Statistica Sinica **21** (2011), 1687-1711
doi:http://dx.doi.org/10.5705/ss.2009.282

# CONVERGENCE OF THE EQUI-ENERGY SAMPLER AND ITS APPLICATION TO THE ISING MODEL

Xia Hua and S. C. Kou

*Massachusetts Institute of Technology and Harvard University*

*Abstract:* We provide a complete proof of the convergence of a recently developed sampling algorithm called the equi-energy (EE) sampler (Kou, Zhou and Wong (2006)) in the case that the state space is countable. In a countable state space, each sampling chain in the EE sampler is strongly ergodic a.s. with the desired steady-state distribution. Furthermore, all chains satisfy the individual ergodic property. We apply the EE sampler to the Ising model to test its efficiency, comparing it with the Metropolis algorithm and the parallel tempering algorithm. We observe that the dynamic exponent of the EE sampler is significantly smaller than those of parallel tempering and the Metropolis algorithm, demonstrating its high efficiency.

*Key words and phrases:* Dynamic exponent, ergodic property, Monte Carlo methods, phase transition, steady-state distribution, temperature, transition kernel.

## 1. Introduction

In Monte Carlo simulation and statistical inference problems it is often important to obtain samples from a given distribution. For instance, in statistical physics problems, the distribution of interest is usually the Boltzmann distribution

$$p(x) = \frac{1}{Z(T)} \exp\Big( - \frac{h(x)}{T} \Big), \tag{1.1}$$

where $h(x)$ is the energy of a state $x$, $T$ is the temperature of the system, and $Z(T)$ is a normalizing constant called the partition function. One wants to study how the system behaves as temperature varies. In pure statistical applications, the starting point is usually *one* given distribution, for example, a distribution on a high-dimensional parameter space. However, we can view it as a special case of (1.1) by defining the energy to be the negative log-density function with the implicit temperature $T = 1$. For simple sampling problems, traditional algorithms such as the Metropolis-Hastings (MH) algorithm and Gibbs sampler work. However, if the sampling distribution is multimodal and the modes are far from each other, which is often the case for practical multidimensional problems, it is well known that these simple algorithms can be easily trapped in local modes without being able to escape the high energy barrier. More advanced algorithms,

such as simulated tempering and parallel tempering (the latter is also known as replica exchange) (Geyer (1991); Geyer and Thompson (1995); Marinari and Parisi (1992)), employ multiple Markov chains running at different temperatures and use high temperature chains swapping with the low temperature chains to improve the mixing rate. They tend to work better, but the swapping operation is still not powerful enough when the distributions is highly rugged. The equi-energy (EE) sampler (Kou, Zhou and Wong (2006)) was recently developed to address local trapping by utilizing a new type of move called the equi-energy jump, which aims to move directly between states with similar energy level. A detailed comparison between the EE sampler and parallel tempering was given in Kou, Zhou and Wong (2006). The proof of the ergodicity and convergence of the EE sampler in the original paper, however, is not complete. The first goal of this article is to provide a complete and rigorous proof of the ergodicity in the case that the state space is countable.

To begin, we briefly review the EE sampler. Let $\mathcal{X}$ denote the state space and $\pi(x)$ be the target distribution on $\mathcal{X}$. The corresponding energy function is $h(x) = -\log(\pi(x))$. The EE sampler employs a sequence of energy levels:

$$H_0 < H_1 < H_2 < \cdots < H_K < H_{K+1} = \infty,$$

such that $H_0$ is below the minimum energy. Associated with the energy levels is a sequence of temperatures

$$1 = T_0 < T_1 < T_2 < \cdots < T_K.$$

The EE sampler considers $K+1$ distributions, each indexed by a temperature and an energy truncation. The $i$th distribution $\pi_i$ $(0 \le i \le K)$ is $\pi_i(x) \propto \exp(-h_i(x))$, where $h_i(x) = \max(h(x), H_i)/T_i$. For each $i$, a sampling chain targeting $\pi_i$ is constructed; $\pi_0 = \pi$ is the initial distribution of interest. The state space $\mathcal{X}$ is partitioned according to the energy levels: $\mathcal{X} = \bigcup_{j=0}^{K} D_j$, where $D_j = \{x : h(x) \in [H_j, H_{j+1})\}$. We call the $D_j$ energy rings. For any $x \in \mathcal{X}$, let $I(x)$ denote the partition index such that $I(x) = j$ if $x \in D_j$.

The EE sampler begins from a Metropolis-Hastings (MH) chain $X^K$ targeting $\pi_K$. After an initial burn-in period, the EE sampler starts constructing the $K$th order empirical energy rings $\hat{D}_j^K$, $0 \le j \le K$, where $\hat{D}_j^K$ contains all the samples $X_n^K$ such that $I(X_n^K) = j$. For each $x \in \hat{D}_j^K$, the empirical distribution $F_n^{K,j}(x)$ is defined to be the number of visits of $X^K$ to $x$ up to time $n$, divided by the number of visits of $X^K$ to $D_j$ up to time $n$. After the chain $X^K$ has been running for $N$ steps (for example, $N$ could be 5 times the burn-in period), the EE sampler starts the second highest order chain $X^{K-1}$ targeting $\pi_{K-1}$, while it keeps on running $X^K$ and updating $\hat{D}_j^K$. The chain $X^{K-1}$

is updated by two operations: the MH move and the equi-energy jump. At each update a coin is flipped: with probability $1 - p_{ee}$ the current state $X_n^{K-1}$ undergoes a MH move (i.e., execute one MH step that targets $\pi_{K-1}$) to give the next state $X_{n+1}^{K-1}$, and with probability $p_{ee}$, $X_n^{K-1}$ goes through an equi-energy jump to yield $X_{n+1}^{K-1}$. In the equi-energy jump, first a state $y$ is chosen randomly from $\hat{D}_j^K$ with respect to the empirical distribution $F_{n_K}^{K,j}(y)$, where $j = I(X_n^{K-1})$ and $n_K$ denotes the number of steps that the $K$th order chain $X^K$ has been running; then the chosen $y$ is accepted to be $X_{n+1}^{K-1}$ with probability $\min\big(1, (\pi_{K-1}(y)\pi_K(X_n^{K-1}))/(\pi_{K-1}(X_n^{K-1})\pi_K(y))\big)$; otherwise $X_{n+1}^{K-1}$ keeps the old value $X_n^{K-1}$. After a burn-in period on $X^{K-1}$, the EE sampler starts the construction of the second highest-order empirical energy rings $\hat{D}_j^{K-1}$ in the same way as that of $\hat{D}_j^K$: $\hat{D}_j^{K-1}$ contains all the samples $X_n^{K-1}$ such that $I(X_n^{K-1}) = j$. Similarly, the empirical distribution $F_n^{K-1,j}(x)$ on $\hat{D}_j^{K-1}$ is defined to be the number of visits of $X^{K-1}$ to $x$ up to time $n$, divided by the number of visits of $X^{K-1}$ to $D_j$ up to time $n$. After updating the chain $X^{K-1}$ for $N$ steps, the EE sampler starts $X^{K-2}$ targeting $\pi_{K-2}$ while it keeps on running $X^{K-1}$ and $X^K \ldots$

The EE sampler successively moves down the energy and temperature ladder until the last distribution $\pi_0$. Other than $X^K$, each chain $X^i$, $0 \le i < K$, is updated by the equi-energy jump and the MH move with probabilities $p_{ee}$ and $1 - p_{ee}$, respectively, at each iteration. The equi-energy jump move proposes a state $y$ randomly from the empirical energy ring $\hat{D}_{I(X_n^i)}^{i+1}$ with respect to the empirical distribution $F_{n_{i+1}}^{i+1,I(X_n^i)}(y)$, where $n_{i+1}$ denotes the number of steps that chain $X^{i+1}$ has been running at the time, and accepts $y$ with probability $\min\big(1, (\pi_i(y)\pi_{i+1}(X_n^i))/(\pi_i(X_n^i)\pi_{i+1}(y))\big)$. For each chain $X^i$, the energy rings $\hat{D}_j^i$ together with the empirical distribution $F_{n_i}^{i,j}$ are constructed after a burn-in period, and are used for the chain $X^{i-1}$ in the equi-energy jump. Figure 1 diagrams the EE sampler.

In the EE sampler each chain $X^i$ utilizes the *full* memory of the previous chain $X^{i+1}$. Consequently, the sampling algorithm is not Markov. It is this non-Markovian feature of the EE sampler that makes a rigorous proof of its convergence challenging. For example, the nice theoretical framework pioneered by Diaconis and coworkers (Diaconis and Stroock (1991)) to study the geometric ergodicity of MCMC algorithms is not directly available here. Neither is the drift-and-minorization approach of Rosenthal and others (Rosenthal (1995)). We note that Andrieu et al. (2008) considered the convergence of the EE sampler. However, they only considered two chains ($K = 1$), and, furthermore, the algorithm they studied is not the original EE sampler.

