L. Ying *Res Math Sci(2025)12:28*
https://doi.org/10.1007/s40687-025-00520-y

Research in the Mathematical Sciences

RESEARCH

Multimodal sampling via approximate symmetries

Lexing Ying

*Correspondence:
lexing@stanford.edu
Department of Mathematics, Stanford University, Stanford, CA 94305, USA

Abstract

Sampling from multimodal distributions is a challenging task in scientific computing. When a distribution has an exact symmetry between the modes, direct jumps among them can accelerate the samplings significantly. However, the distributions from most applications do not have exact symmetries. This paper considers the distributions with approximate symmetries. We first construct an exactly symmetric reference distribution from the target one by averaging over the group orbit associated with the approximate symmetry. Next, we can apply the multilevel Monte Carlo methods by constructing a continuation path between the reference and target distributions. We discuss how to implement these steps with annealed importance sampling and tempered transitions. Compared with traditional multilevel methods, the proposed approach can be more effective since the reference and target distributions are much closer. Numerical results of the Ising models are presented to illustrate the efficiency of the proposed method.

Keywords: Multimodal distributions, Symmetries, Annealed importance sampling, Tempered transition, Ising models

Mathematics Subject Classification: 82B20, 82B80

1 Introduction

Sampling from multimodal distributions is challenging in scientific computing. Standard Markov Chain Monte Carlo (MCMC) methods face difficulties because they tend to mix efficiently locally but fail to introduce moves between different modes. Over the years, several classes of methods have been introduced to address this issue.

The first class is the multilevel or tempering methods, where the target distribution is associated with a fixed temperature. The high-temperature version of the target distribution is typically easy to sample. This class constructs a continuation path between the high-temperature version and the target distribution. The actual sampling is conducted via Metropolis–Hasting moves along the continuation path. Well-known examples of this type include simulated tempering [6,17], parallel tempering [5,10], tempered transitions [19], and annealed importance sampling [20]. The efficiency of these methods depends on the acceptance rate of the MCMC moves or the variance of the importance weights. To keep these under control, the continuation path sometimes needs to be long and discretized with many intermediate temperatures. This could result in high computational

© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2025.

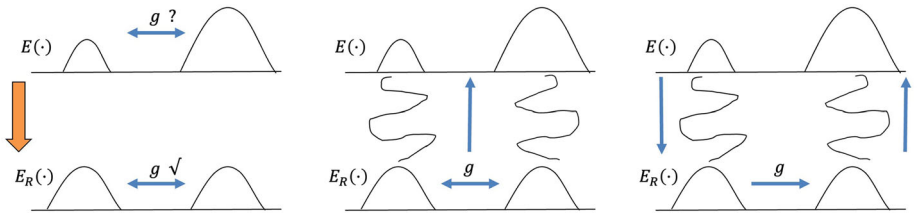

Fig. 1 Main steps of the algorithms. Left: G is the group of symmetries. G is an approximate symmetry for the distribution with log density $E(s)$. Through averaging over the group orbit, we construct a reference log density $E_R(s)$, which is symmetric under G. Middle: annealed importance sampling (AIS), where $g \in G$ allows efficient sampling from the reference $E_R(s)$. Right: tempered transition (TT), where $g \in G$ allows for global jumps within a TT move

and storage costs. In the rest of this paper, we will refer to these methods as continuation methods because the concept of connecting an easy-to-sample distribution with the target distribution is more general than tempering.

A second class of methods uses the symmetry of the target distribution. For many spin systems, an explicit symmetry exists between the spin values. One can accelerate the mixing by introducing a spin value flip over a large cluster of spins. This is, for example, used partly in Swendsen–Wang algorithm [23] and Wolff's method [27], where the global move flips between $+1$ and -1 configurations. Another related example is the Hamiltonian Monte Carlo (or hybrid Monte Carlo) [3,18], where the state space is augmented into a phase space by including momentum variables and the Hamiltonian flow acts as an ancillary symmetry that preserves the distribution over the phase space.

Though distributions with exact symmetry are important for analytical studies and theoretical understanding, many distributions from practical applications often fail to have exact symmetries. A natural question is whether it is still possible to leverage the power of symmetries there.

A key observation is that many high-dimensional multimodal distributions in computational physics can often be *approximately symmetric*. This is quite common at relatively low temperatures because one of the modes will otherwise dominate, effectively reducing the distribution to a single mode. For a distribution to be genuinely multimodal at low temperatures, there is often an approximate symmetry behind the scenes.

1.1 Contribution

This paper introduces a method for sampling from multimodal distributions with approximate symmetries. First, we identify the approximate symmetry. Second, we construct a symmetric reference multimodal distribution by averaging the log density over the group orbit associated with the approximate symmetry. Third, by applying the continuation method between the symmetric reference distribution and the target one, we obtain efficient algorithms for the target multimodal distribution. Figure 1 illustrates these main steps.

Compared with the existing continuation method that uses a high-temperature distribution as the reference, the proposed method can be more efficient. The reason is that the distance between the target distribution and our reference distribution constructed from group action averaging is often much shorter. This results in a few intermediate

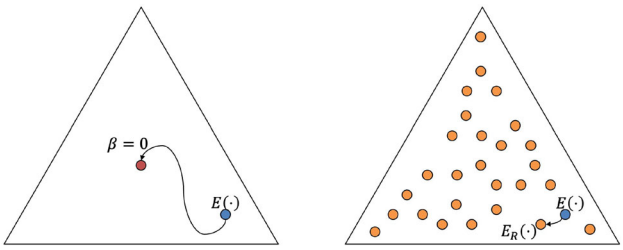

Fig. 2 Triangle stands for the probability simplex defined over the state space. Left: the usual continuation method. The usual reference distribution with $\beta = 0$ has a lot of symmetries and is easy to sample. However, it is often far from the target distribution $E(s)$, so the continuation path is long. Right: the proposed method. Based on the approximate symmetry G of the $E(s)$, we identify a nearby reference distribution $E_R(s)$ with G as the exact symmetry. The orange dots are the distributions that satisfy the exact symmetry in the group G ($\beta = 0$ is merely one of them). The continuation path from $E(s)$ to $E_R(s)$ is much shorter than the one from $E(s)$ to the $\beta = 0$ distribution

steps along the continuation path and less computational cost. Figure 2 offers a cartoon illustration of this point.

1.2 Related work

The field of Monte Carlo methods is vast. The standard references for advanced Monte Carlo strategies are [13,15]. In what follows, we only touch on the references closely related to our work.

In the class of the continuation (or tempering/annealing) methods, several well-known examples include simulated tempering [6,17], parallel tempering [5,10], simulated tempering [19], and annealed importance sampling [20].

