Exact and Approximate Sampling from the Stationary Distribution of a Markov Chain

Michael K. Schneider

November 2, 1998

Abstract

When simulating a physical system with discrete states, one often would like to generate a sample from the stationary distribution of a Markov chain. This report focuses on three sampling methodologies which do not rely on explicitly computing the stationary distribution. Two of these lead to algorithms which can generate an exact sample in finite time. The third yields a sample whose distribution approximates, but is arbitrary close to, the stationary distribution from which one desires a sample. The approximate and one of the exact methodologies are illustrated with examples from statistical mechanics.
Contents

1 Introduction 3

2 Markov Chain Monte Carlo for Statistical Mechanics 5
   2.1 Ising Model ................................................. 5
   2.2 Metropolis and Heat Bath Chains .......................... 8

3 Coupling-From-the-Past 12
   3.1 The Algorithm .............................................. 12
   3.2 Monotone Coupling ......................................... 14
   3.3 Multi-γ Coupling ........................................... 16
   3.4 Run-Time Bounds for Monotone Coupling ................. 19

4 Run-Time Bounds for Forward Simulation 24
   4.1 Inequalities .................................................. 24
   4.2 Bounds for the Subgraphs-World Process ................. 29
   4.3 Countable Spaces and Minorization ...................... 33

5 Stopping Times 35
   5.1 Existence Questions ....................................... 35
   5.2 Randomized Stopping Time Construction .................. 37

6 Conclusion 42
1 Introduction

When studying the behavior of systems arising in the field of statistical mechanics, physicists are often interested in generating a sample from a high-dimensional distribution. A typical method for generating a sample on a computer is Markov chain Monte Carlo (MCMC). The first step of MCMC is to construct an ergodic Markov chain whose stationary distribution is the same as the one from which one would like to sample. Then one starts the chain off in a particular state and simulates the transitions on a computer. After the computer has simulated the chain long enough, the chain will almost be in steady-state. This report addresses exactly how long the chain must be run so that it’s almost in steady-state and the use of variations on this sampling idea that allow one to sample exactly from the stationary distribution of a Markov chain in finite time. The focus is on three specific methodologies.

The first is an algorithm introduced by Propp and Wilson known as the coupling-from-the-past (CFTP) protocol [6P16–18P23]. This algorithm is capable of generating an exact sample from the stationary distribution of a large class of Markov chains. If the state space is large, however, the chain must have some special structure for the algorithm to be efficient. Fortunately, many chains arising in statistical mechanics have this structure; so the algorithm is especially appropriate for applications in this field.

The second methodology has also been successfully applied to a problem in statistical mechanics [7P11P19]. The idea behind this approach is to develop bounds on the number of steps needed for a finite-state Markov chain to be almost in steady-state. The bounds exploit the Laplacian-like structure of the transition matrix of a reversible Markov chain one whose steady-state statistics are the same looking forwards and backwards in time. The bounds are written in terms of the geometry of the underlying Markov chain. Although these bounds are easy to compute for many situations
and have been computed for a problem in statistical mechanics; they are often too difficult to
compute in practice. This is one reason why CFTP would be more attractive for chains which have
the structure necessary to make the algorithm efficient.

However, CFTP is not the only method for generating an exact sample from the stationary
distribution. The third methodology constructs a computable randomized rule to determine when
to stop a finite-state Markov chain so that the current state is exactly from the stationary distri-
bution [3]. The rule makes use of the continuous time Markov process associated with the chain.
Unfortunately, the rule is not practical for chains with large state spaces because it is too com-
putationally intense. However, the result is interesting because it demonstrates the existence of
methodologies other than CFTP for generating an exact sample from a Markov chain.

Sections 3, 4, and 5 present each of these methodologies in detail. Section 2 contains a descrip-
tion of some particular problems in statistical mechanics that are used to illustrate the theory in
Sections 3 and 4.

Throughout the paper, the state space $S$ of the Markov chains is finite unless otherwise noted.
The notation $P(\sigma, \sigma')$ denotes the conditional probability that the chain will move from state $\sigma$ to
state $\sigma'$ in one step and $P^m(\sigma, \sigma')$ denotes the conditional probability that the chain will move from
state $\sigma$ to $\sigma'$ in $m$ steps. For ergodic Markov chains, $\pi(i)$ denotes the probability that the chain is in
state $i$. In order to make precise the notion of a chain being approximately in steady-state, one has
to have a method for measuring the difference between probability distributions. All of the analysis
in this report will make use of the variational distance between two probability distributions. For
two distributions $\mu_1$ and $\mu_2$, the distance $\|\mu_1 - \mu_2\|$ is defined by

$$\|\mu_1 - \mu_2\| = \sup_{A \subseteq S} |\mu_1(A) - \mu_2(A)|$$

(1)

\[ ^1 \text{There is another popular definition of variational distance which is equal to twice the distance defined here.} \]
Figure 1: In the two-dimensional Ising model there is an array of sites. Each site is associate with a spin which can be up or down. The solid lines in the figure connect neighboring spins.

Note that there is an alternate and often more useful characterization of variational distance given by the following proposition. A proof is in the appendix.

**Proposition 1.**

\[
\|\mu_1 - \mu_2\| = \frac{1}{2} \sum_{\sigma \in S} |\mu_1(\sigma) - \mu_2(\sigma)|
\]  

(2)

2 Markov Chain Monte Carlo for Statistical Mechanics

2.1 Ising Model

A popular and widely scrutinized model in statistical mechanics is the two-dimensional Ising model [516]. In the model there is an array of sites (see Figure 1). One associates a spin to each site. Each spin can be in one of two states up or down corresponding to the two possible values for its magnetic moment. In the presence of a magnetic field the total energy of the system depends on the interaction between the field and the individual spins and on the interaction between neighboring
spins (neighboring spins are connected by lines in Figure 1). Specifically, the energy is written as

\[ E_{\text{Ising}} = - \sum_i B_i s_i - \sum_{\{i,j\} \in \mathcal{N}} J_{ij} s_i s_j \]  

where \( s_i = \pm 1 \) depending on whether the spin at site \( i \) is up or down, \( B_i \) is the magnetic field strength at site \( i \), \( J_{ij} \geq 0 \) is the strength of the interaction between spins at sites \( i \) and \( j \), and \( \mathcal{N} \) is the collection of pairs of sites \( (i,j) \) such that sites \( i \) and \( j \) are neighbors. In a physical system at equilibrium, the probability of a given configuration is proportional to \( e^{-E/k_B T} \) where \( E \) is the total energy, \( k_B \) is Boltzmann’s constant, and \( T \) is the absolute temperature. The normalization constant in the distribution, denoted by \( Z \), has physical significance and is called the partition function. The distribution \( e^{-E/k_B T}/Z \) is known as the Gibbs distribution. Thus, one can write the Gibbs distribution for the Ising model as

\[
\frac{1}{Z_{\text{Ising}}} \exp \left( \sum_i B_i s_i + \sum_{\{i,j\} \in \mathcal{N}} J_{ij} s_i s_j \right). \tag{4}
\]

The random cluster model is another statistical mechanical model and has close ties to the Ising model [9,16,20]. One can derive the distribution for the random cluster model from the Ising model distribution as follows. For the purposes of the present discussion, assume that no magnetic field is present. A magnetic field can be introduced into the model using the trick described in [9]. In the absence of a magnetic field, one can rewrite (3) as

\[
E_{\text{Ising}} = \sum_{\{i,j\} \in \mathcal{N}} 2J_{ij}(1 - \delta_{s_is_j}) + c \tag{5}
\]

where \( \delta \) is the Kronecker delta and \( c \) is a constant. The energy function \( E_{\text{Potts}} = E_{\text{Ising}} - c \) corresponds to that of a model known as the two-state Potts model. Note that the two-state Potts
model has the same Gibbs distribution as the Ising model but just a different partition function. Now the Gibbs distribution for the two-state Potts model can be written

\[
\frac{1}{Z_{\text{Potts}}} \prod_{(i,j) \in \mathcal{N}} e^{-2J_{ij}(1-\delta_{n_{ij}})} = \frac{1}{Z_{\text{Potts}}} \prod_{(i,j) \in \mathcal{N}} \left( (p_{ij} + (1 - p_{ij})\delta_{s_i s_j}) \right) \tag{6}
\]

where \( p_{ij} = e^{-2J_{ij}} \). Let \( n_{ij} \) be an auxiliary variable taking values in \( \{0, 1\} \) and marking whether neighboring sites \( i \) and \( j \) must have the same spin. Consider the following joint distribution for variables \( n_{ij} \) and spins \( s_i \):

\[
\frac{1}{Z_{\text{FKSW}}} \prod_{(i,j) \in \mathcal{N}} \left( (1 - p_{ij})\delta_{n_{ij} 0} + p_{ij} \delta_{n_{ij} 1}\delta_{s_i s_j} \right).
\tag{7}
\]

This is the Gibbs distribution of the Fortuin-Kastelyn-Swendsen-Wang model. Summing over all \( n_{ij} \) one arrives at the marginal distribution for the spins which is precisely that of the two-state Potts model. By summing over all spins one arrives at the following marginal distribution on the \( n_{ij} \):

\[
\frac{2^{\mathcal{C}(n)}}{Z_{\text{RC}}} \prod_{(i,j) \in \mathcal{N}} p_{ij} \prod_{n_{ij}=1} \prod_{n_{ij}=0} (1 - p_{ij}) \tag{8}
\]

where \( \mathcal{C}(n) \) denotes the number of clusters of neighboring spins for which \( n_{ij} = 1 \). This distribution is called the random cluster model. Note that conditioned on the \( n_{ij} \) the distribution of the spins in the clusters is uniform and independent. Thus one can sample from the Ising model by first sampling from the random cluster model and then picking spins for each of the clusters. This method has been used to efficiently generate samples from the Ising model distribution [21] and is discussed further in Section 3.

Another model with a close relationship to the Ising model is one which is called the subgraphs-
world process by Jerrum and Sinclair [11]. In this model one associates a variable $\epsilon_{ij}$ with every edge connecting neighboring sites in the Ising model. The variable $\epsilon_{ij}$ can take on the value zero or one. One can associate with each collection $\epsilon$ of $\epsilon_{ij}$ a graph $G(\epsilon)$ whose vertices are the spin sites and whose edges connect those sites for which $\epsilon_{ij} = 1$. The distribution of the configurations is proportional to

$$w(\epsilon) = \mu^{\text{odd}(G(\epsilon))} \prod_{(i,j) \in N} w_{ij}$$

(9)

where $w_{ij} = \tanh J_{ij}/k_B T \leq 1 \Gamma \mu = \tanh(B/k_B T)$ for a constant magnetic field $B$ and $\text{odd}(\cdot)$ is the set of odd vertices in a graph. The connection between this model and the Ising model is that the partition function of the Ising model

$$Z_{\text{Ising}} = \left(2 \cosh \frac{B}{k_B T}\right)^n \prod_{(i,j) \in N} \cosh \frac{J_{ij}}{k_B T} \sum_{\epsilon} w(\epsilon).$$

(10)

A proof of this result is given in [11]. This relationship suggests that one can use simulations of the subgraphs-world process to compute quantities concerning the Ising model. Jerrum and Sinclair present methods for performing such computations in [11]. The remainder of this report will refer to the subgraphs-world process only in the context of simulating it directly and not in the context of performing computations with the Ising model.