Since its introduction, the EE sampler has been applied to a variety of problems, including motif sampling in computational biology, density of states estimation in statistical physics (Kou, Zhou and Wong (2006)), protein folding in

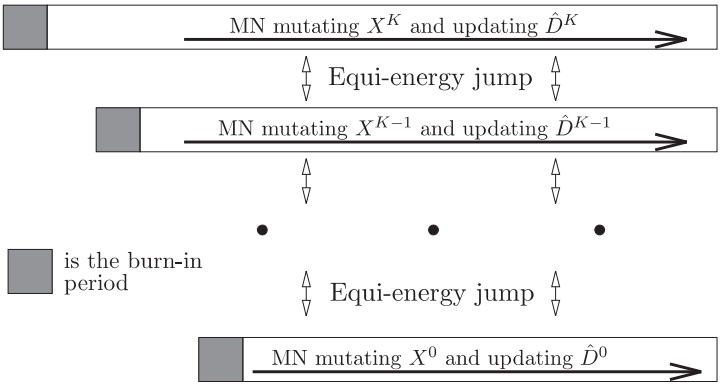


Figure 1. Diagram of the EE sampler.

biophysics (Kou, Oh and Wong (2006)), and characterizing energy landscapes in computational physics (Zhou and Wong (2008, 2009)). Here, in addition to providing a rigorous proof of the convergence and ergodicity of the EE sampler in the case of $\mathcal{X}$ being countable (Sections 2 and 3), we apply it (in Section 4) to the Ising model as a further test of its sampling efficiency. We compare the EE sampler with the MH algorithm and parallel tempering. We calculate the dynamic exponent, which is the benchmark measure of a given Monte Carlo algorithm's efficiency in studying phase transition systems (Newman and Barkema (1999)). We observe that the dynamic exponent of the EE sampler is significantly smaller than those of the MH algorithm and parallel tempering, indicating the EE sampler's high efficiency.

## 2. Notation and Assumptions

To prove the convergence and ergodicity of the EE sampler for countable $\mathcal{X}$, we first introduce some notation and definitions.

The $i$th ($i \le K-1$) chain $X^i$ in the EE sampler is a discrete stochastic process. If the $i$th chain has been running for $n_i$ steps, we write the transition probability as

$$\mathcal{K}_{xy}^{i,n_i} = P(X_{n_i+1}^i = y | X_{n_i}^i = x, X_{n_i-1}^i, \ldots, X_0^i, X_{n_{i+1}}^{i+1}, X_{n_{i+1}-1}^{i+1}, \ldots, X_0^{i+1}, \ldots, X_{n_K}^K, \ldots, X_0^K),$$

where we have $n_{i+k} = n_i + kN$, $1 \le k \le K-i$.

Let $P^i = \{P_{xy}^i\}$, $0 \le i \le K$, denote the transition matrix of the local MH moves on the $i$th chain such that $\pi_i$ is the invariant distribution. Suppose the $i$th ($i \le K-1$) chain has been running for $n_i$ steps and the $(i+1)$th chain has been running for $n_{i+1}$ steps. According to the construction of the EE sampler,

if $X^i_{n_i} = x$, then with probability $1 - p_{ee}$, $X^i_{n_i+1}$ will be drawn via the local MH move. With probability $p_{ee}$ $(0 < p_{ee} < 1)$ the EE sampler attempts to make an equi-energy jump by randomly choosing a state $y$ from $\hat{D}^{i+1}_{I(x)}$ with respect to the empirical distribution $F^{i+1,I(x)}_{n_{i+1}}$ and accepts it with acceptance rate given by

$$\alpha^i_{xy} = \min\left(1, \frac{\pi_i(y)\pi_{i+1}(x)}{\pi_i(x)\pi_{i+1}(y)}\right).$$

Therefore the effective transition probability from state $x$ to state $y$ is:

$$\begin{cases} \mathcal{K}^{K,n_K}_{xy} = P^K_{xy}, \\ \mathcal{K}^{i,n_i}_{xy} = (1-p_{ee})P^i_{xy} + p_{ee}\mathbf{1}_{\hat{D}^{i+1}_{I(x)}}(y)F^{i+1,I(x)}_{n_{i+1}}(y)\alpha^i_{xy}, \qquad 0 \le i \le K-1. \end{cases} \tag{2.1}$$

Note that the $i$th chain is not Markovian by itself because the equi-energy jumps depend on samples generated by higher order chains. However, since $1_{\hat{D}^{i+1}_{I(x)}}(y)F^{i+1,I(x)}_{n_{i+1}}(y)$ only involves the $(i+1)$th chain, the transition probabilities $\mathcal{K}^{i,n_i}$ satisfy

$$\begin{aligned} \mathcal{K}^{i,n_i}_{xy} &= P(X^i_{n_i+1} = y | X^i_{n_i} = x, X^i_{n_i-1}, \ldots, X^i_0, X^{i+1}, X^{i+2}, \ldots, X^K) \\ &= P(X^i_{n_i+1} = y | X^i_{n_i} = x, X^{i+1}). \end{aligned}$$

It follows that for any bounded function $f$ on $\mathcal{X}$,

$$E[f(X^i_{n+m})|X^i_1, \ldots, X^i_n, X^{i+1}] = \mathcal{K}^{i,n+m-1} \ldots \mathcal{K}^{i,n} f(X^i_n).$$

This implies that for any $A \in \sigma(X^i_1, \ldots, X^i_n)$ (the $\sigma$-algebra generated by $X^i_1, \ldots, X^i_n$) and $B \in \sigma(X^i_m, m \ge n)$, $P(A \cap B | X^i_n, X^{i+1}) = P(A|X^i_n, X^{i+1})P(B|X^i_n, X^{i+1})$. We can interpret this as the Markov property of the $i$th chain $X^i$ conditioned on $X^{i+1}$.

Let us also define $p^i_j = \sum_{x \in D_j} \pi_i(x)$ and assume that $p^i_j > 0, \forall i, j$. Then we can define a transition probability matrix $EE^i$ as

$$EE^i_{xy} = \mathbf{1}_{D_{I(x)}}(y)\frac{\pi_{i+1}(y)}{p^{i+1}_{I(x)}}\alpha^i_{xy}, \quad 0 \le i \le K-1.$$

Think of $\mathbf{1}_{D_{I(x)}}(y)(\pi_{i+1}(y)/p^{i+1}_{I(x)})$ as a proposal from $\pi_{i+1}$ but restricted to $D_{I(x)}$, and $\alpha^i_{xy}$ as the corresponding acceptance rate. Finally, define transition probability matrices $S^i$ as

$$\begin{cases} S^K_{xy} = P^K_{xy}, \\ S^i_{xy} = (1-p_{ee})P^i_{xy} + p_{ee}EE^i_{xy}, \quad 0 \le i \le K-1. \end{cases} \tag{2.2}$$

**Assumptions.** We assume the following conditions for the EE sampler:

**(A)** $\pi_i(x)$ defines a genuine probability distribution, i.e., $\sum_x \exp(-h_i(x)) < \infty$.

**(B)** The MH transition matrix $P^K$ targets $\pi_K$ and is irreducible, reversible, and aperiodic. That is, for any $x, y \in \mathcal{X}$, there is an integer $m$ such that $(P^K_{xy})^n > 0\ \forall n > m$, and $\pi_K(x)P^K_{xy} = \pi_K(y)P^K_{yx}$.

**(C)** For $i = 0, \ldots, K-1$, the MH transition matrix $P^i$ targets $\pi_i$ and is reversible. It connects adjacent energy rings in the sense that for any $0 \le j < K$, there exist $x_1, x_2 \in D_j$ with $\pi_i(x_1) > 0, \pi_i(x_2) > 0$, and $y \in D_{j-1}$ if $j \ge 1$, $z \in D_{j+1}$ if $j < K$ with $\pi_i(y) > 0, \pi_i(z) > 0$ such that $P^i_{x_1 y} > 0, P^i_{x_2 z} > 0$.

To study the convergence, we use the following matrix norm throughout this paper.

**Definition 1.** For any real-valued matrix $\{A_{xy}, x, y \in \mathcal{X}\}$, the norm of $A$ is

$$\|A\| = \sup_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} |A_{xy}|. \tag{2.3}$$

It can be shown that the set of matrices with finite norm is complete (Stroock (2005)). Furthermore, $\|A_1 + A_2\| \le \|A_1\| + \|A_2\|$, $\|A_1 A_2\| \le \|A_1\| \, \|A_2\|$ (if the product $A_1 A_2$ exists) and $\|aA\| = |a| \, \|A\|$, $a \in \mathbf{R}$. Accordingly, we use the variation norm for vectors.

**Definition 2.** For any real-valued vector $\{v_x, x \in \mathcal{X}\}$, the variation norm of $v$ is

$$\|v\| = \sum_{x \in \mathcal{X}} |v_x|.$$

For every $\pi_i$, let $\Pi_i$ be the constant row matrix with each row being $\pi_i$. For $x \in \mathcal{X}$, we let

$$L^i_n(x) = \frac{1}{n} \sum_{m=0}^{n-1} 1(X^i_m = x)$$

be the average amount of time the chain $X^i$ spends at $x$ before time $n$.

A Markov chain with transition matrix $P$ is called strongly ergodic if it has a unique stationary probability $\pi$ and $\|P^n - \Pi\| \to 0$, as $n \to \infty$. In the setting of the EE sampler, we call the $i$th chain strongly ergodic if

$$\Big\| \prod_{k=0}^{n} \mathcal{K}^{i,k} - \Pi_i \Big\| \to 0 \ a.s., \ n \to \infty.$$

We say that the $i$th chain has the individual ergodic property if it has a unique steady state $\pi_i$ and for all $x \in \mathcal{X}$

$$L^i_n(x) \to \pi_i(x) \ a.s., \ n \to \infty.$$

The $i$th chain will be said to have the mean ergodic property if for all $x \in \mathcal{X}$

$$E[L_n^i(x)] \to \pi_i(x),\ n \to \infty.$$

## 3. Ergodicity of the EE Sampler

The main theoretical result of this article is the following.

**Theorem 1.** *Under Assumptions* (A)−(C)*, if the state space $\mathcal{X}$ is countable, then $X^i$, $0 \le i \le K-1$, is strongly ergodic with $\pi_i$ as its steady-state distribution. Moreover, all chains satisfy the individual ergodic property and, therefore, the EE sampler produces samples with the desired distributions $\pi_i$.*

To establish it, we first show that the transition probabilities $S^i$ lead to the target distribution $\pi_i$. Then we use induction to show that the time-inhomogeneous transition probabilities $\mathcal{K}^i$ converge to $S^i$ under the matrix norm (2.3), and thus they also target $\pi_i$.