Many methods take advantage of the exact symmetries and group actions. For example, within the clustering approach of spin systems, the Swendsen–Wang and Wolff algorithms [4,23,27] utilize the spin value symmetry in their cluster moves across all spatial scales. As we mentioned earlier, Hamiltonian or hybrid Monte Carlo [3,18] is another example that utilizes the symmetry of the Hamiltonian flow. The Hamiltonian Monte Carlo approach has also been extended to discrete state spaces in [2,12]. Group action has also been applied in [16] to improve Gibbs sampling.

Mode jumping (or mode hopping) is another class of methods for multimodal sampling [1,11,21,22,24,25]. Assuming that the modes are pre-identified, such a method constructs an easy-to-sample local approximation around each mode and uses their sum (typically a mixture of Gaussians) to approximate the overall target distribution. This approximation is then utilized in the Metropolis–Hasting step for sampling. Compared to the methods based on symmetries and approximate symmetries, this class of methods typically requires knowing more information about the modes. The numerical performance depends critically on the accuracy of the local approximation, which can be a non-trivial task.

Another important class is the population-based methods. Two well-known examples of this class are the Snooker algorithm [8] and the evolutionary Monte Carlo [14]. A recent development emphasizing affine invariance can be found in [9].

1.3 Content

The result of the paper is organized as follows. Section 2 describes the general framework and shows how to combine it with two continuation methods (annealed importance sampling and tempered transition). Section 3 focuses on two Ising-type models. Finally, Sect. 4 contains some discussions for future work.

2 Method

2.1 Approximate symmetry

We first describe how exact symmetry can help accelerate the mixing and then how to extend it to multimodal distributions with approximate symmetries.

Let S be state space, $p(s)$ be the target probability at s, and $E(s)$ be its log density (up to a normalization constant), i.e., $p(s) \propto \exp(E(s))$. We define symmetry via a group G that acts on the state space on the left. Following the standard notations in group theory (see [26] for example), we use gs to stand for the state obtained by acting g on s.

Definition 1 The distribution $p(x) \propto \exp(E(s))$ is symmetric with respect to G if

$$E(gs) = E(s) \tag{1}$$

for any $g \in G$ and $s \in S$.

This paper assumes that the symmetry can map from one mode to another. For simplicity, let us assume that G is a discrete group, though the Lie group case is similar. In the simplest case, $G = \{e, g\}$ with $g^2 = e$. Given G, one can define a Markov chain with the transition kernel $T(s, t)$ where

$$T(s,t) = \begin{cases} \frac{1}{|G|-1} & t = gs \text{ for some } g \neq e, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The claim is that T satisfies the detailed balance [15], i.e.,

$$p(s)T(s,t) = p(t)T(t,s).$$

To see this, notice that when t cannot be written as gs for any g, both sides are zero. When $t = s$, both sides are zero too. Finally, when $t = gs$ for some $g \neq e$, $p(s) \cdot \frac{1}{|G|-1} = p(t) \cdot \frac{1}{|G|-1}$ following (1).

The more interesting case is when the symmetry in G is only approximate. A distribution $p(x) \propto \exp(E(s))$ is *approximately symmetric* with respect to G if

$$E(gs) \approx E(s) \tag{3}$$

for any $g \in G$ and $s \in S$. The following assumption is important for the proposed approach.

Assumption 2 The action of G is transitive over the modes of $p(x) \propto \exp(E(x))$.

Definition 3 The *reference* log density $E_R(s)$ is defined by averaging over the group orbit

$$E_R(s) := \frac{1}{|G|} \sum_{g \in G} E(gs). \tag{4}$$

The corresponding reference distribution is $p_R(s) \propto \exp(E_R(s))$.

Three properties of $E_R(s)$ are important for the rest of the paper.

- $E_R(s)$ is symmetric under G, i.e.,

$$E_R(s) = E_R(gs).$$

This is because $E_R(s) = \frac{1}{|G|}\sum_h E(hs) = \frac{1}{|G|}\sum_h E(h \cdot gs) = E_R(gs)$, due to the orbit averaging.

- $E_R(s)$ and $E(s)$ are close:

$$E_R(s) \approx E(s)$$

and this follows from averaging (3).

- G also acts transitively on the modes of $E_R(s)$.

With these three properties, the reference log density $E_R(s)$ can help with the sampling from the multimodal distribution $p(s) \propto \exp E(s)$. The following two subsections describe how this can done in two continuation methods (annealed importance sampling and tempered distribution).

2.2 Annealed importance sampling

2.2.1 Existing method

In annealed importance sampling [20] (AIS), one has $E(s) = -\beta U(s)$ where β is the target inverse temperature and $U(s)$ is the potential. The method introduces the inverse temperature sequence:

$$0 \approx \beta_0 < \beta_1 < \ldots < \beta_{L-1} < \beta_L = \beta.$$

Each level l is associated with a distribution $p_l(s) \propto \exp(-\beta_l U(s))$ at inverse temperature β_l.

Since $\beta_0 \approx 0$, one assumes that $p_0(s) \propto \exp(-\beta_0 U(s))$ is easy to sample. We further assume that, at each β_l, $p_l(s)$ associated with a detailed balanced Markov Chain with a transition kernel $T_l(s,t)$, i.e.,

$$p_l(s)T_l(s,t) = p_l(t)T_l(t,s). \tag{5}$$

Given these building blocks, AIS computes independent weighted samples, one by one, from the following procedure.

(1) Sample a configuration $s_{1/2}$ from $p_0(s) \propto \exp(-\beta_0 U(s))$.
(2) For $l = 1, \ldots, L-1$, take one step (or a few steps) of $T_l(\cdot,\cdot)$ (associated with the distributions β_l) from $s_{l-1/2}$ to $s_{l+1/2}$.
(3) Set $s := s_{L-1/2}$.
(4) Compute the weight

$$w := \frac{p_1(s_{1/2})}{p_0(s_{1/2})} \cdots \frac{p_L(s_{1/2})}{p_{L-1}(s_{1/2})} \propto \frac{\exp(-\beta_1 U(s_{1/2}))}{\exp(-\beta_0 U(s_{1/2}))} \cdot \frac{\exp(-\beta_L U(s_{L-1/2}))}{\exp(-\beta_{L-1} U(s_{L-1/2}))}.$$

The normalization constant hidden in $\propto$ is the same for all independent samples from AIS and hence can be ignored.