### 2.2 Metropolis and Heat Bath Chains

One method for sampling from the distributions discussed so far is the random update Metropolis method [811P13T20]. A sample from each of the distributions presented can be described by a system state $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_k)$ where the $\sigma_i$ would be the $s_i \Gamma n_{ij} \Gamma$ and $\epsilon_{ij}$ in the case of the
Ising model, random cluster model, and subgraphs-world process respectively. Suppose \( \pi \) is the distribution on the \( \sigma \) from which one would like to sample. Now consider the Markov chain on the states \( \sigma \) which uses the following rule to leave state \( \sigma \):

1. Pick an \( i \) uniformly from \( \{1, \ldots, k\} \).

2. Choose \( \sigma'_i \) uniformly from the possible configurations and let

   \[ \sigma' = (\sigma_1, \sigma_2, \ldots, \sigma_{i-1}, \sigma'_i, \sigma_{i+1}, \ldots, \sigma_k). \tag{11} \]

3. If \( \pi(\sigma') \geq \pi(\sigma) \), transit to \( \sigma' \); if \( \pi(\sigma') < \pi(\sigma) \), transit to \( \sigma' \) with probability \( \pi(\sigma') / \pi(\sigma) \) and stay in \( \sigma \) otherwise.

The Metropolis method for sampling from a distribution simulates this chain. One can check that the stationary distribution of this chain is indeed \( \pi \).

In order to verify that the stationary distribution is \( \pi \), one needs to recall the following three definitions and theorem. First, there is the notion of irreducible. A Markov chain is irreducible if for any two states in the state space there is a positive probability of transiting from one to the other after some finite number of steps. Next, there is the notion of aperiodic. A Markov chain is aperiodic if for any two states in the state space there is a number of steps \( m \) such that there is a positive probability of transiting from one state to the other in any number of steps greater than \( m \). Lastly, there is the notion of reversible. An ergodic Markov chain is reversible if the transitions of the chain in steady-state have the same statistics looking forwards and backwards in time. In other words

\[ \pi(\sigma)P(\sigma, \sigma') = \pi(\sigma')P(\sigma', \sigma), \quad \forall \sigma, \sigma' \in S. \tag{12} \]
The following theorem characterizes the stationary distribution of aperiodic, irreducible, and reversible Markov chains [10].

**Theorem 1.** Let \( P(\sigma, \sigma') \) be the transition probabilities of an aperiodic and irreducible Markov chain, and let \( \pi \) be any distribution that satisfies (12). Then, the chain is reversible, and \( \pi \) is the stationary distribution.

One can now apply Theorem 1 to verify that the Metropolis chain has the intended stationary distribution. First, the Metropolis chain is irreducible because there is a positive probability of making a transition from any state to any other state in a number of moves equal to the number of sites in the system being simulated. Next, the chain is aperiodic because there is a positive probability of remaining in the current state at every step. Lastly, the transition probabilities \( P(\sigma, \sigma') \) defined by the chain and the distribution \( \pi \) from which one wants a sample satisfy (12).

To see this, note that \( P(\sigma, \sigma') = 0 \) unless \( \sigma \) and \( \sigma' \) differ only at one site. Now suppose that \( \sigma \) and \( \sigma' \) only differ possibly at site \( i \). Recall that there are \( k \) sites and let \( c \) be the number of possible configurations at a site. Then, if \( \pi(\sigma') = \pi(\sigma) P(\sigma, \sigma') = P(\sigma', \sigma) = 1/kc \). If \( \pi(\sigma') > \pi(\sigma) \) \( P(\sigma, \sigma') = 1/kc \) and \( P(\sigma', \sigma) = \pi(\sigma)/kc \pi(\sigma') \). Likewise, if \( \pi(\sigma) > \pi(\sigma') \) \( P(\sigma, \sigma') = \pi(\sigma')/kc \pi(\sigma) \) and \( P(\sigma', \sigma) = 1/kc \). So, the Metropolis chain is reversible and has stationary distribution \( \pi \).

Another approach to sampling from the distributions under consideration is the cyclic heat bath method [8\( \Pi 16 \Pi 20 \)]. The cyclic heat bath method involves constructing a Markov chain on the configurations in which each site is updated sequentially using the conditional distribution. Specifically, let \( i \) be the site being updated and \( (\sigma_1, \ldots, \sigma_k) \) be the current configuration. The heat bath method selects \( \sigma'_i \) from the conditional distribution

\[
\frac{\pi(\sigma_1, \ldots, \sigma_{i-1}, \sigma'_i, \sigma_{i+1}, \ldots, \sigma_k)}{\sum_{\sigma'_i} \pi(\sigma_1, \ldots, \sigma_{i-1}, \sigma'_i, \sigma_{i+1}, \ldots, \sigma_k)}
\] (13)
where $\bar{\sigma}_i$ and $\sigma'_i$ range over the $c$ possible configurations and $\pi$ is the distribution from which one wants a sample. The Markov chain thus defined is periodic and, strictly speaking, has no steady-state. However, one can define a limiting probability for this chain.

Suppose the $P(\sigma, \sigma')$ are the transition probabilities for an irreducible Markov chain. For a particular state $\sigma$ let the period of $\sigma$ be the greatest common divisor of those $m$ such that the $m$-step transition probabilities from $\sigma$ to $\sigma$ are positive. For an irreducible chain the periods of every state are the same [10]; so let $d$ represent the common period. Now we have the following theorem concerning a periodic Markov chain [20].

**Theorem 2.** Let $P(\sigma, \sigma')$ be the transition probabilities for an irreducible Markov chain with period $d$. Suppose that the distribution $\pi'$ satisfies

$$\sum_{\sigma \in S} \pi'(\sigma)P(\sigma, \sigma') = \pi'(\sigma'), \quad \forall \sigma' \in S.$$  \hspace{1cm} (14)

Then, the $d$ irreducible, aperiodic chains formed by observing the Markov chain every $d$ steps have a stationary distribution $d\pi'$.

For the cyclic heat bath chain the state can be written as $(\sigma_1, \ldots, \sigma_k, i)$ where $(\sigma_1, \ldots, \sigma_k)$ represent the current configuration of the system and $i$ is the site that will be updated at the next transition. This chain has a period equal to the number of sites $k$. Let

$$\pi'(\sigma_1, \ldots, \sigma_k, i) = \frac{1}{k^i} \pi(\sigma_1, \ldots, \sigma_k),$$  \hspace{1cm} (15)
where \( \pi \) is the distribution from which one wants a sample. Then if \( i' \) is the next site to be updated after site \( i \)

\[
\pi'(\sigma_1, \ldots, \sigma_k, i') P \left( (\sigma_1, \ldots, \sigma_{i-1}, \sigma_i', \sigma_{i+1}, \ldots, \sigma_k, i') \right)
\]

\[
= \frac{\pi(\sigma_1, \ldots, \sigma_k) \pi(\sigma_1, \ldots, \sigma_{i-1}, \sigma_i', \sigma_{i+1}, \ldots, \sigma_k)}{\sum_{q_i} \pi(\sigma_1, \ldots, \sigma_{i-1}, \sigma_i', \sigma_{i+1}, \ldots, \sigma_k)}
\]

\[
= \frac{\pi(\sigma_1, \ldots, \sigma_{i-1}, \sigma_i', \sigma_{i+1}, \ldots, \sigma_k) \pi(\sigma_1, \ldots, \sigma_k)}{\sum_{q_i} \pi(\sigma_1, \ldots, \sigma_{i-1}, \sigma_i, \sigma_{i+1}, \ldots, \sigma_k)}
\]

\[
= \pi'(\sigma_1, \ldots, \sigma_{i-1}, \sigma_i', \sigma_{i+1}, \ldots, \sigma_k, i') P \left( (\sigma_1, \ldots, \sigma_{i-1}, \sigma_i', \sigma_{i+1}, \ldots, \sigma_k, i), (\sigma_1, \ldots, \sigma_k, i') \right)
\]

Summing over \( \sigma_i \) shows that (14) holds. Thus the chain consisting of every sweep of \( k \) steps of the heat bath method has the stationary distribution \( \pi \Gamma \) as desired.

Of course there is a randomized version of the heat bath method and a cyclic version of Metropolis. However it is the randomized Metropolis chain and cyclic heat bath chain that will be used to illustrate the ideas presented in Sections 3 and 4 for sampling from the distributions discussed at the beginning of this section.

3 Coupling-From-the-Past

3.1 The Algorithm

Suppose one would like to sample from the stationary distribution of an ergodic Markov chain. One method for doing this would be to start the chain in an arbitrary state in the infinite past \( \Gamma \) simulate the chain to obtain the sequence of states \( \ldots, X(-1), X(0), X(1), \ldots \) and output \( X(0) \). Unfortunately this requires an infinite number of steps to simulate. Now consider a collection of chains (not independent) \( X_1, X_2, \ldots \) each with the same marginal transition probabilities so that \( X_i(0) \) is distributed according to the stationary distribution of the chain for each \( i \). Suppose further
that \( X_i(0) = X_j(0) \) for all \( i \) and \( j \) with probability one. Then one can simulate \( X_i(0) \) by finding a time \( T < 0 \) such that the chains \( X_{i\sigma} \) simulated from time \( T \) starting in state \( \sigma \) are all equal at time zero.

The details of the algorithm that samples in this fashion are as follows. Let the sequence \( \ldots, U(-1), U(0), U(1), \ldots \) be independent identically distributed (IID) random variables taking values in \( \mathcal{U} \) and let \( \phi: \mathcal{U} \times S \to S \) be a function such that \( \Pr\{ \phi(U(i), \sigma) = \sigma' \} = P(\sigma, \sigma') \). The algorithm consists of the following steps:

1. Let \( T = -1 \).

2. If \( \phi\left(U(-1), \phi(U(-2), \cdots, \phi(U(T), \cdot) \cdots\right) \) is a constant, output its value as the sample; otherwise, decrement \( T \) and return to step 2.

Essentially, the algorithm is simulating a set of infinite chains by starting at time zero and working backwards. When a time is found such that all the chains being sampled have the same value at time zero, i.e., have coupled at time zero, this value is output.

The following theorem and proof by Propp and Wilson rigorously establish that the algorithm works under mild conditions [16].