### 3.1. Properties of the transition matrices $S^i$

Comparing (2.1) and (2.2), we see that we need to prove the individual ergodic property of each chain in order to show that the $\mathcal{K}^{i,k}$ converge to $S^i$ as $k \to \infty$. Since the process we encounter is non-Markovian and time inhomogeneous, the classical proof of Markov chains' ergodicity, which uses the Strong Law of Large Numbers for *i.i.d.* random variables, does not work in our case. It turns out that $S^i$ satisfies a Doeblin-type condition, a strong stability condition. Using it, we can establish uniform integrability, which leads to the individual ergodic property of the EE sampler (in the next subsection).

**Lemma 1.** *The transition matrix $S^i$, $0 \le i \le K-1$ is irreducible, aperiodic, and reversible, with $\pi_i$ as a stationary probability distribution. Furthermore, $S^i$ satisfies the Doeblin-type condition: for any fixed $x \in D_K$, there is an integer $M > 0$ and $\epsilon > 0$ such that $(S^i)^m_{yx} \ge \epsilon$ for all $y \in \mathcal{X}$ and all $m \ge M$. Therefore, $\pi_i$ is the unique stationary probability distribution and*

$$\|(S^i)^n - \Pi_i\| \le 2(1-\epsilon)^{[n/M]}.$$

*Consequently, $S^i$ has the individual ergodic property and all states are positive recurrent.*

**Proof.** By definition, for $0 \le i \le K-1$, $S^i_{xy}$ is a mixture of $P^i$ and $EE^i$. Assumption (C) together with the transition matrix $EE^i$ guarantees that each state is accessible from any other state through a combination of transitions via $P^i$ and $EE^i$. Therefore $S^i$ is irreducible. Since there is a positive probability

of retaining a state at each transition, $S^i$ is also aperiodic. Next, it is straightforward to verify that the transition matrix $EE^i_{xy}$ satisfies the detailed balance condition: $\pi_i(x)EE^i_{xy} = \pi_i(y)EE^i_{yx}$. This together with the assumption that $P^i$ is a reversible MH transition matrix implies that for $0 \le i \le K-1$, $S^i$ satisfies the detailed balance.

To see that $S^i$ satisfies the Doeblin-type condition, first note that every $x \in D_j$, $j \le K-1$, satisfies $h(x) < H_{j+1}$, so $\pi_i(x) = c_i \exp(-\max(h(x), H_i)/T_i) \ge c_i \exp(-\max(H_{j+1}, H_i)/T_i)$, where $c_i$ is the normalizing constant that depends only on $i$. Therefore, $D_j$ is finite for all $j \le K-1$.

For $x, y \in D_K$, by definition, $h(x) \ge H_K$, $h(y) \ge H_K$. Since

$$\pi_{i+1}(y) = c_{i+1}\exp(-\frac{h(y)}{T_{i+1}}), \quad \pi_i(y) = c_i \exp\Big(-\frac{h(y)}{T_i}\Big),$$

where $c_i$ and $c_{i+1}$ are normalizing constants, we have

$$\begin{aligned}\frac{\pi_{i+1}(y)}{\pi_i(y)} &= \frac{c_{i+1}}{c_i}\exp\Big[\Big(-\frac{1}{T_{i+1}}+\frac{1}{T_i}\Big)h(y)\Big] \\ &\ge \frac{c_{i+1}}{c_i}\exp\Big[\Big(-\frac{1}{T_{i+1}}+\frac{1}{T_i}\Big)H_K\Big] := b_i > 0.\end{aligned}$$

Thus, $\alpha^i_{yx} = \min\left(1, (\pi_i(x)\pi_{i+1}(y))/(\pi_i(y)\pi_{i+1}(x))\right) \ge \min\left(1, b_i(\pi_i(x)/\pi_{i+1}(x))\right)$,

$$S^i_{yx} \ge p_{ee}\frac{\pi_i(x)}{p^i_K}\alpha^i_{yx} \ge p_{ee}\frac{\pi_i(x)}{p^i_K}\min\left(1, b_i\frac{\pi_i(x)}{\pi_{i+1}(x)}\right) := s_{x,i} > 0,$$

where $s_{x,i}$ depends only on $x$ and $i$. Inductively, we can prove

$$(S^i)^k_{yx} \ge (s_{x,i})^k, \ k \ge 1.$$

Now for any $z \notin D_K$, since $S^i$ is irreducible and aperiodic, it follows that there exists an integer $M_z$ such that $(S^i)^k_{zx} > 0$ for all $k > M_z$. However, as there are only finitely many $z \notin D_K$, we can take $M = \max_{z\notin D_K} M_z$ and have

$$(S^i)^M_{zx} > 0, \text{ for all } z \notin D_K.$$

Let $\epsilon = \min(\min_{z\notin D_K}(S^i)^M_{zx}, \min_{y\in D_K}(S^i)^M_{yx})$. Since $\epsilon \ge \min(\min_{z\notin D_K}(S^i)^M_{zx}, (s_{x,i})^M) > 0$, then $(S^i)^M_{yx} \ge \epsilon$ for all $y \in \mathcal{X}$. It follows that for all $y \in \mathcal{X}$ and all $m \ge M$, $(S^i)^m_{yx} = \sum_z (S^i)^{m-M}_{yz}(S^i)^M_{zx} \ge \epsilon \sum_z (S^i)^{m-M}_{yz} = \epsilon$. This Doeblin-type condition tells us that any fixed $x \in D_K$ can serve as a renewal state. The rest of the statement in the lemma can be proved using Doeblin's well-known arguments and the Strong Law of Large Numbers as in Stroock (2005, pp.28-39).

Next we prove a simple but useful result that shows that in a countable state space, pointwise convergence self-improves to uniform convergence.

**Lemma 2.** *If* $\{a_{m,n},\ m,n \geq 0\}$ *and* $\{b_m,\ m \geq 0\}$ *are two sequences of real numbers such that* $\lim_{n\to\infty} a_{m,n} = b_m$ *for all* $m \geq 0$ *and* $\sum_{m=0}^{\infty} |a_{m,n}| \to \sum_{m=0}^{\infty} |b_m| < \infty$ *as* $n \to \infty$*, then we have*

$$\lim_{n\to\infty} \sum_{m=0}^{\infty} |a_{m,n} - b_m| = 0.$$

**Proof.** By the triangle inequality, we have $\Big||a_{m,n}| - |b_m| - |a_{m,n} - b_m|\Big| \leq 2|b_m|$. Since $\sum_{m=0}^{\infty} |b_m| < \infty$, the Dominated Convergence Theorem implies that, as $n \to \infty$,

$$\sum_{m=0}^{\infty} ||a_{m,n}| - |b_m| - |a_{m,n} - b_m|| \to 0. \tag{3.1}$$

Since $|a_{m,n} - b_m| = |a_{m,n} - b_m| - (|a_{m,n}| - |b_m|) + |a_{m,n}| - |b_m|$, we know

$$\sum_{m=0}^{\infty} |a_{m,n} - b_m| \leq \sum_{m=0}^{\infty} ||a_{m,n} - b_m| - (|a_{m,n}| - |b_m|)| + \sum_{m=0}^{\infty} (|a_{m,n}| - |b_m|).$$

Sending $n \to \infty$, (3.1) and the assumption that $\sum_{m=0}^{\infty} |a_{m,n}| \to \sum_{m=0}^{\infty} |b_m|$ now imply

$$\lim_{n\to\infty} \sum_{m=0}^{\infty} |a_{m,n} - b_m| = 0.$$

### 3.2. Ergodicity of the EE sampler

We use induction to establish the individual ergodic property for each chain in the EE sampler.

**Lemma 3.** *If the individual ergodic property holds for the* $(i+1)$*th chain, that is, for any* $x \in \mathcal{X}$*, as* $n \to \infty$*,* $L_n^{i+1}(x) \to \pi_{i+1}(x)$ *a.s., then we have as* $n \to \infty$*,*

$$\|L_n^{i+1} - \pi_{i+1}\| \to 0 \ a.s., \tag{3.2}$$

$$\|\mathcal{K}^{i,n} - S^i\| \to 0 \ a.s.. \tag{3.3}$$

**Proof.** Equation (3.2) is a direct consequence of Lemma 2 and the fact that

$$\sum_{x\in\mathcal{X}} L_n^{i+1}(x) = \sum_{x\in\mathcal{X}} \pi_{i+1}(x) = 1.$$

Therefore, for any $\epsilon > 0$, there is an integer $N$ such that $\|L_n^{i+1} - \pi_{i+1}\| < \epsilon,\ a.s.$ for all $n > N$. For $0 \leq j \leq K$, define $\epsilon_j = \sum_{z\in D_j} \pi_{i+1}(z) - \sum_{z\in D_j} L_n^{i+1}(z)$. Then

for any $n > N$,

$$\begin{aligned}
&\sum_{y\in D_j}\Big|\frac{L_n^{i+1}(y)}{\sum_{z\in D_j}L_n^{i+1}(z)}-\frac{\pi_{i+1}(y)}{\sum_{z\in D_j}\pi_{i+1}(z)}\Big|\\
&=\sum_{y\in D_j}\Big|\frac{L_n^{i+1}(y)}{\sum_{z\in D_j}L_n^{i+1}(z)}-\frac{L_n^{i+1}(y)}{\sum_{z\in D_j}\pi_{i+1}(z)}+\frac{L_n^{i+1}(y)}{\sum_{z\in D_j}\pi_{i+1}(z)}-\frac{\pi_{i+1}(y)}{\sum_{z\in D_j}\pi_{i+1}(z)}\Big|\\
&\le\sum_{y\in D_j}\Big\{\frac{L_n^{i+1}(y)|\epsilon_j|}{\Big(\sum_{z\in D_j}L_n^{i+1}(z)\Big)\Big(\sum_{z\in D_j}\pi_{i+1}(z)\Big)}+\frac{|L_n^{i+1}(y)-\pi_{i+1}(y)|}{\sum_{z\in D_j}\pi_{i+1}(z)}\Big\}\\
&=\sum_{y\in D_j}\frac{|\epsilon_j|+|L_n^{i+1}(y)-\pi_{i+1}(y)|}{\sum_{z\in D_j}\pi_{i+1}(z)}\le 2\frac{\sum_{y\in D_j}|L_n^{i+1}(y)-\pi_{i+1}(y)|}{\sum_{z\in D_j}\pi_{i+1}(z)}\\
&\le 2\sum_{y\in D_j}|L_n^{i+1}(y)-\pi_{i+1}(y)|\frac{1}{\min_j\sum_{z\in D_j}\pi_{i+1}(z)}.
\end{aligned}$$

