.32$ (that is, near critical temperature). The computer times are per time step in nanoseconds, with the simulation run on a 2010 model Mac Mini.

Part 3. **Effects of Joint Update Schemes**

3.1. TIME-SHIFTED CROSS-CORRELATIONS

In the midst of the analysis of GRG's joint updating scheme, we discovered a strange phenomenon that this algorithm produces. While sampling from a target distribution with independent variables, there are time-shifted cross-correlations between different variables at different times, even though the different variables at the same time are independent, as required by the reversibility of the Markov chain.

3.1.1. **Energy-based systems.** A particularly common type of probability density in applications is of the form

$$\pi(\mathbf{x}) = Z^{-1} \exp\left[-\beta E(\mathbf{x})\right] \tag{3.1.1}$$

This occurs frequently in thermodynamic simulations for fixed temperature, where $E(\mathbf{x})$ represents the energy of a system, β is a constant related to the temperature, and Z is a normalizing constant. If the energy function can be expressed as a sum of single-variable energies $E_j\left(x_j\right)$, the density becomes

$$\pi(\mathbf{x}) = \prod_{j=1}^{N} Z_j^{-1} e^{-\beta E_j(x_j)} \tag{3.1.2}$$

The system's probability distribution is then simply the joint distribution of N independent degrees of freedom.

While we do not intend to explore the physics of such systems in this paper, this template provides a simple method of generating probability distributions with varied properties. To make the degrees of freedom both independent and identically distributed, we need to also require that all the E_j are the same function. With these fairly strong restrictions in place, GRG's proof will hold.

3.1.2. **Measuring efficiency.**

3.1.2.1. *Local and global autocorrelations.* Another important fact about the analysis by GRG is that the proof treats each degree of freedom's evolution as a separate process, optimizing the speed of each degree of freedom's diffusion through its own sample space. However, they do not consider the possibility of cross-correlations between degrees of freedom at different times. For this reason, we consider not only efficiency measures for quantities determined by the entire system, but also efficiency measures for estimating quantities depending only on individual degrees of freedom.

Suppose the random variable $X(\mathbf{x})$ is of the form of (1.3.29), that is, the sum of separate values $X_j(x_j)$ for each degree of freedom in the system. In addition to the MKB correlation functions ϕ_X for the quantity X and ϕ_{X_j} for each quantity X_j, as defined in (1.3.43), we define a local autocorrelation function

$$\lambda_X(\Delta t) := \frac{1}{N}\sum_{j=1}^{N} \phi_{X_j}(\Delta t) = \frac{1}{N}\sum_{j=1}^{N} \frac{\mathsf{Cov}_t\left[X_j(t),\, X_j(t+\Delta t)\right]}{\mathsf{Var}_t\left[X_j\right]} \tag{3.1.3}$$

The function λ_X is the mean of the autocorrelation functions ϕ_{X_j}. If the degrees of freedom are independent and identically distributed, in a joint updating scheme, the true values of ϕ_{X_j} are all equal. Thus the measured local autocorrelation function is an estimator

of the autocorrelation function for each degree of freedom. Originally the autocorrelation ϕ_X, which we will term the global autocorrelation function, and the local autocorrelation function λ_X were considered separately due to the estimation of λ_X being more efficient, as taking the sum of the X_j tends to dampen the fluctuations we are trying to measure. This distinction turns out to have deeper significance.

3.1.2.2. *Relation to cross-correlations.* For $N = 1$, we immediately have $\phi_X = \phi_{X_1} = \lambda_X$. However, for $N > 1$, we may have $\phi_X \neq \lambda_X$. Indeed, we now show that the difference between these functions is related to the time-shifted cross-correlation function we will define, for $N > 1$, as

$$\rho_X(\Delta t) := \frac{2}{N(N-1)} \sum_{i<j} \frac{\mathsf{Cov}_t \left[X_i(t),\, X_j(t+\Delta t)\right]}{\sqrt{\mathsf{Var}_t \left[X_i\right] \mathsf{Var}_t \left[X_j\right]}} \tag{3.1.4}$$

which is simply the mean of correlation coefficients among pairs of distinct degrees of freedom time-shifted by Δt steps. Thus if $X_i(t)$ and $X_j(t+\Delta t)$ are independent for all t, we should have $\rho(\Delta t) = 0$.

Theorem 22. *Let $N \in \mathbb{N} \cap [2, \infty)$ be given and $\{X_i\}_{i=1}^N$ be a collection of N stochastic processes such that*

$$\mathsf{Var}\left[X_i\right] = \mathsf{Var}\left[X_j\right] \neq 0 \, \forall i, j \in \mathbb{N} \cap [1, N] \tag{3.1.5}$$

Then using the definitions above, we have

$$\phi_X(\Delta t) = \frac{\lambda_X(\Delta t) + (N-1)\rho_X(\Delta t)}{1 + (N-1)\rho_X(0)} \tag{3.1.6}$$

Proof. Since the processes have equal variance, we have

$$\lambda_X(\Delta t) = \frac{\sum_i \mathsf{Cov}\left[X_i(t),\, X_i(t+\Delta t)\right]}{N \cdot \mathsf{Var}\left[X_1\right]} \tag{3.1.7}$$

$$\rho_X(\Delta t) = \frac{2\sum_{i<j} \mathsf{Cov}\left[X_i(t),\, X_j(t+\Delta t)\right]}{N(N-1) \cdot \mathsf{Var}\left[X_1\right]} \tag{3.1.8}$$

Thus we can derive the desired relation among these functions:

$$\begin{aligned}
\phi_X(\Delta t) &= \frac{\mathsf{Cov}\left[\sum_i X_i(t),\, \sum_i X_i(t+\Delta t)\right]}{\mathsf{Var}\left[\sum_i X_i\right]} && (3.1.9)\\
&= \frac{\sum_i \mathsf{Cov}\left[X_i(t),\, X_i(t+\Delta t)\right] + 2\sum_{i<j} \mathsf{Cov}\left[X_i(t),\, X_j(t+\Delta t)\right]}{N \cdot \mathsf{Var}\left[X_1\right] + 2\sum_{i<j} \mathsf{Cov}\left[X_i,\, X_j\right]} && (3.1.10)\\
&= \frac{N\lambda_X(\Delta t)\mathsf{Var}\left[X_1\right] + N(N-1)\rho_X(\Delta t)\mathsf{Var}\left[X_1\right]}{N \cdot \mathsf{Var}\left[X_1\right] + N(N-1)\rho_X(0)\mathsf{Var}\left[X_1\right]} && (3.1.11)\\
&= \frac{\lambda_X(\Delta t) + (N-1)\rho_X(\Delta t)}{1 + (N-1)\rho_X(0)} && (3.1.12)
\end{aligned}$$

□

Corollary 23. *Under the conditions of Theorem 22, if* $N > 1$ *and the quantities* $X_i(t)$ *and* $X_j(t)$ *are independent for any* t *and* $i \neq j$*, then*

$$\rho_X(\Delta t) = \frac{\phi_X(\Delta t) - \lambda_X(\Delta t)}{N-1} \tag{3.1.13}$$

Proof. This follows immediately from (3.1.6) by substituting $\rho_X(0) = 0$. □

Under the independence hypothesis of Corollary 23, any difference between the global and local autocorrelation functions indicates the presence of a time-shifted cross-correlation between the degrees of freedom.

3.1.3. **Computing time-shifted correlation coefficient.** Consider a bivariate Gaussian random variable with covariance matrix $\frac{1}{2}\mathbf{I}$. Its density function is

$$\pi(\mathbf{x}) = Z^{-1} \exp\left(-|\mathbf{x}|^2\right) \tag{3.1.14}$$

This is equivalent to an isotropic simple harmonic oscillator in physics, which is the simplest of a common class of physical simulations. Indeed, this is a case of the energy-based system given by (3.1.2) with $c = 1$ and energy random variables $E_j\left(x_j\right) = x_j^2$ for

$j \in \{1,2\}$ and $E(\mathbf{x}) = |\mathbf{x}|^2 = E_1(x_1) + E_2(x_2)$. Suppose that we perform a global update Metropolis simulation with $N = 2$ and proposed moves drawn from a bivariate Gaussian with covariance matrix $s^2\mathbf{I}$. The proposal density is

$$q(\mathbf{x}, \mathbf{x}+\mathbf{y}) = \frac{1}{2\pi s^2} \exp\left(-\frac{|\mathbf{y}|^2}{2s^2}\right) \tag{3.1.15}$$

The energy cross-correlation function with a time difference $\Delta t = 1$ becomes

$$\begin{aligned}
\rho_E(1) &= \mathsf{Cov}_t\left[E_1(x_1,t),\, E_2(x_2,t+1)\right] \\
&= \frac{1}{2\pi^2 s^2} \iint_{\mathbb{R}^2} \iint_{\mathbb{R}^2} x_1^2 \left[x_2^2 + (2x_2+y_2)\, y_2 a(\mathbf{x}, \mathbf{x}+\mathbf{y})\right] \exp\left(-|\mathbf{x}|^2 - \frac{|\mathbf{y}|^2}{2s^2}\right) d\mathbf{y}\, d\mathbf{x} - \mathbb{E}_t[E]^2 \\
&= \frac{1}{2\pi^2 s^2} \iint_{\mathbb{R}^2} x_1^2 \exp\left(-|\mathbf{x}|^2\right) \iint_{\mathbb{R}^2} (2x_2+y_2)\, y_2 a(\mathbf{x}, \mathbf{x}+\mathbf{y}) \exp\left(-\frac{|\mathbf{y}|^2}{2s^2}\right) d\mathbf{y}\, d\mathbf{x}
\end{aligned} \tag{3.1.16}$$

where $a(\mathbf{x}, \mathbf{z})$ is the acceptance probability for an already-proposed move from $\mathbf{x}$ to $\mathbf{z}$ as given in (1.3.3).

We evaluated this integral for s ranging from 0.1 to 3, using Gauss-Hermite quadrature for the $\mathbf{x}$ integrals and Gauss-Legendre quadrature over the square $[x_1 - 5s,\, x_1 + 5s] \times [x_2 - 5s,\, x_2 + 5s]$ for the $\mathbf{y}$ integrals. The correlation coefficient obtained by this classical, deterministic approximation is compared with the results recorded by our MCMC simulation in Table 3.1 and summarized in Figure 3.1a. There is excellent agreement between these independent methods of evaluating the cross-correlation, and both show that $\rho_E(1) \neq 0$, indicating a cross-correlation between the energy of oscillator 1 and that of oscillator 2 one time step later, despite the energies of the oscillators being independent variables in the target distribution.

3.1.4. **Cause of time-shifted cross-correlation.** This effect seems counterintuitive at first, since the Markov process theory guarantees that the oscillator energies are independent at a given time once equilibrium is reached. It is also intuitive to presume that two random variables that are positively correlated to a third must themselves be positively correlated.

sampling std dev	acceptance ratio	quadrature $\rho_E(1)$	Monte Carlo $\rho_E(1)$	Monte Carlo std error $\rho_E(1)$	difference of $\rho_E(1)$ estimates
0.8	0.506	0.0475	0.0443	0.00175	0.0032
1.0	0.422	0.0585	0.0608	0.00276	-0.0023
1.2	0.353	0.0643	0.0644	0.000474	-0.0001
1.4	0.295	0.0659	0.0681	0.00176	-0.0022
1.6	0.251	0.0646	0.0643	0.00300	0.0003

TABLE 3.1. Time-shifted correlation coefficient $\rho_E(1)$ for simple harmonic oscillator energies, as determined by quadrature and MC simulation.