**Theorem 3.** Let \( \ldots, U(-1), U(0), U(1), \ldots \) be IID random variables taking values in \( \mathcal{U} \), and let \( \phi: \mathcal{U} \times S \to S \) be a function such that

\[
\sum_{\sigma \in S} \pi(\sigma) \Pr\{ \phi(U(i), \sigma) = \sigma' \} = \pi(\sigma'), \quad \forall \sigma' \in S, i \in \{ \ldots, -1, 0, 1, \ldots \}.
\]

(20)

Let \( \Phi_T(\cdot) = \phi\left(U(-1), \phi(U(-2), \cdots, \phi(U(T), \cdot) \cdots\right) \), and assume that, with probability one, there exists a \( T < 0 \) for which \( \Phi_T \) is constant. Then, \( \lim_{T \to -\infty} \Phi_T \) exists with probability one and is distributed according to \( \pi \).
Proof. Let \( X^* \) be distributed according to \( \pi \). Then \( \Gamma \Phi_T(X^*) \) is distributed according to \( \pi \) for every \( T \). Thus \( \lim_{T \to -\infty} \Phi_T \) must also be distributed according to \( \pi \).

The function \( \phi \) defines a coupling which in this context \( \Gamma \) is a rule for simulating a collection of Markov chains with the same marginal transition probabilities in such a way that the chains eventually all take on the same value. In order for the CFTP method to be useful one needs couplings \( \phi \) that are easy to compute. One possible coupling is the independent coupling. In this case \( \phi \) is constructed such that the random variables \( \phi(U_i, \sigma) \) are independent for all \( \sigma \in S \). Since the chain is irreducible there exists an \( m \) and a state \( \sigma' \) such that there is a positive probability that the chain will transit from any state \( \sigma \) to \( \sigma' \) in \( m \) steps. Thus the conditions of Theorem 3 are satisfied and the \( \phi \) is valid. Unfortunately the \( \phi \) are difficult to compute when the state space is large. The monotone and multi-\( \gamma \) couplings can sometimes be used under these circumstances.

### 3.2 Monotone Coupling

In the setup of the monotone coupler \([616]\) one makes the following additional assumptions concerning the state space:

1. The state space is partially ordered; that is there is a relation \( \preceq \) among the elements such that \( \sigma_1 \preceq \sigma_2 \) and \( \sigma_2 \preceq \sigma_3 \) imply \( \sigma_1 \preceq \sigma_3 \) for \( \sigma_1, \sigma_2, \sigma_3 \in S \) \([14]\).

2. The state space has a maximal element \( \overline{\sigma} \) and a minimal element \( \underline{\sigma} \) such that \( \underline{\sigma} \preceq \sigma \) and \( \sigma \preceq \overline{\sigma} \) for all \( \sigma \in S \).

Then the coupling is monotone if the function \( \phi(U, \sigma) \) is monotone in \( \sigma \) i.e. \( \sigma \preceq \sigma' \) implies that \( \phi(U, \sigma) \preceq \phi(U, \sigma') \). Note that a coupling has occurred in the previous \( |T| \) steps if and only if

\[
\phi\left( U_{-1}, \phi\left( U_{-2}, \cdots, \phi\left( U_T, \underline{\sigma} \right) \cdots \right) \right) = \phi\left( U_{-1}, \phi\left( U_{-2}, \cdots, \phi\left( U_T, \overline{\sigma} \right) \cdots \right) \right).
\] (21)
This simplifies the structure of CFTP for large state spaces because one need only keep track of the trajectories starting in the maximal and minimal states. The following algorithm will generate a sample from the stationary distribution of a Markov chain using CFTP and monotone coupling.

1. Let $T = -1$.

2. Compute $\phi\left(U_{-1}, \phi\left(U_{-2}, \cdots, \phi\left(U_{T}, \sigma \right) \cdots \right)\right)$ and $\phi\left(U_{-1}, \phi\left(U_{-2}, \cdots, \phi\left(U_{T}, \bar{\sigma} \right) \cdots \right)\right)$.

3. If $\phi\left(U_{-1}, \phi\left(U_{-2}, \cdots, \phi\left(U_{T}, \sigma \right) \cdots \right)\right) = \phi\left(U_{-1}, \phi\left(U_{-2}, \cdots, \phi\left(U_{T}, \bar{\sigma} \right) \cdots \right)\right)$ output the corresponding state; otherwise double $T$ and return to 2.

The process of doubling at every time step will guarantee that one will never simulate more than twice as many transitions of the chain than are necessary.

Many chains arising in statistical mechanics admit a monotone coupling. What follows is a coupling one can use to generate a sample from the stationary distribution of the heat bath chain associated with the random cluster model presented in Section 2 [16]. Write a state $\sigma$ of the chain as $\left(\left\{ n_{kl} \right\}, \left( i, j \right) \right)$ where the $\left\{ n_{kl} \right\}$ are the variables either zero or one marking whether spins $k$ and $l$ must be equal and $\left( i, j \right) \in \mathcal{N}$ specifies that $n_{ij}$ is updated next. Define $\left(\left\{ n_{kl} \right\}, \left( i, j \right) \right) \preceq \left(\left\{ n'_{kl} \right\}, \left( i', j' \right) \right)$ if $n_{kl} \leq n'_{kl}$ for all $\left( k, l \right) \in \mathcal{N}$. Let $\ldots, U(-1), U(0), U(1), \ldots$ be a collection of IID uniform random variables and define $\phi$ by

$$
\phi(U(i), \sigma) = \begin{cases} 
\sigma_1 & \text{if } \frac{\pi(\sigma_1)}{\pi(\sigma_0) + \pi(\sigma_1)} > U(i) \\
\sigma_0 & \text{otherwise}
\end{cases}
$$

(22)

where $\sigma_0 \preceq \sigma_1$ are the two possible states into which the chain can transition from $\sigma$. Suppose that $\sigma \preceq \sigma'$ and that $n_{ij}$ is to be updated next. One would like to verify that $\phi(U(i), \sigma) \preceq \phi(U(i), \sigma')$. 


\[ \frac{\pi(\sigma)}{\pi(\sigma_0) + \pi(\sigma_1)} \leq \frac{\pi(\sigma'_1)}{\pi(\sigma'_0) + \pi(\sigma'_1)}. \] (23)

There are three cases to consider.

1. If nodes $i$ and $j$ are in different clusters in both configurations $\sigma$ and $\sigma'$ then

\[ \frac{\pi(\sigma)}{\pi(\sigma_0) + \pi(\sigma_1)} = \frac{p_{ij}}{2(1 - p_{ij} + p_{ij})} = \frac{\pi(\sigma'_1)}{\pi(\sigma'_0) + \pi(\sigma'_1)}. \] (24)

2. If nodes $i$ and $j$ are in the same clusters in both configurations $\sigma$ and $\sigma'$ then

\[ \frac{\pi(\sigma)}{\pi(\sigma_0) + \pi(\sigma_1)} = \frac{p_{ij}}{(1 - p_{ij} + p_{ij})} = \frac{\pi(\sigma'_1)}{\pi(\sigma'_0) + \pi(\sigma'_1)}. \] (25)

3. If nodes $i$ and $j$ are in separate clusters in configuration $\sigma$ but in the same cluster in configuration $\sigma'$ then

\[ \frac{\pi(\sigma)}{\pi(\sigma_0) + \pi(\sigma_1)} = \frac{p_{ij}}{2(1 - p_{ij} + p_{ij})} \leq \frac{p_{ij}}{(1 - p_{ij} + p_{ij})} = \frac{\pi(\sigma'_1)}{\pi(\sigma'_0) + \pi(\sigma'_1)}. \] (26)

Thus $\phi$ is a monotone coupling. Furthermore the states $\sigma = (\{\bar{n}_{kl}\}, (i, j))$ with all $\bar{n}_{kl} = 1$ and $\sigma = (\{n_{kl}\}, (i, j))$ with all $n_{kl} = 0$ are maximal and minimal respectively. So one can use the monotone version of CFTP to sample from $\pi$.

### 3.3 Multi-$\gamma$ Coupling

Another coupling which has potential for practical use is the multi-$\gamma$ coupling [15]. The multi-$\gamma$ coupling can be used to simulate chains for which there exists a minorization [3]. For a chain with
Figure 2: There exists no minorization for the birth-death process with birth probability $p$ and death probability $1 - p$.

![Diagram of a birth-death process with birth probability $p$ and death probability $1 - p$.]

Figure 3: In this process the state space consists of the non-negative integers. There is a probability of $p$ of increasing by one at every step and $1 - p$ of returning to zero. For this process there exists a minorization transition probabilities $P(\sigma, \sigma') \Gamma$ a minorization is a triple $(m, \lambda, \psi)$ such that the $m$-step transition probabilities of the chain satisfy

$$P^m(\sigma, \sigma') \geq \lambda \psi(\sigma'), \quad \forall \sigma' \in S$$

(27)

where $\lambda > 0$ and $\psi$ is a distribution on $S$. For an illustration of this concept consider the following two chains. Figure 2 depicts a birth-death chain with birth probability $p$ and death probability $1 - p$. The chain is ergodic for $p/(1 - p) < 1$ [10]. However there exists no minorization of this chain because for all $m \Gamma P^m(\sigma, \sigma') = 0$ for $|\sigma - \sigma'| > m$. Now consider the chain depicted in Figure 3. This chain is ergodic, so there exists a stationary distribution $\pi$ [10]. Furthermore $\Gamma P(\sigma, 0) = 1 - p$ for all $\sigma \in S$. Let $\psi(\sigma) = \delta_{\sigma_0}$ and $\lambda = 1 - p$. Then $\Gamma(m, \lambda, \psi)$ satisfy (27). In essence a chain for which there exists a minorization is connected in such a way that one can travel reasonably well between states in a fixed number of steps. This concept will reappear in the next section as well.
The multi-$\gamma$ coupling can be used to simulate any chain for which a minorization is known. Without loss of generality assume the minorization $(m, \lambda, \psi)$ has $m = 1$. Let

$$Q(\sigma, \sigma') = \frac{1}{1 - \lambda} \left( P(\sigma, \sigma') - \lambda \psi(\sigma') \right)$$

be the transition probabilities of a secondary Markov chain. Now suppose

$$\ldots, U_1(-1), U_1(0), U_1(1), \ldots$$

are IID random variables taking values in $U_1$ and $\phi_1 : U_1 \times S \mapsto S$ a function such that

$$\Pr \{ \phi_1(U_1(i), \sigma) = \sigma' \} = Q(\sigma, \sigma').$$

Next let $\ldots, U_2(-1), U_2(0), U_2(1), \ldots$ be IID random variables distributed according to $\psi$. Introduce yet another sequence of IID random variables

$$\ldots, U_3(-1), U_3(0), U_3(1), \ldots$$

taking values in $\{1, 2\}$ and such that $\Pr\{U_3(i) = 2\} = \lambda$. Finally let $U(i) = (U_1(i), U_2(i), U_3(i))$. Define the coupling

$$\phi(U(i), \sigma) = \begin{cases} 
\phi_1(U_1(i), \sigma) & \text{if } U_3(i) = 1 \\
U_2(i) & \text{otherwise.}
\end{cases}$$

Note that this defines a valid coupling because $\Pr\{\phi(U(i), \sigma) = \sigma'\} = P(\sigma, \sigma')$ and the probability that $\phi(U(i), \cdot)$ is a constant function is at least $\lambda$. This leads to the following algorithm for
generating a sample from the stationary distribution of a chain with a minorization.