Claim 4 *The configuration s with weight w samples the target distribution $p(s) \propto \exp E_L(s) = \exp E(s)$.*

Proof To see this, consider the path $(s_{1/2}, \ldots, s_{L-1/2})$. This path is generated with probability

$$p_0(s_{1/2})T_1(s_{1/2}, s_{3/2}) \cdots T_{L-1}(s_{L-3/2}, s_{L-1/2}).$$

Multiplying this with w and using the detailed balance (5) of T_l give

$$\begin{aligned} &p_0(s_{1/2})T_1(s_{1/2}, s_{3/2}) \cdots T_{L-1}(s_{L-3/2}, s_{L-1/2}) \cdot \frac{p_1(s_{1/2})}{p_0(s_{1/2})} \cdots \frac{p_L(s_{L-1/2})}{p_{L-1}(s_{L-1/2})} \\ =&p_L(s_{L-1/2})T_{L-1}(s_{L-1/2}, s_{L-3/2}) \cdots T_1(s_{3/2}, s_{1/2}), \end{aligned}$$

which is the probability of going backward, i.e., starting from a sample $s_{L-1/2}$ of $p_L(\cdot) \equiv p(\cdot)$. Taking the margin of the last slot $s_{L-1/2}$ proves that $s := s_{L-1/2}$ with weight w samples the distribution $p_L(s) \equiv p(s)$. □

2.2.2 Extension using symmetric reference

The proposed method does not change the β value. Instead, we continue between $E(s)$ and the symmetric reference $E_R(s)$.

Let $c(t) : [0, 1] \to [0, 1]$ be a strictly monotone function with $c(0) = 0$ and $c(1) = 1$. Define $E_l(s)$ for $0 \le l \le L$:

$$E_l(s) = (1 - c(l/L))E_R(s) + c(l/L)E(s).$$

In the simplest case, $c(t) = t$. This sequence goes from the reference distribution $E_0(s) = E_R(s)$ to the target one $E_L(s) = E(s)$. Since $E_R(s) \approx E(s)$, the discretization length L can be small.

Since $E_R(s)$ is symmetric and G acts transitively on its modes, it can be sampled efficiently. At each $E_l(s)$, we assume that $p_l(s) \propto \exp(E_l(s))$ is associated with a detailed balanced Markov Chain $T_l(s, t)$.

Given these building blocks, the proposed method computes independent weighted samples from the following procedure.

(1) Sample a configuration $s_{1/2}$ from $E_0(\cdot) = E_R(\cdot)$. This can be done efficiently since $E_R(\cdot)$ is symmetric.
(2) For $l = 1, \ldots, L-1$, take one step (or a few steps) of $T_l(\cdot, \cdot)$ (associated with the log density $E_l(\cdot)$) from $s_{l-1/2}$ to $s_{l+1/2}$.
(3) Set $s := s_{L-1/2}$.
(4) Compute the weight

$$w := \frac{p_1(s_{1/2})}{p_0(s_{1/2})} \cdots \frac{p_L(s_{1/2})}{p_{L-1}(s_{1/2})} \propto \frac{\exp(E_1(s_{1/2}))}{\exp(E_0(s_{1/2}))} \cdots \frac{\exp(E_L(s_{L-1/2}))}{\exp(E_{L-1}(s_{L-1/2}))}.$$

The justification for its correctness is the same as Claim 4.

2.3 Tempered transition

2.3.1 Existing method

In tempered transition [19] (TT), one introduces the following sequence of inverse temperatures:

$$\beta = \beta_0 > \ldots > \beta_L \approx 0 < \ldots < \beta_{2L} = \beta,$$

with $\beta_l = \beta_{2L-l}$ for $0 \le l \le L$.

Since $\beta_L \approx 0$, one assumes that $p_L(s) \propto \exp(-\beta_L U(s))$ can be mixed rapidly. At each β_l, $p_l(s) \propto \exp(-\beta_l U(s))$ is again associated with a detailed balanced Markov Chain with a kernel $T_l(s,t)$. In addition, $T_l(s,t) = T_{2L-l}(s,t)$.

The tempered transition (TT) combines the usual local moves at β with the following more sophisticated one.

(1) Start from a configuration s at β. Set $s_{1/2} = s$.
(2) For $l = 1, \ldots, L-1$, take one step (or a few steps) of $T_l(\cdot,\cdot)$ (associated with the distributions at β_l) from $s_{l-1/2}$ to $s_{l+1/2}$.
(3) At $l = L$, take one step (or a few steps) of $T_L(\cdot,\cdot)$ (associated with the distributions β_L) from $s_{L-1/2}$ to $s_{L+1/2}$. Since $\beta_L \approx 0$, this step can move efficiently between different modes.
(4) For $l = L+1, \ldots, 2L-1$, take one step (or a few steps) of $T_l(\cdot,\cdot)$ (associated with the distributions β_l) from $s_{l-1/2}$ to $s_{l+1/2}$.
(5) Set $t := s_{2L-1/2}$.
(6) Compute the amplitude

$$\begin{aligned} A &= \min\left(1, \frac{p_1(s_{1/2})}{p_0(s_{1/2})} \cdots \frac{p_{2L}(s_{2L-1/2})}{p_{2L-1}(s_{2L-1/2})}\right) \\ &= \min\left(1, \frac{\exp(-\beta_1 U(s_{1/2}))}{\exp(-\beta_0 U(s_{1/2}))} \cdots \frac{\exp(-\beta_{2L} U(s_{2L-1/2}))}{\exp(-\beta_{2L-1} U(s_{2L-1/2}))}\right), \end{aligned}$$

where the normalization constants cancel out.
(7) Sample α uniformly from $[0,1]$. If $\alpha < A$, set $s = t$. Otherwise, keep s unchanged.

Claim 5 *The tempered transition satisfies the detailed balance.*

Proof Let us consider the total transition between s and t. Each transition is implemented with a path $s_{1/2}, \ldots, s_{2L-1/2}$. Consider such a path one by one.The probability flux from s to t is given by

$$p_0(s_{1/2}) T_1(s_{1/2}, s_{3/2}) \cdots T_{2L-1}(s_{2L-3/2}, s_{2L-1/2}),$$

while the probability flux from t to s is given by

$$\begin{aligned} &p_{2L}(s_{2L-1/2}) T_1(s_{2L-1/2}, s_{2L-3/2}) \ldots T_{2L-1}(s_{1/2}, s_{3/2}) \\ =&p_{2L}(s_{2L-1/2}) T_{2L-1}(s_{2L-1/2}, s_{2L-3/2}) \ldots T_1(s_{1/2}, s_{3/2}), \end{aligned}$$

due to the symmetry $T_l = T_{2L-l}$. The ratio of the latter over the former is

$$\frac{p_{2L}(s_{2L-1/2}) T_{2L-1}(s_{2L-1/2}, s_{2L-3/2}) \ldots T_1(s_{1/2}, s_{3/2})}{p_0(s_{1/2}) T_1(s_{1/2}, s_{3/2}) \cdots T_{2L-1}(s_{2L-3/2}, s_{2L-1/2})} = \frac{p_1(s_{1/2})}{p_0(s_{1/2})} \cdots \frac{p_{2L}(s_{2L-1/2})}{p_{2L-1}(s_{2L-1/2})}.$$

Therefore, the Metropolis–Hasting move described above satisfies the detailed balance. □

2.3.2 Extension using symmetric reference

Instead of changing the value of β, we continue between $E(\cdot)$ and its symmetric reference $E_R(\cdot)$.