A positive correlation between a degree of freedom and its future values (the well-known autocorrelation effect) coupled with a positive correlation between a degree of freedom and the future values of another degree of freedom (this unexpected cross-correlation effect) would then seem to force a positive correlation between the two degrees of freedom at the same time in the future. However, there are widely-known counterexamples to this presumption; it is even possible to have a negative correlation between two variables both positively correlated to a third variable.[7]

Indeed, there is a sensible explanation for this phenomenon in the case of joint updating. Consider the case with two simple harmonic oscillators and a random walk Metropolis joint update kernel, with a symmetric proposal distribution, such as the Gaussian with mean 0 or a uniform distribution on $[-r, r]$. If both oscillators have energy 0, and the proposed moves have length δ_1 and δ_2, then the expectation of the post-update energy of the second oscillator is

$$\begin{aligned} \mathbb{E}\left[E_2(t+1)|E_1(t)=0\right] &= \delta_2^2 e^{-\Delta E_1 - \Delta E_2} \\ &= \delta_2^2 e^{-\delta_1^2 - \delta_2^2} \end{aligned}$$

On the other hand, if oscillator 1 has displacement $\varepsilon > 0$ while oscillator 2 is at displacement 0, and $\delta_1 \leq \varepsilon$, then the expectation of E_2 after the update will be

$$\begin{aligned}\mathbb{E}\left[E_2(t+1)|E_1(t) = \varepsilon^2\right] &= \frac{\delta_2^2}{2}\left(e^{-(\varepsilon-\delta_1)^2+\varepsilon^2-\delta_2^2} + e^{-(\varepsilon+\delta_1)^2+\varepsilon^2-\delta_2^2}\right) \\ &= \delta_2^2 e^{-\delta_1^2-\delta_2^2}\cosh(2\varepsilon\delta_1)\end{aligned}$$

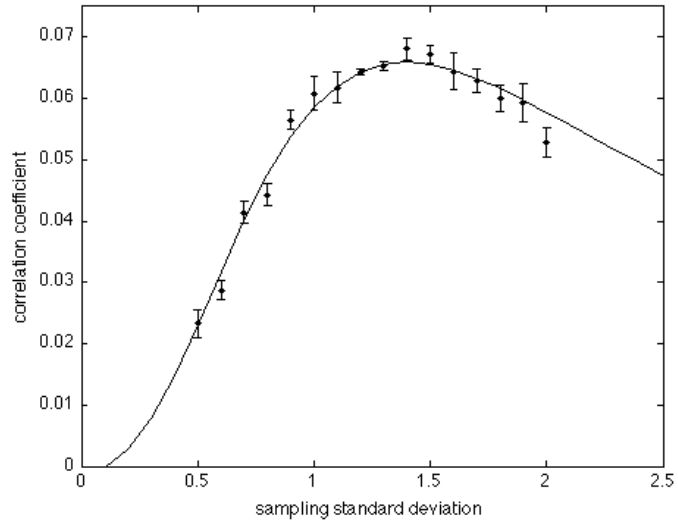
since the update to oscillator 1 is equally likely to propose a move to $\varepsilon - \delta_1$ or $\varepsilon + \delta_1$. The ratio of these expectations is

$$\frac{\mathbb{E}\left[E_2(t+1)|E_1(t) = \varepsilon^2\right]}{\mathbb{E}\left[E_2(t+1)|E_1(t) = 0\right]} = \cosh(2\varepsilon\delta_1)$$

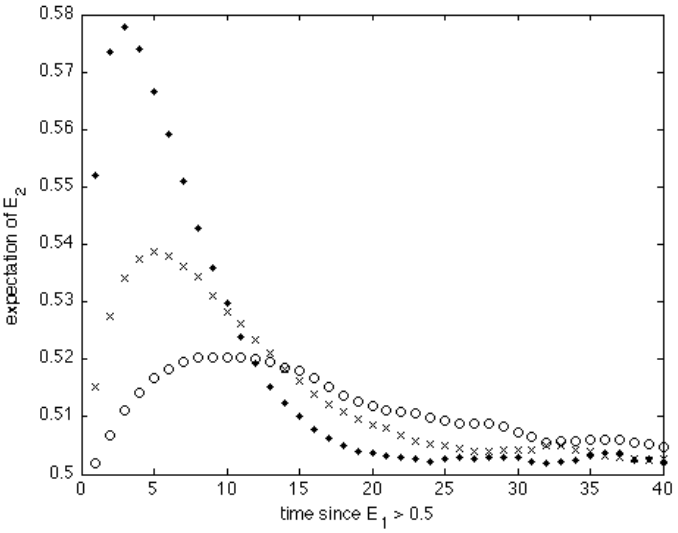
For $\varepsilon > 0$ this ratio will be larger than 1, so one oscillator being in a high-energy state makes a move by the other oscillator into a high-energy state more likely, despite the fact that the oscillators are supposed to be independent under the target distribution Π.

To test this explanation, we ran a simulation of a system with $N = 2$ for the simple harmonic oscillator system with joint updating, and measured the conditional expectation of the energy of the second oscillator given that the first oscillator had energy above its mean value. This produced a peak of about 0.58, as seen in Table 3.2 and Figure 3.1b. This is significantly higher than 0.50, the expectation of the oscillator energy. The effect on a single oscillator diminishes as N increases, but from Corollary 23 and the approximately linear increase of both global and local energy correlation times for the simple harmonic oscillator system using global updating, it appears that the effect is constant for $N > 1$ when the entire system is taken into account rather than a single oscillator.

This effect does not produce cross-correlations for quantities probabilistically symmetric with respect to their mean; that is, quantities X with $\mu = \mathbb{E}[X]$, for which $\mathsf{Pr}[\mu - b < X < \mu - a] = \mathsf{Pr}[\mu + a < X < \mu + b]$ for all $0 \leq a < b$. In that case a move to a lower probability region is equally likely to cause an increase or decrease of the same magnitude in the symmetric quantity, and thus will fail to bias the quantity. The numerical



(A)



(B)

FIGURE 3.1. In (a) we show the correlation coefficient between the energy of different oscillators at a time difference of 1, for a system of two independent simple harmonic oscillators updated using the global update kernel. The curve is the result of numerically integrating the exact expression, while the error bars indicate the results from compiling statistics from an actual MCMC simulation. In (b) we show the conditional expectation of the energy of oscillator 2 given that oscillator 1 had high energy at some fixed number of time steps in the past. Closed circles represent $N = 2$, crosses represent $N = 4$, and open circles represent $N = 8$.

Δt	conditional mean of E_2	standard error
1	0.5521	0.0027
2	0.5736	0.0030
3	0.5777	0.0031
4	0.5739	0.0029
6	0.5593	0.0030
8	0.5429	0.0028
10	0.5298	0.0028
12	0.5193	0.0027
16	0.5077	0.0024
20	0.5033	0.0027

TABLE 3.2. Means of oscillator 2 energy Δt time steps after oscillator 1 had energy above 0.5 in the simple harmonic oscillator system with $N = 2$.

experiments of Gelman et al. in 1996 treated only Gaussian random variables, which are probabilistically symmetric, so they could not have observed this cross-correlation effect.

3.2. EFFICIENCY OF SIMULATIONS USING JOINT UPDATE SCHEME

The global update scheme assumed by GRG's proof of 0.234 as the optimal acceptance ratio for systems of independent variables as $N \to \infty$ is extremely inefficient compared to single-variable update scheme applied to the same systems. Indeed, the autocorrelation times for important random variables of such systems increase by a factor of N when using the global update, while the same autocorrelation times remain constant when using a single-variable update.[7] In the following analysis, note that a single-variable update for any value of N has the same efficiency behavior as the global update with $N = 1$.

3.2.1. **Optimization results before GRG.** In the early 1990's, Bouzida, et al.[4, 5, 6] investigated conditions for optimization of MCMC step sizes and proposed adaptive algorithms for determining them from simulation data. They found that for an MCMC simulation of a one-dimensional simple harmonic oscillator using a uniform distribution for the trial moves, the optimal acceptance ratio at which τ_E is minimized is very nearly 0.5. This confirmed a rule of thumb that had long be used in MCMC simulations and brought it out of the realm of "folklore," as it was characterized five years later.[1] Other potentials were found to have other optimal acceptance ratios, although they were close to one-half for most single-well potentials.

As we will show below, the most efficient simulations of symmetric potentials are performed with one-dimensional moves; however, there are still important reasons for using two-, three-, and even four-dimensional moves in certain applications. In particular, models of biological molecules are characterized by highly inhomogeneous local structure. Each particle has a unique environment, which can also be highly anisotropic. It is the anisotropic local potentials that make one-dimensional MCMC moves substantially less efficient than well-optimized two- and three-dimensional moves that reflect the local

[7] For systems of independent variables, which are the only systems GRG's proof applies to.

anisotropy. In considering angular moves along the backbone of a protein, even four-dimensional moves can be needed. The Acceptance Ratio Method (ARM) and Dynamically Optimized Monte Carlo (DOMC) were developed to implement efficient MCMC simulations for such problems.[4, 5, 6]

For MCMC simulations of simple harmonic oscillators in two and three dimensions, the optimal acceptance ratio was found to decrease,[4, 5] anticipating the trend reported by Gelman and coworkers.[1, 2]

3.2.2. **Numerical results.** We use essentially the same definition of efficiency as GRG used in [1], the correlation time defined in (1.3.44). Of course, computing the exact value of this quantity for a particular simulation, even if we accept an estimate $\hat{\phi}_Y(s)$ of the true autocorrelation function $\phi_Y(s)$ for each $s \in \mathbb{N}$, would require running the simulation for infinite time steps. Even computing the finite sum $\sum_{s=1}^{T} \hat{\phi}_Y(s)$ for a simulation of T time steps would be impractical, as the determination of the $\hat{\phi}_Y(s)$ would require, for each $1 \leq t \leq T$, computing the product $Y(t)Y(s)$ for $1 \leq s \leq T - t$ during the tth equilibrium time step, increasing the complexity by a factor of T, which is typically on the order of 10^4 or 10^5. Thus we rely on two approximation techniques, which require only determining $\hat{\phi}_Y(s)$ for $1 \leq s \leq T_* \ll T$. The first is the truncated correlation time

$$\tau_Y^{\mathsf{trunc}} := \frac{1}{2} + \sum_{s=1}^{\min\{T_-, T_*\}} \hat{\phi}_Y(s) \tag{3.2.1}$$

where

$$T_- := \min\left\{s \in \mathbb{N} : \hat{\phi}_Y(s) \leq 0\right\} \tag{3.2.2}$$

Note that while the true autocorrelation function $\phi_Y(s) \geq 0$ for all $s \in \mathbb{N}_0$, the simulation estimate $\hat{\phi}_Y(s)$ may be negative if $\phi_Y(s)$ is near zero. Thus τ_Y^{trunc} estimates the correlation time by ignoring the statistical noise that dominates estimates of small $\phi_Y(s)$. This is, of course, a biased estimator, biased low, and only sensible to use if $T_- \leq T_*$. Thus we only use it for cases where the correlation time is low.