1. Let $T = -1$.

2. Let $t = T$.

3. If $U_3(t) = 1$ increment $t$ by one and return to step 3 unless $t = 0$. If $t = 0$ double $T$ and return to step 2.

4. Let $X = U_2(t)$.

5. Increment $t$ by one. If $t = 0$ go to step 6 otherwise assign $X$ the value of $\phi(U(i), X)$.

6. Output $X$.

Note that the state space of the chain may be countably infinite and even uncountably infinite [15]. However the method is efficient only if one can sample efficiently from $Q(\cdot, \cdot)$ and $\psi(\cdot)$.

3.4 Run-Time Bounds for Monotone Coupling

One can ask whether the use of CFTP requires more, fewer or the same number of steps as simply running the chain forward long enough that the initialization bias is negligible. The answer is "roughly the same" [16]. To be precise one needs a notion of how close a chain is to steady-state after a fixed number of steps $m$. Define

$$d(m) = \max_{\sigma, \sigma'} \|P^m(\sigma \cdot) - P^m(\sigma' \cdot)\|. \quad (33)$$

The mixing time of the chain $T_{\text{mix}}$ is the smallest $T$ such that $d(T) \leq 1/e$. For a monotone coupling let $T_\phi$ be the smallest $T$ such that

$$\phi\left(U(-1), \phi\left(U(-2), \ldots, \phi\left(U(T), \sigma \right), \ldots \right)\right) = \phi\left(U(-1), \phi\left(U(-2), \ldots, \phi\left(U(T), \sigma' \right), \ldots \right)\right). \quad (34)$$
The goal here is to relate $ET_s$ and $T_{\text{mix}}$.

First, one needs to demonstrate the submultiplicity of $\tilde{d}$.

**Proposition 2.** For any positive integers $m_1$ and $m_2$,

\[ \tilde{d}(m_1 + m_2) \leq \tilde{d}(m_1)\tilde{d}(m_2). \] (35)

The proposition is stated without proof in [1]. One can prove the proposition using a coupling argument. One part of the argument requires the following lemma [12].

**Lemma 1.** Let $X_1$ and $X_2$ be two random variables taking values in the same space with distributions $\mu_1$ and $\mu_2$. Then,

\[ \|\mu_1 - \mu_2\| \leq \Pr\{X_1 \neq X_2\}. \] (36)

**Proof.** For any subset $A$,

\[
\Pr\{X_1 \neq X_2\} \geq \Pr\{X_1 \in A, X_1 \neq X_2\} - \Pr\{X_2 \in A, X_1 \neq X_2\}
= \Pr\{X_1 \in A, X_1 \neq X_2\} - \Pr\{X_2 \in A, X_1 \neq X_2\} + \\
\Pr\{X_1 \in A, X_1 = X_2\} - \Pr\{X_2 \in A, X_1 = X_2\}
= \mu_1(A) - \mu_2(A). \]

The foundation of the proof of Proposition 2 requires a $\gamma$-coupling construction [12]. A $\gamma$-coupling between two discrete random variables $X_1$ and $X_2$ with marginal distributions $\mu_1$ and $\mu_2$
is a joint distribution \( \mu \) defined by

\[
\mu(x_1, x_2) = \frac{1}{1 - \gamma} \left( \mu_1(x_1) - \min(\mu_1(x_1), \mu_2(x_1)) \right) \left( \mu_2(x_2) - \min(\mu_1(x_2), \mu_2(x_2)) \right) + \min(\mu_1(x_1), \mu_2(x_2)) \delta_{x_1, x_2}
\]

(38)

where \( \gamma = \sum_{x_1, x_2} \min(\mu_1(x_1), \mu_2(x_2)) \delta_{x_1, x_2} \). Note that \( \sum_{x_2} \mu(x_1, x_2) = \mu_1(x_1) \) and \( \sum_{x_1} \mu(x_1, x_2) = \mu_2(x_2) \). Furthermore,

\[
||\mu_1 - \mu_2|| = \frac{1}{2} \sum_x |\mu_1(x) - \mu_2(x)|
\]

\[
= \frac{1}{2} \sum_x (\mu_1(x) - \min(\mu_1(x), \mu_2(x))) + \frac{1}{2} \sum_x (\mu_2(x) - \min(\mu_1(x), \mu_2(x)))
\]

\[
= 1 - \sum_x \min(\mu_1(x), \mu_2(x))
\]

\[
= 1 - \sum_x \mu(x, x)
\]

\[
= \Pr\{X_1 \neq X_2\}
\]

(39)

where \( X_1 \) and \( X_2 \) are jointly distributed according to \( \mu \).

Proof of Proposition 2. For \( \sigma, \sigma' \in \mathcal{S} \) let \( X_\sigma(m_1) \) and \( X_{\sigma'}(m_1) \) be jointly distributed according to the \( \gamma \)-coupling between \( P^{m_1}(\sigma, \cdot) \) and \( P^{m_1}(\sigma', \cdot) \) and let \( X_\sigma(m_2) \) and \( X_{\sigma'}(m_2) \) be jointly distributed according to the \( \gamma \)-coupling between \( P^{m_2}(\sigma, \cdot) \) and \( P^{m_2}(\sigma', \cdot) \). For \( \gamma \)-coupling between \( P^{m_1}(\sigma, \cdot) \) and \( P^{m_1}(\sigma', \cdot) \) and let \( X_\sigma(m_1 + m_2) \) and \( X_{\sigma'}(m_1 + m_2) \) be jointly distributed according to the \( \gamma \)-coupling between \( P^{m_1}(\sigma, \cdot) \) and \( P^{m_1}(\sigma', \cdot) \) and let \( X_\sigma(m_1 + m_2) \) and \( X_{\sigma'}(m_1 + m_2) \) be jointly distributed according to the \( \gamma \)-coupling between \( P^{m_2}(\sigma, \cdot) \) and \( P^{m_2}(\sigma', \cdot) \).
Let \( d_2 \) be jointly distributed according to the concatenation of the two couplings. Then

\[
\tilde{d}(m_1)\tilde{d}(m_2) = \max_{\sigma_1, \sigma_2 \in S} \Pr\{X_{\sigma_1}(m_1) \neq X_{\sigma_2}(m_1)\} \max_{\sigma_3, \sigma_4 \in S} \Pr\{X_{\sigma_3}(m_2) \neq X_{\sigma_4}(m_2)\}

= \max_{\sigma_1 \neq \sigma_2} \Pr\{X_{\sigma_1}(m_1) \neq X_{\sigma_2}(m_1)\} \max_{\sigma_3 \neq \sigma_4} \Pr\{X_{\sigma_3}(m_2) \neq X_{\sigma_4}(m_2)\}

\geq \max_{\sigma_1 \neq \sigma_2} \sum_{\sigma_3 \neq \sigma_4} \Pr\{X_{\sigma_1}(m_1) = \sigma_3, X_{\sigma_2}(m_1) = \sigma_4\} \Pr\{X_{\sigma_3}(m_2) \neq X_{\sigma_4}(m_2)\} \tag{40}

\geq \tilde{d}(m_1 + m_2)

where the last inequality follows from Lemma 1. \( \square \)

Propp and Wilson prove the following theorem relating \( \tilde{d} \) and \( T_s \) [16]. A key step in the proof is that \( T_s \) is distributed according to \( T^* \Gamma \) the smallest \( m \) such that the value of the forward run chain at time \( m \Gamma = \phi(U(m - 1), \phi(U(m - 2), \cdots, \phi(U(0), \cdots) \cdots) \) is a constant.

**Theorem 4.** Let \( \tau \) be the length of the longest totally ordered subset in the partially ordered state space \( S \). Then

\[
\frac{\Pr\{T_s > m\}}{\tau} \leq \tilde{d}(m). \tag{41}
\]

**Proof.** For \( \sigma \in S \Gamma \) let \( h(\sigma) \) denote the length of the longest chain whose last element is \( \sigma \). Let

\[
X_\sigma(m) = \phi(U(m - 1), \phi(U(m - 2), \cdots, \phi(U(0), \sigma), \cdots)
\]

and

\[
X_{\sigma}(m) = \phi(U(m - 1), \phi(U(m - 2), \cdots, \phi(U(0), \sigma), \cdots)
\]
2), \cdots, \phi(U(0), \sigma) \) \). If \( X_\sigma(m) \neq X_\sigma(m) \) then \( h(X_\sigma(m)) + 1 \leq h(X_\sigma(m)) \). So

\[
\Pr\{T_* > m\} = \Pr\{T^* > m\} \\
= \Pr\{X_\sigma(m) \neq X_\sigma(m)\} \\
\leq E\left[h(X_\sigma(m)) - h(X_\sigma(m))\right] \\
= \sum_\sigma P^m(\sigma, \sigma) h(\sigma) - \sum_\sigma P^m(\sigma, \sigma) h(\sigma) \tag{42} \\
\leq \tau \sum_{P^m(\sigma, \sigma) \geq P^m(\sigma, \sigma)} (P^m(\sigma, \sigma) - P^m(\sigma, \sigma)) \\
\leq \tau \|P^m(\sigma, \cdot) - P^m(\sigma, \cdot)\| \\
\leq \tilde{d}(m) \tau.
\]

\[\square\]

Now fix \( M = \left[ T^{\frac{1}{1 + \log \tau}} \right] \). Then \( \tilde{d}(M) \leq 1/e \tau \) since \( \tilde{d} \) is submultiplicative. Furthermore,

\[
ET_* = ET^* \\
\leq M + M \Pr\{T^* > M\} + M \Pr\{T^* > 2M\} + \cdots \tag{43} \\
\leq M + M \Pr\{T^* > M\} + M \left(\Pr\{T^* > M\}\right)^2 + M \left(\Pr\{T^* > M\}\right)^3 + \cdots \\
\leq \frac{M}{1 - \frac{1}{e}} \\
= \frac{e}{e - 1} T_{\text{mix}} (1 + \log \tau)
\]

where the second inequality follows from the fact that the probability of coupling in an interval of length \( m_1 + m_2 \) is at least that of coupling in independent intervals of lengths \( m_1 \) and \( m_2 \). Thus the expected value of \( T_* \) is a constant multiple of the mixing time for Markov chains where the length of the longest totally ordered subset is fixed.
4 Run-Time Bounds for Forward Simulation

4.1 Inequalities

Although CFTP will allow one to generate an exact sample from the stationary distribution of a Markov chain, the method is impractical if the state space is large and no special structure (e.g., monotonicity) can be taken advantage of. In such situations one may want to simply run the Markov chain, starting in an arbitrary state, forward long enough that the initialization bias is negligible. For reversible Markov chains, Diaconis and Stroock derive bounds expressed in terms of geometrical properties of the chain on the variational distance between the stationary distribution and the distribution after a finite number of steps [7]. These bounds exploit the Laplacian-like structure of $P$ that exists under these conditions. Specifically, turn $P$ into a linear operator on the space of functions on the state space by letting $(Pf)(\sigma) = \sum_{\sigma'} P(\sigma, \sigma') f(\sigma')$ for a function $f$ on the state space. Furthermore, with respect to the $L^2(\pi)$ inner product defined by $(f, g) = \sum_{\sigma} \pi(\sigma) f(\sigma) g(\sigma) \Gamma P$ is self-adjoint (i.e., $(f, P g) = (P f, g)$) because the chain is reversible. Thus, the operator $L = I - P$ has many characteristics like that of a Laplacian and these are exploited in the subsequent bounds.