Let $c=\min_j\sum_{z\in D_j}\pi_{i+1}(z)>0$, which depends only on $i$. It follows that for $n>N$,

$$\sum_{j=0}^{K}\sum_{y\in D_j}\Big|\frac{L_n^{i+1}(y)}{\sum_{z\in D_j}L_n^{i+1}(z)}-\frac{\pi_{i+1}(y)}{\sum_{z\in D_j}\pi_{i+1}(z)}\Big|\le\frac{2}{c}\|L_n^{i+1}-\pi_{i+1}\|\le\frac{2\epsilon}{c},\ a.s.$$

Hence, for any $x\in\mathcal{X}$,

$$\sum_{y\in\mathcal{X}}|\mathcal{K}_{xy}^{i,n}-S_{xy}^{i}|=\sum_{y\in D_{I(x)}}\Big|\frac{L_n^{i+1}(y)}{\sum_{z\in D_{I(x)}}L_n^{i+1}(z)}-\frac{\pi_{i+1}(y)}{\sum_{z\in D_{I(x)}}\pi_{i+1}(z)}\Big|\alpha_{xy}^{i}\le\frac{2\epsilon}{c},\ a.s.,$$

which implies that $\|\mathcal{K}^{i,n}-S^{i}\|=\sup_x\sum_{y\in\mathcal{X}}|\mathcal{K}_{xy}^{i,n}-S_{xy}^{i}|\le 2\epsilon/c$, $a.s.$ for all $n>N$.

With the establishment of Lemma 3, we can prove that the transition probabilities $\mathcal{K}^i$ lead to the target distribution $\pi_i$, and, as a consequence, the mean ergodic property holds.

**Lemma 4.** *If equation* (3.3) *holds for* $X^i$, *then as* $n\to\infty$,

$$\Big\|\prod_{k=0}^{n}\mathcal{K}^{i,k}-\Pi_i\Big\|\to 0\ a.s. \tag{3.4}$$

*and, for any* $x\in\mathcal{X}$,

$$E[L_n^i(x)|X^{i+1}]\to\pi_i(x)\ a.s.,\ n\to\infty. \tag{3.5}$$

**Proof.** If equation (3.3) holds for $X^i$, then for any fixed $n$, we have

$$\Big\| \prod_{k=m}^{m+n} \mathcal{K}^{i,k} - (S^i)^n \Big\| \to 0 \ a.s., \ \text{as } m \to \infty.$$

By Lemma 1, $\|(S^i)^n - \Pi_i\| \to 0$, as $n \to \infty$. Furthermore $\prod_{k=0}^{m-1} \mathcal{K}^{i,k}$ is a probability matrix (i.e., each row of the matrix is a probability vector) and $\Pi_i$ has constant rows $\pi_i$. Therefore,

$$\begin{aligned}\Big\| \prod_{k=0}^{m-1} \mathcal{K}^{i,k}(S^i)^n - \Pi_i \Big\| &= \Big\| \prod_{k=0}^{m-1} \mathcal{K}^{i,k}(S^i)^n - \prod_{k=0}^{m-1} \mathcal{K}^{i,k}\Pi_i \Big\| \\ &\le \left( \Big\| \prod_{k=0}^{m-1} \mathcal{K}^{i,k} \Big\| \right) \|(S^i)^n - \Pi_i\| = \|(S^i)^n - \Pi_i\| \to 0, \ n \to \infty.\end{aligned}$$

Hence,

$$\begin{aligned}\Big\| \prod_{k=0}^{m+n} \mathcal{K}^{i,k} - \Pi_i \Big\| &\le \Big\| \prod_{k=0}^{m-1} \mathcal{K}^{i,k} \prod_{k=m}^{m+n} \mathcal{K}^{i,k} - \prod_{k=0}^{m-1} \mathcal{K}^{i,k}(S^i)^n \Big\| + \Big\| \prod_{k=0}^{m-1} \mathcal{K}^{i,k}(S^i)^n - \Pi_i \Big\| \\ &\le \Big\| \prod_{k=m}^{m+n} \mathcal{K}^{i,k} - (S^i)^n \Big\| + \Big\| \prod_{k=0}^{m-1} \mathcal{K}^{i,k}(S^i)^n - \Pi_i \Big\| \to 0, \ a.s.\end{aligned}$$

as $m, n \to \infty$. This proves (3.4). To prove (3.5), let

$$A_n^i = \frac{1}{n} \sum_{m=0}^{n-1} \mathcal{K}^{i,0} \dots \mathcal{K}^{i,m}.$$

From (3.4) we have, as $n \to \infty$,

$$\|A_n^i - \Pi_i\| \le \frac{1}{n} \sum_{m=0}^{n-1} \|\mathcal{K}^{i,0} \dots \mathcal{K}^{i,m} - \Pi_i\| \to 0, \ a.s..$$

Consequently, for any $x \in \mathcal{X}$,

$$\begin{aligned}E[L_n^i(x) | X_0^i, X^{i+1}] &= \frac{1}{n} \sum_{m=0}^{n-1} E[\mathbf{1}_x(X_m^i) | X_0^i, X^{i+1}] \\ &= \frac{1}{n} \sum_{m=0}^{n-1} \mathcal{K}^{i,0} \dots \mathcal{K}^{i,m} \mathbf{1}_x(X_0^i) \\ &= A_n^i \mathbf{1}_x(X_0^i) \to \pi_i(x) \ a.s., \ n \to \infty.\end{aligned}$$

It follows that for any $x \in \mathcal{X}$, as $n \to \infty$,

$$E[L_n^i(x)|X^{i+1}] \to \pi_i(x) \text{ a.s., and } E[L_n^i(x)] \to \pi_i(x).$$

Lemma 4 shows that if chain $i+1$ has the individual ergodic property, then chain $i$ will have the mean ergodic property conditioning on chain $i+1$. Next we show that the mean ergodic property can be improved to the individual ergodic property.

**Lemma 5.** *If the individual ergodic property holds for the chain $X^{i+1}$, then for any $x \in \mathcal{X}$, $L_n^i(x) \to \pi_i(x)$ a.s., $n \to \infty$.*

**Proof.** Recall that conditioning on $X^{i+1}$, the $i$th chain has the Markov property: $\forall A \in \sigma(X_1^i, \ldots, X_n^i)$ and $\forall B \in \sigma(X_m^i, m \geq n)$, $P(A \cap B|X_n^i, X^{i+1}) = P(A|X_n^i, X^{i+1})P(B|X_n^i, X^{i+1})$. Suppose the initial state of the $i$th chain is some $x_0^i \in \mathcal{X}$. Let $\rho_m^i(x)$ be the time of the $m$th return to $x$ of the $i$th chain $X^i$. In other words, $\rho_0^i(x) := 0$ and $\rho_m^i(x) = \inf\{n > \rho_{m-1}^i(x) : X_n^i = x\}$. ($\inf(\emptyset)$ is interpreted to be $\infty$.) Then by the Markov property, $\{\rho_m^i(x) - \rho_{m-1}^i(x), m \geq 1\}$ is a sequence of independent random variables conditioning on $X^{i+1}$ and $X_0^i = x_0^i$.

By assumption, $\lim_{n\to\infty} L_n^{i+1}(x) = \pi_{i+1}(x)$ a.s., so we can focus on realizations $X^{i+1} = \hat{X}^{i+1}$ such that $\lim_{n\to\infty}(1/n)\sum_{m=0}^{n-1} 1_{\{x\}}(\hat{X}_m^{i+1}) = \pi_{i+1}(x)$. For notational convenience, we write $\delta_m = \rho_m^i(x) - \rho_{m-1}^i(x)$ and $\mu_m = E[\delta_m|X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i]$. We will show that

**c1.** $\mu_m < \infty$ for all $m$ and $\lim_{m\to\infty} \mu_m = 1/\pi_i(x)$,

**c2.** $\delta_m$ are uniformly integrable conditioned on $X^{i+1} = \hat{X}^{i+1}$ and $X_0^i = x_0^i$, that is,

$$\lim_{R\to\infty} \sup_{m\geq 1} E\Big[\delta_m \mathbf{1}_{(R,\infty)}(\delta_m)|X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i\Big] = 0.$$

Then it easily follows that

$$\lim_{R\to\infty} \sup_{m\geq 1} E\Big[(\delta_m - \mu_m)\mathbf{1}_{(R,\infty)}(\delta_m - \mu_m) \,\Big|\, X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i\Big] = 0,$$

and hence by the Strong Law of Large Numbers under uniform integrability (Landers and Rogers (1985)) we have

$$P\Big(\lim_{n\to\infty} \frac{1}{n}\sum_{m=0}^{n-1}(\delta_m - \mu_m) = 0 \,\Big|\, X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i\Big) = 1.$$