Another useful estimate, which complements the truncated correlation time fairly well, is the fitted correlation time. It uses the fact that $\phi_Y(s)$ approaches a decaying exponential function as s grows large [18]. Thus we define

$$\tau_Y^{\mathsf{fit}} := \frac{1}{2} + \sum_{s=1}^{T_*} \hat{\phi}_Y(s) + \int_{T_*}^{\infty} Ce^{-\gamma s} ds \tag{3.2.3}$$

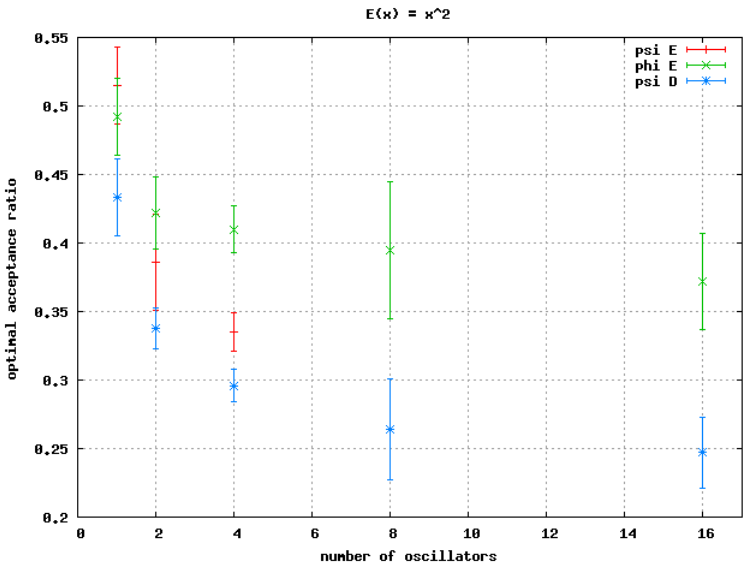
$$\approx \frac{1}{2} + \sum_{s=1}^{T_*} \hat{\phi}_Y(s) + \frac{\hat{\phi}_Y(T_*)}{\gamma} \tag{3.2.4}$$

The exponent γ is estimated using regression for $s \lesssim T_*$ or some other technique. This estimate will be more accurate when T_- defined in (3.2.2) is high, giving a larger range of s to use for regression. Thus, we use fitted correlation time for cases where correlation time is high.

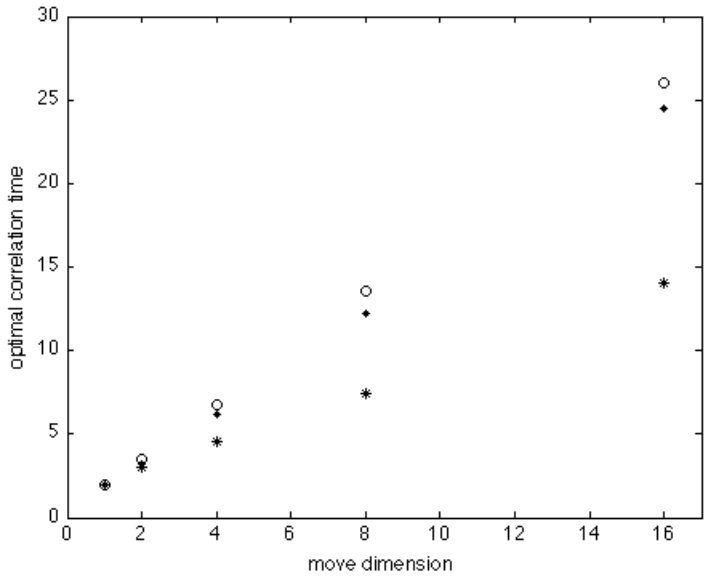
3.2.2.1. *Simple harmonic oscillators.* We now turn to evaluating the efficiency of joint and single updating simulations of a system of independent and identical simple harmonic oscillators (or a multivariate Gaussian) treated in Section 3, except now in an arbitrary number of dimensions N. The probability density for the displacement vector $\mathbf{x} \in \mathbb{R}^N$ of the oscillators now becomes

$$\pi_N^{\mathsf{sho}}(\mathbf{x}) = Z^{-1} \exp\left(-|\mathbf{x}|^2\right) \tag{3.2.5}$$

Results obtained for simulations with global update kernels are summarized in Table 3.3. The single updating scheme occurs for $N = 1$, while the joint updating scheme is present for $N > 1$. The correlation time increases roughly linearly with N for larger N. This can also be found in Gelman et al. [1] in Table 1.1, where the inverse quantity is termed "efficiency" and decreases as N^{-1}. In the case of optimizing both local correlation times, the optimal acceptance ratio seems consistent with convergence to a value near 0.234, while the optimal acceptance ratio for estimating energy as a global quantity exhibits slower convergence.



(A)



(B)

FIGURE 3.1. Optimal (a) acceptance ratio and (b) correlation time for energy and displacement for a system of N (move dimension) simple harmonic oscillators simulated using random walk Metropolis with the global update. GRG predict that optimal acceptance ratio for the global update approaches 0.234 as $N \to \infty$, which appears correct, but the correlation times increase roughly linearly with N, indicating global update is a poor choice of kernel. In (b), open circles represent global correlation time for energy, closed circles represent local correlation time for energy, and asterisks represent correlation time for displacement (local and global are identical).

N	acc ratio	τ_x (local)	acc ratio	τ_E (global)	acc ratio	τ_E (local)
1	0.431	1.7	0.495	1.8	0.495	1.8
2	0.338	3.5	0.425	3.6	0.387	2.7
4	0.295	5.9	0.407	6.1	0.336	4.6
8	0.261	12.5	0.395	13.8	0.275	7.5
16	0.249	24.6	0.373	25.5	0.258	13.7

TABLE 3.3. Optimal acceptance ratio and correlation time for displacement and energy in simple harmonic oscillators system. Global and local correlation times are roughly equal for displacement. The single-variable update for this system is equivalent to global update with $N = 1$.

3.2.2.2. *Symmetric anharmonic oscillators.* While it was not tested by Gelman et al., another interesting system with practical applications consists of independent, identical symmetric anharmonic oscillators with single-variable energy $E_j(x_j) = x_j^4 - x_j^2$ with $\mathbf{x} \in \mathbb{R}^N$ again representing displacement. This is called a double-well potential due to the shape of the graph of E_j, with local minima at $x_j = \pm\frac{\sqrt{2}}{2}$ and a local maximum at $x_j = 0$. The density function is of the form (3.1.2), specifically:

$$\pi_N^{\mathsf{dw}}(\mathbf{x}) = Z_N^{-1} \exp\left(-\sum_{j=1}^{N}\left(x_j^4 - x_j^2\right)\right) \tag{3.2.6}$$

exhibits a double-peak with the same relevant values of x_j. If the proposed move distribution favors very small moves, it is difficult for an oscillator to move from the area of one local minimum of E to the other, as the small size of the moves would make the displacement near 0 and are likely to be rejected due to the relatively high energy there. If large moves are likely to be proposed, this problem is averted.

Here we focused on relaxation time rather than the correlation time. We measured how long it takes the system to move from the local energy minimum where $x_j = 1/\sqrt{2}$ for each oscillator j, to a state where the mean displacement among the N oscillators reaches $\sqrt{2}/4$, half the initial value. The expectation of mean displacement at equilibrium is 0, so this is an indicator of how quickly oscillators are moved from one energy minimum to the other. In Table 3.4, we see that this decay time of the mean displacement increases roughly linearly with N when the simulation uses global updating. This indicates that

N	time to halve mean displacement
4	32
16	248
64	2119
256	7920
1024	29412

TABLE 3.4. Time to halve mean displacement for symmetric anharmonic oscillators system with all oscillators initialized to displacement $1/\sqrt{2}$, using global update kernel. For each N, the acceptance ratio was the optimal value for measuring displacement. While the equivalent of the single-variable update (global update with $N = 1$) was not measured, it is clear that the displacement-halving time is smaller for smaller values of N.

larger numbers of degrees of freedom make it more difficult to move from one well to the other, and thus require longer simulations to accurately sample the entire distribution.

3.2.2.3. *Asymmetric anharmonic oscillators.* To illustrate that the displacement also can be affected by spurious cross-correlations in the global update kernel, we considered a slightly different anharmonic potential $E(x) = 2\left(4x^4 - x^3 - 3x^2\right)$. This is also a double-well potential with local maximum at $x = 0$, with E locally minimized at x near 0.713 and -0.526. Its density function is

$$\pi_N^{\mathsf{asym}}(\mathbf{x}) = Z_N^{-1} \exp\left(-2\sum_{j=1}^{N}\left(4x^4 - x^3 - 3x^2\right)\right) \tag{3.2.7}$$

In this system we observed a difference between the global and local correlation times for displacement as well as energy. In Part 3 we explained that this happens because E is no longer symmetric about $x = 0$. Data for the two types of correlation time for displacement are given in Table 3.5.

3.2.3. **Effect of proposed move distribution.**

3.2.3.1. *Direct effect on efficiency.* In all the foregoing, we used a Gaussian as the proposal distribution, since that is the distribution that Gelman et al. worked with [2, 1]. However it is possible, and often desirable, to use other distributions for proposed moves. Many factors affect this decision, the foremost being the shape of the probability landscape of

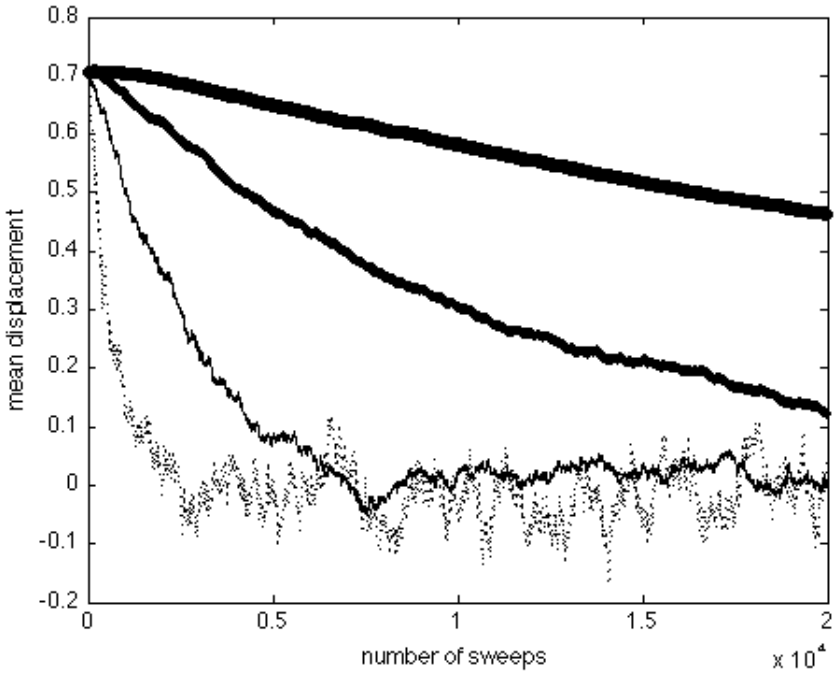


FIGURE 3.2. Time evolution of mean displacement for first 20,000 time steps for global update Metropolis simulation of symmetric anharmonic oscillators, with $N = 16, 64, 256, 1024$ and all oscillators initialized to displacement $\sqrt{2}/2$. Thicker lines indicate larger N. In equilibrium this system has an expected value of displacement equal to 0, indicating that equilibrium is much slower to develop for larger N. A single-variable update would behave as if $N = 1$ regardless of how many oscillators were present in the system, so equilibrium is reached much more quickly for a single-variable update.