The first step in establishing bounds in terms of the geometric properties of the chain is to relate the distance from stationarity to the eigenvalues $\Gamma 1 = \beta_0 > \beta_1 \geq \cdots \geq \beta_{|S|-1} \geq 1$ of the operator $P$. The expressions are actually easier to write for the associated continuous-time Markov process with unit transition intensity [4]. The probability that the associated continuous-time process transitions from state $\sigma$ to $\sigma'$ over an interval of length $t$ is given by

$$P_t(\sigma, \sigma') = \sum_{m=0}^\infty \frac{e^{-t}t^m}{m!} P^m(\sigma, \sigma'). \quad (44)$$
The following proposition is presented in [7]. The proof is provided here for the continuous-time case. The discrete-time proof is very similar and is provided in [7].

**Proposition 3.** For an ergodic and reversible Markov chain with transition probabilities \( P(\sigma, \sigma') \), stationary distribution \( \pi(\sigma) \), and eigenvalues \( 1 = \beta_0 > \beta_1 \geq \cdots \geq \beta_{|S|-1} > -1 \),

\[
4||P^m(\sigma, \cdot) - \pi(\cdot)|| \leq \frac{1}{\pi(\sigma)} \beta_s 2^m \quad \text{for } \beta_s = \max(\beta_1, |\beta_{|S|-1}|) \\
\frac{4 ||P_t(\sigma, \cdot) - \pi(\cdot)||}{4} \leq \frac{1}{\pi(\sigma)} e^{-2(1-\beta_t)} t. 
\]

**Proof.**

\[
4||P_t(\sigma, \cdot) - \pi||^2 = \left( \sum_{\sigma'} \left| P_t(\sigma, \sigma') - \pi(\sigma') \right| \right)^2 \\
= \left( \sum_{\sigma'} \sqrt{\frac{\pi(\sigma')}{\pi(\cdot)} |P_t(\sigma, \sigma') - \pi(\sigma')|} \right)^2 \\
\leq \sum_{\sigma'} \frac{1}{\pi(\sigma')} \left( (P_t(\sigma, \sigma'))^2 - 2P_t(\sigma, \sigma')\pi(\sigma') + (\pi(\sigma'))^2 \right) \\
\]

where the inequality follows from applying Cauchy-Schwartz. Now note that

\[
\sum_{\sigma'} \frac{(P_t(\sigma, \sigma'))^2}{\pi(\sigma')} = \sum_{\sigma'} \frac{1}{\pi(\sigma')} \left( e^{-t} \sum_{m=0}^{\infty} \frac{1}{m!} P^m(\sigma, \sigma') \right)^2 \\
= \sum_{\sigma'} \frac{1}{\pi(\sigma')} e^{-2t} \sum_{l,m=0}^{\infty} \frac{l^m + m}{l! m!} P^l(\sigma, \sigma') P^m(\sigma, \sigma') \\
= e^{-2t} \sum_{l,m=0}^{\infty} \frac{l^m + m}{l! m!} \sum_{\sigma'} \frac{1}{\pi(\sigma')} P^l(\sigma, \sigma') P^m(\sigma, \sigma') \\
= e^{-2t} \sum_{l,m=0}^{\infty} \frac{l^m + m}{l! m!} \frac{P^{l+m}(\sigma, \sigma)}{\pi(\sigma)}
\]

where the last equality uses reversibility of the chain.

Let the \(|S| \times |S|\) diagonal matrix \( D \) have diagonal entries equal to \( \sqrt{\pi(\sigma)} \). Viewing \( P \) as a matrix indexed by \((\sigma, \sigma') \in S \times S\) with entries \( P(\sigma, \sigma') \) one observes that \( DP D^{-1} \) is symmetric
because the chain is reversible. Let \( O \) be an orthogonal matrix such that \( O^TDPD^{-1}O = B \) where \( B \) is the diagonal matrix with the eigenvalues of \( P \) on the diagonal. Suppose \( \bar{\sigma} \) marks the entry of \( B \) such that \( B_{\bar{\sigma}\bar{\sigma}} = 1 \). Then the \( \bar{\sigma} \) column of \( O \) is \( D \) times the vector of ones: \( \left( \cdots \sqrt{\pi(\sigma)} \cdots \right)^T \).

So \( \Gamma P^{l+m} = D^{-1}OB^{l+m}O^T \Delta \Gamma \) and

\[
P^{l+m}(\sigma, \sigma) = \pi(\sigma) + \sum_{\sigma' \neq \bar{\sigma}} B^{l+m}_{\sigma'\sigma} O^{2}_{\sigma\sigma'},
\]

Continuing (47)\(\Gamma\)

\[
\sum_{\sigma'} \frac{(P_l(\sigma, \sigma'))^2}{\pi(\sigma')} = e^{-2t} \sum_{l,m} \frac{t^{l+m} \pi(\sigma)}{l!m! \pi(\sigma)} + e^{-2\beta_1} \sum_{\sigma' \neq \bar{\sigma}} B^{l+m}_{\sigma'\sigma} O^{2}_{\sigma\sigma'} \\
= 1 + e^{-2t} \sum_{\sigma' \neq \bar{\sigma}} \sum_{l,m} (tB_{\sigma'\sigma})^{l+m} \frac{O^{2}_{\sigma\sigma'}}{l!m! \pi(\sigma)} \\
= 1 + e^{-2t} \sum_{\sigma' \neq \bar{\sigma}} \frac{e^{2tB_{\sigma'\sigma}} O^{2}_{\sigma\sigma'}}{\pi(\sigma)} \\
\leq 1 + \frac{e^{-2(1-\beta_1)}(1 - \pi(\sigma))}{\pi(\sigma)}
\]

\(\square\) since \( B_{\sigma'\sigma'} \leq \beta_1 \Gamma \) and \( \sum_{\sigma' \neq \bar{\sigma}} O^{2}_{\sigma\sigma'} = 1 - O^{2}_{\bar{\sigma}} = 1 - \pi(\sigma) \).

One can verify using standard minimization arguments that

\[
1 - \beta_1 = \inf \left\{ \frac{\langle f, Lf \rangle}{\text{Var}(f)} : f \text{ is not constant} \right\}.
\]
This characterization is very useful if one writes \( \text{Var}(f) \) and \((f, Lf)\) in a particular form. Specifically,

\[
\text{Var}(f) = \sum_{\sigma} f^2(\sigma)\pi(\sigma) - \left( \sum_{\sigma} f(\sigma)\pi(\sigma) \right)^2
\]

\[
= \frac{1}{2} \sum_{\sigma} f^2(\sigma)\pi(\sigma) - \sum_{\sigma, \sigma'} f(\sigma)f(\sigma')\pi(\sigma)\pi(\sigma') + \frac{1}{2} \sum_{\sigma'} f^2(\sigma')\pi(\sigma')
\]

\[
= \frac{1}{2} \sum_{\sigma, \sigma'} \left( f(\sigma) - f(\sigma') \right)^2 \pi(\sigma)\pi(\sigma').
\]

Likewise,

\[
(f, Lf) = \sum_{\sigma, \sigma'} \pi(\sigma) f(\sigma) \left( \delta_{\sigma, \sigma'} - P(\sigma, \sigma') \right) f(\sigma')
\]

\[
= \sum_{\sigma} f^2(\sigma)\pi(\sigma) - \sum_{\sigma, \sigma'} f(\sigma)f(\sigma')R(\sigma, \sigma')
\]

\[
= \sum_{\sigma, \sigma'} f^2(\sigma)\pi(\sigma)P(\sigma, \sigma') - \sum_{\sigma, \sigma'} f(\sigma)f(\sigma')R(\sigma, \sigma')
\]

\[
= \frac{1}{2} \sum_{\sigma, \sigma'} f^2(\sigma)R(\sigma, \sigma') + \frac{1}{2} \sum_{\sigma, \sigma'} f^2(\sigma')R(\sigma, \sigma') - \sum_{\sigma, \sigma'} f(\sigma)f(\sigma')R(\sigma, \sigma')
\]

\[
= \frac{1}{2} \sum_{\sigma, \sigma'} \left( f(\sigma) - f(\sigma') \right)^2 R(\sigma, \sigma')
\]

\[
= \mathcal{E}(f, f)
\]

where \( R(\sigma, \sigma') = \pi(\sigma)P(\sigma, \sigma') \) the last equality is a definition, and the fourth equality follows from reversibility.

The geometry in the bounds is introduced via a graph whose vertices are labeled by the states of the chain and whose edges consist of those pairs \((\sigma, \sigma')\) such that \( R(\sigma, \sigma') > 0 \). For each pair of vertices \( \sigma, \sigma' \in S \) let \( \gamma_{\sigma\sigma'} \) denote a path from \( \sigma \) to \( \sigma' \) (a path is a walk on the graph with no repeated edges). Let \( \Gamma \) denote a collection of paths for each pair of states. For a path \( \gamma_{\sigma\sigma'} \in \Gamma \) let the length of the path be given by

\[
|\gamma_{\sigma\sigma'}|_R = \sum_{e \in \gamma_{\sigma\sigma'}} \frac{1}{R(e)}
\]

(53)
where the sum is over all edges $e$ in the path. Diaconis and Stroock give the following proposition and proof to bound $\beta_1$ [7].