But $\lim_{m\to\infty} \mu_m = 1/\pi_i(x)$, so it follows that

$$P\Big(\lim_{n\to\infty} \frac{1}{n}\sum_{m=0}^{n-1}\delta_m = \frac{1}{\pi_i(x)} \,\Big|\, X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i\Big) = 1.$$

By elementary manipulations (e.g. those on page 39 of Stroock (2005)) we can show that for $m = [n\pi_i(x)]$,

$$|L_n^i(x) - \pi_i(x)| \le \frac{2}{n} + 3\pi_i(x)\Big|\frac{\rho_m^i(x)}{m} - \frac{1}{\pi_i(x)}\Big|.$$

But $(1/m)\sum_{i=0}^{n-1}\delta_i = \rho_m^i(x)/m$. Therefore,

$$P\Big(\lim_{n\to\infty} L_n^i(x) = \pi_i(x) | X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i\Big) = 1,$$

which holds for almost all realizations $\hat{X}^{i+1}$ of $X^{i+1}$. It follows that

$$P\Big(\lim_{n\to\infty} L_n^i(x) = \pi_i(x)\Big) = 1.$$

We first prove **c1** for $x \in D_K$. By Lemma 3, $\|\mathcal{K}^{i,k} - S^i\| \to 0$ as $k \to \infty$ given $\hat{X}^{i+1}$ such that $\lim_{n\to\infty}(1/n)\sum_{m=0}^{n-1} 1_{\{x\}}(\hat{X}_m^{i+1}) = \pi_{i+1}(x)$. For $x \in D_K$, we have shown in Lemma 1 that there is an integer $M$ and some $\epsilon > 0$ such that $((S^i)^M)_{yx} \ge \epsilon$ for all $y \in \mathcal{X}$. By Lemma 4, there is an integer $N > 0$ that depends on $\hat{X}^{i+1}$ such that $\|\prod_{k=n}^{n+M}\mathcal{K}^{i,k} - (S^i)^M\| < \epsilon/2$ for all $n \ge N$. Therefore, for all $y \in \mathcal{X}$ and $n \ge N$,

$$\Big(\prod_{k=n}^{n+M}\mathcal{K}^{i,k}\Big)_{yx} \ge \frac{\epsilon}{2}.$$

It follows that for any integer $0 \le n' < N$ and any $y \in \mathcal{X}$,

$$\Big(\prod_{k=n'}^{N+M}\mathcal{K}^{i,k}\Big)_{yx} = \Big(\prod_{k=n'}^{N-1}\mathcal{K}^{i,k}\prod_{k=N}^{N+M}\mathcal{K}^{i,k}\Big)_{yx} \ge \frac{\epsilon}{2}.$$

Let $m > 1$ be any integer. If $\rho_{m-1}^i(x) = t$ for some $1 \le t < N$, then

$$\begin{aligned} &P(\rho_m^i(x) \le M + N \,|\, X^{i+1} = \hat{X}^{i+1}, \rho_{m-1}^i(x) = t, X_0^i = x_0^i) \\ &\quad \ge \big(\mathcal{K}^{i,t}\ldots\mathcal{K}^{i,N+M}\big)_{xx} \ge \frac{\epsilon}{2}. \end{aligned}$$

Therefore, $P(\rho_m^i(x) > M + N | X^{i+1} = \hat{X}^{i+1}, \rho_{m-1}^i(x) = t, X_0^i = x_0^i) \le 1 - \epsilon/2$. Inductively, suppose $P(\rho_m^i(x) > nM + N \,|X^{i+1} = \hat{X}^{i+1}, \rho_{m-1}^i(x) = t, X_0^i =$

$x_0^i) \le (1-\epsilon/2)^n$, then

$$\begin{aligned}
&P(\rho_m^i(x) > (n+1)M+N | X^{i+1} = \hat{X}^{i+1}, \rho_{m-1}^i(x) = t, X_0^i = x_0^i)\\
&= \sum_{z\in\mathcal{X}, z\neq x} P(\rho_m^i(x) > nM+N, X_{nM+N}^i = z \,|\, X^{i+1} = \hat{X}^{i+1}, \rho_{m-1}^i(x) = t, X_0^i = x_0^i)\\
&\quad \times P(X_k^i \neq x, nM+N \le k \le (n+1)M+N \,|\\
&\quad X_{nM+N}^i = z, X^{i+1} = \hat{X}^{i+1}, \rho_{m-1}^i(x) = t, X_0^i = x_0^i)\\
&\le \sum_{z\in\mathcal{X}} P(\rho_m^i(x) > nM+N, X_{nM+N}^i = z \,|\, X^{i+1} = \hat{X}^{i+1}, \rho_{m-1}^i(x) = t, X_0^i = x_0^i)\\
&\quad \times \left(1 - \left(\mathcal{K}^{i,nM+N} \dots \mathcal{K}^{i,(n+1)M+N}\right)_{zx}\right)\\
&\le (1-\frac{\epsilon}{2})^n \cdot (1-\frac{\epsilon}{2}) = (1-\frac{\epsilon}{2})^{n+1}.
\end{aligned}$$

Summing over all $1 \le t < N$ gives for any integer $n \ge 0$,

$$P(\rho_m^i(x) > nM+N \,|\, X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i, 1 \le \rho_{m-1}^i(x) < N) \le (1-\frac{\epsilon}{2})^n,$$

and thus

$$\begin{aligned}
&P(\rho_m^i(x) - \rho_{m-1}^i(x) > nM+N \,|\, X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i, 1 \le \rho_{m-1}^i(x) < N)\\
&\quad \le (1-\frac{\epsilon}{2})^n. \qquad (3.6)
\end{aligned}$$

If $\rho_{m-1}^i(x) = t$ for some $t \ge N$ then

$$P(\rho_m^i(x) \le M+t \,|\, X^{i+1} = \hat{X}^{i+1}, \rho_{m-1}^i(x) = t, X_0^i = x_0^i) \ge (\mathcal{K}^{i,t} \dots \mathcal{K}^{i,M+t})_{xx} \ge \frac{\epsilon}{2}.$$

Therefore, $P(\rho_m^i(x) > M+t \,|\, X^{i+1} = \hat{X}^{i+1}, \rho_{m-1}^i(x) = t, X_0^i = x_0^i) \le 1-\epsilon/2$. Using induction again, we can verify $P(\rho_m^i(x) > nM+t \,|\, X^{i+1} = \hat{X}^{i+1}, \rho_{m-1}^i(x) = t, X_0^i = x_0^i) \le (1-\epsilon/2)^n$. Summing over all $t \ge N$ gives, for any $n \ge 0$,

$$P(\rho_m^i(x) - \rho_{m-1}^i(x) > nM \,|\, X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i, N \le \rho_{m-1}^i(x) < \infty) \le (1-\frac{\epsilon}{2})^n. \qquad (3.7)$$

(3.6) and (3.7) together imply that for any $m > 1$,

$$\begin{aligned}
&P(\rho_m^i(x) - \rho_{m-1}^i(x) > nM+N | X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i, 1 \le \rho_{m-1}^i(x) < \infty)\\
&\quad \le (1-\frac{\epsilon}{2})^n. \qquad (3.8)
\end{aligned}$$

Therefore, for any $p > 0$

$$
\begin{aligned}
&E[(\rho_m^i(x) - \rho_{m-1}^i(x))^p | X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i, 1 \le \rho_{m-1}^i(x) < \infty] \\
&= \sum_{n=N+2M+1}^{\infty} n^p P(\rho_m^i(x) - \rho_{m-1}^i(x) = n | X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i, 1 \le \rho_{m-1}^i(x) < \infty) \\
&\quad + \sum_{n=1}^{N+2M} n^p P(\rho_m^i(x) - \rho_{m-1}^i(x) = n \mid X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i, 1 \le \rho_{m-1}^i(x) < \infty) \\
&\le \sum_{k=[N/M]+2}^{\infty} (k+1)^p M^p \sum_{n=kM+1}^{(k+1)M} P(\rho_m^i(x) - \rho_{m-1}^i(x) = n | X^{i+1} = \hat{X}^{i+1}, \\
&\quad X_0^i = x_0^i, 1 \le \rho_{m-1}^i(x) < \infty) + \sum_{n=1}^{N+2M} n^p \\
&\le \sum_{k=[N/M]+2}^{\infty} (k+1)^p M^p P(\rho_m^i(x) - \rho_{m-1}^i(x) > kM | X^{i+1} = \hat{X}^{i+1}, \\
&\quad X_0^i = x_0^i, 1 \le \rho_{m-1}^i(x) < \infty) + \sum_{n=1}^{N+2M} n^p \\
&\le M^p \sum_{i=1}^{\infty} (i + \left[\frac{N}{M}\right] + 2)^p (1 - \frac{\epsilon}{2})^{i+1} + \sum_{n=1}^{N+2M} n^p < \infty.
\end{aligned}
$$

In particular, for $p = 1$ we have

$$
E[\rho_m^i(x) - \rho_{m-1}^i(x) | X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i, 1 \le \rho_{m-1}^i(x) < \infty] < \infty \tag{3.9}
$$

for all $m > 1$. For $m = 1$, note that $\rho_0^i(x) = 0$ by definition. Therefore

$$
P(\rho_1^i(x) \le M + N \mid X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i) \ge (\mathcal{K}^{i,0} \dots \mathcal{K}^{i,N+M})_{x_0^i x} \ge \frac{\epsilon}{2}.
$$

Using induction we can show that

$$
P(\rho_1^i(x) > nM + N \mid X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i) \le (1 - \frac{\epsilon}{2})^n, \tag{3.10}
$$

$$
E[\rho_1^i(x)^p | X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i] < \infty, \ \forall p > 0. \tag{3.11}
$$