N	acc ratio	τ_x (global)	acc ratio	τ_D (local)
1	0.482	1.75	0.467	1.49
2	0.350	3.71	0.350	3.58
4	0.251	10.9	0.251	10.4
8	0.210	30.4	0.190	29.3
16	0.195	79.6	0.195	77.7
32	0.222	187	0.205	183
64	0.225	398	0.225	396

TABLE 3.5. Displacement correlation time for asymmetric anharmonic oscillators system with $\alpha = 2$. The single-variable update exhibits the same efficiency behavior for any N as the global update does with $N = 1$.

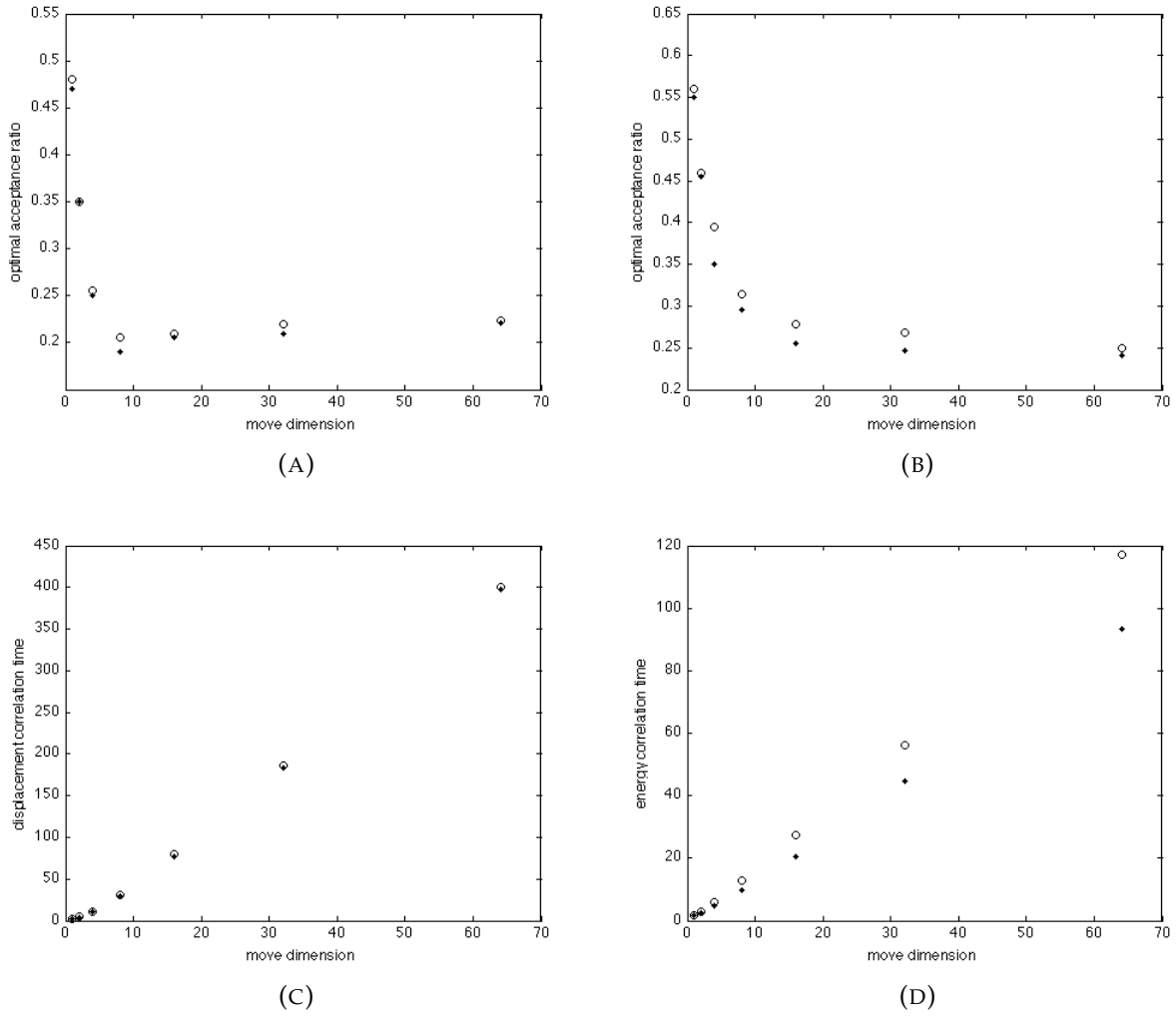


FIGURE 3.3. Optimal acceptance ratios for (a) displacement and (b) energy, and optimal correlation times for (c) displacement and (d) energy in the system of asymmetric anharmonic oscillators, using the global update as move dimension N increases. Open circles represent global correlation times and closed circles represent local correlation times (and their associated optimal acceptance ratios). GRG predict that optimal acceptance ratio for the global update approaches 0.234 as $N \to \infty$, and these data agree with that prediction, however the correlation time of the global update increases linearly with N making it an unattractive choice of kernel.

the problem. The most common alternative distributions are uniform in some region of space, be it a hypercube or a hypersphere (which may be uniform in volume or uniform in radius). Each random number sampled from the multivariate Gaussian distribution

requires the generation of a uniform random number and evaluation of three slow functions, so there is a computer time penalty in choosing Gaussian proposed moves over uniform. Also, our experiments summarized in Figure 3.4 indicate that sampling from a Gaussian produces longer correlation times than sampling from a symmetric uniform distribution even when the same updating scheme is used.

3.2.3.2. *Effect on length of proposed moves.* Maintaining the "optimal" acceptance ratio for the joint updating scheme with Gaussian proposed moves requires that the sampling standard deviation s_m for m-dimensional moves must decrease proportionally to $\sqrt{m}$. This actually leads to proposed moves being sampled almost exclusively from the surface of a hypersphere of constant radius as m grows large, causing the RMS length of each dimension's proposed moves to decrease proportionally to $\sqrt{m}$.

Theorem 24. *Suppose that for each $m \in \mathbb{N}$, Q_m is an m-dimensional Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $s_m^2\mathbf{I}$ for some $s_m > 0$. Let r_m be a random variable representing the length of a proposed move drawn from Q_m. Then*

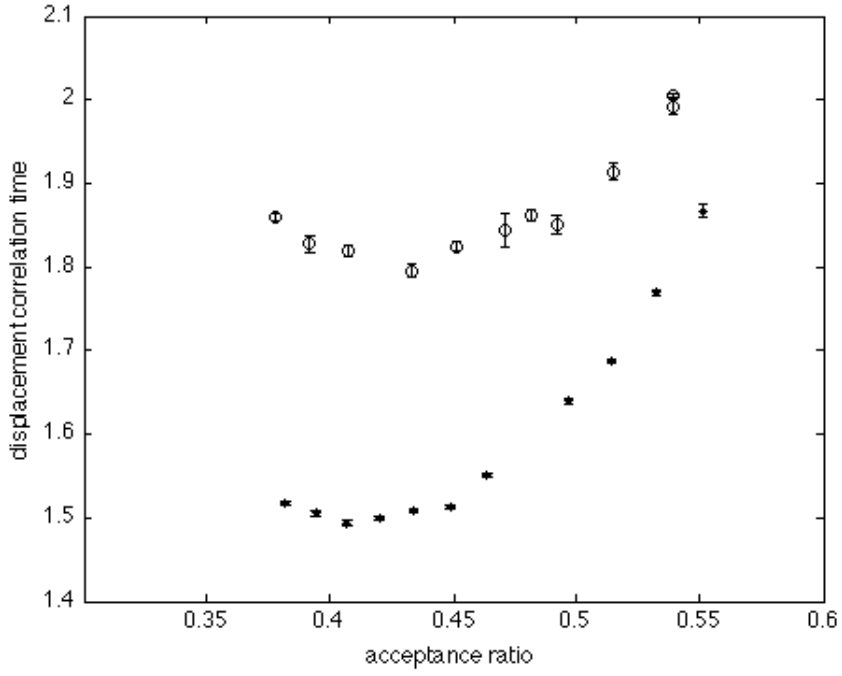
$$\mathbb{E}\left[r_m^2\right] = ms_m^2 \tag{3.2.8}$$

and for the limit of the relative width of the distribution of r_m we have

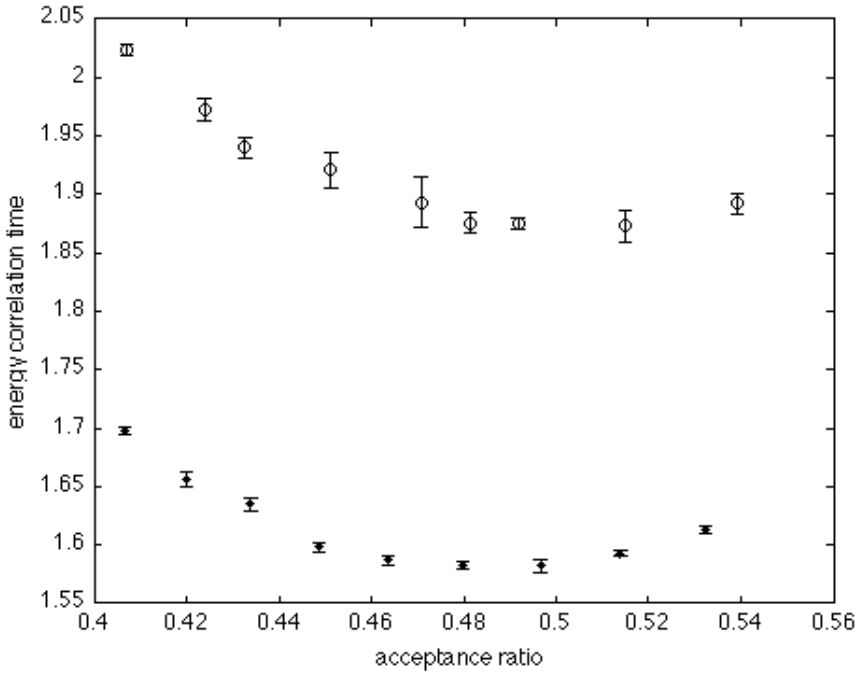
$$\lim_{m\to\infty} \frac{\mathsf{Var}\left[r_m\right]^{1/2}}{\mathbb{E}\left[r_m\right]} = 0 \tag{3.2.9}$$

Proof. For m-dimensional moves, the probability density of the proposed move length r generated by the multivariate Gaussian sampler with covariance matrix $s_m^2\mathbf{I}$ is given by

$$f_m(r) = Z_m^{-1} r^{m-1} \exp\left(-\frac{r^2}{2s_m^2}\right) \tag{3.2.10}$$



(A)



(B)

FIGURE 3.4. Correlation times for (a) displacement and (b) energy in simulating a single simple harmonic oscillator. The open circles represent the results for Gaussian proposal distributions, while the closed circles represent those for uniform proposal distributions.

with the normalization

$$\begin{aligned} Z_m &= \int_0^\infty r^{m-1} \exp\left(-\frac{r^2}{2s_m^2}\right) dr \\ &= \Gamma\left(\frac{m}{2}\right) 2^{\frac{m}{2}-1} s_m^m \end{aligned} \tag{3.2.11}$$