Let $c(t) : [0,1] \to [0,1]$ again be a strictly monotone function with $c(0) = 0$ and $c(1) = 1$. Define $E_l(s)$ for $0 \le l \le 2L$. For $0 \le l \le L$

$$E_l(s) = (1 - c(l/L))E(s) + c(l/L)E_R(s),$$

and for $L \le l \le 2L$

$$E_l(s) = E_{2L-l}(s).$$

In the simplest case, $c(t) = t$. This sequence starts from the target distributions $E_0(s) = E(s)$, goes to the reference one $E_L(s) = E_R(s)$, and returns to $E_{2L}(s) = E(s)$. Since $E_R(s) \approx E(s)$, the discretization length $2L$ is small.

Since $E_R(s)$ is symmetric under G, it enjoys the group transition (2) between multiple modes. At each $E_l(s)$, we assume that $p_l(s) \propto \exp(E_l(s))$ is associated with a detailed balanced Markov chain $T_l(s, t)$.

The proposed method combines the usual local moves at $E(\cdot)$ with the following more sophisticated move that takes advantage of the symmetric reference $E_R(s)$.

(1) Start from the current configuration s at $E(\cdot)$. Set $s_{1/2} = s$.
(2) For $l = 1, \ldots, L-1$, take one step (or a few steps) of $T_l(\cdot, \cdot)$ (associated with the distribution at E_l) from $s_{l-1/2}$ to $s_{l+1/2}$.
(3) At $l = L$, take a random $g \neq e$ uniformly from G. Let $s_{L+1/2} = g s_{L-1/2}$ be the resulting configuration (see (2)).
(4) For $l = L+1, \ldots, 2L-1$, take one step (or a few steps) of $T_l(\cdot, \cdot)$ (associated with the distribution at E_l) from $s_{l-1/2}$ to $s_{l+1/2}$.
(5) Set $t := s_{2L-1/2}$.
(6) Compute the amplitude

$$\begin{aligned} A &= \min\left(1, \frac{p_1(s_{1/2})}{p_0(s_{1/2})} \cdots \frac{p_{2L}(s_{2L-1/2})}{p_{2L-1}(s_{2L-1/2})}\right) \\ &= \min\left(1, \frac{\exp(E_1(s_{1/2}))}{\exp(E_0(s_{1/2}))} \cdots \frac{\exp(E_{2L}(s_{2L-1/2}))}{\exp(E_{2L-1}(s_{2L-1/2}))}\right). \end{aligned}$$

(7) Sample α uniformly from $[0, 1]$. If $\alpha < A$, set $s = t$. Otherwise, keep s unchanged.

The justification for its correctness is the same as Claim 5.

2.4 Comments

We have described the method for AIS and TT. The proposed approach works also for simulated tempering (ST) and parallel tempering (PT). Both ST and PT use the same continuation sequence as AIS. In principle, the approach can be combined with even the simple Metropolis–Hasting, with the transition (2) used as a proposal. However, this would require $E_R(s)$ to be very close to $E(s)$ because otherwise, this would not be effective.

The traditional tempering methods utilize the high-temperature distribution as the reference (or at least as the limiting one). It is trivially symmetric and easy to sample. However, this high-temperature reference can be quite far from the target distribution. The main observation of our approach is that any distributions with exact symmetries are easy to sample, and there are many more of them. For a target distribution $p(s)$ with approximate symmetry, our method identifies a nearby reference distribution $p_R(s)$ with exact symmetry. The distance between $p(s)$ and $p_R(s)$ is much closer to the one between $p(s)$ and the high-temperature distribution. Therefore, the continuation path can be much shorter.

The idea of using a reference distribution without temperature adjustment has been explored before. In [28], for the Ising model with boundary condition, the reference

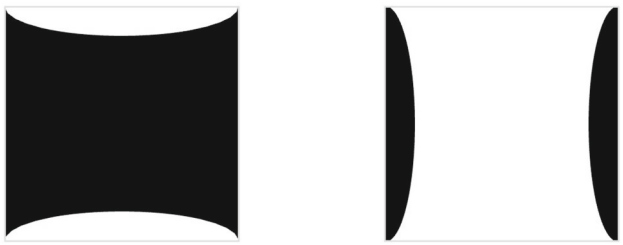

Fig. 3 Two dominant macroscopic profiles of the Ising model in Example 1. White stands for $+1$ while black for -1. The two dominant macroscopic profiles are a -1 cluster linking the left/right sides and a $+1$ cluster linking the top/bottom sides

distribution is set to be the one with the boundary condition turned off. Still, the distance between the target distribution and the reference one is quite significant.

3 Examples

3.1 Example 1

Consider the Ising model on a square lattice of size $n \times n$. Let I be the set of nodes and write $i \sim j$ when a bond links i and j. The state space consists of all $\{\pm 1\}^I$ configurations. β is a fixed inverse temperature. A forcing term is specified on the four sides as follows.

$$f_i = \begin{cases} -1 + Z_i & i \text{ on the left/right sides,} \\ +1 + Z_i & i \text{ on the top/bottom sides,} \\ 0 & \text{otherwise.} \end{cases}$$

where $Z_i \sim N(0, 1)$ are independent standard Gaussian.

Given a configuration $s = (s_i)_{i \in I}$, the log density $E(s)$ is defined via

$$\beta^{-1} E(s) = \sum_{i \sim j} s_i s_j + \sum_i f_i s_i, \quad \text{or} \quad \beta^{-1} E(s) = \frac{1}{2} s^T A s + f^T s,$$

where the matrix A is the connectivity matrix of the lattice, , i.e., $A_{ij} = 1$ iff $i \sim j$ and otherwise 0.

The forcing term tends to pin down the spins at the left/right sides to be -1 and the ones at the top/bottom sides to be $+1$. An essential feature of this Ising model is that the distribution exhibits macroscopically different profiles below the critical temperature. The two dominant macroscopic profiles are a -1 cluster linking the left/right sides and a $+1$ cluster linking the top/bottom sides, as shown in Fig. 3.