Combining (3.9) and (3.11), we obtain $P(\rho_m^i(x) < \infty) = 1$ and $E[\rho_m^i(x) - \rho_{m-1}^i(x) | X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i] < \infty$ for all $m \ge 1$. Hence the first part of **c1** is true for $x \in D_K$. Furthermore, (3.8) can be combined with (3.10) to give

$$
P(\rho_m^i(x) - \rho_{m-1}^i(x) > nM + N | X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i) \le (1 - \frac{\epsilon}{2})^n, \ \forall \ m \ge 1. \tag{3.12}
$$

For the second part of **c1**, note that for any $k \geq 1$ and any $t \geq 0$, it is easy to check by induction that $P(\rho_m^i(x) - \rho_{m-1}^i(x) > k \,|\, X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i, \rho_{m-1}^i(x) = t)$ is a polynomial of degree at most $k$ in variables $(\prod_{j=l}^{l+n} \mathcal{K}^{i,j})_{xx}$, where $l = t, t+1, \ldots, t+k-1$ and $n \leq t+k-l-1$. For example, $P(\rho_m^i(x) - \rho_{m-1}^i(x) > 2 \,|\, X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i, \rho_{m-1}^i(x) = t) = 1 - \mathcal{K}_{xx}^{i,t} - (\mathcal{K}^{i,t}\mathcal{K}^{i,t+1})_{xx} + \mathcal{K}_{xx}^{i,t}\mathcal{K}_{xx}^{i,t+1}$. Let $\lambda(x)$ be the return time of the Markov chain $Y_n$ under the transition matrix $S^i$ with initial state $Y_0 = x$, i.e., $\lambda(x) = \inf\{n > 0; Y_n = x\}$. Then $P(\lambda(x) > k \,|\, Y_0 = x)$ is a polynomial of degree at most $k$ in variables $(S^i)_{xx}^j$, $j = 0, \ldots, k$. Using Lemma 3, for any fixed $k$,

$$\begin{aligned}&\lim_{t\to\infty} P\Big(\rho_m^i(x) - \rho_{m-1}^i(x) > k \,|\, X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i, \rho_{m-1}^i(x) = t\Big)\\ &\quad = P(\lambda(x) > k \,|\, Y_0 = x).\end{aligned}$$

Thus,

$$\begin{aligned}&\lim_{m\to\infty} P(\rho_m^i(x) - \rho_{m-1}^i(x) > k \,|\, X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i)\\ &\leq \lim_{m\to\infty} \sup_{t\geq m-1} P(\rho_m^i(x) - \rho_{m-1}^i(x) > k \,|\, X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i, \rho_{m-1}^i(x) = t)\\ &= P(\lambda(x) > k \,|\, Y_0 = x).\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\lim_{m\to\infty} P(\rho_m^i(x) - \rho_{m-1}^i(x) > k \;|X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i)\\ &\geq \lim_{m\to\infty} \inf_{t\geq m-1} P(\rho_m^i(x) - \rho_{m-1}^i(x) > k \;|X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i, \rho_{m-1}^i(x) = t)\\ &= P(\lambda(x) > k \;\,|\, Y_0 = x).\end{aligned}$$

Therefore, $\lim_{m\to\infty} P(\rho_m^i(x) - \rho_{m-1}^i(x) > k \;\,|\, X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i) = P(\lambda(x) > k \,|\, Y_0 = x)$.

By (3.12), for all $m \geq 1$, $P(\rho_m^i(x) - \rho_{m-1}^i(x) > k \,|\, X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i)$ is bounded above by an integrable function $g$: $g(k) = (1-\epsilon)^{[(k-N)/M]}$ for $k \geq N$ and $g(k) = 1$ for $k < N$. Therefore, by the Dominated Convergence Theorem,

$$\begin{aligned}&\lim_{m\to\infty} E[\rho_m^i(x) - \rho_{m-1}^i(x) | X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i]\\ &\quad = \lim_{m\to\infty} \Big(\sum_{k=0}^{\infty} P(\rho_m^i(x) - \rho_{m-1}^i(x) > k \,|\, X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i)\Big)\\ &\quad = \sum_{k=0}^{\infty} P(\lambda(x) > k \,|\, Y_0 = x) = E(\lambda(x) | Y_0 = x) = \frac{1}{\pi_i(x)}.\end{aligned}$$

Thus, we have proved **c1** for $x \in D_K$.

For **c2**, note that for any integer $r > 0$, any $p > 0$, and any $m \geq 1$, by (3.12),

$$
\begin{aligned}
&E[(\rho_m^i(x) - \rho_{m-1}^i(x))^p \mathbf{1}_{(rM+N,\infty)}(\rho_m^i(x) - \rho_{m-1}^i(x)) | X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i] \\
&= \sum_{n=rM+N+1}^{\infty} n^p P(\rho_m^i(x) - \rho_{m-1}^i(x) = n | X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i) \\
&\leq \sum_{k=r}^{\infty} ((k+1)M+N)^p \sum_{n=kM+1}^{(k+1)M} P(\rho_m^i(x) - \rho_{m-1}^i(x) = n+N | X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i) \\
&\leq \sum_{k=r}^{\infty} ((k+1)M+N)^p P(\rho_m^i(x) - \rho_{m-1}^i(x) > kM+N | X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i) \\
&\leq \max(M,N)^p \sum_{k=r}^{\infty} (k+2)^p (1 - \frac{\epsilon}{2})^k.
\end{aligned}
$$

In particular, for $p = 1$,

$$
\begin{aligned}
&E[(\rho_m^i(x) - \rho_{m-1}^i(x)) \mathbf{1}_{(rM+N,\infty)}(\rho_m^i(x) - \rho_{m-1}^i(x)) | X^{i+1} = \hat{X}^{i+1}, X_0^i = x_0^i] \\
&\quad \leq 4 \max(M,N) \frac{1}{\epsilon^2} \left( (r+2)(1 - \frac{\epsilon}{2})^r - (r+1)(1 - \frac{\epsilon}{2})^{r+1} \right),
\end{aligned}
$$

which is independent of $m$ and tends to 0 as $r \to \infty$. Thus **c2** holds for $x \in D_K$.

Finally for $x \in D_j$, $j \neq K$, since $S^i$ is irreducible, for any fixed $x_0 \in D_K$ there is an integer $M'$ such that $\left((S^i)^{M'}\right)_{x_0 x} \geq \epsilon'$ for some $\epsilon' > 0$. From Lemma 1, there exists $M_0$ such that for all $k \geq M_0$ and $y \in \mathcal{X}$, $\left((S^i)^k\right)_{y x_0} \geq \epsilon_0$ for some $\epsilon_0 > 0$, which implies $\left((S^i)^k\right)_{yx} \geq \left((S^i)^{k-M'})_{yx_0} ((S^i)^{M'}\right)_{x_0 x} \geq \epsilon' \epsilon_0$ for $k \geq M' + M_0$. Thus we can find an integer $N'$ a.s. that depends on $\hat{X}^{i+1}$ such that, for $m \geq N'$ and all $y \in \mathcal{X}$,

$$
\left( \prod_{k=m}^{m+M'+M_0} \mathcal{K}^{i,k} \right)_{yx} \geq \frac{\epsilon' \epsilon_0}{2}.
$$

With this result we note that the same argument that we used to prove **c1** and **c2** for $x \in D_K$ applies for general $x \in D_j$.

Now we are ready to prove the main theorem.

**Proof of Theorem 1.** By assumption, the highest order chain $X^K$ has transition probabilities $P^K$ that targets $\pi_K$ and is irreducible and aperiodic. Therefore, standard Markov theory tells us that $\pi_K$ is the unique stationary probability and the individual ergodicity holds: as $n \to \infty$, $L_n^K(x) \to \pi_K(x)$ a.s. Inductively, if the $(i+1)$th $(i \leq K-2)$ chain is strongly ergodic a.s. and satisfies

$L_n^{i+1}(x) \to \pi_{i+1}(x)$ $a.s.$, $n \to \infty$, then by Lemmas 3 and 4,

$$\Big\| \prod_{k=0}^{n} \mathcal{K}^{i,k} - \Pi_i \Big\| \to 0 \ a.s., \ n \to \infty.$$

Therefore, the $i$th chain is strongly ergodic, and by Lemma 5, $L_n^i(x) \to \pi_i(x)$ $a.s.$, $n \to \infty$. Therefore, the $i$th chain also has the individual ergodic property. This induction result completes the proof.

**Remark.** It is worth pointing out that our theoretical results can be easily extended to cover more general settings of the EE sampler. The energy function $h_i$ for the $i$th distribution, $\pi_i(x) \propto \exp(-h_i(x))$, can be any function that depends only on $T_i$ and $H_i$; for example, $h_i(x) = h(x)/T_i$.

## 4. Applying the EE Sampler to the Ising Model

In this section, we apply the EE sampler to the Ising model and compare its efficiency with the MH algorithm and the parallel tempering algorithm, a well-known all-purpose Monte Carlo simulation method.

### 4.1. The Ising model

The Ising model is a simple model of a magnet in which spins $s_i$ are placed on the sites $i$ of a lattice (Newman and Barkema (1999)). Each spin can take either of two values: $+1$ (up) and $-1$ (down). If there are $N$ sites on the lattice, then the system can be in any of $2^N$ states and the energy of any particular state is given by the Ising Hamiltonian:

$$H = -J \sum_{i \sim j} s_i s_j,$$

where $J$ is an interaction energy between nearest neighbor spins $i$ and $j$.