We then see that

$$\begin{aligned} \mathbb{E}\left[r^2\right] &= Z_m^{-1} \int_0^\infty r^{m+1} \exp\left(-\frac{r^2}{2s_m^2}\right) dr \\ &= m s_m^2 \end{aligned} \tag{3.2.12}$$

and

$$\mathbb{E}[r] = Z^{-1} \int_0^\infty r^m \exp\left(-\frac{r^2}{2s_m^2}\right) dr \tag{3.2.13}$$

$$= \left[\frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)}\right] s_m \sqrt{2} \tag{3.2.14}$$

Using Stirling's formula we obtain

$$\begin{aligned} \lim_{m\to\infty} \mathbb{E}[r] &= \lim_{m\to\infty} \left[\frac{\sqrt{\pi(m-1)}\left(\frac{m-1}{2}\right)^{\frac{m-1}{2}} e^{-\frac{m-1}{2}}}{\sqrt{\pi(m-2)}\left(\frac{m-2}{2}\right)^{\frac{m-2}{2}} e^{-\frac{m-2}{2}}} s_m \sqrt{2}\right] \\ &= \lim_{m\to\infty} \left[\left(1+\frac{1}{m-2}\right)^{\frac{m-2}{2}} e^{-\frac{1}{2}} \sqrt{\frac{m-1}{m-2}} s_m \sqrt{m-1}\right] \end{aligned} \tag{3.2.15}$$

So we have for the limit of the relative width of the distribution

$$\begin{aligned}\lim_{m\to\infty}\frac{\mathrm{Var}[r]^{1/2}}{\mathbb{E}[r]} &= \left[\lim_{m\to\infty}\frac{\mathbb{E}\left[r^2\right]}{\mathbb{E}[r]^2}-1\right]^{1/2}\\ &= \left[\lim_{m\to\infty}\left(\left(1+\frac{1}{m-2}\right)^{-(m-2)}e\left(1-\frac{1}{(m-1)^2}\right)\right)-1\right]^{1/2}\\ &= 0 \qquad (3.2.16)\end{aligned}$$

□

Since the sampling distribution is spherically symmetric, we can conclude that for large m, the multivariate Gaussian distribution proposes moves almost exclusively from the surface of a hypersphere of fixed radius R_m for each m.[8]

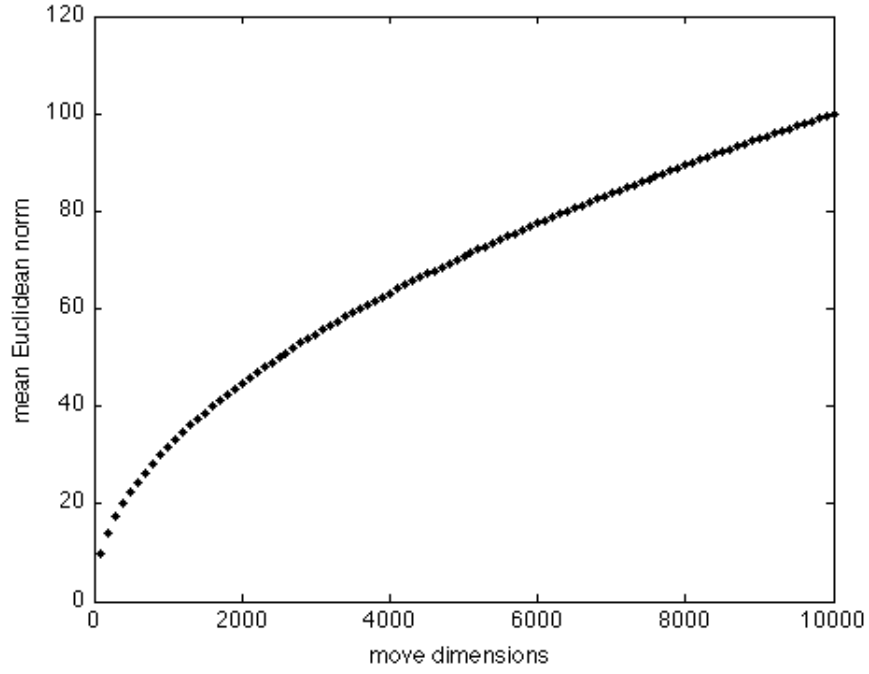
In addition, our experiments confirm the results of Roberts et al. that to maintain the acceptance ratio in the neighborhood of 0.234, we must set $s_m \approx Cm^{-1/2}$ for a constant C, so in view of (3.2.12) we have

$$\lim_{m\to\infty}\mathbb{E}\left[r_m^2\right] = C \qquad (3.2.17)$$

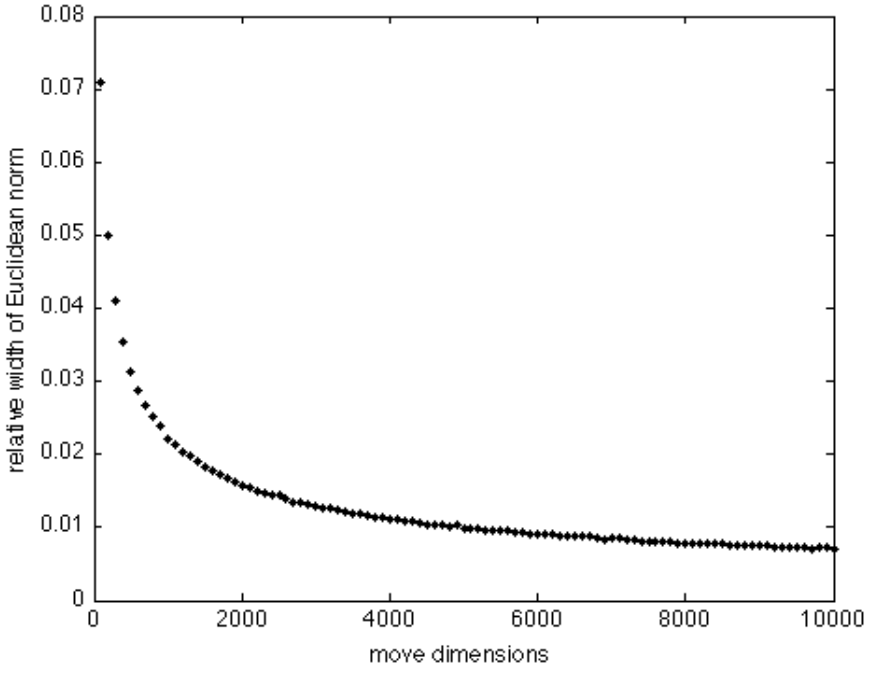
This is confirmed by our results summarized in Figure 3.5. Thus the proposed change for each variable in the update set, in the sense of root mean square of the length, decreases with $m^{-1/2}$.

3.2.3.3. *Comparison to volume-uniform sampling in hypersphere.* For large m, sampling proposed moves uniformly in volume from a fixed hypersphere displays similar behavior to Gaussian sampling. For fixed $0 < \varepsilon < 1$ the proportion of the volume of a hypersphere of radius $(1-\varepsilon)R$ to that of a hypersphere of radius R is $(1-\varepsilon)^{-m}$, which of course

[8]Of course, the highest probability density of the multivariate Gaussian distribution always occurs at the origin, but the region near the surface of the hypersphere of radius R_m contains an increasing share of the proposed moves as m increases. Despite the lower probability density, the volume near this surface increases as m increases.



(A)



(B)

FIGURE 3.5. Behavior of (a) mean Euclidean norm and (b) relative width of Euclidean norm as the number of dimensions in each Gaussian proposed move increase.

approaches 0 as $m \to \infty$.[9] Since both distributions are spherically symmetric and heavily concentrated on the surface of a hypersphere as $m \to \infty$, it seems plausible that an analogous theorem to the original 0.234 proof by Gelman et al. can be investigated using this type of uniform distribution, though the same issues with the global update scheme would limit its direct relevance to practical calculations.

3.3. CONCLUSIONS

Although Gelman and co-workers have created an interesting mathematical proof of an optimal acceptance ratio, the result has no practical applications that we have been able to find. The proof assumes a global update scheme requiring that all variables be updated together. For a system with N independent variables, the global update scheme decreases efficiency by a factor of N compared to a single-variable update scheme. The computer time consumed by a single sweep of the single updating scheme cannot be more than twice that used by a sweep of the joint updating scheme on the same system, and usually the computer time difference per sweep is negligible. So the use of a joint updating scheme is inadvisable for large systems.

While investigating the algorithm required by GRG, we discovered an interesting side effect of the joint update scheme, which in contrast to the above general effect is significant even when the move dimension is small. While it can be proved that no correlations between the simultaneous values of independent random variables are introduced by any MCMC algorithm that satisfies detailed balance, joint updating with move dimension $m > 1$ can introduce spurious correlations between the values of these random variables at different times. This counterintuitive result is important in evaluating the efficiency of any algorithm that uses joint updating with multi-dimensional MCMC moves. Joint updating should be avoided in favor of the single updating scheme with $m = 1$ unless there is a specific reason for their use when dealing with a particular simulation. Such exceptions may occur when dealing with a highly anisotropic model, but even in those

[9] An analogous phenomenon occurs with the volume-uniform distribution on a hypercube, but there is not spherical symmetry in that case, making it less likely that a similar proof technique to that used to arrive at Gelman et al.'s optimal acceptance ratio would work.

cases where joint updating is crucial the move dimension m should be kept as low as possible.

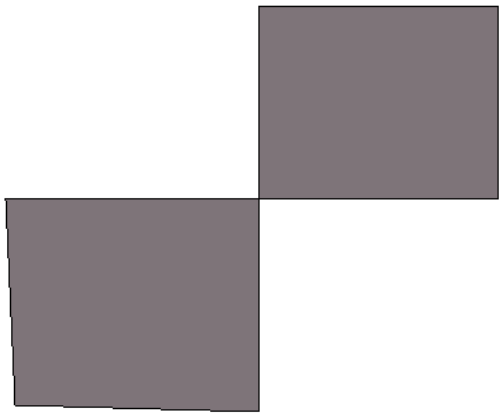


FIGURE 4.1. If $\mathsf{supp}(\Pi) \ll \mathcal{L}^N$ and is the gray area in this picture, a global update using a Gaussian or uniform proposal distribution will be irreducible since the support is connected. However, the single-variable update using horizontal and vertical moves only will fail to be irreducible; moving from one square to the other first requires moving to the edge of the square, which has probability 0.

Part 4. Conditions for Π-irreducibility and aperiodicity

4.1. Fixed cycle subset update

Chan and Geyer in [10] showed that the single-variable update using a fixed cycle (which they call variable-at-a-time Metropolis) is Π-irreducible provided that every one of the subkernels is irreducible with respect to the conditional distribution of Π given the values of the other variables. While this is a useful result, and can be trivially extended to our partition updates, the hypothesis can be weakened somewhat.

If $\Omega = \mathbb{R}^N$ and $\mathsf{supp}(\Pi)$ is open and connected but not convex, the result of Chan and Geyer would require that each subset proposal distribution q_i be able to "jump the gap" between points in $\mathsf{supp}(\Pi)$ which differ only in variables in the set U_i, but are separated by points outside $\mathsf{supp}(\Pi)$. In practice, this would mean that the width of a uniform proposal distribution, might have to be larger than we would like. Large widths lower the acceptance ratio. So we will weaken the hypotheses of Chan and Geyer in the following.