3.1.1 Identifying approximate symmetry

Due to the randomness in f, there is no exact symmetry. However, there is an approximate symmetry. Let $G = \{e, g\}$ be the group acting on the nodes I, where g acts on a node i by flipping along the 45-degree line. The key observation is that, for a boundary node $i \in I$,

$$f_{gi} \approx -f_i. \tag{6}$$

Based on this group action on the nodes, we lift the action to the configurations: for any configuration s,

$$(es)_i = s_i, \quad (gs)_{gi} = -s_i.$$

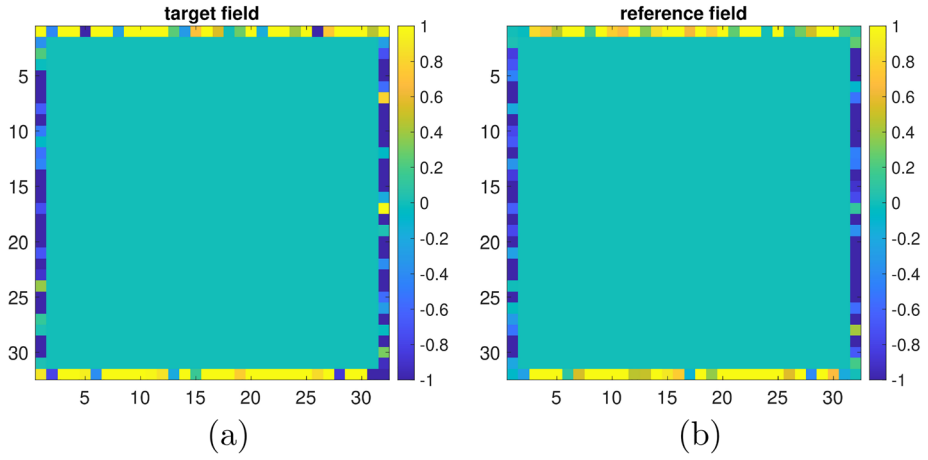

Fig. 4 Example 1. **a** The forcing term f of $E(s)$. **b** The forcing term $\frac{1}{2}(f - Pf)$ of $E_R(s)$

The group element e does not change the configuration as expected. The element g flips the lattice nodes along the 45-degree line and then the spin value. The minus sign is essential since it would otherwise not match the forcing term applied on the boundary nodes. We refer to the action of g as the *double-flip*, following [29].

We claim that $E(gs) \approx E(s)$. To show this,

$$\begin{aligned}\beta^{-1}E(gs) &= \sum_{i\sim j}(gs)_i(gs)_j + \sum_i f_i(gs)_i = \sum_{gi\sim gj}(gs)_{gi}(gs)_{gj} + \sum_i f_{gi}(gs)_{gi} \\ &\approx \sum_{i\sim j}(-s_i)(-s_j) + \sum_i(-f_i)(-s_i) = \beta^{-1}E(s),\end{aligned}$$

where the approximation uses (6).

Based on that, the reference distribution is defined via

$$\beta^{-1}E_R(s) = \frac{1}{2}(E(s) + E(gs)) = \sum_{i\sim j} s_i s_j + \frac{1}{2}\sum_i (f_i - f_{gi})s_i,$$

or in terms of the connectivity matrix A

$$\beta^{-1}E_R(s) = \frac{1}{2}s^T As + \frac{1}{2}(f - Pf)^T s.$$

Here, P is the permutation matrix associated with the action of g on the nodes. Figure 4 gives the forcing terms: f of $E(s)$ and $\frac{1}{2}(f - Pf)$ of $E_R(s)$ for $n = 32$.

In what follows, we present the results with $n = 32$ and $\beta = 0.8$. Before showing the results for AIS and TT, we first perform a simple MCMC with Glauber dynamics interspersed with proposals from approximate symmetry. Figure 5 plots the average spin value across a total of 10,000 MCMC moves, interspersed with 100 proposals. Notice that there is only one single accepted proposal (i.e., a jump in the average spin value), demonstrating that this simple MCMC is inefficient in this case.

3.1.2 AIS with symmetric reference

Applying the method described in Sect. 2.2.2 to this example requires specifying the sampling of $E_R(\cdot)$ and the transitions $T_l(\cdot,\cdot)$. Both use the Glauber dynamics done in parallel: we partition the nodes into a collection of independent subsets. These subsets

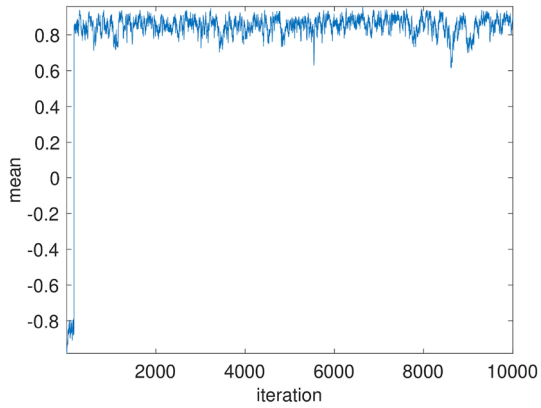

Fig. 5 Average spin value along a simple MCMC simulation with Glauber dynamics interspersed with proposals from approximate symmetry. There is only one single accepted proposal between the $+1$ and -1 dominant configurations

are visited in a random order. The updates for all nodes in a single independent subset are performed in parallel. In what follows, we will refer to this as a *Glauber sweep* over the nodes.

First, in order to obtain one sample from the symmetric $E_R(\cdot)$, we choose an initial configuration uniform randomly and then perform a few Glauber sweeps of $E_R(\cdot)$. The uniform random initial configuration ensures that the two modes of $E_R(\cdot)$ are visited equally, while the Glauber dynamics mix rapidly in each mode [7]. Second, the transition $T_l(\cdot,\cdot)$ at each level l is also implemented with a single Glauber sweep of $E_l(\cdot)$.

The linear function $c(t) = t$ is used to interpolate between $E_R(\cdot)$ and $E(\cdot)$, with $L = 64$. A total number of $K = 10{,}000$ independent weighted samples are collected, with each denoted as $(s^{(k)}, w^{(k)})$ for $1 \le k \le K$.

To monitor the variance of the algorithm, we record the weight history within each level l:

$$w_l^{(k)} := \frac{p_1(s_{1/2}^{(k)})}{p_0(s_{1/2}^{(k)})} \cdot \cdots \cdot \frac{p_l(s_{l-1/2}^{(k)})}{p_{l-1}(s_{l-1/2}^{(k)})}$$

for $l = 1, \ldots, L$. These weights are then normalized at each level l

$$w_l^{(k)} \Leftarrow w_l^{(k)} / \left(K^{-1} \sum_{g=1}^{K} w_l^{(g)} \right). \tag{7}$$

The sampling efficiency, a quantity between 0 and 1, is defined as $(1 + \mathrm{Var}(\{w_L^{(k)}\}_{k=1}^K))^{-1}$, following [20].

Figure 6 plots the logarithms of the normalized weights of each sample as a function of the level l. The sampling efficiency is 0.086, which translates to 400 Glauber sweeps per independent sample. The probabilities at the $+1$ and -1-dominant configurations are estimated to be 95% vs. 5%, respectively.