**Proposition 4.** The second largest eigenvalue of $P$ satisfies

$$\beta_1 \leq 1 - \frac{1}{\kappa} \quad (54)$$

where

$$\kappa = \max_e \sum_{\gamma_{\sigma\sigma'} \ni e} |\gamma_{\sigma\sigma'}| R \pi(\sigma) \pi(\sigma'). \quad (55)$$

**Proof.**

$$\text{Var}(f) = \frac{1}{2} \sum_{\sigma, \sigma'} (f(\sigma) - f(\sigma'))^2 \pi(\sigma) \pi(\sigma')$$

$$= \frac{1}{2} \sum_{\sigma, \sigma'} \left( \sum_{e \in \gamma_{\sigma\sigma'}} \sqrt{\frac{R(e)}{R(e)}} f(e) \right)^2 \pi(\sigma) \pi(\sigma')$$

$$\leq \frac{1}{2} \sum_{\sigma, \sigma'} |\gamma_{\sigma\sigma'}| R \pi(\sigma) \pi(\sigma') \sum_{e \in \gamma_{\sigma\sigma'}} R(e) f^2(e)$$

$$= \frac{1}{2} \sum_e R(e) f^2(e) \sum_{\gamma_{\sigma\sigma'} \ni e} |\gamma_{\sigma\sigma'}| R \pi(\sigma) \pi(\sigma')$$

$$\leq \kappa \mathcal{E}(f; f) \quad (56)$$

where $f(e) = f(\sigma) - f(\sigma')$ for $e = (\sigma, \sigma')\Gamma$ and the first inequality follows from an application of Cauchy-Schwartz. \qed

Proposition 4 is intriguing because it bounds $\beta_1$ in terms of paths on the graph of the chain. One can derive other bounds in terms of similar quantities. In particular consider the geometric...
quantity

\[ \eta = \max_e \frac{1}{R(e)} \sum_{\gamma_{\sigma\sigma'} \geq e} \pi(\sigma) \pi(\sigma'). \]  \hspace{1cm} (57)

Then the following proposition proved in [7] bounds \( \beta_1 \) in terms of \( \eta \).

**Proposition 5.** The second largest eigenvalue, \( \beta_1 \), of \( P \) satisfies

\[ \beta_1 \leq 1 - \frac{1}{8\eta^2}. \]  \hspace{1cm} (58)

It is this second bound which has actually proven useful in the context of statistical mechanics.

To illustrate the application of this bound \( \eta \) will be subsequently bounded for the Metropolis chain applied to the subgraphs-world process as described in Section 2.

### 4.2 Bounds for the Subgraphs-World Process

Jerrum and Sinclair bound \( \eta \) for the random update Metropolis chain applied to the subgraphs-world process described in Section 2 [11]. Their construction of paths between any two configurations relies on viewing a particular configuration as a graph. The following two facts concerning any graph play an important role in the construction of the paths [22]. The first is that the number of vertices of odd degree is even because the sum of the degrees of all vertices is twice the number of edges and is even. The second is that any connected graph will all vertices of even degree can be covered by a single closed circuit (walk with no repeated edges starting and stopping in the same place). For configurations \( \sigma \) and \( \sigma' \), let \( \Delta = \sigma \oplus \sigma' \) denote the configuration with an edge between two neighboring sites if that edge is in one of \( \sigma \) and \( \sigma' \) and not the other. Let \( 2l \) be the number of odd degree vertices in \( \Delta \). Decompose \( \Delta \) into \( C_1, \ldots, C_l \) open trails (walks with no repeated edges and different starting and stopping points) and \( C_{l+1}, \ldots, C_r \) closed circuits. That one can do this
follows from the subsequent proposition proved in the appendix. Then the path from \( \sigma \) to \( \sigma' \Gamma \gamma_{\sigma \sigma'} \Gamma \) consists of the sequence of configurations \( \Gamma \) starting with \( \sigma \Gamma \) in which one removes or adds edges in sequence along the paths and circuits \( C_1, \ldots, C_r \). The order in which these paths and circuits are traversed is assumed to be fixed for each \( C_i \).

**Proposition 6.** A graph \( \Delta \) with \( 2l \) vertices of odd degree can be decomposed into \( C_1, \ldots, C_l \) paths and \( C_{l+1}, \ldots, C_r \) circuits.

The bound on \( \eta \) in [11] makes use of an injective map \( \iota(\tau, \tau') : \{ \gamma_{\sigma \sigma'} \ni (\tau, \tau') \} \to S \) for each transition between states \( \tau \) and \( \tau' \) in the state space \( S \). The map is defined by the relation
\[
\iota(\tau, \tau')(\gamma_{\sigma \sigma'}) = \sigma \oplus \sigma' \oplus (\tau \cup \tau') \Gamma
\]
which consists of edges that have been deleted up to and including the current transition \( \Gamma \) those that need to be added after the current transition \( \Gamma \) and those in both \( \sigma \) and \( \sigma' \). To see that \( \iota \) is injective, note that \((\sigma \oplus \sigma' \oplus (\tau \cup \tau')) \oplus (\tau \cup \tau') = \sigma \oplus \sigma' \). Furthermore, \( \sigma \) and \( \sigma' \) can be reconstructed from \( \sigma \oplus \sigma' \Gamma \) knowledge of the states \( \tau \) and \( \tau' \) before and after the transition \( \Gamma \) and the unique sequence of transitions defined by the paths and circuits associated with \( \sigma \oplus \sigma' \). The function \( \iota \) also satisfies the following bound:
\[
\pi(\iota(\tau, \tau')(\gamma_{\sigma \sigma'})) R(\tau, \tau') \geq \frac{1}{2k} \mu^4 \pi(\sigma) \pi(\sigma')
\] (59)
where \( k \) is the number of neighboring pairs of states.

To verify (59) note that (59) holds if and only if
\[
\prod_{(i,j) \in \tilde{\tau}} w_{ij} \min \left( \prod_{(i,j) \in \tau} w_{ij}, \prod_{(i,j) \in \tau'} w_{ij} \right) \geq \prod_{(i,j) \in \sigma} w_{ij} \prod_{(i,j) \in \sigma'} w_{ij}
\] (60)
\[
|\text{odd}(\tilde{\tau})| + |\text{odd}(\tau)| - |\text{odd}(\sigma)| - |\text{odd}(\sigma')| \leq 4
\] (61)
\[
|\text{odd}(\tilde{\tau})| + |\text{odd}(\tau')| - |\text{odd}(\sigma)| - |\text{odd}(\sigma')| \leq 4
\] (62)
where \( \bar{\tau} = \lambda_{(\tau,\tau')}(\sigma,\sigma') \). The equivalence between the sets of inequalities follows from the definition of \( \pi \) in (9) and the fact that \( R(\tau,\tau') = \min(\pi(\tau), \pi(\tau'))/2k \) for the Metropolis chain applied to the subgraphs-world process. Inequality (60) holds because

\[
\prod_{(i,j) \in \bar{\tau}} w_{ij} \prod_{(i,j) \in \tau \cup \tau'} w_{ij} = \prod_{(i,j) \in \bar{\tau} \cup \tau'} w_{ij} \prod_{(i,j) \in \bar{\tau} \cap (\tau \cup \tau')} w_{ij} = \prod_{(i,j) \in \sigma} w_{ij} \prod_{(i,j) \in \sigma'} w_{ij}
\]

and

\[
\prod_{(i,j) \in \tau \cup \tau'} w_{ij} \leq \min \left( \prod_{(i,j) \in \tau} w_{ij}, \prod_{(i,j) \in \tau'} w_{ij} \right),
\]

where the inequality follows from \( w_{ij} \leq 1 \). To verify (61) let \( i \) be a vertex in the graph and consider the following two cases for which \( 1_{\text{odd}(\bar{\tau})}(i) + 1_{\text{odd}(\tau)}(i) - 1_{\text{odd}(\sigma)}(i) - 1_{\text{odd}(\sigma')}(i) > 0 \) where \( 1_{A}(\cdot) \) is the indicator function of a set \( A \).

1. \( 1_{\text{odd}(\bar{\tau})}(i) + 1_{\text{odd}(\tau)}(i) - 1_{\text{odd}(\sigma)}(i) - 1_{\text{odd}(\sigma')}(i) = 1 \). Recall that \( \sigma \oplus \sigma' = \bar{\tau} \oplus (\tau \cup \tau') \) and \( \tau \) and \( \tau' \) differ at exactly one spot: the edge being added or deleted. At any vertex \( j \) not an endpoint of the edge being modified, the sum of the degrees of \( j \) in \( \bar{\tau} \) and \( \tau \) must differ from the sum of the degrees of \( j \) in \( \sigma \) and \( \sigma' \) by an even number. Thus \( 1_{\text{odd}(\bar{\tau})}(j) + 1_{\text{odd}(\tau)}(j) - 1_{\text{odd}(\sigma)}(j) - 1_{\text{odd}(\sigma')}(j) \) is not odd. The implication is that \( i \) must be a vertex of the edge being added or deleted.

2. \( 1_{\text{odd}(\bar{\tau})}(i) + 1_{\text{odd}(\tau)}(i) - 1_{\text{odd}(\sigma)}(i) - 1_{\text{odd}(\sigma')}(i) = 2 \). Then \( i \) is even in both \( \sigma \) and \( \sigma' \) and odd in both \( \tau \) and \( \bar{\tau} \). A vertex which is even in both \( \sigma \) and \( \sigma' \) will also be even in the intermediate configuration \( \tau \) unless \( i \) is the starting vertex of the closed circuit currently
being traversed or \( i \) is the starting vertex of the edge currently being added or deleted.

Thus

\[ |\text{odd}(\tilde{\tau})| + |\text{odd}(\tau)| - |\text{odd}(\sigma)| - |\text{odd}(\sigma')| = \sum_i 1_{\text{odd}(\tilde{\tau})(i)} + 1_{\text{odd}(\tau)(i)} - 1_{\text{odd}(\sigma)(i)} - 1_{\text{odd}(\sigma')(i)} \leq 5 \quad (65) \]

since there are only three positive terms in the right-most sum and they can only take on the values two, one, and either one or two respectively. Furthermore the sum is even because the number of odd-degree vertices in a graph is even. So (61) holds. By symmetry (62) holds as well. The conclusion is that (59) holds.

These arguments produce the following bound on \( \eta \):

\[ \eta = \max_{[\tau,\tau']} \frac{1}{R(\tau,\tau')} \sum_{\gamma_{\sigma,\sigma'} \in [\tau,\tau']} \pi(\sigma)\pi(\sigma') \]

\[ \leq \max_{[\tau,\tau']} \frac{1}{R(\tau,\tau')} \sum_{\gamma_{\sigma,\sigma'} \in [\tau,\tau']} R(\tau,\tau')\pi(i_{\tau,\tau'}(\gamma_{\sigma,\sigma'}))2k\mu^{-1} \]

\[ \leq 2k\mu^{-1} \quad (66) \]

where the first inequality follows from (59) and the second from the injectivity of \( i \). Proposition 5 implies that

\[ \beta_1 \leq 1 - \frac{\mu^8}{32k^2}. \quad (67) \]

Furthermore all the eigenvalues are non-negative because the holding probability in the Metropolis chain is greater than or equal to one half. (A diagonally dominant matrix is positive semi-definite). Lastly \( \pi(0) \geq \pi(\sigma) \) for all configurations \( \sigma \); so \( \pi(0) \geq 2^{-k} \). Thus the bound on \( \beta_1 \) in (67) implies a bound on the distance that the current state of the Metropolis chain started in state 0 is from
steady-state via Proposition 3.