Define the subspace distance function for $\mathbf{x}, \mathbf{y} \in \Omega$ by

$$d_i(\mathbf{x}, \mathbf{y}) := \left[\sum_{j \in U_i} \left(y_j - x_j \right)^2 \right]^{1/2} \tag{4.1.1}$$

and the subspace ball

$$B_i(\mathbf{x}, \rho) := \{ \mathbf{y} \in \Omega | d_i(\mathbf{x}, \mathbf{y}) < \rho \} \tag{4.1.2}$$

To show Π-irreducibility, given $\mathbf{x} \in \Omega$ and $A \in \mathcal{A}$ with $\Pi(A) > 0$, we seek to determine a finite sequence of sets, each with positive Π-probability, such that a fixed cycle subset update kernel is able to move from $\mathbf{x}$ to some state in A by moving from one set in the sequence to the next in a finite number of time steps.

First we require a general result about product measures.

Lemma 25. *Let $(\mu, U, \mathcal{U})$ and $(\nu, V, \mathcal{V})$ be measure spaces and let $A \subseteq U \times V$ be measurable with respect to the product measure $\mu \times \nu$. Define*

$$A_{\mathbf{u}} := \{ (\mathbf{u}, \mathbf{v}) \in A | \mathbf{v} \in V \} \tag{4.1.3}$$

$$A_{\mathbf{v}} := \{ (\mathbf{u}, \mathbf{v}) \in A | \mathbf{u} \in U \} \tag{4.1.4}$$

and

$$A_U := \{ \mathbf{u} \in U | \nu \left(A_{\mathbf{u}} \right) > 0 \} \tag{4.1.5}$$

$$A_V := \{ \mathbf{v} \in V | \mu \left(A_{\mathbf{v}} \right) > 0 \} \tag{4.1.6}$$

Then if $\mu \left(A_U \right) = 0$ or $\nu \left(A_V \right) = 0$, we must have $(\mu \times \nu)(A) = 0$.

Proof. The product measure $\mu \times \nu$ can be written as

$$(\mu \times \nu)(A) = \int_U \nu \left(A_{\mathbf{u}} \right) \mu(d\mathbf{u}) = \int_V \mu \left(A_{\mathbf{v}} \right) \nu(d\mathbf{v}) \tag{4.1.7}$$

Suppose WoLoG that $\mu(A_U) = 0$. By the hypothesis of the theorem we have

$$(\mu \times \nu)(A) = \int_U \nu(A_{\mathbf{u}})\, \mu(d\mathbf{u}) = \int_{A_U} \nu(A_{\mathbf{u}})\, \mu(d\mathbf{u}) = 0 \tag{4.1.8}$$

□

Armed with this result, we can now approach the main Π-irreducibility results. Suppose that, for some $\rho > 0$, each proposal distribution Q_i has $B_i(\mathbf{x}, \rho) \subseteq \mathsf{supp}\,(Q_i(\mathbf{x}, \cdot))$ for all $\mathbf{x} \in \Omega$. Suppose also that $\Pi \ll \mathcal{L}^N$ and $\Pi(\mathsf{int}(\mathsf{supp}(\Pi))) = 1$. Then we prove that for any $\mathbf{x} \in \Omega$ and set $A \in \mathcal{A}$ with $\Pi(A) > 0$, the set A is accessible from $\mathbf{x}$ in a finite number of time steps

- when $\mathbf{x}$ and A are both inside a "small" rectangle inside the support of Π,
- when $\mathbf{x}$ and A are both inside an open connected set inside the support of Π,
- for any $A \in \mathcal{A}$ with $\Pi(A) > 0$.

Note that the condition on the support's interior does exclude some distributions with $\Pi \ll \mathcal{L}^N$. If $\Omega = \mathbb{R}^N$, $\Pi \ll \mathcal{L}^N$, and the boundary of $\mathsf{supp}(\Pi)$ is of dimension $N - 1$ or less, then we immediately have that $\Pi(\mathsf{int}(\mathsf{supp}(\Pi))) = \Pi(\mathsf{supp}(\Pi)) = 1$, so this condition is violated only for supports with fractal boundaries. Let $F \subseteq \mathbb{R}$ be a set with $\mathcal{L}^1(F) > 0$ and $\mathsf{int}(\mathsf{cl}(F)) = \emptyset$, that is, a nowhere-dense set with positive measure, such as the "Fat Cantor set" discussed in [19]. Then the uniform distribution for F, as defined in (1.1.11), has $\mathsf{Unif}_F \ll \mathcal{L}^1$ and $\mathsf{supp}\,(\mathsf{Unif}_F) \subseteq \mathsf{cl}(F)$. But F is nowhere-dense, so

$$\mathsf{Unif}_F\,(\mathsf{int}\,(\mathsf{supp}\,(\mathsf{Unif}_F))) = \mathsf{Unif}_F(\emptyset) = 0 \tag{4.1.9}$$

Fortunately, such distributions do not often arise in physical applications.

Lemma 26. *Suppose $\Pi \ll \mathcal{L}^N$. Suppose that for every $i \in \mathbb{N} \cap [1, \ell]$, there is some $\rho_i > 0$ such that for every $\mathbf{x} \in \Omega$ we have $q_i(\mathbf{x}, \mathbf{z}) > 0$ for all $\mathbf{z} \in B_i(\mathbf{x}, \rho_i)$. Let $R \subseteq \mathsf{supp}(\Pi)$ be a*

rectangle with side length s_j *in the* j*th dimension for* $j \in \mathbb{N} \cap [1, N]$. *Define*

$$s := \max_j s_j \tag{4.1.10}$$

$$\rho := \min_i \rho_i \tag{4.1.11}$$

If $s\sqrt{N} < \rho$, *then for any* $\mathbf{y} \in R$ *and any set* $A \in \mathcal{A}$ *with* $A \subseteq R$ *and* $\Pi(A) > 0$, *we have*

$$P(\mathbf{y}, A) > 0 \tag{4.1.12}$$

Proof. We can define for $i \in \mathbb{N} \cap [1, \ell]$

$$\mu_i := \nu_1 \times \nu_2 \times \cdots \times \nu_i \tag{4.1.13}$$

$$W_i := V_1 \times V_2 \times \cdots \times V_i \tag{4.1.14}$$

and then write for $i \in \mathbb{N} \cap [1, \ell]$

$$0 < \mathcal{L}^N(A) = (\mu_{\ell-1} \times \nu_\ell)(A) \tag{4.1.15}$$

From the contrapositive of Lemma 25, we see that

$$\mu_{\ell-1}\left(A_{W_{\ell-1}}\right) > 0 \tag{4.1.16}$$

and we have that $P_\ell(\mathbf{z}, A) > 0$ for $\mathbf{z} \in A_{W_{\ell-1}}$.

Analogously, let $i \in \mathbb{N} \cap [1, \ell - 2]$ be given. Write

$$\mu_{i+1}\left(A_{W_{i+1}}\right) = (\mu_i \times \nu_{i+1})\left(\left[A_{W_{i+1}}\right]_{W_i}\right) \tag{4.1.17}$$

so by the contrapositive of Lemma 25, if $\mu_{i+1}\left(A_{W_{i+1}}\right) > 0$, we have that $P_i\left(\mathbf{z}, A_{W_{i+1}}\right) > 0$ for all $\mathbf{z} \in A_{W_i}$ and $i \in \mathbb{N} \cap [1, \ell - 1]$. Thus $P(\mathbf{y}, A) > 0$. □

Theorem 27. *Suppose* $\Pi \ll \mathcal{L}^N$. *Let* $C \subseteq \mathsf{supp}(\Pi)$ *be an open connected set. Suppose that for some* $\rho > 0$, *we have* $q_i(\mathbf{z}, \mathbf{y}) > 0$ *for all* $i \in \mathbb{Z} \cap [0, \ell)$, $\mathbf{y} \in B_i(\mathbf{z}, \rho)$ *and* $\mathbf{z} \in \Omega$. *Then there is some* $n \in \mathbb{N}_0$ *for which*

$$P^n(\mathbf{x}, A) > 0 \tag{4.1.18}$$

for any $t \in \mathbb{N}_0$, $\mathbf{x} \in C$, *and* $A \subseteq C$ *with* $\Pi(A) > 0$.

Proof. Let $\mathbf{x} \in C$ and $A \subseteq C$ with $\Pi(A) > 0$ be given. Find an open rectangle $R^* \subseteq C$ with side length smaller than ρ, and $\Pi(R^* \cap A) > 0$. Let $\mathbf{y}$ be the centroid of R^*. Since C is open and connected, it is pathwise connected. Find a continuous path function $\gamma : [0,1] \to C$ such that $\gamma(0) = \mathbf{x}$ and $\gamma(1) = \mathbf{y}$.

The details of the proof from this point are a bit tedious, but this is the general outline (and we also refer the reader to Figure 4.2): we cover the path from $\mathbf{x}$ to $\mathbf{y}$ in C with open rectangles small enough for Lemma 26 to apply, and which are entirely inside the open set C. The path is compact as it is the image of a compact set under a continuous function, so we find a finite subset of these open rectangles which still cover the path, and order them according to the largest value in $[0,1]$ which is mapped into the rectangle by γ. Since these are open sets they must overlap with some other member of the subset, and these intersections (if nonempty) must have positive measure. In this way we can use Lemma 26 to guarantee a positive probability of moving from one rectangle to another without repetition. Thus we can eventually reach a rectangle which contains $\mathbf{y}$ in a finite number of steps. From there we apply Lemma 26 to R^* to ensure positive probability of moving from the final rectangle in the cover to the destination set A.

For each $u \in [0,1]$, let $R_u \subseteq C$ be an open rectangle containing $\gamma(u)$ and with side length less than ρ. This is possible because C is open. The image set $\Gamma = \gamma([0,1])$ is compact and covered by $\{R_u\}_{u \in [0,1]}$, so we can find a finite sequence $\{u_s\}_{s=1}^k$ among $[0,1]$ such that Γ is covered by $\bigcup_{s=1}^k R_{u_s}$. For each s, let v_s be the least upper bound of the parameter which is mapped into the rectangle R_{u_s} by γ. That is,

$$v_s := \sup \{u \in [0,1] | \gamma(u) \in R_{u_s}\} \tag{4.1.19}$$

WoLoG assume that v_s is nondecreasing with respect to s (this can be ensured by reordering the rectangles' indices). For any $s \in \mathbb{Z} \cap [1,k)$ with $v_s < 1$, we have $\gamma(v_s) \in R_{u_w}$ for some w_s with $s < w_s \leq k$ and $R_{u_s} \cap R_{u_{w_s}} \neq \emptyset$. The fact that these open rectangles intersect inside $\mathrm{supp}(\Pi)$ means that $\Pi\left(R_{u_s} \cap R_{u_{w_s}}\right) > 0$, so by Lemma 26 we have $P\left(\mathbf{z}, R_{u_{w_s}}\right) > 0$ for any $\mathbf{z} \in R_{u_s}$. Now we have one of two possibilities.