3.1.3 TT with symmetric reference

The method described in Sect. 2.3.2 is embedded in an MCMC run. Among the moves, 99% are the (local) parallel Glauber sweeps of $E(s)$, while the other 1% are TT moves described

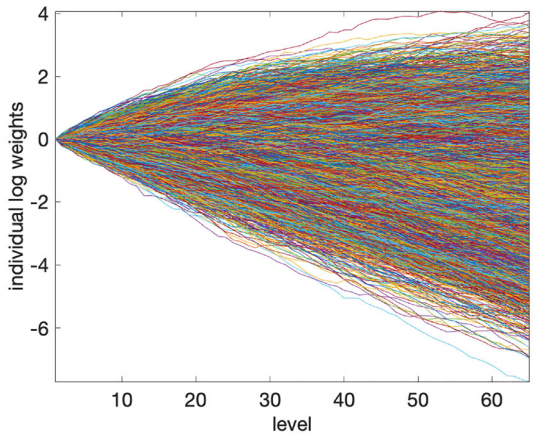

Fig. 6 Logarithms of the normalized weights (7) of each sample over all levels

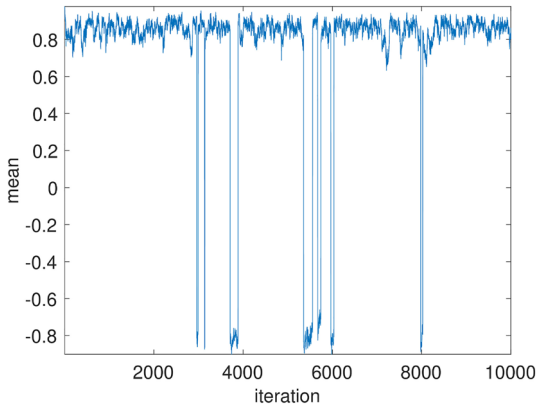

Fig. 7 Average spin value along the MCMC simulation. Each jump stands for a successful TT move between the $+1$ and -1 dominant configurations

in Sect. 2.3.2. The percentage is chosen to more or less balance the work between the two parts.

For the TT move, we adopt again the linear interpolation from $E(s)$ to $E_R(s)$ and back, with $L = 64$. The transition $T_l(\cdot, \cdot)$ at each level l is also implemented with a single Glauber sweep of $E_l(s)$. A total of 10,000 MCMC moves are performed, and about 100 TT moves are attempted among them.

In order to monitor the performance, we record the average spin value $(\sum_{i\in I} s_i)/|I|$. Figure 7 plots this average as a function of time. There are 15 successful transitions between the two modes, and the probabilities at $+1$ and -1-dominant configurations are estimated to be 96% vs. 4%, respectively.

3.2 Example 2

Consider the Ising model on a rectangular lattice of size $n_1 \times n_2$ with $n_1 \neq n_2$. A forcing term is again specified on the four sides:

$$f_i = \begin{cases} -1 + c, & i \text{ on the left/right sides,} \\ +1 + c, & i \text{ on the top/bottom sides,} \\ 0, & \text{otherwise,} \end{cases}$$

where the constant c is chosen so that f_i has zero mean. Ensuring $\sum_i f_i = 0$ helps to make both $+1$-dominant and -1-dominant profiles have non-negligible probabilities.

Given a configuration $s = (s_i)_{i \in I}$, the log probability $E(s)$ is defined via

$$\beta^{-1} E(s) = \sum_{i \sim j} s_i s_j + \sum_i f_i s_i, \quad \text{or} \quad \beta^{-1} E(s) = \frac{1}{2} s^T A s + f^T s.$$

Similar to Example 1, the distribution with the specified forcing term exhibits macroscopically different profiles below the critical temperature.

3.2.1 *Identifying approximate symmetry*

Since $n_1 \neq n_2$, there is no exact symmetry. However, there is an approximate symmetry. The group is still $G = \{e, g\}$. We start by identifying the action of g on the nodes. We embed the nodes I into the $[-1, 1]^2$ box, with stepsize $\frac{2}{n_1 - 1}$ in the horizontal direction and $\frac{2}{n_2 - 1}$ in the vertical one. For each i, denote $\vec{x}_i$ as the associated geometric location and correspondingly define $\vec{y}_i$ as

$$\vec{y}_i = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \vec{x}_i.$$

The goal is to find a map $g : I \to I$ such that two conditions: $g^2 = e$ and

$$\vec{y}_i \approx \vec{x}_{gi}. \tag{8}$$

The motivations are twofold. First, such g would ensure $f_{gi} \approx -f_i$. Second, if $i \sim j$, then $\vec{x}_i \approx \vec{x}_j$ and $\vec{y}_i \approx \vec{y}_j$. (8) implies that $\vec{x}_{gi} \approx \vec{x}_{gj}$, which further correlates with $gi \sim gj$, i.e., the nodes gi and gj are linked in the Ising model. We adopt the following heuristic procedure proposed in [29] to achieve these two conditions.

(1) Order the interior vertices $\{\vec{x}_j\}_{j \in I}$ based on their distances to the origin in the decreasing order. The distance is typically chosen to be either the ℓ_∞ norm or the ℓ_2 norm.
(2) Mark all vertices $j \in I$ as unpaired.
(3) Scan the interior vertices in this ordered list. For each $\vec{x}_j$, if j is already paired, then skip. If not, find the unpaired i such that $\vec{y}_i$ is closet to $\vec{x}_j$, pair i and j

$$gi := j, \quad gj := i,$$

and mark both i and j as paired.

The heuristic is that, by following the order of decreasing distance to the origin, the remaining unpaired vertices are forced to cluster near the center of the domain, thus reducing the distance between them.

We then lift the action of g from the nodes to the configuration s. Given s, we define gs by

$$(es)_i = s_i, \quad (gs)_{(gi)} = -s_i.$$

The minus sign here is again essential for g to be an approximate symmetry.

The claim is $E(gs) \approx E(s)$. To see this,

$$\begin{aligned}\beta^{-1}E(gs) &= \sum_{i\sim j}(gs)_i(gs)_j + \sum_i f_i(gs)_i \approx \sum_{gi\sim gj}(gs)_{gi}(gs)_{gj} + \sum_{gi} f_{gi}(gs)_{gi} \\ &\approx \sum_{i\sim j}(-s_i)(-s_j) + \sum_i(-f_i)(-s_i) = \beta^{-1}E(s),\end{aligned}$$

where the first approximation uses the fact that $i \sim j$ is correlated with $gi \sim gj$ and the second approximation uses $f_{gi} \approx -f_i$.

Based on that, the reference distribution $E_R(s)$ is defined as

$$\beta^{-1}E_R(s) = \frac{1}{2}\left(\sum_{ij} s_i s_j + \sum_i f_i s_i\right) + \frac{1}{2}\left(\sum_{ij}(gs)_i(gs)_j + \sum_i f_i(gs)_i\right).$$

or in terms of the connectivity matrix A

$$\beta^{-1}E_R(s) = \frac{1}{2}s^T(A + P^T AP)s + \frac{1}{2}(f - Pf)^T s.$$

Here, P is again the permutation matrix associated with the action of g on the nodes. Figure 8 plots the g action between the nodes, the connectivity of $E_R(s)$, and the forcing terms of $E(s)$ and $E_R(s)$, for $n_1 = 30$, $n_2 = 32$. Notice that in this example, $P^T AP$ differs from A; therefore, the connectivity of $E_R(s)$ has more bonds.