4.3 Countable Spaces and Minorization

One can imagine trying to derive bounds that for every $\varepsilon > 0$ will provide one with a time $T(\varepsilon)$ such that

$$
\|P^{T(\varepsilon)}(\sigma, \cdot) - \pi(\cdot)\| < \varepsilon \quad \forall \sigma \in S
$$

(68)

for a countably infinite state space $S$. As it turns out, the class of processes for which there exist such times $T(\varepsilon)$ is limited. The following proposition is stated in [3].

**Proposition 7.** Let $P(\sigma, \sigma')$ be the transition probabilities of an ergodic Markov chain on a countably infinite state space. Suppose that for every $\varepsilon > 0$, there exists a time $T(\varepsilon)$ such that (68) holds. Then, there exists a minorization of $P$, i.e. a triplet such that (27) holds.

**Proof.** Let $\pi$ be the stationary distribution and let $\sigma^*$ be such that $\pi(\sigma^*) > 0$. Set $\varepsilon = \pi(\sigma^*)/4$. Then

$$
\|P^{T(\varepsilon)}(\sigma, \cdot) - \pi(\cdot)\| = \frac{1}{2} \sum_{\sigma'} |P^{T(\varepsilon)}(\sigma, \sigma') - \pi(\sigma')| < \varepsilon \quad \forall \sigma \in S.
$$

(69)

So $|P^{T(\varepsilon)}(\sigma, \sigma^*) - \pi(\sigma^*)| < 2\varepsilon = \pi(\sigma^*)/2$ and $P^{T(\varepsilon)}(\sigma, \sigma^*) \geq \pi(\sigma^*)/2$ for all $\sigma$. Let $\psi(\sigma) = \delta_{\sigma, \sigma^*}$. Then

$$
P^{T(\varepsilon)}(\sigma, \sigma') \geq \frac{\pi(\sigma^*)}{2} \psi(\sigma') \quad \forall \sigma \in S,
$$

(70)

and $(T(\varepsilon), \pi(\sigma^*)/2, \psi)$ is a minorization. □
Furthermore, one can use a known minorization to derive a time function \( T(\varepsilon) \). The following proposition illustrates how \( [3] \). The proof presented here makes use of the multi-\( \gamma \) coupler in Section 3.3.

**Proposition 8.** Suppose \((m, \lambda, \psi)\) is a minorization of \( P(\sigma, \sigma') \), the transition probabilities of an ergodic Markov chain with stationary distribution \( \pi \) on a countably infinite state space. Then,

\[
\left\| P^\varepsilon(\sigma, \cdot) - \pi(\cdot) \right\| \leq (1 - \lambda) \left\lfloor \frac{\varepsilon}{m} \right\rfloor.
\]

**Proof.** Let \( \ldots, U(-1), U(0), U(1), \ldots \) and \( \phi \) be as in the multi-\( \gamma \) coupling construction of Section 3.3 but for the \( m \)-step transition probabilities \( P^m(\sigma, \sigma') \). Let \( X^* \) be distributed according to \( \pi \Gamma X^*(i) = \phi \left( U(i - 1), \phi \left( U(i - 2), \cdots, \phi( U(0), \sigma) \cdots \right) \right) \Gamma \) and \( X^*(i) = \phi \left( U(i - 1), \phi \left( U(i - 2), \cdots, \phi(U(0), X^*) \cdots \right) \right) \). Then Lemma 1 implies

\[
\left\| P^\varepsilon(\sigma, \cdot) - \pi(\cdot) \right\| \leq \Pr \left\{ X^*(\lfloor l/m \rfloor) \neq X^*(\lfloor l/m \rfloor) \right\}
\leq (1 - \lambda) \left\lfloor \frac{\varepsilon}{m} \right\rfloor.
\]

Thus, one can find deterministic times for countably infinite state spaces that will guarantee negligible initialization bias if and only if there exists a minorization.
5 Stopping Times

5.1 Existence Questions

Section 4 focuses on determining a $T(\varepsilon)$ for every $\varepsilon > 0$ such that

$$\|P_{T(\varepsilon)}(\sigma, \cdot) - \pi(\cdot)\| < \varepsilon \quad \forall \sigma \in S. \quad (73)$$

One can broaden the focus to explore the possibility of whether there exist randomized stopping times $T$ such that the distribution of the state of the chain at time $T$ is equal to $\pi$ or at least arbitrarily close independent of the initial state $\sigma$ and of the transition probabilities $P$. In this context, a randomized stopping rule is a collection of binary-valued (the domain is $\{0, 1\}$) functions $r_t$. The arguments of the $r_t$ are the Markov chain states $X(0), X(1), \ldots, X(t)$ and possibly an independent process $U(0), U(1), \ldots, U(t)$. Furthermore, the $r_t$ are defined in such a way that there exists a $t$ such that $r_t = 1$ with probability one. A stopping time associated with a randomized stopping rule is the smallest time $T$ such that $r_T = 1$. Unfortunately, one can not guarantee the existence of a randomized stopping time with the desired property for state spaces which are countably infinite.

To see this, consider the following counterexample based on arguments in [3]. For $\varepsilon > 0$, suppose $T(\varepsilon)$ is a randomized stopping time such that $\|\Pr(X(T(\varepsilon)) \in \cdot | X(0) = \sigma) - \pi(\cdot)\| < \varepsilon$ independent of the transition probabilities $P(\cdot, \cdot)$ and starting state $\sigma$. Let $P(\sigma, \sigma')$ be the transition probabilities of the birth-death process pictured in Figure 2. Represent the maximum state visited before the stopping time by

$$M = \max_{0 \leq i \leq T(\varepsilon)} X(i). \quad (74)$$
Since $M$ and $X(T(\varepsilon))$ are finite random variables there exists an $m < \infty$ such that

\[ \Pr(X(T(\varepsilon)), M \leq m|X(0) = 0) \geq 2\varepsilon. \]  

(75)

Now create the modified birth-death chain $P_\alpha(\sigma, \sigma')$ by setting

\[ P_\alpha(m + 1, m) = \alpha \]

\[ P_\alpha(m + 1, m + 1) = 1 - P(m + 1, m + 2) - \alpha \]

(76)

\[ P_\alpha(m + 1, m + 2) = P(m + 1, m + 2) \]

where $0 < \alpha \leq P(m + 1, m + 2)$. Let $\pi_\alpha$ denote the stationary distribution of this modified chain.

By local balance one can conclude that for $\sigma \leq m\Gamma$

\[ \pi_\alpha(\sigma) = \left( \frac{1 - p}{p} \right)^{m - \sigma} \frac{\alpha}{p} \pi_\alpha(m + 1) \leq \left( \frac{1 - p}{p} \right)^{m - \sigma} \frac{\alpha}{p}. \]  

(77)

So

\[ \pi_\alpha\{0, \ldots, m\} \leq \frac{\alpha}{p} \frac{1 - \left( \frac{1 - p}{p} \right)^{m + 1}}{1 - \frac{1 - p}{p}}. \]  

(78)

Fix $\alpha$ so that $\pi_\alpha\{0, \ldots, m\} \leq \varepsilon$. If $X_{\alpha}(0), X_{\alpha}(1), \ldots$ represent the states visited by the modified
chain and $M_a = \max_{0 \leq i \leq T(\varepsilon)} X_a(i)$ then

\[
\|\Pr(X_a(T(\varepsilon)) \in \cdot | X_a(0) = 0) - \pi_a(\cdot)\| \geq \|\Pr(X_a(T(\varepsilon)) \leq m | X_a(0) = 0) - \pi_a(0, \ldots, m)\|
\]

\[
= \Pr(X_a(T(\varepsilon)) \leq m | X_a(0) = 0) - \pi_a(0, \ldots, m)
\]

\[
\geq \Pr(X_a(T(\varepsilon)), M_a \leq m | X_a(0) = 0) - \pi_a(0, \ldots, m)
\]

\[
= \Pr(X(T(\varepsilon)), M \leq m | X(0) = 0) - \pi_a(0, \ldots, m)
\]

\[
\geq 2\varepsilon - \varepsilon = \varepsilon.
\]

(79)

Note that this counterexample relies on the state space being infinite. If the state space is restricted to be finite, one can construct randomized stopping times with the desired properties.

### 5.2 Randomized Stopping Time Construction

The goal is to construct a randomized stopping time $T$ such that $\Pr(X(T) \in \cdot) = \pi(\cdot)$ for arbitrary ergodic transition probabilities $P(\sigma, \sigma')$ on finite state spaces. Furthermore, one would like $T$ to be non-trivial in the sense that it does not involve explicitly computing $\pi$. One could use CFTP with independent coupling to generate such a time. Another construction, given in [3], is summarized here. The construction relies on the continuous-time Markov process associated with the chain [4]. The continuous process can be thought of changing state according to $P(\sigma, \sigma')$ at the arrival times of a Poisson process with unit intensity.

Consider a two-state Markov process with $S = \{1, 2\}$. Let $T_1$ be the amount of time first spent in state 1 and $T_2$ in state 2. Then $T_1$ and $T_2$ are exponentially distributed with parameters $P(1, 2)$ and $P(2, 1)$ respectively. Thus

\[
P(T_1 > T_2) = \frac{P(2, 1)}{P(2, 1) + P(1, 2)} = \pi(1).
\]

(80)
The following algorithm uses this fact to construct a stopping time.

1. Let $T_1 = 0$.

2. Simulate an exponential random variable with unit intensity and increment $T_1$ by the exponential random variable’s value.

3. Let $X$ be a sample from $P(1, \cdot)$. If $X = 1$ return to step 2.

4. Let $T_2 = 0$.

5. Simulate an exponential random variable with unit intensity and increment $T_2$ by the exponential random variable’s value.

6. Let $X$ be a sample from $P(2, \cdot)$. If $X = 2$ return to step 5.

7. If $T_1 > T_2$ return state 1; otherwise return state 2.

Note that this algorithm is simply simulating two continuous-time Markov processes and outputting the sample at the time the processes couple. The randomized stopping time $T$ corresponding to this algorithm uses the states of an observed chain $X(0), X(1), \ldots$ and IID exponential random variables to generate a state $X^*$ distributed according to $\pi$. Then $\Gamma T$ is the next smallest time such that $X(T) = X^*$.