It is known that there is a critical temperature $T_c$ at which a phase transition occurs. Below the critical temperature, the system forms into large clusters of predominantly up- or down-pointing spins and magnetization develops. Above the critical temperature, the spins tend to be randomly arranged and the average magnetization is zero. For the two-dimensional Ising model, the exact value of $T_c$ is known:

$$T_c = \frac{2J}{\log(1 + \sqrt{2})} \approx 2.269J.$$

As the system approaches $T_c$, the typical size of the clusters, termed correlation length, diverges. These clusters contribute significantly to both the magnetization $m = \sum_i s_i$ and the energy $E = H$ of the system so that, as they flip from

one orientation to another, they produce large fluctuations in $m$ and $E$, a phenomenon termed critical fluctuations. As the typical size of the clusters diverges as $T \to T_c$, the variation of $E$, termed specific heat, and the variation of $m$, termed susceptibility, diverge as well.

### 4.2. Simulation result

Ising models on square lattices of finite sites have been extensively studied by Monte Carlo methods from the MH algorithm to parallel tempering to the more sophisticated cluster algorithms such as the Swendsen-Wang (Swendsen and Wang (1987)) and Wolff (1989) algorithms. The MH algorithm is known to have a very long correlation time at the critical temperature. More specifically, if we take a square lattice of $L \times L$ sites and let $\tau$ be the integrated correlation time of magnetization $m(t)$ obtained from an algorithm at the critical temperature $T_c$, then we typically have

$$\tau \sim L^z,$$

where $z$ is called the dynamic critical exponent. It is a key characteristic of an algorithm's efficiency in studying phase transition systems (Nightingale and Blote (1996)): a small $z$ is much preferred. For the MH algorithm, the best available dynamic exponent measurement is $z_{MH} = 2.1665 \pm 0.0012$ (Nightingale and Blote (1996)).

To test the efficiency of the EE sampler, we apply it to the Ising model and measure the dynamic exponent at the critical temperature. We used $h_i(x) = h(x)/T_i$, and five chains with temperature levels 2.47, 2.41, 2.35, 2.3, 2.269, and $p_{ee} = 0.05$. Each chain was burned for 100,000 steps per site to ensure the system was in equilibrium, then was run for 180,000 steps per site for sampling. From the Monte Carlo samples, we estimated the integrated correlation times $\hat{\tau}$ to be $19.87 \pm 0.88$, $23.88 \pm 1.59$, $30.97 \pm 2.45$, $37.46 \pm 3.27$, $44.20 \pm 3.14$, $51.474 \pm 4.144$ for $L = 24, 32, 48, 64, 80, 96$, respectively. Figure 2 plots $\log \hat{\tau}$ versus $\log L$. Regressing $\log \hat{\tau}$ on $\log L$, we measure the dynamic exponent to be $z_{EE} = 0.678 \pm 0.054$.

The integrated correlation time $\hat{\tau}$ and its standard error were calculated by

$$\hat{\tau} = \sum_{t=0}^{W} \frac{A(t)}{A(0)},$$
$$\mathrm{Var}(\hat{\tau}) = \frac{2(2W+1)}{n}\hat{\tau}^2,$$

where $n = 180{,}000$ is the sample size, $A(t) = (1/(n-t))\sum_{i=1}^{n-t}(m(i) - \bar{m})(m(i+t) - \bar{m})$ is the sample autocorrelation function, and $W$ satisfying $\hat{\tau} \ll W \ll n$ is

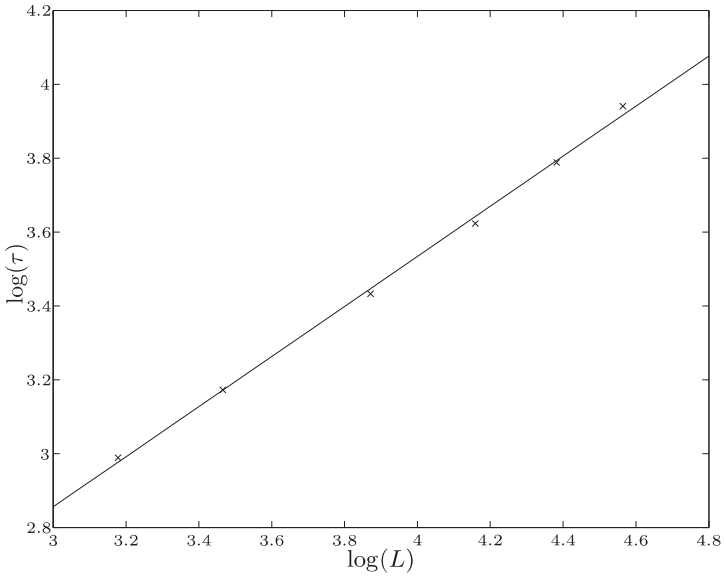


Figure 2. The estimated correlation times $\hat{\tau}$ (measured in Monte Carlo steps per lattice site) at $L = 24, 32, 48, 64, 80, 96$ from the EE sampler output, which uses five chains at temperatures 2.47, 2.41, 2.35, 2.3, 2.269. The linear regression fit gives $z_{EE} = 0.678 \pm 0.054$.

estimated as in Sokal (1989). For large sample size, $\hat{\tau}$ is approximately normally distributed (Wei (1990)). The dynamic exponent is estimated as

$$\hat{z} = \frac{\sum_i (\log L_i - \overline{\log L})(\log \hat{\tau}_i - \overline{\log \hat{\tau}})}{\sum_i (\log L_i - \overline{\log L})^2},$$

and the standard deviation of our estimate of $z$ is

$$\frac{\sqrt{\sum_i (\log L_i - \overline{\log L})^2 \mathrm{Var}(\log \hat{\tau}_i)}}{\sum_i (\log L_i - \overline{\log L})^2},$$

where the $\mathrm{Var}(\log \hat{\tau}_i)$ are estimated by numerical integration under normal approximation (Wei (1990)). Tables 1 and 2 report the energy ladders used in the simulation and the acceptance rates of the EE jumps.

For comparison, we also estimated the dynamic exponent of parallel tempering. We also used five chains. In general, for optimal performance, the temperature levels in the parallel tempering should be different from those of the EE sampler. After numerous trials, we found that, for parallel tempering, taking the five temperature levels to be 2.41, 2.365, 2.33, 2.3, 2.269, and swapping frequency $p = 0.15$, appeared to offer good performance. Each chain was burned for

Table 1. Energy ladders for $L = 24, 32, 48, 64, 80, 96$ in the EE sampling.

| $L$ | $H_0$ | $H_1$ | $H_2$ | $H_3$ | $H_4$ |
|---|---|---|---|---|---|
| 24 | -1152 | -850 | -800 | -720 | -650 |
| 32 | -2048 | -1556 | -1444 | -1334 | -1222 |
| 48 | -4608 | -3500 | -3250 | -3000 | -2750 |
| 64 | -8192 | -5800 | -5400 | -4800 | -4200 |
| 80 | -12800 | -9000 | -8500 | -7800 | -7200 |
| 96 | -18432 | -13000 | -11800 | -10800 | -9800 |

Table 2. Acceptance rates of EE jumps for $L = 24, 32, 48, 64, 80, 96$.

| $L$ | 0th chain | 1st chain | 2nd chain | 3rd chain |
|---|---|---|---|---|
| 24 | 0.88 | 0.82 | 0.79 | 0.79 |
| 32 | 0.84 | 0.75 | 0.71 | 0.70 |
| 48 | 0.73 | 0.60 | 0.55 | 0.54 |
| 64 | 0.59 | 0.43 | 0.30 | 0.43 |
| 80 | 0.50 | 0.31 | 0.26 | 0.29 |
| 96 | 0.40 | 0.10 | 0.15 | 0.24 |

Table 3. Acceptance rates of swap moves for $L = 32, 40, 48, 64, 80, 96$ in the parallel tempering sampling.

| $L$ | 0th chain | 1st chain | 2nd chain | 3rd chain |
|---|---|---|---|---|
| 32 | 0.66 | 0.69 | 0.67 | 0.62 |
| 40 | 0.58 | 0.60 | 0.59 | 0.56 |
| 48 | 0.50 | 0.54 | 0.52 | 0.47 |
| 64 | 0.36 | 0.42 | 0.40 | 0.36 |
| 80 | 0.23 | 0.33 | 0.31 | 0.25 |
| 96 | 0.14 | 0.23 | 0.24 | 0.17 |

200,000 steps per site to ensure the system was in equilibrium, then was run for 350,000 steps per site for sampling. Table 3 reports the acceptance rates of swap moves in the parallel tempering sampling. We measured $\hat{\tau}$ to be $409.7 \pm 90.1$, $660.9 \pm 220.8$, $1029.3 \pm 296.3$, $1524.7 \pm 857.5$, $2828.9 \pm 1406.2$, $3604.0 \pm 2142.4$ for $L = 32, 40, 48, 64, 80, 96$, respectively. Figure 3 plots $\log \hat{\tau}$ versus $\log L$. The dynamic exponent of the parallel tempering was measured to be $z_{PT} = 1.98 \pm 0.59$. The large standard error in our estimation reflects the very long correlation time in parallel tempering.