In what follows, we present the results for $n_1 = 30$, $n_2 = 32$, and $\beta = 0.8$. Before showing the results for AIS and TT, we perform again a simple MCMC with Glauber dynamics interspersed with proposals from approximate symmetry. Figure 9 summarizes the average spin over 10,000 MCMC moves, interspersed with 100 proposals. Notice that there is only one accepted proposal, showing that this simple MCMC is inefficient.

3.2.2 AIS with symmetric reference

We adopt a linear interpolation between $E_R(\cdot)$ and $E(\cdot)$, with $L = 64$. The choice of the initial configuration and the transition $T_l(\cdot,\cdot)$ are the same as Sect. 3.1.2. A total of 10,000 samples are collected in this test.

Figure 10 gives the logarithm of the normalized weights for each sample across all levels. The sampling efficiency is 0.49, which translates to 129 Glauber sweeps per independent sample. The probabilities at $+1$ and -1-dominant configurations are estimated to be 98% vs. 2%. This difference between the probability masses of the two modes originates from the slight discrepancy between n_1 and n_2 but is enhanced by the relatively large β value. About half of the samples are devoted to the 98% $+1$-dominant configurations and the other half to the 2% -1-dominant ones. Given prior knowledge about this imbalance, this inefficiency can be improved by lowering the probability but increasing the weights for the initial samples obtained from $E_R(s)$.

3.2.3 TT with symmetric reference

The method described in Sect. 2.3.2 is embedded in an MCMC run. Among the moves, 99% are the parallel Glauber sweeps of $E(s)$, while the remaining 1% are TT moves.

We use the linear interpolation for the TT move from $E(\cdot)$ to $E_R(\cdot)$ and back, with $L = 128$. The transition $T_l(\cdot,\cdot)$ at each level l is implemented also with a Glauber sweep of $E_l(\cdot)$. A total of 10,000 moves are carried out, with about 100 TT moves being attempted.

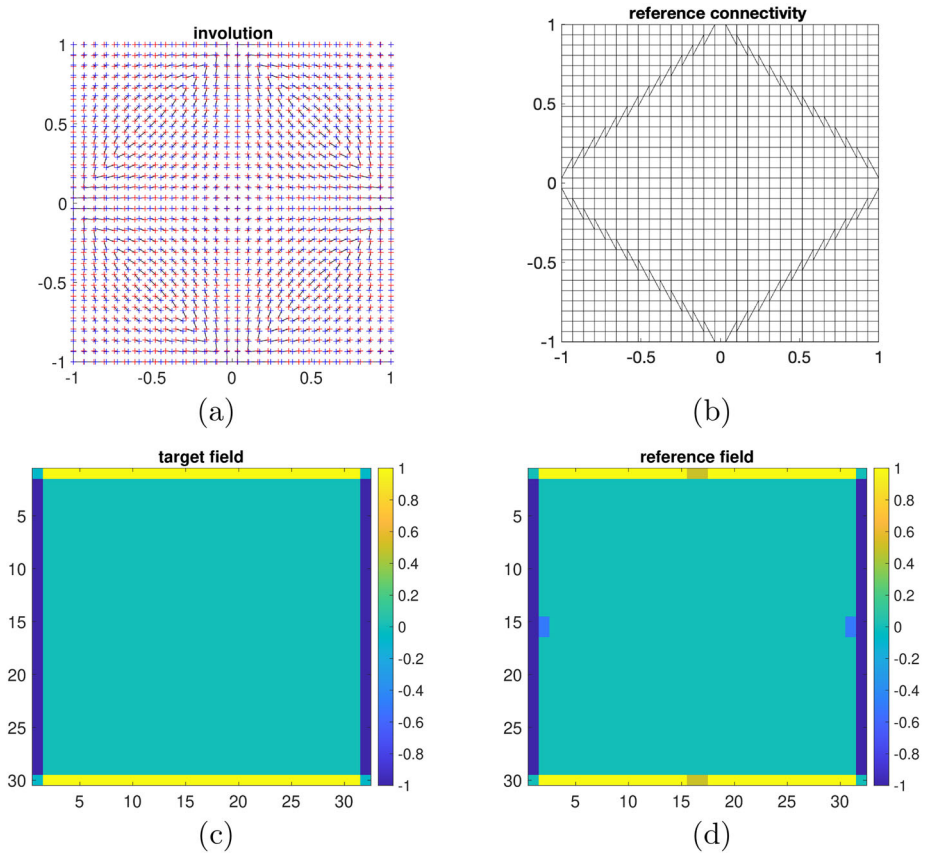

Fig. 8 Example 2. **a** The map $g : I \to I$. The points $\{\vec{x}_i\}$ and $\{\vec{y}_i\}$ are marked in blue and red, respectively. A black segment is linked between $\vec{y}_i$ and $\vec{x}_j$ if and only if $gj = i$ and $gi = j$. **b** The connectivity pattern of $\frac{1}{2}(A + P^T AP)$. Notice that it contains a few more edges than the connectivity pattern of A due to the averaging over the group orbit. **c** The forcing term f of $E(s)$. **d** The forcing term $\frac{1}{2}(f - Pf)$ of $E_R(s)$

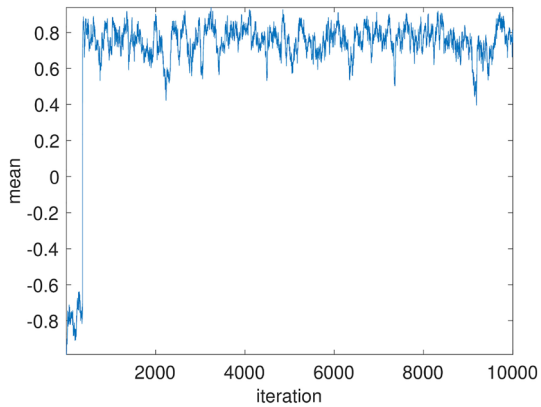

Fig. 9 Average spin value along a simple MCMC simulation with Glauber dynamics interspersed with proposals from approximate symmetry. There is only a single jump between the +1 and −1 dominant configurations