One can extend this construction to arbitrary finite state spaces. Partition $S$ into two disjoint subsets $S_0$ and $S_1$. Then $\Gamma \pi(\cdot) = \pi(\cdot|S_0)\pi(S_0) + \pi(\cdot|S_1)\pi(S_1)\Gamma$ where $\pi(\cdot|A)$ denotes the conditional distribution given $A$. So one can generate a sample from $S$ if one can generate samples from $\pi(\cdot|S_0)$ and $\pi(\cdot|S_1)$ and also a Bernoulli random variable with parameter $\pi(S_0)$. To generate samples from $\pi(\cdot|S_0)\Gamma$ partition $S_0$ further into two disjoint subsets $S_{00}$ and $S_{01}$. Then one can generate a sample from $\pi(\cdot|S_0) = \pi(\cdot|S_{00})\pi(S_{00}|S_0) + \pi(\cdot|S_{01})\pi(S_{01}|S_0)$ if one can generate samples from $\pi(\cdot|S_{00})$ and $\pi(\cdot|S_{01})$ and also a Bernoulli random variable with parameter $\pi(S_{01}|S_0)$. Continue partitioning
Figure 4: In the construction of the randomized stopping time, the state space is recursively partitioned into two subsets. The relations among the partitions are illustrated here for a grouping of eight subsets.

sets into two subsets to generate a collection \( \{S_b\} \) indexed by binary decimals \( b \Gamma \) as illustrated in Figure 4. The recursive partitioning should stop at a set that has one element. Then \( \Gamma \) for every \( \sigma \in S \Gamma \) there exists an index \( b_\sigma \) such that \( S_{b_\sigma} = \{\sigma\} \). The distribution \( \pi(\sigma'|S_{b_\sigma}) = \delta_{\sigma\sigma'} \) is easy to sample from.

To sample from a Bernoulli random variable with parameter \( \pi(S_{b0}|S_b) \) for some set \( S_b \), consider the following Markov chain on the state space \( S_b \). This new chain is formed by starting the original chain in a state \( \sigma \in S_b \) and transitioning to the next state in \( S_b \) which the original chain visits. Essentially the sequence of states visited by the new chain is the subsampling of states visited by the original chain which lie in \( S_b \). Then one can verify that the stationary distribution \( \pi_{S_b}(\cdot) \) of this new chain is \( \pi(\cdot|S_b) \) [2]. Let \( P_{S_b} \) denote the transition probabilities of this new chain. Furthermore, let \( T_{S_{b0}}^{S_b} \) be exponentially distributed with parameter \( \sum_{\sigma \in S_{b0}} \pi(\sigma|S_{b0})P_{S_b}(\sigma, S_{b1}) \) and \( T_{S_{b1}}^{S_b} \Gamma \) with
parameter $\sum_{\sigma \in S_0} \pi(\sigma | S_0) P_{S_0}(\sigma, S_{b0})$. Then

$$
\Pr \left\{ T_{S_{b0}}^{S_b} > T_{S_{b0}}^{S_b} \right\} = \frac{\sum_{\sigma \in S_{b0}} \pi(\sigma | S_{b0}) P_{S_0}(\sigma, S_{b1})}{\sum_{\sigma \in S_{b0}} \pi(\sigma | S_{b0}) P_{S_0}(\sigma, S_{b1}) + \sum_{\sigma \in S_{b1}} \pi(\sigma | S_{b1}) P_{S_0}(\sigma, S_{b0})}
$$

$$
= \frac{1}{1 + \frac{\sum_{\sigma \in S_{b0}} \pi(\sigma | S_{b0}) P_{S_0}(\sigma, S_{b1})}{\sum_{\sigma \in S_{b1}} \pi(\sigma | S_{b1}) P_{S_0}(\sigma, S_{b0})}}
$$

$$
= \frac{1}{1 + \frac{\pi(S_{b0})}{\sum_{\sigma \in S_{b0}} \pi(\sigma | S_{b0}) P_{S_0}(\sigma, S_{b0})}}
$$

$$
= \frac{\pi(S_{b0})}{\pi(S_{b0}) + \pi(S_{b1})} = \pi(S_{b0} | S_b)
$$

(81)

where the fourth equality follows from local balance.

Noting that an exponential random variable with parameter $\sum_{\sigma \in S_{b0}} \pi(\sigma | S_{b0}) P(\sigma, S_{b1})$ is distributed like the minimum of exponential random variables with parameters $\pi(\sigma | S_{b0}) P(\sigma, S_{b1})$, one arrives at the following algorithm for generating a sample from $\pi(\cdot | S_b)$.

1. If $S_b = \{\sigma\}$ for some $\sigma$ return $\sigma$.

2. Realize $T_{S_{b0}}^{S_b}$ by performing the following steps assuming that $S_{b0} = \{\sigma_1, \ldots, \sigma_l\}$:

   (a) Let $T_{S_{b0}}^{S_b} = \infty$ and $t_1 = t_2 = \cdots = t_l = 0$.

   (b) Cycle through the states in $S_{b0}$. For each state $\sigma_i$:

      i. Simulate the continuous-time Markov process associated with state space $S$ starting in $\sigma_i$ and and stopping when the process makes a transition to a state $\sigma' \in S_b$.

      ii. Increment $t_i$ by the time it takes to make the first transition.

      iii. If $\sigma' \in S_{b1}$ generate a sample $\sigma''$ from $\pi(\cdot | S_{b0})$ (a recursive call).

      iv. If $\sigma'' = \sigma_i$ let $T_{S_{b0}}^{S_{b1}} = \min \left( T_{S_{b0}}^{S_b}, t_i \right)$.

   (c) If $T_{S_{b0}}^{S_{b1}} < \min(t_1, \ldots, t_i)$ go to step 3; otherwise return to 2b.

3. Realize $T_{S_{b1}}^{S_b}$ in a similar manner.
4. If $T_{S_{b_0}}^S > T_{S_{b_1}}^S$, return a sample from $\pi(\cdot|S_{b_0})$; otherwise return a sample from $\pi(\cdot|S_{b_1})$.

A sample is generated from $\pi$ by running this algorithm on the set $S$. A stopping time can be constructed as in the two-state case.

Unfortunately, the complexity of this algorithm can be quite large for certain chains and choices of partitions. Consider the following example [3]. Let $S = \{1, \ldots, l\}$ and suppose

$$P(\sigma, \sigma') = \begin{cases} \frac{1}{l-1} & \sigma' \neq \sigma \\ 0 & \sigma' = \sigma. \end{cases}$$

(82)

Generate the partitions by letting one subset of each partition be a one-point set. Let $m_\ell$ denote the expected number of steps needed to generate a sample from a subset $S_b$ of size $\ell'$. Write $S_b = S_{b_0} \cup S_{b_1}$ where $|S_{b_1}| = 1$. The algorithm will cycle through each point in $S_{b_0}$ until there is a transition to $S_{b_1}$. When there is a transition to $S_{b_1}$ the algorithm samples from $\pi(\cdot|S_{b_0})$ which is uniform on $S_{b_0}$. If the sample is not equal to the state from which a transition to $S_{b_1}$ occurred, the process is restarted. The expected number of samples needed from $\pi(\cdot|S_{b_0})$ is $1/\pi(\sigma|S_{b_0}) = \ell' - 1$. Thus $m_\ell > m_{\ell-1}(\ell' - 1)$ and $m_1 > (l - 1)!$.

This construction demonstrates the existence of a randomized stopping rule that can produce an exact sample from the stationary distribution of a Markov chain and is different in nature than CFTP with independent coupling. Life CFTP however, the complexity of the algorithm outlined in this section can be quite large. In particular, it is unlikely that this algorithm could be applied to any of the sampling problems from statistical physics discussed in Section 2. The state space is simply too large and there is no easy method for partitioning the space.
6 Conclusion

In principle, one can always generate an exact sample from the stationary distribution of a finite-state Markov chain without computing the distribution first. Algorithms for doing this include the coupling-from-the-past protocol presented in Section 3 and the stopping time algorithm of Section 5. Both of these algorithms are widely applicable but are only efficient when certain structure is present. When that structure is not present, one may still be able to make use of the bounds in Section 4 to determine a bound on the number of steps that a Markov chain needs to be simulated so that initialization bias is negligible. However, the application of these bounds to a particular problem is often difficult.

These approaches leave many open areas for further research. With regards to CFTP, one can consider designing other coupling mechanisms appropriate for simulating models different from the ones already considered. If one would rather make use of a randomized stopping time, one can consider the construction of particular partitions for use with the algorithm in Section 5. One could also construct stopping times that would be more efficient for certain problems. Lastly, alternate bounds might be derived that would allow one to determine when initialization bias is negligible.
Appendix

Proof of Proposition 1. Let \( A_1^* = \{ \sigma \in S : \mu_1(\sigma) > \mu_2(\sigma) \} \) the set of states whose probability is greater under the first distribution and let \( A_2^* = \{ \sigma \in S : \mu_2(\sigma) \geq \mu_1(\sigma) \} \) the set of states whose probability is greater under the second distribution. Then

\[
\|\mu_1 - \mu_2\| = \sup_{A \subseteq S} |\mu_1(A) - \mu_2(A)|
\]

\[
= \max \left( \sup_{A \subseteq S} \mu_1(A) - \mu_2(A), \sup_{A \subseteq S} \mu_2(A) - \mu_1(A) \right)
\]

\[
= \max (\mu_1(A_1^*) - \mu_2(A_1^*), \mu_2(A_2^*) - \mu_1(A_2^*)).
\]

Now one can rewrite

\[
\mu_1(A_1^*) - \mu_2(A_1^*) = \frac{1}{2} \left( \mu_1(A_1^*) - \mu_2(A_1^*) + (1 - \mu_1(A_2^*)) - (1 - \mu_2(A_2^*)) \right)
\]

\[
= \frac{1}{2} \sum_{\sigma \in S} |\mu_1(\sigma) - \mu_2(\sigma)|.
\]

Likewise \( \mu_1(A_2^*) - \mu_2(A_2^*) = 1/2 \sum_{\sigma \in S} |\mu_1(\sigma) - \mu_2(\sigma)|. \) So \( \|\mu_1 - \mu_2\| = 1/2 \sum_{\sigma \in S} |\mu_1(\sigma) - \mu_2(\sigma)|. \)

\[\square\]

Proof of Proposition 6. The proof is by induction. If \( l = 0 \) then \( \Delta \) can be covered by a single circuit [22]. Now suppose the proposition holds for \( l \leq m \). Then if there are \( 2(m + 1) \) odd-degree vertices one can form a path \( C_1 \) between two of these odd degree vertices. The graph formed by removing the edges in \( C_1 \) from \( \Delta \) has \( 2m \) odd-degree vertices which by hypothesis can be covered by \( m \) paths and a set of circuits.

\[\square\]
References