Case 1: $v_1 = 1$. In this case, $\mathbf{x}, \mathbf{y} \in R_{u_1}$ and thus $R_{u_1} \cap R^* \neq \emptyset$. Then

$$P^2(\mathbf{x}, A) \geq P^2(\mathbf{x}, A \cap R^*) \tag{4.1.20}$$

$$\geq \int_{R_{u_1}} P(\mathbf{x}, d\mathbf{z}) P(\mathbf{z}, A \cap R^*) \tag{4.1.21}$$

$$> 0 \tag{4.1.22}$$

Case 2: $v_1 < 1$. Then we can construct a finite sequence in the following manner: let $t_1 := 1$, and if $v_{t_w} < 1$, let t_{w+1} satisfy $t_w < t_{w+1} \leq k$ and $R_{u_{t_w}} \cap R_{u_{t_{w+1}}} \neq \emptyset$. That is, t_{w+1} is the index of a rectangle containing part of the image curve with higher parameter values which intersects R_{t_w}. Since $v_k = 1$, we know that there will be some $b \in \mathbb{Z} \cap [1,k]$, with $v_{t_b} = 1$ and thus $\mathbf{y} \in R_{t_b}$. This implies that $P\left(\mathbf{z}^{(b)}, A \cap R^*\right) > 0$ in view of Lemma 26. Then

$$P^{b+1}(\mathbf{x}, A) \geq P^{b+1}(\mathbf{x}, A \cap R^*) \tag{4.1.23}$$

$$\geq \int_{R_{t_b}} \int_{R_{t_{b-1}}} \cdots \int_{R_{t_1}} P\left(\mathbf{x}, \mathbf{z}^{(1)}\right) \left[\prod_{w=2}^{b} P\left(\mathbf{z}^{(w-1)}, d\mathbf{z}^{(w)}\right)\right] P\left(\mathbf{z}^{(b)}, A \cap R^*\right) \tag{4.1.24}$$

$$> 0 \tag{4.1.25}$$

□

Since $\mathrm{supp}(\Pi)$ has finite Π-measure, it can have only countably many connected components with positive Π-measure. Thus to have irreducibility, we only need Theorem 27

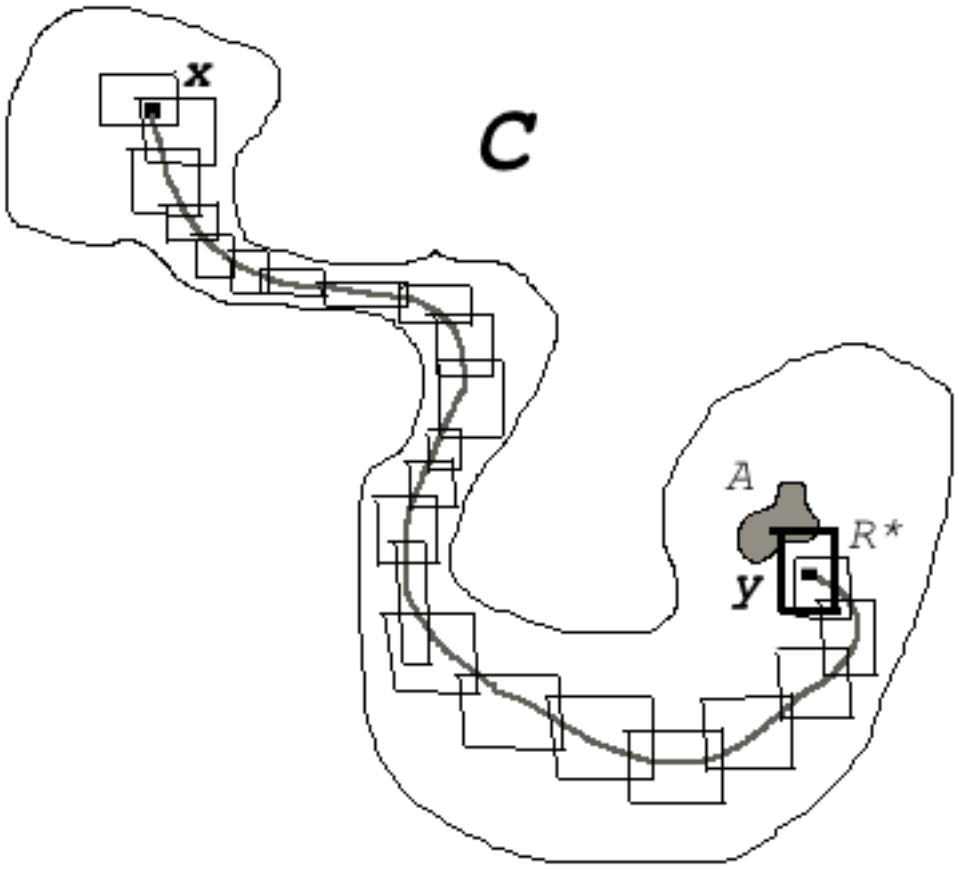


FIGURE 4.2. Illustration of the proof of Theorem 27. It is possible to move from the point x to a rectangle R^* containing both part of the set A and the point y by moving through the rectangles covering the path between those points in succession. Once that rectangle is reached, Lemma 26 guarantees we can reach the intersection $A \cap R^* \subseteq A$.

to hold for each connected component, and to have the various subkernels able to jump from one connected component to another. This gives the following.

Theorem 28. *Suppose* $\Pi \ll \mathcal{L}^N$, *with* $\Pi(\mathsf{int}(\mathsf{supp}(\Pi))) = 1$ *and density* $\pi(\mathbf{x}) = 0$ *for all* $\mathbf{x} \in \Omega \backslash \mathsf{supp}(\Pi)$, *and there is a sequence* $\{C_j\}_{j \in \mathbb{N}}$ *of connected components of* $\mathsf{int}(\mathsf{supp}(\Pi))$ *including each. Suppose there are sequences* $\{D_j\}_{j \in \mathbb{N}}, \{E_j\}_{j \in \mathbb{N}}$ *of sets in* $\mathcal{A}$, $\{i_j\}_{j \in \mathbb{N}}$ *with* $i_j \in \mathbb{N}_0 \cap [0, \ell)$, *and* $\rho > 0$ *such that, for each* $j \in \mathbb{N}$, *we have* $D_j \subseteq C_j$, $E_j \subseteq C_j$, $\mathcal{L}^N(D_j) >$

0, $\mathcal{L}^N\left(E_j\right) > 0$ *and* $d_{i_j}(\mathbf{x},\mathbf{y}) < \rho$ *for all* $\mathbf{x} \in D_j$, $\mathbf{y} \in E_{j+1}$. *Suppose also that we have* $q_i(\mathbf{x},\mathbf{y}) > 0$ *for all* $\mathbf{y} \in B_i(\mathbf{x},\rho)$. *Then* P *is* Π*-irreducible.*

Proof. Let $\mathbf{x} \in \Omega$ with $\pi(\mathbf{x}) > 0$ and $A \in \mathcal{A}$ with $\Pi(A) > 0$ be given. We seek to show that $P^n(\mathbf{x}, A) > 0$ for some $n \in \mathbb{N}_0$. First a bit of a roadmap for the proof (and we refer the reader to Figure 4.3). We will construct a sequence of sets forming a path from $\mathbf{x}$ to somewhere in A. Every transition in the path will be either between subsets of the same connected component, or from connected component C_j to C_{j+1}, and specifically from the departure region $D_j \subseteq C_j$ to the entrance region $E_{j+1} \subseteq C_{j+1}$.

Since $\Pi\left(B \cap \mathsf{supp}(\Pi)\right) = \Pi(B)$ for any $B \in \mathcal{A}$, we can find $s_A \in \mathbb{N}$ such that $\Pi\left(A \cap C_{s_A}\right) > 0$. We must have $\mathbf{x} \in \mathsf{supp}(\Pi)$ by hypothesis, so we can find $s_x \in \mathbb{N}$ such that $\mathbf{x} \in C_{s_x}$. Note that s_x is the index of the connected component containing x and s_A is that of a connected component including A. If $s_x = s_A$, then they are in the same connected component. By Theorem 27 we would then have $P^n(\mathbf{x}, A) > 0$ for some $n \in \mathbb{N}_0$ and be done.

So suppose WoLoG that $s_x < s_A$. Let $s \in \mathbb{N}_1 \cap [s_x, s_A)$ be given. Let $\mathbf{z} \in D_s$ be given with $\pi(\mathbf{z}) > 0$. Then

$$P\left(\mathbf{z}, E_{s+1}\right) = \int_{E_{s+1} \cap V_i} q_{i_s}(\mathbf{z}, \mathbf{z}+\mathbf{y}) \min\left\{1, \frac{\pi(\mathbf{z}+\mathbf{y})}{\pi(\mathbf{z})}\right\} \nu_i(d\mathbf{y}) > 0 \tag{4.1.26}$$

and thus $P\left(D_s, E_{s+1}\right) > 0$ for all $i \in \mathbb{N}$. From Theorem 27 we know that $P^{n_s}\left(E_s, D_s\right) > 0$ for some $n_s \in \mathbb{N}_0$ and $P^{n_A}\left(E_{s_A}, A \cap C_{s_A}\right) > 0$ for some $n_{s_A} \in \mathbb{N}_0$. Thus

$$P^n(\mathbf{x}, A) \geq P^{n_{s_x}}\left(\mathbf{x}, D_{s_x}\right) P^{n_{s_A}}\left(E_{s_A}, A \cap C_{s_A}\right) \prod_{s=s_x+1}^{s_A-1} P\left(D_{s-1}, E_s\right) P^{n_s}\left(E_s, D_s\right) \tag{4.1.27}$$