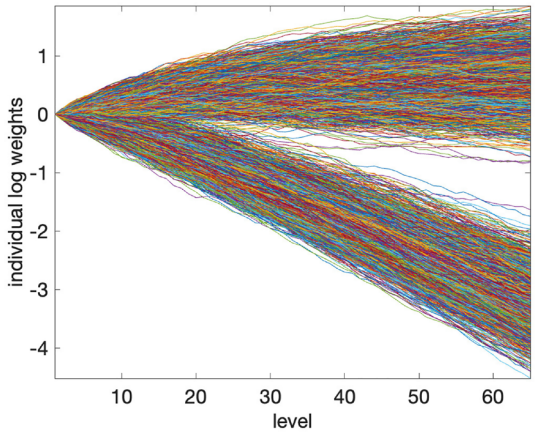

Fig. 10 Logarithms of the normalized weights (7) of each sample over all levels

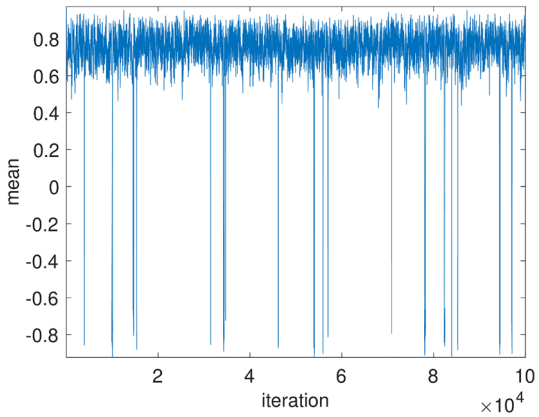

Fig. 11 Average spin value along the MCMC simulation. Each jump stands for a successful transition between the $+1$ and -1 dominant configurations

Figure 11 plots the average spin value as a function of time. There are 36 successful transitions between the modes, and the probabilities at $+1$ and -1-dominant configurations are estimated to be 98% vs. 2%.

4 Discussion

This paper proposes a new approach for sampling multimodal distributions. The main requirement is identifying the approximate symmetry of the target distribution. Once the approximate symmetry is identified, we construct a nearby symmetric reference distribution and combine it with two continuation methods, AIS and TT. This framework can also be combined with simulated tempering (ST) and parallel tempering (PT).

The examples considered are all simple spin systems, though with non-trivial boundary conditions. In principle, the method can be applied to distributions over continuous state spaces equipped with approximate symmetries. When the group is discrete, the proposed method should require relatively little change. On the other hand, when the group is a Lie group, the evaluation $E_R(s)$ would average over a continuous group action with respect to the Haar measure, which can itself be a non-trivial task. In the long run, the hope is that

this framework can contribute to sampling multimodal molecular configurations under binding or external constraints.

Acknowledgements
The author thanks professors Persi Diaconis and Wing Hung Wong for suggestions and comments.

Data availability
Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Received: 4 January 2024 Accepted: 29 March 2025
Published online: 16 April 2025

References

1. Andricioaei, I., Straub, J.E., Voter, A.F.: Smart darting Monte Carlo. J. Chem. Phys. **114**(16), 6994–7000 (2001)
2. Diaconis, P., Holmes, S., Neal, R.M.: Analysis of a nonreversible Markov chain sampler. Ann. Appl. Probab. **10**, 726–752 (2000)
3. Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D.: Hybrid Monte Carlo. Phys. Lett. B **195**(2), 216–222 (1987)
4. Edwards, R.G., Sokal, A.D.: Generalization of the Fortuin–Kasteleyn–Swendsen–Wang representation and Monte Carlo algorithm. Phys. Rev. D **38**(6), 2009 (1988)
5. Geyer, C.J.: Markov chain Monte Carlo maximum likelihood (1991)
6. Geyer, C.J., Thompson, E.A.: Annealing Markov chain Monte Carlo with applications to ancestral inference. J. Am. Stat. Assoc. **90**(431), 909–920 (1995)
7. Gheissari, R., Sinclair, A.: Low-temperature Ising dynamics with random initializations. In: Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing, pp. 1445–1458 (2022)
8. Gilks, W.R., Roberts, G.O., George, E.I.: Adaptive direction sampling. J. R. Stat. Soc. Ser. D (The Statistician) **43**(1), 179–189 (1994)
9. Goodman, J., Weare, J.: Ensemble samplers with affine invariance. Commun. Appl. Math. Comput. Sci. **5**(1), 65–80 (2010)
10. Hukushima, K., Nemoto, K.: Exchange Monte Carlo method and application to spin glass simulations. J. Phys. Soc. Jpn. **65**(6), 1604–1608 (1996)
11. Ibrahim, A.I.N.: New methods for mode jumping in Markov chain Monte Carlo algorithms. Ph.D. thesis (2009)
12. Kapfer, S.C., Krauth, W.: Irreversible local Markov chains with rapid convergence towards equilibrium. Phys. Rev. Lett. **119**(24), 240603 (2017)
13. Liang, F., Liu, C., Carroll, R.: Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples, vol. 714. Wiley, London (2011)
14. Liang, F., Wong, W.H.: Evolutionary Monte Carlo for protein folding simulations. J. Chem. Phys. **115**(7), 3374–3380 (2001)
15. Liu, J.S.: Monte Carlo Strategies in Scientific Computing, vol. 10. Springer, Berlin (2001)
16. Liu, J.S., Sabatti, C.: Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. Biometrika **87**(2), 353–369 (2000)
17. Marinari, E., Parisi, G.: Simulated tempering: a new Monte Carlo scheme. Europhys. Lett. **19**(6), 451 (1992)
18. Neal, R.M: Monte Carlo implementation. In: Bayesian Learning for Neural Networks, pp. 55-98 (1996)
19. Neal, R.M.: Sampling from multimodal distributions using tempered transitions. Stat. Comput. **6**, 353–366 (1996)
20. Neal, R.M.: Annealed importance sampling. Stat. Comput. **11**(2), 125–139 (2001)
21. Sminchisescu, C., Welling, M.: Generalized darting Monte Carlo. In: Artificial Intelligence and Statistics, pp. 516–523 (2007)
22. Sminchisescu, C., Welling, M., Hinton, G.: A mode-hopping MCMC sampler. Technical Report CSRG-478, University of Toronto, submitted to Machine (2003)
23. Swendsen, R.H., Wang, J.-S.: Nonuniversal critical dynamics in Monte Carlo simulations. Phys. Rev. Lett. **58**(2), 86 (1987)
24. Tjelmeland, H., Eidsvik, J.: On the use of local optimizations within Metropolis–Hastings updates. J. R. Stat. Soc. Ser. B Stat. Methodol. **66**(2), 411–427 (2004)
25. Tjelmeland, H., Hegstad, B.K.: Mode jumping proposals in MCMC. Scand. J. Stat. **28**(1), 205–223 (2001)
26. Vinberg, E.B.: A Course in Algebra. American Mathematical Society, Providence (2003)
27. Wolff, U.: Collective Monte Carlo updating for spin systems. Phys. Rev. Lett. **62**(4), 361 (1989)
28. Ying, L.: Annealed importance sampling for Ising models with mixed boundary conditions. J. Comput. Math. **41**(3), 542–550 (2023)
29. Ying, L.: Double flip move for Ising models with mixed boundary conditions. J. Comput. Math. **41**(5), 1003–1016 (2023)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.