Contrasting the dynamic exponents — $z_{MH} = 2.1665 \pm 0.0012$, $z_{PT} = 1.98 \pm 0.59$, $z_{EE} = 0.678 \pm 0.054$ — of the three algorithms, we see that the EE sampler was much more efficient than both the MH algorithm and the parallel tempering algorithm in studying the Ising model. Parallel tempering does not seem to significantly improve the MH algorithm even with high swapping acceptance rates. This is probably because, when the parallel tempering performs a swap,

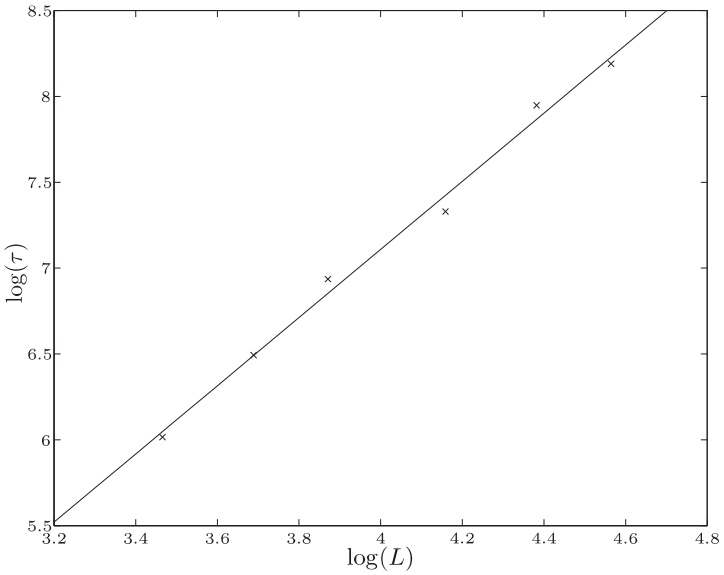


Figure 3. The estimated correlation times $\hat{\tau}$ (measured in Monte Carlo steps per lattice site) at $L = 32, 40, 48, 64, 80, 96$ from the output of the parallel tempering with five chains at temperatures 2.41, 2.365, 2.33, 2.3, 2.269. The linear regression fit gives $z_{PT} = 1.9846 \pm 0.5858$.

the new state is not significantly different from the old state, whereas the EE sampler can reach all previously visited states through the EE jumps. We note that the Wolff algorithm with $z = 0.25 \pm 0.01$ (Coddington and Baillie (1992)) is still more efficient than the EE sampler. This is not surprising as the Wolff algorithm is specifically designed for the Ising model (to tackle the critical slow down), whereas the EE sampler is a universal Monte Carlo algorithm and can be combined with virtually any simulation algorithm as its local update.

Using the energy rings constructed by the EE sampler, we can also estimate the density of states and calculate the Boltzmann averages of various functions of the Ising system, as described in Kou, Zhou and Wong (2006). Figures 4 and 5 show the specific heat $C(T)$ and the expectation of the absolute magnetization $E\{|m(T)|\}$ estimated at temperatures around $T_c$ for $L = 96$, using the energy rings constructed by an EE sampler, where the temperature ladder is taken to be 2.5, 2.4, 2.35, 2.305, 2.28, 2.25, 2.2, and the energy ladder is the same as in Table 1.

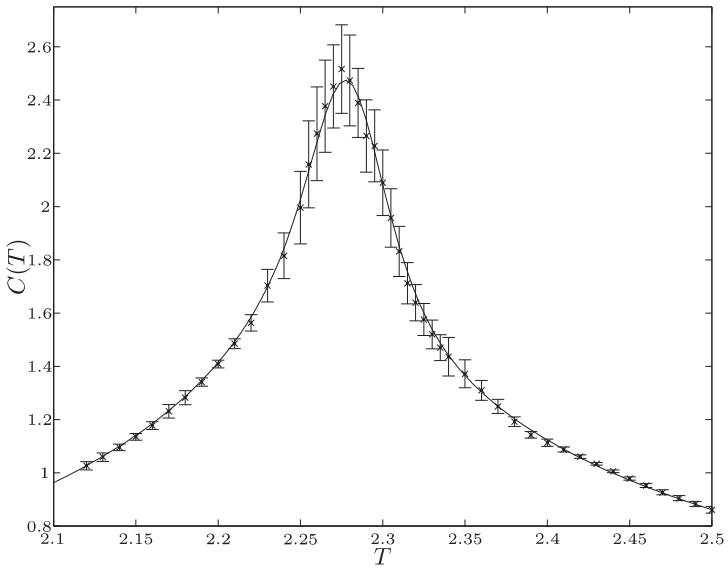


Figure 4. Temperature dependence of the specific heat around the critical temperature. The solid line is the exact curve for the $96 \times 96$ Ising lattice (Ferdinand and Fisher (1969)). Error bars represent the standard deviation obtained from ten independent runs.

## 5. Conclusion

In this paper, we presented a rigorous proof of the convergence and ergodicity of the EE sampler in the case of countable state spaces. We then applied it to the Ising model as a further test of its sampling efficiency. The simulation showed that the dynamic exponent of the EE sampler is significantly smaller than those of the MH algorithm and parallel tempering, indicating the EE sampler's high efficiency. An important open problem is to study the convergence rate of the EE sampler. Although many empirical studies support the EE sampler as a highly effective general purpose Monte Carlo algorithm, a rigorous theoretical investigation of its convergence rate is very desirable. Such results not only would provide the theoretical underpinning of the empirical observations, but might also lead to new methods for analyzing general non-Markov algorithms.

## Acknowledgement

The authors are grateful to Professor Richard Dudley and Professor Daniel Stroock for helpful comments. This work is supported in part by the NSF grant DMS-0449204 and the NIH/NIGMS grant R01GM090202-01.

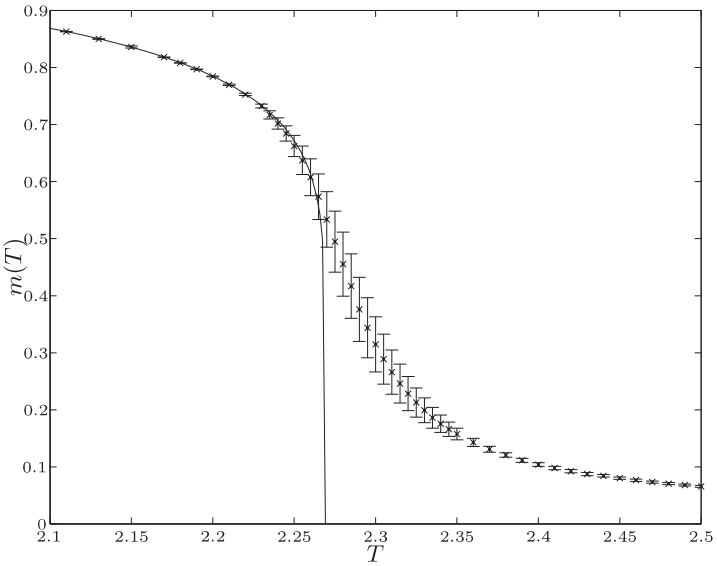


Figure 5. Temperature dependence of $E|m|$ around the critical temperature. The solid line is the exact curve for the infinite Ising lattice (Ferdinand and Fisher (1969)). Similar deviations from the exact curve are observed in Newman and Barkema (1999) and Landau (1976). Error bars represent the standard deviation obtained from ten independent runs.

## References

Andrieu, C., Jasra, A., Doucet, A. and Del Moral, P. (2008). A note on convergence of the equi-energy sampler. *Stochastic Analysis and Applications*, **26**, 298-312.

Coddington, P. D. and Baillie, C. F. (1992). Empirical relations between static and dynamic exponents for Ising model cluster algorithm. *Phys. Rev. Lett.* **68**, 962-965.

Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Prob.* **1**, 36-61

Ferdinand, A. E. and Fisher, M. (1969). Bounded and inhomogeneous Ising models. I. specific-heat anomaly of a finite lattice. *Phys. Rev.* **185**, 832-846.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the* 23*rd Symposium on the Interface*. (Edited by E. M. Keramigas), 156-163. Interface Foundation, Fairfax, VA.

Geyer, C. J. and Thompson, E. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Stat. Assoc.* **90**, 909-920.

Kou, S. C., Oh, J. and Wong, W. H. (2006). A study of density of states and ground states in hydrophobic-hydrophilic protein folding models by equi-energy sampling. *J. Chemical Physics* **124**, 244930.

Kou, S. C., Zhou, Q. and Wong, W. H. (2006). Equi-energy sampler with applications in statistical inference and statistical mechanics (with discussion). *Ann. Statist.* **34**, 1581-1652.

Landau, D. P. (1976). Finite-size behavior of the Ising square lattice. *Phys. Rev. B* **13**, 2997-3011.

Landers, D. and Rogers, L. (1985). Laws of large numbers for pairwise independent uniformly integrable random variables. *Math. Nachr.* **130**, 189-192.

Marinari, E. and Parisi, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhysics Lett.* **19**, 451-458.

Newman, M. E. J. and Barkema, G. T. (1999). *Monte Carlo Methods in Statistical Physics*. Oxford University Press, New York.

Nightingale, M. P. and Blote, H. W. J. (1996). Dynamic exponent of the two-dimensional Ising model and Monte Carlo computation of the subdominant eigenvalue of the stochastic matrix. *Phys. Rev. Lett.* **76**, 4548-4551.

Rosenthal, J. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Stat. Assoc.* **90**, 558-566.

Stroock, D. W. (2005). *An Introduction to Markov Processes*. Springer-Verlag, Berlin.

Swendsen, R. H. and Wang, J. S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58**, 86-88.

Sokal, A. D. (1989). Monte Carlo methods in statistical mechanics: foundations and new algorithms. Cours de Troisieme Cycle de la Physique en Suisse Romande 15, Lausanne.

Wei, W. W. S. (1990). *Time Series Analysis*. Addison-Wesley, Redwood City, California.

Wolff, U. (1989). Collective Monte Carlo updating for spin systems. *Phys. Rev. Lett.* **62**, 361-364.

Zhou, Q. and Wong, W. H. (2008). Reconstructing the energy landscape of a distribution from Monte Carlo samples. *Ann. Appl. Statist.* **2**, 1307-1331.

Zhou, Q. and Wong, W. H. (2009). Energy landscape of a spin-glass model: exploration and characterization. *Physical Rev. E* **79**, 051117.

Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Avenue Cambridge, MA 02139-4307, USA.

E-mail: xia@math.mit.edu

Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138-2901, USA.

E-mail: kou@stat.harvard.edu

(Received October 2009; accepted April 2010)