$$> 0 \tag{4.1.28}$$

for $n := s_A - s_x + \sum_{s=s_x}^{s_A} n_s$. □

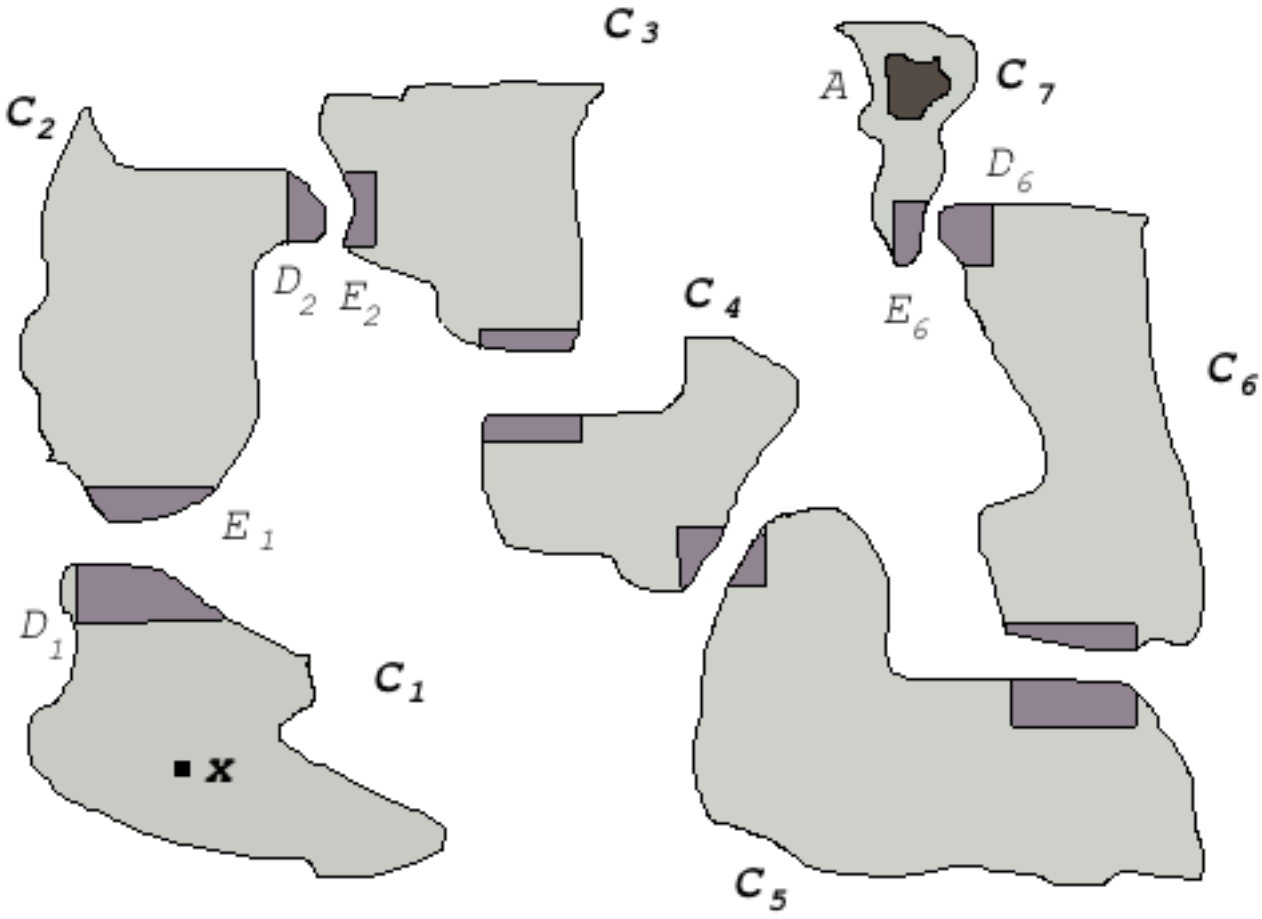


FIGURE 4.3. Illustration of the proof of Theorem 28. The light gray regions are the connected components C_i of $\mathrm{supp}(\Pi)$. Moving from the point x to the set A requires jumping (using horizontal or vertical moves only) between the departure regions D_i and the entrance regions E_i, in dark gray. Theorem 28 guarantees that we can move from the entrance region to the next departure region through the open, connected set C_i.

Next we move on to show aperiodicity of a fixed cycle subset update kernel. Recall that it is sufficient to show that there is some $n_0 \in \mathbb{N}$ for which

$$\Pr\left[\mathbf{X}(t+n) \in A | \mathbf{X}(t) \in A\right] > 0 \tag{4.1.29}$$

for any $n \in \mathbb{N} \cap [n_0, \infty]$.

Theorem 29. *Suppose* $P(\mathbf{x}, \cdot) \ll \mathcal{L}^N$ *for every* $\mathbf{x} \in \Omega$ *and there is* $\rho > 0$ *such that, for each* $0 \leq i < \ell$*, we have* $q_i(\mathbf{x}, \mathbf{x}+\mathbf{y}) > 0$ *for all* $\mathbf{y} \in B_i(\mathbf{0}, \rho)$*. If* P *is* Π*-irreducible, then* P *is aperiodic.*

Proof. Let $A \subseteq \Omega$ be given such that $\Pi(A) > 0$. Let $\mathbf{x} \in A$ be given.

Case 1: Suppose A has a region of positive Lebesgue measure within a distance ρ of $\mathbf{x}$, that is, $\nu_i\left(B_i(\mathbf{x}, \rho) \cap A\right) > 0$.

$$\begin{aligned} P_i(\mathbf{x}, A) &= 1 - \int_{V_i \setminus (A-\mathbf{x})} q_i(\mathbf{x}, \mathbf{x}+\mathbf{y}) a(\mathbf{x}, \mathbf{y}) \nu_i(d\mathbf{y}) \\ &\geq 1 - \int_{V_i \setminus (A-\mathbf{x})} q_i(\mathbf{x}, \mathbf{x}+\mathbf{y}) \nu_i(d\mathbf{y}) \\ &= \int_{A-\mathbf{x}} q_i(\mathbf{x}, \mathbf{x}+\mathbf{y}) \nu_i(d\mathbf{y}) \\ &\geq \int_{(A-\mathbf{x}) \cap B_i(\mathbf{0}, \rho)} q_i(\mathbf{x}, \mathbf{x}+\mathbf{y}) \nu_i(d\mathbf{y}) \qquad (4.1.30) \\ &> 0 \qquad (4.1.31) \end{aligned}$$

and

$$\begin{aligned} P(\mathbf{x}, A) &= \int_A \int_\Omega \cdots \int_\Omega P_1\left(\mathbf{x}, d\mathbf{z}^{(1)}\right) \left[\prod_{i=2}^{\ell-1} P_i\left(\mathbf{z}^{(i-1)}, d\mathbf{z}^{(i)}\right)\right] P_\ell\left(\mathbf{z}^{(\ell-1)}, d\mathbf{y}\right) \\ &\geq \int_A \int_A \cdots \int_A P_1\left(\mathbf{x}, d\mathbf{z}^{(1)}\right) \left[\prod_{i=2}^{\ell-1} P_i\left(\mathbf{z}^{(i-1)}, d\mathbf{z}^{(i)}\right)\right] P_\ell\left(\mathbf{z}^{(\ell-1)}, d\mathbf{y}\right) \\ &> 0 \qquad (4.1.32) \end{aligned}$$

so for any $n \geq 1$, we have $P^n(\mathbf{x}, A) > 0$.

Case 2: Suppose $\nu_i\left(\mathbf{x}, B_i(\mathbf{x}, \rho) \cap A\right) = 0$. Since $\Pi(A) > 0$, we can find a point $\mathbf{y} \in A$ such that $\Pi(B(\mathbf{y}, \rho) \cap A) > 0$. By Π-irreducibility, we know that there is some $n \in \mathbb{N}$ such that

$$P^n\left(\mathbf{x}, B(\mathbf{y}, \rho)\right) > 0 \qquad (4.1.33)$$

and Case 1 applies to $\mathbf{y}$. So for any $k \geq n$, we have $P^k(\mathbf{x}, A) > 0$. □

4.2. GLOBAL UPDATE AND MIXTURE OR MIXED CYCLE SUBSET UPDATE

Ergodicity of the global update Metropolis algorithm is shown in [9] to require only that the proposal distribution $Q(\mathbf{x}, \cdot) \ll \mathcal{L}^N$ and $Q(\mathbf{x}, A)$ be positive on every open set A with $\mathbf{x} \in A$. For mixtures, we know from Chan and Geyer [10] that Π-irreducibility and aperiodicity of each $P_i(\mathbf{x}, \cdot)$ on the translated subspace $V_i + \mathbf{x}$ is sufficient to prove Π-irreducibility and aperiodicity for the full mixture updating algorithm.

REFERENCES

[1] Gelman, A.; Roberts, G. O.; Gilks, W. R. "Efficient Metropolis jumping rules." Bayesian Statistics 5, 599-608. (Clarendon, Oxford, 1996)

[2] Roberts, G. O.; Gelman, A.; Gilks, W. R. "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms." The Annals of Applied Probability, Vol. 7, No. 1 (Feb 1997), pp. 110-120.

[3] Müller-Krumbhaar, H.; Binder K. "Dynamic properties of the Monte Carlo method in statistical mechanics", Journal of Statistical Physics, Vol. 8, No. 1, 1973.

[4] D. Bouzida, S. Kumar, and R.H. Swendsen,"Almost Markov Processes." in *Computer Simulation Studies in Condensed Matter Physics III* (Athens, Georgia, 12-16 February 1990), Edited by Landau, D. P.; Mon, K. K.; Schüttler, H.-B. Springer-Verlag, Berlin-Heidelberg (1991), pp. 193-196.

[5] Bouzida, D.; Kumar, S.; Swendsen, R. H. "Efficient Monte Carlo Methods for Computer Simulation of Biological Molecules." Phys. Rev. A, 45, 8894-8901 (1992)

[6] Bouzida, D.; Kumar, S.; Swendsen, R. H. "A Simulated Annealing Approach for Probing Biomolecular Structures." *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences 1993*. Edited by Mudge, T.N.; Milutinovic, V.; Hunter, L. IEEE Computer Society Press (1993), pp.736-742.

[7] Langford, Eric; Schwertman, Neil; Owens, Margaret. "Is the Property of Being Positively Correlated Transitive?" The American Statistician, Vol. 55, No. 4 (Nov 2001), pp. 322-325

[8] Glauber, Roy J. "Time Dependent Statistics of the Ising Model." J. Math. Phys. 4, 294 (1963)

[9] Tierney, Luke. "Markov Chains for Exploring Posterior Distributions." The Annals of Statistics, Vol 22, No 4 (1994), pp 1701-1728.

[10] Chan, K.S.; Geyer, C.J. Discussion of Tierney's paper. The Annals of Statistics, Vol 22, No 4 (1994), pp1747-1758.

[11] Athreya, Krishna B.; Doss, Hani; Sethuraman, Jayaram. "A proof of convergence of the Markov chain simulation method" Technical Report 868, Dept. of Statistics, Florida State University (1992).

[12] Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. (1953). "Equations of State Calculations by Fast Computing Machines". Journal of Chemical Physics 21 (6): 1087–1092.

[13] Schilling, Rene. *Measures, Integrals, and Martingales*. Cambridge University Press (2006).

[14] Leoni, Giovanni. *A First Course In Sobolev Spaces*. American Mathematical Society (2009).

[15] Schwartz, Russell. *Biological Modeling and Simulation*: *A Survey of Practical Models, Algorithms, and Numerical Methods*. MIT Press (2008).

[16] Nummelin, Esa. *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press (1984).

[17] Manousiouthakis, Vasilios I.; Deem, Michael W. "Strict detailed balance is unnecessary in Monte Carlo simulation." J. Chem. Phys. 110, 2753 (1999).

[18] Reynolds, John F. "Some Theorems on the Transient Covariance of Markov Chains." Journal of Applied Probability, Vol 9, No 1 (Mar 1972), pp 214-218.

[19] Gelbaum, Bernard R.; Olmsted, John M. H. *Counterexamples in Analysis.* Dover Publications (2003).

$\ll$ (absolutely continuous), 4
$\rightleftarrows$ (reverse), 37
$\perp$ (singular), 5
$[[\cdot]]$ (positive modulus), 35

A

absolutely continuous, 4
almost every, 2
aperiodic, 9, 11
autocorrelation function, 27
  global, 47
  local, 46

B

basis measure, 4

C

complete, 2
completion, 2
correlation time, 27
counting measure, 1

D

density function, 4
detailed balance, 9, 13

E

equilibrium distribution, 9
ergodic, 9

G

global update, 20

H

H_t, 7

I

invariant measure, 11
irreducible, 9, 11

J

joint update, 20

K

kernel
  off-diagonal, 12
  product of, 6
  stochastic, 5
  substochastic, 12
  transition, 6

L

Lebesgue measure, 2, 3

M

master equation, 8, 10
measurable, 1
measure, 1
  probability, 1

N

$\mathbb{N}_0$ (natural numbers with 0), 5
nowhere-dense, 71

P

partition update, 20
probability space, 1

product space, 3

S

σ-algebra, 1

σ-finite, 3

singular, 5

subset update, 19

support, 3

T

time-homogeneous, 7

total balance, 9, 11

total variation, 7

U

uniform distribution, 4

update set, 17